

VQArt-Bench: A semantically rich VQA Benchmark for Art and Cultural Heritage

Andrea Alfarano*[†] Lorenzo Venturoli* Darío Negueruela del Castillo
University of Zurich, Max Planck Society

andrea.alfaran@uzh.ch lorenzo.venturoli@uzh.ch dario.neguerueladelcastillo@uzh.ch

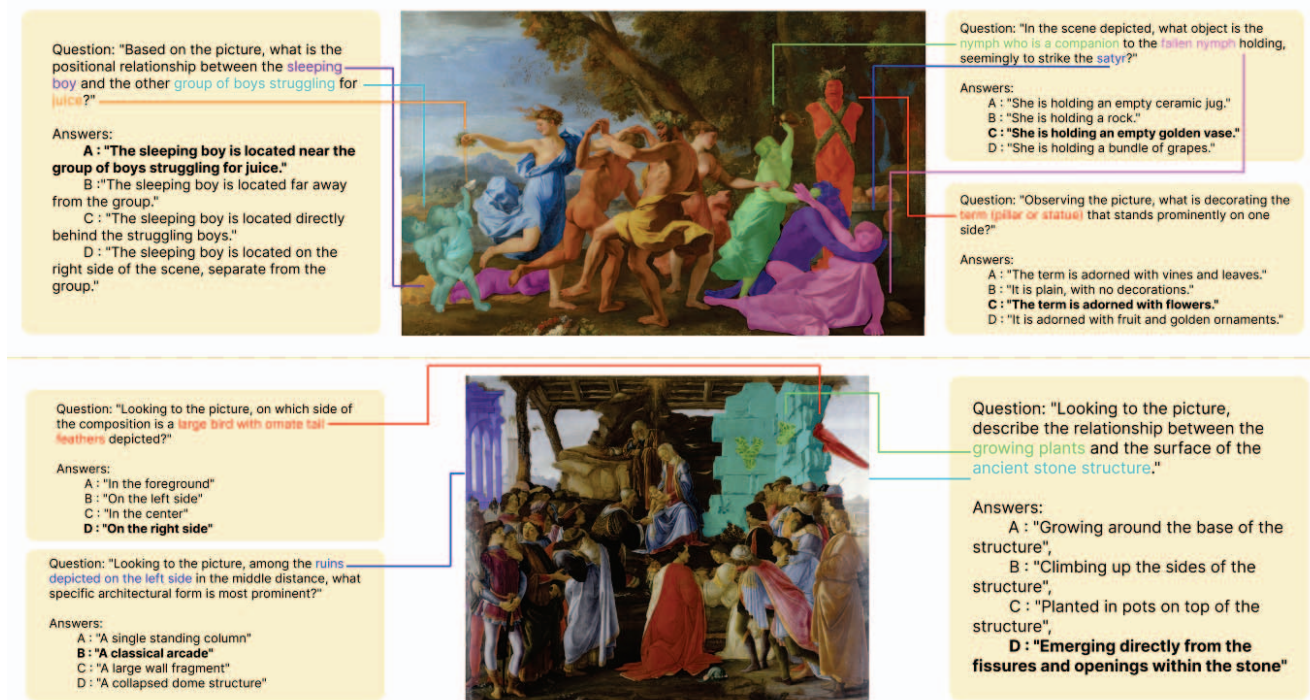


Figure 1. Examples from our VQArt-bench with highlighted related subjects for better visualization. VQArt-bench is deeply grounded in the related images and refers to specific elements or areas of the artwork. Each question requires a profound visual understanding to be answered correctly.

Abstract

Multimodal Large Language Models (MLLMs) have demonstrated significant capabilities in joint visual and linguistic tasks. However, existing Visual Question Answering (VQA) benchmarks often fail to evaluate deep semantic understanding, particularly in complex domains like visual art analysis. Confined to simple syntactic structures

and surface-level attributes, these questions fail to capture the diversity and depth of human visual inquiry. This limitation incentivizes models to exploit statistical shortcuts rather than engage in visual reasoning. To address this gap, we introduce VQArt-Bench, a new, large-scale VQA benchmark for the cultural heritage domain. This benchmark is constructed using a novel multi-agent pipeline where specialized agents collaborate to generate nuanced, validated, and linguistically diverse questions. The resulting benchmark is structured along relevant visual understanding dimensions that probe a model's ability to interpret symbolic meaning, narratives, and complex visual relationships. Our

*Equal contributors.

[†]Corresponding author: andrea.alfarano@uzh.ch

¹Dataset link: <https://github.com/AlfaranoAndrea/VQArt-Bench>.

evaluation of 14 state-of-the-art MLLMs on this benchmark reveals significant limitations in current models, including a surprising weakness in simple counting tasks and a clear performance gap between proprietary and open-source models. Our dataset is available here¹

1. Introduction

In recent years, MLLMs [22] have demonstrated exceptional capabilities in image-text comprehension, understanding, reasoning, and generation. VQA [5, 27] has emerged as a critical benchmark for evaluating such models’ capabilities. This task requires a model to answer a textual query about an image, a process that necessitates both linguistic and visual understanding.

Advancing these systems toward a true visual interpretation of art requires benchmarks that reflect the complex skills necessary for artistic analysis [6, 16]. This task is particularly challenging due to the significant domain shift between the images and textual information used to train large models and the unique language of visual art. Art, unlike many other visual domains, demands a sophisticated interpretive engagement. The relation between image and meaning in artistic contexts is complex, multilayered, and inherently non-linear. Despite significant progress in multimodal vision-language systems, most approaches still rely on relatively reductive assumptions about visual content: being primarily optimized for object detections, image captioning, or scene classification.

