Do Multimodal Large Language Models Have Good Taste? An Evaluation of MLLMs via the Visual Aesthetic Sensitivity Test

Yi Yao
City University of Macau
yiyao.tansy.gmail.com

Zhuang Qiu
City University of Macau
zhuangqiu.cityu.edu.mo

Abstract

The surge of large language models (LLMs) has sparked an ongoing debate on the extent to which LLMs emulate human cognition, and the emergence of multimodal large language models (MLLMs) opens a new window for benchmarking machine capabilities, including the processing of aesthetic input. Although MLLMs have the ability to process images, research focusing specifically on their ability to process abstract visual aesthetic input remains limited. In this study, we subjected five state-of-the-art MLLMs, namely, idefics2-8, llava-1.5-7b-hf, gemma-3n-E2B-it, moondream2, and llama-3.1-Nemotron-Nano-VL-8B-V1, to the Visual Aesthetic Sensitivity Test (VAST, Goetz et al., 1979) using zero-shot prompting to evaluate their ability to process abstract visual aesthetic stimuli. We found that MLLMs' processing of abstract visual aesthetic input is influenced by the interaction between prompt formulation and the model type. A noticeable gap exists between human and MLLMs responses in the VAST task, as reflected by their differing accuracy rates. Overall, while MLLMs show promising potential in aesthetic processing, their behavior differs noticeably from that of humans.

1 Introduction

In recent years, large language models (LLMs) have advanced at an unprecedented pace, achieving remarkable performance across a wide range of natural language understanding and generation tasks (Abe et al., 2024; Qin et al., 2024; Qiu et al., 2024; Wang et al., 2024). With the emergence of multimodal large language models (MLLMs), these systems have expanded into domains once regarded as uniquely human, including artistic creation and aesthetic evaluation (Hakopian, 2024; Yeo and Um, 2025; Khadangi et al., 2025). Although LLMs often generate outputs that appear human-like superficially, their underlying processing mechanisms may be fundamentally different from those of human cognition. There is an ongoing debate about whether these models truly emulate human-like thinking or simply mimic human patterns in response to prompts. Increasing amounts of research have tackled this debate head-on with an empirical approach, subjecting LLMs to many psychological experiments (Binz and Schulz, 2023; Kosinski, 2024; Cai et al., 2023; Qiu et al., 2025). For example, Binz and Schulz (2023) subjected GPT-3 to psychological experiments originally designed to study aspects of human cognition, such as decision-making, information search, and causal reasoning. They found that GPT-3 exhibited human-like or even better-than-human performance in tasks like gamble decisions and multiarmed bandit tasks, with signs of model-based reinforcement learning. In this study, we focused on another important aspect of human cognition: the ability to process the aesthetics of visual stimuli, such as appreciating the beauty of artworks, natural landscapes, and other visually pleasing patterns.

In philosophy and psychology, aesthetic judgment is a crucial subject (Beardsley, 1981; Martindale, 1988), and has been deeply explored by renowned figures like Immanuel Kant (Kant, 2024) and David Hume (Hume, 2017). In Critique of Judgment, Kant conceptualized the judgment of taste as the foundation of aesthetics, grounded in universal human sensibility. In contrast, Hume emphasized that aesthetics involves both cognitive and bodily factors. He suggested that aesthetic preferences are shaped by practice and by the ability to make nuanced comparisons. These philosophical perspectives laid the foundation for cognitive models of aesthetic judgment, aiming to explain how such judgments operate in human cognition. Building on these philosophical insights, Leder et al. (2004) introduced a cognitive model that outlines the five processes by which humans engage with aesthetics, including percep-

tual analysis, implicit memory activation, evaluative categorization, and emotional response and evaluation. This approach emphasizes the relationship between visual cues and learned cognitive responses, providing a framework for examining the processing of aesthetic stimuli. Further research shows that key features such as symmetry, complexity, and harmony are important determinants of aesthetic judgments in humans (Enquist and Arak, 1994). Eysenck (1940) proposed a new concept to explain the individual ability to make aesthetic judgment called aesthetic sensitivity. Aesthetic sensitivity refers to the ability to recognize and respond to subtle design elements that define beauty. These studies provide theoretical foundations for investigating AI systems that simulate human aesthetic perception.

