BannerBench: Benchmarking Vision Language Models for Multi-Ad Selection with Human Preferences

Hiroto Otake[†] **Peinan Zhang**[‡], **Yusuke Sakai**[†],
Masato Mita[§], **Hiroki Ouchi**[†‡], **Taro Watanabe**[†]

[†]Nara Institute of Science and Technology, [‡]CyberAgent, [§]The University of Tokyo
otake.hiroto.od2@naist.ac.jp
{sakai.yusuke.sr9, hiroki.ouchi, taro}@is.naist.jp
zhang_peinan@cyberagent.co.jp
mita@g.ecc.u-tokyo.ac.jp

Abstract

Web banner advertisements, which are placed on websites to guide users to a targeted landing page (LP), are still often selected manually because human preferences are important in selecting which ads to deliver. To automate this process, we propose a new benchmark, **BannerBench**, to evaluate the human preference-driven banner selection process using vision-language models (VLMs). This benchmark assesses the degree of alignment with human preferences in two tasks: a ranking task and a best-choice task, both using sets of five images derived from a single LP. Our experiments show that VLMs are moderately correlated with human preferences on the ranking task. In the best-choice task, most VLMs perform close to chance level across various prompting strategies. These findings suggest that although VLMs have a basic understanding of human preferences, most of them struggle to pinpoint a single suitable option from many candidates. Our benchmark dataset is available at https://huggingface.co/datasets/cyberagent/BannerBench.

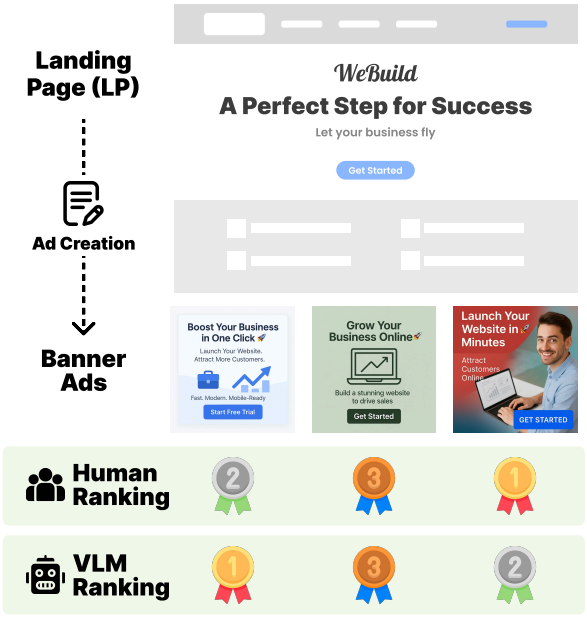

Figure 1: Automation of the banner evaluation process using VLMs: From an LP, multiple banners are created based on its content and persuasive appeal. The banner actually delivered is selected from these candidates through manual evaluation. A VLM that understands human preferences can autonomously perform this selection and evaluation process.

1 Introduction

Online advertisements (ads) are designed to engage consumers by appealing to their interests and preferences. As advertising techniques have matured, users frequently encounter ads on the internet. In particular, banner ads are placed on websites to direct users to a targeted landing page (LP) when clicked. Figure 1 illustrates a common workflow; multiple banners are created based on a single LP, and the task is to choose the most effective one for delivery. For effective ad delivery, it is crucial to select banners based on human preferences to encourage clicks. Selecting banners that match human preferences is important for delivering effective banners and encouraging clicks. Additionally, the banners should match the content of the LP to capture users' interest and encourage actual purchases or service sign-ups. Therefore, selecting appropriate banners for delivery is a crucial step in maximizing ad effectiveness. At the same time, recent progress in automated ad creation (Mishra et al., 2020; Chen et al., 2021a,b, 2025; Wang et al., 2025b,c; Gao et al., 2025), has made it easier to create many banners from a single LP. However, the selection process requires manual effort to choose the most appropriate banner. The large number of banners increases manual workload, highlighting the importance of developing fast and scalable methods for selecting the best banners to deliver.

To facilitate the automation of this process using vision-language models (VLMs), we propose a

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 24145–24159
November 4-9, 2025

new benchmark, **BannerBench**, which evaluates the ability of VLMs to identify the banner that best matches human preferences from a set of candidates. Specifically, BannerBench includes two tasks. The first is a ranking task, where five banners created from a single LP are sorted based on human preference. The second is a best-choice task, where the model selects the single most preferred banner, reflecting situations when only a single ad slot is available. These tasks are designed to evaluate VLMs' banner understanding capabilities from the perspective of human-aligned preference selection and their application in the ad delivery process.

Our experiments show that VLMs perform reasonably well on the ranking task, and the results are moderately correlated with human preferences. On the best-choice task, their performance drops, and the agreement rates are close to chance level regardless of different prompting strategies. With fewer options, some models show improved performance on the best-choice task. These findings suggest that although VLMs demonstrate a basic understanding of human preferences, most of them still struggle to make precise selections in special situations when required to select the single best option from many candidates.

2 Background and Related Work

2.1 Ads Deliver Process

To encourage specific user actions, such as making a purchase or submitting an inquiry, LPs are designed to provide essential information, including product details, service descriptions, pricing, and other elements that support the user's decision-making process. They also include actionable components like purchase buttons, inquiry forms, and sign-up links, which serve as clear pathways to promote user engagement. Banners are created based on the LP, ensuring consistency with its messaging and visual appeal. This alignment helps maximize the effectiveness of the banners when delivered to potential customers. Banner creatives are developed and delivered in accordance with the content and design of the LP. At the same time, insights gained from banner performance are continuously used to refine the LP. As the LP is updated, new creatives are subsequently produced to match the revised content, establishing a continuous improvement cycle between the LP and the banners. This process contributes to increased user engagement and the overall effectiveness of the campaign.

The content of banners is determined based on the target audience and the product or service being promoted, and multiple candidates are created before the final delivery. Figure 1 illustrates the process of evaluating these candidates to select the most suitable banner for delivery. Through this process, the final banner is chosen and delivered. The selection process involves heuristic evaluations, such as those based on human preferences, which necessitate human labor.

Furthermore, non-contextual ad delivery systems[1] are more common than context-specific tailoring, where the display context is determined automatically by the system rather than being manually tailored to each platform or audience. Advertisers are typically required only to prepare a limited number of variations aligned with general user preferences, while the platform handles placement dynamically based on topics or keywords. Our focus in this study is on the mainstream workflow of non-contextual delivery.

