VISUAL PROMPTING METHODS FOR GPT-4V BASED ZERO-SHOT GRAPHIC LAYOUT DESIGN GENERATION

Kunal Singh, Mukund Khanna, Ankan Biswas, Pradeep Moturi & Shivam
Fractal Analytics, India

ABSTRACT

Graphic layout design generation is a challenging problem in computer vision. The key aspect of the challenge is ensuring coherent placement of textual elements on the background image to ensure aesthetic appeal and avoiding occlusion of key visual elements. Although prior methods have made attempts to solve this multi-modal problem, they couldn't perfect it. Owing to the complexity required in understanding the relationship between visual and text elements in the aforementioned task, we investigate GPT-4-Vision(GPT-4V), a large multimodal models(LMMs), to do zero-shot graphic layout design generation in a versatile manner. Our approach explores various off-the-shelf segmentation/superpixel methods to identify and mark the key regions to visually augment the image to enhance GPT-4V's spatial reasoning capability . The results of our comprehensive experiments on a self-curated dataset demonstrates the efficacy of our proposed visual prompting methods, showing improvement over standard GPT-4V prompting method and also performing at par and even better, for some techniques, than state-of-the-art specialist model.The code and data is available at this link.

1 INTRODUCTION AND LITERATURE REVIEW

Graphic layout design involves the strategic arrangement of textual and visual elements to convey a message and is a crucial aspect of visual communication and marketing. This process requires a thoughtful understanding of the structure of each element and its relationship among the elements. Automating this task has immense real-world applications and remains a critical problem in vision. Efforts have been made using generative models but they do not take into account the text content and do unconditional generation Cao et al. (2022), Cheng et al. (2023), Guo et al. (2021), Gupta et al. (2021), Zhou et al. (2022), Zheng et al. (2019). Researchers have also tried by formulating this as an object detection task where they predict a bounding box for an input text string on an input image by using text-guided object detection conditioned on multi-modal inputs Yu et al. (2022) but it still struggles with inaccurate placements leading to lower aesthetic performance and occlusion issues. Unlike the supervised approaches, this paper explores using a generalist LMM to do zero-shot layout designing due to their ability to reason and analyse diverse modalities simultaneously.

Recent works have demonstrated GPT-4V's Yang et al. (2023b) potential across various vision tasks like medical image VQA Buckley et al. (2023) Yan et al. (2023), object detection Cao et al. (2023), segmentation Yang et al. (2023a), anomaly detection Zhang et al. (2023) and autonomous driving Wen et al. (2023), and have demonstrated its powerful generation capabilities. Some research has also focused on improving the visual understanding of GPT-4V by using segmentation algorithms to mark different sub-regions for grounding tasks Yang et al. (2023a). Understanding the accurate placement and alignment of the foreground textual content on the background image requires global and local comprehension of the overall image and sub-sections of the image respectively. In this paper we use GPT-4V to do zero-shot layout generation for text element, and explore various visual prompting methods to improve the generation quality by improving the understanding of GPT-4V of the image. Specifically, we propose our grid and super-pixel clustering Lowekamp et al. (2018) based labelling methods which show better performance than previously known visual prompting methods. We also compare with LayoutDetr which is a DeTr Carion et al. (2020) based model and show that our visual prompting approach helps GPT-4V surpass its performance.

Table 1: Best IoU scores across 3 runs & IoU for visual prompting methods & LayoutDetr

Method	IoU	Best of 3 IoU
GPT-4V	0.068	0.088
GPT-4V + SoM	0.072	0.090
GPT-4V + Grid	0.125	**0.192**
GPT-4V + SLIC	0.107	**0.186**
GPT-4V + Watershed	0.113	0.151
GPT-4V + Reference Image	0.110	0.137
LayoutDetr	0.165	0.165

2 Dataset, Method, Experiments and Results

Dataset: To evaluate our approaches, we compiled a dataset of 50 images from Canva.com, featuring advertisements, posters, and cards. Each image features multiple text elements; we identified and labeled the largest text as the primary element for ground truth. During inference, we used images stripped of all text to assess the performance in text placement of the primary text's content. More information on why and how we curated a new dataset can be found in the Appendix section.