Constructing such representative benchmarks is fraught with challenges, and existing VQA benchmarks often fail, particularly in specialized domains like art and cultural heritage [6, 16]. These failures largely stem from the way large-scale VQA datasets for visual art analysis are constructed. Rule-based generation programmatically creates question-answer pairs by taking structured metadata (like image captions or object labels) and inserting them into a fixed set of sentence templates. Although this methodology is highly scalable, its reliance on rigid templates has severe limitations, especially in the boundless artistic domain, filled with its unique subjects, complex allegorical relationships, and symbolic actions. First, it leads to a lack of linguistic diversity as the questions are confined to a narrow range of syntactic structures. Second, it produces shallow questions failing to capture context, symbolism, or implicit relationships. Figure 2 exemplifies these limitations by comparing questions generated by our proposed method with those from the rule-based benchmark AQUA [16], which is the first and only publicly available benchmark for art-focused VQA [7]. To avoid cherry-picking, we report the first question from the AQUA test and validation sets and generate a question for the same image using the same captions from SemArt [15]. As shown, the rule-

based approach lacks correct terminology (e.g., referring to the Madonna as a “woman”) and can even introduce factual inaccuracies, such as hallucinating an “animal on the shirt” due to flawed metadata parsing.

Rule-based VQA approaches are hindered by their tendency to create questions and answers focused on simplistic, high-frequency patterns [18, 37]. This problem arises from generating queries automatically from metadata, leading to an overabundance of questions based on overrepresented common patterns like “What is near X?”. In turn, the answers become dominated by common terms; for instance, in the AQUA test set, “Human” and “Person” alone account for more than 30% of all correct replies. This skewed distribution encourages models to exploit linguistic priors instead of performing true visual analysis [1, 19]. A robust visual art benchmark must move beyond this, testing a model’s capacity to recognize the complex actions, rare subjects, and symbolic content—such as historical or cultural motifs—that define an artwork and are systematically ignored by standard datasets.

To address these challenges we move beyond the superficial inquiries of rule-based systems and propose a novel agentic pipeline [12] for visual art question generation. Our framework deconstructs the complex task of question creation into a sequential, multi-agent workflow comprising an agentic data cleaning and four question crafting agents: a Topic Selector, a Question Generator, a Question Refiner, and a final Judge. Initially, the Topic Selector analyzes the image caption to identify candidate question categories that can be posed given the available information. The Question Generator then crafts nuanced, open-ended questions based on these candidates. Subsequently, the Question Refiner converts these into a challenging multiple-choice format with plausible distractors, and finally, every question is validated by the Judge, which makes sure that each question is non-trivial, unambiguously answerable from the image, and linguistically correct.

The main contributions of this work are as follows:

- We point out current limitations of existing VQA benchmarks for art, demonstrating how rule-based approaches fail to evaluate deep semantic understanding.
- The introduction of a novel, multi-agent pipeline for generating questions that are linguistically complex, context-aware, and designed to test for nuanced visual reasoning.
- VQArt-bench, a new large-scale VQA dataset featuring semantically rich questions to rigorously benchmark genuine visual literacy in the cultural heritage domain.
- An extensive evaluation of 14 state-of-the-art (SOTA) models on our new benchmark across multiple dimensions, showcasing their capabilities in visual art analysis.

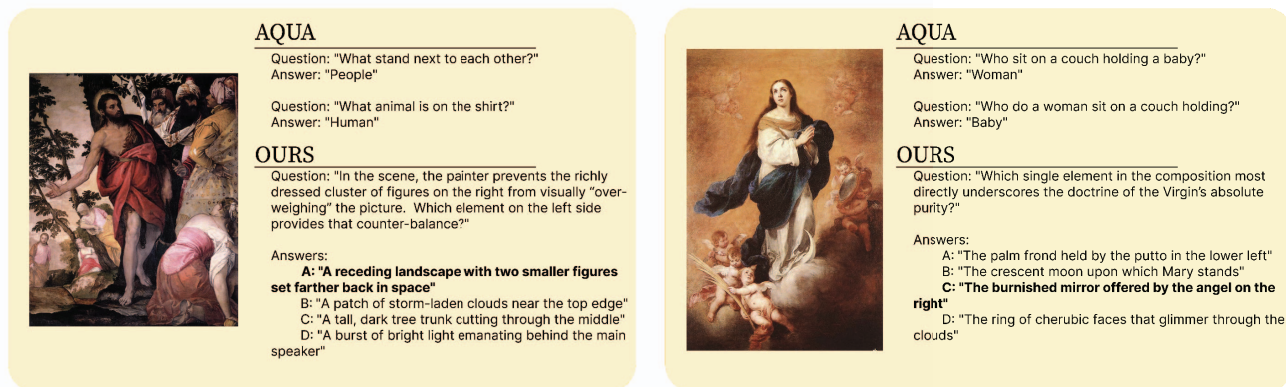


Figure 2. Demonstration of question quality improvement using our agentic pipeline (*Ours*) over a rule-based system (*AQUA*). All the questions have been generated from the same data source. While the rule-based questions are often shallow, lack nuance, and can be factually inconsistent, our method produces context-aware questions appropriate for fine-grained analysis. The rule-based approach produces semantically shallow questions that lack correct terminology (e.g., referring to the Madonna as a "woman") and can introduce factual inaccuracies, such as hallucinating an "animal on the shirt". In contrast, our agentic pipeline leverages LLMs to generate questions that are both challenging and precise. It preserves the nuances of the source material, formulating sophisticated questions that require a deep understanding of artistic composition (left example) and complex symbolism (right example). Correct answers are reported in bold text.

