Bringing Rome to Life: Evaluating Historical Image Generation

Phillip B. Ströbel[1,2,*,†], Zejie Guo[1,†], Ülkü Karagöz[1,†], Eva Maria Willi[2] and Felix K. Maier[2]

[1]Department of Computational Linguistics, University of Zurich, Andreasstrasse 15, 8050 Zurich, Switzerland
[2]Department of History, University of Zurich, Karl Schmid-Strasse 4, 8006 Zurich, Switzerland

Abstract

This study evaluates the potential of AI image generation for visualising historical events, focusing on two ancient Roman scenarios: the Roman triumph and the *Lupercalia* festival. Using DALL-E 3, we generated 600 images based on 100 prompts derived from scientific texts. We then conducted a two-part evaluation: (1) A human evaluation by 21 history students, who compared image pairs and rated individual images on accuracy and prompt alignment, and (2) two automated analyses, one modelled after the human evaluation protocol and one using visual question-answering (VQA) techniques.

Our results reveal both the promise and limitations of AI in historical visualisation. While DALL-E 3 produced many convincing images, there were notable discrepancies between human and automated assessments. We found that Large Language Models tend to rate images more favourably than human evaluators.

We contribute a novel dataset for historical image generation, initial human and automated evaluation protocols, and insights into the challenges of using AI for historical visualisation, which is incredibly important for historians to reconstruct past events. Our findings highlight the need for refined evaluation methods and underscore the complexity of assessing historical accuracy in AI-generated imagery. This study lays the groundwork for future research on improving AI models for historical visualisation and developing more robust evaluation frameworks.

Keywords

Digital Humanities, image generation, human evaluation, automatic evaluation, history, image dataset

1. Introduction

Historians, akin to criminologists, analyse primary sources and eyewitness accounts to extract meaning and understand the motives and circumstances of historical events. However, unlike criminologists, who can re-enact events, historians face the challenge of studying occurrences that cannot be replicated or reproduced in experiments. This presents a significant challenge in their work.

Criminologists have developed methods to mitigate the uncertainties involved. Re-enacting crucial moments of an action or crime using real people or AI-based simulations has become

CHR 2024: Computational Humanities Research Conference, December 4–6, 2024, Aarhus, Denmark
*Corresponding author.
†These authors contributed equally.
phillip.stroebel@uzh.ch (P. B. Ströbel); zejie.guo@uzh.ch (Z. Guo); uelkue.karagoez@uzh.ch (Ü. Karagöz); evamaria.willi@uzh.ch (E. M. Willi); felix.maier@hist.uzh.ch (F. K. Maier)
0000-0003-2063-5495 (P. B. Ströbel); 0000-0002-5578-723X (F. K. Maier)

one of the most potent tools in criminology. These re-enactments allow us to visualise the action, providing a cinematic perspective that clarifies a crime's sequence and spatial dynamics. This process enhances our understanding by enabling us to perceive previously unseen details and prompting further questions. By experiencing events as if we were witnesses, we gain a newfound clarity.

Surprisingly, despite the similar challenges faced by both professions, historians have yet to embrace this re-enactment approach fully. Our project aims to change that. Leveraging rapid advancements in AI development, we aim to introduce an innovative platform to redefine how we perceive and comprehend historical events. Our goal is to develop a web application that generates storyboards or individual images from historical texts.

This 're-experiencing' of history will empower users to recapture seemingly 'lost' historical moments. Users can model specific actions or occasions from diverse perspectives by employing various scenarios with AI support. This approach will uncover performative dynamics, potentially revealing previously undisclosed aspects of historical events, much like the re-enactment in criminology.

However, we must question the suitability of such image-generation models. They require testing for historical accuracy before we can employ them for the previously mentioned purposes. We chose two specific Roman scenarios to test the capacity of DALL-E 3 to create historically accurate and engaging images: the Roman procession and the *Lupercalia* festival.

1.1. Our Contribution

Our study focuses on these two events due to their significance in Roman culture and the varying levels of textual and visual documentation available for each. The Roman triumph, a well-documented celebration of military victory, provides a rich base of textual descriptions. In contrast, the Lupercalia, an ancient fertility festival, offers a more challenging scenario with fewer detailed contemporary accounts.

To assess DALL-E 3's capabilities in this domain, we generated 600 images – 450 for the triumph and 150 for the *Lupercalia* (see Section 3). Our evaluation process is twofold:

1. **Human evaluation**: We conducted a comprehensive review involving 21 advanced history students to assess the images' historical accuracy.
2. **Automated analysis**: We employed computer vision techniques to analyse the images for prompt alignment.