Method: GPT-4V can understand the global semantic relationship between the textual and visual elements of a graphic but is unable to visually link the text with a sub-section in the background image. Therefore, we focus on vision prompting approaches where we partition an image into semantically meaningful segments to improve spatial understanding of the visual elements for GPT-4V. We believe that this visual grounding will allow GPT-4V to better localize the placement of the text content on the background image. To this end we try the following approaches- *Set-of-Marks*: All images were processed using SoM method before beeing fed to GPT-4V. SoM is a visual prompting method which partitions an image into regions at different levels of granularity using models like SAM Kirillov et al. (2023), and overlays these regions with a set of marks. *Grid segmentation*: All images are divided into a 3x3 grid. *Superpixel segmentation*: All images are segmented into 9 segments using various superpixel algorithms like watershed, and SLIC. We numerically mark the centre of each region generated using the methods discussed: SoM, grid, SLIC and watershed.*Reference Image*: We try a few-shot approach where we introduce a randomly sampled reference image, an example of perfectly placed text on an image, to the model to improve its performance. Refer 1 for pipeline.

Experiments: For all experiments we resize the images to 768x768 to maintain consistency. We explore various region labelling methods on the images which serve as the input for GPT-4V. We prompt GPT-4V to provide the bounding box coordinates for the position it deems to be the best for the placement of the text within the image, judging by the positioning of other elements present in the image. We also prompted it to provide the ideal font size of the text and choose an appropriate font colour between black and white, to maintain readability when placed on the image background.

Results: We run each experiment 3 times to account for the variability that rises from using a largely non-deterministic model like GPT-4V. For each experiment we calculate the IoU using 2 different approaches, first where we use the average over the entire dataset and second where we first select the best score per image across the three trials before averaging and we present these results in Table 1. We also compare our approach with LayoutDeTr which is trained in a supervised way to do graphic layout design. Our results show that all our approaches beat base GPT-4V and set-of-marks prompting method. Our grid and SLIC based approach outperform LayoutDetr (on Best of 3 IoU).

3 Conclusion

We test the capabilities of GPT-4V on zero-shot graphic layout designing and augment its capabilities with various visual prompting approaches. Our empirical findings substantiate that providing supplementary visual cues to GPT-4V enhances the model's performance relative to its baseline configuration. We also compare our approach with other supervised detection methods and show that our grid based and SLIC based visual prompting method matches and in some iterations even surpasses LayoutDetr.

URM Statement

The authors acknowledge that at least one key author of this work meets the URM criteria of ICLR 2024 Tiny Papers Track

References

Thomas Buckley, James A Diao, Adam Rodman, and Arjun K Manrai. Accuracy of a vision-language model on challenging medical cases. *arXiv preprint arXiv:2311.05591*, 2023.

Yunkang Cao, Xiaohao Xu, Chen Sun, Xiaonan Huang, and Weiming Shen. Towards generic anomaly detection and understanding: Large-scale visual-linguistic model (gpt-4v) takes the lead. *arXiv preprint arXiv:2311.02782*, 2023.

Yunning Cao, Ye Ma, Min Zhou, Chuanbin Liu, Hongtao Xie, Tiezheng Ge, and Yuning Jiang. Geometry aligned variational transformer for image-conditioned layout generation. In *Proceedings of the 30th ACM International Conference on Multimedia*, pp. 1561–1571, 2022.

Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pp. 213–229. Springer, 2020.

Chin-Yi Cheng, Forrest Huang, Gang Li, and Yang Li. Play: Parametrically conditioned layout generation using latent diffusion. *arXiv preprint arXiv:2301.11529*, 2023.

Shunan Guo, Zhuochen Jin, Fuling Sun, Jingwen Li, Zhaorui Li, Yang Shi, and Nan Cao. Vinci: An intelligent graphic design system for generating advertising posters. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 2021. URL `https://api.semanticscholar.org/CorpusID:233987851`.

