

Can GPTs Evaluate Graphic Design Based on Design Principles?

Daichi Haraguchi
CyberAgent
Japan

Naoto Inoue
CyberAgent
Japan

Wataru Shimoda
CyberAgent
Japan

Hayato Mitani
Kyushu University
Japan

Seiichi Uchida
Kyushu University
Japan

Kota Yamaguchi
CyberAgent
Japan

"Please rate between 1 to 10 points"

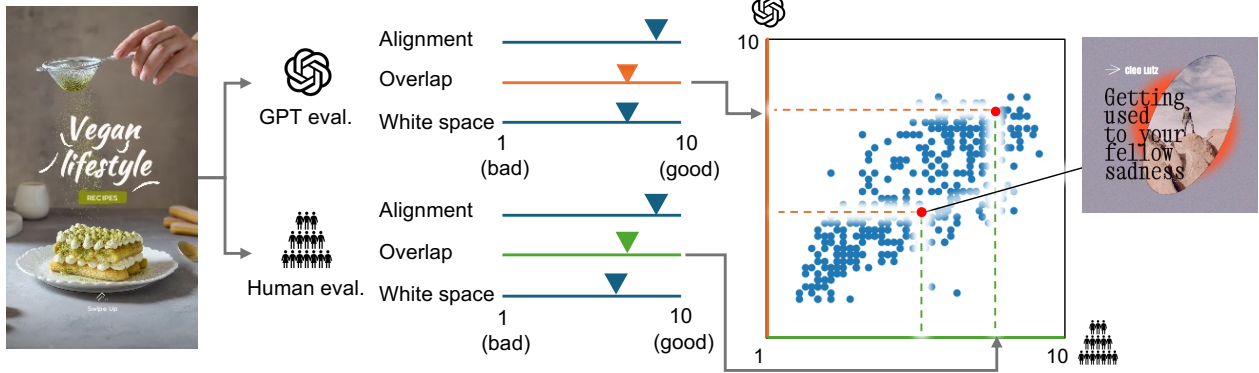


Figure 1: Overview of our study. We investigate GPT-based evaluation ability for graphic designs in three design principles “alignment,” “overlap,” and “white space.”

Abstract

Recent advancements in foundation models show promising capability in graphic design generation. Several studies have started employing Large Multimodal Models (LMMs) to evaluate graphic designs, assuming that LMMs can properly assess their quality, but it is unclear if the evaluation is reliable. One way to evaluate the quality of graphic design is to assess whether the design adheres to fundamental graphic design principles, which are the designer’s common practice. In this paper, we compare the behavior of GPT-based evaluation and heuristic evaluation based on design principles using human annotations collected from 60 subjects. Our experiments reveal that, while GPTs cannot distinguish small details, they have a reasonably good correlation with human annotation and exhibit a similar tendency to heuristic metrics based on design principles, suggesting that they are indeed capable of assessing the quality of graphic design. Our dataset is available at <https://cyberagentailab.github.io/Graphic-design-evaluation/>.

CCS Concepts

• **Computing methodologies** → **Computer vision**; **Computer graphics**.

Keywords

Large Multimodal Model, Large Language Model, Graphic Design, Graphic Design Evaluation

ACM Reference Format:

Daichi Haraguchi, Naoto Inoue, Wataru Shimoda, Hayato Mitani, Seiichi Uchida, and Kota Yamaguchi. 2024. Can GPTs Evaluate Graphic Design Based on Design Principles?. In *Proceedings of SIGGRAPH Asia (SA '24)*. ACM, New York, NY, USA, 6 pages. <https://doi.org/XXXXXXX.XXXXXXX>

1 Introduction

Foundation models learn from a large-scale corpus and exhibit remarkable generalization capability across various tasks, and the same is true for graphic design tasks [Chen et al. 2024; Cheng et al. 2024; Inoue et al. 2024; Jia et al. 2023]. While successful in certain graphic design tasks, it is still not apparent whether foundation models, such as GPT-4o, can reliably judge the quality of graphic design. Generally, high-quality graphic designs tend to follow design principles that are the designer’s common practices, such as alignment or repetition, as described in [Graham 2002; Williams 2014]. While humans can judge whether a given design follows these principles, human evaluation is time-consuming and not scalable. One of the early attempts in automatic evaluation is [O’Donovan

