Caption generation in Cultural Heritage: Crowdsourced Data and Tuning Multimodal Large Language Models

Artem Reshetnikov
Barcelona Supercomputing Center
Barcelona, Spain
a.reflesh@gmail.com

Maria-Cristina Marinescu
School of Management, IQS
Universitat Ramon Llull
Barcelona, Spain
cristina.marinescu@iqs.url.edu

Abstract

Automated caption generation for paintings enables enhanced access and understanding of visual artworks. This work introduces a novel caption dataset, obtained by manual annotation of about 7500 images from the publicly available DEArt dataset for object detection and pose estimation. Our focus is on describing the visual scenes rather than the context or style of the artwork - more common in other existing captioning datasets. The dataset is the result of a crowdsourcing initiative spanning 13 months, with volunteers adhering to explicit captioning guidelines reflecting our requirements. We provide each artwork in the dataset with five captions, created independently by volunteers to ensure diversity of interpretation and increase the robustness of the captioning model.

In addition, we explore using the crowdsourced dataset for fine-tuning Large Language Models with vision encoders for domain-specific caption generation. The goal is to improve the performance of multimodal LLMs in the context of cultural heritage, a domain with "small data" which often struggles with the nuanced visual analysis and interpretation required for cultural objects such as paintings. The use of crowdsourced data in the domain adaptation process enables us to incorporate the collective perceptual insights of diverse annotators, resulting in an exploration of visual narratives and a observing a reduction in hallucinations otherwise created by these large language models.

1 Introduction

To offer innovative methods for engaging with and understanding visual artefacts at scale, many systems rely on rich metadata - for instance, in the form of captions or descriptions. Having access to good captions of artworks not only facilitates broader public access to these artifacts but also fosters a deeper appreciation for their cultural significance. However, the automatic generation of captions is not without challenges. Artworks often present scenes with intricate symbolism and complex narratives, where the most important elements can be hard to identify and demand nuanced caption beyond simple object recognition.

In this paper we introduce a novel dataset of captions of the visual content of artworks and showcase how it can help in the domain adaptation of state-of-art approaches such as Multimodal Large Language Models (mLLMs) (Liu et al., 2023) for the task of caption generation. The image dataset was sourced from the publicly available DEArt object detection and pose estimation dataset (Reshetnikov et al., 2022b), a curated assemblage of paintings spanning diverse European cultures, centuries and artistic movements.

Our motivation for collecting this new dataset was twofold. First, good models rely on the existence of large amounts of quality data. For reasons that are (1) technical - small data with a large variety between the representation of objects - real or imaginary, depiction of actions usually not captured in photographs, etc - but also (2) the relatively low interest in cultural heritage - which results in limited effort and financing, there is still a considerable gap between how precise multimodal LLMs perform for photographs and artworks. This gap could be narrowed by new quality datasets. Secondly, we chose to focus on the visual scene because we believe that it is necessary to identify all/most of the elements to be able to assign cultural meaning to a work; additionally, in those cases where a visual setup can consistently derive further meaning could be inferred more reliably top-down (from domain knowledge) rather than being generated based on a limited dataset.

The model adaptation work was motivated by the experiments we ran that use mMMLs to generate captions for cultural heritage (CH) artifacts; the results underlined some apparent shortcomings: on the one hand, content unrelated to the visual

scene (i.e., mistaken identity for both objects and relationships between objects), and on the other, missing elements. Our hypothesis was that these models could effectively leverage domain knowledge from datasets like ours, to overcome some of their apparent limitations.

Our crowdsourcing campaign was hosted on the Zooniverse platform and involved volunteers from various backgrounds and expertise levels, who created detailed caption annotations for about 7500 DEArt paintings during a year-long period. Given that a high percentage of the images are non-iconic (Berg and Berg, 2009), gathering 5 different annotations per image allows for a diversity of perspectives and interpretations, which can make the trained model more robust. Based on this data, we use parameter-efficient fine-tuning techniques(Xu et al., 2023) and demonstrate the possibility of mitigating hallucinations in LLM-generated captions.