2.2 Ads Understanding

Several studies have introduced multi-faceted evaluation frameworks aligned with real-world deployment scenarios (Zhang et al., 2025), as well as methods that incorporate LP information into the evaluation process (Kamigaito et al., 2024). Furthermore, Murakami et al. (2025) investigated linguistic properties of ad text that influence human preferences. In the image processing field, image understanding has been explored through post-view behavior prediction tasks (Hussain et al., 2017). These tasks often leverage text embedded in images (Ye and Kovashka, 2018; Zhang et al., 2020; Kalra et al., 2020), and some methods employing VLMs have shown strong performance (Jia et al., 2023). However, while most existing studies focus on ad text or a single ad image, the handling of multiple ad images still lags behind.

2.3 Visual Alignment with Human Intuition

In general image understanding, several tasks have been proposed for multi-image reasoning, including dialogue evaluation across images (Huang et al., 2024) and comprehensive benchmarks for MLLM understanding in multi-image scenarios (Liu et al., 2024; Meng et al., 2025; Wang et al., 2025a; Chen et al., 2024; Ge et al., 2025). These tasks primarily

[1] https://support.google.com/google-ads/answer/2470108?hl=en

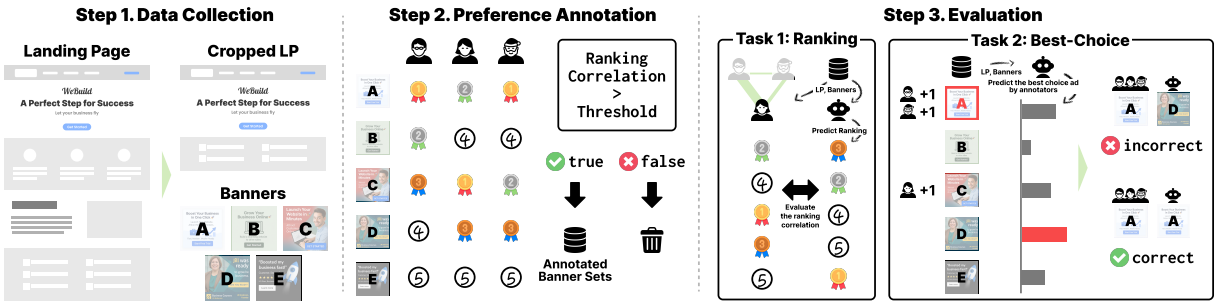

Figure 2: The dataset construction procedure consists of three steps. Step 1, we collected sets of one landing page (LP) and five banners (Banner Sets; BSs), cropping only the top portion of each LP. Step 2, we annotated each banner with human preference information and excluded any banner sets whose rank correlation score was below 0.6. Step 3, we designed two evaluation tasks: the ranking task, which evaluates the model by measuring the rank correlation between its output and human judgments, and the best-choice task, formulated as a selection task where the correct answer is defined as the banner most frequently preferred by human annotators, and the model is evaluated based on its accuracy in selecting that image. Further details are provided in Section 3.

evaluate abstract reasoning, such as temporal relations and dialogue consistency. IRR (Hayashi et al., 2025), an image-text task for evaluating model understanding of diverse perspectives, and Engagement Arena (Khurana et al., 2025), a benchmark for evaluating text-to-image models based on how engaging their outputs are. However, tasks involving multiple images and requiring human preference, such as banner selection, have not received enough attention.

3 BannerBench

We introduce **BannerBench**, a novel benchmark designed to evaluate the extent to which VLMs can align with human preferences in the banner selection process. As shown in Figure 2, we first collect sets of banners that share the same promotional content. Next, we annotate preference-based rankings for each set. Finally, we formalize these annotated sets into benchmark tasks specifically designed to assess how well VLMs align with human preferences in multiple banner selection scenarios.

3.1 Data Collection

Banners are designed to guide users to a specific LP. Therefore, the banners derived from the LP would share the same promotional messages and appeal points. We curated sets of five banners derived from a single LP (Banner Sets; BSs). Inspired by Hoßfeld et al. (2018), we cropped the LP images to the above-the-fold area[2] as shown in Figure 2-1. After cropping, LPs with low visual clarity or poor visibility were removed (see Appendix A), and all remaining LPs were manually reviewed for quality assurance. As a result, we obtained 900 LPs and 4,500 corresponding banners.

[2]Above-the-fold is used in web development to refer to the visible portions of a webpage without further scrolling or clicking: https://en.wikipedia.org/wiki/Above_the_fold.

Instruction for annotators

We present a landing page image and five banner images corresponding to the landing page.
Please rank from one to five in order that the contents inferred from the image of the landing page and the content that the advertisement image claims are compatible. However, note that the numbers from one to five are rankings, not scores, and the smaller they are, the more they match. In addition, there should be no duplication in the ranking, and be sure to use the ranking from one to five at a time.

- A landing page is the first dedicated web page that users who visit from advertisements and search results, and with a simple design and clear message, it is made for the purpose of encouraging concrete actions such as purchases and inquiries.

Figure 3: Details of instruction for annotators

3.2 Preference Annotation

In our annotation process, we decided to use human preference rather than actual ad delivery statistics such as Click-through Rate (CTR). Preference and CTR are fundamentally different metrics (Yang and Zhai, 2022). In practice, CTR is heavily influenced by external factors such as incentives and temporal trends, making it difficult to accurately assess the

inherent quality of an advertisement. This understanding is widely shared among professionals in banner selection, where preference-based evaluation is often considered more reliable than CTR due to the latter's sensitivity to such external factors. Moreover, we envision practical use cases where models are employed to help prioritize which ads to serve. In these scenarios, improving alignment with human preferences would directly contribute to better real-world performance. For these reasons, we evaluate banner selection performance from the perspective of user preference, which provides a more faithful and practical approach that aligns with real-world ad delivery workflows.

To evaluate whether the VLM can select images that align with human preferences from multiple banners, we annotated human preferences to the BSs. We hired five annotators per BS via the Lancers[3]. The annotators were presented with an LP and five randomly ordered banners. They ranked the banners, and to validate the consistency of the rankings, they were also asked to select the most and least preferable ones in combination with the given LP.