This dual approach allows us to measure the generated images' subjective impact on human viewers and their objective alignment with historical data. Our research contributes to the broader discussion of AI's potential in historical visualisation and its limitations and contains the following items:

1. A novel, automatically generated dataset comprising 100 prompts and 600 images for historical image generation.
2. An initial human evaluation of a subset of these automatically generated images.
3. An initial automatic evaluation of the same subset.
4. An assessment of how well human and automatic evaluation correlate.

2. Related Work

The evaluation of automatically generated images has recently gained traction, mainly due to the increasingly sophisticated image generation models. Otani, Togashi, Sawai, Ishigami, Nakashima, Rahtu, Heikkilä, and Satoh [24] contemplated, based on an extensive analysis of 37 papers, that human evaluation protocols are often not reproducible and lack a clear description. Moreover, evaluation usually relies on automatic measures that poorly align with human scores.

The advantage of human feedback is that it can improve text-to-image models, e.g., with reinforcement learning from human feedback (as used in Natural Language Processing [28]). Xu, Liu, Wu, Tong, Li, Ding, Tang, and Dong [33] exploited a dataset of 8,878 prompts and 136,892 image comparisons to fine-tune a reward model that aligns more closely with human preferences. Liang, He, Li, Li, Klimovskiy, Carolan, Sun, Pont-Tuset, Young, Yang, Ke, Dvijotham, Collins, Luo, Li, Kohlhoff, Ramachandran, and Navalpakkam [17] used human feedback concerning *Plausibility*, *Aesthetics*, *Text-image Alignment*, and an *Overall* impression to predict human feedback scores. Due to the successful integration of human feedback in the model fine-tuning by Xu, Liu, Wu, Tong, Li, Ding, Tang, and Dong [33], we created an evaluation scenario which allows us to integrate such feedback directly in future work (see Section 4.1).

While Xu, Liu, Wu, Tong, Li, Ding, Tang, and Dong [33] focused on prompt-to-image alignment, other image properties are open for evaluation. Lee, Yasunaga, Meng, Mai, Park, Gupta, Zhang, Narayanan, Teufel, Bellagente, Kang, Park, Leskovec, Zhu, Li, Wu, Ermon, and Liang [16] worked on holistic image evaluation and identified twelve aspects among which we find *Alignment*, *Quality*, *Aesthetics*, and *Originality* (among others). Evaluating each aspect calls for different measures, some of them human, some of them automated. They created a holistic image evaluation benchmark for existing datasets and reported scores for all aspects and 26 models. While such an evaluation effort is valuable and provides a helpful oversight, we focus on prompt-to-image alignment evaluation in this work.

The research mentioned above has had access to large and heterogeneous datasets and results from extensive evaluation campaigns. In the context of historical image generation, such work does not yet exist. One exception is the investigation of Fareed, Bou Nassif, and Nofal [8] who tested the usage of Leonardo[1] for teaching purposes in the field of "History of Architecture". They evaluated the usability of Leonardo with a questionnaire after a workshop, which generally showed a need for the evaluation of AI-generated images for usage in the historical domain.

3. Data Collection with DALL-E 3

Next, we outline the methodology for data collection using DALL-E 3 to generate images related to triumphal processions and the *Lupercalia*, which included the following steps:

1. **Collecting Historical Documents**: We collected resources (i.e., academic papers, books, and other relevant documents) about the triumph and the *Lupercalia* in ancient

[1] See https://leonardo.ai.

Rome. Specifically, we included five documents related to the *Lupercalia* [32, 29, 20, 7, 10] and 15 documents focused on triumphal processions [27, 22, 3, 2, 15, 14, 18, 23, 12, 25, 13, 19, 9, 1, 30].

2. **Creating Prompts from Documents**: For each document, we manually derived five prompts. Each prompt was designed to capture a specific scene described in the texts. E.g., a document on triumphal processions could include prompts about the attire Romans wore, the types of vehicles used, or the procession sequence. In total, we created 100 prompts.
3. **Image Generation with DALL-E 3**: We used each prompt to generate six images using DALL-E 3 [4] via the OpenAI API.[2] The 100 prompts resulted in 150 generated images for the *Lupercalia* and 450 for the triumphal processions.[3]

Note that we did not force the model to produce realistic images. This led to a great variety of image styles, some of which are indeed life-like, while others are more in the style of a Renaissance painting or a black-and-white pencil sketch. All prompts, however, are based on scientific literature. See Figure 1 for example images and prompts from the dataset.[4]

4. Evaluating Automatically Generated Data

The following sections focus on the different evaluation scenarios employing human annotators and automatic evaluation measures.

4.1. Human Evaluation

4.1.1. Human Evaluation Setup

We generated two evaluation scenarios to obtain feedback from human annotators.

