


Understanding Museum Exhibits using Vision-Language Reasoning

Ada-Astrid Balauca^{1,*},  Sanjana Garai^{1,3,*} Stefan Balauca¹ Rasesh Udayakumar Shetty³
 Naitik Agrawal³ Dhwanil Subhashbhai Shah³ Yuqian Fu¹ Xi Wang^{1,2}
 Kristina Toutanova^{1,4} Danda Pani Paudel¹ Luc Van Gool¹

¹INSAIT, Sofia University “St. Kliment Ohridski”, Bulgaria ²ETH Zürich, Switzerland

³Indian Institute of Technology, Varanasi (IIT BHU) ⁴Google DeepMind

Abstract

Museums serve as repositories of cultural heritage and historical artifacts from diverse epochs, civilizations, and regions, preserving well-documented collections that encapsulate vast knowledge, which, when systematically structured into large-scale datasets, can train specialized models. Visitors engage with exhibits through curiosity and questions, making expert domain-specific models essential for interactive query resolution and gaining historical insights. Understanding exhibits from images requires analyzing visual features and linking them to historical knowledge to derive meaningful correlations. We facilitate such reasoning by (a) collecting and curating a large-scale dataset of 65M images and 200M question-answer pairs for exhibits from all around the world; (b) training large vision-language models (VLMs) on the collected dataset; (c) benchmarking their ability on five visual question answering tasks, specifically designed to reflect real-world inquiries and challenges observed in museum settings. The complete dataset is labeled by museum experts, ensuring the quality and the practical significance of the labels. We train two VLMs from different categories: BLIP [41] with vision-language aligned embeddings, but lacking the expressive power of large language models, and the LLaVA [46] model, a powerful instruction-tuned LLM enriched with vision-language reasoning capabilities. Through extensive experiments, we find that while both model types effectively answer visually grounded questions, large vision-language models excel in queries requiring deeper historical context and reasoning. We further demonstrate the necessity of fine-tuning models on large-scale domain-specific datasets by showing that our fine-tuned models significantly outperform current SOTA VLMs in answering questions related to specific attributes, highlighting their limitations in handling complex, nuanced queries. Our dataset, benchmarks, and source code are available at: insait-institute/Museum-65.

1. Introduction

We release a high-quality, large-scale dataset and demonstrate through experiments that training large VLMs on it enables museum artifact understanding, for visually understanding exhibit images through visual question answering. VLMs like CLIP [64], Gemini [75], and LLaVA [46] have demonstrated strong capabilities in learning from large-scale noisy image-text data, improving visual understanding through natural language and bridging the gap between textual annotations and images [14, 45, 55, 63, 65, 66, 80, 81, 85]. However, these models [41, 46] struggle in domains like museums, which require detailed, interdisciplinary knowledge and structured attribute prediction, such as age, origin, material, and cultural relevance [7, 54, 60]. While pre-trained VLMs are effective for tasks like object detection [6, 30, 91] and semantic segmentation [20, 40, 88], more complex multi-modal tasks demand advanced reasoning across visual and textual domains [34, 59, 66, 77].

Visual Question Answering (VQA) is a key multi-modal task explored in works like [3, 7, 9, 52, 68, 70, 92]. In the cultural heritage domain, VQA can enhance museum engagement, but a large-scale dataset covering diverse artifacts with both visual and textual data is lacking. Existing datasets primarily focus on art [67, 72, 82] and are often used for image generation and style transfer [19, 27, 38, 65], failing to capture deeper exhibit-context relationships.

In this work, we collect a novel large-scale **multilingual dataset** MUSEUM-65 with high-quality images and extensive textual information for a wide range of museum artifacts, totaling 65M images and 200M question-answer pairs. We curate and use it to fine-tune VLMs, BLIP and LLaVA, to enable a better understanding of museum exhibits. The textual information of MUSEUM-65 reflects the viewpoint of knowledgeable museum experts, providing both depth and breadth for effective AI training. We further design 5 real-world tasks: general VQA, category-wise VQA, MultiAngle – questions using images from different

*Equal Contribution  astrid.mokanu@gmail.com

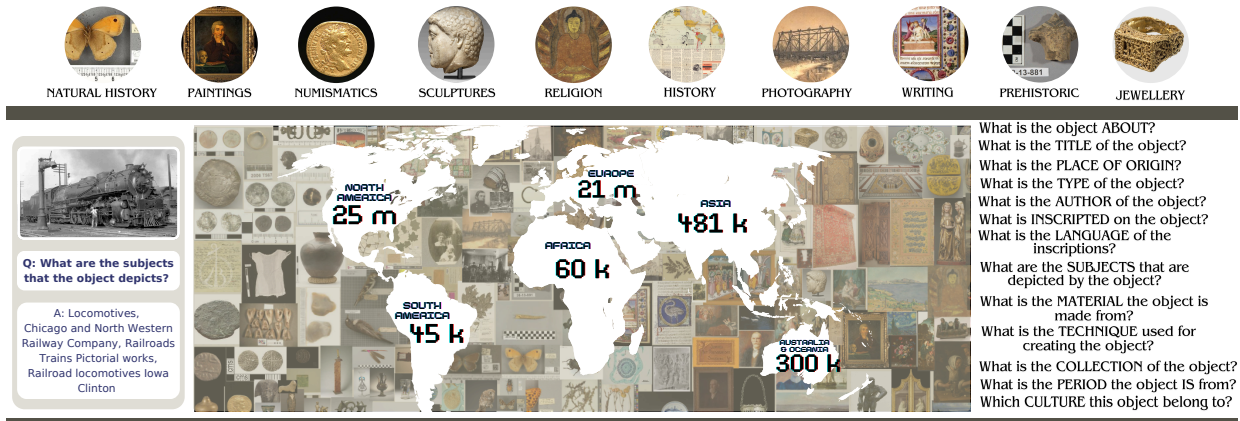


Figure 1. **Dataset composition.** MUSEUM-65 covers a wide range of exhibit categories (top), e.g arts, historical/pre-historical, natural sciences, and contains a **large number of images from around the globe**. Each image is paired with multiple questions exploring subjects like Title, Creator, Period, Techniques, Culture, Inscriptions, etc. (right). A sample image with a question and answer is shown on the left.

viewpoints, Visually Unanswerable Questions – complex questions requiring the use of general knowledge, and MultiLanguage – questions in languages other than the English. We also conduct an ablation study on place of origin to assess potential regional biases. We also show in Tab. 1 that the existing vision-language models perform poorly, using questions for place of origin and object title.

Model Name	Zero-Shot	Attribute	
		Title	Place
GPT-4o	✓	22.03	33.33
Claude-3-7-sonnet	✓	21.89	40.43
Llama-3.2-90b-vision	✓	16.84	29.58
Gemini 1.5B flash	✓	27.08	32.98
LLaVa nofinetune	✓	10.13	23.42
LLaVa-ours (20mn 1ep)	✗	57.00	70.00
BLIP nofinetune	✓	3.00	5.00
BLIP-ours (20mn 5ep)	✗	52.00	61.00

Table 1. **Zero-Shot SOTA vs. our Fine-Tuned MUSEUM-65 Models.** The results demonstrate that fine-tuning significantly improves accuracy over zero-shot SOTA models.