2. Related works

The Evolution of VQA and Multimodal Models Visual Question Answering (VQA) was introduced as a benchmark task to measure a machine’s ability to reason about visual content in response to natural language queries [34, 42]. The field has then grown rapidly, leading to the development of numerous datasets, such as COCO-QA [35] and VQA [5], and a variety of methods focusing on the joint embedding of visual and textual features [20, 39]. A parallel line of work has focused on grounding linguistic concepts to visual data, with large-scale datasets like Visual Genome [26] aiming to create dense, fine-grained alignments between images and text. More recently, the landscape has been reshaped by the success of a new class of MLLMs [22, 28, 40, 41, 43]. These models leverage the powerful reasoning and generation capabilities of LLMs to achieve state-of-the-art performance in multimodal comprehension.

Benchmarking MLLMs The rapid development of MLLMs has created an urgent need for benchmarks that can rigorously evaluate their capabilities. Recent benchmarks have been proposed, with a notable trend toward more objective evaluation formats. For instance, MME [13], MM-Bench [32], and SeedBench [27] introduced multiple choice and true/false questions to mitigate the ambiguity and cost associated with human evaluation or LLM-based scoring. However, many existing benchmarks face limitations: some are constrained to a narrow set of visual understanding skills, while others are relatively small in scale, which can

lead to unreliable performance metrics [27].

VQA in the Specialized Domain of Cultural Heritage

The translation of VQA into specialized domains such as cultural heritage presents a unique set of challenges [6, 8, 16]. Artworks contain visual information, such as painting techniques, iconographic symbols and historical styles, that is systematically absent from the natural image datasets typically used to train MLLMs. Existing art-specific VQA datasets, such as AQUA [16] and VISCOUNTH [6], tried to tackle the presented challenges relying on rule-based or template-based methods, but ultimately falling short due to the inherent limitations of their methods, as shown in Figure 2, explained in Sec. 1 and later in this section.

While efforts like AQUA and VISCOUNTH have introduced art-specific VQA datasets, their reliance on rule-based or template driven generation results in narrow, often superficial linguistic structures and symbolic blindness which arise from lack of grounding in the interpretive frameworks that have long structure art historical reasoning. The complexity of artistic images has been systematically addressed in scholarly traditions for more than a century. Wölfflin’s analysis of visual systems, Riegl’s notion of *Kunstwollen* and Warburg’s iconographic atlas (*Mnemosyne*) highlight how form, gesture and motif encode deep cultural memory. Erwin Panofsky’s layered theory of interpretation formalizes this complexity in three levels: iconographic recognition of motifs and forms, iconographic identification of subjects, narratives and symbols, and iconological analysis of cultural, ideological and psy-

chological models embedded in artworks. Computational ontology initiatives (ICON, IICONGRAPH) take up this challenge and offer data models and symbolic representations of these Panofskian levels of analysis, which enable semantic enrichment and structured querying. However, these are not designed to test the interpretative capacities of deep learning models in open-ended, dynamic settings. Our proposed benchmark complements these ontology-driven approaches, shifting the focus from structured representation to active reasoning.

Current limits in Automatic Question Generation

Methodologies Early methodologies for creating large-scale VQA datasets relied on rule-based systems that applied templates to an image’s structured semantic annotations, such as scene graphs [21, 24]. While highly scalable and logically grounded in image content, this process resulted in a lack of linguistic diversity and incentivized models to exploit statistical shortcuts rather than engaging in genuine visual reasoning [3, 17]. To address this limited expressiveness, subsequent research turned to Neural Question Generation (NQG) models, which promised greater linguistic variety [33] but introduced new challenges, including a critical propensity for “hallucinating” factually ungrounded content [23]. The introduction of LLMs in this pipeline marks the latest shift, demonstrating a remarkable ability to generate more diverse and seemingly faithful questions [29]. However, these powerful models do not completely overcome the shortcomings of their predecessors and keep introducing subtle factual inconsistencies or hallucinations that are harder to detect due to their high fluency [44]; furthermore, their monolithic nature makes the generation process difficult to control or verify.

3. Methods

3.1. Evaluation dimensions

To ensure that our benchmark evaluates a comprehensive range of visual understanding skills, we use a well-established hierarchy of seven distinct reasoning dimensions derived from [27]. These dimensions, inspired by cognitive science, structure the evaluation from basic perception to complex inference, allowing for a granular analysis of a model’s capabilities.

- **Instance Identity:** Involves identifying a specific instance, including its existence or class, based on visual evidence.
- **Instance Attribute:** Relates to the specific visual attributes of an instance, such as its color, shape, texture, or material.
- **Instance Location:** Concerns the absolute or relative position of a specified instance within the image frame.

- **Instance Counting:** Requires the model to accurately count all occurrences of a specific object class.
- **Spatial Relation:** Tests the ability to recognize the relative spatial relationship between two or more distinct objects.
- **Instance Interaction:** Requires recognizing actions or relationships between two or more subjects or objects.
- **Visual-Inspired Reasoning:** Evaluates a model’s ability to perform common-sense or causal reasoning based on the visual scene (e.g., inferring intent, cause, or future outcomes).

It is possible to see an example for each category in Fig. 3.

