Can Large Language Models Grasp Concepts in Visual Content? A Case Study on YouTube Shorts about Depression

Jiaying "Lizzy" Liu*
jiayingliu@utexas.edu
School of Information, The University of Texas at Austin
USA

Yiheng Su*
sam.su@utexas.edu
School of Information, The University of Texas at Austin
USA

Praneel Seth
praneelseth@utexas.edu
Computer Science Department, The University of Texas at Austin
USA

ABSTRACT

Large language models (LLMs) are increasingly used to assist computational social science research. While prior efforts have focused on text, the potential of leveraging multimodal LLMs (MLLMs) for online video studies remains underexplored. We conduct one of the first case studies on MLLM-assisted video content analysis, comparing AI's interpretations to human understanding of abstract concepts. We leverage LLaVA-1.6 Mistral 7B to interpret four abstract concepts regarding video-mediated self-disclosure, analyzing 725 keyframes from 142 depression-related YouTube short videos. We perform a qualitative analysis of MLLM's self-generated explanations and found that the degree of operationalization can influence MLLM's interpretations. Interestingly, greater detail does not necessarily increase human-AI alignment. We also identify other factors affecting AI alignment with human understanding, such as concept complexity and versatility of video genres. Our exploratory study highlights the need to customize prompts for specific concepts and calls for researchers to incorporate more human-centered evaluations when working with AI systems in a multimodal context.

KEYWORDS

Multimodal Information, Large Language-and-Vision Assistant (LLaVA), User Generated Content, Content Analysis

ACM Reference Format:
Jiaying "Lizzy" Liu, Yiheng Su, and Praneel Seth. 2018. Can Large Language Models Grasp Concepts in Visual Content? A Case Study on YouTube Shorts about Depression. In *CSCW 24*. ACM, New York, NY, USA, 10 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

Video-sharing platforms such as YouTube [31], TikTok [41], and Instagram [2] are rich data sources for research in human-computer interaction and computational social sciences. However, traditional methods for analyzing videos, like digital ethnography [24] and content analysis [13], are labor-intensive with limited scalability [4]. Consequently, there is a rising demand for automated approaches to analyze multimodal (visual, textual, audio) content [5].

One successful strategy is leveraging LLMs to augment text-based content analysis, improving open coding efficiency [10] and enabling collaborative coding frameworks [16, 48]. Emerging Multimodal LLMs (MLLMs) like LLaVA [29] and GPT-4 [38] demonstrate promise for understanding visual information at scale [46]. However, few works have investigated how MLLMs can best assist content analysis of videos [44, 50]. Preliminary work [34] suggests that MLLMs may struggle to capture abstract visual concepts, such as video presentation style [32], limiting their applications beyond objective entity or action recognition in video analysis [1, 8].

This case study thus aims to explore the capability of MLLMs to understand abstract concepts in multimodal contexts. Specifically, we investigate how LLaVA-1.6 Mistral 7B interprets four concepts related to depression and self-disclosure behaviors in short YouTube videos, assessing the MLLM's alignment with human understanding. We aim to explore:

- RQ1: How can social concepts be operationalized to guide MLLMs in interpreting video content?
- RQ2: What factors affect MLLM's alignment with human interpretations of social concepts in videos?

Echoing the emerging trend of LLM-assisted content analysis, our case study is one of the earliest efforts to leverage MLLMs for video content analysis: 1) We experiment with harnessing an MLLM for annotating abstract visual concepts with structured and explainable outputs; 2) We examine the MLLM's explanations and reveal contextual factors that affect MLLM's alignment with human understanding of abstract social concepts. 3) We discuss implications for designing robust, human-centered workflows for future MLLM-assisted video content analysis.

2 CONTEXT: MENTAL HEALTH DISCLOSURE THROUGH VIDEOS

Individuals increasingly use digital platforms to share their mental health experiences and seek support online [14]. While prior research has extensively focused on text-based platforms like Twitter [11] and Reddit [40], visual-based platforms like Instagram [3] and YouTube [21, 33] are growing in popularity for self-disclosure documentation.

Similar to the influence of linguistic features on engagement for text-based social media posts, prior studies have highlighted the significant role of visual representations in shaping audience perception [20] and engagement [28]. However, which specific features of the visual representations and how they influence audience engagement remain unclear. This study is part of a larger project

*Both authors contributed equally to this research.

Conference acronym 'XX, Woodstock, NY

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/XXXXXXX.XXXXXXX

to investigate how visual features moderate the relationship between self-disclosure and video engagement (e.g., likes, comments) in depression-related YouTube shorts, which can inform the design of more supportive communities on video-sharing platforms. Given the challenges of manual annotation for large-scale video content analysis, we leverage MLLMs for assistance.

We selected four concepts (Table 1) that shape video-mediated self-disclosure. Presenting and interacting styles represent distinct approaches to structuring and delivering video narratives, which influence audience engagement [2, 25]. Visual diversity and arousal are unique for video-based communication, influencing viewers' attention and perception of content engagement [36, 42]. These visual characteristics are indicative cues to determine how effectively mental health content resonates with and engages viewers.

3 METHODOLOGY

3.1 Dataset

Using the query "depression" with the YouTube Data API, we collected the metadata (e.g., title, channel, duration) of 3,892 videos uploaded by February 2024. We randomly selected 150 videos and downloaded them using `YoutubeDownloader`[1]. Following , due to computational constraints and current MLLM's limited context window to process videos [19], we applied FFmpeg [45] to extract keyframes instead, resulting in 800 keyframes. After labeling and filtering out low-quality frames (e.g., transitional frames, blurry, black screens), we obtained 725 keyframes across 142 videos.