Our contribution aims to facilitate the development of AI models that can handle complex cross-disciplinary questions in a truthful and comprehensive manner, enabling museums to serve as dynamic educational platforms that enrich visitor experience and deepen understanding across diverse cultural, historical, and scientific domains, as we show by fine-tuning BLIP [41] and LLaVA.[46] BLIP aligns images with descriptive text effectively, generating accurate captions that enhance its question-answering capabilities. Still, BLIP’s smaller text encoder/decoder (*BERT-base*, 110M params.) limits its ability to handle complex instructions. LLaVA, powered by the larger *Llama-7B* LLM, excels in instruction comprehension and vision-language reasoning, making it capable of performing complex tasks. We provide insights into the nuanced and detailed understanding and real-world applications required for museum exhibits, presenting comparisons of the two models on multiple met-

Dataset	Domain	#images	#questions	Public
Sheng et al. [69]	Archaeology	160	800	✗
AQUA [26]	Art	21K	80K	✓
iMet [86]	Art, History	155K	155K	✓
VISCOUNTH [7]	Art	500K	6.5M	✗
MUZE [5]	Art, History	210K	1.5M	✓
MUSEUM-65 (ours)	Art, History, Nat. Sciences	65M	200M	✓

Table 2. **Literature comparison.** MUSEUM-65 v.s. related datasets from literature based on data domains, size and structure.

rics. We show both can handle questions well when answers can be directly derived from visual features. However, for questions requiring the integration of visual features with broader human knowledge, large VLMs attain higher accuracy, performing the reasoning needed for such inquiries. For instance, they can answer questions that link visual details to historical facts or explain connections to related events or figures not directly depicted. The major contributions of the paper are:

- **Dataset and fine-tuned models:** We introduce a dataset of 65M images and 200M question-answer pairs for museum exhibits suitable to build new vision-language models and to fine-tune existing ones (e.g. BLIP, LLaVA)
- **Benchmark:** We propose 5 tasks derived from our dataset, setting directions for research in real-world AI for cultural heritage, along with the metrics to evaluate them.
- **Results and insights:** We offer several insights about the collected dataset as well as the real-world tasks proposed.

2. Related Work

Vision language pre-training models and VQA. Models like CLIP [64], BLIP [41] and LLaVA [46], pre-trained on large-scale datasets, have shown remarkable versatility in both unimodal and multimodal tasks [12, 13, 32, 36, 42, 43, 47, 49, 93], incl. zero-shot recognition [87, 89, 90], image segmentation [20, 39, 88], object detection [6, 30, 91], etc. They offer a broad understanding of general concepts and

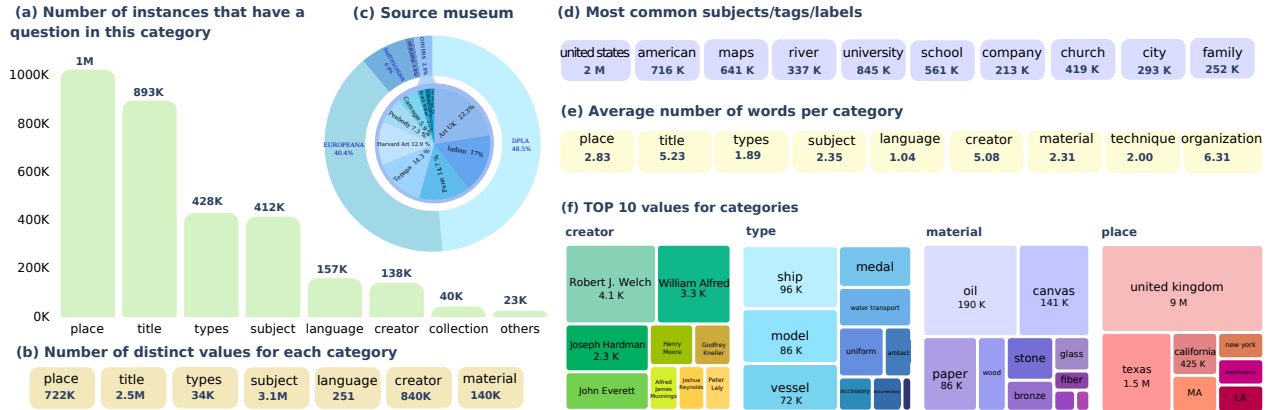


Figure 2. **Dataset statistics.** (a) distribution of questions, categorized by *type*: the most common question is about the objects’ *place of origin*, (b) number of distinct values of each category: the most varied category is *subject* (c) data sources of each contributing museum, (d) the most common subjects/tags associated with the exhibits: objects coming from historical museums, like maps, items related to the United States, or personal themes, (e) average number of words (length) of each category value: *organization* has the most words and (f) the most frequent values across different question categories: the objects’ *types* include ships, models, vessels, medals, and pieces of art.

can become valuable for specialized fields like cultural heritage and museums. Previous studies on VQA have largely focused on images or videos, some works extending VQA by integrating external general knowledge [53, 79, 83] or knowledge tailored to specific datasets [24, 78].

Digital humanities and cultural heritage. In cultural heritage, achieving qualitative supremacy in visual understanding requires both informative images and reliable textual information. However, the required expertise is a major challenge in data collection [15, 26, 50, 69, 76]. Multiple approaches for art understanding exist, including tasks such as cross-modal retrieval [2], image captioning [4, 48, 67], classifying [11, 56, 58, 74] or recognizing [17, 35] artworks. Previous attempts leverage existing cultural heritage data, approaching it from a multi-modal perspective [4, 7, 22, 31, 48, 73] but usually without using VLMs. MUZE [5] achieves strong results on fill-in-the-gaps tasks by leveraging CLIP’s multi-modal representations. However, its design relies on separate attention heads for individual attributes, making it both computationally expensive and challenging to scale for a large, diverse dataset like ours. Moreover, it does not align well with the direct Q&A needs of our dataset, limiting its applicability to our tasks.

Domain-Specific datasets. General-purpose datasets [18, 44] are vast but lack domain-specific capabilities for cultural artifacts and scientific exhibits. For history and natural sciences [57, 71], datasets are scarce and often rely on external knowledge bases. In the Art domain, multiple datasets [67, 72, 82] exist but mainly focus on artistic images with limited text and others [1, 8, 10, 16, 23, 25, 28, 33, 51] combine visual and textual data but are either small, lack diversity, or rely on synthetic sources. VISCOUNT [7] has 500K images and 6.5M questions only covering paintings and sculptures, while MUZE [5] has 210K images and 1.5M texts in art and history (see

Tab. 2). Our dataset of 65M images and 200M questions strikes a balance between scale and domain-specificity. It offers both the diversity and depth needed for a more comprehensive exploration of art, history and natural sciences VQA tasks, including data from museums used by previously mentioned works. We perform a benchmark comparison, evaluating the performance of our best BLIP model against BLIP trained on the MUZE dataset, showing that our dataset offers superior utility and effectiveness over existing alternatives with the experiment results being highlighted in section 4.1 of the Supp. Mat.

3. Dataset

We built MUSEUM-65, a multi-modal dataset containing 65M images and 200M question-answer pairs in multiple languages, ensuring cultural diversity, see Fig. 1.