2 Related work

Early efforts in image captioning, such as (Vinyals et al., 2015), laid the groundwork for later advancements. The distinctive challenges posed by cultural heritage artworks demand specialized solutions due to several important features not present in everyday pictures: anachronic objects, imaginary beings, actions not present in photographs - eg decapitations, etc. Several works have made significant contributions in the area of captioning for cultural heritage, of which we briefly present those directly related to our task - visual content captioning.

(Cetinic, 2021) highlights the complexity of describing artworks with multiple levels of interpretation and develops a captioning model based on a large-scale dataset of artwork images annotated with concepts from the Iconclass classification system. The model is fine-tuned using a transformer-based vision-language pre-trained model. Results suggest that the model could generate meaningful captions that exhibit a stronger relevance to the visual art context than those generated by the baseline (pre-trained) model.

(Bai et al., 2021) introduces a multi-topic and knowledgeable art description framework (Bai et al., 2021) which models the generated sentences according to three artistic perspectives and enhances each caption with external knowledge (from Wikipedia). The framework is validated through an exhaustive analysis, both quantitative and qualitative, as well as a comparative human evaluation.

(Stefanini et al., 2019) addresses the problem of cross-modal retrieval of images and sentences coming from the artistic domain. The authors collect and manually annotate the Artpedia dataset that contains paintings and textual sentences describing both the visual content of the paintings and other (contextual) information. They then devise a visual-textual model that jointly addresses the challenge of the retrieval of images and sentences by exploiting the visual and textual chunks.

More recently, the ArtCap dataset(Lu et al., 2024) provides 3,606 paintings, each annotated with five captions, showcasing high-quality annotations and effectiveness in benchmarking painting captioning models. The SemArt dataset(Garcia and Vogiatzis, 2018), designed for semantic art understanding, includes fine-art paintings with attributes and textual artistic comments. It also introduces the Text2Art challenge, a multi-modal retrieval task linking artistic texts and paintings.

The DEArt dataset(Reshetnikov et al., 2022a) focuses on object detection and pose estimation for 15K images of European artwork between the 12th and the 18th centuries. It includes 69 object classes, many of which are specific to cultural heritage, but does not include caption annotations. Recognizing this gap and considering the rich variety of non-iconic images in DEArt, we decided to leverage a subset to create a caption generation dataset.

Recent advances in the field of Large Language Models (LLMs) (OpenAI, 2023) have seen the successful integration of visual information into these models, giving rise to a new generation of mLLMs. Notable among these is LLaVA (Liu et al., 2023), which, along with other models such as MiniGPT4 (OpenAI, 2023) and Instruct-BLIP (Dai et al., 2023), have shown impressive image captioning and question-answering capabilities.

Like LLMs and unlike most of the ArtCap and SemiArt works, our approach relies on crowdsourcing data. This has the advantage of training the model with a variety of interpretations of paintings, coming from volunteers with different levels of expertise in cultural heritage. We believe that this can make the trained model more flexible and accurate.

Other works in metadata generation for cultural heritage exist, but they at least partly focus on the generation of style and context information (artwork's history, author's biography etc.), which introduces noise in the captions.

3 Guidelines for caption generation

To create effective guidelines[1], we drew inspiration from established practices such as (Starr, 2022), and we discussed our proposal with several cultural heritage experts. After deciding to use Zooniverse as a platform, we received expert advice from one of their shepherds.

Our guidelines emphasize the requirements of clarity, simplicity, and objectivity. We encourage annotators to start captions with the most crucial elements, progressing from foreground to background. We recommend avoiding assumptions, e.g., the identity of characters, events or places, assumptions about time periods (which, e.g. may bias the choice of object names), or professional jargon. The focus should always be on what is visually present in the image, avoiding implications or intentions. Named entities should be identified, but only if they are clearly recognizable or convey important information. Guidelines provide specific instructions for spatial orientation, using absolute positioning and limiting the use of "background/foreground" to essential details. They also advocate for concise annotations, restricting captions to 250 characters, while encouraging multiple sentences for clarity and simplicity. The language should be straightforward and avoid comparative constructs (e.g. larger, smallest), pronouns, and unnecessary punctuation. The annotation interface included examples to illustrate the preferred style and promote clear and informative annotations in a standardized manner.