3.2. Data Curation and Pre-processing

Popular VQA benchmark curation methods propose to directly generate VQA from pictures by tasking an LLM to generate a relevant question given the picture [27]. We found that in visual art analysis, this can lead to superficial questions which assess basic visual appearance due to the difficulty of fully operationalizing what is *relevant*. An established option is to ground VQA questions on expert-curated descriptions and metadata often present in art repositories, so that the questions cover the same topics and relevant aspects that the curator decided to discuss. A key challenge in using such data is that it often does not contain only the visual information of the target artwork; rather, it usually interleaves the desired visual analysis with non-visual contextual information, such as artist biographies or historical background. Additionally, the contextual information for one artwork is often based on the descriptions of other artworks (such as preceding artworks by the same artists). This leads to the necessity of isolating visually relevant text, while also distinguishing which visual elements refer to the target work and which refer to external artworks. Such elements, if not filtered out, may lead to hallucinated questions in the following steps. For this purpose, we implemented a pre-processing step using an LLM that, unlike a rigid, rule-based filter that might be erroneously confused by text ambiguity, is able to leverage contextual understanding to parse the raw article and extract only the most relevant sentences that refer to the target artwork. This step yields a clean, relevant textual signal for our generation pipeline.

Our question generation pipeline is designed to be source-agnostic. For this work, we selected Wikipedia as our primary data source due to its vast repository of images accompanied by human-generated collaborative descriptions [38]; however, other resources can be used effectively. The initial phase involved downloading approximately 30,000 images and their corresponding descriptions. To ensure each description was sufficiently rich for generating detailed questions, we applied a length filter, discarding any image-text pair where the article contained fewer than 400 words. Each image text pair has been cleaned by an

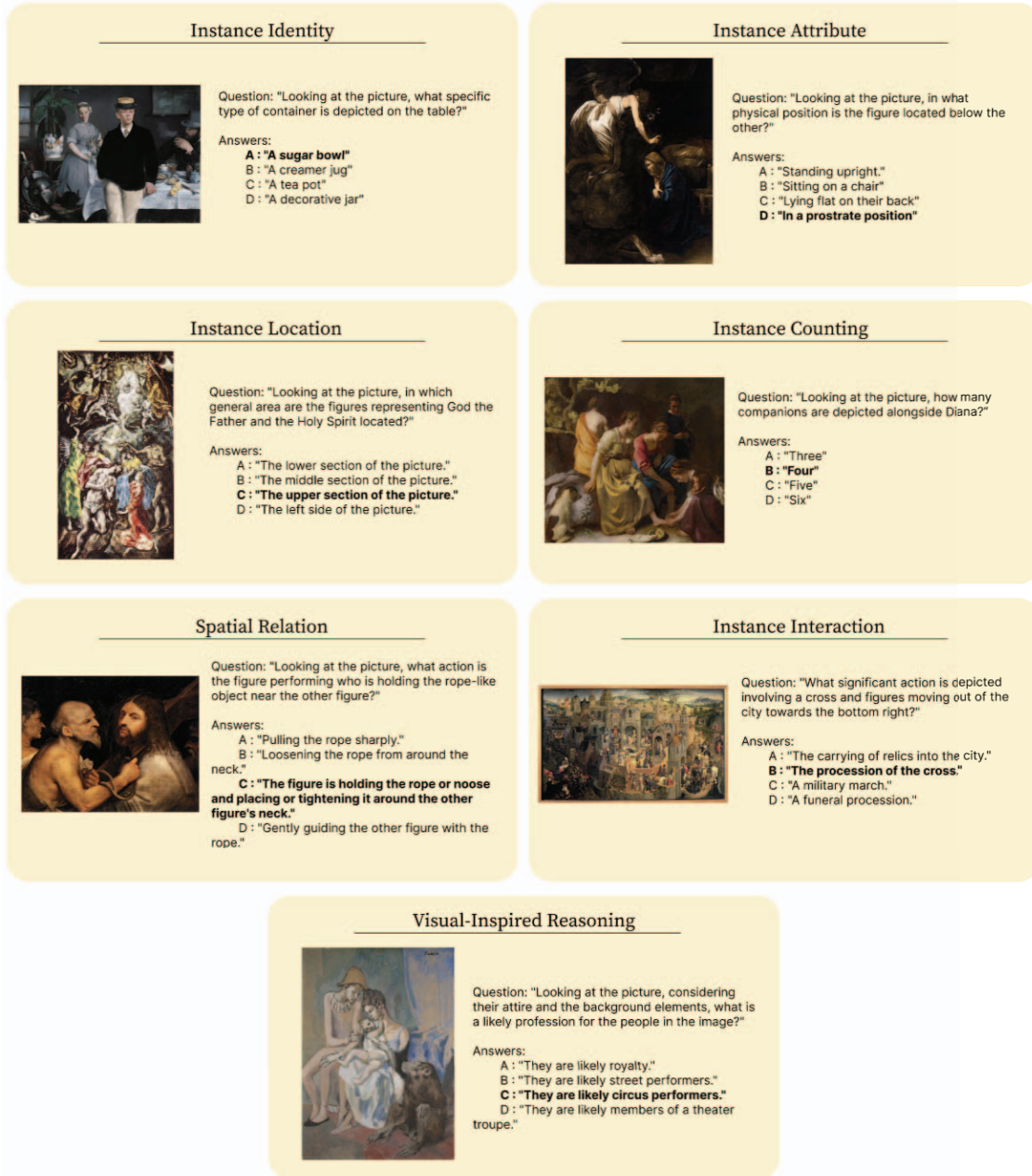


Figure 3. Examples from our *VQArt-Bench* dataset, categorized by our seven core evaluation dimensions. Our benchmark is designed to test a spectrum of visual reasoning skills. It challenges models on foundational abilities like identifying objects and their properties (*Instance Identity*, *Instance Attribute*), locating them in the scene (*Instance Location*), and quantifying them (*Instance Counting*). The evaluation then progresses to more complex compositional tasks, such as understanding *Spatial Relation* and *Instance Interaction*, and high-level tasks that require inferring context and causality (*Visual-Inspired Reasoning*). Correct answers are highlighted in Bold text.

