Combining LLM and Topic Modeling for Automated Video Analysis: A Case Study

Howanitz, Gernot
gernot.howanitz@uibk.ac.at
Universität Innsbruck, Österreich

Kaltseis, Magdalena
magdalena.kaltseis@uibk.ac.at
Universität Innsbruck, Österreich

Sulzhytski, Ilya
ilya.sulzhytski@uibk.ac.at
Universität Innsbruck, Österreich

Introduction

With the advancement of computer vision techniques, digital humanities scholars are increasingly interested in the study of visual data, including images (O'Halloran et al., 2014; Maiwald et al., 2017; Emanuel, 2018; Munster et al., 2019) and videos (Kuhn et al., 2014; Jakubowski et al., 2017; Bakels et al., 2020; Pustu-Iren et al., 2020; El-Keilany et al., 2022). The possibilities for using computational methods in the digital humanities have expanded considerably with the recent boom in generative AI, including monomodal textual Large Language Models (LLMs) and Multimodal Large Language Models (MLLMs) (Aguiar & Araújo, 2024; Chun & Elkins, 2023; Guo, 2024).

In this respect, we agree with Thomas Smits and Melvin Wevers that “multimodal models have the potential to cause a multimodal turn in DH research” (Smits & Wevers, 2023, p. 1268). However, despite the extensive capabilities associated with image understanding tasks offered by recent MLLMs, we know little about how non-human agents ‘view’ visual media (Arnold & Tilton, 2023, pp. 9–31) and whether the results of such ‘viewing’ can be used to understand images and videos with complex cultural, social and behavioural contexts and dynamics.

With this complexity in mind, we analyse the capabilities of MLLMs in understanding video content within Taylor Arnold’s and Lauren Tilton’s concept of ‘Distant Viewing’ (Arnold & Tilton, 2023). We used two multimodal MLLMs: the closed-source *GPT-4o mini* from OpenAI (OpenAI, 2024) and the open-source *Large Language and Vision Assistant* (LLaVA, Liu et al., 2023). Both models can process textual and visual data; and while they cannot work directly with the video format, they can produce aggregated descriptions by splitting videos into sequences of frames.

In this paper, we present our case study that focuses on the visual and narrative features of Olga Abramchik’s 2021 documentary *Мы не знали друг друга до этого лета* [*We Didn't Know Each Other Before This Summer*] (Abramchik, 2021) by using two different MLLMs. First, we introduce our case study. Afterwards, we describe the image analysis pipeline that combines MLLM annotation with topic modelling. We then discuss our main findings by comparing the annotation results of the two models, focusing on the interpretability of the topics extracted from these annotations. Finally, we discuss the strengths and limitations of both MLLMs in the image annotation task and their use in understanding the analysed video as a whole.

Multimodal and Video Large Language Models

As outlined in Section 1, recent advances in LLMs have led to the development of MLLMs, which combine the reasoning of LLMs with visual processing (Li et al., 2023; Yin et al., 2023). Notable MLLMs with image-to-text capabilities such as GPT-4o / GPT-4o mini (OpenAI, 2024), LLaVA (Liu et al., 2023), Llama 3 (Dubey et al., 2024), Pixtral (Agrawal et al., 2024) and others have demonstrated significant progress in image understanding through various architectural innovations and training strategies (Yin, et al., 2023).

The evolution from image-focused MLLMs to LLMs with the ability of video understanding (Vid-LLMs) advances multimodal reasoning and addresses challenges such as temporal dynamics and long-range context dependencies. Recent models in this area include Video-LLaMA (Zhang et al., 2023) and Video-ChatGPT (Maaz et al., 2023). According to Tang et. al, Vid-LLMs deal with three main tasks: abstract understanding, temporal understanding and spatiotemporal understanding (Tang et al., 2023, pp. 4–5). Currently, the main challenge lies in understanding long context video data (Tang et al., 2023, pp. 13–14, Yin, et al., 2023, p. 13).