4 Crowdsourcing process and preprocessing of the annotated data

After a thorough assessment of various platforms, which included short tests for quality of annotations, we decided to use the Zooniverse platform. Zooniverse's established reputation in supporting citizen science projects with quality metadata, the possibility of hosting caption annotation tasks, and the reality that volunteers make for better annotators (possibly due to the inherent interest in the project), were decisive factors in our decision. Due to GDPR and other law restrictions, Zooniverse platform doesn't allow the collecting of data about volunteers. However, In March 2015, the Zooniverse team conducted a survey to better understand their volunteer community. The survey, part of a Master's thesis by Victoria Homsy at Oxford University, gathered responses from approximately 300 active participants. Key findings revealed a gender distribution of 60% male and 40% female volunteers. Age-wise, the community was diverse, with a slight underrepresentation of older individuals. Geographically, the user base was primarily from English-speaking countries, notably the UK and the US, each contributing about a third of the participants, while only 2% hailed from developing nations. Employment data indicated that around half of the volunteers were employed, 15% were retired, 10% unemployed, and 4% unable to work due to disability. The survey also highlighted a wide range of occupations among volunteers, including roles such as professor, administrator, guard, and various technical positions.

The crowdsourcing campaign was initiated with the design of a user-friendly interface (UI) to facilitate efficient interaction between volunteers and the paintings. We ran the campaign in batches to try to get 4-5 good annotations per image for increasingly larger subsets of DEArt, while at the same time keeping a balance between diversity and thematic consistency.

Concretely, we included images with different styles and from different time periods, while excluding most portraits and other iconic images with limited interest from a captioning perspective (e.g.images that weren't iconographic or that had low expected variability for the captions). This process was iterative and involved: (1) the gradual decrease of the size of the batches to increase the motivation of the volunteers to complete the work, and (2) the adaptation of the image selection process to propose paintings of complexity that had lead to good captions in previous batches.

The multiple captions generated for each painting by the different volunteers reflect diverse artistic interpretations and visual insights and thus help us train a more robust captioning model.

At the end of each batch annotation process, we ran a data health check; rigorous quality control mechanisms were applied to manually verify the adherence of captions to guidelines and to maintain thematic alignment. Corrections and clarifications were incorporated into our guidelines and User Interface to enhance annotation accuracy. This iterative batch approach enabled us to capitalize on the collective contributions of volunteers while preserving dataset integrity.

The total number of uploaded images was 7543.

[1] Link to guidelines

Dataset	Images	Captions per Image	Total Captions
Our Dataset	7,543	**4.57**	**34,535**
SemArt	**21,383**	1	21,383
ArtPedia	2,930	3.1	9,173

Table 1: Comparison of our dataset versus state-of-the-art caption datasets in CH. Our dataset features a balanced mix of images and captions per image, achieving the highest total caption count among the datasets.

The dataset health check was based on several rules:

1. Filter out captions with fewer than two tokens.
2. Filter out captions containing specific words. E.g. when presented with an image, some users introduce a caption of an image in the guidelines rather than the one that corresponds to the dataset image. Another (general) case was due to our campaign not allowing volunteers to skip images they didn't want to annotate.
3. Eliminate annotations by users who either didn't read the guidelines properly or intentionally chose not to follow them. Some examples we identified that fall in this class are "aN oLd DrAwInG!!!!!!!!!!!!!!!!!!!!", "bad example".
4. Users seem to be remarkably consistent in providing useless, or high-quality, annotations. This provided us with yet another criteria to eliminate all captions from specific users.