Figure 3 shows the instructions for annotators to rank the images. Each annotator was given 20 BSs, and they needed to pass the quality control questions. The detailed annotation process is described in the Appendix B. The total annotation cost was around 250 dollars. The inter-annotator agreement was evaluated using pair-wise Cohen's kappa, yielding a score of 0.499 ± 0.005, which indicates a moderate level of agreement in annotator preferences. Furthermore, for each set, we extracted the top half of all possible annotator pairs based on rank correlation, computed their average, and retained only those sets with an average value of at least 0.6.

3.3 Task Organization

To efficiently evaluate model capabilities using the annotated banners in different situations, we propose two subtasks: **Ranking** and **Best-Choice**.

Ranking. To assess whether VLMs can perform standard selection of banners, we treat the VLM as a "sixth annotator" and measure the highest correlation between its rankings and those of an arbitrary human annotator. We include only examples where the majority of annotators (three or more out of five) selected the same banner as the most preferred, in order to reduce noise from individual annotator variation. This results in a total of 547 examples for this subtask.

Best-Choice. To simulate a more specific banner selection process, we regard the banner chosen by the majority of annotators as the widely preferred one, and task the VLM with selecting it from a set of five. This also includes 547 examples.

4 Experimental Setting

VLMs. We benchmarked well-known VLMs, including GPT-4o-mini (OpenAI et al., 2024), Gemini-1.5-flash (Team et al., 2024), Gemini-2.0-flash (Deepmind, 2025), Qwen2-VL (Wang et al., 2024b), Qwen2.5-VL (Bai et al., 2025), LLaVA-NeXT (Li et al., 2024), Mantis (Jiang et al., 2024), and Llama3.2 (Grattafiori et al., 2024). More detailed model information is described in Appendix C.

Prompting. To account for positional bias (Chen et al., 2024; Wang et al., 2024a), we randomized the order of banner inputs before feeding them into the VLMs. We also analyze the impact of input order on model outputs in the Appendix D. To assess model capability in banner evaluation, we used three prompting strategies: **zero-shot**, **one-shot**, and **CoCoT** (Zhang et al., 2024), designed for multi-image understanding. For detailed experimental settings, such as prompt templates, decoding strategy, please refer to Appendix E.

Evaluation. In the **Ranking** task, we calculated all rank correlations and then took the average for the report. Note that the 2nd to 4th positions were treated as tied ranks. We average the rank correlations of the top 50 % annotator pairs per set, yielding a human baseline of 0.600. In the **Best-Choice** task, accuracy is measured by whether the model-selected banner from five candidates matches the human majority choice. We computed a weighted average over cases with 3, 4, and 5 annotator agreements, using weights of 0.600, 0.800, and 1.000, respectively, as the human baseline.

5 Experimental Results and Discussions

Ranking. Table 1 shows the average rank correlations between the VLMs and annotators when using three different prompting strategies. Across all prompting strategies, the average rank correlation was approximately **0.4**, indicating a moderate

[3]A crowdsourcing platform: https://www.lancers.jp.

	Ranking			Best-Choice			Best-Choice (Ablation: 2 ads)		
	zero-shot	one-shot	CoCoT	zero-shot	one-shot	CoCoT	zero-shot	one-shot	CoCoT
Chance Rate	0.000	0.000	0.000	0.200	0.200	0.200	0.500	0.500	0.500
Human Baseline	0.600	0.600	0.600	0.724	0.724	0.724	0.723	0.723	0.723
Gemini-1.5-flash	0.385	**0.440**	0.418	0.188	0.207	0.201	0.518	0.513	0.487
Gemini-2.0-flash	0.413	0.420	**0.444**	**0.221**	0.208	**0.257**	**0.729**	**0.745**	**0.719**
GPT-4o-mini	0.433	0.416	0.424	0.208	0.223	0.192	0.484	0.501	0.479
LLaVA-NeXT(Mistral-7B)	0.425	0.363	0.423	0.185	0.186	0.201	0.400	0.443	0.493
Mantis(CLIP-Llama3-8B)	0.436	0.416	0.426	0.192	0.172	0.177	0.475	0.482	0.491
Mantis(SigLIP-Llama3-8B)	**0.437**	0.415	0.429	0.194	**0.229**	0.179	0.496	0.499	0.491
Qwen2.5-VL(7B)	0.399	0.421	0.404	0.208	0.190	0.192	0.526	0.513	0.491
Qwen2-VL(7B)	0.411	0.415	0.425	0.197	0.150	0.197	0.577	0.627	0.614
Llama-3.2(11B)	0.422	0.423	0.411	0.199	0.214	0.199	0.498	0.504	0.495

Table 1: Benchmark results of various VLMs on BannerBench. The table on the right shows the results of a binary classification ablation study for the Best-Choice task. Ranking performance is reported using Cohen's kappa coefficient, while Best-Choice is evaluated by accuracy. **Bold** is the highest value, underline is the second one.

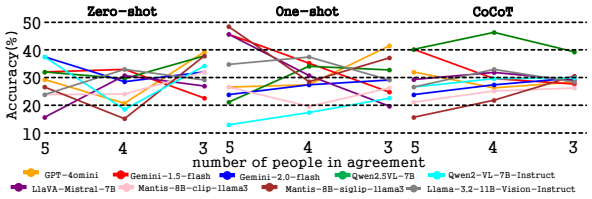

Figure 4: Accuracy by the number of annotators who selected the same banner as the most preferred.

correlation between rankings generated by models and annotators. These results suggest that the model tends to align with basic human preferences.

Best-Choice. Table 1 shows the agreement rates between the model outputs and human evaluations. For all prompting strategies, the agreement rate was approximately equal to the chance level of **0.200**. Figure 4 shows how the model's agreement rate varies based on the number of annotators who selected the same banner as the most preferred. For models, the agreement rate did not consistently vary with the number of annotators who agreed. These results suggest that when selecting a single preferred option, the model's choices do not align well with human preferences.

Ablation for Best-Choice. To simplify the **Best-Choice** task and assess whether models can identify the human-preferred banner naively, we conducted an ablation by extracting only the annotators' best and worst choices and framing it as a binary choice task. The right side of Table 1 shows that, across all prompting strategies, the agreement rates remained around the chance level for most models. In contrast, Gemini-2.0-flash and Qwen2-VL consistently exceeded **0.5** across all prompt strategies. Furthermore, the result of Gemini-2.0-flash was comparable to the human baseline, demonstrating a strong understanding of human preferences. These results show that fewer options help some models select the most suitable choice when human preferences are clearly differentiated.

