Localized Text-to-Image Generation For Free via Cross Attention Control

Yutong He, Ruslan Salakhutdinov & J. Zico Kolter
Carnegie Mellon University
{yutonghe,rsalakhu,zkolter}@cs.cmu.edu

Abstract

Despite the tremendous success in text-to-image generative models, *localized* text-to-image generation (that is, generating objects or features at specific locations in an image while maintaining a consistent overall generation) still requires either explicit training or substantial additional inference time. In this work, we show that localized generation can be achieved by simply controlling cross attention maps during inference. With no additional training, model architecture modification or inference time, our proposed cross attention control (CAC) provides *new* open-vocabulary localization abilities to standard text-to-image models. CAC also enhances models that are already trained *for* localized generation when deployed at inference time. Furthermore, to assess localized text-to-image generation performance *automatically*, we develop a standardized suite of evaluations using large pretrained recognition models. Our experiments show that CAC improves localized generation performance with various types of location information ranging from bounding boxes to semantic segmentation maps, and enhances the compositional capability of state-of-the-art text-to-image generative models.

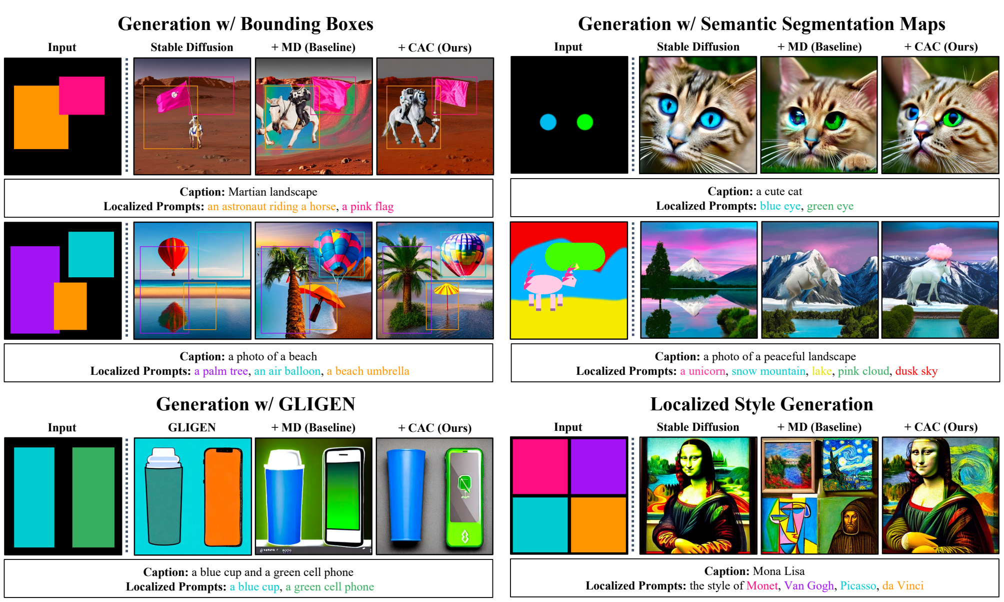

Figure 1: CAC as a plugin to existing methods for localized text-to-image generation. CAC improves upon diverse types of localization (bounding boxes, semantic segmentation maps and localized styles) with different base models (Stable Diffusion and GLIGEN). [2]

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

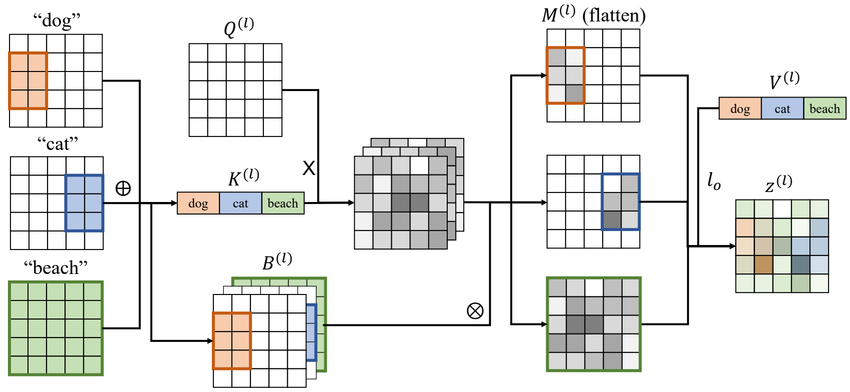

Figure 2: The illustration of CAC for localized generation. CAC uses localized text descriptions and spatial constraints to manipulate the cross attention maps.