Following the dataset health check, 34535 captions were retained. Our crowdsourced dataset stands out for its richness (i.e. number of annotations per images and total annotations) and diversity (i.e. different annotator views, given by the number of annotations per image) in comparison to the datasets that are the largest and most relevant in cultural heritage, as indicated by the metrics in Table 1. While it contains fewer images than SemArt, our dataset offers an average of 4.57 captions per image. The higher caption diversity is crucial to train more nuanced models, as it reflects how a visual scene can be described differently - which increases the power of generalization. In contrast, SemArt provides only one caption per image (although of museum-expert quality), which may limit the range of insights available for each artwork. Although ArtPedia offers a moderate number of captions per image (3.1 on average), its total image count is significantly lower, leading to a smaller pool of captions overall (9173).

This comparative analysis highlights the balance achieved in our dataset between the quantity of images and the variety of annotations. The emphasis on obtaining multiple captions per image enriches the dataset by incorporating a variety of descriptive styles, subjective interpretations, and visual details, thus providing a comprehensive base for fine-tuning models. The iterative process of data collection and quality checks ensures that our dataset maintains both breadth and depth, allowing the generation of high-quality, diverse painting captions.

To measure diversity, we calculated three metrics and compared them with the ArtCap dataset. Our choice is due to the similarity in dataset structure (e.g. multiply captions per image, focus on visual content). Diversity was measured using the following metrics:

- Lexical Diversity: Counts unique words across captions (e.g., type-token ratio).
- Semantic Diversity: Measures how semantically different the captions are using embeddings.
- Edit Distance/Overlap: Measure by counting the minimum number of operations required to transform one string into the other.

Results are shown in Table 2. We will release the caption dataset after publication.

5 Model architecture and training process

For model training and caption generation experiments, we chose an open-sourced model, the LLAVA (Large Language and Vision Assistant) llava-v1.5-7b. LLAVA is a novel end-to-end large multimodal model that combines a vision encoder and an LLM for general-purpose visual and language understanding. It represents a significant

Metric	Our dataset	ArtCap	Observation
Lexical Diversity	**0.5831**	0.4765	Our dataset has higher lexical diversity, meaning it uses a larger variety of unique words relative to its total word count. This suggests that captions in our dataset are more varied in vocabulary compared to ArtCap data.
Semantic Diversity	0.8094	**0.8236**	Both datasets exhibit high semantic diversity, but ArtCap dataset is slightly more diverse. Captions in ArtCap likely describe the images using different structure of sentences more often than in our dataset.
Edit Distance	**174.73**	44.07	Our dataset has a much higher edit distance, indicating its captions are structurally more distinct. Captions in ArtCap dataset are more similar in word arrangement and structure.

Table 2: Comparative analysis of caption diversity metrics between our dataset and ArtCap. Higher values indicate greater diversity.

advancement in the field of multimodal AI, demonstrating impressive multimodal chat capabilities - sometimes of similar quality of captions as those generated by the multimodal GPT-4 - and setting a new state-of-the-art accuracy standard for QA(Rodrigues et al., 2024).

The LLAVA pre-trained visual encoder and the LLM connect using a simple projection matrix. This setup allows the model to convert images into a word embedding space, while textual input is also transformed into the same space. The image and word tokens are then passed to a LLaMA (Touvron et al., 2023) decoder, which produces output.

Retraining or even fine-tuning LLMs typically demands extensive datasets and significant GPU hours. This process not only consumes considerable computational resources but also carries the risk of catastrophic forgetting, where the model loses the knowledge it previously acquired when too many layers of the network update their weights. To address these challenges, one of the parameter-efficient fine-tuning (PEFT) approaches (Xu et al., 2023) has been used for domain adaptation of the LLAVA model. PEFT methods are designed to adjust only a small subset of the model's parameters while keeping the majority of them fixed. This makes the fine-tuning process more efficient and less resource-intensive. By focusing on a limited number of parameters, PEFT techniques significantly reduce the computational load and the amount of data required, enabling quicker and more cost-effective adaptation to new domains. This can lead to a more agile and scalable deployment of LLMs for specialized application domains, ensuring that the model remains both accurate and efficient.