Image Comparison (IC) The first scenario asks annotators to decide which of two images better reflects the prompt. This is a cognitively easier task. Much in the manner of Xu, Liu, Wu, Tong, Li, Ding, Tang, and Dong [33], we plan to use these ratings for fine-tuning models to produce more faithful images. The participants are instructed not to judge the image style. We only compared images generated with the same prompt, which, based on the formula $\frac{n(n-1)}{2}$ to find unique pairings, results in 15 pairs per prompt (as mentioned in the previous section, we generated six images per prompt). Multiplied with the 100 total prompts in the dataset, we arrive at 1,500 comparisons.

Image Rating (IR) The second task requires the participants to rate an image on a 5-point Likert scale with the following options:

1. The image does not match the prompt at all.

[2] See https://openai.com/index/openai-api.
[3] The image generation costs amount to $48.06.
[4] The whole dataset (images and prompts) is available on GitHub. See https://github.com/AncientHistory-UZH/CHR2024_prompt-and-image-dataset.

Figure 1: Four example images for the two scenarios generated with DALL-E 3. Top row (**a** and **a'**), triumphal procession, prompt: *Generate an image of Trajan's Triumph as it passes through the Circus Maximus from the point of view of one of the around 150,000 to 250,000 spectators.*
Bottom row (**b** and **b'**), *Lupercalia*, prompt: *Create a historical image of a group of Luperci running about naked and holding thongs made of goat hides during the Lupercalia ritual in 44 BCE at the foot of the Palatine Hill. As they run past people they strike them with the thongs. They are laughing, larking about the exchanging obscenities with those who attended the ritual. People seem to be happy with what's going on.*

2. The image barely contains aspects of the prompt.
3. The image catches some aspects of the prompt, but it is not very accurate.
4. The image catches most of the aspects of the prompt.
5. The image completely matches the prompt.

Additionally, we asked the users to describe which aspects of the image did *not* correspond to the prompt in a text field. In this scenario, which demands more time and effort, we need 600 ratings for one complete dataset annotation.

We set up a Prodigy interface,[5] which we used to obtain the assessment of the annotators. See Figure 2 to get an impression of the annotation environment. We recruited 21 advanced history students for the annotations. We did not ask the participants to annotate a specific number of pairs. They were compensated with book vouchers of a value of $30. An online meeting was organised to explain the guidelines, emphasising that in the first scenario, they should judge based on the alignment of the images with the prompts rather than their visual appeal. They should consider visual features only if the two images reflect the prompts equally. The students spent approximately one afternoon annotating the data in both scenarios.

[5]See https://prodi.gy.

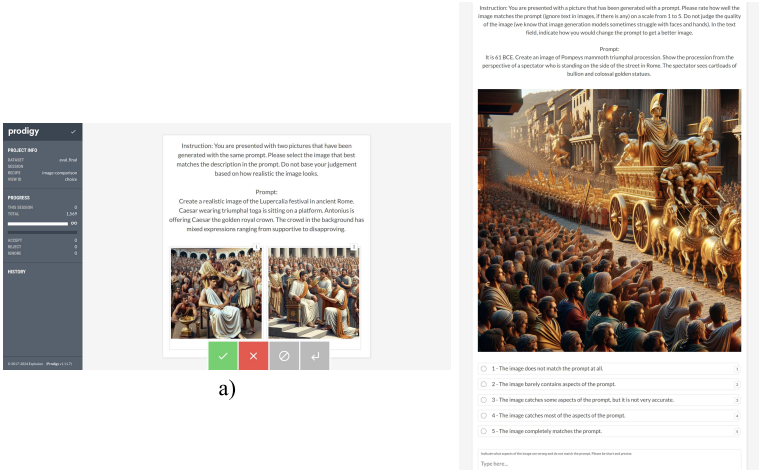

Figure 2: Parts of the Prodigy interface to obtain human assessments: **a)** the interface for image comparison with the side panel with an overview of how many image pairs have been annotated, **b)** the interface for the rating scenario with the 5-point Likert scale and a text comment field.

Table 1
Overview of results in the IC and the IR scenarios.

Evaluation scenario	Total assessments	Multiple annotations	Excluded	After exclusions
Image Comparison (IC)	1,569	103	64	1,505
Image Rating (IR)	568	29	24	544

4.1.2. Results of Human Evaluation

Table 1 gives an overview of the results from the human evaluation. In the IC setting, we received 1,569 comparisons. 103 samples were annotated more than once. For unknown reasons, 64 data points did not contain the human assessment, so we excluded them from further analysis. On average, each participant compared 74.71 (SD 43.32) image pairs.

The IR scenario received less feedback since the participants provided written feedback in a text field besides their rating. We obtained 568 ratings, of which 29 were double ratings—24 feedbacks without scores needed to be excluded.