Previous studies have explored the capability of MLLMs in visual recognition using large datasets of human labeled art (Huang et al., 2024; Fumanal-Idocin et al., 2023; Murray et al., 2012). For example, AesBench (Huang et al., 2024) offers a benchmark specifically designed to evaluate the ability of MLLMs to perceive and assess the aesthetic quality of images. However, all the above studies focus on concrete images, which limits their generalizability to highly abstract aesthetic content. In contrast, the Visual Aesthetic Sensitivity Test (VAST) aims to assess individual sensitivity to abstract visual forms. Over the years, VAST has been extensively used in human-centered studies to evaluate aesthetic preferences, demonstrating its reliability and validity as a measure of aesthetic sensitivity (Eysenck et al., 1984; Fróis and Eysenck, 1995; Myszkowski and Storme, 2017; Gear, 1986; Chan et al., 1980). Although extensively used in human-centered research, the potential application of VAST in artificial intelligence, especially in evaluating the aesthetic sensitivity of MLLMs, remains underexplored.

This study represents the first attempt to use well-validated psychological tests to investigate MLLMs' ability to process abstract visual input, offering new insights into whether AI systems can demonstrate abstract taste or aesthetic sensitivity. The research questions explored in this study are as follows:

- To what extent do MLLMs process abstract visual aesthetics in a way similar to humans?
- How do prompts influence MLLMs interpretation and evaluation of abstract visual aesthetics?

Model A	idefics2-8b
Developer	Hugging Face
Size	8B
Description	Instructional image-text model with OCR and visual reasoning ability.
Model B	llava-1.5-7b-hf
Developer	LLaVA Community
Size	7B
Description	Chat model based on LLaMA/Vicuna.
Model C	gemma-3n-E2B-it
Developer	Google
Size	∼2B effective
Description	Lightweight multimodal model for text, image, audio, video; 32K context.
Model D	moondream2
Developer	Vikhyat Korrapati
Size	∼1.9B
Description	Tiny, edge-optimized model for vision-language tasks like VQA and captioning.
Model E	llama-3.1-Nemotron-Nano-VL-8B-V1
Developer	NVIDIA
Size	8B
Description	Document-intelligent model with OCR, summarization, long-context support.

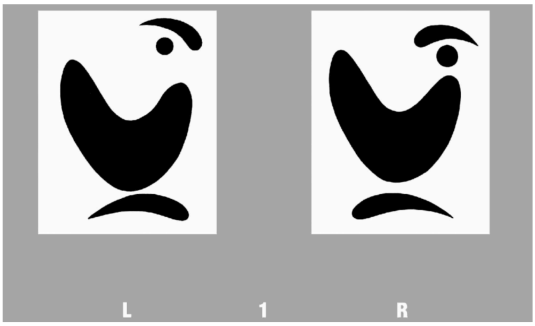
Table 1: Vertically arranged summary of the five MLLMs used in this study.

Our main contributions are as follows.

- Methodological Innovation: We introduce the first adaptation of the VAST for LLM-based evaluation, which enables the evaluation of aesthetic sensitivity in these models.
- Model Evaluation: We evaluate five MLLMs, including IDEFICS2-8, Llava-HF/Llava-1.5-7B-HF, Gemma-3N-E2B-IT, Moondream2, and Llama-3.1-Nemotron-Nano-VL-8B-V1, across 50 VAST items, and compared their responses with human normative data.
- Cognitive Insight: Building on Leder et al. (Leder et al., 2004), our approach provides new insights into how LLMs may simulate the cognitive processes behind aesthetic judgment.

2 Methodology

To address our research questions, we applied the visual aesthetic sensitivity test (VAST, Goetz et al., 1979) to obtain judgment data from five different MLLMs. The VAST is a reliable and well-validated instrument for assessing individual sensitivity to abstract visual forms. In the revised version of Goetz's study, participants viewed 50 pairs



There are two pictures in this image: the one on the left is labeled L, and the one on the right is labeled R. Describe both pictures, then tell me which one is the better design. You must make a clear and unambiguous choice, and justify it. Keep your response under 200 words.

Figure 1: An example VAST trial showing two abstract rooster-shaped images arranged side by side.

of non-representational pictures, of which one image in each pair had been intentionally altered by incorporating certain design faults. Participants were required to select the more aesthetically pleasing option, producing a quantitative score of individual aesthetic sensitivity. We selected five models for the VAST task based on their accessibility, cost efficiency, and ratings among the Hugging Face Image-Text-To-Text models. Table 1 provides a summary of the models included in our research. These models were accessed via the Hugging Face Hub and run on Colab GPUs.