Kamal Gupta, Justin Lazarow, Alessandro Achille, Larry S Davis, Vijay Mahadevan, and Abhinav Shrivastava. Layouttransformer: Layout generation and completion with self-attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1004–1014, 2021.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023.

Bradley C Lowekamp, David T Chen, Ziv Yaniv, and Terry S Yoo. Scalable simple linear iterative clustering (sslic) using a generic and parallel approach. *arXiv preprint arXiv:1806.08741*, 2018.

Licheng Wen, Xuemeng Yang, Daocheng Fu, Xiaofeng Wang, Pinlong Cai, Xin Li, Tao Ma, Yingxuan Li, Linran Xu, Dengke Shang, et al. On the road with gpt-4v (ision): Early explorations of visual-language model on autonomous driving. *arXiv preprint arXiv:2311.05332*, 2023.

Zhiling Yan, Kai Zhang, Rong Zhou, Lifang He, Xiang Li, and Lichao Sun. Multimodal chatgpt for medical applications: an experimental study of gpt-4v. *arXiv preprint arXiv:2310.19061*, 2023.

Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. Set-of-mark prompting unleashes extraordinary visual grounding in gpt-4v. *arXiv preprint arXiv:2310.11441*, 2023a.

Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9:1, 2023b.

Ning Yu, Chia-Chih Chen, Zeyuan Chen, Rui Meng, Gang Wu, Paul Josel, Juan Carlos Niebles, Caiming Xiong, and Ran Xu. Layoutdetr: Detection transformer is a good multimodal layout designer. *arXiv preprint arXiv:2212.09877*, 2022.

Jiangning Zhang, Xuhai Chen, Zhucun Xue, Yabiao Wang, Chengjie Wang, and Yong Liu. Exploring grounding potential of vqa-oriented gpt-4v for zero-shot anomaly detection. *arXiv preprint arXiv:2311.02612*, 2023.

Xinru Zheng, Xiaotian Qiao, Ying Cao, and Rynson W. H. Lau. Content-aware generative modeling of graphic design layouts. *ACM Transactions on Graphics (TOG)*, 38:1 – 15, 2019. URL `https://api.semanticscholar.org/CorpusID:196834740`.

Min Zhou, Chenchen Xu, Ye Ma, Tiezheng Ge, Yuning Jiang, and Weiwei Xu. Composition-aware graphic layout gan for visual-textual presentation designs. *arXiv preprint arXiv:2205.00303*, 2022.

A Appendix

A.1 Dataset

A.1.1 Why we made our own Dataset

We made a conscious decision to avoid using the ad banner dataset that LayoutDetr introduced and is trained on. Mainly due to the concern that they used diffusion-based image inpainting process to remove the layout elements to prepare the input image for the experiments. Also, the images in the dataset are not stored in file format that would allow manual editing of text elements on a graphic design tool. The inpainting process used to remove the text elements, ended up introducing artifacts (blurred sections) as seen in Figure 2 and potentially biased the evaluation/comparison. To test out the biases in the LayoutDetr dataset we ran zero-shot evaluation on base GPT-4V and using two of our segmentation approaches – SLIC and Grid with the task of detecting primary text, on a randomly sampled set of 500 images selected from the layoutDetr test set. For each sampled image in the test set we selected the primary text based on whichever box had the highest bounding box area to text length ratio. We found exceptionally low IoU values as reported in Table 2. This proves the issue that arises from artifacts in the image as they appear to be visual elements on the images. So, laying a textual element on top of it would be essentially blocking the artifact and will be a bad place to put the text and hence GPT-4V rightly avoids it.