Text-to-image generative models have shown strong performance in recent years: models like Stable Diffusion [1] and Dall-E [2] are capable of generating high quality and diverse images from arbitrary text prompts. However, a significant challenge faced by these models is that they rely *solely* on text prompts alone for content control over the generation process, which is inadequate for many applications. Specifically, one of the most intuitive and user-friendly ways to exert control over the generation is to provide *localization information*, which guides the models on where to generate specific elements within the image. Unfortunately, current pretrained models face limitations in their capability to perform localized generation. These limitations arise not only from their inability to incorporate location information as input but also from the inherent difficulties associated with compositionality, which is a known challenge for many multimodal foundation models [3].

Existing methods addressing this issue typically fall into three main categories: training entirely new models [4, 5], fine-tuning existing models with additional components such as task-specific encoders [6], or strategically combining multiple samples into one [7, 8]. All of these approaches often demand a substantial amount of training data, resources, and/or extended inference time, rendering them impractical for real-life applications due to their time and resource-intensive nature. On the other hand, in a separate but related line of work, [9] proposed Prompt-to-Prompt Image Editing, which edits generated images based on modified text prompts by manipulating cross attention maps in text-to-image generative models. Notably, this work also shows that cross attention layers play a pivotal role in controlling the spatial layout of generated objects associated with specific phrases in the prompts.

In this work, we propose to use cross attention control (CAC) to provide pretrained text-to-image models with better open-vocabulary localization abilities. As illustrated in Figure 1, given a caption and localization information, such as bounding boxes and semantic segmentation maps, along with their corresponding text descriptions, we first construct a new text input by concatenating the caption and all prompts associated with the location information. We then compute the cross attention maps from this new text prompt and apply localization constraints to the cross attention maps according to the localization information. Our method does not require any additional training or model architecture modification like designing task-specific encoders. It also does not impose any language restrictions such as using a fixed set of vocabulary or a language parser. Moreover, it is highly portable and can be easily integrated into a single forward pass in any cross attention based text-to-image generation framework with only a few lines of code, thus demanding no extra inference time.

We develope a standardized suite of evaluation metrics for localized text-to-image generation tasks using off-the-shelf large pretrained recognition models [10, 11, 12, 13]. We apply CAC to various state-of-the-art baseline text-to-image generative models and experiment with different forms of localization information including bounding boxes and semantic segmentation maps. We demonstrate that CAC endows pretrained standard text-to-image models with new localized generation abilities, and furthermore, improves upon models specifically trained for localized generation. In addition, we show that with simple heuristics that spatially separate the components within text prompts, our method can significantly improve the compositional ability of text-to-image generative models.

[2] All shades of pink in the middle right example correspond to the prompt "unicorn".

Acknowledgement

This work is supported by fundings from the Bosch Center for Artificial Intelligence and in part by ONR N000142312368. We would like to thank Joshua Williams for his supports on human evaluations, Minji Yoon and Jing Yu Koh for proof reading this paper, and Ellie Haber, Yiding Jiang, Jeremy Cohen, Yuchen Li and Samuel Sokota for their helpful feedback and discussions.

References

[1] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021.

[2] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation, 2021.

[3] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *CVPR*, 2022.

[4] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

[5] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. *CVPR*, 2017.

[6] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. *CVPR*, 2023.

[7] Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XVII*, pages 423–439. Springer, 2022.

[8] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: Fusing diffusion paths for controlled image generation. *arXiv preprint arXiv:2302.08113*, 2023.

[9] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. 2022.

[10] Glenn Jocher, Ayush Chaurasia, and Jing Qiu. YOLO by Ultralytics, January 2023.

[11] Jiaqi Chen, Zeyu Yang, and Li Zhang. Semantic segment anything. `https://github.com/fudan-zvg/Semantic-Segment-Anything`, 2023.

[12] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything, 2023.

[13] Liunian Harold Li*, Pengchuan Zhang*, Haotian Zhang*, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *CVPR*, 2022.