3.1. Data Collection

MUSEUM-65 covers 50M objects with questions in English and 15M with questions in **37 languages from Europe and Asia** (*French, Spanish, German, etc*). List of languages in Supp. Mat. MUSEUM-65 is created by scraping museum websites of 3 prime international aggregators (DPLA, Europeana, Smithsonian), covering museums from Europe and North America and 12 other individual museums (see Supp. Mat.) spread over the other continents. Some museums consist multiple images of the same object from different angles. We collected the web urls of all the images. We show more details about the data origin in Fig. 2. We will make the dataset publicly available under the same license museums use, CC BY-NC-4.0.

3.2. Data Curation

A total of 10 experts worked over 3 months, 2 experts cross-checked for quality to clean and curate the entire data. Tabular representation in the form of attribute-value pairs

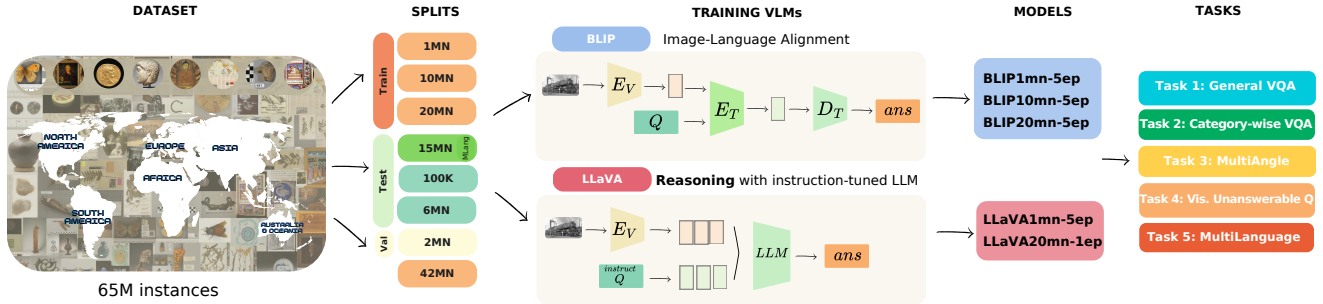


Figure 3. **Workflow.** Using smaller subsets of the dataset (1mn, 10mn and 20mn), we fine-tune BLIP and LLaVA models. **BLIP**, an encoder-decoder based model, **aligns language and image** in the same space while **LLaVA**, built on an instruction-tuned LLM is **directly reasons** based on the language.

is the usual format for museum exhibit information. Each museum has a unique set of attributes. After extracting the attributes, we reformulate them as questions and associated values become answers. For detailed method, see Supp mat section 2.5. The overview of the method is as below.

Separating into attribute-value pairs. Information about exhibits takes 2 forms: (a) attribute-value pairs, scraped using museums APIs; (b) single strings, otherwise. We determine separators to obtain the attribute-value pairs when object information is retrieved as a complete string.

Filtering attributes. The object attributes also include *display site in museum*, *catalog number*, *inventory date*, *dimensions*, and more. These are redundant for our goals and we excluded them from the main dataset. The remaining attributes were again divided into 2 types: (a) medium length attributes (with a length less than 100 words) (b) long length attributes, the rest. The reason is the restriction to 512 input tokens for BLIP. Despite LLaVA allowing for more input tokens, the final dataset on which our models have been trained was limited to the *medium attributes*, thus ensuring a fair comparison of BLIP vs. LLaVA. When referring to our dataset in terms of training, validation or testing, we refer to the one with *medium attributes* only. However, we will make the complete dataset along with the filtered and long-length attributes publicly available as the raw version.

Creating questions from attributes. We structure attribute data for visual question answering separately for each museum, adapting to their format differences. Questions are manually crafted (63 unique questions, listed in Supp. Mat) with attribute’s value serving as answer. Humans formulated the questions to ensure diversity, having slightly varied questions for the same attributes across different museums, mimicking natural human curiosity to phrase questions in varied ways. For example, for the attribute *material*, two varied questions were: *Which primary material is the object made of?* vs. *What is the material used in the object?*

Creating the final dataset. We download all 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). Each question’s answer is a list, since multiple answers may apply. See Supp. Mat. for an example.

3.3. Data Statistics and Bias Analysis

We analyzed the dataset by examining question distribution, category diversity, sources, common subjects, word counts per category, and frequent question types (See Fig. 2).

While bias-free datasets are unattainable [21], we ensure **our dataset is bias-aware**. 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. Our dataset includes 5M+ objects from other continents. Nevertheless, results clearly show that finetuning on MUSEUM-65 causes benefits to distribute evenly despite regional biases (see Tab. 8). 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, and more. To mitigate **language bias**, we include 15M samples across 37 languages beyond English, with ongoing expansions. 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.

To help researchers analyze and address biases, we will release MUSEUM-65 with tools for large-scale dataset exploration. These tools will enable image retrieval via text or image queries, aiding systematic bias detection and mitigation. By making the source code and essential routines publicly available, we aim to support customized dataset curation while fostering transparency and inclusivity. Additionally, we encourage users to explore the dataset and, in the future, report undetected biases and model behaviors through a planned public portal, improving data curation and solidifying MUSEUM-65 as a real-world dataset.

For applications requiring a minimally biased dataset, debiasing techniques such as model-agnostic training or specialized architectures will be commended [29, 61, 84].

3.4. Societal Impact of Dataset

Our dataset supports training multimodal models that enhance cultural accessibility, educational tools, and virtual heritage exploration, while promoting multilingual data and cross-cultural appreciation by enabling global artifact comparison. Inspection of images and text reveals that museums, as reputable institutions, curate collections to address controversies—such as historical disputes, privacy, religious issues, and racial bias—and tag inappropriate content, ensuring dataset safety and quality. While origin bias remains a concern, we aim to mitigate it through collaborations and diversification, hoping broader museum digitization will further enhance diversity. In the current form, we consider this dataset a research artifact and strongly advocate **academic use only** and advise careful investigation of downstream model biases (further analysis in Supp. Mat.).

3.5. Data Splits

We split the data (English) in train, val and test, having 42M, 2M and 6M images, with an average of 3.5 questions per image (15M instances in other languages are in a separate test split). We create multiple smaller train subsets of 1M, 10M, 20M, and a smaller subset of the test dataset, with 10K instances, which we use during experiments and evaluation. The stratification is done to meet different computational needs. For more details about the splits, as well as the data format and examples, see Supp. Mat.

4. Evaluation

We compute two types of metrics: (1) traditional uni-gram and n-gram-based numeric metrics that rely on lexical overlaps, and (2) deeper semantic-based metrics that leverage word embeddings for a more nuanced evaluation.

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.

4.1. Numeric metrics

We compute the commonly used **precision, recall, and BLEU scores**. To simplify evaluation, we introduce Complete Precision, which is the percentage of questions where the answer fully matches the ground truth (precision = 1.0). Similarly, Partial Precision is the percentage where there is at least some overlap (precision > 0.0). Complete Recall and Partial Recall are defined analogously. The BLEU score [62] measures the fraction of word n-grams in the model’s prediction that appear in at least one valid answer, with a brevity penalty to discourage short responses. We scale BLEU scores between 0 and 100, reporting results for

BLEU1 (1-gram) and BLEU2 (2-gram). For detailed explanation of metrics, refer Supp. Mat.