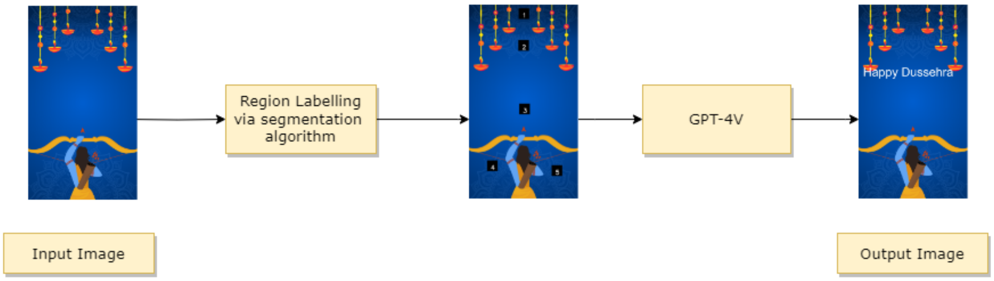

Figure 1: Overview of our proposed pipeline for zero-shot layout generation. For our reference image based prompting approach, which is one shot, we pass an additional reference image along with the input image to GPT-4V.

A.1.2 Dataset creation and Labelling

Hence, to ensure fair and accurate evaluation we aimed to manually create a high-quality dataset and compile a diverse set of graphics and to this end included advertisements, greeting posters, and invitation cards, to cover a broad range of graphic layout design scenarios. To avoid inpainting, we resorted to using graphics/templates present on graphic design tools/platforms (this overcomes the limitation of the existing open-source dataset faced - as textual elements can be edited out in graphics/templates on graphic design platforms like Canva). Canva.com is a widely popular graphic design tool and contains sample/pre-made top rated graphics of previously mentioned categories. We used Canva.com to identify and collect the templates and then perform editing to generate input images (without any editable textual elements) and ground truth images (with only primary/main

Table 2: Performance of base GPT-4V and our segmentation approach on LayoutDeTr dataset

Method	Average IoU	Best of 3 IoU
GPT-4V	0.0023	0.004
GPT-4V + Grid	0.0044	0.0034
GPT-4V + SLIC	0.0044	0.0066

Table 3: Best and average IoU across 3 runs, for proposed visual prompting methods and Layout-Detr on alternate dataset

Method	Average IoU	Best of 3 IoU
GPT-4V	0.063	0.092
GPT-4V + SoM	0.060	0.089
GPT-4V + Grid	0.101	0.142
GPT-4V + SLIC	0.119	0.172
GPT-4V + Watershed	0.065	0.084
GPT-4V + Reference Image	0.105	0.135
LayoutDetr	0.161	0.161

text element). This was followed by annotation of the ground truth images to get the bounding box coordinates for the primary text. The distribution of our dataset was as follows - 40% greeting posters, 24% invitation cards and 36% advertisements, with size totalling to 50. We have shared the dataset as well along with the code. We show sample graphics from all the 3 categories in Figure 3