6 Conclusion

We proposed BannerBench, a novel benchmark to evaluate whether VLMs can align with human preferences in the banner selection process. It covers two tasks: a ranking task and a best-choice task, using sets of five banners associated with a single LP. Our experiments show that although VLMs demonstrate a basic understanding of human preferences, most of them struggle to pinpoint a single suitable option. We hope this benchmark contributed to future research on developing and evaluating VLMs in the real-world ad delivery process.

7 Limitations

Dataset scale and diversity. This study constructs the dataset based on 900 LPs and 4,500 banners that were actually used in ad delivery, all of which were collected with proper permission. Since BannerBench is intended solely for evaluation purposes, it is not designed for training use; the benchmark focuses on assessing the inductive capabilities of VLMs. While the dataset could potentially be expanded, the primary goal of this work is to examine whether VLMs can identify banners that align with human preferences. Although the dataset size may seem relatively limited, it is comparable to many existing benchmarks for language model evaluation, where each task or subtask typi-

cally includes a few hundred examples (Maia Polo et al., 2024). Therefore, we consider the current scale of BannerBench appropriate for evaluation purposes. Moreover, while the dataset is collected from an advertising company, the data is a rare and high-quality resource based on real ad delivery logs. Such data, which was actually used in real ad delivery, holds significant value as evaluation data. Further scaling is certainly worth exploring in future work. However, the main novelty and contribution of this short paper lie in presenting the first benchmark that ambitiously targets the VLM-based automation of the banner selection process. As such, concerns about dataset size fall out of scope in this study.

VLMs and promptings. This study evaluates nine VLMs under three prompting strategies using our benchmark dataset. While it is certainly possible to include additional models or prompting methods, such expansions would be open-ended and are beyond the scope of this work. The primary objective of this study is to construct a benchmark dataset that supports the use of VLMs in real-world applications, particularly for banner selection. Therefore, we did not pursue extensive comparisons across a broader set of models or prompts. We believe the results obtained across the selected models provide adequate and meaningful insights.

Human preference annotations. BannerBench represents an ambitious initiative aimed at automating the banner selection step, a known bottleneck in the ad delivery process traditionally handled by human labor. This work revisits the ad delivery workflow, identifies its critical bottlenecks, and is the first to construct a dataset grounded in real-world advertising materials used in actual delivery scenarios. Although the dataset does not directly serve as a replacement for the entire ad delivery process, we believe it provides a valuable benchmark for exploring the potential of VLMs in this context. Importantly, rather than relying on historical delivery outcomes, human preference annotations were newly collected for this study. This decision was also made in accordance with a condition set by our data provider, which stated that actual delivery data, such as CTR, could only be shared if it was not made publicly available. Additionally, the banner selection process we target precedes performance-based feedback metrics such as CTR, which are influenced by multiple factors, including search engine optimization. Due to such confounding variables, these metrics are not typically used directly for banner selection. Considering this, our dataset offers a reasonable and practical first step toward real-world applicability. We publicly release as much information as possible within the bounds of data usage agreements.

Applicability of real-world scenarios. While BannerBench is grounded in the specific domain of banner selection, the underlying problem it addresses, aligning visual model preferences with nuanced human judgments when evaluating multiple image candidates, is not limited to advertising. We believe that the methodology of the benchmark construction, which involves ranking and selecting the best image from a set based on human preference, is transferable and relevant to a variety of other domains, such as applications in e-commerce (product image selection), content curation (social media posts), and design evaluation. All of these require models to understand and predict human visual preferences among multiple options. Therefore, while our dataset is from advertising, the principles and evaluation framework of BannerBench provide a foundation for addressing similar human preference alignment challenges in a broader range of visual application contexts.

Explainable evaluation. BannerBench is based on human preferences. Therefore, collecting annotators' reasons for their preferences could have enabled more fine-grained analysis. While such reasoning could provide additional insights, we chose not to adopt this approach in BannerBench, as it would significantly increase annotation costs and complicate its integration into our evaluation benchmark. We leave the integration of such reasoning into benchmark design to future work.

8 Ethical Considerations

Licence. The data used in this study were provided by an advertising company under a formal data-sharing agreement. Due to contractual restrictions, we were prohibited from using or disclosing performance-related metrics such as CTR. As such, our dataset includes only LPs and the corresponding banners, with no additional metadata or delivery logs used in the study. The dataset is distributed under a license restricting use to research purposes only, in accordance with the agreement with the data provider.

Annotation. For annotation, we ensured that all rights to the artifacts were transferred to the authors through explicit agreements. Annotators were recruited through a crowdsourcing platform, where compensation terms were clearly communicated and accepted before participation. All annotators were paid appropriately for their contributions.

Others. We carefully verified that the dataset does not contain any personally identifiable information, harmful content, or inappropriate material that would conflict with the data provider's guidelines. Therefore, we believe that this study fully complies with the ethical standards and considerations required for responsible research conduct.

References

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *Preprint*, arXiv:2502.13923.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. 2024. MLLM-as-a-judge: Assessing multimodal LLM-as-a-judge with vision-language benchmark. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 6562–6595. PMLR.

Jin Chen, Tiezheng Ge, Gangwei Jiang, Zhiqiang Zhang, Defu Lian, and Kai Zheng. 2021a. Efficient optimal selection for composited advertising creatives with tree structure. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(5):3967–3975.

Jin Chen, Ju Xu, Gangwei Jiang, Tiezheng Ge, Zhiqiang Zhang, Defu Lian, and Kai Zheng. 2021b. Automated creative optimization for e-commerce advertising. In *Proceedings of the Web Conference 2021*, WWW '21, page 2304–2313, New York, NY, USA. Association for Computing Machinery.

Xingye Chen, Wei Feng, Zhenbang Du, Weizhen Wang, Yanyin Chen, Haohan Wang, Linkai Liu, Yaoyu Li, Jinyuan Zhao, Yu Li, Zheng Zhang, Jingjing Lv, Junjie Shen, Zhangang Lin, Jingping Shao, Yuanjie Shao, Xinge You, Changxin Gao, and Nong Sang. 2025. Ctr-driven advertising image generation with multimodal large language models. In *Proceedings of the ACM on Web Conference 2025*, WWW '25, page 2262–2275, New York, NY, USA. Association for Computing Machinery.