4.2. Semantic metrics

Some attributes like subject and short description, where textual variations in answers are equally valid, make numeric metrics insufficient. Semantic metrics offer a deeper evaluation of the model’s domain understanding by capturing contextual meaning rather than relying solely on exact word matching. Results (Tab. 3) show that fine-tuning on MUSEUM-65 significantly improves these metrics.

METEOR Score. METEOR aligns words using synonyms, stemming, and paraphrasing, making it more robust than pure n-gram overlap metrics. The final score considers precision, recall, and a fragmentation penalty to account for word order. We scale the score between 0 and 100 and average it across all instances.

Word Mover’s Distance Score. We also report Word Mover’s Distance (WMD) based top-1 accuracy [37], which measures the minimum cumulative distance required to transform the predicted response into the ground truth in the Word2Vec embedding space. The most probable class is determined based on the smallest WMD score and accuracy of determining the ground truth class is calculated.

Model	BLIP	BLIP	LLaVA	LLaVA
	nofinetune	20mn 5e	nofinetune	20mn 1e
METEOR	3.24	37.45	2.96	58.85
WMD Acc.	35.54	74.02	54.5	87.02

Table 3. **Semantic Evaluation results.** Results demonstrate a substantial enhancement in domain understanding after fine-tuning

5. Experiments

We introduce a comprehensive benchmark for MUSEUM-65, evaluating general and specific tasks across different metrics by exploring multiple VQA-based tasks, including general question VQA, category-wise VQA, and three more challenging tasks designed to address real-world problems relevant to Museum LLMs. We also perform an ablation experiment on the place of origin to check if our models are biased to giving more accurate results to objects that belong from a specific region. This benchmark standardizes comparisons across methods, guiding future research toward effective models and identifying areas for improvement. See Fig. 3 for an overview.

5.1. Experimental Setup

In our experiments we use two models known for VQA tasks, LLaVA [46] and BLIP [41] using our dataset. We train multiple model configurations with varying amounts of data and training epochs to analyze the impact of training time and data size on the results. We evaluate their performance using multiple scores (precision, recall, BLEU), and discuss their behavior. For further details and why we choose BLIP and LLaVA models, as well as our code and dataset please refer to Supp. Mat.

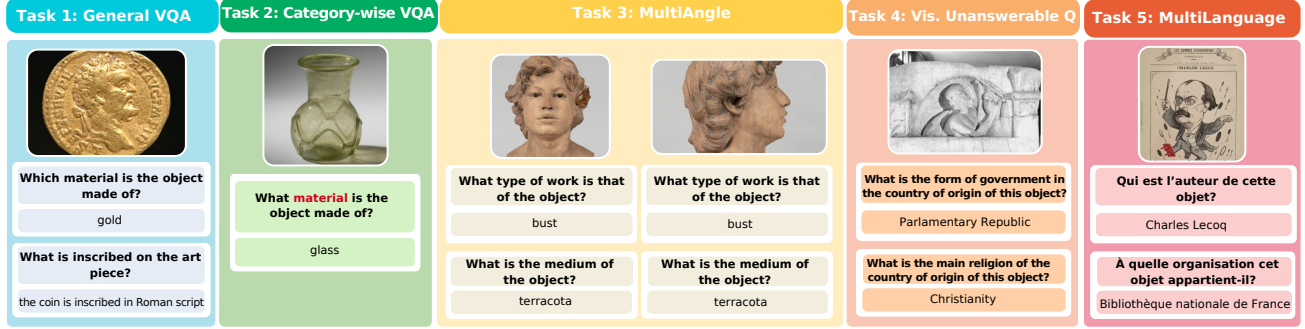


Figure 4. **Benchmarked tasks.** (1) **general VQA**, (2) **category-wise VQA**, (3) **MultiAngle** - measures the adaptability to different angle images of the same object, (4) **Visually Unanswerable Questions** - observes the response to new common knowledge questions derived from dataset’s available information for an exhibit, (5) **MultiLanguage** - checks the ability to use languages like French and German

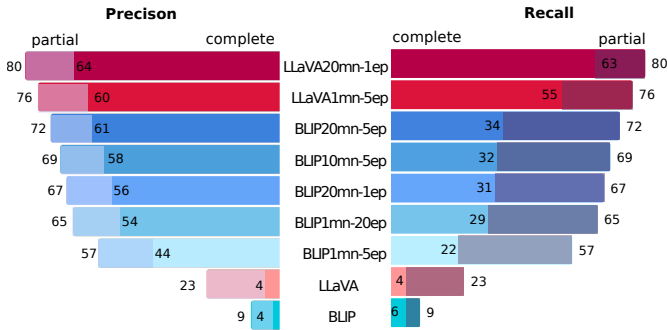


Figure 5. **General VQA results.** Comparison of fine-tuned and non-fine-tuned models on precision and recall. Models fine-tuned with the 20mn dataset perform best, with **LLaVA20mn-1ep** achieving 80% partial precision and 64% complete precision. **LLaVA models also outperform BLIP in recall**, indicating their predictions more often contain or are contained in the ground truth.

Training on our dataset. We fine-tune LLaVA and BLIP using the same image-question pairs, choosing for every image one random question-answer pair every epoch. In each case, the front view image of an object is used.

Finetuning BLIP. In our experiments we use BLIP, with the configuration available as *blip-vqa*. We fine-tune three main versions of BLIP, using: (a) 1mn train dataset for 5 epochs, extended up to 20 epochs (independently fine-tuned), (b) 10mn train dataset, 5 epochs, (c) 20mn train dataset, 5 epochs referring to them as BLIP1mn-5ep, BLIP10mn-5ep, and BLIP20mn-5ep respectively. We also fine-tune a 20mn train dataset version for exactly 1 epoch to have a fairer comparison for LLaVA20mn-1ep. During fine-tuning we use a batch size of 512, mainly following the fine-tuning scheme of [41]. More details in Supp. Mat.

Finetuning LLaVA. For finetuning LLaVA, we assure the use of the same object-question pairs and the same order as for BLIP experiments. We fine-tune two versions of LLaVA, (a) using 1mn train dataset for 5 epochs, and (b) using the 20mn dataset for 1 epoch. We will refer to them as LLaVA1mn-5ep and LLaVA20mn-1ep. We use a batch

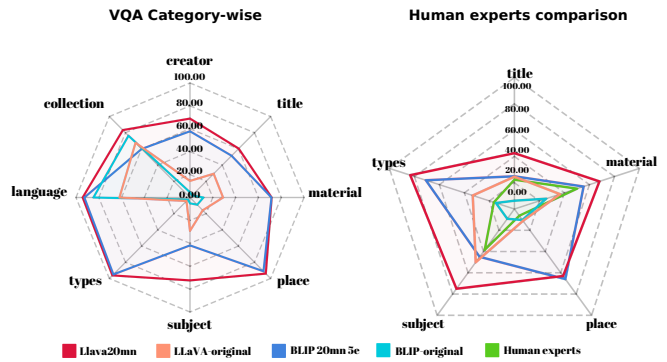


Figure 6. **VQA category-wise results.** Comparison of models with human experts (right) and fine-tuned vs. original models (left). **Fine-tuned models outperform in all categories, LLaVA20mn performing best.** Fine-tuned models exceed human performance. The originals excel in *language* and *collection* due to common knowledge answers and fewer related instances.

size of 512. We evaluate all models on the VQA tasks.