We subjected each MLLM to the VAST task in a way similar to that of a human study. Each trial of the task represented a conversational session in which a PNG file containing two horizontally arranged pictures was presented, and the MLLM was prompted to select the better picture based on their visual aesthetic properties. Figure 1 illustrates an example trial of this experiment. In this trial, two abstract images that mimic the shape of a rooster were presented as visual input, while the text prompt read: 'There are two pictures in this image: the one on the left is labeled L and the one on the right is labeled R. Describe both pictures, then tell me which one is the better design. You must make a clear and unambiguous choice and justify it. Keep your response under 200 words.'

To explore our second research question, we designed five text prompts with slightly different wording and focus. These are listed in Table 2. Among them, Prompt 3 is adapted from the original formulation of the human study by Goetz et al. (1979, 796), which instructed participants to identify 'the most harmonious' design. Prompt 1 retains the original structure but removes the explicit reference to harmony, instead focusing on a general aesthetic comparison. Prompt 2 simplifies Prompt 1 by removing the requirement for picture descriptions. Prompt 4 replaces 'visual harmony' with 'aesthetic appeal' to shift the evaluative focus. Prompt 5 asks MLLMs to explicitly rely on their 'sensitivity to aesthetics', emphasizing subjective evaluation over descriptive analysis. We elicited the models' judgments in a zeroshot setting, where each model received only one text prompt and one PNG file in one conversational session. Each model analyzed all 50 VAST items five times, each with a different prompt. This led to 1250 pieces of model outputs from five models in total.

Prompt 1:
There are two pictures in this image: the one on the left is labeled L, and the one on the right is labeled R. Describe both pictures, then tell me which one is the better design. You must make a clear and unambiguous choice, and justify it. Keep your response under 200 words.
Prompt 2:
There are two pictures in this image: the one on the left is labeled L, and the one on the right is labeled R. Tell me which one is the better design. You must make a clear and unambiguous choice, and justify it. Keep your response under 200 words.
Prompt 3:
There are two pictures in this image: the one on the left is labeled L, and the one on the right is labeled R. Describe both pictures, and compare them in terms of their visual harmony. Then tell me which one is the better design. You must make a clear and unambiguous choice, and justify it. Keep your response under 200 words.
Prompt 4:
There are two pictures in this image: the one on the left is labeled L, and the one on the right is labeled R. Describe both pictures, and compare them in terms of their aesthetic appeal. Then tell me which one is the better design. You must make a clear and unambiguous choice, and justify it. Keep your response under 200 words.
Prompt 5:
There are two pictures in this image: the one on the left is labeled L, and the one on the right is labeled R. Describe both pictures, using your sensitivity to aesthetics to evaluate them. Then tell me which one is the better design. You must make a clear and unambiguous choice, and justify it. Keep your response under 200 words.

Table 2: The five prompts used to instruct MLLMs during the VAST trials.

To code the model output, we adopted a meta-evaluation approach in which each raw MLLM response was judged by two separate LLMs, namely Mistral 7B Instruct and LLaMA 3 8B Instruct, to classify the choice as left (L), right (R), or unclear (N). The judging models' decisions were parsed from their outputs and recorded as L, R, or NA.

Model	Count	Accuracy
A	210	0.538
B	241	0.481
C	250	0.468
D	241	0.452
E	238	0.542

Table 3: Count and accuracy by model. A: idefics2-8b, B: llava-1.5-7b-hf, C: gemma-3n-E2B-it, D: moondream2, E: llama-3.1-Nemotron-Nano-VL-8B-V1

Prompt	Count	Accuracy
1	238	0.496
2	238	0.487
3	235	0.506
4	233	0.476
5	236	0.508

Table 4: Count and accuracy by prompt.