LoRA (Low-Rank Adaptation of Large Language Models) (Hu et al., 2021) is one of the PEFT techniques to train LLMs on specific tasks or domains. This technique introduces trainable rank decomposition matrices into each layer of transformer architecture and also reduces the number of trainable parameters for downstream tasks while keeping the pre-trained weights frozen.

To further optimize resource usage and fine-tuning efficiency, we employed QLoRA (Quantized Low-Rank Adaptation) instead of traditional LoRA. QLoRA was the most optimal choice because it reduces the memory footprint even further by leveraging 4-bit quantization, allowing for the fine-tuning of LLMs on consumer-grade hardware without sacrificing model performance. The use of QLoRA enables efficient memory utilization, allowing us to fine-tune larger models with fewer hardware resources, significantly lowering both the cost and time required for adaptation (Han et al., 2024).

Our QLoRA(Table 3) configuration is characterized by several key parameters such as the rank and the alpha value, which contribute to better convergence and scalability. Additionally, the use of mixed-precision training with bfloat16 (BF16) and TensorFlow32 (TF32) enables faster computation while minimizing memory requirements. To ensure

Model Architecture			
LoRA Rank (r)	128	LoRA Alpha	256
Vision Tower	clip	MM Projector Type	mlp2xgelu
MM Projector LR	2e-5	Vision Select Layer	-2
Quantization Bits	4	Image Aspect	pad
Model Max Length	2048		
Training Configuration			
Train Batch Size	4	Eval Batch Size	4
Grad. Accum. Steps	16	DataLoader Workers	4
Learning Rate	2e-4	Weight Decay	0.0
Warmup Ratio	0.03	LR Scheduler	cosine
Training Epochs	10		

Table 3: LoRA fine-tuning hyperparameters organized by model architecture and training configuration.

effective utilization of resources, the data loading process is optimized with lazy preprocessing and efficient parallelism (Rasley et al., 2020) using multiple dataloader workers. The LLAVA architecture we implement utilizes Vicuna-7B as LLM (Zheng et al., 2023) and the ViT vision transformer (Dosovitskiy et al., 2021) from OpenAI's CLIP model (Dai et al., 2023), which incorporates advanced features like multimodal projection layers and gradient checkpointing (See Figure1). See more details about model parameters and training configuration in Table 3.

6 Evaluation

We employed multiple evaluation metrics to assess the quality of the image captions generated by the baseline (LLAVA) and fine-tuned models, including Rouge1 (R1), Rouge2 (R2), RougeL (RL), and RougeLsum (RLsum), which measure n-gram overlap between generated and reference captions. Additionally, we included Meteor, Cider, and ClipScore, providing a more comprehensive view of the captioning performance. Rouge metrics are particularly useful for evaluating fluency and structure through n-gram and subsequence overlaps, while Meteor and Cider provide insights into the semantic accuracy; ClipScore assesses the alignment between the generated captions and the visual content.

Table 4 presents the comparison between results with the baseline LLAVA model and its fine-tuned version using QLoRA - for our dataset and the SemArt dataset. Fine-tuning on our dataset led to significant improvements over the baseline; for instance, the Rouge1 score increased from 0.31 to 0.43, and Rouge2 rose from 0.09 to 0.18, indicating a stronger overlap with reference captions. RougeL and RougeLsum similarly improved from 0.21 to 0.31 and 0.21 to 0.32, respectively, reflecting enhanced structural consistency and coherence of generated captions. The fine-tuned LLAVA model also demonstrated notable gains in Meteor and Cider scores, with Cider improving from 0.28 to 0.48, suggesting a better match with the overall reference data. Additionally, ClipScore increased from 0.31 to 0.42, indicating a higher alignment between captions and the visual content of the images.

However, the results on SemArt were more modest. Fine-tuning improved Rouge1 from 0.19 to 0.21 and Rouge2 from 0.027 to 0.11, while the gains in RougeL and RougeLsum were similarly limited (0.14 to 0.16). The lower ClipScore of 0.315 for the fine-tuned LLAVA on SemArt, compared to 0.42 on our dataset, indicates that the captions generated for SemArt images were less contextually aligned with the visual content. This disparity suggests that the model's ability to generate highly relevant captions is influenced by the characteristics of the dataset used for training, with our dataset providing a better foundation for capturing the nuanced relationship between text and imagery.