Hardware. We train and evaluate our models using 64×NVIDIA H100 GPUs.

5.2. Task 1: VQA on general questions

The task involves using all the questions associated with each image and producing the individual scores described in Sec. 4. We compute the average score over all image-question pairs for each metric, to observe the model’s general VQA capability and adaptability across a diverse range of visual and linguistic contexts, providing the performance on any kind of question addressed by the user.

While evaluating the fine-tuned LLaVA and BLIP on all the questions we observe that the LLaVA models are always receiving better results than their BLIP counterpart (See Fig. 5). LLaVA20mn trained 1 epoch receives the best results having for 80% of the predictions at least a part in common with the ground truth, and 63% perfect match (prediction and ground truth are equal). We observe that the LLaVA models (fine-tuned 1mn or 20mn, and original LLaVA) have usually a close result between precision and

Model	Angle	Partial Prec.	Complete Prec.	Partial Recall	Complete Recall	BLEU1
LLaVA20mn-1ep	Original	58.09	46.09	58.12	41.04	42.14
	Alternative	56.14	44.89	56.15	40.01	41.02
LLaVA no finetune	Original	24.35	0.09	24.35	11.25	1.61
	Alternative	23.56	0.02	23.56	10.85	1.54
BLIP-	Original	52.78	42.51	52.78	35.29	38.31
	Alternative	51.75	41.87	51.75	34.59	37.62
BLIP no finetune	Original	13.82	9.70	13.82	5.22	6.52
	Alternative	12.86	8.71	12.86	4.72	5.92

Table 4. **MultiAngle results.** Fine-tuned LLaVA20mn-1ep and BLIP20mn-5ep v.s. no fine-tune models. **The models are stable w.r.t. the viewpoint changes.** (Please refer to Fig. 7-3rd col.).

recall, while the BLIP models (fine-tuned and original) have a big decrease in complete recall (the ground truth is completely present in the prediction).

5.3. Task 2: VQA category-wise

Questions are grouped by attributes (eg: *title*, *creator*, *technique*, *subjects/labels*). For each category, relevant questions are compiled (e.g asking about the title, denomination, or object name collected under *title*). The model then answers each question, generating individual scores, which are then aggregated to compute an average score for each category, allowing for a detailed analysis of the model’s strengths and weaknesses across different categories, revealing areas where it may excel or struggle. All questions attributed to one category, in Supp. Mat.

For this experiment, we compare the partial precision. We see in Fig. 6 that LLaVA and BLIP original have very low results for most of the categories. We notice LLaVA fine-tuned having significantly better results than BLIP fine-tuned on *subject* and *collection*. The lowest result for all models are for *title*, which is also very difficult for humans.

Human experts evaluation on VQA category-wise task.

We randomly selected 850 question-answer pairs covering different attributes and conducted an experiment with 10 museum experts, who answered the same questions as our models. Their responses were evaluated across the categories *types*, *title*, *place*, *material*, and *subject* using the same methodology as for the models. The results (see Fig. 6) reveal that certain categories, such as *place* and *types*, are particularly challenging for humans. Notably, fine-tuned models outperform human experts across all categories, especially in *subjects*, *place*, and *types*, highlighting the need for specialized models with domain-specific knowledge. For *materials*, performance is comparable, as these can be determined by simply observing the object.

IAA metrics: We approximate Fleiss’s Kappa by simulating categorical behavior for free-form answers. Five experts independently answered a set of 63 unique questions (from test set). The “best” response was chosen via majority voting, and agreement was measured as proportion of the remaining four matching it, yielding **52.7%** agreement—well above the **6.25%** = 0.5^4 expected by chance.

5.4. Task 3: Multi Angles

To assess the model’s resilience to viewpoint changes, we evaluate it using images captured from different angles or perspectives, available in our dataset. By substituting these viewpoint-varied images for the originals, we can directly compare these scores with those from the initial baseline images to observe any shifts in accuracy or relevance.

For this task we select a subset of $\approx 5K$ exhibits from the test dataset with multiple images taken from different angles (e.g. 2nd column of Fig. 7). In total we evaluate on $\approx 22K$ questions. All our models (Tab. 4) show consistent scores when presented with images from different angles, suggesting a strong capacity for generalization and an ability to recognize objects despite variations in angle or orientation, providing insights into the model’s ability to maintain performance stability when faced with real-world variability in image capture. The slight performance drop can be attributed to a decrease in image information (e.g. pictures of statues from the side are generally harder to recognize).

5.5. Task 4: Visually Unanswerable Questions

We introduce a set of specialized questions to assess the model’s contextual understanding, focusing on an object’s country of origin or creator. These carefully designed questions require a deeper level of contextual or associative reasoning. For example, instead of simply asking about characteristics that may be linked with a visual pattern (assuming that the painters’ style can be visually recognized - “*Who is the painter of this painting?*”), these questions may ask, “*Who was the mentor of the painter of this painting?*” or “*What is the nationality of the painter of this painting?*”.

We manually generate 5-6 questions for exhibits, related either to the creator or country and search for answers online (e.g. 3rd column of Fig. 7). We obtain 510 and 515 image-question-answer pairs from the train and test dataset respectively. This approach evaluates not only whether the model can correctly identify or infer the country of origin or creator based on visual cues but also tests its ability to correlate these features with general knowledge or cultural information, addressing beyond surface-level visual details. The full list of questions is available in Supp. Mat. According to results in Tab. 5, both original and fine-tuned LLaVA have much higher reasoning capabilities than BLIP, due to LLaVA’s higher model size and larger pre-training dataset. Moreover, fine-tuning LLaVA enhances its ability to reason about museum exhibits, esp. when considering the precision of its answers. On the other hand, BLIP’s performance on this complex task drops after fine-tuning, hinting at BLIP’s limited model capacity causing forgetting of prior knowledge in order to accommodate the new training data. The consistency of results across the test dataset further supports LLaVA20mn-1ep’s ability to reason beyond visual features even on unseen images (see Tab. 6).

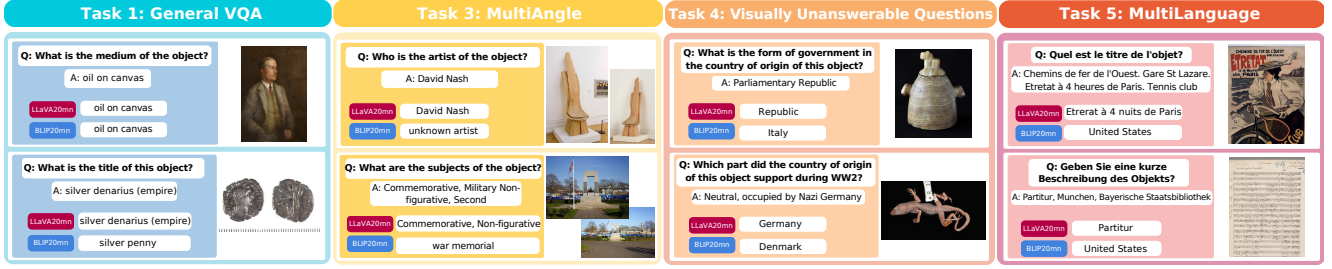


Figure 7. **Examples.** LLaVA20mn-1ep and BLIP20mn-5ep behaviour on different tasks, General VQA (1st column), MultiAngle (2nd column), Visually Unanswerable Questions (3rd column) and MultiLanguage (4th column). We observe **more precise answers for LLaVA20mn** than for BLIP20mn on all the tasks. Also the **last two tasks seem to be impossible for BLIP20mn**.

