Automatic Layout Planning for Visually-Rich Documents with Instruction-Following Models

Wanrong Zhu[¶], **Jennifer Healey**[§], **Ruiyi Zhang**[§], **William Yang Wang**[¶], **Tong Sun**[§]
[¶]UC Santa Barbara, [§]Adobe Research
{wanrongzhu,william}@cs.ucsb.edu, {jehealey,ruizhang,tsun}@adobe.com

Abstract

Recent advancements in instruction-following models have made user interactions with models more user-friendly and efficient, broadening their applicability. In graphic design, non-professional users often struggle to create visually appealing layouts due to limited skills and resources. In this work, we introduce a novel multimodal instruction-following framework for layout planning, allowing users to easily arrange visual elements into tailored layouts by specifying canvas size and design purpose, such as for book covers, posters, brochures, or menus. We developed three layout reasoning tasks to train the model in understanding and executing layout instructions. Experiments on two benchmarks show that our method not only simplifies the design process for non-professionals but also surpasses the performance of few-shot GPT-4V models, with mIoU higher by 12% on Crello (Yamaguchi, 2021). This progress highlights the potential of multimodal instruction-following models to automate and simplify the design process, providing an approachable solution for a wide range of design tasks on visually-rich documents.

1 Introduction

The creation of visually-rich documents (e.g., posters, brochures, book covers, digital advertisements, etc) using available visual components, poses a significant challenge for both professionals and amateurs in the design field. Central to this challenge is the task of arranging these components in an efficient and aesthetically pleasing manner, a process known to be both tedious and time-consuming. Existing toolkits such as Adobe Express[1], Canva[2], and PicsArt[3], usually provide fixed templates to users. These templates, while useful, often fail to fully accommodate the varied and evolving design needs of users, thereby potentially limiting creative expression. Existing research on automatic layout planning (Hsu et al., 2023; Yamaguchi, 2021; Inoue et al., 2023) often requires detailed annotations and poses addition constraints on fixed canvas ratios, thereby diminishing user-friendliness and adaptability.

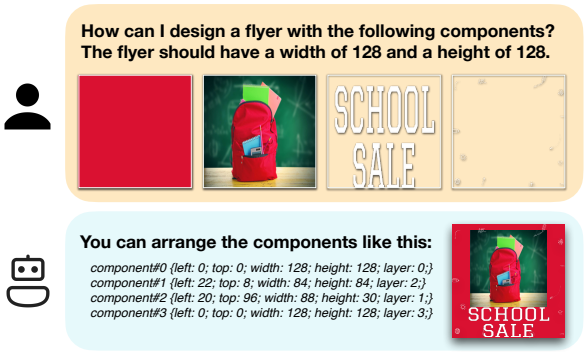

Figure 1: An example of a model conducting automatic layout planning following human-provided instructions and arranging visual contents for design purpose.

Recent advancements in large language models (LLMs) have showcased their remarkable ability to follow human instructions and execute specified tasks (Brown et al., 2020; Ouyang et al., 2022; OpenAI, 2023a), introducing a new level of flexibility and control in human-computer interaction. Alongside these developments, we have witnessed the emergence of instruction-tuned multimodal models (Ye et al., 2023; Li et al., 2023a,b; Awadalla et al., 2023; OpenAI, 2023b), extending the capabilities of LLMs to understand and process information across both textual and visual domains. This progression naturally raises the question of the potential application of instruction-following models in the complex domain of multimodal layout planning. However, employing these models for layout planning presents significant challenges, as the task requires intricate reasoning abilities, including but not limited to, cross-referencing multiple images and performing numerical calculations.