Overall, the evaluation demonstrates that fine-tuning using QLoRA can significantly improve the performance of mMMLs when training for specific domains, especially when these domains do not (or cannot) have extensive datasets. Moreover, the richer and more diverse the manual annotations, the higher the quality of the generated captions, as reflected by the lower ClipScore.

7 Limitations and discussion

Given the widespread excitement surrounding LLM capabilities and despite the improvements our fine-tuned model brings, we questioned whether these quantitative results also reflect a better quality of the generated captions from a *human viewpoint*. We thus embarked on an empirical exploration; our experiments with the baseline LLAVA model and the improvements that the fine-tuned LLAVA model achieved point to limitations in terms of the effectiveness of general-purpose mLLMs in the absence of domain-specific adaptations.

1. Hallucinations: One of the most notable limitations observed was the baseline model's tendency to hallucinate (invent details not present in the actual artwork). E.g., in the caption of "Palas Athena in Fight against Centaurs" (Figure 2c), the baseline LLAVA model generated incorrect elements, such

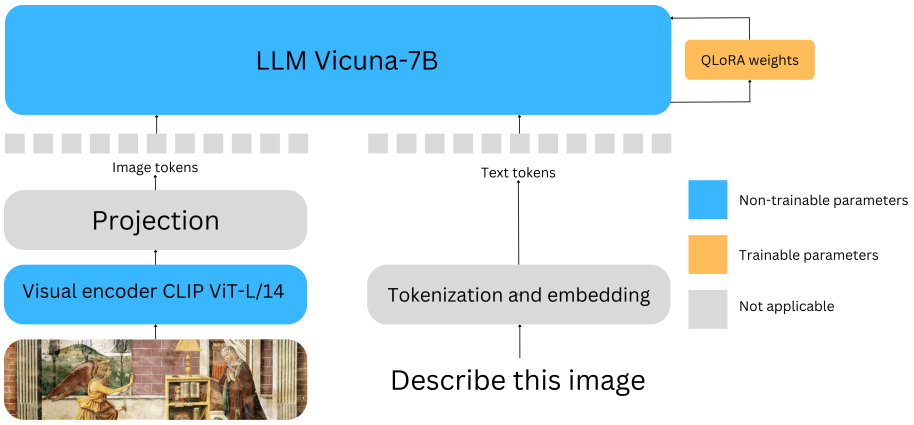

Figure 1: Architecture of LLAVA model with QLoRA layer

Model	R1	R2	RL	RLsum	Meteor	Cider	ClipScore
Baseline LLAVA (Our Dataset)	0.31	0.09	0.21	0.21	0.25	0.28	0.31
Fine-tuned LLAVA (Our Dataset)	0.43	0.18	0.31	0.32	0.31	0.48	0.42
Baseline LLAVA (SemArt)	0.19	0.027	0.14	0.14	0.12	0.19	0.21
Fine-tuned LLAVA (SemArt)	0.21	0.11	0.16	0.16	0.19	0.24	0.315

Note: R = ROUGE (R1 = ROUGE-1, R2 = ROUGE-2, RL = ROUGE-L, RLsum = ROUGE-L summary)

Table 4: Evaluation metrics for LLAVA models fine-tuned using QLoRA on two datasets (Our Dataset and SemArt).

as a dog and a bird, which do not exist in the painting. Similarly, for "Jupiter and Bellerophon" (Figure 2a), it inaccurately describes a scene involving angels when the painting actually features a man and a winged horse. This may also be interpreted to some extent as a case of mistaken identity in the case of the horse, whose wings made the baseline model believe it is an angel. On the other hand, the man on the left does not have wings, and the baseline model hallucinates angel instead. Finally, in "Annunciation" (Figure 2b), the basic model hallucinates a baby and a potted plant; this last could also be interpreted as mistaken identity since we assume the wings are interpreted as greenery. Both models hallucinate a man in white.