Model	partial prec.	complete prec.	partial recall	complete recall	BLEU1
LLaVA20mn-1ep	31.37	25.1	31.37	12.94	15.16
LLaVA no finetune	24.27	0.58	24.27	6.21	1.74
BLIP 20mn-5ep	2.35	0.2	2.35	0.2	0.63
BLIP no finetune	6.08	5.69	6.08	2.75	2.95

Table 5. **Visually Unanswerable Questions results on train images.** Clearly, **LLaVA20mn-1ep**, performs best, especially for complete precision and complete recall, showing the **ability** to visually link the objects with the corresponding dataset information and **to respond to visually unanswerable questions**.

Model	partial prec.	complete prec.	partial recall	complete recall	BLEU1
LLaVA 20mn-1ep	29.7	25.83	29.7	10.29	12.67
LLaVA no finetune	27.18	1.55	27.18	6.41	3.08
BLIP 20mn-5ep	3.3	0.78	3.3	0.19	0.73
BLIP no finetune	5.44	5.24	5.44	2.33	2.58

Table 6. **Visually Unanswerable Questions results on test images.** Results are consistent with those of the train split indicating the capability of the model to generalise well on unseen images.

5.6. Task 5: Multiple Languages

We evaluate the model’s zero-shot performance on non-English questions, including *French*, *German*, *Spanish*, and others available in the *multilanguage* dataset split. Questions are formulated in the respective languages using collected attributes. This assesses the model’s ability to link visual content with multilingual queries, recognizing objects, actions, or scenes without relying on English training biases, which is crucial for real-world multilingual use. Lastly, we evaluate our models on 500 images with textual data in French and German, for a total of 2864 question-answer pairs (e.g. in Fig. 7 - 4th column). In Tab. 7 we can observe that both variants of LLaVA achieve better results than BLIP. However, our fine-tuned LLaVA seems to have partially forgot its abilities to answer in foreign languages due to it being only fine-tuned with english data. Although the original LLaVA easily answers questions in different languages (it has high partial precision and recall), it mostly fails to give perfect answers. Further fine-tuning the models using multilingual data from MUSEUM-65 should improve their performance. A small-scale dataset was curated for Tasks 4 & 5 to ensure quality, given the significant manual effort. We plan to continue scaling this curation.

Model	French		German		Average	
	partial prec.	complete prec.	partial prec.	complete prec.	BLEU1	BLEU2
LLaVA20mn-1ep	10.37	0.54	9.72	1.17	1.36	0.27
LLaVA no finetune	41.81	0.4	18.41	0.15	1.46	0.13
BLIP20mn-5ep	4.02	0.4	0.73	0.15	0.21	0.01
BLIP no finetune	2.01	0.6	0.8	0.29	0.13	0

Table 7. **Multi-Language results.** LLaVA models perform better than BLIP ones. LLaVA20mn-1ep **slightly forgets 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 slightly better than the no fine-tune versions.

5.7. Place of Origin Ablation

We curated 1K images per continent and evaluated our best models on it in Tab. 8. Despite the bias in place of origin, the benefits distribute evenly.

Model	Europe	N. America	S. America	Asia	Africa	Oceania
LLaVA 20mn-1ep	85.2	79.6	86.6	67.4	86.7	99.2
LLaVA	8.6	43.57	20.3	23.4	20.79	52.4
BLIP-20mn-5ep	79.1	73.1	76.4	65.5	76.4	49.7
BLIP	4.3	15.2	19.7	9.3	19.7	6.6

Table 8. **Continent-wise Partial Precision.** Despite of training data imbalance, the training on our dataset benefits all continents.

6. Conclusion

We present a large, specialized dataset for VQA on museum exhibits, designed to bridge visual content and text-based queries. This dataset encompasses millions of images paired with varied questions, enabling models to deliver answers about a broad range of cultural artifacts. We fine-tune two VLMs, BLIP and LLaVA, to compare their performance on this museum VQA dataset. LLaVA, in particular, excels at answering visually unanswerable questions through reasoning and general knowledge. Additionally, cross-lingual tests confirm the adaptability of these models in multilingual contexts, highlighting their potential for use in diverse cultural and linguistic settings. This dataset and our experiments open doors for future applications in museum experiences. Models trained on MUSEUM-65 could support interactive virtual tours, where users ask detailed questions in their own languages. They could power digital curators, providing rich cultural insights, or integrate with AR to offer real-time, on-site interpretation, creating immersive learning experiences for museum visitors globally.

Acknowledgements

We highly appreciate Pratyush Sinha, Krishnav Bajoria, Mohit Sharma, Anshuman Biswal, Rishabh Varshney, Anjali Roy, Raluca Mocanu, Reni Paskaleva and Nora Paskaleva for their help in gathering and curating the data, and for all the support, ideas and relevant discussions during the project. This research was partially funded by the Ministry of Education and Science of Bulgaria (support for INSAIT, part of the Bulgarian National Roadmap for Research Infrastructure). We thank the Bulgarian National Archaeological Institute with Museum for the support and guidance. We thank all institutions included in European, Digital Public Library of America (DPLA), Smithsonian Institution, Ariadne Project and also to the aggregators themselves for providing open access to their data. We also thank to Carnegie Museums of Pittsburgh, Modern and Contemporary Art Museum Korea, Harvard Museums US, Peabody Museum US, ArtUK Project, Hermitage Museum Russia, South Wales Museum Australia, The Indian Museum Project, Colbase Project Japan, The Museum of New Zealand Te Papa Tongarewa and Penn Museum US for the access to their data that made this research possible. We thank Google DeepMind which provided vital support and resources for this research.