In this regard, despite the recent developments in Vid-LLMs, the analysis of videos through frame-based processing using image-focused MLLMs remains a compelling alternative to end-to-end video understanding (Huang, et al., 2024). Although Vid-LLMs theoretically provide more complete temporal analyses for short videos, they are computationally expensive for long videos and have problems with long-range dependencies (Tang et al., 2023, pp. 13–14). In this regard, Meinardus et al. demonstrated the efficacy of image-text MLLM for video moment retrieval (Mr. BLIB), emphasising the benefits of a frame-based approach without the need for resource-intensive pretraining, precise localisation of relevant moments, and flexible processing of longer videos through the use of key frames (Meinardus, et al., 2024).

Therefore, frame-based analysis remains a more reliable approach for long video understanding. The following sections (3 and 4) will provide a more detailed examination of this approach, with a demonstration of how the complex visual dynamic of protest videos can be effectively captured through the analysis of sequential frames and the synthesis of the resulting data using topic modelling.

Methodology

General considerations

In building our analysis pipeline, we had to consider several issues. First, *LLMs are rarely trained to work directly on video input*, so we decided to operate on individual frames rather than on the video as a whole. In particular, Chen (2023) suggests this approach for analysing video with the GPT-4 model family. Second, *closed-source LLMs often outperform open-source LLMs*. While many promising multimodal networks are available online, the GPT-4 model family still outperforms them on various metrics (Fu et al., 2023; Yin et al., 2023; Fu et al., 2024), but it uses a subscription model and has to be paid for. Therefore, we decided to compare the performance of GPT-4o mini with that of an open-source network. Third, *automated image analysis still has substantial hardware requirements*, most of which are not feasible to run outside of HPC clusters. After some experimentation, we chose LLaVA-v1.5 with 13 billion parameters and 4-bit encoding, as this network offers reasonable performance during inference and can be run on a single A100 GPU. Fourth, *LLMs are notoriously black boxes*, so we used them as little as possible. While LLMs can be used to create full-text descriptions of videos (e.g. Chen, 2023), we limited the automated work to keyword annotation. This approach also allowed us to compare and process the annotations more easily.

Sample Case

This article is part of the project ‘Kaleidoscopic Patterns of Protest’, which analyses visual and textual (self-)representations of East Slavic protest cultures. As an ideal candidate for testing a video annotation task that would meet the project’s goals, we chose the documentary *We Didn't Know Each Other Before This Summer* (Abramchik 2021). The film focuses on the protests in Belarus following the presidential elections on 9 August 2020. It provides a detailed audiovisual documentation of symbols, behaviours, police actions and other features of the protests. Constructed from footage of participants and witnesses, the film offers an unfiltered, day-by-day narrative of collective actions and police responses, without outside commentary or music. It also highlights the regime’s human rights abuses and public reactions. The documentary offers a valuable resource for studying visual representations of protests and testing LLMs ability to annotate such material, as it provides unique, unedited grassroots footage with uninterrupted protest action sequences.

Image analysis pipeline

We used GPT-4o mini and LLaVA to automatically annotate the documentary. First, we extracted one frame per second from the documentary using the OpenCV library for Python. Next, we evaluated 20 different prompt variations for image annotation using a random sample of 30 video frames. Through manual evaluation of descriptive accuracy and contextual completeness, we created two optimal prompts for each model. These prompts can be considered zero-shot prompts (Kojima et al., 2022), which provide no contextual information, as we wanted the LLM to ‘interpret’ as little as possible and instead describe the visual images in general terms:

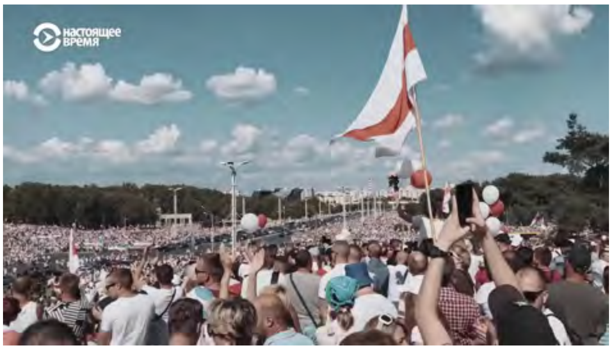

Table 1. Comparison of prompts (top) and keywords (below) extracted by GPT-4o mini (left) and LLaVA (right) for the sample image shown above (Source: *We Didn't Know Each Other Before This Summer*, 00:56:44)

