User Study And Evaluation Of Saliency-guided Image Generation Result

Nan Wu

Motivation and Goals

Diffusion models offer unprecedented image generation capabilities given just a text prompt. While emerging control mechanisms have enabled users to specify the desired spatial arrangements of the generated content[1][2][3], they cannot predict or control where viewers will pay more attention due to the complexity of human vision. Recognizing the critical necessity of attention controllable image generation in practical applications, there has been work on presenting a saliency-guided framework to incorporate the data priors of human visual attention into the generation process. Given a desired viewer attention distribution, the control module conditions a diffusion model to generate images that attract viewers' attention toward desired areas.

In this project, I propose an eye-tracked user study saliency analysis to assess the efficacy of this approach, where we set up an eye tracking device, show users a collection of generated images generated by the saliency-guided model and record users' gaze paths during their viewing. The results will then be used to compare with the conditioned saliency models to evaluate the efficacy of the model.

Related work

Controllable Diffusion Models

Popular approaches to controlled image[1] generation include lightweight adaptation modules built around a foundation model. The adapter networks are usually conditioned by depth maps, semantic segmentation masks, body poses, or bounding boxes[2][3] to control the spatial layout of an image or video.

Human Visual Attention Models and Saliency Prediction

Researchers have attempted to develop saliency models in a bottom-up fashion from image space statistical features[4], as well as conducting large-scale user studies [5] and utilizing deep neural networks to measure compounded influences affecting human visual attention [6].

References

[1] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023b. **Adding conditional control to text-to-image diffusion models.** In Proceedings of the IEEE/CVF International Conference on Computer Vision. 3836–3847.
[2] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. 2023b. **Gligen: Open-set grounded text-to-image generation.** In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 22511–22521.
[3] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. 2023. **Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models.** arXiv preprint arXiv:2308.06721 (2023).
[4] Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba. 2009. **Learning to predict where humans look.** In 2009 IEEE 12th international conference on computer vision. IEEE, 2106–2113.
[5] Xun Huang, Chengyao Shen, Xavier Boix, and Qi Zhao. 2015. **Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks.** In Proceedings of the IEEE international conference on computer vision. 262–270.
[6] Sen Jia and Neil DB Bruce. 2020. **Eml-net: An expandable multi-layer network for saliency prediction.** Image and vision computing 95 (2020), 103887.