Our study qualifies for exemption under our Institutional Review Board guidelines. Nevertheless, recognizing the sensitive nature of mental health topics, we safeguard video creators' privacy by anonymizing their identities through obscuring facial features. Further discussion of ethical considerations can be found in Appendix C.

3.2 MLLM Concept Annotation: Models and Prompts

We select `llava-v1.6-mistral-7b-hf`[2] [30] for analysis, and will henceforth refer to this model as (the) MLLM for convenience. To investigate the MLLM's comprehension of abstract visual concepts (Table 1), operationalizing these concepts is essential for articulating them effectively. To address RQ1 and explore how to operationalize the concepts for MLLM prompt configuration, we tested four strategies to evaluate their effectiveness. Specifically, we implemented four prompting configurations with progressively increasing levels of operational guidance to strike a balance between clarity and flexibility. See Appendix A for all prompt configurations.

- Naive: The MLLM is directly queried for the presence or extent of the concept without any additional contexts.
- Simple: A short definition is added to the naive query.
- Detailed: A detailed definition with three abstract manifestations is added to the naive query.
- Open-minded: Similar to the detailed prompt, but also explicitly encourages the MLLM to consider other scenarios not already stated.

[1]https://github.com/Tyrrrz/YoutubeDownloader

[2]https://huggingface.co/llava-hf/llava-v1.6-mistral-7b-hf

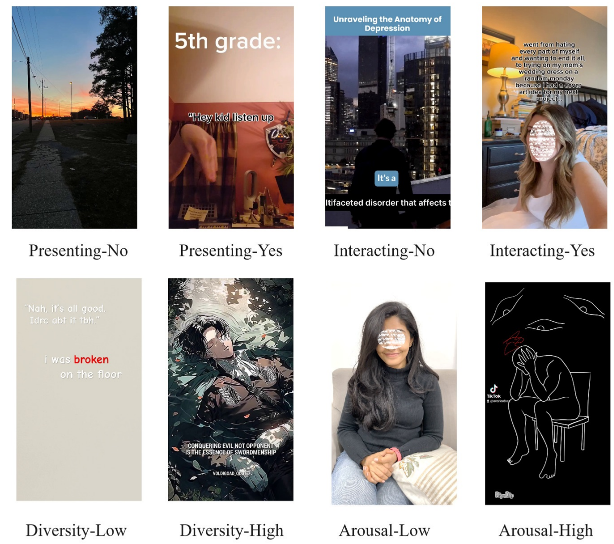


Figure 1: Examples of human interpretations of the four selected concepts. We annotate Yes/No for *presenting* and *interacting,* High/Low for *diversity* and *arousal.* We then compare human interpretations with the MLLM interpretations to evaluate human-AI alignment.

Our prompting configurations are informed by established practices in prompt engineering. For instance, the Detailed configuration aligns with in-context learning by incorporating prototypical examples to serve as implicit "demonstrations" [37]. Similarly, the Open-minded configuration is inspired by chain-of-thought, which uses directives like "think step-by-step" to encourage more flexible interpretations [23]. We do not experiment with advanced configurations such as in-context learning or fine-tuning [12, 18], as we are interested in assessing the MLLM's off-the-shelf capabilities.

We tasked the MLLM with annotating each keyframe across four concepts: Yes/No for *interacting* and *presenting*, and High/Low for *arousal* and *diversity*. To ensure consistency, we prompted the MLLM to provide both interpretations and explanations simultaneously, reducing the likelihood of generating contradictory or hallucinated explanations. Keyframes were queried in temporal order for each video, while the order of prompt configurations and associated concepts was randomized per keyframe to mitigate potential biases. Occasionally, the MLLM combines annotations (e.g., Yes/No) with explanations [32]. To isolate explicit annotations, we utilized `Llama-3.1-8B-Instruct`[3] to parse the MLLM's interpretations. Following this, we manually reviewed all extracted annotations to verify the accuracy of the parsing process.

3.3 Human Annotation Process

To obtain human interpretations, two authors independently coded a random sample of 200 keyframes, with a third author providing an additional vote to resolve disagreements. Figure 1 illustrates examples of human interpretations. After discussing disputes in a group meeting and ensuring that Intercoder Reliability (ICR) [39]

[3]https://huggingface.co/meta-llama/Meta-Llama-3.1-8B-Instruct

Table 1: Four abstract visual concepts shaping video-mediated self-disclosure.

Concept	Definition
Presenting	Presenting style involves the delivery of information, typically accompanied by visual aids like slides or graphics [22].
Interacting	Interacting refers to creators establishing a simulated interpersonal relationship with their audience, fostering a sense of engagement and connection [22].
Diversity	Diversity of an image includes varied scenes, color variation, compositional complexity, and originality of the image [42].
Arousal	Arousal refers to the degree of alertness or excitement elicited by the stimulus such as dynamic visual elements and emotional intensity [36].

is higher than 75%, the three coders split the remaining keyframes and coded them separately. We dropped ambiguous keyframes and low-quality images (e.g., transitional frames, blurry, black screens) from further analysis. Ultimately, we obtain 725 frames across 142 videos with human concept annotations.

3.4 Data Analysis

Quantitative Comparisons. To compare the four prompt configurations, we quantify human-AI (mis)alignment as the consistency between a prompt-concept pair and the corresponding human annotations. We then employ the bootstrapping approach from Berg-Kirkpatrick et al. [6] to assess how human-AI alignment differs across configurations per concept. Please see Appendix B for details. We discuss quantitative comparisons in Section 4.1.

Qualitative Analysis. To investigate the underlying factors behind human-AI (mis)alignments, we first curated a focused dataset of instances where the MLLM's annotations diverged when using different prompting configurations. Two authors then independently conducted thematic analysis [7] on the MLLM's explanations for these keyframes. They met weekly to discuss emerging themes and patterns in the data, resolve any coding discrepancies through detailed discussion, and iterate on the coding scheme to establish definitions for each thematic category. The analysis focused on several key dimensions, including the nature and patterns of annotation changes, the MLLM's reasoning and justification for modifications, contextual factors that appeared to influence changes, and the relationship between prompting configuration and annotation stability. We summarize recurring themes and patterns in Section 4.2.