LLM. We found that while passing both the picture and the text may improve the ability of the LLM to discriminate which content is actually in the picture, this also can potentially lead to hallucinate descriptions not contained in the text, and that the contextual understanding of the LLM is still sufficient for the task, leaving hallucination checks to

next steps in the VQA generation pipeline.

3.3. Agent-Based Question Generation

From the cleaned descriptions of 3.2 we generate questions using a sequential pipeline of four specialist agents. This division-of-labor approach ensures questions grounded in

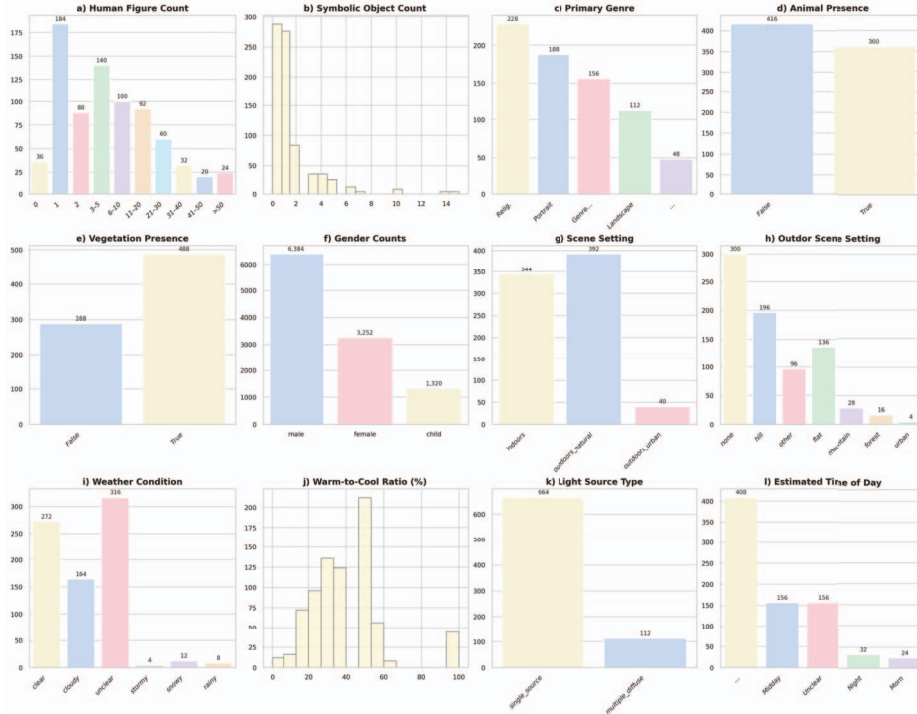


Figure 4. Exploitative LLM based statistical evaluation of key attributes in our *VQArt-Bench*. The dataset shows broad diversity in terms of **compositional elements**, including (a) a wide range of human figure counts, from individuals to large crowds; (b) varied numbers of symbolic objects per scene; and (f) representation across different genders. The collection spans multiple **genres and settings**, covering (c) primary artistic genres like religious and portraiture; (d, e) a balance of scenes with and without animal or vegetation presence; and (g, h) a mix of indoor and diverse outdoor environments. Finally, the dataset captures rich **atmospheric and stylistic variations**, including (l) different times of day; (i) various weather conditions; (k) simple and complex lighting sources; and (j) a full spectrum of warm-to-cool color ratios, indicating diverse visual moods.

the metadata, reduce hallucinations and relevance regarding the desired evaluations dimension of Sec. 3.1

Step 1: Topic Selection and Grounding The pipeline begins with the **Topic Selector** Agent, which is tasked to analyze the pre-processed visual description of each artwork, and propose a list of candidate topics that may be relevant for that picture. We found that this step is really prone to hallucinate facts to support the relevant questions. To better ground each question, for each candidate topic the agents output must cite the minimal text snippet from the cleaned description containing the supporting answer, ensuring that all subsequent generation is based on factual information.

Step 2: Open-Ended Question Formulation The grounded topics are then passed to the **Question Generator**. This agent takes the proposal from Step 1 and formulates nuanced, open-ended questions. While the topic agents primary goals is to ground the question and reduce hallucinations, the open ended agent goal is to formulate the question in the most relevant way, leaving the formulation

of the closed ended option to the next agents. The agent is tasked to keep the question *informed* by the text, and answerable *by observing the image* [36], avoiding questions that would be answerable only by looking to the metadata.

Step 3: Multiple-Choice Refinement and Distractor Generation The open-ended questions are then given to the **Question Refiner**, which converts them into a challenging multiple-choice format. This agent is explicitly tasked to design highly plausible distractors by anticipating possible visual misinterpretations grounded on the specific image, incorporating subtle details, and using contextually relevant but incorrect information [14]. This step is crucial for creating questions that demand deep and precise visual analysis.

Step 4: Final Judgment and Validation Finally, every multiple-choice question is evaluated by the **Judge**. This agent acts as a quality gatekeeper, ensuring that each question is unambiguously answerable from the image, is non-trivial, follows the provided evaluation dimensions and is

linguistically sound [45]. Only questions that pass this stringent assessment are included in the benchmark.