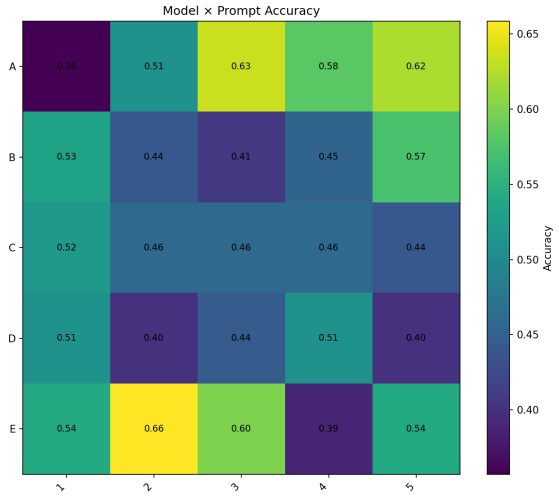


Figure 2: Heatmap of Accuracy for ModelPrompt Pairs

Empty or invalid responses were also coded as NA. The coded results, along with the original responses, were stored for further analysis. Disagreements between the two judging models were explicitly marked and resolved by two human researchers. The script and data used for the statistical analysis in the following section are publicly available via GitHub[1]

3 Results

Of the 1250 model responses to the VAST task, 70 responses were discarded due to the absence of an explicit choice between the right and left picture, leaving 1180 valid data points for analysis. Across all valid trials, the overall accuracy was 49.5%. Accuracy varied across models, ranging from 45.2% for Model D (moondream2) to 54.2% for Model E (llama-3.1-Nemotron-Nano-VL-8B-V1). Model A (idefics2-8b) and Model E achieved the highest accuracies, while Model D and Model C (gemma-3n-E2B-it) were the least accurate (see Table 3). Accuracy by prompt was relatively stable, ranging from 47.6% (Prompt 4) to 50.8% (Prompt 5)(see Table 4). The model-prompt breakdown revealed considerable variability. As shown in Figure 2, Model A performed best with Prompt 3 (63.4%) but poorly with Prompt 1 (35.7%). Conversely, Model E achieved its highest accuracy with Prompt 2 (65.9%) but its lowest with Prompt 4 (38.8%).

To assess the influence of model and prompt on accuracy, we fit a mixed-effects logistic regression model with random intercepts by image. The model included dummy-coded predictors for the model (reference: Model A) and prompt (reference: Prompt 1)[2]. Results from the GLMM showed that, relative to Model A, Model B ($\beta = 0.32$, 95% CI $= [0.06, 0.58]$), Model C ($\beta = 0.27$, 95% CI $= [0.01, 0.52]$), and Model E ($\beta = 0.38$, 95% CI $= [0.11, 0.64]$) had significantly higher accuracy. Model D did not differ significantly from Model A (95% CI $= [-0.03, 0.49]$). An omnibus Wald test showed that while the main effect of the model ($p = 0.04$) and the interaction effect ($p < 0.0001$) were significant, the main effect of prompt did not reach significance ($p > 0.05$). The comparison between model accuracy and human accuracy on the VAST task is shown in Table 5.

4 Discussion

This study investigated whether MLLMs possess an aesthetic sensitivity comparable to that of humans. We adopted a cognitive test originally designed to measure human visual aesthetic sensitivity (VAST, Goetz et al., 1979), and subjected MLLMs to the test. Our analysis revealed some noticeable response patterns across five leading MLLMs under five distinct prompt conditions. First, the result was significantly influenced by the interaction between prompt formulation and

[1] https://github.com/PON2020/vast

[2] Following a reviewer's suggestion, we constructed another GLMM, adding the interaction between the model and prompt. We confirmed the significance of the interaction effect; however, a significant interaction generally makes the main effects difficult to interpret. We included the script and results of that GLMM in the published GitHub repository for readers who are interested.

System	Mean Accuracy	Range
Human	0.6 (children), 0.7 (adults)	–
Model A	0.54	0.36–0.63
Model B	0.48	0.41–0.57
Model C	0.47	0.44–0.52
Model D	0.45	0.40–0.51
Model E	0.54	0.39–0.66