4 FINDINGS

4.1 Quantative Evaluation of MLLM-Human Alignment

Figure 2 shows the distribution of bootstrapped alignment scores across prompt configurations for each concept. The MLLM demonstrates varying capabilities: no single prompt configuration consistently achieves the highest alignment.

The MLLM excels at abstract concepts like arousal and diversity but exhibits lower alignment and more variance for performative concepts like interacting and presenting. Under the naive approach, the MLLM performs well for concepts like interacting, arousal, and diversity, suggesting that the MLLM's prior knowledge of these concepts (derived from pre-trained data) aligns well with corresponding human conceptions. We only observe substantial alignment gains with more operationalization guidance for presenting. However, this effect is not monotonic (e.g., a more detailed prompt does not always lead to better alignment) and does not generalize to other concepts. Adding definitions may decrease alignment for presenting and interacting, restricting the MLLM's capabilities. We discuss the factors that impact annotations in Section 4.2.

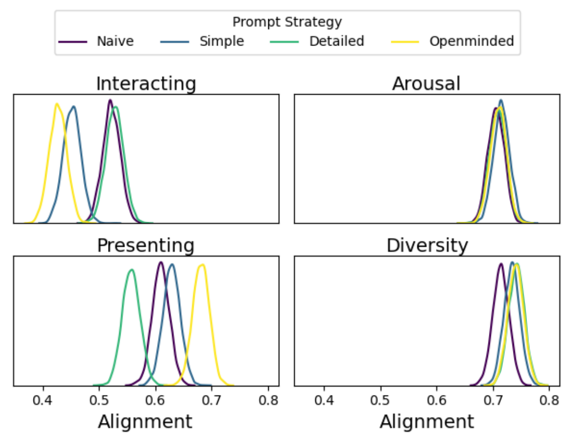
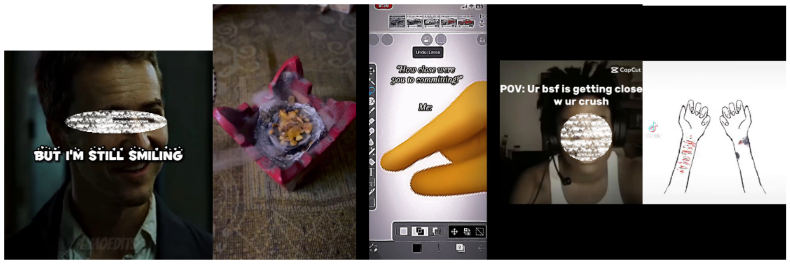


Figure 2: Distribution of bootstrap alignment scores across prompt configurations and concepts. The MLLM demonstrates varying capabilities: no single prompt configuration consistently achieves the highest alignment across all concepts.

4.2 Factors Affecting MLLM-Human (Mis)Alignment

Evidently, concept operationalization is a key factor influencing human-AI alignment. By analyzing the MLLM's explanations, we offer qualitative insights into how and why operationalization impacts alignment. Additionally, we identify two further factors contributing to human-AI (mis)alignment: concept complexity and the diversity of genres.

4.2.1 ***Varying Concept Specification*** Concept specification refers to the amount of detail in the prompts. For interacting and presenting (Figure 2), auxiliary definitions may inadvertently prioritize "what is in the prompt" over the holistic context of the image, causing the MLLM to be less aligned with human perceptions. In contrast, the naive approach shows greater flexibility in capturing novel categories of presenting and interacting communication styles.



(a) (b) (c) (d) (e)

Figure 3: Problematic MLLM Annotations.

Figure 3-(a) illustrates the variability in the MLLM's interpretation of presenting style. When prompted naively, the MLLM correctly identifies (a) as presenting, stating that the superimposed caption is "*a common technique used in presentations*", complemented by "*the person's facial expression, which appears to be a smile*". Conversely, when prompted with simple or detailed configurations, the MLLM misclassifies (a), citing "*no visible slide or graphic that would be associated with a presentation*" as evidence. This misclassification occurred because the detailed prompts explicitly exemplified presenting styles as "slides or graphics," limiting the MLLM from considering informal contexts of presenting style. In contrast, the openminded configuration correctly identifies (a), further underscoring that additional details can enhance clarity but reduce alignment if not carefully operationalized.

Without definitional constraints, the naive configuration can better capture nuanced social dynamics. In Figure 3-(b), the MLLM accurately described the interactive potential, noting the "*dynamic and engaging*" style of the image to "*[invite] the viewer to observe and possibly speculate about what is happening.*" However, when prompted with a detailed configuration, the MLLM incorrectly claims that the image "*is a still photograph*" with "*no indication of a simulated interpersonal relationship or engagement with an audience*".

We consistently observe this pattern of contradictory decisions for presenting and interacting queries, where explanations often highlight the absence of explicit elements outlined in the prompt. For example, keyframes without human presence or overt conversational styles (Figure 3-(c)) were misclassified as non-interactive despite employing engaging nontraditional styles such as memes.

4.2.2 ***Varying Complexity of Concepts*** The complexity and scope of the four analyzed concepts vary, making some more challenging for the MLLM. For example, diversity is relatively straightforward, as it involves identifying and counting visual categories, a common pre-training task for MLLMs. Figure 1 illustrates this: the low-diversity image shows a plain background with simple text overlays, while the high-diversity image features a vibrant anime figure. Similarly, the MLLM effectively recognizes arousal levels through visual cues like facial expressions, body language, and visual intensity. In Figure 1, the low-arousal image depicts a calm individual with relaxed features, while the high-arousal image shows an abstract figure with intense body language indicating distress.

