NO MORE BLAH-BLAH: EMBRACING REAL TEXT IN THE IMAGE SYNTHESIS WORLD

Aref Tabatabaei, Zahra Dehghanian, Negar Movaghatian, Maryam Amirmazlaghani
Amirkabir University of Technology
No. 350, Hafez Ave, Tehran, Iran
{aref3463,z.dehghanian,n.movaghatian,mazlaghani}@aut.ac.ir

ABSTRACT

The integration of text onto objects within an image frequently results in an unnatural appearance, where the text appears overlaid rather than seamlessly embedded. Existing text-to-image models often encounter challenges in accurately incorporating text within an image context. This paper introduces a novel conditional inpainting technique aimed at overcoming these limitations. Our proposed model showcases exceptional effectiveness in addressing this issue, achieving compelling results by integrating text into images, thus significantly enhancing the naturalness and coherence of the final compositions.

1 INTRODUCTION

Text-to-image models such as DALL·E-3 Betker et al. (2023), Stable Diffusion (SD) Rombach et al. (2022) and Stable Diffusion XL (SDXL) Podell et al. (2023) excel in creating visuals from text. However, these models face challenges in accurately transcribing the input text onto the image, a process referred to as text-into-image. This limitation constrains their ability to precisely integrate textual elements onto specific objects within the generated images. As the usage of these models continues to expand, the importance of accurately incorporating real text into images is on the rise. Notably, we stand as pioneers in introducing this specific problem.

Our proposed strategy enhances image generation accuracy by incorporating specific details. This includes Canny edge Canny (1986) representations of the specified text, capturing its visual nuances. This step is pivotal as it serves as a condition image for the ControlNet model by Zhang et al. (2023). This model introduces noise to the entire image and subsequently ensuring the preservation of the Canny image's shape within the original image during the denoising process. Additionally, we have developed two innovative evaluation metrics tailored for this task, setting a new standard for assessing text integration in images.

2 METHODOLOGY

In our methodology, we use three primary inputs. The first input is the original image designated for editing. Accompanying this is a mask, that precisely outlines the area within the image where the overlaying element is intended to be placed. The third input is the actual text, logo, chosen shape, or image of the text that needs to be integrated into the specified area of the input image.

The process commences with the extraction of two distinct features. Initially, we apply the Canny edge detection algorithm to the image of the target text or shape. This step is crucial as it helps delineate the edges in the image, providing a structural outline that guides the placement and integration of the text. Following this, we focus on extracting detailed information about the context and color of the background where the text is to be placed. This is achieved in two phases. First, the SAM model Kirillov et al. (2023) is employed to effectively segment an object based on the input mask. Notably, even in cases where the mask partially obscures the object, the SAM model proficiently retrieves the complete object. Subsequently, the BLIP model Li et al. (2022) is utilized to extract pertinent information about the identified target object, encompassing details such as its name and color. This acquired information is then employed as a textual prompt for our proprietary

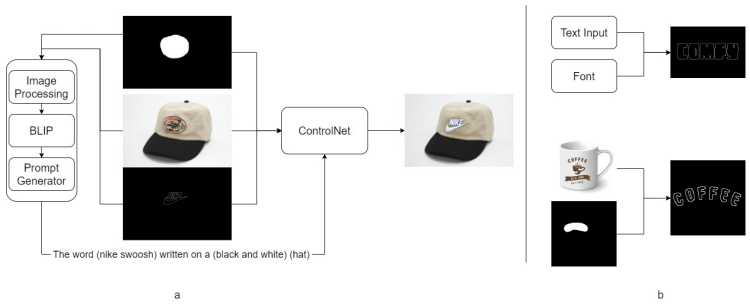

Figure 1: a: Proposed Model Architecture: Our proposed model architecture seamlessly integrates source image, mask, and canny image inputs. Through advanced image processing, we align the canny image with the mask, segment the target object, and employ the BLIP model to inquire about its identity. Harnessing this information, we generate a prompt and input it into the ControlNet model, along with the input images. b: Dual Canny Edge Generation: Comparison Between Text-Driven and Image-Driven Approaches

model. This information is pivotal in ensuring that the text blends in seamlessly, both in terms of visual aesthetics and contextual relevance.

In the final stage of our methodology, extracted features are inputted into the ControlNet model as a textual prompt along with the Canny image as the conditioning image, ensuring the preservation of abstract characteristics to a greater extent during the denoising process. Figure 1 shows our proposed model to inject text into an image. In the next section, we will see the resultant output.

3 Experiments and Results

In this section, the performance of the proposed model is scrutinized in Figure 2. It is evident that the output demonstrates notable improvement in both linguistic quality and meaningfulness of the generated text. In our experimental setup, employing SD or SDXL proved insufficient for generating accurate images solely from prompts. Augmenting SD with the Canny image of the text demonstrated improved results, further enhanced when combined with the prompt. More results, ablation study and numerical evaluation metrics are stated in the appendix.