Google Deepmind. 2025. Gemini 2.0 flash. https://deepmind.google/technologies/gemini/flash/.

Yifan Gao, Zihang Lin, Chuanbin Liu, Min Zhou, Tiezheng Ge, Bo Zheng, and Hongtao Xie. 2025. Postermaker: Towards high-quality product poster generation with accurate text rendering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8083–8093.

Wentao Ge, Shunian Chen, Hardy Chen, Nuo Chen, Junying Chen, Zhihong Chen, Wenya Xie, Shuo Yan, Chenghao Zhu, Ziyue Lin, Dingjie Song, Xidong Wang, Anningzhe Gao, Zhang Zhiyi, Jianquan Li, Xiang Wan, and Benyou Wang. 2025. MLLM-bench: Evaluating multimodal LLMs with per-sample criteria. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4951–4974, Albuquerque, New Mexico. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Kazuki Hayashi, Kazuma Onishi, Toma Suzuki, Yusuke Ide, Seiji Gobara, Shigeki Saito, Yusuke Sakai, Hidetaka Kamigaito, Katsuhiko Hayashi, and Taro Watanabe. 2025. IRR: Image review ranking framework for evaluating vision-language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9939–9956, Abu Dhabi, UAE. Association for Computational Linguistics.

Tobias Hoßfeld, Florian Metzger, and Dario Rossi. 2018. Speed index: Relating the industrial standard for user perceived web performance to web qoe. In *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6.

Yupan Huang, Zaiqiao Meng, Fangyu Liu, Yixuan Su, Nigel Collier, and Yutong Lu. 2024. Sparkles: Unlocking chats across multiple images for multimodal instruction-following models. In *ICLR 2024 Workshop on Navigating and Addressing Data Problems for Foundation Models*.

Zaeem Hussain, Mingda Zhang, Xiaozhong Zhang, Keren Ye, Christopher Thomas, Zuha Agha, Nathan Ong, and Adriana Kovashka. 2017. Automatic understanding of image and video advertisements. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhiwei Jia, Pradyumna Narayana, Arjun Akula, Garima Pruthi, Hao Su, Sugato Basu, and Varun Jampani. 2023. KAFA: Rethinking image ad understanding with knowledge-augmented feature adaptation of vision-language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 5: Industry Track)*, pages

772–785, Toronto, Canada. Association for Computational Linguistics.

Dongfu Jiang, Xuan He, Huaye Zeng, Cong Wei, Max Ku, Qian Liu, and Wenhu Chen. 2024. Mantis: Interleaved multi-image instruction tuning. *Transactions on Machine Learning Research*.

Kanika Kalra, Bhargav Kurma, Silpa Vadakkeeveetil Sreelatha, Manasi Patwardhan, and Shirish Karande. 2020. Understanding advertisements with BERT. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7542–7547, Online. Association for Computational Linguistics.

Hidetaka Kamigaito, Soichiro Murakami, Peinan Zhang, Hiroya Takamura, and Manabu Okumura. 2024. Generating attractive ad text by facilitating the reuse of landing page expressions. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 597–608, Tokyo, Japan. Association for Computational Linguistics.

Varun Khurana, Yaman Kumar Singla, Jayakumar Subramanian, Changyou Chen, Rajiv Ratn Shah, zhiqiang xu, and Balaji Krishnamurthy. 2025. Measuring and improving engagement of text-to-image generation models. In *The Thirteenth International Conference on Learning Representations*.

Bo Li, Kaichen Zhang, Hao Zhang, Dong Guo, Renrui Zhang, Feng Li, Yuanhan Zhang, Ziwei Liu, and Chunyuan Li. 2024. Llava-next: Stronger llms supercharge multimodal capabilities in the wild.

Haowei Liu, Xi Zhang, Haiyang Xu, Yaya Shi, Chaoya Jiang, Ming Yan, Ji Zhang, Fei Huang, Chunfeng Yuan, Bing Li, and Weiming Hu. 2024. MIBench: Evaluating multimodal large language models over multiple images. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 22417–22428, Miami, Florida, USA. Association for Computational Linguistics.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. 2024. tinyBenchmarks: evaluating LLMs with fewer examples. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 34303–34326. PMLR.

Fanqing Meng, Jin Wang, Chuanhao Li, Quanfeng Lu, Hao Tian, Tianshuo Yang, Jiaqi Liao, Xizhou Zhu, Jifeng Dai, Yu Qiao, Ping Luo, Kaipeng Zhang, and Wenqi Shao. 2025. MMIU: Multimodal multi-image understanding for evaluating large vision-language models. In *The Thirteenth International Conference on Learning Representations*.

Shaunak Mishra, Manisha Verma, Yichao Zhou, Kapil Thadani, and Wei Wang. 2020. Learning to create better ads: Generation and ranking approaches for ad creative refinement. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, CIKM '20, page 2653–2660, New York, NY, USA. Association for Computing Machinery.

Soichiro Murakami, Peinan Zhang, Hidetaka Kamigaito, Hiroya Takamura, and Manabu Okumura. 2025. AdParaphrase: Paraphrase dataset for analyzing linguistic features toward generating attractive ad texts. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1426–1439, Albuquerque, New Mexico. Association for Computational Linguistics.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, Soroosh Mariooryad, Yifan Ding, Xinyang Geng, Fred Alcober, Roy Frostig, Mark Omernick, Lexi Walker, Cosmin Paduraru, Christina Sorokin, and 1118 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

Fei Wang, Xingyu Fu, James Y. Huang, Zekun Li, Qin Liu, Xiaogeng Liu, Mingyu Derek Ma, Nan Xu, Wenxuan Zhou, Kai Zhang, Tianyi Lorena Yan, Wenjie Jacky Mo, Hsiang-Hui Liu, Pan Lu, Chunyuan Li, Chaowei Xiao, Kai-Wei Chang, Dan Roth, Sheng Zhang, and 2 others. 2025a. Muirbench: A comprehensive benchmark for robust multi-image understanding. In *The Thirteenth International Conference on Learning Representations*.

Haohan Wang, Wei Feng, Yaoyu Li, Zheng Zhang, Jingjing Lv, Junjie Shen, Zhangang Lin, and Jingping Shao. 2025b. Generate e-commerce product background by integrating category commonality and personalized style. In *ICASSP 2025 - 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Heng Wang, Yotaro Shimose, and Shingo Takamatsu. 2025c. Banneragency: Advertising banner design with multimodal llm agents. *Preprint*, arXiv:2503.11060.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024a. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450, Bangkok, Thailand. Association for Computational Linguistics.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin

Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024b. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *Preprint*, arXiv:2409.12191.