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

SA '24, December 03–06, 2024, Tokyo, Japan

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-XXXX-X/18/06

<https://doi.org/XXXXXXX.XXXXXXX>

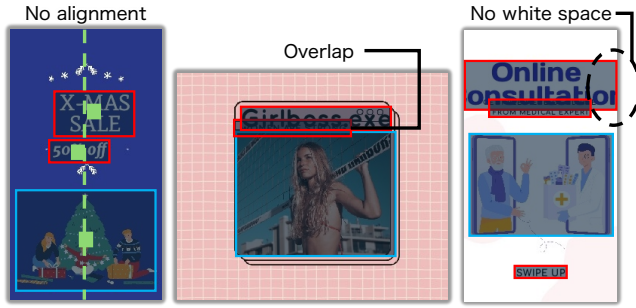


Figure 2: Negative examples of three design principles.

et al. 2014], which introduces hand-crafted metrics for optimization. Recent studies of graphic design generation have employed Large Multimodal Models (LMMs), particularly GPT-4 [Achiam et al. 2023], to directly estimate the quality. However, these studies do not consider the viewpoints of design principles for the evaluation.

In this paper, we quantitatively study the behavior of GPTs with respect to human evaluators in assessing graphic designs, as illustrated in Figure 1. To investigate the performance of GPTs, we employ heuristic metrics as baseline. We compare these approaches across three representative design principles, alignment, overlap, and white space, by manipulating graphic designs. We curate a dataset of graphic banner designs from an online service and perturb original designs to artificially generate low-quality designs for evaluation. Then, we ask human subjects to rate those designs in terms of the design principles. This human rate annotation allows us to assess which method – heuristic evaluation or GPTs – better aligns with human evaluation. We also discuss the qualitative comparison of heuristic and GPT-based evaluation.

We summarize our contributions in the following.

- We empirically study the relationship between hand-crafted and GPT-based evaluation for graphic designs with respect to human evaluation.
- We build a human-rated dataset of graphic designs that have varying degrees of quality to study the design quality.
- We find a higher correlation between human annotation and GPTs than heuristic metrics, suggesting that GPTs can give reliable judgment for graphic design quality under certain conditions.

2 Graphic design principles

Graphic design principles are the designer’s common practices to create aesthetic work [Graham 2002; Williams 2014]. In this paper, we employ representative design principles used in several studies of graphic design generation [Kong et al. 2022; O’Donovan et al. 2014]: *alignment*, *overlap*, and *white space*. Alignment and overlap are commonly used in evaluation for layout generation, such as [Li et al. 2020]. White space is also a critical factor in graphic design generation approaches [Kong et al. 2022]. We follow three design principles based on [O’Donovan et al. 2014]. Below, we briefly describe each principle.

Alignment. We consider the arrangement of elements such that their edges line up along common rows or columns to express a sense of order and structure.

- (1) Alignment along with the horizontal and vertical direction is considered.
- (2) The elements that align at a glance but slight misalignment are penalized because it is visually displeasing.
- (3) Larger alignment groups (i.e., aligned elements distant from each other) are preferred as they produce simpler designs with more unity between elements.

Overlap. Inappropriate overlap reduces readability. We consider the following aspects.

- (1) The three types of overlap, the overlap of elements on text, the overlap of text on graphics, and the overlap of graphics on other graphics, are considered.
- (2) Hard-to-read text because of insufficient color contrast between a text and the background color is penalized.
- (3) The graphic design that includes elements extending past the boundaries is also penalized.

White space. White space is for the appropriate amount of space in a design for better readability.

- (1) A large ratio of white space that is not covered by design elements (e.g., graphics and texts) is preferred.
- (2) However, the graphic design with a too large region of empty white space on the image is undesirable.
- (3) The greater distance between each element is preferred.
- (4) Uniformed vertical spacing of each text element is preferred.
- (5) Wider border margins (i.e., the white space at the edges of the image) for each element are preferred.