Figure 2: The comparison between the output of SD, SDXL, and ours

4 Discussion and Conclusion

Our paper introduces a pioneering solution to the novel text-into-image problem. As the first to address this challenge, we propose a method that goes beyond conventional approaches by incorporating the output of the Canny edge detector as an image condition, alongside textual prompts. This innovative dual-condition approach enhances the linguistic authenticity of generated text and preserves the visual quality of outputs. In our study, we aimed beyond mere text imposition onto input images, particularly addressing challenges posed by non-planar objects. It can replace previous content beneath the mask and generate a completely new design in its place. Our work not only breaks new ground in text-into-image synthesis but also opens avenues for future advancements in the field.

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

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2023.

John Canny. A computational approach to edge detection. *IEEE Transactions on pattern analysis and machine intelligence*, (6):679–698, 1986.

JaidedAI. Easyocr: Ready-to-use ocr with 80+ supported languages. `https://github.com/JaidedAI/EasyOCR`, 2024. Accessed: 17 January 2024.

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.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pp. 12888–12900. PMLR, 2022.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

John W Ratcliff, David Metzener, et al. Pattern matching: The gestalt approach. *Dr. Dobb's Journal*, 13(7):46, 1988.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3836–3847, 2023.

A NUMERICAL EVALUATION

In this paper, we have introduced two distinct metrics for evaluating our text-into-image synthesis model: a subjective metric and an objective metric. The subjective metric involves human evaluators who are tasked with selecting the best option out of three models. The evaluation focuses on assessing naturalness and correctness in terms of context and placement. We called this metric, Individual Perception Score (IPS). This approach is crucial in capturing human perceptual judgments, offering insights into the model's effectiveness from a viewer's perspective. The evaluators assess not just the visual appeal but also the contextual appropriateness of the text within the image.
On the other hand, we employ a nuanced approach for the objective metric, integrating the use of the Easy OCR JaidedAI (2024) algorithm with the Gestalt Pattern Matching (GPM) score Ratcliff et al. (1988). Initially, the Easy OCR algorithm is applied to detect text within the synthesized images. Subsequently, the recognized text strings are subjected to the GPM score evaluation. The GPM score, rooted in a string-matching algorithm, is adept at assessing the similarity between the detected text and the original input text. This method provides a sophisticated measure of the text's accuracy and integrity post-synthesis. By comparing the OCR-detected text with the intended text, the GPM score quantifies the degree to which the synthesized text maintains its original content and structure. This metric, therefore, offers a valuable objective assessment, capturing not just the visibility and clarity of the text, but also its fidelity to the source text. The results of these scores are shown in table1.

Table 1: Results of different Evaluation Metrics

	SD	SDXL	Ours
IPS	0.08	0.14	0.78
GPM	0.21	0.28	0.95

Together, these metrics form a comprehensive evaluation framework, effectively capturing both human qualitative assessments and quantitative algorithmic analyses. This dual approach ensures a balanced and thorough evaluation of our model, addressing both aesthetic and functional aspects and providing a robust understanding of the model's capabilities in the realm of text-into-image synthesis.

B MORE VISUAL RESULTS

In this section, we first present an expanded set of results to further demonstrate the efficacy and versatility of our proposed model in Figure 3.

Through these experiments, we identified three pivotal elements crucial for precise text-into-image conversion: the target text, the designated target object and its corresponding color, and the Canny image edges of the text (or image). In figure 4 we do an ablation study, by visually illustrating the exclusion of any of these components, emphasizing their collective significance in this process.

Exploring various conditions, our findings indicate that incorporating the Canny edge consistently yielded superior results than the other edge detection techniques, as depicted in Figure 5.

Furthermore, Figure 6 illustrates the conditioning scale's impact, a parameter that regulates the influence ratio of the Canny edge.

Our investigation revealed that the material of an object has a noteworthy influence on how accurate and realistic the generated images appear. Notably, we found fascinating outcomes when working with fabric fibers. These outcomes demonstrated exceptional integration of the text into the fibers, creating a detailed appearance akin to a sewn pattern. Additionally, we consistently noticed that using higher resolution images led to better results, highlighting the importance of having high-quality images in this process.