2. Incompleteness and mistaken identity: In several instances, the model produced captions that lacked crucial details. For example, in "Annunciation" (Figure 2b), the baseline model's caption mentions a woman and a child, omitting - or making the mistake - that the second figure is an angel and he is holding a flower; the baseline model does not see the wings, nor the flower. Both aspects are part of a significant religious interaction that is central to the meaning of the painting. Likewise, in Figure 2c, the centaurs are interpreted by the baseline model as horses; in Figure 2a, the model does not identify a horse at all.

The baseline model also mistakenly identifies crucial actions. In Figure 2a, the man *falls* rather than *being lifted*, while in Figure 2c, the woman is *killing* a beast rather than *standing* on it. Both actions are critical to understand the true meaning embedded in the artpiece, and its symbolism.

These limitations suggest that foundation models like LLAVA struggle with both accuracy and completeness in complex visual art contexts. Developing a domain-specific dataset and fine-tuning mLLM models like LLAVA for art interpretation is a more reliable approach.

8 Conclusion

This work introduces a novel dataset for caption generation of paintings sourced from a publicly available dataset. Our caption dataset was created through a year-long crowdsourcing campaign on Zooniverse, with volunteers adhering to explicit captioning guidelines. We have shown that this is a valuable resource for research and development in the field of automated caption generation for visual art. It can be used to fine-tune mLLMs - such as LLAVA - for domain adaptation, resulting in significant improvements in the quality of generated captions compared to the baseline models. The LoRA approach to fine-tuning mLLMs appears to be particularly effective in enhancing the model's ability to generate captions that are more coherent, contextually aligned, and better capture the nuances of

(a) "Jupiter and Bellerophon", unknown.

Baseline LLAVA: *The image depicts a painting of a man being lifted into the air by two angels. The man is positioned in the center of the scene, with one angel on his left side and the other on his right side. The angels are holding the man, and he appears to be flying through the sky. In the background, there are two people standing, possibly observing the scene or waiting for their turn to be lifted. The painting captures a moment of wonder and awe.*

Fine-tuned LLAVA: *The image features a painting of a man and a winged horse flying through the sky. The man falls from the horse, which is positioned towards the right side of the scene. The painting captures the dynamic and dramatic moment of the man and the horse soaring through the air. In the background, there is another man visible, on the left side of the painting. The scene appears to be set in a mountainous landscape, with a mountain visible in the background. The overall atmosphere of the painting is one of adventure and freedom.*

(b) "Annunciation", D. Ghirlandaio.

Baseline LLAVA: *The image depicts a scene from a painting, featuring a woman and a child. The woman is kneeling down, holding a baby, while the child is standing nearby. The woman appears to be dressed in a yellow gown, and the child is wearing a white gown. The painting is set in a room with a window in the background. In the room, there is a chair positioned towards the right side of the scene, and a potted plant is located on the left side. Additionally, there is a bowl placed on the floor, close to the woman and the child.*

Fine-tuned LLAVA: *The image features a painting of a woman sitting in front of an angel, who is kneeling and holding a white flower. They appear to be engaged in a conversation or a religious ceremony. The woman is wearing a yellow dress, while the man is dressed in white. In the background, there is a window with a curtain, which adds to the atmosphere of the scene. The painting is likely a religious or historical depiction, as it captures the interaction between the two individuals.*

(c) "Palas Athena in Fight against Centaurs", Milan Thomka Mitrovský.

Baseline LLAVA: *The image features a woman dressed in a warrior costume, standing on a horse and holding a shield. She appears to be engaged in a battle or a mock fight. The horse is positioned in the center of the scene, with the woman standing on its back. There are two other people in the background, one on the left side and the other on the right side of the image. The scene also includes a dog, which is located on the left side of the image, and a bird, which is flying in the sky above the woman and the horse.*

Fine-tuned LLAVA: *The image depicts a painting of a woman killing a centaur. The woman is wearing a helmet and is holding a shield, while the man appears to be holding a sword. The scene takes place in a lush green field, with a body of water in the background. There are several other centaurs in the painting, some of whom are closer to the water and others further away. The overall atmosphere of the painting suggests a battle or a moment of intense action.*

Figure 2: Comparison of generated captions between baseline and fine-tuned LLAVA models on different artworks.

artistic interpretation. We believe that this research will contribute to further advancements in automated caption generation for paintings and other forms of visual art, ultimately enhancing accessibility and understanding of these cultural artifacts.