References

- [1] Panos Achlioptas, Maks Ovsjanikov, Kilichbek Haydarov, Mohamed Elhoseiny, and Leonidas J Guibas. Artemis: Affective language for visual art. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11569–11579, 2021. 3
- [2] Amith Ananthram, Olivia Winn, and Smaranda Muresan. Feelingblue: A corpus for understanding the emotional connotation of color in context. *Transactions of the Association for Computational Linguistics*, 11:176–190, 2023. 3
- [3] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 1
- [4] Zechen Bai, Yuta Nakashima, and Noa Garcia. Explain me the painting: Multi-topic knowledgeable art description generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5422–5432, 2021. 3
- [5] Ada-Astrid Balaucă, Danda Pani Paudel, Kristina Toutanova, and Luc Van Gool. Taming clip for fine-grained and structured visual understanding of museum exhibits. *arXiv preprint arXiv:2409.01690*, 2024. 2, 3
- [6] Hanoona Bangalath, Muhammad Maaz, Muhammad Uzair Khattak, Salman H Khan, and Fahad Shahbaz Khan. Bridging the gap between object and image-level representations for open-vocabulary detection. *Advances in Neural Information Processing Systems*, 35:33781–33794, 2022. 1, 2
- [7] Federico Becattini, Pietro Bongini, Luana Bulla, Alberto Del Bimbo, Ludovica Marinucci, Misael Mongiovì, and Valentina Presutti. Viscounth: A large-scale multilingual visual question answering dataset for cultural heritage. *ACM Transactions on Multimedia Computing, Communications and Applications*, 2023. 1, 2, 3
- [8] Simone Bianco, Luigi Celona, Paolo Napoletano, and Raimondo Schettini. Predicting image aesthetics with deep learning. In *Advanced Concepts for Intelligent Vision Systems: 17th International Conference, ACIVS 2016, Lecce, Italy, October 24-27, 2016, Proceedings 17*, pages 117–125. Springer, 2016. 3
- [9] Jeffrey P Bigham, Chandrika Jayant, Hanjie Ji, Greg Little, Andrew Miller, Robert C Miller, Robin Miller, Aubrey Tatarowicz, Brandyn White, Samuel White, et al. Vizwiz: nearly real-time answers to visual questions. In *Proceedings of the 23rd annual ACM symposium on User interface software and technology*, pages 333–342, 2010. 1
- [10] Pietro Bongini, Federico Becattini, Andrew D Bagdanov, and Alberto Del Bimbo. Visual question answering for cultural heritage. In *IOP Conference Series: Materials Science and Engineering*, page 012074. IOP Publishing, 2020. 3
- [11] Eva Cetinic, Tomislav Lipic, and Sonja Grgic. Fine-tuning convolutional neural networks for fine art classification. *Expert Systems with Applications*, 114:107–118, 2018. 3
- [12] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020. 2
- [13] Marcos V Conde and Kerem Turgutlu. Clip-art: Contrastive pre-training for fine-grained art classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3956–3960, 2021. 2
- [14] Peng Cui, Dan Zhang, Zhijie Deng, Yinpeng Dong, and Jun Zhu. Learning sample difficulty from pre-trained models for reliable prediction. *Advances in Neural Information Processing Systems*, 36, 2024. 1
- [15] E Dataset. Novel datasets for fine-grained image categorization. In *First Workshop on Fine Grained Visual Categorization, CVPR. Citeseer. Citeseer*. Citeseer, 2011. 3
- [16] Riccardo Del Chiaro, Andrew D Bagdanov, and Alberto Del Bimbo. Noisyart: A dataset for webly-supervised artwork recognition. In *VISIGRAPP (4: VISAPP)*, pages 467–475, 2019. 3
- [17] Riccardo Del Chiaro, Andrew D Bagdanov, and Alberto Del Bimbo. Webly-supervised zero-shot learning for artwork instance recognition. *Pattern Recognition Letters*, 128:420–426, 2019. 3
- [18] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 3
- [19] Yingying Deng, Fan Tang, Weiming Dong, Chongyang Ma, Xingjia Pan, Lei Wang, and Changsheng Xu. Stytr2: Image style transfer with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11326–11336, 2022. 1
- [20] Jian Ding, Nan Xue, Guisong Xia, and Dengxin Dai. Decoupling zero-shot semantic segmentation. 2022 IEEE. In

- CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11573–11582, 2021. 1, 2
- [21] 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
- [22] Haoyuan Gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei Xu. Are you talking to a machine? dataset and methods for multilingual image question. *Advances in neural information processing systems*, 28, 2015. 3
- [23] Noa Garcia and George Vogiatzis. How to read paintings: semantic art understanding with multi-modal retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018. 3
- [24] Noa Garcia, Mayu Otani, Chenhui Chu, and Yuta Nakashima. Knowit vqa: Answering knowledge-based questions about videos. In *Proceedings of the AAAI conference on artificial intelligence*, pages 10826–10834, 2020. 3
- [25] Noa Garcia, Chentao Ye, Zihua Liu, Qingtao Hu, Mayu Otani, Chenhui Chu, Yuta Nakashima, and Teruko Mitamura. A dataset and baselines for visual question answering on art. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 92–108. Springer, 2020. 3
- [26] Noa Garcia, Chentao Ye, Zihua Liu, Qingtao Hu, Mayu Otani, Chenhui Chu, Yuta Nakashima, and Teruko Mitamura. A dataset and baselines for visual question answering on art. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 92–108. Springer, 2020. 2, 3
- [27] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423, 2016. 1
- [28] Koustav Ghosal, Aakanksha Rana, and Aljosa Smolic. Aesthetic image captioning from weakly-labelled photographs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 3
- [29] 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
- [30] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv preprint arXiv:2104.13921*, 2021. 1, 2
- [31] Darryl Hannan, Akshay Jain, and Mohit Bansal. Many-modalqa: Modality disambiguation and qa over diverse inputs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7879–7886, 2020. 3
- [32] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021. 2
- [33] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910, 2017. 3
- [34] Chen Ju, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie. Prompting visual-language models for efficient video understanding. In *European Conference on Computer Vision*, pages 105–124. Springer, 2022. 1
- [35] Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Ákos Kádár, Adam Trischler, and Yoshua Bengio. Figureqa: An annotated figure dataset for visual reasoning. *arXiv preprint arXiv:1710.07300*, 2017. 3
- [36] Aishwarya Kamath, Mannat Singh, Yann LeCun, Gabriel Synnaeve, Ishan Misra, and Nicolas Carion. Mdetrm: modulated detection for end-to-end multi-modal understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1780–1790, 2021. 2
- [37] Matt Kusner, Yu Sun, Nicholas Kolkin, and Kilian Weinberger. From word embeddings to document distances. In *International conference on machine learning*, pages 957–966. PMLR, 2015. 5
- [38] Gihyun Kwon and Jong Chul Ye. Clipstyler: Image style transfer with a single text condition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18062–18071, 2022. 1
- [39] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. In *International Conference on Learning Representations*, 2022. 2
- [40] Boyi Li, Kilian Q. Weinberger, Serge J. Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. *CoRR*, abs/2201.03546, 2022. 1
- [41] 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. 1, 2, 5, 6
- [42] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019. 2
- [43] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 121–137. Springer, 2020. 2
- [44] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 3
- [45] Weizhe Lin, Jinghong Chen, Jingbiao Mei, Alexandru Coca, and Bill Byrne. Fine-grained late-interaction multi-modal