For better understanding, we show some visually unappealing examples because they violate the principles in Figure 2.

3 Evaluation approach

In this paper, we compare the evaluation metrics of the heuristic approach and the GPT-based approach. In both approaches, we input a graphic design and then obtain the score of the input. We set the lower and upper bounds for both metrics, where the higher values indicate better quality.

3.1 Heuristic evaluation metrics

We employ the hand-crafted formulation in [O’Donovan et al. 2014] as heuristic evaluation metrics, which were originally designed for graphic design optimization. We adopt the formulas related to the alignment, overlap, and white space as evaluation metrics. These metrics directly consider the size and coordinates of text or other graphic elements. We normalize the metric range from 0 to 1.

3.2 GPT-based evaluation metrics

We give prompts to GPT-4o¹ and ask for scores in terms of alignment, overlap, and white space. We made the prompt based on explanations of design principles and formulas described in [O’Donovan et al. 2014]. See the Appendix for more details. We render and rasterize the original design to create an input prompt. We show an

¹<https://openai.com/index/hello-gpt-4o/>

Example of instruction:
Please rate between 1 to 10 points. Assess the graphic design in terms of [the name of design principle] from the following perspectives. [The design principles. (See subsection for each design principle)]



Figure 3: How to rate graphic designs by GPT and humans. The detailed input prompts for the GPT is described in the Appendix.

example in Figure 3, where GPT-4o gives scores from 1 to 10 to the input based on a specific aspect.

4 Dataset

We collect graphic design templates from VistaCreate², which hosts a large number of banner or poster designs. The templates contain coordinate and size information for graphic and text elements within a design. We randomly sampled one hundred templates from VistaCreate and perturbed coordinate and size parameters to create aesthetically inferior samples. We apply two kinds of perturbations in our experiments: x -coordinate and font size. We perturb the x -coordinate of the text position to evaluate alignment and the font size of text elements to evaluate overlap and white space. We give three ranges of perturbation, small, medium, and large, to investigate the sensitivity of metrics. We combined the 100 original samples and 600 perturbed samples to build a dataset of 700 samples.

For each design, we collect human scores for alignment, overlap, and white space. We recruited 60 participants via crowdsourcing and collected five annotations per graphic design, where we asked participants for scores in the 1 to 10 range, as shown in Figure 3. We use the average score of five annotations in our experiments.

5 Experiments

5.1 Quantitative evaluation

Setup For a fair comparison with human evaluation and detailed analyses, we conduct the GPT-based evaluation five times and use the average score for the GPT score. We set a sampling temperature, which controls the randomness of the GPT, to 1 (default value). Higher temperature (e.g., 0.9) shows more diverse output. Therefore, our experimental setting takes the diversity into account.

Correlation to human evaluation To analyze the correlation between human annotation and the scores by automatic evaluation, we prepare two types of scatter plots of evaluation scores: one comparing heuristic metrics with human evaluation and the other comparing GPT-4o with human evaluation for each design principle, as shown in Figure 4. Interestingly, GPT-4o shows a better correlation with human evaluation than heuristic metrics. This may

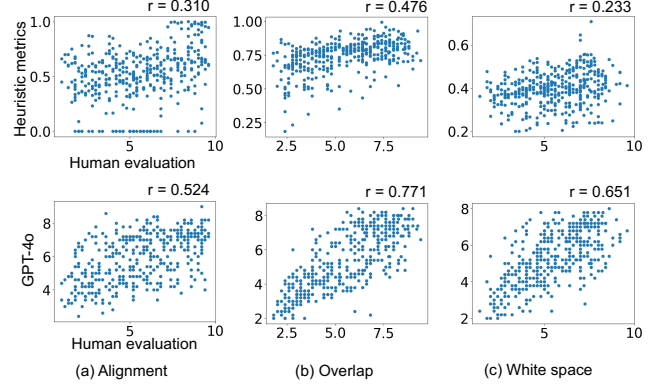


Figure 4: Correlation between human annotation and heuristic or GPT scores. The r is the Pearson correlation coefficient.