We must note here that, due to a wrong parameter setting of Prodigy in both scenarios, the data samples to be evaluated were presented to the participants in sequential instead of a random order. This led to only marginal annotation overlap. For this reason, we cannot compute inter-annotator agreements (IAA) yet. However, since we plan to improve the models with the feedback obtained from the participants, we will have further evaluation rounds during which we can take care of this limitation. Still, to the best of our knowledge, this is the first "large-scale" evaluation campaign dedicated to historical image generation. We can still analyse and compare the results obtained with the limitations in mind (see Section 4.1.3).

However, since previous studies reported low IAA in human evaluation scenarios (cf. [16]),

Table 2
Agreement of human evaluation with GPT-4o's assessment.

Score Agreement	Count	Percentage
TRUE	864	57.41%
FALSE	641	42.59%
Total	1,505	

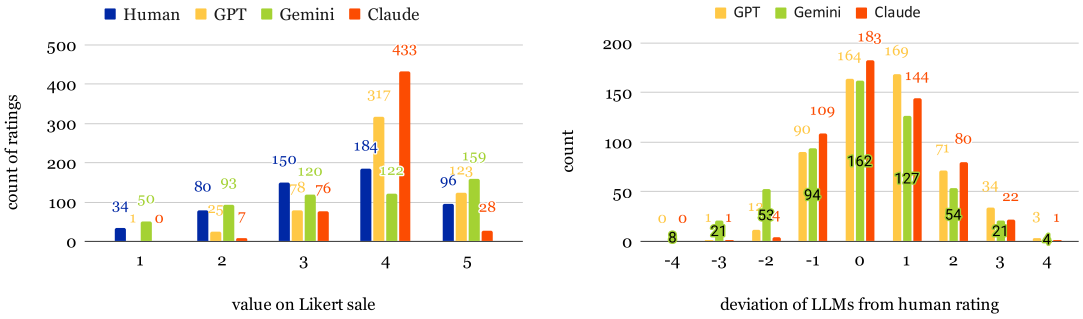

Figure 3: Left: Aggregation and comparison of scores of human ratings vs. LLM ratings. **Right:** Deviation of LLM scores from human ratings.

we hypothesise a similar outcome on our dataset.

4.1.3. Comparison of Human Results with Large Language Model (LLM) Evaluation

To mitigate the missing information on IAA and to evaluate the suitability of multimodal LLMs for scoring tasks, we employed GPT-4o [21], Gemini 1.5 Pro [26] and Claude 3.5 Sonnet.[6] We let the LLMs solve the same tasks as the annotators, i.e., we applied them to the IC (only GPT-4o) and IR (all three) evaluation scenarios.[7]

For IC, Table 2 shows the agreements of the human comparisons with GPT-4o's comparisons. We see that in 57.51% of the cases, human annotators and GPT-4o agree on which of the two images better corresponds to the prompt.

Figure 3 summarises the results for the IR setting. The left graph shows the differences between the human and the LLM ratings. The tendency is that LLMs rate images higher than human annotators. The right graph shows the LLM's deviations from the human scores. E.g., in 164 (30.15%) ratings, GPT-4o agrees with the human scores. In 169 (31.07%) cases, GPT-4o scores one point higher on the Likert scale than the human annotators (i.e., GPT-4o had rated an image a 3 when the human annotator rated it at 2). We see that Claude tends to rate images higher, especially. Overall, the deviations seem normally distributed, a fact that might be exploited for future evaluations.

Choosing two scenarios to evaluate allows us to test for differences in assessing images

[6]See https://www.anthropic.com/news/claude-3-5-sonnet.
[7]This generated costs of $19.14 for GPT-4o, $2.49 for Gemini and $5.27 for Claude.

Table 3
Data statistics and results of Welch's *t*-test.

Ratings	Human		GPT		Gemini		Claude	
	Triumph	*Lupercalia*	Triumph	*Lupercalia*	Triumph	*Lupercalia*	Triumph	*Lupercalia*
# of samples	404	140	404	140	404	140	404	140
Average score	3.46	3.31	4.09	3.68	3.60	3.03	3.91	3.81
SD	1.12	1.14	0.70	0.82	0.90	0.74	0.46	0.52
p-value	0.18		0.0000002		0.00004		0.052	

Table 4
Inter-annotator agreement between different groups using Krippendorff's alpha. We used the same 544 images for which we have computed the *t*-test..