In contrast, concepts like interacting and presenting are more challenging because they require situating visual cues within context. For instance, in the "Presenting-Yes" image (Figure 1), while the hand gesture might initially suggest interaction, the gesture is not directed at the audience but instead presents the scenario encoded in the text overlay ("*5th grade: Hey kid listen up*"). In multimodal contexts, the meaning of one element (e.g., a visual cue) can influence, support, or contradict another (e.g., text). This demand to interpret co-dependent features holistically poses a novel challenge absent in text-only settings.

When MLLM's pre-trained knowledge diverges from human conceptions, naive queries often result in misalignment. We observe this quantitatively, as the Naive alignment for presenting is very low (Figure 2). Qualitatively, in the "Presenting-Yes" image (Figure 1), the MLLM incorrectly states that the image does not show presentation style, citing the absence of expected behaviors like "*a speaker standing at a podium or a lectern*" and "*a slide or a graphic*". The MLLM fails to contextualize the informal setting and gesture as a valid presentation style, thus struggling to adapt to novel communicative contexts outside pretraining. Prompt engineering can help, as the MLLM correctly identifies this image for all other configurations besides naive.

4.2.3 ***Versatile Video Genres*** The versatility of videos can challenge the MLLM's ability to understand social concepts. We identify two genres with relatively low alignment, highlighting the complexities of interpreting diverse content.

Mixture of textual and visual elements. Short videos often combine visuals with overlaying text, as shown in Figure 3-(a, c, d). When visual signals conflict with textual information, MLLMs (typically) prioritize textual over visual cues (since they were pre-trained with more text data), potentially leading to misinterpretations. For example, in Figure 3-(d), the MLLM reasons that the image "*does not directly portray an interacting style...as it is static*" but the text overlay "*implies a narrative or a message that is meant to convey a sense of interaction.*" Effectively synthesizing two potentially conflicting sources of information—visual and textual—is a unique and open challenge for MLLMs.

Non-human genres. Resonating Zhong and Baghel [52], the MLLM struggles to interpret non-human video genres such as cartoons, memes, and abstract art, which often require cultural, emotional, or other contextual knowledge for accurate interpretations. For example, Figure 3-(e) depicts a hand-drawn image of self-harm behaviors, potentially signaling interaction intentions such as a call for help. However, the MLLM failed to recognize implicit interaction cues and explained that "*the drawing...does not exhibit any conversational language or behaviors that would suggest an interacting style*". Fine-tuning or more sophisticated prompt engineering is likely needed to educate the MLLM on a broader range of visual storytelling techniques and cultural references.

5 DISCUSSION AND FUTURE WORK

We conduct an exploratory study with a single model, limited samples, and simple prompts, so our findings may not be generalizable. Computational constraints further prevented the inclusion of temporal context in videos, which may limit our findings. Despite these limitations, our study offers insights into the potential of leveraging Multimodal Large Language Models (MLLMs) to assist visual content analysis. Recognizing the inherent subjectivity of social concepts (even with high intercoder reliability), we use "alignment" rather than "accuracy" and contextualize our quantitative statistics with qualitative insights. Our analysis illuminates key factors contributing to MLLM's misalignment from human understanding, including concept specifications, concept complexity, and versatility of video genres, which must be considered carefully when engineering prompts for MLLMs-assisted video content analysis.

5.1 Harnessing MLLMs for Large-Scale Multimodal Content Analysis: Opportunities and Challenges

MLLMs show potential in scaling visual content analysis. With appropriate operationalization, our results show that the MLLM can align highly with human perceptions, even for abstract concepts like presentation style. By expediting manual labeling, which is often time-intensive and costly [13], MLLM can enable more comprehensive analyses of large datasets, potentially uncovering rare communication patterns that might otherwise go unnoticed in small-sample qualitative studies [35]. Furthermore, MLLMs can enhance data quality by serving as a proxy for human intervention. In our pipeline, the MLLM accurately labeled low-quality frames as "Not Applicable," distinguishing them from frames that genuinely lacked the desired concept. This capability can help researchers filter noisy inputs by inspecting ambiguous model outputs and explanations.

Despite their potential, MLLMs can be misaligned with human perceptions. Our findings indicate that operationalizing abstract concepts with greater detail can enhance alignment. However, it may also risk constraining the MLLM's ability to uncover novel social dynamics beyond the specified criteria. This contrasts with typical in-context or few-shot learning scenarios, where multiple demonstrations help the model infer task structure and reduce ambiguity by leveraging patterns recognized during pretraining [37]. In diverse social media content, models must balance consistency with flexibility to adapt to dynamic contexts. Additionally, when applying MLLM to analyze videos in the wild, style diversity is a crucial factor impacting model alignment. The short videos in our study are predominantly informal and casually filmed in everyday settings. They differ from vlogs, tutorials, streams, or product reviews, typically more structured and polished. Our findings show that the MLLM can struggle to capture and interpret unconventional visual cues, such as the novel yet subtle suggestion of suicide depicted in Figure 3 (c). Developing and evaluating models that can effectively navigate such ambiguity while maintaining alignment on more structured formats remains essential for advancing multimodal analysis across diverse platforms.

5.2 Future Directions

We emphasize three directions to improve human-AI alignment in (M)LLM-assisted visual content analysis: human-centered auditing, multimodal synthesis, and temporality incorporation.