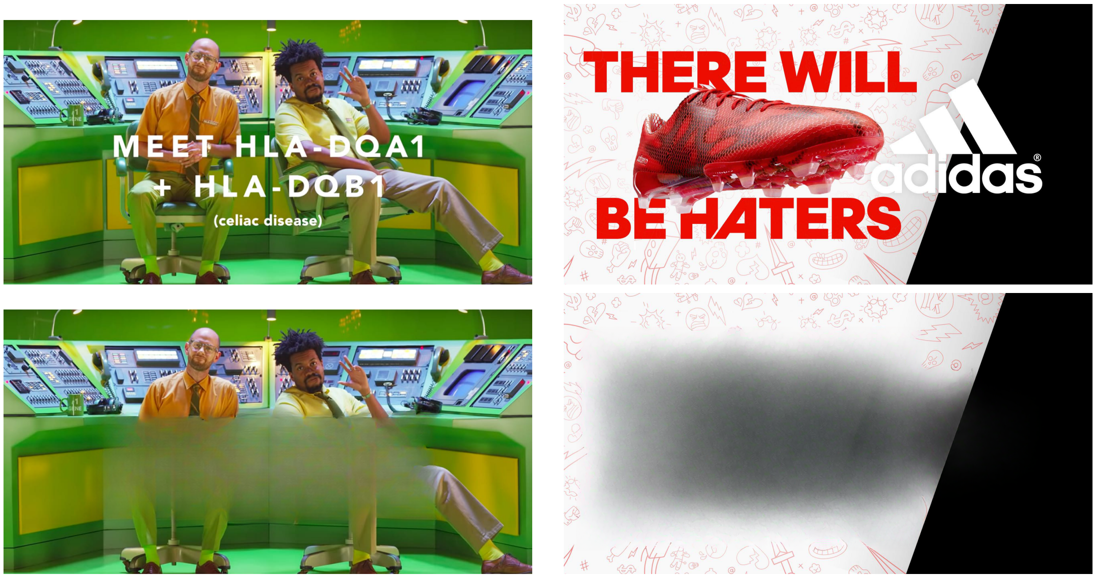

Figure 2: First row left to right -ground truth image from LayoutDeTr dataset. Second row left to right - 1x Inpainted images from LayoutDeTr dataset.

A.2 Ablation Studies

In this section we provide some ablation studies that we ran. Our hypothesis is that GPT-4V is able to understand the semantic relationship between the elements, but it struggles with finding the appropriate coordinates in the image. To test our hypothesis, we try our proposed methods on an alternate version of our ground-truth images, in which we only remove the text from the graphics that we want to detect and keep the rest of the text intact in the image. For this experiment we present our results in Table 3 and as we can observe there is not much difference proving that our hypothesis was right. GPT-4V can identify the correct place for the text, but it struggles with this task as it is not able to identify the appropriate bounding box coordinates.

A.3 PROMPT

We are sharing the prompt that we designed to get the best response from GPT-4V. We tried multiple iterations of this prompt and found that providing it the image size and explicitly telling it how to position the text improved its performance. We use the same prompt for every approach except for the reference image experiment in which we add a one line description about the reference image -

*You are an expert design consultant with a creative mind. You are provided with an image of dimensions 768*768 which is to be used for an advertisement banner. Your task is to suggest the best place within the image to place the text of the advertisement 'text'.*

You must keep in mind the following things: 1. Position Selection: The given text is the most important text or the header of an advertisement. Therefore, it has to positioned strictly in a plain area and must not overlap any other objects present in the image. 2. Font Size Selection: The text must have a font size which is large enough to garner attention, yet at the same time not overlap any of the objects in the image. So, estimate the font size for the header accordingly. 2. Font Color Selection: Identify and select the ideal font color as black or white according to the background. The font should be clearly visible and in contrast with the background.

You need to provide the ideal bounding box coordinates for the text box. Calculate the area text would take and then give coordinates, as text should not go out of the image. The width of the box would be x2-x1, so specify the coordinates according to the width it would need. Do remember that You are capable enough and have all the relevant information present. Process through the image and provide the results.

DO remember that you have all the information needed for getting the results.

Your response must be in the following format: Line1: 'Font Size = p' where p is the font size suggestion. Line2: 'Coordinates = [x1,y1] where x1, y1 are the top left coordinates of the bounding box. Line3: 'Font Color = k' where k = 0,0,0 if font color is black and k = 256,256,256 if font color is white. Line4: 'Bottom = [p1,q1]' where p1, q1 are the bottom right coordinates of the bounding box. Just print these 4 lines and nothing else. There should be no space after any comma while printing coordinates. Strictly print it in the above format.