GPT-4-o mini	LLaVA
Prompt: Please conduct an exhaustive analysis of the provided image. Focus on identifying and listing every visible element in the image. Write a description in the form of keywords. Be specific and extract as many unique keywords as possible. Output format: keyword 1, keyword 2, …	*Prompt:* List every visible element in the image. Do not use full sentences, use only keywords. Use as many keywords as necessary, but not more than 20.
Result: crowd, protest, flag, Belarusian flag, people, balloons, blue sky, clouds, city skyline, trees, cheering, smartphones, outdoor event, summer, cultural gathering, democratic movement, gathering, celebration, unity, red and white colors	*Result:* people, flags, banners, cars, trucks, buses, boats, kites, balloons, cell phones, handbags, backpacks, ties, sunglasses, hats, shirts, pants, towels, umbrellas, benches

We used OpenAI’s Batch API (OpenAI, n.d.) to obtain the list of keywords for each frame generated by GPT-4o mini; LLaVA, on the other hand, was run locally. As a result, we extracted 3730 frames from the documentary and obtained two lists of keywords for each frame. We then used topic modelling to process the keywords. This approach avoided some of the pitfalls of LLMs as black boxes, because topic modelling is an established and mathematically proven technique. We used Gensim’s implementation of Ensemble LDA (Brigl, 2019) to run multiple topic models simultaneously, keeping only the topics consistent across several topic models, thereby reducing the number of incoherent topics.

The keywords for a single frame were then considered a ‘document’; each keyword (which could consist of compounds, e.g., ‘army truck’) was considered a ‘word’. As a control for the method and the automated analysis, we did a close reading of the topics. We then applied the topic model to each list of keywords, resulting in a topic distribution for each frame. For a larger video corpus, each video could be identified by its topic distribution. This makes the videos easily comparable, using established distance metrics, such as PCA or Isomap, to map them (Howanitz, 2020, 92–107).

Results

A first-look comparison of the annotations showcased clear differences in the keyword generation patterns of the two models. While GPT-4o mini provided more consistent and typically more detailed annotations, LLaVA showed a higher variability in the results and a capacity for both very sparse and extremely wordy annotations. In terms of annotation quality, further human observations revealed more nuanced differences between the two models. GPT-4o mini showed higher contextual accuracy, correctly identifying and including event-specific keywords such as 'Minsk', demonstrating its ability to accurately capture the implicit context of the images. In addition, GPT-4o mini excelled in OCR, particularly in recognising and correctly interpreting text within images, especially Cyrillic script. In contrast, LLaVA struggled with Cyrillic letters, often failing to recognise or accurately annotate the Russian or Belarusian text. Both models have difficulties with blurry night images due to poor lighting and low visibility.

Topic modelling revealed further differences in the interpretability of annotations between the two models, especially when understanding video content from aggregated descriptions of individual frames. For example, both GPT-4o mini and LLaVA produced an ‘intertitle’ topic. However, GPT-4o mini was more specific and detailed, identifying the actual logo of the YouTube channel (“current time” [настоящее время]) with keywords such as ‘black background’, ‘white text’, ‘Russian language’, ‘Cyrillic script’, ‘logo’, ‘настоящее время logo’ or ‘political content”. In contrast, LLaVA produced more common and less diverse keywords, such as ‘Russian’, ‘text’, ‘foreign language’, ‘black’ or ‘foreign’, and also contained errors (e.g. by identifying the Russian text as ‘Bulgarian’). While the keywords produced by GPT-4o mini were more concise and allowed a better interpretation of the topic, the disadvantage of LLaVA was somewhat mitigated when we applied the topic model to the whole documentary. Both ‘intertitle’ topics allowed us to immediately identify the intertitles in the documentary and, thus, served their purpose well (see Figure 1).