Table 5: Comparison of human and model accuracy on the VAST task. The human data was reported in Goetz et al. (1979). Range refers to the lowest and highest accuracy achieved. Model A: idefics2-8b, Model B: llava-1.5-7b-hf, Model C: gemma-3n-E2B-it, Model D: moondream2, Model E: llama-3.1-Nemotron-Nano-VL-8B-V1

the model type. This pattern indicates that current models may prioritize surface-level linguistic cues rather than engage in deeper forms of aesthetic reasoning. Second, we observed a significant variance in prompt robustness across models. Idefics2-8b and llama-3.1-Nemotron-Nano-VL-8B-V1 exhibited the highest variability across prompt conditions, suggesting a weaker generalization ability in aesthetic tasks. In contrast, Gemma-3n-E2B-it maintained relatively consistent performance, implying more stable internal representations. Third, the overall aesthetic judgment capabilities of MLLMs remained below human baselines. As shown in Table 5, human participants in the VAST task achieved an average accuracy of 0.7 for adults and 0.6 for children (Goetz et al., 1979). Although there was a clear variability among participants, human mean performance remained noticeably above chance, indicating that most human participants could perform the task with a fair degree of reliability. In contrast, our best-performing models reached a mean accuracy of 0.54, falling short of human means and never approaching the highest human scores (about 80% accuracy). This suggests that current MLLMs still lack the nuanced processing required for abstract visual evaluation.

Our findings are consistent with previous studies such as AesBench (Huang et al., 2024), showing a noticeable difference between MLLMs and human performance in aesthetic tasks. While AesBench evaluated models' performance in labeling and rating realistic photographs, our study leveraged the VAST framework, focusing on the cognitive ability of processing the aesthetics in abstract shapes. The observed differences between humans and MLLMs in the VAST could be explained by the cognitive model of aesthetic appreciation proposed by Leder et al. (2004), which describes aesthetic judgment as a process involving multiple stages. Human observers engage in this process by combining perceptual experience with cultural priors, personal background, and memory-based associations; by contrast, the processing of aesthetic images by MLLMs relies on statistical regularities of the input without accessing sensory or other embodied experiences. This distinction explains the patterns we observed in this study and offers valuable insight into ongoing debates on whether LLMs exhibit human-like cognitive mechanisms.

In conclusion, MLLMs processing of abstract visual aesthetic input is significantly influenced by the interaction between prompt formulation and the model type. A noticeable gap exists between human and MLLM responses in the processing of abstract visual input, as reflected in their different accuracy rates in the VAST task. In general, while MLLMs show promising potential in aesthetic processing, current MLLMs' behavior differs noticeably from that of humans.

4.1 Limitations and Future Work

We adopted the original VAST (Goetz et al., 1979) as a benchmark of aesthetic processing because of its standardized format and demonstrated reliability and validity in previous research. However, the normative human data was collected in 1979, raising concerns about its ability to reflect a contemporary aesthetic landscape. To address these gaps, future research should collect up-to-date human data to capture mainstream aesthetic preferences and then compare contemporary human responses with model data. It is also informative to test whether human responses are prompt-dependent, which allows researchers to claim more confidently whether the effect of the prompt in this study is unique to models or shared with humans. Last but not least, it is worth exploring the performance of frontier closed-source models (e.g., GPT-4V) on the same task to learn about upper bounds and transferability.

References

Yoshia Abe, Tatsuya Daikoku, and Yasuo Kuniyoshi. 2024. Assessing the aesthetic evaluation capabilities of gpt-4 with vision: Insights from group and individual assessments. In *38 (2024)*, pages 2Q1IS301–2Q1IS301. .

Monroe C Beardsley. 1981. *Aesthetics, problems in the philosophy of criticism*. Hackett Publishing.

Marcel Binz and Eric Schulz. 2023. Using cognitive psychology to understand gpt-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120.

Zhenguang G Cai, Xufeng Duan, David A Haslett, Shuqi Wang, and Martin J Pickering. 2023. Do large language models resemble humans in language use? *arXiv preprint arXiv:2303.08014*.

J Chan, Hans J Eysenck, and Karl O Götz. 1980. A new visual aesthetic sensitivity test: Iii. cross-cultural comparison between hong kong children and adults, and english and japanese samples. *Perceptual and Motor Skills*, 50(3_suppl):1325–1326.

Magnus Enquist and Anthony Arak. 1994. Symmetry, beauty and evolution. *Nature*, 372(6502):169–172.

Hans J Eysenck, KO Götz, Han Yee Long, DKB Nias, and M Ross. 1984. A new visual aesthetic sensitivity testiv. cross-cultural comparisons between a chinese sample from singapore and an english sample. *Personality and Individual Differences*, 5(5):599–600.

Hans Jurgen Eysenck. 1940. The general factor in aesthetic judgements 1. *British Journal of Psychology. General Section*, 31(1):94–102.