Take a deep breath. Think step by step and answer

A.4 VISUAL RESULTS

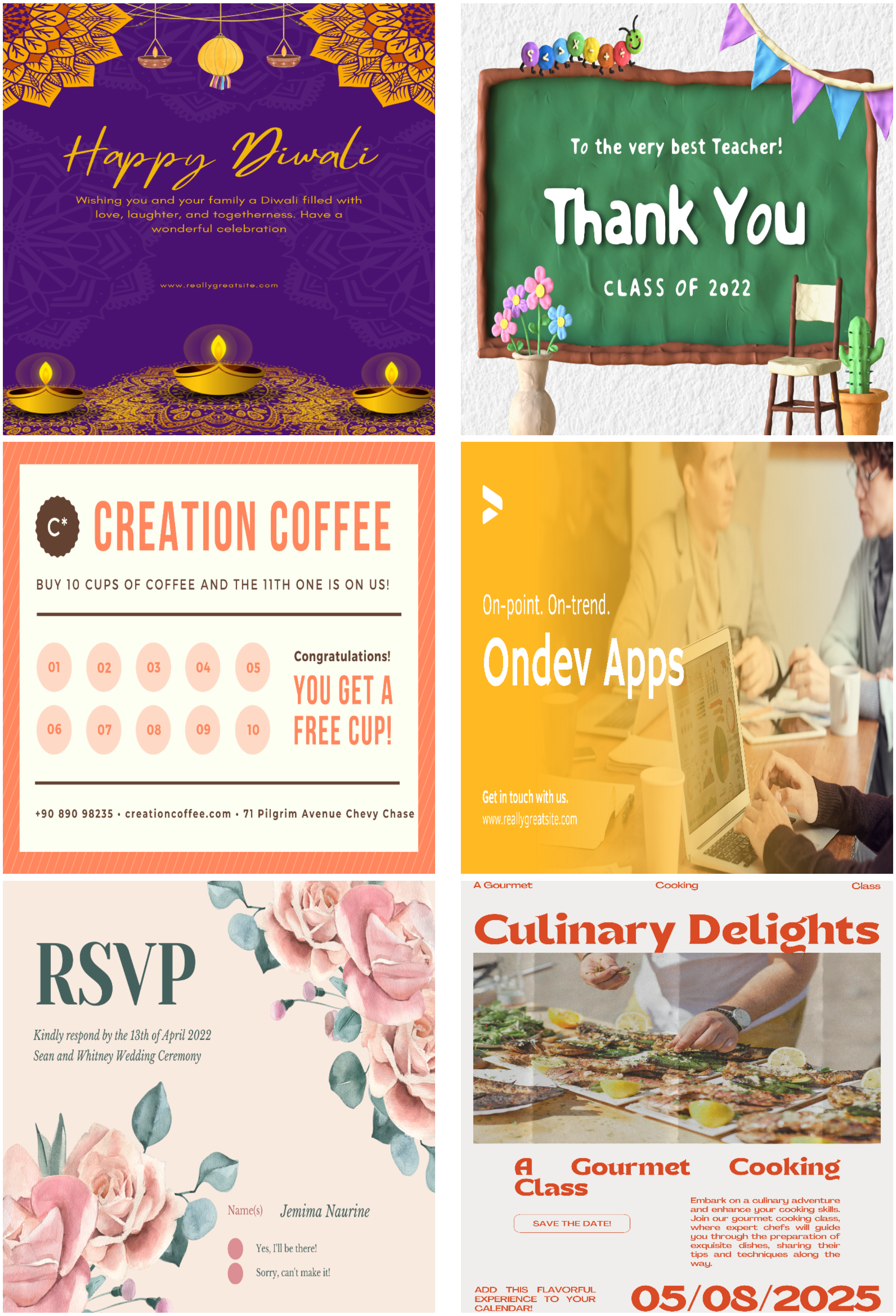
Happy Diwali
Wishing you and your family a Diwali filled with love, laughter, and togetherness. Have a wonderful celebration
www.reallygreatsite.com
To the very best Teacher!
Thank You
CLASS OF 2022
C*
CREATION COFFEE
BUY 10 CUPS OF COFFEE AND THE 11TH ONE IS ON US!
01 02 03 04 05
06 07 08 09 10
Congratulations!
YOU GET A FREE CUP!
+90 890 98235 • creationcoffee.com • 71 Pilgrim Avenue Chevy Chase
On-point. On-trend.
Ondev Apps
Get in touch with us.
www.reallygreatsite.com
RSVP
Kindly respond by the 13th of April 2022
Sean and Whitney Wedding Ceremony
Name(s) Jemima Naurine
Yes, I'll be there!
Sorry, can't make it!
A Gourmet Cooking Class
Culinary Delights
A Gourmet Cooking Class
SAVE THE DATE!
Embark on a culinary adventure and enhance your cooking skills. Join our gourmet cooking class, where expert chefs will guide you through the preparation of exquisite dishes, sharing their tips and techniques along the way.
ADD THIS FLAVORFUL EXPERIENCE TO YOUR CALENDAR!
05/08/2025

Figure 3: Few samples of images for some of the categories present in the dataset. Row-wise from top to bottom we have Greeting cards, Advertisements and Invitations respectively.