[1] https://www.adobe.com/express/
[2] https://www.canva.com/
[3] https://picsart.com/

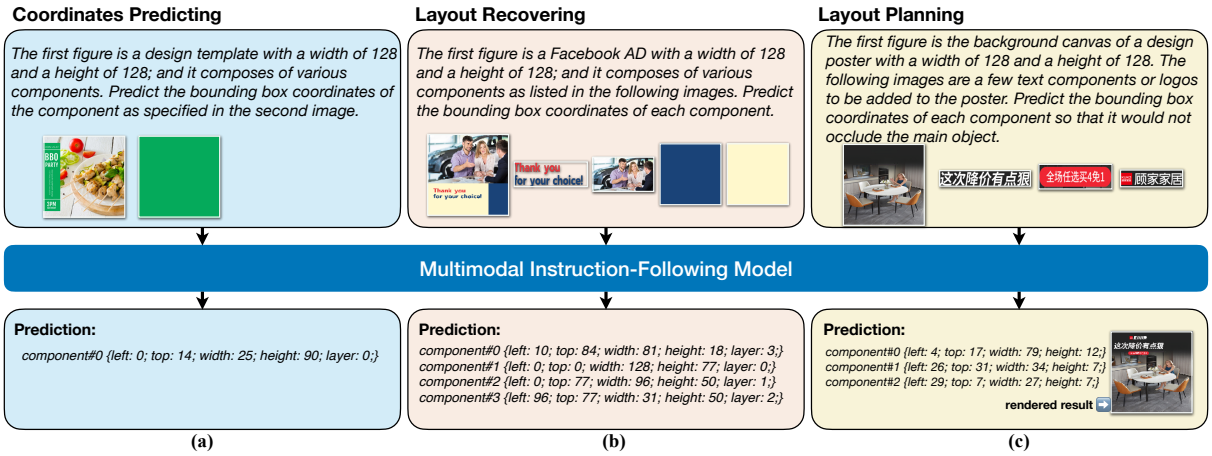

Figure 2: Example inputs and outputs of the three layout reasoning tasks. (a) and (b) are examples from Crello (Yamaguchi, 2021), while (c) is an example from PosterLayout (Hsu et al., 2023).

In this study, we propose DocLap, aiming to address the challenge of visually-rich document layout planning using instruction-following models. To equip these models with the necessary knowledge beyond their primary focus on natural language processing, we have devised three instruction-following tasks focusing on layout reasoning. We evaluated our instruction-tuned DocLap model across two benchmark datasets, and the findings reveal that our approach not only succeeds in this novel application but also outperforms the baseline established by few-shot GPT-4(V). Our main contributions are:

- We propose a novel method for solving the layout planning task using instruction-following models, opening new avenues for research in design automation.
- We develop an instruction dataset featuring three layout reasoning tasks, aiming to enrich the resources available for future research.
- Through experiments on two benchmark datasets, we validate the feasibility of our approach and demonstrate its competitive performance against few-shot GPT-4(V) models.

2 Instruction-Guided Layout Planning for Visually-Rich Documents

Task Definition Visually-rich documents consist of diverse design elements distributed across a canvas. To maintain the integrity of original text designs, text content is converted into images in our setup. The layout planning task involves arranging these design components, provided as a sequence of images $\boldsymbol{i}_1, \boldsymbol{i}_2, ...\boldsymbol{i}_n$, where n represents the component count, onto a canvas for specific application scenarios $\boldsymbol{a}$ (e.g., posters, Instagram posts, book covers) with defined dimensions $\boldsymbol{w}$ (width) and $\boldsymbol{h}$ (height). The canvas may either be blank or have a predefined background.

Instruction-Following Format To offer a more adaptable solution and enhance user experience, we approach this visually-rich layout planning task in an instruction-following manner (Ye et al., 2023; Li et al., 2023a,b; Awadalla et al., 2023; OpenAI, 2023b). The model, in addition to receiving the sequence of design components $\boldsymbol{i}_1, \boldsymbol{i}_2, ...\boldsymbol{i}_n$, will also be given instructions $\mathcal{I}$ detailing the application scenarios $\boldsymbol{a}$ and the canvas size $(\boldsymbol{w}, \boldsymbol{h})$. It is tasked with predicting the layout of each component in a structured format (Feng et al., 2023; Lin et al., 2023). We adopt CSS to encapsulate layout properties including `top`, `left`, `width`, `height`, and another property `layer` that manages the stacking order of potentially overlapping elements.