	GPT vs. Gemini vs. Claude		GPT vs. human	
	Triumph	*Lupercalia*	Triumph	*Lupercalia*
α	0.079	-0.044	-0.008	-0.005

between the triumph and the *Lupercalia* scenario. Our null hypothesis H_0 is that there is no difference in the ratings of human annotators and, e.g., GPT-4o in the two historical scenarios. Table 3 shows the results of two Welch's *t*-test [31], which we chose because of (i) unequal variation and (ii) unequal sample sizes. For the human evaluation (unifying the assessment results but excluding invalid samples), the *p*-value does not allow us to reject H_0. The GPT and Gemini ratings show another picture. The *p*-values show a highly significant difference between ratings of the triumph and the *Lupercalia* images. Claude's *p*-value is on the brink of showing a statistically significant difference. The, on average, lower ratings by LLMs of the *Lupercalia* images could indicate DALL-E's difficulties in generating adequate imagery. Firstly, since the *Lupercalia* are not so much a described nor illustrated phenomenon, it is reasonable that images portraying the festival are not on the same standard as those generated for the triumphal procession. Secondly, the automatic evaluation poses problems for LLMs because they do not "know" as much as they do for the triumph.

Although we cannot provide IAA scores for the human evaluation yet, we can do so for the automatically generated ratings by the LLMs. Table 4 shows the results when we compare the ratings for the LLMs (again split into triumph- and *Lupercalia*-related scores). The scores are all around 0, indicating low overlap, IAA. Unifying all human scores and comparing them against the ratings obtained via GPT-4o also shows low overlap. These results hint at the very different rating "strategies" of the LLMs. We need further evaluation to shed more light on the origins of the discrepancies.

4.2. Automatic Evaluation

4.2.1. Automatic Evaluation Setup

For a further fully automatic evaluation procedure, we employed the Question Generation and Answering (QG/A) [11, 6] framework for automatic image evaluation. The first step in this

framework involves using a pre-trained language model to generate a set of questions based on a given prompt and question-generation instructions via few-shot learning. In the second step, a pre-trained multimodal model generates answers given the image and the generated set of questions.

Question Generation (QG) In our study, we utilised GPT-3.5 [5] for QG employing the Davidsonian Scene Graph (DSG) [6] method. DSG serves as an evaluation framework grounded in formal semantics. This method's main advantage is its ability to generate atomic and unique questions structured in dependency graphs, which (i) ensure comprehensive semantic coverage and (ii) avoid inconsistencies in responses. Cho, Hu, Garg, Anderson, Krishna, Baldridge, Bansal, Pont-Tuset, and Wang [6] empirically demonstrated that DSG addresses the challenges of hallucinations, duplications, and omissions in QG.

Visual Question Answering (VQA) We employed GPT-4o for the VQA task. The following prompt instruction guides the model: "You are a helpful assistant. Please answer the question only with 'Yes' or 'No'. Do not give other outputs. Question: `{question}`." To ensure precise control over the output, specifically responding with either 'Yes' or 'No', we set the parameter `logit_bias` to 100 for both 'Yes' and 'No' tokens. Logit bias modifies the likelihood of specified tokens appearing in the model-generated output. We also set the `top_p` (nucleus sampling) parameter to 0.1 to restrict the model's consideration to a subset of tokens (the nucleus) whose cumulative probability mass reaches a designated threshold (top-p). In the context of a 0.1 `top_p` setting, the model exclusively considers tokens constituting the top 10% of the probability mass for the subsequent token. The combination of `logit_bias` and `top_p` configurations enables the outputs to adhere to predefined patterns ('Yes' and 'No'), rendering the model more deterministic and particularly suitable for our image evaluation task.[8] We assign a score of 1 for 'Yes' and 0 for 'No' and then compute an average score for each image. We observe that GPT occasionally generates questions such as "Is there an image?" or "Can you visualize a scene?" which are invalid in our context, as the input consistently includes an image and a set of questions. We excluded the scores of these invalid questions from our analysis.

4.2.2. Results of VQA

Figure 4 shows a histogram of the results of the VQA scores for all 600 images. We find most scores between 0.5 and 0.9, with over 60 images obtaining a perfect score of 1. This means that each 'Yes-or-No' question was answered with 'Yes'. When we look at three results as presented in Figure 5 in Appendix A, we find that VQA attributes a low score of 0.05 for image **a)**. The human evaluator and GPT, however, have scored this image with a 4 in the IR scenario. In **b)**, we have a medium VQA score of 0.61, a human score of 5 and a GPT score of 4. Lastly, **c**) shows an image with a VQA score of 1, but a human annotator scored this image a 3 and GPT a 4. We already see discrepancies between the different scores from these three examples only. A comparison of VQA between the 450 images from the triumphal procession and the 150 images

[8]This evaluation scenario cost us $7.69. The whole experiment, i.e., image generation, LLM evaluation in the two scenarios from Section 4.1.3 and the one mentioned in this section totalled at $80.67.