Brandon T. Willard and Rémi Louf. 2023. Efficient guided generation for large language models. *Preprint*, arXiv:2307.09702.

Yanwu Yang and Panyu Zhai. 2022. Click-through rate prediction in online advertising: A literature review. *Information Processing & Management*, 59(2):102853.

Keren Ye and Adriana Kovashka. 2018. Advise: Symbolism and external knowledge for decoding advertisements. In *Computer Vision – ECCV 2018*, pages 868–886, Cham. Springer International Publishing.

Daoan Zhang, Junming Yang, Hanjia Lyu, Zijian Jin, Yuan Yao, Mingkai Chen, and Jiebo Luo. 2024. Cocot: Contrastive chain-of-thought prompting for large multimodal models with multiple image inputs. *CoRR*, abs/2401.02582.

Huaizheng Zhang, Yong Luo, Qiming Ai, Yonggang Wen, and Han Hu. 2020. Look, read and feel: Benchmarking ads understanding with multimodal multitask learning. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 430–438, New York, NY, USA. Association for Computing Machinery.

Peinan Zhang, Yusuke Sakai, Masato Mita, Hiroki Ouchi, and Taro Watanabe. 2025. AdTEC: A unified benchmark for evaluating text quality in search engine advertising. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7672–7691, Albuquerque, New Mexico. Association for Computational Linguistics.

A Image Processing and Filtering

A.1 Above-the-Fold Cropping of LPs

We collected LP and corresponding banner sets. Web pages are progressively loaded over time as they are being displayed. To enhance user experience, it is better to place important information near the top of the page (Hoßfeld et al., 2018). Based on this, we assumed that key information is often placed in the upper sections of web pages and decided to focus on the top portion rather than analyzing the entire page. We cropped only the area that is visible on a monitor without scrolling. Assuming that web pages are rarely displayed in full screen, we cropped the upper part of each LP image to a size smaller than a typical display resolution, specifically, 1200 pixels wide and 800 pixels tall, and used these cropped images during the annotation process.

VLMs	HuggingFace ID / API Name
Gemini-1.5-flash	Gemini/gemini-1.5-flash
Gemini-2.0-flash	Gemini/gemini-2.0-flash
GPT-4o-mini	OpenAI/gpt-4o-mini-2024-07-18
LLaVA-NeXT(Mistral7B)	llava-hf/llava-v1.6-mistral-7b-hf
Mantis(CLIP-Llama3-8B)	TIGER-Lab/Mantis-8B-clip-llama3
Mantis(SigLIP-Llama3-8B)	TIGER-Lab/Mantis-8B-siglip-llama3
Qwen2.5-VL(7B)	Qwen/Qwen2.5-VL-7B-Instruct
Qwen2-VL(7B)	Qwen/Qwen2-VL-7B-Instruct
Llama3.2(11B)	meta-llama/Llama-3.2-11B-Vision-Instruct

Table 2: Details of the VLMs for the experiments.

A.2 Filtering Inappropriate LPs for Dataset

Some of the collected LP images had poor visibility in the upper portion of the screen, so we decided to include only those with sufficient visual clarity. The filtering of LP images was conducted systematically first, followed by manual inspection. For the systematic filtering, we applied a blurring process and excluded images in which a large portion was occupied by a single color after blurring. After this automatic filtering step, we manually removed LP images for which the promoted product or service could not be identified.

B Detailed Human Annotation Process

We asked annotators to rank five banners corresponding to an LP image in order of preference. As a quality check, we also asked them to select the most and least preferred banner. Figure 5 shows the annotation interface used for this task.

C Details of the VLMs

In the experiments in Section 5, we used publicly available VLMs, including Qwen2-VL (Wang et al., 2024b), Qwen2.5-VL (Bai et al., 2025), LLaVA-NeXT (Li et al., 2024), Mantis (Jiang et al., 2024), and Llama3.2 (Grattafiori et al., 2024), and GPT-4o-mini (OpenAI et al., 2024) and Gemini (Team et al., 2024; Deepmind, 2025) API. Table 2 shows the details of the VLMs.

D Analysis of input order for VLMs' output

We investigated the influence of position bias (Chen et al., 2024; Wang et al., 2024a) in the best-choice task. Specifically, we examined whether the selection results of models differ depending on whether

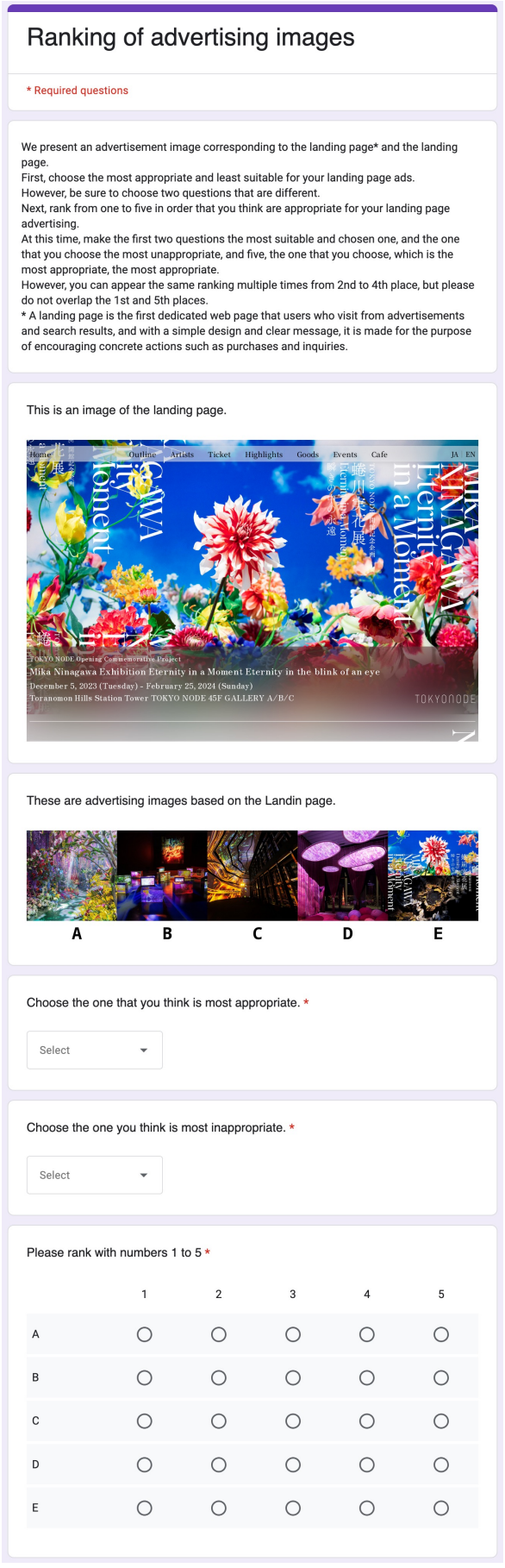

Figure 5: An example of the annotation interface.