Instruction-Following Format The task of layout planning encompasses challenges such as following instructions, cross-modal understanding, and numerical reasoning. To equip the model with essential knowledge, we designed three interrelated tasks, as illustrated in Figure 2: (a) *Coordinates Predicting*, where the model predicts the coordinates of a specific component within a given design template; (b) *Layout Recovering*, which involves predicting the coordinates of each component in a template given a sequence of components; and (c) *Layout Planning*, where the model arranges a sequence of components on a canvas by predicting their coordinates. During preprocessing, components smaller than 5% of the canvas size are ex-

		Express	Crello	PosterLayout
Train	Coordinates Predicting	581k	57k	26k
	Layout Recovering	160k	18k	9k
	Layout Planning	160k	18k	9k
Val	Design Layout	-	1493	591

Table 1: Number of examples contained in each training or validation tasks for the datasets used in this study.

cluded, and all templates are resized to ensure the longest edge does not exceed 128. While all three tasks contribute to model training, only the *Layout Planning* task is evaluated during inference.

Model DocLap extends mPLUG-Owl (Ye et al., 2023), a multimodal framework integrating an LLM, a visual encoder, and a visual abstractor module. Specifically, it employs Llama-7b v1 (Touvron et al., 2023) as the LLM and CLIP ViT-L/14 (Radford et al., 2021) as the visual encoder. The visual abstractor module converts CLIP's visual features into 64 tokens that match the dimensionality of text embeddings, allowing for the simultaneous processing of multiple visual inputs. We extended the Llama v1 vocabulary with numerical tokens ranging from 0 to 128. The embeddings of the extended tokens are randomly initialized, and then tuned in further instruction tuning.

3 Experimental Setup

Datasets We conduct experiments on layout planning for visually-rich documents with the following two benchmarks: (1) Crello (Yamaguchi, 2021) is built upon design templates collected from online service. This task begins with an empty canvas, challenging the model to organize the layouts of the provided visual components. (2) PosterLayout (Hsu et al., 2023) starts from non-empty canvas (background image for posters), and requires the model to strategically place text, labels, and logos. Our training data is supplemented with design templates from Adobe Express. Detailed dataset statistics are available in Table 1. To ensure fair comparison, validation examples are limited to no more than 4 images, aligning with the input constraints of GPT-4V at the time of our submission. Illustrative examples from both datasets are presented in Figure 2.

Baselines For Crello, we compare with CanvasVAE (Yamaguchi, 2021) and FlexDM (Inoue et al., 2023). For PosterLayout, we compare with DS-GAN (Hsu et al., 2023). Additionally, we include

	Model	mIoU	Left	Top	Width	Height
#1	CanvasVAE	42.39	29.31	30.97	27.58	29.99
#2	FlexDM	50.08	34.98	34.03	30.04	33.08
#3	GPT-4 0-shot	30.75	24.36	24.07	13.63	15.11
#4	GPT-4 1-shot	29.97	26.09	23.71	13.94	13.33
#5	GPT-4V 0-shot	28.81	19.96	18.09	10.45	10.08
#6	GPT-4V 1-shot	35.17	22.77	20.90	13.16	14.11
#7	DocLap (Ours)	43.75	33.46	35.61	19.18	22.79

Table 2: Automatic evaluation results on Crello showing mIoU and the accuracy for left, top, width and height.

	Model	Occ.↓	Uti.↑	Rea.↓
#1	DS-GAN	21.57	23.92	20.16
#2	GPT-4 0-shot	50.61	43.09	25.87
#3	GPT-4 1-shot	47.92	38.00	25.34
#4	GPT-4V 0-shot	36.67	33.26	24.39
#5	GPT-4V 1-shot	36.39	20.24	26.03
#6	DocLap (Ours)	23.01	22.46	21.00

Table 3: Evaluation results on PosterLayout. *Occ.*: occlusion rate; *Uti.*: utility rate; *Rea.*: unreadability.

comparative evaluations with text-only versions of GPT-4 and GPT-4V (OpenAI, 2023a,b,c; gpt, 2023) across both tasks. For the text-only GPT-4 evaluations, visual components are not directly supplied. Instead, we employ BLIP-2 (Li et al., 2023c) to generate textual descriptions of each component.