3.4. Human Validation and Quality Assurance

Although our automated pipeline is designed for high fidelity, we performed a human validation study to rigorously quantify the final quality of our benchmark. We randomly sampled 25% of the generated image-question pairs for manual review. Expert annotators were tasked with verifying that each question was factually grounded in the corresponding image. This process confirmed that over the 98% of the reviewed questions were free from hallucinations, underscoring the reliability of our agentic approach. To make sure that our questions capture the essence of the source descriptions, we compare the generated questions against the original text. The analysis found that in most of the cases, the questions successfully represented most of the salient information. This result demonstrates that our pipeline not only produces linguistically diverse questions but also achieves high information coverage, rivaling the thoroughness of more rigid methods.

3.5. Final Benchmark Composition

The complete **VQArt-Bench** benchmark consists of 14,463 high-quality, multiple-choice questions that span the seven reasoning dimensions presented above. The final distribution of questions across these categories is detailed in Table 1. This balanced composition ensures that the models are evaluated across a comprehensive spectrum of visual skills.

To provide an approximate quantitative overview of key attributes within VQArt-Bench, we automatically analyzed the dataset’s statistical distribution of compositional elements using Gemini 2.5 [40] as State-Of-The-Art visual LLM [29]. This analysis reveals the dataset’s broad diversity across several categories and can be seen in Figure 4. The collection shows significant variation in its compositional elements, including: (a) a wide range of human figure counts, with a focus on scenes depicting a small number of individuals (1-3 figures) but also including large crowds; (b) a varied representation across men, women, and children, with a majority presence of male figures

The dataset also spans multiple genres and settings, covering: (c) a distribution across primary artistic genres, with Religious art, Portraiture, being the most prominent categories; (d, e) a significant presence of vegetation, while most scenes do not feature animals, providing distinct contexts for analysis; and (g, h) a mix of indoor and outdoor environments, with indoor settings and natural landscapes being more common than urban scenes, and most outdoor scenes set during the day.

Finally, the dataset captures rich atmospheric and stylistic variations, including: (l) a variety of estimated times of day, with a focus on midday and afternoon; (i) a range of weather

conditions, predominantly featuring clear or cloudy skies; (k) a strong emphasis on artworks with a single, clear light source over more complex, diffuse lighting; and (j) a full spectrum of warm-to-cool color ratios, with a tendency towards balanced or cooler color palettes, indicating diverse visual moods.

Table 1. Distribution of Questions in the VQArt-Bench benchmark.

Reasoning Dimension	Number of Questions
Instance Identity	2031
Instance Attribute	2598
Instance Location	2100
Instance Counting	1710
Spatial Relation	2067
Instance Interaction	1794
Visual-Inspired Reasoning	2163
Total	14463

4. Evaluation Results

Evaluated models We evaluated 14 models, including 3 variants of Gemma3 [41] in order to test how performances change at scale. We evaluate other open source models as Aria [28], Aya Vision [11], Kimi-VL [25], Phi-4 [2], Pixtral 12B [4], LLaVA [30], LLaVA-NeXT [31], InstructBLIP-Vicuna [10, 46], as well as some SOTA closed source models as Gemini 2.5 [40], GPT-4o and GPT-4o mini [22].

4.1. Analysis

The evaluation results of different models on VQArt-Bench are listed in Table 2, where accuracy refers to the proportion of correctly answered multiple-choice questions relative to the total number of questions. We have observed a number of findings that can inform and guide future work.

Most MLLMs still exhibit limited performance As depicted in Tab. 2, most of the models analyzed fail to achieve particularly high overall scores, displaying limited capabilities in art understanding. Although most of the models achieve better results than random guessing (four choices: ~25%), most of them struggle to reach 50% accuracy.

MLLMs struggle with enumerating instances, while overperform in reasoning Table 2 shows that all evaluated models perform significantly below their overall accuracy in Instance Counting, yet excel in Visual-Inspired Reasoning. This outcome is counterintuitive, as a human would find counting elements much simpler than reasoning about an entire artwork. We can explain this phenomenon by looking at the skills required: Visual-Inspired Reasoning demands less specific knowledge about individual instances and a greater capacity for generalization, which is a key

Table 2. Evaluation results of different models on VQArt-Bench. Seven spatial–reasoning dimensions and overall accuracy.

Source	Model	Evaluation Dimensions								
		Instance Attribute	Instance Localization	Instance Counting	Spatial Relation	Instance Interaction	Instance Identity	Visual-Inspired Reasoning	Overall Acc.	
Closed Source	Gemini 2.5	0.73	0.73	0.66	0.72	0.75	0.74	0.80	0.71	
	GPT-4o	0.66	0.66	0.59	0.65	0.68	0.67	0.72	0.64	
	GPT-4o mini	0.58	0.58	0.53	0.58	0.60	0.59	0.64	0.57	
Open Source	Aria [28]	0.60	0.59	0.54	0.51	0.61	0.60	0.77	0.58	
	Aya Vision [11]	0.59	0.53	0.53	0.56	0.58	0.61	0.75	0.57	
	Kimi-VL [25]	0.69	0.70	0.64	0.66	0.67	0.72	0.83	0.67	
	Phi4 [2]	0.59	0.52	0.48	0.56	0.55	0.55	0.72	0.54	
	Pixtral [4]	0.65	0.63	0.57	0.62	0.61	0.63	0.81	0.62	
	InstructBLIP-Vicuna [10, 46]	0.25	0.21	0.10	0.21	0.24	0.47	0.26	0.24	
	LLaVA [30]	0.44	0.36	0.26	0.37	0.37	0.53	0.50	0.39	
	LLaVA-NeXT [31]	0.51	0.48	0.31	0.45	0.47	0.52	0.66	0.46	
	Gemma 3 27B [41]	0.41	0.39	0.39	0.41	0.47	0.54	0.54	0.42	
	Gemma 3 12B [41]	0.38	0.33	0.33	0.33	0.38	0.42	0.47	0.36	
	Gemma 3 4B [41]	0.35	0.27	0.26	0.30	0.39	0.39	0.47	0.33	

strength of current state-of-the-art MLLMs. Furthermore, Instance Counting questions are constructed with far more likely distractors, which easily trick “more naive” models, as it happens with InstructBLIP-Vicuna [10, 46].