Figure 5: Graphic designs with their alignment scores.

be because heuristic metrics are designed to optimize a graphic design by comparing the graphic design before and after optimization but not for comparing two totally different designs.

We show an example of completely different graphic designs and their scores in Figure 5. According to the heuristic metrics, the order of better design is (a), (b), and (c). However, the order according to GPT-4o and human evaluation is (a), (c) (or (a) equal (c)), and (b). The heuristic evaluation quite degrades the score between (a) and (c) despite the fact that both graphic designs are created by human designers. On the other hand, GPT-4o and human evaluation are not much different. The results suggest that GPT-4o could be a suitable approach for quality evaluation for recent LMM-based generators since they generate significantly different designs by altering sampling parameters (or random seeds).

Sensitivity We investigate whether GPT-based evaluation is effective across designs of diverse qualities by calculating the correlation coefficients for each perturbation level, as shown in Figure 6. GPT scores exhibit a stronger correlation with human annotation compared to heuristic scores across all metrics and perturbation levels. Additionally, the correlation of GPT scores tends to increase as perturbation levels rise. This implies that when perturbations are significant, such as in the case of a graphic design with a noticeably

²<https://create.vista.com>

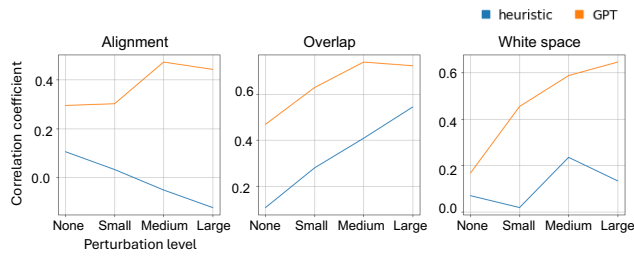


Figure 6: Correlation coefficient of the scores between human evaluation and each method.

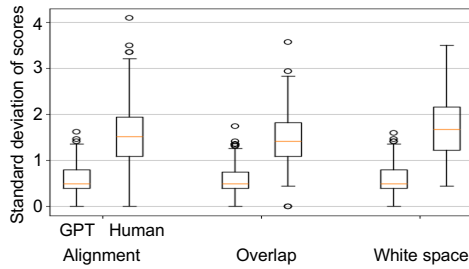


Figure 7: Box plot of the standard deviation of scores.

poor appearance, achieving a stable evaluation becomes easier and more efficient than humans.

Reliability We investigate how stable the GPT scores are since running GPT multiple times may give different scores. We show the standard deviation of the GPT scores and human annotation in Figure 7. The standard deviation of GPT scores is lower than that of the human evaluation scores. This result suggests that GPT scores with a single run for each principle are practical and cost-effective evaluation metrics.

5.2 Qualitative analysis

We show typical cases where GPT-based evaluation succeeds while heuristic evaluation fails and vice versa. We show cases where only GPT-based evaluation succeeds in Figure 8. The design includes objects in the background, but heuristic evaluation considers the background as white space. Heuristic evaluation is difficult when assessing graphic designs that include an object embedded in the background.

We also show cases where only heuristic evaluation succeeds in Figure 9. Since heuristic metrics are vector-based, they can capture slight differences in the design directly from the vector values. However, GPT-based evaluation struggles to detect such differences accurately. A similar limitation of GPT has also been reported in another study [You et al. 2023].

6 Conclusion and Future Work

In this paper, we investigated the appropriateness of using heuristic evaluation metrics and GPT-4o for evaluating graphic designs focusing on design principles. To achieve this, we collected the large-scale human annotation and analyzed the correlation between the annotation and each evaluation metric. Our experiment showed that GPT-based evaluation better correlates with human annotation.

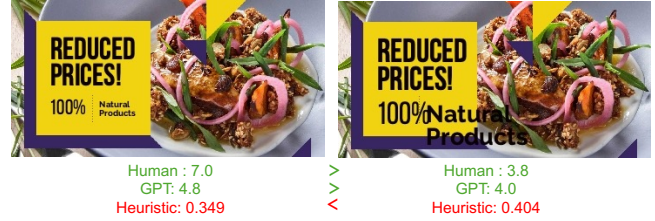


Figure 8: A sample with a correct white space assessment by GPT.