Implementing MLLM response auditing. In our case study, MLLM interpretations often diverged from human concept understanding due to factors like concept complexity and the diversity of video genres. Specifically, the MLLM may systematically misunderstand the visual cues of videos of specific genres, such as cartoons and memes, as suggested in Section 4.2.3. Thus, it is crucial to implement human-centered post hoc audits [47, 51]. Shen et al. [43] developed a framework to audit the value alignment of humans and language models to improve transparency and ethical use of AI in social research. Future work can explore incorporating human-centered evaluation as a standard step in MLLM-assisted content analysis workflows [9, 17, 27, 49]. Such measures can facilitate the iterative refinement of concept operationalization and prompt engineering to address known biases in an AI's understanding of social concepts.

Synthesizing multimodal inputs. In our current workflow, we decode videos into keyframes and prompt the MLLM to annotate concepts given isolated images. However, we can also incorporate audio or transcripts to provide a more comprehensive analysis, though interpreting signals from multiple input sources remains challenging. Additionally, as discussed in Section 4.2.3, conflicting information across different modalities can complicate interpretations. Developing more sophisticated methods for synthesizing multimodal inputs is thus a promising avenue for future research.

Incoporating video temporality. Some concepts require temporal context for accurate interpretations. For example, concepts like emotional valence and genre often depend on a holistic understanding of the video's overall narrative [32], which isolated keyframes cannot capture. Future work could explore MLLMs that directly interpret videos or a sequence of keyframes to provide more contextual information.

5.3 Ethical Considerations

Our findings suggest that human-AI misalignment may result in systematic biases. Previous studies have reported LLMs' biases towards minorities and underrepresented populations, including people with disabilities [15] and socially subordinate groups [26]. Future studies can work on identifying the potential biases in LLMs.

6 CONCLUSION

We conduct one of the earliest case studies on leveraging Multimodal Large Language Models (MLLMs) to interpret abstract social concepts in video data. Our results underscore the importance of post-hoc auditing and human oversight to ensure agreement between AI outputs and human understanding. Future work should explore the integration of multimodal inputs and experiment with fine-tuning or in-context learning to enhance the model's ability to understand more complex social interactions.

REFERENCES

[1] Dana Alsagheer, Rabimba Karanjai, Weidong Shi, Nour Diallo, Yang Lu, Suha Beydoun, and Qiaoning Zhang. 2024. Evaluating Irrationality in Large Language Models and Open Research Questions. (2024).

[2] Nazanin Andalibi. 2017. Self-disclosure and Response Behaviors in Socially Stigmatized Contexts on Social Media: The Case of Miscarriage. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems (CHI EA '17)*. Association for Computing Machinery, New York, NY, USA, 248–253. doi:10.1145/3027063.3027137

[3] Nazanin Andalibi, Pinar Ozturk, and Andrea Forte. 2017. Sensitive Self-disclosures, Responses, and Social Support on Instagram: The Case of #Depression. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*. ACM, Portland Oregon USA, 1485–1500. doi:10.1145/2998181.2998243

[4] Ava Bartolome and Shuo Niu. 2023. A Literature Review of Video-Sharing Platform Research in HCI. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 790, 20 pages. doi:10.1145/3544548.3581107

[5] Ava Bartolome and Shuo Niu. 2023. A Literature Review of Video-Sharing Platform Research in HCI. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–20.

[6] Taylor Berg-Kirkpatrick, David Burkett, and Dan Klein. 2012. An Empirical Investigation of Statistical Significance in NLP. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, Jun'ichi Tsujii, James Henderson, and Marius Paşca (Eds.). Association for Computational Linguistics, Jeju Island, Korea, 995–1005. https://aclanthology.org/D12-1091

[7] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[8] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2024. A Survey on Evaluation of Large Language Models. *ACM Trans. Intell. Syst. Technol.* 15, 3, Article 39 (mar 2024), 45 pages. doi:10.1145/3641289

[9] Xiang Chen, Chenxi Wang, Yida Xue, Ningyu Zhang, Xiaoyan Yang, Qiang Li, Yue Shen, Lei Liang, Jinjie Gu, and Huajun Chen. 2024. Unified Hallucination Detection for Multimodal Large Language Models. arXiv:2402.03190 [cs.CL] https://arxiv.org/abs/2402.03190

[10] Robert Chew, John Bollenbacher, Michael Wenger, Jessica Speer, and Annice Kim. 2023. LLM-Assisted Content Analysis: Using Large Language Models to Support Deductive Coding. doi:10.48550/arXiv.2306.14924 Issue: arXiv:2306.14924 arXiv:2306.14924 [cs, stat].

[11] Munmun De Choudhury and Sushovan De. 2014. Mental Health Discourse on reddit: Self-Disclosure, Social Support, and Anonymity. (May 2014). https://www.scinapse.io/papers/2182854643

[12] Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Baobao Chang, Xu Sun, Lei Li, and Zhifang Sui. 2024. A Survey on In-context Learning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 1107–1128. doi:10.18653/v1/2024.emnlp-main.64

[13] James W. Drisko and Tina Maschi. 2016. *Content Analysis*. Oxford University Press. Google-Books-ID: 07GYCgAAQBAJ.