João Pedro Fróis and Hans J Eysenck. 1995. The visual aesthetic sensitivity test applied to portuguese children and fine arts students. *Creativity Research Journal*, 8(3):277–284.

Javier Fumanal-Idocin, Javier Andreu-Perez, Oscar Cordón, Hani Hagras, and Humberto Bustince. 2023. Artxai: Explainable artificial intelligence curates deep representation learning for artistic images using fuzzy techniques. *IEEE Transactions on Fuzzy Systems*, 32(4):1915–1926.

Jane Gear. 1986. Eysenck's visual aesthetic sensitivity test (vast) as an example of the need for explicitness and awareness of context in empirical aesthetics. *Poetics*, 15(4-6):555–564.

Karl O Goetz, R Lynn, A. R. Borisy, and Hans J Eysenck. 1979. A new visual aesthetic sensitivity test: I. construction and psychometric properties. *Perceptual and Motor Skills*, 49(3):795–802.

Mashinka Firunts Hakopian. 2024. Art histories from nowhere: on the coloniality of experiments in art and artificial intelligence. *AI & SOCIETY*, 39(1):29–41.

Yipo Huang, Quan Yuan, Xiangfei Sheng, Zhichao Yang, Haoning Wu, Pengfei Chen, Yuzhe Yang, Leida Li, and Weisi Lin. 2024. Aesbench: An expert benchmark for multimodal large language models on image aesthetics perception. *arXiv preprint arXiv:2401.08276*.

David Hume. 2017. Of the standard of taste. In *Aesthetics*, pages 483–488. Routledge.

Immanuel Kant. 2024. *Critique of judgment*, volume 10. Minerva Heritage Press.

Afshin Khadangi, Amir Sartipi, Igor Tchappi, and Gilbert Fridgen. 2025. Cognartive: Large language models for automating art analysis and decoding aesthetic elements. *arXiv preprint arXiv:2502.04353*.

Michal Kosinski. 2024. Evaluating large language models in theory of mind tasks. *Proceedings of the National Academy of Sciences*, 121(45):e2405460121.

Helmut Leder, Benno Belke, Andries Oeberst, and Dorothee Augustin. 2004. A model of aesthetic appreciation and aesthetic judgments. *British Journal of Psychology*, 95(4):489–508.

Colin Martindale. 1988. Aesthetics, psychobiology, and cognition. In Frank H. Farley and Ronald W. Neperud, editors, *The Foundations of Aesthetics, Art, and Art Education*, pages 7–42. Praeger, New York.

Naila Murray, Luca Marchesotti, and Florent Perronnin. 2012. Ava: A large-scale database for aesthetic visual analysis. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 2408–2415. IEEE.

Nils Myszkowski and Martin Storme. 2017. Measuring good taste with the visual aesthetic sensitivity test-revised (vast-r). *Personality and Individual Differences*, 117:91–100.

Libo Qin, Qiguang Chen, Xiachong Feng, Yang Wu, Yongheng Zhang, Yinghui Li, Min Li, Wanxiang Che, and Philip S Yu. 2024. Large language models meet nlp: A survey. *arXiv preprint arXiv:2405.12819*.

Zhuang Qiu, Xufeng Duan, and Zhenguang G Cai. 2025. Grammaticality representation in chatgpt as compared to linguists and laypeople. *Humanities and Social Sciences Communications*, 12(1):1–15.

Zhuang Qiu, Peizhi Yan, and Zhenguang Cai. 2024. Large language models for second language english writing assessments: An exploratory comparison. In *Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation*, pages 363–370.

Jiaqi Wang, Hanqi Jiang, Yiheng Liu, Chong Ma, Xu Zhang, Yi Pan, Mengyuan Liu, Peiran Gu, Sichen Xia, Wenjun Li, Yutong Zhang, Zihao Wu, Zhengliang Liu, Tianyang Zhong, Bao Ge, Tuo Zhang, Ning Qiang, Xintao Hu, Xi Jiang, Xin Zhang, Wei Zhang, Dinggang Shen, Tianming Liu, and Shu Zhang. 2024. A comprehensive review of multimodal large language models: Performance and challenges across different tasks. *Preprint*, arXiv:2408.01319.

Yunha Yeo and Daeho Um. 2025. Can ai recognize the style of art? analyzing aesthetics through the lens of style transfer. *arXiv preprint arXiv:2504.14272*.