Moreover, Figure 7 accentuates the significance of aligning the Canny edge within the mask region to optimize outcomes. This alignment is crucial as variations in the size of the mask directly impact the fitting of the target text. A mask that is too small might result in the target text being trun-

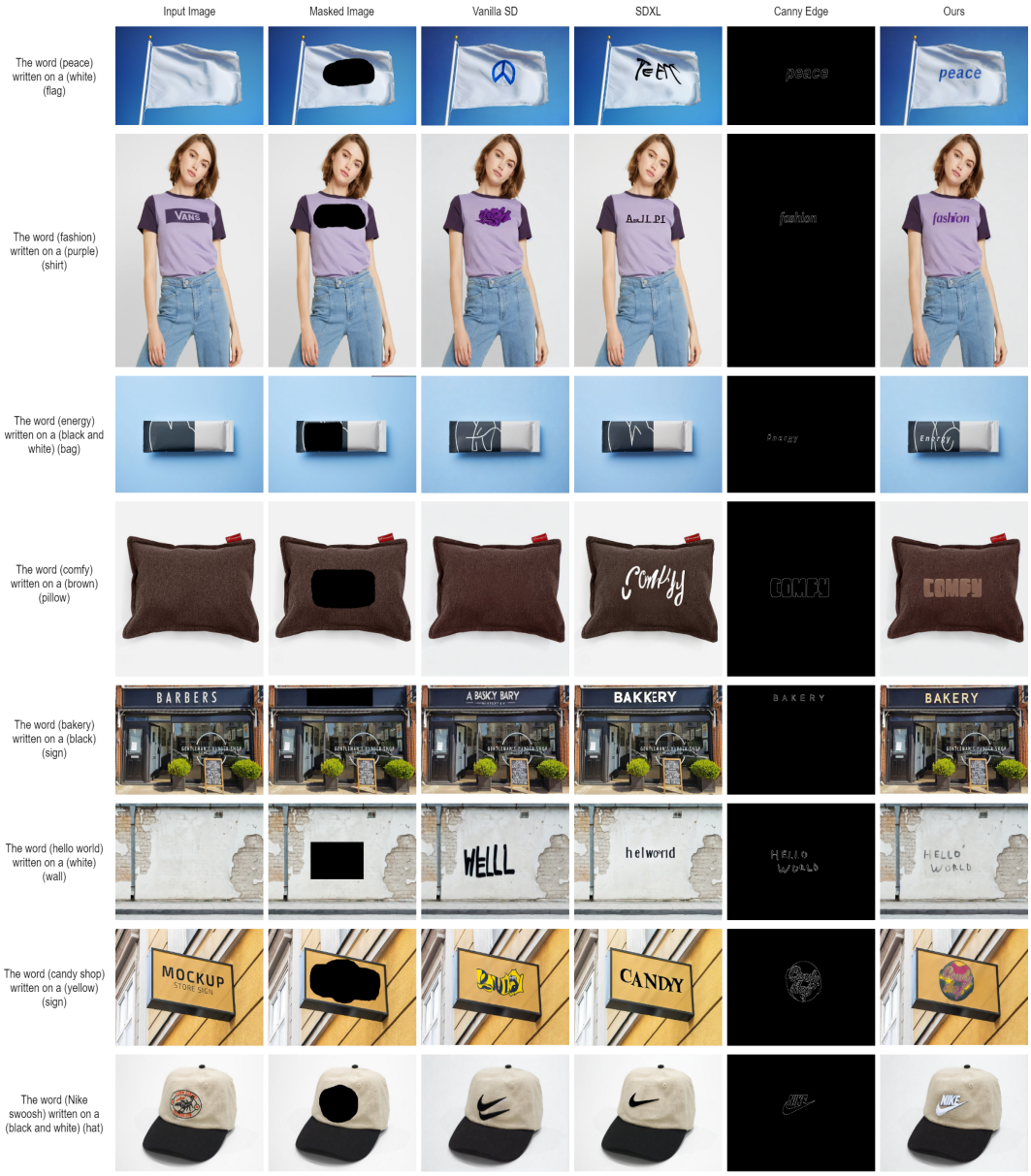

Figure 3: The Output of Our Proposed Model Compared to SD and SDXL.

cated, whereas an oversized mask can lead to the inclusion of unwanted elements. Therefore, this alignment is conducted to attain superior results.

Also, to further illustrate the robustness and versatility of our model, we have included an additional comparative analysis in Figure 8 This figure presents a side-by-side comparison of outputs generated by SD, SDXL, and our proposed model, using six different seeds. This comparative analysis clearly demonstrates the natural visual quality, linguistic accuracy, and overall robustness inherent in our model.