[14] Jessica L. Feuston and Anne Marie Piper. 2019. Everyday Experiences: Small Stories and Mental Illness on Instagram. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–14. doi:10.1145/3290605.3300495

[15] Vinitha Gadiraju, Shaun Kane, Sunipa Dev, Alex Taylor, Ding Wang, Emily Denton, and Robin Brewer. 2023. "I wouldn't say offensive but...": Disability-Centered Perspectives on Large Language Models. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 205–216. doi:10.1145/3593013.3593989

[16] Jie Gao, Yuchen Guo, Toby Jia-Jun Li, and Simon Tangi Perrault. 2023. CollabCoder: A GPT-Powered WorkFlow for Collaborative Qualitative Analysis. In *Companion Publication of the 2023 Conference on Computer Supported Cooperative Work and Social Computing (CSCW '23 Companion)*. Association for Computing Machinery, New York, NY, USA, 354–357. doi:10.1145/3584931.3607500

[17] Katy Ilonka Gero, Chelse Swoopes, Ziwei Gu, Jonathan K. Kummerfeld, and Elena L. Glassman. 2024. Supporting Sensemaking of Large Language Model Outputs at Scale. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–21. doi:10.1145/3613904.3642139

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

[19] Bo He, Hengduo Li, Young Kyun Jang, Menglin Jia, Xuefei Cao, Ashish Shah, Abhinav Shrivastava, and Ser-Nam Lim. 2024. MA-LMM: Memory-Augmented Large Multimodal Model for Long-Term Video Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[20] Pengwei Hu, Chenhao Lin, Jiajia Li, Feng Tan, Xue Han, Xi Zhou, and Lun Hu. 2023. Making the Implicit Explicit: Depression Detection in Web across Posted Texts and Images. In *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. 4807–4811. doi:10.1109/BIBM58861.2023.10385590 ISSN: 2156-1133.

[21] Jina Huh, Leslie S. Liu, Tina Neogi, Kori Inkpen, and Wanda Pratt. 2014. Health Vlogs as Social Support for Chronic Illness Management. *ACM Trans. Comput.-Hum. Interact.* 21, 4 (Aug. 2014), 23:1–23:31. doi:10.1145/2630067

[22] Margot Kelly-Hedrick, Paul H. Grunberg, Felicia Brochu, and Phyllis Zelkowitz. 2018. "It's Totally Okay to Be Sad, but Never Lose Hope": Content Analysis of Infertility-Related Videos on YouTube in Relation to Viewer Preferences. *Journal of Medical Internet Research* 20, 5 (May 2018), e10199. doi:10.2196/10199 Number: 5.

[23] Takeshi Kojima, Shixiang (Shane) Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large Language Models are Zero-Shot Reasoners. In *Advances in Neural Information Processing Systems*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (Eds.), Vol. 35. Curran Associates, Inc., 22199–22213. https://proceedings.neurips.cc/paper_files/paper/2022/file/8bb0d291acd4acf06ef112099c16f326-Paper-Conference.pdf

[24] Sebastian Kubitschko and Anne Kaun (Eds.). 2016. *Innovative Methods in Media and Communication Research*. Springer International Publishing, Cham. doi:10.1007/978-3-319-40700-5

[25] E Megan Lachmar, Andrea K Wittenborn, Katherine W Bogen, and Heather L McCauley. 2017. #MyDepressionLooksLike: Examining Public Discourse About Depression on Twitter. *JMIR MENTAL HEALTH* (2017), 11.

[26] Messi H.J. Lee, Jacob M. Montgomery, and Calvin K. Lai. 2024. Large Language Models Portray Socially Subordinate Groups as More Homogeneous, Consistent with a Bias Observed in Humans. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24)*. Association for Computing Machinery, New York, NY, USA, 1321–1340. doi:10.1145/3630106.3658975

[27] Florian Leiser, Sven Eckhardt, Valentin Leuthe, Merlin Knaeble, Alexander Mädche, Gerhard Schwabe, and Ali Sunyaev. 2024. HILL: A Hallucination Identifier for Large Language Models. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3613904.3642428

[28] Shuailin Li, Shiwei Wu, Tianjian Liu, Han Zhang, Qingyu Guo, and Zhenhui Peng. 2024. Understanding the Features of Text-Image Posts and Their Received Social Support in Online Grief Support Communities. *Proceedings of the International AAAI Conference on Web and Social Media* 18 (May 2024), 917–929. doi:10.1609/icwsm.v18i1.31362

[29] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved Baselines with Visual Instruction Tuning. arXiv:2310.03744 [cs.CV]

[30] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved Baselines with Visual Instruction Tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 26296–26306.

[31] Jiaying Liu and Yan Zhang. 2024. Modeling Health Video Consumption Behaviors on Social Media: Activities, Challenges, and Characteristics. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW1 (April 2024), 208:1–208:28. doi:10.1145/3653699

[32] Jiaying Lizzy Liu, Yunlong Wang, Yao Lyu, Yiheng Su, Shuo Niu, Xuhai Orson Xu, and Yan Zhang. 2024. Harnessing LLMs for Automated Video Content Analysis: An Exploratory Workflow of Short Videos on Depression. In *arXiv.org*. doi:10.1145/3678884.3681850

[33] Leslie S. Liu, Jina Huh, Tina Neogi, Kori Inkpen, and Wanda Pratt. 2013. Health vlogger-viewer interaction in chronic illness management. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, Paris France, 49–58. doi:10.1145/2470654.2470663

[34] Lydia Manikonda and Munmun De Choudhury. 2017. Modeling and Understanding Visual Attributes of Mental Health Disclosures in Social Media. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. Association for Computing Machinery, New York, NY, USA, 170–181. doi:10.1145/3025453.3025932

[35] Ryan McGrady, Kevin Zheng, Rebecca Curran, Jason Baumgartner, and Ethan Zuckerman. 2023. Dialing for Videos: A Random Sample of YouTube. *Journal of Quantitative Description: Digital Media* 3 (2023).

[36] Nikos Metallinos. 2013. *Television aesthetics: Perceptual, cognitive and compositional bases*. Routledge.

[37] Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the Role of Demonstrations: What Makes In-Context Learning Work?. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (Eds.). Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 11048–11064. doi:10.18653/v1/2022.emnlp-main.759

[38] OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]

[39] Cliodhna O'Connor and Helene Joffe. 2020. Intercoder reliability in qualitative research: debates and practical guidelines. *International journal of qualitative methods* 19 (2020), 1609406919899220.