Metrics For Crello evaluation, we measure mean Intersection-over-Union (mIoU) between predicted and actual bounding boxes, along with accuracy in width, height, left, and top dimensions following FlexDM (Inoue et al., 2023). Accuracy is quantified by assigning a score of 1 if the predicted value falls into the same 64-bin quantized range as the ground truth; otherwise, it scores 0. In assessing PosterLayout, we follow DS-GAN (Hsu et al., 2023) and employ content-aware metrics, including (1) occlusion rate↓, indicating the percentage of primary objects obscured by design elements; (2) utility rate↑, reflecting the extent to which design components cover non-primary object areas; and (3) unreadability↓, measuring the uniformity of areas where text-containing elements are placed.

4 Results & Analysis

Quantitative Results Table 2 shows the automatic evaluation results on Crello dataset. The first two lines are results from models that are trained with supervised learning. Line #3–#6 show few-shot GPT-4(V) results, in which we notice that GPT-4V surpasses text-only GPT-4, and that providing demonstrative examples leads to better results com-

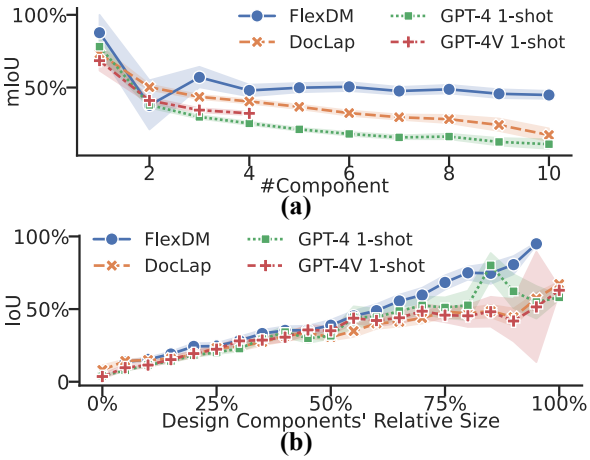

Figure 3: (a) mIoU variation with the number of visual components in design templates. (b) IoU correlation with the relative size of a single visual component. Both plots pertain to Crello.

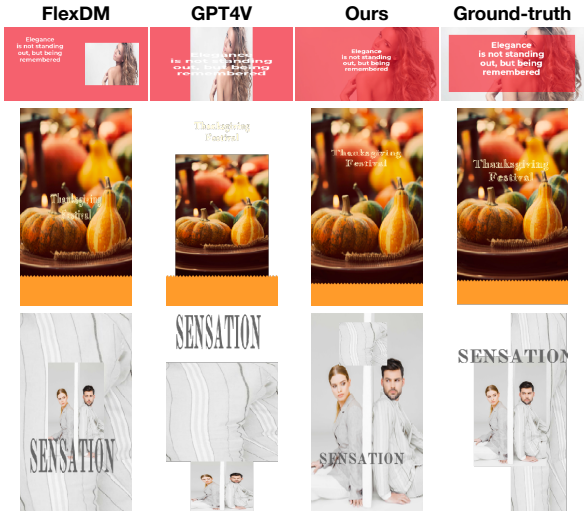

Figure 4: Qualitative comparisons for layout planning results on Crello. GPT-4V w/ 1-shot learning.

Figure 5: Qualitative comparisons for layout planning results on PosterLayout. GPT-4V w/ 1-shot learning.

pared to zero-shot prompting. Our DocLap's performance (#7) surpass the few-shot GPT-4(V) on both mIoU and aspect accuracies, but still falls behind a bit compared to FlexDM (#2).

Table 3 presents the PosterLayout evaluation results, which reveals a trade-off between occlusion rate and utility rate across models. GPT-4(V) models (#2–#5) exhibit high occlusion and utility rates, indicating a propensity for predicting larger bounding boxes. Our DocLap shows a reduced occlusion rate, accompanied by a decrease in utility rate. Regarding unreadability, DocLap outperforms GPT-4(V), though DS-GAN (#1) achieves the highest performance, underscoring the efficacy of supervised models in this context.

Effects of #Component Figure 3(a) reveals that all listed models exhibit high mIoU for templates with a single component. FlexDM's mIoU shows slight fluctuations, stabilizing around 50%. In contrast, mIoU for DocLap and GPT-4(V) decreases as the number of components increases, indicating that more complex scenarios involving more visual components might pose challenges to current instruction-following models.