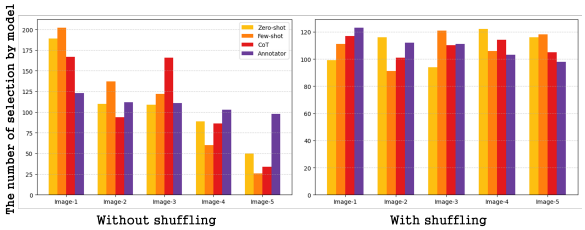

Figure 6: Gemini-1.5-flash's output distribution with five images (w/o shuffling)

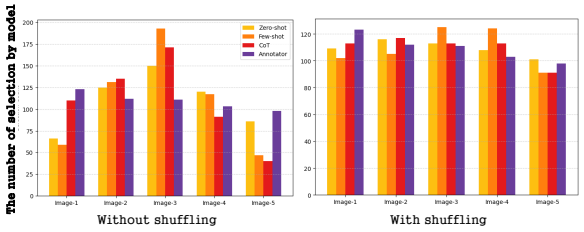

Figure 7: GPT-4o-mini's output distribution with five images (w/o shuffling)

the image order is shuffled, using Gemini-1.5-flash (Team et al., 2024) and GPT-4o-mini (OpenAI et al., 2024). Figure 6 shows the selection distribution of Gemini-1.5-flash under two conditions: with and without image shuffling. Without shuffling, we observed that images in the 4th and 5th positions were consistently less likely to be selected, regardless of the prompt strategy. In contrast, when shuffling was applied, all images, including those close to the annotator's preferred choice, were selected more evenly. Figure 7 shows the selection results of GPT-4o-mini under the same two conditions. Without shuffling, the image in the 3rd position was disproportionately likely to be selected across various prompting strategies. However, with shuffling, all images near the annotator's preferred choice were selected more uniformly. These preliminary findings suggest that shuffling the input image order mitigates the effects of position bias. Therefore, in this study, we adopt a shuffled input strategy when feeding images into the models.

E Details of the Experimental Settings

Details for prompt. We provided the models with three types of prompts *zero-shot*, *one-shot*, and *CoCoT* (Zhang et al., 2024), and fixed the seed values as much as possible. In the *zero-shot* setting, we give only a description of the task to models. In the *one-shot* setting, we provided the same instructions as in the *zero-shot* setting, along with an example of the conversation that reflected

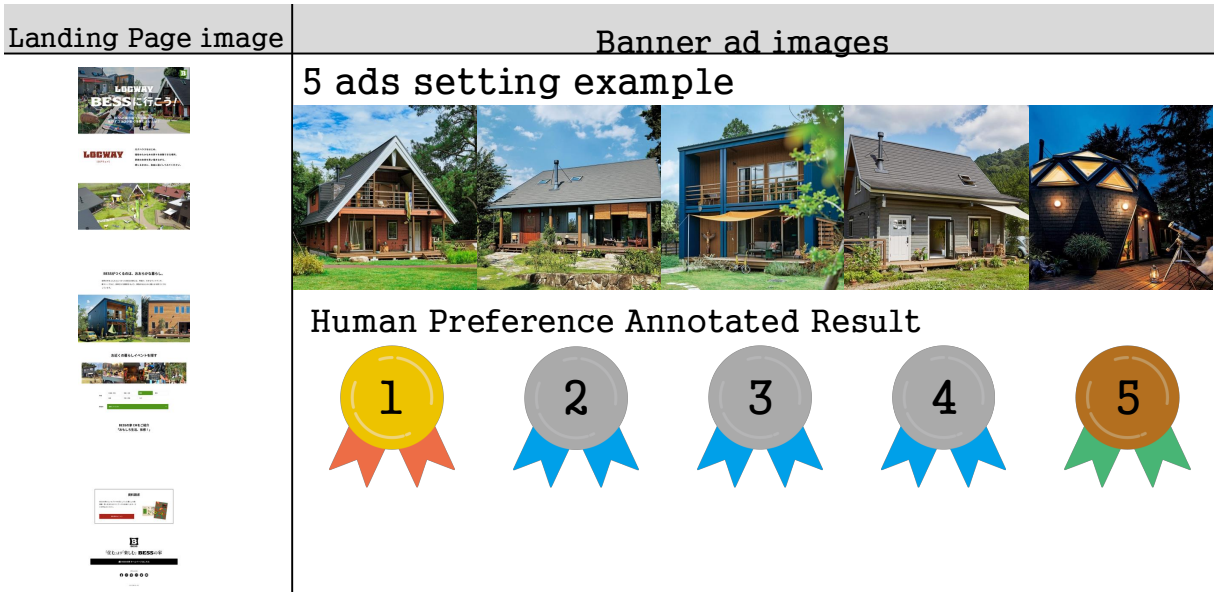

Figure 8: Example images for *one-shot* setting. We provided these LP, banner images, and human evaluation results to the model.

Model	Max Dimension (pixels)
Mantis	1000
Llama-3.2	1000
Qwen2-VL	3000
Qwen2.5-VL	3000
Llava-NeXT	5000
GPT-4o-mini	8000

Table 3: Maximum image dimension limits set for each model.

human evaluation results. Figure 8 shows the example set used in the *one-shot* prompt. This set is selected from the set in which all five annotators unanimously agreed on the most preferred banner.