Artistic visual analysis is more challenging than standard benchmarks While standard image benchmarks also feature naturalistic landscapes and human interactions, artistic scenes present a unique challenge. They often depict a broader range of subjects beyond everyday life, including historical, religious and fictional figures. The actions portrayed can be far more complex and uncommon than those typically found in online images. Furthermore, the visual appearance of art is fundamentally different due to diverse artistic styles and historical contexts, a stark contrast to the more uniform distribution of images found on the internet, from which large-scale benchmarks are typically sourced. This effect can be seen by looking at table 2, more specifically at the results achieved by InstructBLIP-Vicuna, which obtained a score just below the “random guessing” value of 25%, while achieving the best overall score in the SEED-Bench dataset [27]: 59% with similar evaluation.

Closed source models lead the performances Our evaluation demonstrates a clear performance disparity between closed-source and open-source models in the domain of artistic visual analysis. Notably, Gemini 2.5 surpasses all current baselines on every evaluation metric. Although open-source models like Kimi-VL are showing promising capabilities, they have not yet reached the same level of performance. This finding underscores the ongoing need for the open-source research community to intensify efforts in developing more capable models tailored to the unique challenges of artistic interpretation.

Kimi-VL is the best Open Source Model From the results in table 2 it appears that Kimi-VL [25] is the best model among all the open source MLLMs, even outperforming closed source models like GPT-4o and GPT-4o

mini. This may be due to its native-resolution vision encoder (MoonViT) and its novel training strategy, which directly targets visual reasoning instead of more standard tasks like image captioning. Kimi-VL performances suggest that MoE models [9] can be effective in improving performance while reducing active parameters.

Larger models perform generally better: We experimented with how the amount of parameters influences the performance by testing our benchmark against three different versions of Gemma 3 (4B, 12B, 27B) [41]. We observed that, as expected, the same model achieved better results at larger scales, without being subject to fine-tuning of any kind: Gemma 3 - 27B improves by +9% compared to the 4B version and by +6% compared to the 12B version. Another example of such is the improved score of GPT-4o when compared to its *mini* version.

5. Conclusion

In this work, we addressed the critical limitations of rule-based VQA for art. Such methods, constrained by rigid templates, produce linguistically and semantically shallow questions, leading to skewed data distributions that prevent genuine visual evaluation. To solve this, we introduced VQArt-bench, a new benchmark built with a novel agentic pipeline to generate semantically rich and challenging questions. Our evaluation of 14 state-of-the-art models revealed significant performance limitations, highlighting a surprising weakness in instance counting alongside a strength in visual reasoning. While closed models like Gemini 2.5 currently lead, the promising results from open-source models such as Kimi-VL show a path forward. VQArt-bench effectively exposes current model weaknesses and sets a more rigorous standard for developing AI with genuine visual literacy for art VQA.

References

- [1] Sherif Abdelkarim, Aniket Agarwal, Panos Achlioptas, Jun Chen, Jiaji Huang, Boyang Li, Kenneth Church, and Mohamed Elhoseiny. Exploring long tail visual relationship recognition with large vocabulary. In *Proceedings of the*

- IEEE/CVF International Conference on Computer Vision*, pages 15921–15930, 2021. 2
- [2] Marah Abdin et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024. 7, 8
- [3] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Anirudha Kembhavi. Don’t just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 4
- [4] Pravesh Agrawal et al. Pixtral 12b. *arXiv preprint arXiv:2410.07073*, 2024. 7, 8
- [5] 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 (ICCV)*, 2015. 2, 3
- [6] Federico Becattini, Pietro Bongini, Luana Bulla, Alberto Del Bimbo, Ludovica Marinucci, Misael Mongiovì, and Valentina Presutti. Viscount: a large-scale multilingual visual question answering dataset for cultural heritage. *ACM Transactions on Multimedia Computing, Communications and Applications*, 19(6):1–20, 2023. 2, 3
- [7] Tibor Bleidt, Sedigheh Eslami, and Gerard de Melo. Artquest: Countering hidden language biases in artvqa. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 7326–7335, 2024. 2
- [8] 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
- [9] Weilin Cai, Juyong Jiang, Fan Wang, Jing Tang, Sunghun Kim, and Jiayi Huang. A survey on mixture of experts. *arXiv preprint arXiv:2407.06204*, 2024. 8
- [10] Wenliang Dai, Junnan Li, et al. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500*, 2023. 7, 8
- [11] Saurabh Dash et al. Aya vision: Advancing the frontier of multilingual multimodality. *arXiv preprint arXiv:2505.08751*, 2025. 7, 8
- [12] Zane Durante, Qiuyuan Huang, Naoki Wake, Ran Gong, Jae Sung Park, Bidipta Sarkar, Rohan Taori, Yusuke Noda, Demetri Terzopoulos, Yejin Choi, Katsushi Ikeuchi, Hoi Vo, Li Fei-Fei, and Jianfeng Gao. Agent ai: Surveying the horizons of multimodal interaction, 2024. 2
- [13] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, et al. Mme: a comprehensive evaluation benchmark for multimodal large language models. *corr abs/2306.13394 (2023)*, 2023. 3
- [14] Chen-Kai Gao, Chieh-Hsin Tan, Ke-Jia Chen, and Hung-Yu Kao. Generating plausible distractors for reading comprehension questions with generation-based and retrieval-based methods. In *Proceedings of the 3rd Workshop on Neural Generation and Translation*, 2019. 6
- [15] 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. 2
- [16] 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, 2020. 2, 3
- [17] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017. 4
- [18] Tao He, Lianli Gao, Jingkuan Song, Jianfei Cai, and Yuanfang Li. Learning from the scene and borrowing from the rich: Tackling the long tail in scene graph generation. *arXiv preprint arXiv:2006.07585*, 2020. 2
- [19] Tao He, Lianli Gao, Jingkuan Song, Jianfei Cai, and Yuanfang Li. Learning from the scene and borrowing from the rich: Tackling the long tail in scene graph generation, 2020. 2
- [20] Hexiang Hu, Wei-Lun Chao, and Fei Sha. Learning answer embeddings for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5428–5436, 2018. 3
- [21] Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 4
- [22] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 2, 3, 7
- [23] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38, 2023. 4
- [24] 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 (CVPR)*, 2017. 4
- [25] Kimi Team. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*, 2025. 7, 8
- [26] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *corr. arXiv preprint arXiv:1602.07332*, 2016. 3
- [27] Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension, 2023. 2, 3, 4, 8
- [28] Dongxu Li, Yudong Liu, Haoning Wu, et al. Aria: An open multimodal native mixture-of-experts model. *arXiv preprint arXiv:2410.05993*, 2024. 3, 7, 8
- [29] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 4, 7