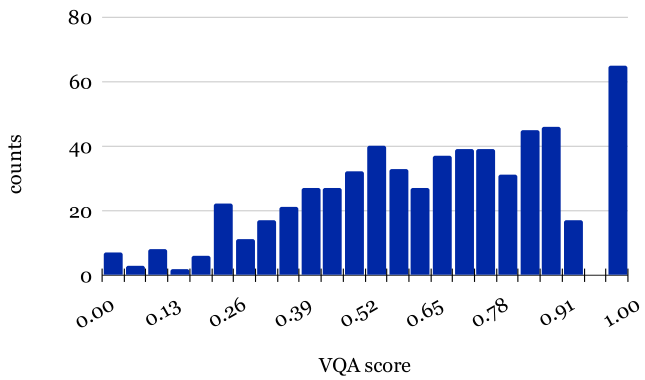

Figure 4: Histogram of scores obtained with the VQA evaluation.

from the *Lupercalia* based on Welch's t-test shows no significant differences between the two ratings ($p = 0.88$). From this, we conclude that ratings based on VQA produce more reliable results than those produced with a Likert scale.

5. Limitations and Outlook

The most significant limitation of our work is the missing IAA scores. For future evaluation rounds, we will set up the evaluation to allow for their computation. In this way, we get reliable measures of how demanding the task of assessing the alignment of historical images and the prompts they produced is. However, we argue that the results we obtained from the human evaluation are still valuable and allow for fine-tuning models based on human feedback (preferences in the IC and textual input in the IR scenario), albeit in a low-resource setting.

Moreover, we will employ more models to generate images in future experiments. This approach enables us to decide which models are the most suitable for historical image generation. The stable prompt base also allows for comparable results. Still, the significant number of images we will generate in future endeavours also calls for automatic evaluation methods.

6. Conclusion

In conclusion, our study provides valuable insights into the potential and challenges of using AI for historical image generation. The evaluation of 600 AI-generated images of triumphal processions and the *Lupercalia* revealed both promising capabilities and significant limitations.

Our findings hint at the discrepancies between human and automated assessments, underscoring the complexity of evaluating historical accuracy in AI-generated imagery. Ultimately, this study serves as a stepping stone towards more sophisticated use of AI in historical recreation and education while cautioning against over-reliance on automated systems for historical interpretation.

This research contributes a novel dataset and evaluation framework to the field, enabling future studies. As AI continues to evolve, our work suggests that while it holds promise for enhancing historical visualisation and understanding, it requires careful human oversight and interpretation.

References

[1] A. Algül. "The Roman Triumph: Participation, Historiography and Remembrance". In: (2018). URL: https://www.academia.edu/43295099/The%5C%5FRoman%5C%5FTriumph%5C%5FParticipation%5C%5FHistoriography%5C%5Fand%5C%5FRemembrance.

[2] J. Armstrong. "Claiming Victory: The Early Roman Triumph". In: *Spalinger, Anthony John 1947- (edt), Armstrong, Jeremy (edt), Rituals of triumph in the Mediterranean world*. Culture and History of the Ancient Near East 63. Leiden u.a.: Brill, 2013, pp. 7–22.

[3] M. Beard. *The Roman Triumph*. Harvard University Press, 2007.

[4] J. Betker, G. Goh, L. Jing, T. Brooks, J. Wang, L. Li, L. Ouyang, J. Zhuang, J. Lee, Y. Guo, et al. "Improving Image Generation with Better Captions". In: (2023). URL: https://cdn.openai.com/papers/dall-e-3.pdf.

[5] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. "Language Models are Few-Shot Learners". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., 2020, pp. 1877–1901. URL: https://proceedings.neurips.cc/paper%5C%5Ffiles/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[6] J. Cho, Y. Hu, R. Garg, P. Anderson, R. Krishna, J. Baldridge, M. Bansal, J. Pont-Tuset, and S. Wang. "Davidsonian Scene Graph: Improving Reliability in Fine-grained Evaluation for Text-to-Image Generation". In: (2023). URL: http://arxiv.org/abs/2310.18235.

[7] D. Š. Erker. "Das Lupercalia-Fest im augusteischen Rom: Performativität, Raum und Zeit". In: *Archiv für Religionsgeschichte* 11.1 (2009), pp. 145–178. DOI: doi:10.1515/9783110208962.2.145.

[8] M. W. Fareed, A. Bou Nassif, and E. Nofal. "Exploring the Potentials of Artificial Intelligence Image Generators for Educating the History of Architecture". In: *Heritage* 7.3 (2024), pp. 1727–1753. DOI: 10.3390/heritage7030081.

[9] "Der römische Triumph in Prinzipat und Spätantike: Probleme – Paradigmen – Perspektiven". In: *Der römische Triumph in Prinzipat und Spätantike*. Ed. by F. Goldbeck and J. Wienand. Berlin, Boston: De Gruyter, 2017, pp. 1–26. DOI: doi:10.1515/9783110448009-003.