Details for setting for inference. For GPT-4o-mini, Gemini-1.5-flash, and Gemini-2.0-flash, we set the temperature and top-p to 0 to obtain outputs as deterministic as possible. For Llava-NeXT, Qwen2-VL, Qwen2.5-VL, Mantis, and Llama3.2, we used greedy decoding. Additionally, to ensure consistent output formatting, we employed `outlines` (Willard and Louf, 2023) for the Co-CoT settings of Mantis, LLama-3.2, Qwen2-VL, and Llava-NeXT. To address issues such as out-of-memory errors and API request failures, we did not feed the original image sizes directly into GPT-4o-mini, Llava-NeXT, Qwen2-VL, Qwen2.5-VL, Mantis, and Llama-3.2. Instead, the images were resized before input. The resizing was done while preserving the original aspect ratio, ensuring that neither the height nor the width exceeded the specified pixel limits.

We resized the input images so that neither the width nor the height exceeded the predefined maximum for each model, as summarized in Table 3.

We conducted the experiments on an NVIDIA RTX 6000 Ada, RTX A6000, RTX3090 GPU.

E.1 Prompt template for ranking task

Zero-shot prompt template for ranking task We used the following prompt for the ranking task with *zero-shot* settings.

Ranking task prompt with Zero-shot setting

Task: You will be shown 6 images. Five of these are advertisement images and one is the landing page (LP) image. Do not include the LP image in your ranking.
Please rank the five ad images from most to least preferred, and output your results as a JSON object in the following format.
`JSON Format: {"Image-1": num, "Image-2": num, "Image-3": num, "Image-4": num, "Image-5": num}`

One-shot prompt template for ranking task This is the example for the ranking task with *one-shot* settings. We used the same prompt as *zero-shot*, but we provided an example of a conversation along with images. We displayed example images in Figure 8.

Ranking task prompt with One-shot setting

User: **Task: You will be shown 6 images. Five of these are advertisement images and one is the landing page (LP) image. Do not include the LP image in your ranking.**
Please rank the five ad images from most to least preferred, and output your results as a JSON object in the following format.
`JSON Format: {"Image-1": num, "Image-2": num, "Image-3": num, "Image-4": num, "Image-5": num}` ASSISTANT: “json
`{"Image-1": 1, "Image-2": 2, "Image-3": 3, "Image-4": 4, "Image-5": 5}`
“‘
User: **Task: You will be shown 6 images. Five of these are advertisement images and one is the landing page (LP) image. Do not include the LP image in your ranking.**
Please rank the five ad images from most to least preferred, and output your results as a JSON object in the following format.
`JSON Format: {"Image-1": num, "Image-2": num, "Image-3": num, "Image-4": num, "Image-5": num}`

CoCoT prompt template for ranking task We used the following prompt for the ranking task with *CoCoT* (Zhang et al., 2024) settings.

Ranking task prompt with CoCoT setting

Task:
You will be shown 6 images. Five of these are advertisement images and one is the landing page (LP) image. Do not include the LP image in your ranking.

```
1. Generate Descriptions (Captions)
For each of the five ad images, write a concise caption (1–2 sentences) that:
- Describes its main visual features.
- Highlights what makes it particularly compelling or effective.
2. Rank the Ads
Contrastively compare your five captions, noting why one image stands out as
the most appropriate choice.
Strict Output Requirements:
- First JSON object: Contains exactly five key–value pairs mapping each ad
image label to its caption.
- Second JSON object: Contains exactly one key–value pair mapping "Best" to
the chosen image label.
- Formatting rules:
- Each JSON object must begin with { and end with }.
- The two objects must be separated by a single line break (i.e. \ n). - No
merging of the two objects into one.

  JSON Format: {"Image-1": "Caption for Image-1", "Image-2": "Caption for
Image-2", "Image-3": "Caption for Image-3", "Image-4": "Caption for Image-4",
"Image-5": "Caption for Image-5"}
{"Image-1": num, "Image-2": num, "Image-3": num, "Image-4":num, "Image-5": num}
```

E.2 Prompt template for best-choice task

Zero-shot prompt template for best-choice task We used the following prompt for the best-choice task with *zero-shot* settings.

Best-choice task prompt with Zero-shot setting

Task: You will be shown 6 images.
Five of these are advertisement images and one is the landing page (LP) image. Do not consider the LP image when making your selection.
Please choose the single most appropriate advertisement image and output your choice as a JSON object in the following format.
`JSON Format: {"Best": "Image-(1|2|3|4|5)"}`

One-shot prompt template for best-choice task This is the example for the best-choice task with *one-shot* settings. We used the same prompt as *zero-shot*, but we gave an example of a conversation and images, similar to the ranking task. We displayed example images in Figure 8.

Best-choice task prompt with One-shot setting

User: **Task: You will be shown 6 images.**
Five of these are advertisement images and one is the landing page (LP) image. Do not consider the LP image when making your selection.
Please choose the single most appropriate advertisement image and output your choice as a JSON object in the following format.
`JSON Format: {"Best": "Image-(1|2|3|4|5)"}` ASSISTANT: “‘json
`{"Best": "Image-1"}`
“‘
User: **Task: You will be shown 6 images.**
Five of these are advertisement images and one is the landing page (LP) image. Do not consider the LP image when making your selection.
Please choose the single most appropriate advertisement image and output your choice as a JSON object in the following format.
`JSON Format: {"Best": "Image-(1|2|3|4|5)"}`

CoCoT prompt template for best-choice task We used the following prompt for the ranking task with *CoCoT* settings.

Best-choice task prompt with CoCoT setting

Task:
You will be shown 6 images. Five of these are advertisement images and one is the landing page (LP) image. Do not consider the LP image when making your selection.

```
1. Generate Descriptions (Captions)
For each of the five ad images, write a concise caption (1–2 sentences) that:
- Describes its main visual features.
- Highlights what makes it particularly compelling or effective.
2. Select the Best Ad
Contrastively compare your five captions, noting why one image stands out as
the most appropriate choice.
Strict Output Requirements:
- First JSON object: Contains exactly five key–value pairs mapping each ad
image label to its caption.
- Second JSON object: Contains exactly one key–value pair mapping "Best" to
the chosen image label.
- Formatting rules:
- Each JSON object must begin with {and end with }.
- The two objects must be separated by a single line break (i.e. \ n).
- No merging of the two objects into one.

 JSON Format:
{"Image-1": "Caption for Image-1", "Image-2": "Caption for Image-2", "Image-3":
"Caption for Image-3", "Image-4": "Caption for Image-4", "Image-5": "Caption
for Image-5"}
{"Best": "Image-(1|2|3|4|5)"}
```