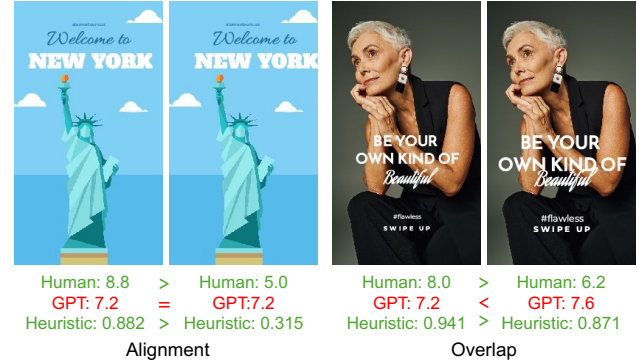


Figure 9: The examples of correctly assessed samples by heuristic evaluation.

For future work, we will jointly evaluate several design principles to assess the overall goodness of graphic designs. Additionally, we plan to evaluate graphic designs beyond design principles, such as font choices.

Acknowledgments

A part of this work was supported by JST ACT-X (Grant No. JPM-JAX22AD).

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).
- Jingye Chen, Yupan Huang, Tengchao Lv, Lei Cui, Qifeng Chen, and Furu Wei. 2024. TextDiffuser-2: Unleashing the Power of Language Models for Text Rendering. In *ECCV*.
- Yutao Cheng, Zhao Zhang, Maoke Yang, Nie Hui, Chunyuan Li, Xinglong Wu, and Jie Shao. 2024. Graphic Design with Large Multimodal Model. *arXiv preprint arXiv:2404.14368* (2024).
- L. Graham. 2002. *Basics of Design: Layout and Typography for Beginners*. Delmar Cengage Learning.
- Naoto Inoue, Kento Masui, Wataru Shimoda, and Kota Yamaguchi. 2024. OpenCOLE: Towards Reproducible Automatic Graphic Design Generation. In *CVPRW*.
- Peidong Jia, Chenxuan Li, Zeyu Liu, Yichao Shen, Xingru Chen, Yuhui Yuan, Yinglin Zheng, Dong Chen, Ji Li, Xiaodong Xie, et al. 2023. COLE: A Hierarchical Generation Framework for Graphic Design. *arXiv preprint arXiv:2311.16974* (2023).
- Wenyuan Kong, Zhaoyun Jiang, Shizhao Sun, Zhuoning Guo, Weiwei Cui, Ting Liu, Jianguang Lou, and Dongmei Zhang. 2022. Aesthetics++: Refining graphic designs by exploring design principles and human preference. *IEEE TVCG* 29, 6 (2022).
- Jianan Li, Jimei Yang, Jianming Zhang, Chang Liu, Christina Wang, and Tingfa Xu. 2020. Attribute-conditioned layout gan for automatic graphic design. *IEEE TVCG* 27, 10 (2020).
- Peter O'Donovan, Aseem Agarwala, and Aaron Hertzmann. 2014. Learning layouts for single-page graphic designs. *IEEE TVCG* 20, 8 (2014).
- Robin Williams. 2014. *Non-Designer's Design Book*. Peachpit Press.

Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. 2023. Ferret: Refer and ground anything anywhere at any granularity. *arXiv preprint arXiv:2310.07704* (2023).

A Details of design principles

We summarize the design principles used in our experiments. These design principles are based on [O'Donovan et al. 2014].

A.1 Alignment

This metric assesses the alignment of graphic and text elements and consists of three major components. First, it assesses whether there is alignment between neighboring elements. Second, it evaluates the alignment error between elements intended to be precisely aligned within a neighborhood. If the elements appear to be aligned but exhibit slight misalignment, the evaluation is based on the degree of this misalignment. Third, the metric considers alignment among distant elements. If the elements between two separated elements are aligned with these two elements, both distant elements are also considered aligned and are evaluated accordingly.