[40] Sachin R. Pendse, Neha Kumar, and Munmun De Choudhury. 2023. Marginalization and the Construction of Mental Illness Narratives Online: Foregrounding Institutions in Technology-Mediated Care. *Proceedings of the ACM on Human-Computer Interaction* 7, CSCW2 (Oct. 2023), 346:1–346:30. doi:10.1145/3610195

[41] Anastasia Schaadhardt, Yue Fu, Cory Gennari Pratt, and Wanda Pratt. 2023. "Laughing so I don't cry": How TikTok users employ humor and compassion to connect around psychiatric hospitalization. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23)*. Association for Computing Machinery, New York, NY, USA, 1–13. doi:10.1145/3544548.3581559

[42] Mirjam Seckler, Klaus Opwis, and Alexandre N Tuch. 2015. Linking objective design factors with subjective aesthetics: An experimental study on how structure and color of websites affect the facets of users' visual aesthetic perception. *Computers in Human Behavior* 49 (2015), 375–389.

[43] Hua Shen, Tiffany Knearem, Reshmi Ghosh, Yu-Ju Yang, Tanushree Mitra, and Yun Huang. 2024. Valuecompass: A framework of fundamental values for human-ai alignment. *arXiv preprint arXiv:2409.09586* (2024).

[44] Yunlong Tang, Jing Bi, Siting Xu, Luchuan Song, Susan Liang, Teng Wang, Daoan Zhang, Jie An, Jingyang Lin, Rongyi Zhu, et al. 2023. Video understanding with large language models: A survey. *arXiv preprint arXiv:2312.17432* (2023).

[45] Suramya Tomar. 2006. Converting video formats with FFmpeg. *Linux journal* 2006, 146 (2006), 10.

[46] Jiaqi Wang, Hanqi Jiang, Yiheng Liu, Chong Ma, Xu Zhang, Yi Pan, Mengyuan Liu, Peiran Gu, Sichen Xia, Wenjun Li, et al. 2024. A Comprehensive Review of Multimodal Large Language Models: Performance and Challenges Across Different Tasks. *arXiv preprint arXiv:2408.01319* (2024).

[47] Ziang Xiao, Wesley Hanwen Deng, Michelle S. Lam, Motahhare Eslami, Juho Kim, Mina Lee, and Q. Vera Liao. 2024. Human-Centered Evaluation and Auditing of Language Models. In *Extended Abstracts of the 2024 CHI Conference on Human Factors in Computing Systems (CHI EA '24)*. Association for Computing Machinery, New York, NY, USA, 1–6. doi:10.1145/3613905.3636302

[48] Ziang Xiao, Xingdi Yuan, Q. Vera Liao, Rania Abdelghani, and Pierre-Yves Oudeyer. 2023. Supporting Qualitative Analysis with Large Language Models: Combining Codebook with GPT-3 for Deductive Coding. In *Companion Proceedings of the 28th International Conference on Intelligent User Interfaces (IUI '23 Companion)*. Association for Computing Machinery, New York, NY, USA, 75–78. doi:10.1145/3581754.3584136

[49] Yakir Yehuda, Itzik Malkiel, Oren Barkan, Jonathan Weill, Royi Ronen, and Noam Koenigstein. 2024. InterrogateLLM: Zero-Resource Hallucination Detection in LLM-Generated Answers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 9333–9347. https://aclanthology.org/2024.acl-long.506

[50] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *National Science Review* 11, 12 (11 2024), nwae403. doi:10.1093/nsr/nwae403 arXiv:https://academic.oup.com/nsr/article-pdf/11/12/nwae403/61201557/nwae403.pdf

[51] He Zhang, Chuhao Wu, Jingyi Xie, Yao Lyu, Jie Cai, and John M. Carroll. 2023. Redefining Qualitative Analysis in the AI Era: Utilizing ChatGPT for Efficient Thematic Analysis. doi:10.48550/arXiv.2309.10771 Issue: arXiv:2309.10771 arXiv:2309.10771 [cs].

[52] Yang Zhong and Bhiman Kumar Baghel. 2024. Multimodal Understanding of Memes with Fair Explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. 2007–2017.

A LLAVA PROMPTS

Table 2: LLaVA Prompts

Concept	Strategy	Prompt
Interacting	Prompt 0 - Naive	"<image> USER: Does this picture portray an interacting style, yes or no? Explain your answer. ASSISTANT:"
	Prompt 1 - Simple Definition	"<image> USER: Interacting style refers to creators establishing a simulated interpersonal relationship with their audience, fostering a sense of engagement and connection. Does this picture portray an interacting style, yes or no? Explain your answer. ASSISTANT:"
	Prompt 2 - Detailed Definition	"<image> USER: Interacting style refers to creators establishing a simulated interpersonal relationship with their audience, fostering a sense of engagement and connection. This involves behaviors such as directly addressing the audience, using conversational language, or acknowledging comments or questions from viewers. Does this picture portray an interacting style, yes or no? Explain your answer. ASSISTANT:"
	Prompt 3 - Openminded	"<image> USER: Interacting style refers to creators establishing a simulated interpersonal relationship with their audience, fostering a sense of engagement and connection. This involves behaviors such as directly addressing the audience, using conversational language, or acknowledging comments or questions from viewers. These are just several examples, so be open-minded to other potential scenarios of interacting style. Does this picture portray an interacting style, yes or no? Explain your answer. ASSISTANT:"
Presenting	Prompt 0 - Naive	"<image> USER: Does this picture communicate in a presenting style, yes or no? Explain your answer. ASSISTANT:"
	Prompt 1 - Simple Definition	"<image> USER: Presenting style involves the delivery of information, typically accompanied by visual aids like slides or graphics. Does this picture communicate in a presenting style, yes or no? Explain your answer. ASSISTANT:"
	Prompt 2 - Detailed Definition	"<image> USER: Presenting style involves the delivery of information, typically accompanied by visual aids like slides or graphics, such as a businessman presenting slides, a student giving a speech on a topic, or a general rallying troops for war. Does this picture communicate in a presenting style, yes or no? Explain your answer. ASSISTANT:"
	Prompt 3 - Openminded	"<image> USER: Presenting style involves the delivery of information, typically accompanied by visual aids like slides or graphics, such as a businessman presenting slides, a student giving a speech on a topic, or a general rallying troops for war. These are just several examples, so be open-minded to other potential scenarios of presenting style. Does this picture communicate in a presenting style, yes or no? Explain your answer. ASSISTANT:"
Arousal	Prompt 0 - Naive	"<image> USER: What level of arousal does this image communicate, low, moderate, or high? Explain your answer. ASSISTANT:"