- [30] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 7, 8
- [31] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 7, 8
- [32] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player?, 2024. 3
- [33] Nasar Mostafazadeh, Ishan Misra, Jacob Devlin, Margaret Mitchell, Xiaodong He, and Lucy Vanderwende. Generating questions from images. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2016. 4
- [34] Anupam Pandey, Deepjyoti Bodo, Arpan Phukan, and Asif Ekbal. The quest for visual understanding: A journey through the evolution of visual question answering. *arXiv preprint arXiv:2501.07109*, 2025. 3
- [35] Mengye Ren, Ryan Kiros, and Richard Zemel. Exploring models and data for image question answering. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2015. 3
- [36] David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, Bontu Fufa Balcha, Chenxi Whitehouse, Christian Salamea, Dan John Velasco, David Ifeoluwa Adelani, David Le Meur, Emilio Villa-Cueva, Fajri Koto, Fauzan Farooqui, Federico Belcavello, Ganzorig Batnasan, Gisela Vallejo, Grainne Caulfield, Guido Ivetta, Haiyue Song, Henok Biadglign Ademtew, Hernán Maina, Holy Lovenia, Israel Abebe Azime, Jan Christian Blaise Cruz, Jay Gala, Jiahui Geng, Jesus-German Ortiz-Barajas, Jinheon Baek, Jocelyn Dunstan, Laura Alonso Alemany, Kumaranage Ravindu Yasas Nagasinghe, Luciana Benotti, Luis Fernando D’Haro, Marcelo Viridiano, Marcos Esteche-Garitagotia, Maria Camila Buitrago Cabrera, Mario Rodríguez-Cantelar, Mélanie Jouitteau, Mihail Mihaylov, Mohamed Fazli Mohamed Imam, Muhammad Farid Adilazuarda, Munkhjargal Gochoo, Munkh-Erdene Otgonbold, Naome Etori, Olivier Niyomugisha, Paula Mónica Silva, Pranjal Chitale, Raj Dabre, Rendi Chevi, Ruochen Zhang, Ryandito Diandaru, Samuel Cahyawijaya, Santiago Góngora, Soyeong Jeong, Sukannya Purkayastha, Tatsuki Kuribayashi, Teresa Clifford, Thanmay Jayakumar, Tiago Timponi Torrent, Toqeer Ehsan, Vladimir Araujo, Yova Kementchedzhieva, Zara Burzo, Zheng Wei Lim, Zheng Xin Yong, Oana Ignat, Joan Nwatu, Rada Mihalcea, Tamar Solorio, and Alham Fikri Aji. Cvqa: Culturally-diverse multilingual visual question answering benchmark, 2024. 6
- [37] Lingyun Song, Chengkun Yang, Xuanyu Li, and Xuequn Shang. A robust dual-debiasing vqa model based on counterfactual causal effect. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4242–4252, 2024. 2
- [38] Vignesh Srinivasan, Sudeep Garg, Chaitanya Boyapaty, Andrew Tomkins, Deepti Ghadiyaram, and Filip Pavetic. WIT: Wikipedia-based image text dataset for multimodal pretraining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 4
- [39] Zhou Su, Chen Zhu, Yinpeng Dong, Dongqi Cai, Yurong Chen, and Jianguo Li. Learning visual knowledge memory networks for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7736–7745, 2018. 3
- [40] 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. 3, 7
- [41] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Pettrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday

- Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 technical report, 2025. [3](#), [7](#), [8](#)
- [42] Qi Wu, Damien Teney, Peng Wang, Chunhua Shen, Anthony Dick, and Anton Van Den Hengel. Visual question answering: A survey of methods and datasets. *Computer Vision and Image Understanding*, 163:21–40, 2017. [3](#)
- [43] S Yin, C Fu, S Zhao, K Li, X Sun, T Xu, and E Chen. A survey on multimodal large language models. arxiv 2023. *arXiv preprint arXiv:2306.13549*. [3](#)
- [44] Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Enhong Chen, and Gaoang Wang. Woodpecker: Hallucination correction for multimodal large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023. [4](#)
- [45] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. [7](#)
- [46] Lianmin Zheng, Wei-Lin Chiang, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90 Tech. report, LMSYS, 2023. [7](#), [8](#)