Effects of Component Size Figure 3(b) demonstrates a linear correlation between the relative size of a single visual component and the IoU of the model prediction with the ground truth for all models assessed. This suggests that smaller visual components pose a greater challenge for precise placement in accordance with the ground truth during layout planning. Typically, these small components, such as logos, small text boxes, or decorative elements, have a degree of positional flexibility, allowing for multiple valid placements.

Demonstrative Examples Figure 4 shows examples from Crello while Figure 5 shows examples from PosterLayout.

5 Conclusion

This study demonstrates the potential of instruction-following models in addressing the intricate task of layout planning for visually rich documents. The positive outcomes observed from our experiments on two distinct benchmarks affirm the viability and effectiveness of our methodology. This research paves the way for future explorations into the application of instruction-following models across various domains, highlighting their potential to revolutionize tasks that require a nuanced understanding of both language and visual elements.

Limitations

This study, while pioneering in its approach to simplifying the graphic design process through instruction-following models, acknowledges several limitations. First, the performance of our model, DocLap, and GPT-4(V) diminishes as the complexity of the layout increases, particularly with the addition of more visual components. This suggests a need for improved model robustness and adaptability in handling more intricate design scenarios. Additionally, the evaluation metrics, such as mIoU and the binary accuracy measurement for layout attributes, may not fully capture the nuances of aesthetic and functional design quality. The reliance on these metrics might overlook the subjective and context-specific nature of effective design, indicating a potential area for developing more comprehensive evaluation frameworks.

Ethics Statement

Our work on instruction-following models for layout planning, while innovative, introduces potential risks including over-reliance on automation, which may impede the development of design skills and creativity. Importantly, our model does not generate new visual content; all predictions are based on existing components provided by users. The outputs are solely layouts in text formats, mitigating risks related to copyright infringement and originality. However, the reliance on automated tools could lead to a homogenization of design aesthetics and potentially amplify biases present in the input data. Addressing these challenges requires careful consideration of the ethical implications of automated design tools and the promotion of responsible usage to complement human creativity. Noted here that we utilize ChatGPT to polish the writing and ensure clarity and conciseness in the presentation of our research, without altering the fundamental nature of the work or its implications.

References

2023. Chatgpt can now see, hear, and speak. https://openai.com/blog/chatgpt-can-now-see-hear-and-speak.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Yitzhak Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. 2023. Openflamingo: An open-source framework for training large autoregressive vision-language models. *ArXiv*, abs/2308.01390.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Weixi Feng, Wanrong Zhu, Tsu-Jui Fu, Varun Jampani, Arjun Reddy Akula, Xuehai He, S Basu, Xin Eric Wang, and William Yang Wang. 2023. LayoutGPT: Compositional visual planning and generation with large language models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Hsiao-An Hsu, Xiangteng He, Yuxin Peng, Hao-Song Kong, and Qing Zhang. 2023. Posterlayout: A new benchmark and approach for content-aware visual-textual presentation layout. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6018–6026.

Naoto Inoue, Kotaro Kikuchi, Edgar Simo-Serra, Mayu Otani, and Kota Yamaguchi. 2023. Towards flexible multi-modal document models. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14287–14296.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, C. Li, and Ziwei Liu. 2023a. Mimic-it: Multi-modal in-context instruction tuning. *arXiv preprint arXiv:2306.05425*.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Jingkang Yang, and Ziwei Liu. 2023b. Otter: A multi-modal model with in-context instruction tuning. *arXiv preprint arXiv:2305.03726*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023c. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*.

Jiawei Lin, Jiaqi Guo, Shizhao Sun, Zijiang James Yang, Jian-Guang Lou, and Dongmei Zhang. 2023. Layoutprompter: Awaken the design ability of large language models. In *Thirty-seventh Conference on Neural Information Processing Systems*.

OpenAI. 2023a. Gpt-4 technical report.

OpenAI. 2023b. Gpt-4v(ision) system card.

OpenAI. 2023c. Gpt-4v(ision) technical work and authors. https://cdn.openai.com/contributions/gpt-4v.pdf.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971.

Kota Yamaguchi. 2021. Canvasvae: Learning to generate vector graphic documents. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5461–5469.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yi Zhou, Junyan Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qiang Qi, Ji Zhang, and Feiyan Huang. 2023. mplug-owl: Modularization empowers large language models with multimodality. *ArXiv*, abs/2304.14178.