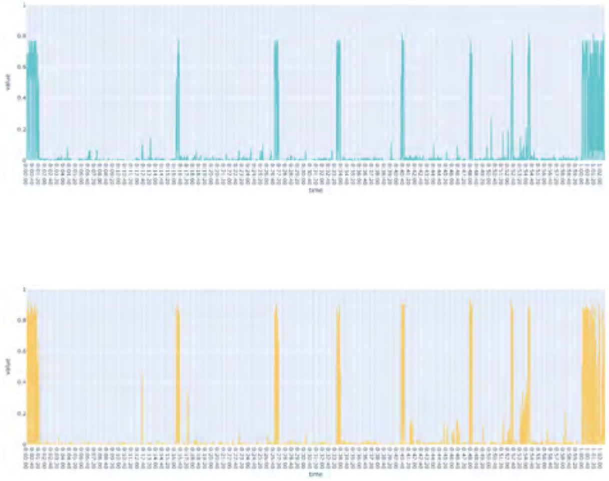

Figure 1. Distribution of GPT-4o mini ‘intertitle’ topic (top) and LLaVA ‘intertitle’ topic (bottom) for the documentary. X-axis is time, y-axis is topic probability. The intertitles are clearly visible and agree for both LLMs, e.g. at the beginning and the end of the documentary and around min. 00:16:00.

Topic modelling also revealed thematic trends, for example, when it produced similar police-related topics in both models (see Table 2). Keywords pertaining to riot control police are combined with ‘nighttime’ (GPT-4o mini) or ‘night’ (LLaVA). This combination is significant because it was mainly at night that police brutality occurred during the protests, while peaceful demonstrations happened during daytime—a fact that the documentary shows very clearly. Therefore, both models recognized this contrast in protest actions, which changed dramatically depending on the time of the day.

Table 2. Two topics about police and violent protests during the nighttime, based on GPT-4o mini annotations (left) and LLaVA annotations (right).

GPT-4o mini	LLaVA
law enforcement, crowd control, pavement, urban environment, outdoor, tension, uniform, public space, protective gear, urban setting, uniformed officers, street, surveillance, building, public safety, riot gear, **nighttime**, confrontation, body armor, intervention	police, security, crowd, protest, law enforcement, safety, demonstration, **night**, public safety, man, riot gear, helmet, car, public order, fence, crowd control, police officers, police car, riot, gathering

We can also see a narrative pattern: If we plot the topic distribution for the ‘riot police by night’ topics, we see that night shots (and police brutality) are more present in the first half of the documentary and then gradually decrease in the second half (see Figure 2). However the topic distributions do not match as clearly as in the case of the ‘intertitle’ topics in Figure 1. This could be due to the fact that the topic ‘riot police by night’ is not as clearly delineated as the ‘intertitle’ topics and is also less visually distinguishable as the documentary also features night shots without any police.

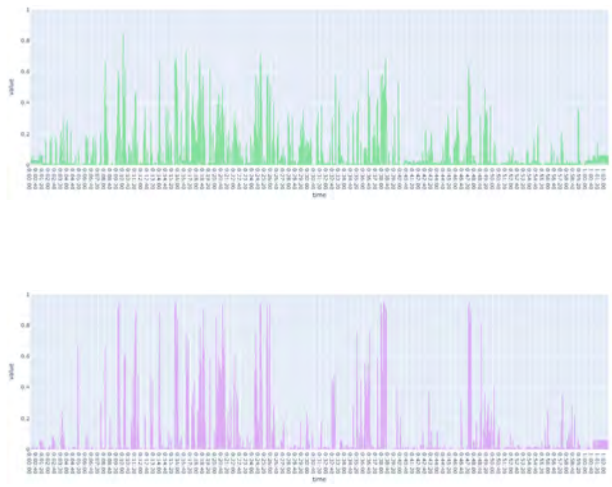

Figure 2. ‘Riot police by night’ topic distribution, based on keywords by GPT-4o mini and LLaVA (bottom).

Discussion

The proposed combination of zero-shot LLM annotation and topic modelling has several advantages. On a technical level, it allows the automatic analysis of a video corpus with reasonable resources. Furthermore, GPT-4o mini produces concise annotations but is too expensive for several thousand videos. Compared to GPT-4o mini, the LLaVA annotations proved not as accurate, but this difference did not affect the efficiency of our pipeline, precisely the output of the Ensemble LDA topic modelling. Another inconvenience of LLaVA is its relatively low speed: processing the 60-minute documentary took approximately 72 hours on a single A100 GPU, while GPT-4o mini completed the same task in only 3 hours and 30 minutes through the cloud API accessed via Jupyter Notebook. Possible speedups include reducing the frame size from Full HD to a lower resolution and using more focused frame sampling techniques.