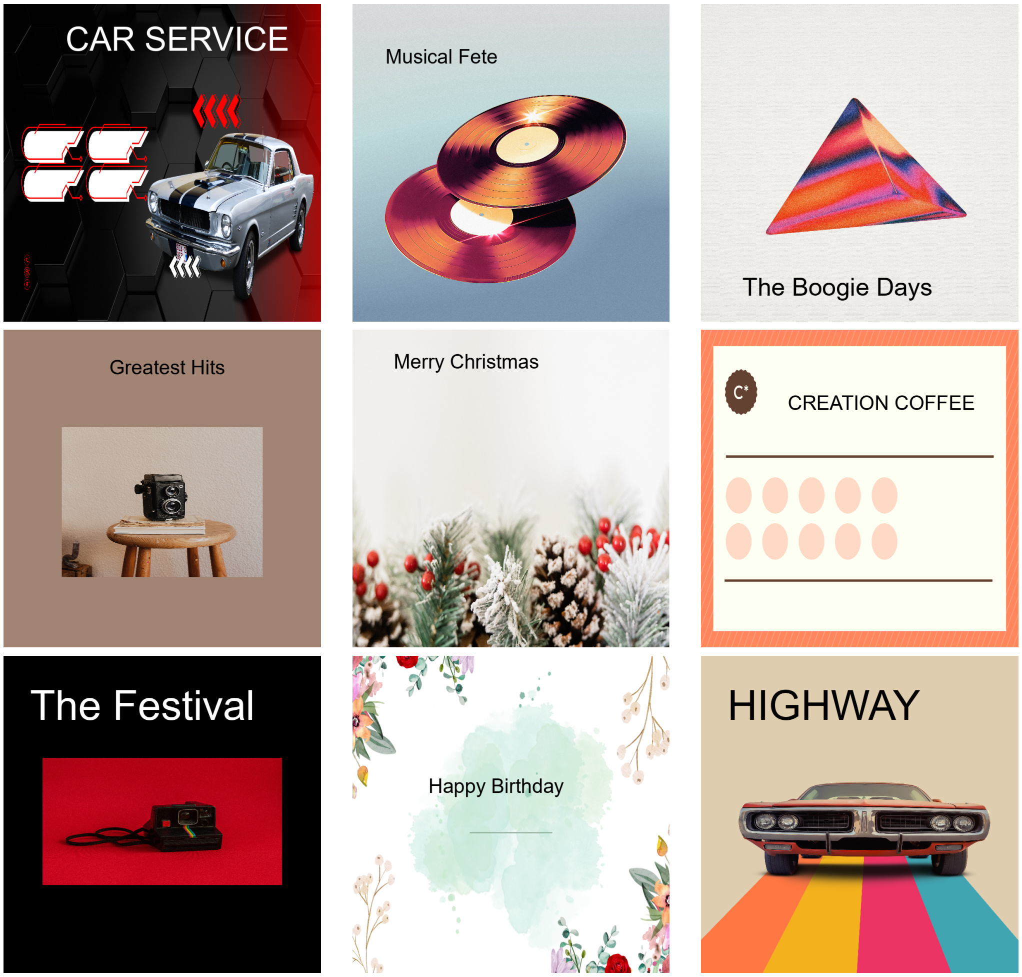

Figure 4: Sample results from GPT-4V with SLIC segmentation approach.