- retrieval for retrieval augmented visual question answering. *Advances in Neural Information Processing Systems*, 36, 2024. 1
- [46] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 1, 2, 5
- [47] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019. 2
- [48] Yue Lu, Chao Guo, Xingyuan Dai, and Fei-Yue Wang. Data-efficient image captioning of fine art paintings via virtual-real semantic alignment training. *Neurocomputing*, 490:163–180, 2022. 3
- [49] Muhammad Maaz, Hanoona Rasheed, Salman Khan, Fahad Shahbaz Khan, Rao Muhammad Anwer, and Ming-Hsuan Yang. Class-agnostic object detection with multimodal transformer. In *European Conference on Computer Vision*, pages 512–531. Springer, 2022. 2
- [50] Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko, and Andrea Vedaldi. Fine-grained visual classification of aircraft. *arXiv preprint arXiv:1306.5151*, 2013. 3
- [51] Mateusz Malinowski and Mario Fritz. A multi-world approach to question answering about real-world scenes based on uncertain input. *Advances in neural information processing systems*, 27, 2014. 3
- [52] Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz. Ask your neurons: A neural-based approach to answering questions about images. In *Proceedings of the IEEE international conference on computer vision*, pages 1–9, 2015. 1
- [53] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204, 2019. 3
- [54] Paul F Marty and Katherine Burton Jones. *Museum informatics: People, information, and technology in museums*. Taylor & Francis, 2008. 1
- [55] Fanqing Meng, Wenqi Shao, Zhanglin Peng, Chonghe Jiang, Kaipeng Zhang, Yu Qiao, and Ping Luo. Foundation model is efficient multimodal multitask model selector. *arXiv preprint arXiv:2308.06262*, 2023. 1
- [56] Thomas Mensink and Jan Van Gemert. The rijksmuseum challenge: Museum-centered visual recognition. In *Proceedings of international conference on multimedia retrieval*, pages 451–454, 2014. 3
- [57] Thomas Mensink, Jasper Uijlings, Lluís Castrejon, Arushi Goel, Felipe Cadar, Howard Zhou, Fei Sha, André Araujo, and Vittorio Ferrari. Encyclopedic vqa: Visual questions about detailed properties of fine-grained categories. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3113–3124, 2023. 3
- [58] Federico Milani and Piero Fraternali. A dataset and a convolutional model for iconography classification in paintings. *Journal on Computing and Cultural Heritage (JOCCH)*, 14(4):1–18, 2021. 3
- [59] Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. Expanding language-image pretrained models for general video recognition. In *European Conference on Computer Vision*, pages 1–18. Springer, 2022. 1
- [60] Ikrom Nishanbaev, Erik Champion, and David A McMeekin. A survey of geospatial semantic web for cultural heritage. *Heritage*, 2(2):1471–1498, 2019. 1
- [61] 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
- [62] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, 2002. Association for Computational Linguistics. 5
- [63] Hieu Pham, Zihang Dai, Golnaz Ghiasi, Kenji Kawaguchi, Hanxiao Liu, Adams Wei Yu, Jiahui Yu, Yi-Ting Chen, Minh-Thang Luong, Yonghui Wu, et al. Combined scaling for zero-shot transfer learning. *Neurocomputing*, 555:126658, 2023. 1
- [64] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 2
- [65] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021. 1
- [66] Hanoona Rasheed, Muhammad Uzair Khattak, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Fine-tuned clip models are efficient video learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6545–6554, 2023. 1
- [67] Dan Ruta, Andrew Gilbert, Pranav Aggarwal, Naveen Marri, Ajinkya Kale, Jo Briggs, Chris Speed, Hailin Jin, Baldo Faieta, Alex Filipkowski, et al. Stylelabel: Artistic style tagging and captioning. In *European Conference on Computer Vision*, pages 219–236. Springer, 2022. 1, 3
- [68] Fereshteh Sadeghi, Santosh K Kumar Divvala, and Ali Farhadi. Viske: Visual knowledge extraction and question answering by visual verification of relation phrases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1456–1464, 2015. 1
- [69] Shurong Sheng, Luc Van Gool, and Marie-Francine Moens. A dataset for multimodal question answering in the cultural heritage domain. In *Proceedings of the COLING 2016 Workshop on Language Technology Resources and Tools for Digital Humanities (LT4DH)*, pages 10–17. ACL, 2016. 2, 3
- [70] Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326, 2019. 1
- [71] Samuel Stevens, Jiaman Wu, Matthew J Thompson, Elizabeth G Campolongo, Chan Hee Song, David Edward Carlyn,

- Li Dong, Wasila M Dahdul, Charles Stewart, Tanya Berger-Wolf, et al. Bioclip: A vision foundation model for the tree of life. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19412–19424, 2024. 3
- [72] Gjorgji Strezoski and Marcel Worring. Omniart: a large-scale artistic benchmark. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 14(4):1–21, 2018. 1, 3
- [73] Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hannaneh Hajishirzi, and Jonathan Berant. Multimodalqa: Complex question answering over text, tables and images. *arXiv preprint arXiv:2104.06039*, 2021. 3
- [74] Wei Ren Tan, Chee Seng Chan, Hernán E Aguirre, and Kiyoshi Tanaka. Ceci n’est pas une pipe: A deep convolutional network for fine-art paintings classification. In *2016 IEEE international conference on image processing (ICIP)*, pages 3703–3707. IEEE, 2016. 3
- [75] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 1
- [76] C Wah, S Branson, P Welinder, P Perona, and S Belongie. The caltech-ucsd birds-200–2011 dataset. technical report california institute of technology. *Technical re-port California Institute of Technology*, 2011. 3
- [77] Mengmeng Wang, Jiazheng Xing, and Yong Liu. Actionclip: A new paradigm for video action recognition. *arXiv preprint arXiv:2109.08472*, 2021. 1
- [78] Peng Wang, Qi Wu, Chunhua Shen, Anton van den Hengel, and Anthony Dick. Explicit knowledge-based reasoning for visual question answering. *arXiv preprint arXiv:1511.02570*, 2015. 3
- [79] Peng Wang, Qi Wu, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. Fvqa: Fact-based visual question answering. *IEEE transactions on pattern analysis and machine intelligence*, 40(10):2413–2427, 2017. 3
- [80] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. *arXiv preprint arXiv:2210.10163*, 2022. 1
- [81] Yixuan Wei, Han Hu, Zhenda Xie, Ze Liu, Zheng Zhang, Yue Cao, Jianmin Bao, Dong Chen, and Baining Guo. Improving clip fine-tuning performance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5439–5449, 2023. 1
- [82] Michael J Wilber, Chen Fang, Hailin Jin, Aaron Hertzmann, John Collomosse, and Serge Belongie. Bam! the behance artistic media dataset for recognition beyond photography. In *Proceedings of the IEEE international conference on computer vision*, pages 1202–1211, 2017. 1, 3
- [83] Qi Wu, Peng Wang, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. Ask me anything: Free-form visual question answering based on knowledge from external sources. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4622–4630, 2016. 3
- [84] Desen Yuan. Language bias in visual question answering: A survey and taxonomy. *arXiv preprint arXiv:2111.08531*, 2021. 5
- [85] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18123–18133, 2022. 1
- [86] Chenyang Zhang, Christine Kaeser-Chen, Grace Vesom, Jennie Choi, Maria Kessler, and Serge Belongie. The imet collection 2019 challenge dataset. *arXiv preprint arXiv:1906.00901*, 2019. 2
- [87] Renrui Zhang, Rongyao Fang, Wei Zhang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free clip-adapter for better vision-language modeling. *arXiv preprint arXiv:2111.03930*, 2021. 2
- [88] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *European Conference on Computer Vision*, pages 696–712. Springer, 2022. 1, 2
- [89] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825, 2022. 2
- [90] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 2
- [91] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *European Conference on Computer Vision*, pages 350–368. Springer, 2022. 1, 2
- [92] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4995–5004, 2016. 1
- [93] Yichen Zhu, Minjie Zhu, Ning Liu, Zhicai Ou, Xiaofeng Mou, and Jian Tang. Llava- ϕ : Efficient multi-modal assistant with small language model. *arXiv preprint arXiv:2401.02330*, 2024. 2