We transcribe the alignment principle to the prompt below.

Correct alignment is an important aspect of design that has been modeled in other layout applications. Text and graphic elements are aligned on the page to indicate organizational structure and aesthetics. Please evaluate the alignment of the input graphic design considering the following points.

1. Alignment along with the horizontal and vertical direction is considered.
2. The elements that align at a glance but slight misalignment are penalized because it is visually displeasing.
3. Larger alignment groups (i.e., aligned elements that are distant from each other) are preferred as they produce simpler designs with more unity between elements.

A.2 Overlap

This metric primarily assesses the overlap of graphic and text elements within the background. It consists of three major components. First, it measures the color change between the target area before and after rendering texts, considering whether the text is rendered against a background of the different color. Second, it calculates the percentage of overlap for each element, accounting for overlaps not only between text elements but also between text and graphics. Third, it assesses the percentage of the elements' area that extends beyond the canvas, measuring the extent to which text and other elements are appropriately contained within the background.

We transcribe the overlap principle to the prompt below.

Overlapping elements are common in many designs and absent from others. Less or proper overlapping might be considered aesthetically pleasing, but others are not. Please consider the following points to evaluate the overlap.

1. The three types of overlap, the overlap of elements on text, the overlap of text on graphics, and the overlap of graphics on other graphics, are considered.
2. Hard-to-read text because of insufficient color contrast between a text and the background color is penalized.
3. The graphic design that includes elements extending past the boundaries is also penalized.

A.3 White space

This metric evaluates the appropriate amount of white space in a design and consists of five main components. First, it measures the percentage of white space in the image, determined by the proportion except for graphic and text elements relative to the overall image area. Generally, a larger amount of natural white space is considered better. Second, it includes a metric that penalizes designs where certain parts of the image have excessively large areas of white space. In contrast to the first component, this metric helps ensure a balanced distribution of white space by penalizing areas with unnatural white space. Third, the metric evaluates white space based on the distance between each element. This component considers the spacing between elements as a measure of white space adequacy. Fourth, it assesses the variance in the distance between text elements. Consistent spacing between texts is considered aesthetically pleasing and indicates uniform white space. Finally, the metric evaluates the white space at the edges of the image (i.e., border margin). An image lacking adequate white space at the edges is considered aesthetically poor.

We transcribe the white space principle to the prompt below.

White space in graphic designs is fundamental for readability and aesthetics. Element distance is also closely related to the principle of proximity, as elements placed near each other may appear to be related. White space also influences the overall design style; many modern designs use significant white space. White space 'trapped' between elements can also be distracting. Evaluate the white space considering the following points.

1. A large ratio of white space that is not covered by design elements (e.g., graphics and texts) is preferred.
2. However, the graphic design with a too large region of empty white space on the image is undesirable.
3. The greater the distance between each element is preferred.
4. Uniformed vertical spacing of each text element is preferred.
5. Wider border margins for each element are preferred.

B Detailed prompts for GPT evaluation

We create the prompt based on design principles described in [O'Donovan et al. 2014]. Note that our prompts include a part of the prompt used in [Jia et al. 2023]. In Section C, we also conduct a pairwise evaluation that directly compares two graphic designs and determines the superior one. Therefore, we describe the prompt for the absolute and pairwise evaluation here.

The prompt of absolute evaluation

You are an autonomous AI Assistant who aids designers by providing insightful, objective, and constructive critiques of graphic design projects. Your goals are: "Deliver comprehensive and unbiased evaluations of graphic designs based on the following design principles."

Grade seriously. The range of scores is from 1 to 10. A flawless design can earn 10 points, a mediocre design can only earn 7 points, a design with obvious shortcomings can only earn 4 points, and a very poor design can only earn 1-2 points.

[A design principle]

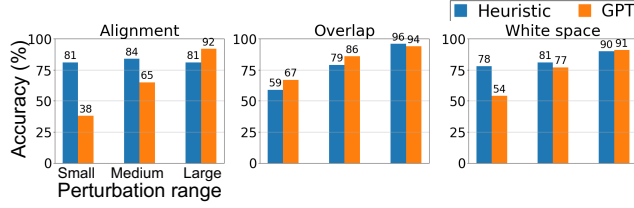


Figure 10: Comparison of pairwise evaluations.