Concept	**Strategy**	**Prompt**
	Prompt 1 - Simple Definition	"<image> USER: Low arousal is associated with calmness, relaxation, or drowsiness. Moderate arousal is a balanced state of alertness and engagement without overstimulation. High arousal is characterized by heightened physiological and emotional activity. What level of arousal does this image communicate, low, moderate, or high? Explain your answer. ASSISTANT:"
	Prompt 2 - Detailed	"<image> USER: Low arousal is associated with calmness, relaxation, or drowsiness. For example, feeling fatigued or viewing a peaceful landscape or a calm, monochromatic image. Moderate arousal is a balanced state of alertness and engagement without overstimulation, often linked with optimal performance and involves minimal physiological activation. For example, feeling attentive or focused; engaging in a conversation or viewing a moderately complex image. High arousal is characterized by heightened physiological and emotional activity. For example, feeling excited, anxious, or stressed; or viewing a dynamic or chaotic scene with bright colors or intense stimuli. What level of arousal does this image communicate, low, moderate, or high? Explain your answer? \nASSISTANT:"
	Prompt 3 - Open Minded	"<image> USER: Low arousal is associated with calmness, relaxation, or drowsiness. For example, feeling fatigued or viewing a peaceful landscape or a calm, monochromatic image. Moderate arousal is a balanced state of alertness and engagement without overstimulation, often linked with optimal performance and involves minimal physiological activation. For example, feeling attentive or focused; engaging in a conversation or viewing a moderately complex image. High arousal is characterized by heightened physiological and emotional activity. For example, feeling excited, anxious, or stressed; or viewing a dynamic or chaotic scene with bright colors or intense stimuli. These are just several examples so be open-minded to other potential scenarios of arousal levels. What level of arousal does this image communicate, low, moderate, or high? Explain your answer? \nASSISTANT:"
Diversity	Prompt 0 - Naive	"<image> USER: What level of diversity does this image communicate, low, moderate, or high? Explain your answer? \nASSISTANT:"
	Prompt 1 - Definition	"<image> USER: The diversity of an image includes the color variation, compositional complexity, and originality of the image. What level of diversity does this image communicate, low, moderate, or high? Explain your answer? \nASSISTANT:"
	Prompt 2 - Detailed	"<image> USER: The diversity of an image includes the color variation, compositional complexity, and originality of the image. Color variation involves assessing the range of colors across the image. Compositional complexity involves the arrangement of diverse elements within the image. Originality assesses whether the image presents a new or uncommon perspective. What level of diversity does this image communicate, low, moderate, or high? Explain your answer? \nASSISTANT:"
	Prompt 3 - Open Minded	"<image> USER: The diversity of an image includes the color variation, compositional complexity, and originality of the image. Color variation involves assessing the range of colors across the image. Compositional complexity involves the arrangement of diverse elements within the image. Originality assesses whether the image presents a new or uncommon perspective. These are just several examples so be open-minded to other instances of diversity. What level of diversity does this image communicate, low, moderate, or high? Explain your answer? \nASSISTANT:"

B BOOTSTRAPING DETAILS

To assess how human-AI alignment differs across configurations for each concept, we employ a bootstrapping approach inspired by the methodology outlined in Berg-Kirkpatrick et al. [6]. We first collect a pool of generated annotations for each concept and prompt configuration to compute an initial alignment score. However, relying on a single measure fails to capture the variability inherent in the data, and observed differences across configurations may arise purely by chance. This limitation makes it challenging to draw reliable conclusions about the relative alignment of different prompt configurations.

The bootstrapping approach addresses this issue by repeatedly resampling the data to estimate the variability in alignment scores. Specifically, we generate N resampled datasets, each of size K, by randomly drawing annotations with replacement from the original pool. An alignment score is computed for each resampled dataset, resulting in a distribution of N scores for each concept-prompt pair. This distribution reflects the variability in alignment and enables us to assess, on average, how reliably each prompting configuration aligns with human perceptions across the selected social concepts beyond random chance. We visualize these score distributions in Figure 2) and discuss findings in Section 4.1.

C ETHICS STATEMENT

We are committed to conducting ethically responsible research, ensuring content creators' privacy, and safeguarding research team members' well-being. Since this study analyzes data on publicly available platforms like YouTube, it qualifies for human subjects exemption under our university's Institutional Review Board (IRB) guidelines, posing minimal risk to content creators (or individuals present in the video). Nevertheless, we acknowledge that creators could not provide explicit consent for the inclusion or exclusion of their content. To respect the creator's privacy, we implemented additional protections, such as anonymizing individuals in the video by obscuring facial features in any snapshots in this paper. Additionally, we do not collect personally identifiable metadata about the creators or individuals presented in the videos.

Another ethical factor is the well-being of researchers exposed to potentially distressing material, particularly during qualitative analyses involving sensitive topics like depression. To mitigate potential emotional harm, we provided team members access to university mental health resources, encouraged breaks during data analysis, and fostered an environment of open communication about the work's emotional impact.