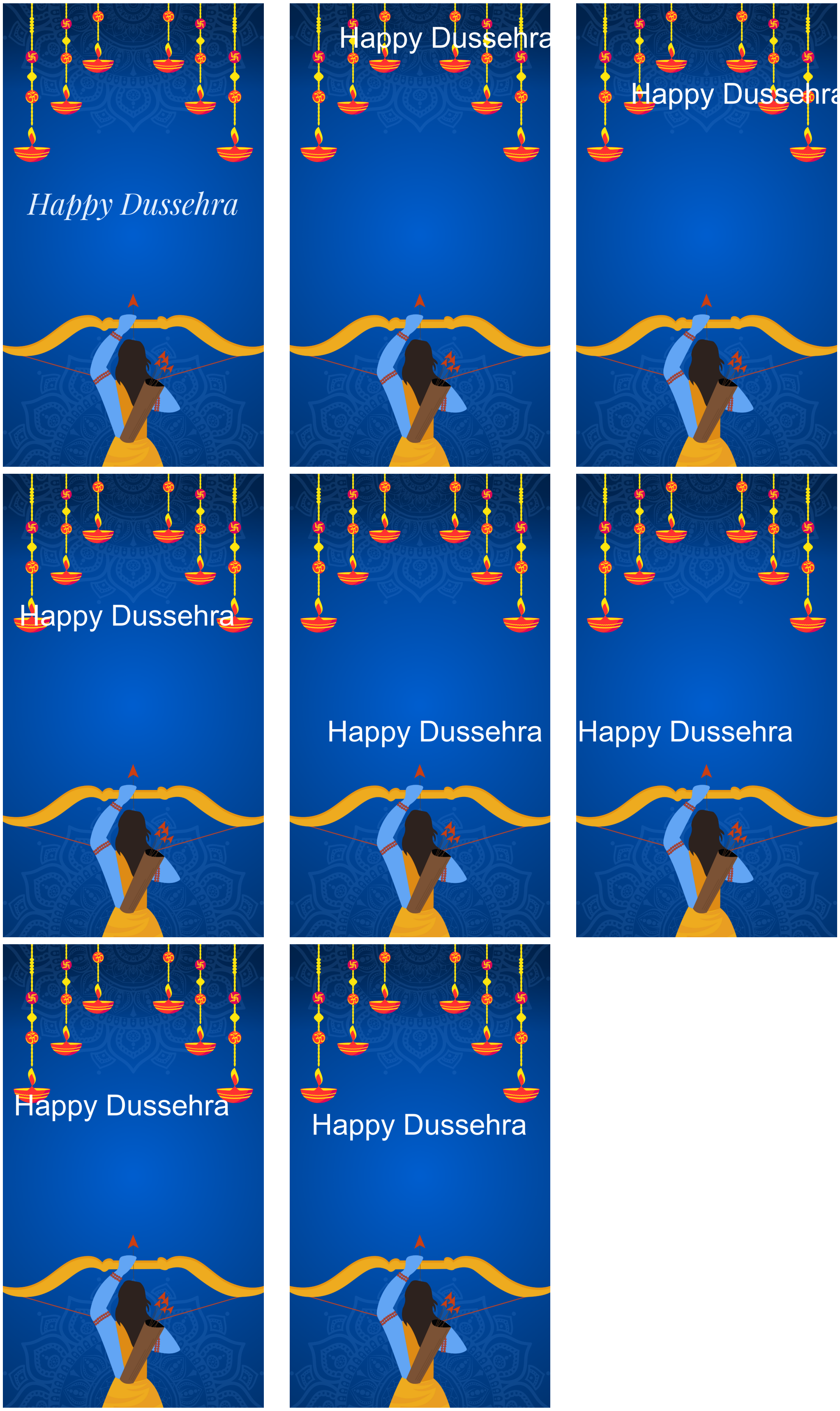

Figure 5: First row: left to right -ground truth image, output from GPT-4V, output from GPT-4V +SoM respectively. Second row: left to right - output from GPT-4V +SLIC, output from GPT-4V +watershed, output from GPT-4V +reference image respectively. Last row: left to right - output from LayoutDeTr, output from GPT-4V +Grid respectively.