If the output is too long, it will be truncated. Only respond in JSON format, no other information. Example of output for a better graphic design: {"score": 6, explanation: "(Please concisely explain the reason of the score.)"}
Please score the following images. [image]

Here, [A design principle] indicates the prompt described in Section A, [image] indicates placeholder of the input image.
The prompt of pairwise evaluation

You are an autonomous AI Assistant who aids designers by providing insightful, objective, and constructive critiques of graphic design projects. Your goals are: "Deliver comprehensive and unbiased evaluations of graphic designs based on the following design principles."
[A design principle]
If the output is too long, it will be truncated. Only respond in JSON format, no other information. Example of output for a better graphic design (a): {"better_design": "a", explanation: "(Please concisely explain the reason of choice.)"}
If both images are the same quality, answer {"better_design": "both", explanation: "(Please concisely explain the reason of choice.)"}
Which of the following graphic designs has better quality regarding the above-described points? (a) [image] (b) [image]

C Pairwise evaluation

We also conduct the pairwise evaluation, which directly assesses which graphic design is better, inputting two images. Two graphic designs are input to GPT-4o, and then GPT-4o answers which graphic design is better from the perspective of a specific design principle. The choices are not limited to “yes” and “no” but also include “not sure.” Each pair is evaluated five times, and the final result is determined by voting. We also conduct a similar evaluation task for humans to obtain the annotation.

We compare pairwise evaluation by GPT-4o with the heuristic evaluation. In heuristic evaluation, we compare the scores of before and after perturbation and determine the better one.

Figure 10 presents the pairwise evaluation results for each design principle. Across all metrics, the accuracy of GPT-4o increases progressively with the range of perturbation. For medium and large perturbations, GPT-4o performance is comparable to that of heuristic metrics, indicating that GPT-4o can distinguish between significantly poorer designs and others. However, GPT-4o performs worse than heuristic metrics for small perturbations in alignment and white space. There are two reasons for this. First, heuristic evaluation metrics are used to compare the two graphic designs before and after optimization. This is similar to the comparison of before and after perturbation designs. Second, as shown in the

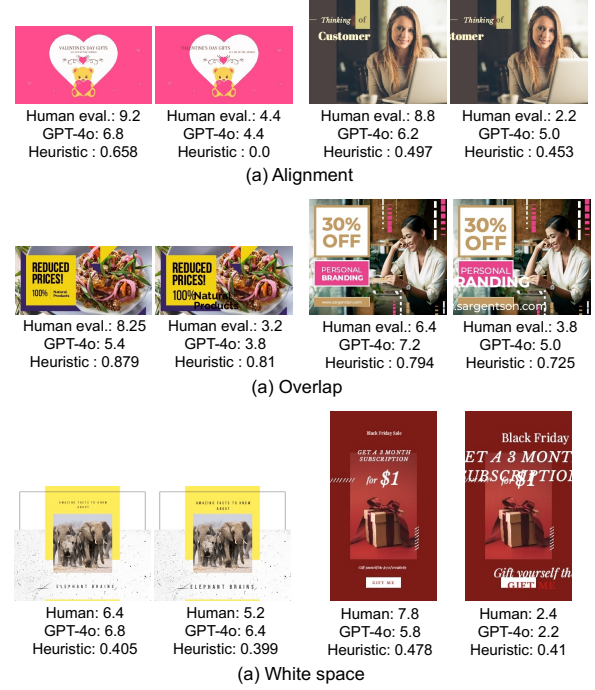


Figure 11: Examples of evaluation scores.

absolute evaluation, GPT-4o struggles to capture subtle differences in detailed design.

D Examples of evaluation scores

We show the examples of evaluation scores in Figure 11. From these examples, GPT-4o scores are more similar to those of a heuristic evaluation. In contrast, heuristic evaluation can be used to compare the before and after perturbation; however, comparing the different graphic designs is difficult, as described in Section 5.1.