[10] D. Guarisco. "Augustus, the Lupercalia and the Roman identity". In: *Acta Antiqua Academiae Scientiarum Hungaricae* 55.1-4 (2015), pp. 223–228. DOI: 10.1556/068.2015.55.1-4.16. URL: https://akjournals.com/view/journals/068/55/1-4/article-p223.xml.

[11] Y. Hu, B. Liu, J. Kasai, Y. Wang, M. Ostendorf, R. Krishna, and N. A. Smith. "TIFA: Accurate and Interpretable Text-to-Image Faithfulness Evaluation with Question Answering". In: (2023). URL: http://arxiv.org/abs/2303.11897.

[12] C. Lange. "Mock the Triumph: Cassius Dio, Triumph and Triumph-Like Celebrations". In: *Cassius Dio*. Brill's Historiography of Rome and Its Empire Series. Brill, 2016, pp. 92–114. DOI: 10.1163/9789004335318_007.

[13] C. H. Lange. "The Late Republican Triumph: Continuity and Change". In: *Der römische Triumph in Prinzipat und Spätantike*. Berlin, Boston: De Gruyter, 2017, pp. 29–58. DOI: doi:10.1515/9783110448009-004. URL: https://doi.org/10.1515/9783110448009-004.

[14] C. Lange. "The Triumph outside the City: Voices of Protest in the Middle Republic". In: *The Roman Republican Triumph*. Ed. by C. Hjort Lange and F. Vervaet. Analecta Romana Instituti, Suppl. Quasar, 2014, pp. 67–81.

[15] C. H. Lange. "Triumph and Civil War in the Late Republic". In: *Papers of the British School at Rome* 81 (2013), pp. 67–90. DOI: 10.1017/s0068246213000056.

[16] T. Lee, M. Yasunaga, C. Meng, Y. Mai, J. S. Park, A. Gupta, Y. Zhang, D. Narayanan, H. Teufel, M. Bellagente, M. Kang, T. Park, J. Leskovec, J.-Y. Zhu, F.-F. Li, J. Wu, S. Ermon, and P. S. Liang. "Holistic Evaluation of Text-to-Image Models". In: *Advances in Neural Information Processing Systems*. Ed. by A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine. Vol. 36. Curran Associates, Inc., 2023, pp. 69981–70011. URL: https://proceedings.neurips.cc/paper%5C%5Ffiles/paper/2023/file/dd83eada2c3c74db3c7fe1c087513756-Paper-Datasets%5C%5Fand%5C%5FBenchmarks.pdf.

[17] Y. Liang, J. He, G. Li, P. Li, A. Klimovskiy, N. Carolan, J. Sun, J. Pont-Tuset, S. Young, F. Yang, J. Ke, K. D. Dvijotham, K. M. Collins, Y. Luo, Y. Li, K. J. Kohlhoff, D. Ramachandran, and V. Navalpakkam. "Rich Human Feedback for Text-to-Image Generation". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2024, pp. 19401–19411. URL: https://openaccess.thecvf.com/content/CVPR2024/papers/Liang%5C%5FRich%5C%5FHuman%5C%5FFeedback%5C%5Ffor%5C%5FText-to-Image%5C%5FGeneration%5C%5FCVPR%5C%5F2024%5C%5Fpaper.pdf.

[18] J. Madsen. "The Loser's Prize: Roman Triumphs and Political Strategies during the Mithridatic Wars". In: *The Roman Republican Triumph Beyond the Spectacle*. Analecta Romana Instituti Danici. Supplementa Xlv. Quasar, 2014, pp. 117–130.

[19] P. F. Mittag. "Die Triumphatordarstellung auf Münzen und Medaillons in Prinzipat und Spätantike". In: *Der römische Triumph in Prinzipat und Spätantike*. Berlin, Boston: De Gruyter, 2017, pp. 419–452. DOI: doi:10.1515/9783110448009-017.

[20] J. A. North. "Caesar at the Lupercalia". In: *Journal of Roman Studies* 98 (2008), pp. 144–160. DOI: 10.3815/007543508786239210.

[21] OpenAI. *Hello GPT-4o*. 2024. URL: https://openai.com/index/hello-gpt-4o.

[22] I. Östenberg. *Staging the World: Spoils, Captives, and Representations in the Roman Triumphal Procession*. Oxford: Oxford University Press, 2009. DOI: 10.1093/acprof:oso/9780199215973.001.0001.

[23] I. Östenberg. “Triumph and spectacle. Victory celebrations in the Late Republican civil wars”. In: *The Roman Republican Triumph Beyond the Spectacle*. 2014, pp. 181–193.