The advantage of the proposed combination on a conceptual level is that we did not have to introduce contextual information at an early stage and were therefore less prone to overlook certain phenomena since we did not precondition the pipeline to find only what we expected, which is a typical and bias-prone strategy in object detection/recognition. Notwithstanding the existence of a number of quantitative metrics and benchmarks for the evaluation of multimodal models with image understanding capabilities (Yang et al., 2023; Bubeck et al., 2023; Wang et al. 2023; Yue et al. 2024), there is a notable lack of quantitative metrics, benchmarks, or datasets that have been specifically designed for the assessment of the performance of these models in the analysis of protest images, particularly those from Eastern Europe. In light of the aforementioned gap, a promising alternative approach is to employ grounded theory to identify potential errors, hallucinations, and ambiguities that may arise when using MLLMs to describe protest images using keywords. We plan to further develop this idea in accordance with the recent paper of Hwang et al. (2023). Adapting this framework for analysing protest images will allow us to systematically uncover the strengths and limitations of multimodal models in our particular task. After these improvements, we aim to test the method on a larger corpus (~1000 videos) and quantitatively compare the videos based on their topic distributions to uncover narrative strategies based on visual content.

The presented approach has its limitations in that it focuses on basic content labelling by LLMs and therefore captures only surface-level visual information. It does not address deeper aspects of cinematic language, such as image syntax, semantics and compositional elements. This early-stage work is therefore positioned as a first step in automated analysis of protest images and videos, and as contributing to the broader discussion of computational approaches to understanding visual narratives. In the future, we aim to explore the integration of formal visual elements and more complex multimodal analysis techniques (Sommer, 2021) into a coherent analytical framework.

Acknowledgements

The work on this paper was financed by the Austrian Academy of Sciences’ *go!digital 3.0* program.

Data availability

Data and scripts to reproduce the results of this paper are available on Github: https://github.com/ghowa/llm-and-topic-modeling.

Bibliographie

Abramchik, Olga. 2021. "We Didn't Know Each Other Before This Summer [My ne znali drug druga do etogo leta]." Belarus: Current Time [Nastoiashchee Vremia]. https://www.youtube.com/watch?v=6vU9GtE75ZA

Agrawal, Pravesh , Szymon Antoniak , Emma Bou Hanna , Baptiste Bout , Devendra Chaplot , Jessica Chudnovsky, ... and Sophia Yang. 2024. “Pixtral 12B.” *arXiv preprint* arXiv:2410.07073.

Aguiar, Micaela, and Sílvia Araújo. 2024. “Final Thoughts: Digital Humanities Looking at Generative AI.” In *Digital Humanities Looking at the World: Exploring Innovative Approaches and Contributions to Society*, 367-380. Cham: Springer Nature Switzerland.

Arnold, Taylor, and Lauren Tilton. 2023. *Distant Viewing: Computational Exploration of Film and Media*. MIT Press. https://direct.mit.edu/books/oa-monograph/5674/Distant-ViewingComputational-Exploration-of.

Bakels, Jan-Hendrik, Matthias Grotkopp, Thomas Scherer and Jasper Stratil. “Matching Computational

Analysis and Human Experience: Performative Arts and the Digital Humanities.” *Digit. Humanit. Q.* 14 (2020): n. Pag.

Brigl, Tobias. 2023. “Extracting Reliable Topics Using Ensemble Latent Dirichlet Allocation.” *ResearchGate.*

Bubeck, Sébastien, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, ... and Yi Zhang. 2023. “ Sparks of artificial general intelligence: Early experiments with gpt-4.” *arXiv preprint arXiv:2303.12712.*

Chen, Kai. 2023. “Processing and Narrating a Video with GPT's Visual Capabilities and the TTS API.” *OpenAI Cookbook*. https://cookbook.openai.com/examples/gpt_with_vision_for_video_understanding.

Chun, Jon, and Katherine Elkins. 2023. “The Crisis of Artificial Intelligence: A New Digital Humanities Curriculum for Human-Centred AI.” *International Journal of Humanities and Arts Computing* 17 (2): 147-167.