9 Ethical and broader impact of the work

Participation in the annotation campaign was voluntary. Annotators were informed about the purpose, benefits, risks, and funding behind the study before participating. The dataset we used as a source of images has a Creative Commons license and is openly available. We pseudo-anonymized the collected data based on identifiers. We did not collect any personally identifiable data beyond user names on the Zooniverse platform. We recognize no additional potential for harm in our work beyond those already incurred by LLMs (e.g. bias), and our approach fine-tunes one such mLLM model to make it more accurate for the cultural heritage domain. AI assistants were not used in this work. Upon publication, we will release the dataset publicly for research use, which is classified as a "not high-risk" according to the EU Artificial Intelligence Act. We are not aware of any other possible ethical consequences of the proposed dataset and fine-tuned model.

10 Acknowledgement

This publication uses data generated via the Zooniverse.org platform, development of which is funded by generous support, including a Global Impact Award from Google, and by a grant from the Alfred P. Sloan Foundation.

This research has been supported by Saint-George-on-a-Bike (project 2018-EU-IA-0104), co-financed by the Connecting Europe Facility of the European Union.

References

Zechen Bai, Yuta Nakashima, and Noa Garcia. 2021. Explain me the painting: Multi-topic knowledgeable art description generation. *Preprint*, arXiv:2109.05743.

Tamara L. Berg and Alexander C. Berg. 2009. Finding iconic images. *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–8.

Eva Cetinic. 2021. Towards generating and evaluating iconographic image captions of artworks. *Journal of Imaging*, 7:123.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Preprint*, arXiv:2305.06500.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. *Preprint*, arXiv:2010.11929.

Noa Garcia and George Vogiatzis. 2018. How to read paintings: Semantic art understanding with multi-modal retrieval. *Preprint*, arXiv:1810.09617.

Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang. 2024. Parameter-efficient fine-tuning for large models: A comprehensive survey. *Preprint*, arXiv:2403.14608.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Preprint*, arXiv:2304.08485.

Yue Lu, Chao Guo, Xingyuan Dai, and Fei-Yue Wang. 2024. Artcap: A dataset for image captioning of fine art paintings. *IEEE Transactions on Computational Social Systems*, 11(1):576–587.

OpenAI. 2023. Gpt-4v(ision) system card.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 3505–3506, New York, NY, USA. Association for Computing Machinery.

Artem Reshetnikov, Maria-Cristina Marinescu, and Joaquim More Lopez. 2022a. Deart: Dataset of european art. *Preprint*, arXiv:2211.01226.

Artem Reshetnikov, Sergio Mendoza, and Maria-Cristina Marinescu. 2022b. Deart: Dataset of european art.

Luiz Rodrigues, Filipe Dwan Pereira, Luciano Cabral, Dragan Gašević, Geber Ramalho, and Rafael Ferreira Mello. 2024. Assessing the quality of automatic-generated short answers using gpt-4. *Computers and Education: Artificial Intelligence*, 7:100248.

Ruth Starr. 2022. Cooper hewitt guidelines for image description.

Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Massimiliano Corsini, and Rita Cucchiara. 2019. Artpedia: A new visual-semantic dataset with visual and contextual sentences in the artistic domain.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. *Preprint*, arXiv:1411.4555.

Lingling Xu, Haoran Xie, Si-Zhao Joe Qin, Xiaohui Tao, and Fu Lee Wang. 2023. Parameter-efficient fine-tuning methods for pretrained language models: A critical review and assessment. *Preprint*, arXiv:2312.12148.

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. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.