[24] M. Otani, R. Togashi, Y. Sawai, R. Ishigami, Y. Nakashima, E. Rahtu, J. Heikkilä, and S. Satoh. “Toward Verifiable and Reproducible Human Evaluation for Text-to-Image Generation”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 14277–14286. URL: https://cvpr2023.thecvf.com/virtual/2023/poster/22014.

[25] M. L. Popkin. *The Architecture of the Roman Triumph: Monuments, Memory, and Identity*. Cambridge University Press, 2016.

[26] M. Reid, N. Savinov, D. Teplyashin, D. Lepikhin, T. Lillicrap, J.-b. Alayrac, R. Soricut, A. Lazaridou, O. Firat, J. Schrittwieser, et al. “Gemini 1.5: Unlocking Multimodal Understanding Across Millions of Tokens of Context”. In: *arXiv preprint arXiv:2403.05530* (2024). DOI: https://doi.org/10.48550/arXiv.2403.05530.

[27] S. T. Schipporeit. “Wege des Triumphes. Zum Verlauf der Triumphzüge im spätrepublikanischen und augusteischen Rom”. In: *Triplici invectus triumpho : Der römische Triumph in augusteischer Zeit.* (2008), pp. 95–136. URL: https://zenon.dainst.org/Record/001069375.

[28] N. Stiennon, L. Ouyang, J. Wu, D. Ziegler, R. Lowe, C. Voss, A. Radford, D. Amodei, and P. F. Christiano. “Learning to Summarize with Human Feedback”. In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc., 2020, pp. 3008–3021. URL: https://proceedings.neurips.cc/paper%5C%5Ffiles/paper/2020/file/1f89885d556929e98d3ef9b86448f951-Paper.pdf.

[29] P. Tennant. “The Lupercalia and the Romulus and Remus Legend”. In: *Acta Classica* 31 (1988), pp. 81–93. URL: http://www.jstor.org/stable/24591847.

[30] L. Webb and L. Brännstedt. “Gendering the Roman Triumph: Elite Women and the Triumph in the Republic and Early Empire”. In: *Gendering Roman Imperialism*. Leiden, The Netherlands: Brill, 2022, pp. 58–95. DOI: 10.1163/9789004524774_005.

[31] B. L. Welch. “The Generalization of Student’s Problem when Several Different Population Variances are Involved”. In: *Biometrika* 34.1-2 (1947), pp. 28–35. DOI: 10.1093/biomet/34.1-2.28. URL: https://doi.org/10.1093/biomet/34.1-2.28.

[32] K.-W. Welwei. “Das Angebot des Diadems an Caesar und das Luperkalienproblem”. In: *Historia: Zeitschrift für Alte Geschichte* 16.1 (1967), pp. 44–69. URL: http://www.jstor.org/stable/4434966.

[33] J. Xu, X. Liu, Y. Wu, Y. Tong, Q. Li, M. Ding, J. Tang, and Y. Dong. “ImageReward: Learning and Evaluating Human Preferences for Text-to-Image Generation”. In: *Advances in Neural Information Processing Systems*. Ed. by A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine. Vol. 36. Curran Associates, Inc., 2023, pp. 15903–15935. URL: https://proceedings.neurips.cc/paper%5C%5Ffiles/paper/2023/file/33646ef0ed554145eab65f6250fab0c9-Paper-Conference.pdf.

A. Additional Figures

a) b) c)

Figure 5: Examples of the VQA ratings. **a)** from the triumphal procession based on Mittag [19], scored 0.05 based on 22 questions, prompt: *"There exist coins minted in 326 CE which show Emperor Constantinus I. on an elephant quadriga during the celebrations of his viceannalia (20 years on the throne). Although textual sources do not confirm that elephant quadrigas were in use, create an image that shows Constantinus I. together with his son Constantius II. on a chariot pulled by four elephants during the vicennalia in Nicomedia. The chariot is accompanied by two lictores. The elephants are guided by Mahouts and Constantinus the I. wears the laurel wreath."*, scored a 4 by both human evaluators and GPT, **b)** from the *Lupercalia* based on Erker [7], scored 0.61 based on 18 questions, prompt: *"Create an image that shows high-ranking magistrates of ancient Rome, dressed in loincloths. They are emerging from a cave of the Paletine Hill to start the traditional run of the Lupercalian festival. They are running on a rugged terrain under a blue sky."*, scored a 5 by a human annotator and a 4 by GPT, **c)** from the triumphal procession based on Madsen [18], scored 1.00 based on 19 questions, prompt: *"Create a historical image of the spectacle of Pompey's triumph in 61 BC. Pompey adorned in triumphal regalia, parades through the streets of Rome atop his chariot, with captured treasures and defeated foes on display. Imagine the jubilation among the crowds as they celebrate Pompey's military prowess and the expansion of Roman territories under his command."*, scored a 3 by a human annotator and a 4 by GPT.