Current Time TV. 2021. “Мы не знали друг друга до этого лета [We Did Not Know Each Other Before This Summer].” *Current Time*. Accessed February 5, 2021. https://www.currenttime.tv/a/my-ne-znali-drug-druga-do-etogo-leta-premiera/31091846.html.

Dubey, Abhimanyu, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, ... and Raj Ganapathy. 2024. “The llama 3 herd of models.” *arXiv preprint arXiv:2407.21783.*

Emanuel, Jeffrey P. 2018. “Stitching Together Technology for the Digital Humanities with the International Image Interoperability Framework (IIIF).” In *Digital Humanities, Libraries, and Partnerships*, 125-135. Chandos Publishing.

El-Keilany, Alina, Thomas Schmidt, and Christian Wolff. 2022. “Distant Viewing of the Harry Potter Movies via Computer Vision.” In *DHNB 2022,* 33-49. https://ceur-ws.org/Vol-3232/paper03.pdf

Fu, Chaoyou, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and others. 2023. “MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models.” In *arXiv preprint arXiv:2306.13394*.

Fu, Chaoyou, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, and others. 2024. “Video-MME: The First-Ever Comprehensive Evaluation Benchmark of Multi-Modal LLMs in Video Analysis.” In *arXiv preprint arXiv:2405.21075*.

Guo, Qiang. 2024. “Prompting Change: ChatGPT’s Impact on Digital Humanities Pedagogy–A Case Study in Art History.” *International Journal of Humanities and Arts Computing* 18 (1): 58-78.

Huang, Suyuan, Haoxin Zhang , Yan Gao , Yao Hu , and Zengchang Qin. 2024. “From Image to Video, what do we need in multimodal LLMs?.” *arXiv preprint arXiv:2404.11865* .

Howanitz, Gernot. 2020. *Leben Weben. (Auto-)Biographische Praktiken Russischer Autorinnen und Autoren im Internet*. Transcript. https://www.transcript-verlag.de/978-3-8376-5132-4/leben-weben/?number=978-3-8394-5132-8.

Hwang, Alyssa, Andrew Head and Chris Callison-Burch. 2023. “Grounded Intuition of GPT-Vision's Abilities with Scientific Images.” *arXiv preprint arXiv:2311.02069* .

Jakubowski, Kelly, Tuomas Eerola, Paolo Alborno, Gualtiero Volpe, Antonio Camurri, and Martin Clayton. 2017. “Extracting Coarse Body Movements from Video in Music Performance: A Comparison of Automated Computer Vision Techniques with Motion Capture Data.” *Frontiers in Digital Humanities* 4: 9.

Kojima, Takeshi, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, Yusuke Iwasawa. 2022. "Large Language Models Are Zero-Shot Reasoners." *Advances in Neural Information Processing Systems* 35: 22199-22213.

Kuhn, Virginia, Michael Simeone, Luigi Marini, Dave Bock, Alan B. Craig, Liana Diesendruck, and Sandeep Puthanveetil Satheesan. 2014. “MOVIE: Large Scale Automated Analysis of MOVing ImagEs.” In *Proceedings of the 2014 Annual Conference on Extreme Science and Engineering Discovery Environment*, 1-3.

Li, Chunyuan, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, ... and Jianfeng Gao. 2024. “Llava-med: Training a large language-and-vision assistant for biomedicine in one day.” *Advances in Neural Information Processing Systems , 36* .

Li, Junnan, Li, Dongxu Li , Silvio Savarese , and Steven Hoi. 2023. “Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models.” In *International conference on machine learning* (pp. 19730-19742). PMLR.

Liu, Haotian, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023. “Improved Baselines with Visual Instruction Tuning.” *arXiv*. https://arxiv.org/abs/2310.03744.

Luo, Sha, San Jung Kim, Zening Duan, and Kaiping Chen. 2024. “A Sociotechnical Lens for Evaluating Computer Vision Models: A Case Study on Detecting and Reasoning about Gender and Emotion.” In *arXiv preprint arXiv:2406.08222*.

Maaz, Muhammad, Hanoona Rasheed , Salman Khan , and Fahad Shahbaz Khan. 2023. “Video-chatgpt: Towards detailed video understanding via large vision and language models.” *arXiv preprint* arXiv:2306.05424.