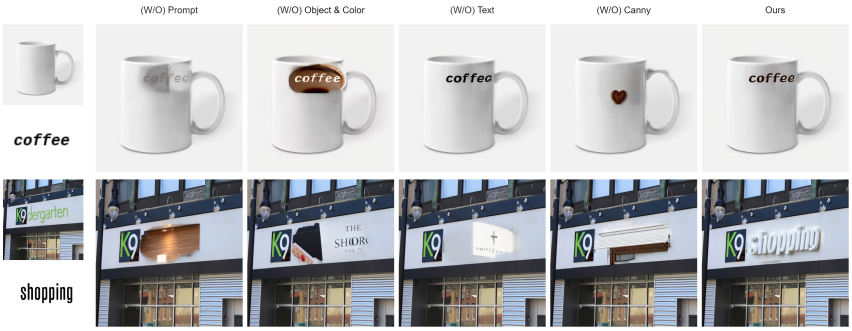

Figure 4: Visual Impact of Key Elements: Eliminating Individual Components Highlights Their Critical Role in Shaping the Final Image Representation

Figure 5: ControlNet Model Evaluation: Canny Edge Detection Prevails Across Varied Conditions

Figure 6: Effect of Conditioning Scale: Comparison of ControlNet's Performance Across Different Scale Parameters

Figure 7: The Significance of Aligning the Canny Edge Within the Mask Region

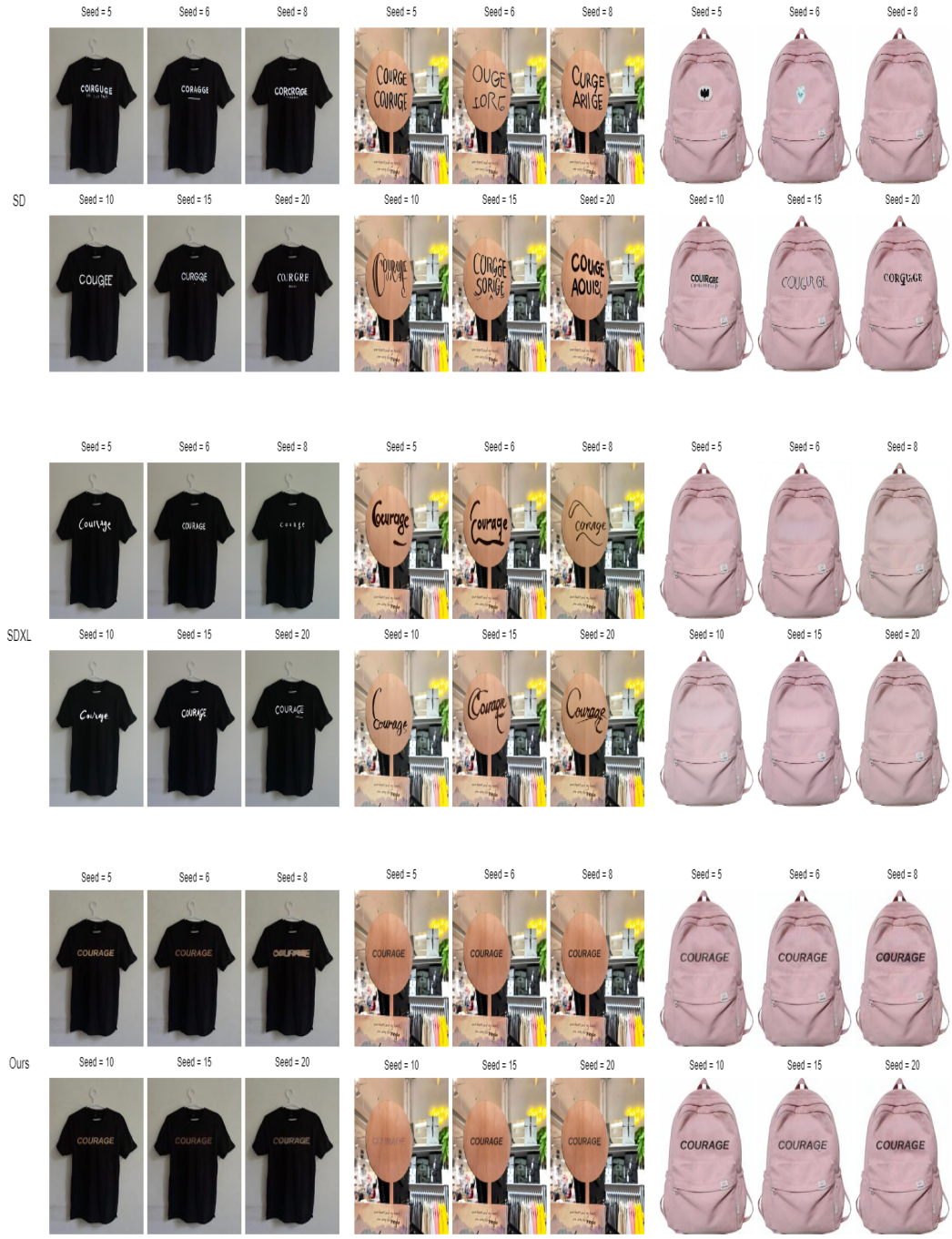

Figure 8: The variation of outputs within random seeds