Maiwald, Ferdinand, Theresa Vietze, Danilo Schneider, Frank Henze, Sander Münster, and Florian Niebling. 2017. “Photogrammetric Analysis of Historical Image Repositories for Virtual Reconstruction in the Field of Digital Humanities.” *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 42: 447-452.

Meinardus, Boris, Anil Batra , Anna Rohrbach , Marcus Rohrbach. 2024. “The surprising effectiveness of multimodal large language models for video moment retrieval.” *arXiv preprint arXiv:2406.18113* .

Münster, Sander, Fabrizio I. Apollonio, Peter Bell, Piotr Kuroczynski, Isabella Di Lenardo, Fulvio Rinaudo, and Rosa Tamborrino. 2019. “Digital

Cultural Heritage Meets Digital Humanities." *International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* 42 (2/W15): 813-820.

Nayak, Shravan, Kanishk Jain, Rabiul Awal, Siva Reddy, Sjoerd van Steenkiste, Lisa Anne Hendricks, and others. 2024. "Benchmarking Vision Language Models for Cultural Understanding." *arXiv preprint arXiv:2407.10920*.

O'Halloran, Kay, Alvin Chua, and Alexey Podlasov. 2014. "The Role of Images in Social Media Analytics: A Multimodal Digital Humanities Approach." In *Visual Communication*, edited by David Machin, 565-588. De Gruyter. https://doi.org/10.1515/9783110255492.565.

OpenAI. 2024. "GPT-4o Mini: Advancing Cost-Efficient Intelligence." OpenAI. Accessed November 27, 2024. *https://openai.com/index/gpt-4o mini-advancing-cost-efficient-intelligence/* .

OpenAI. n.d. "Batch Processing." OpenAI. Accessed November 27, 2024. https://platform.openai.com/docs/api-reference/batch.

Pustu-Iren, Kader, Julian Sittel, Roman Mauer, Oksana Bulgakowa, and Ralph Ewerth. 2020. "Automated Visual Content Analysis for Film Studies: Current Status and Challenges." *DHQ: Digital Humanities Quarterly* 14 (4).

Smits, Thomas, and Melvin Wevers. 2023. "A Multimodal Turn in Digital Humanities: Using Contrastive Machine Learning Models to Explore, Enrich, and Analyze Digital Visual Historical Collections." *Digital Scholarship in the Humanities* 38 (3): 1267-1280. https://doi.org/10.1093/llc/fqad008.

Sommer, Vivien. 2021. "Multimodal analysis in qualitative research: Extending grounded theory through the lens of social semiotics." *Qualitative Inquiry* , *27* (8-9), 1102-1113.

Tang, Yunlong, Jing Bi , Siting Xu , Luchuan Song , Susan Liang , Teng Wang, ... and Chenliang Xu. 2023. "Video understanding with large language models: A survey." *arXiv preprint* arXiv:2312.17432.

Wang, Yi, Yinan He , Yizhuo Li , Kunchang Li , Jiashuo Yu , Xin Ma, ... & Yu Qiao. 2023. "Internvid: A large-scale video-text dataset for multimodal understanding and generation." *arXiv preprint* arXiv:2307.06942.

Yang, Zhengyuan, Linjie Li , Kevin Lin , Jianfeng Wang , Chung-Ching Lin , Zicheng Liu , and Lijuan Wang. 2023. "The dawn of lmms: Preliminary explorations with gpt-4v (ision)." *arXiv preprint arXiv:2309.17421* , *9* (1), 1.

Yin, Shukang, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2023. "A Survey on Multimodal Large Language Models." *arXiv preprint arXiv:2306.13549*.

Yue, Xiang, Yuansheng Ni , Kai Zhang , Tianyu Zheng , Ruoqi Liu , Ge Zhang, ... and Wenhu Chen. 2024. "Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 9556-9567).

Zhang, Hang, Xin Li , and Lidong Bing. 2023. "Video-llama: An instruction-tuned audio-visual language model for video understanding." *arXiv preprint* arXiv:2306.02858.

Zhou, Junjie, Yan Shu , Bo Zhao , Boya Wu , Shitao Xiao , Xi Yang , Yongping Xiong, … and Zheng Liu. 2024. "MLVU: A Comprehensive Benchmark for Multi-Task Long Video Understanding." *arXiv preprint arXiv:2406.04264* .