Statement of Purpose

Jing Yu Koh (PhD applicant, Fall 2022)

Research interests: People build a coherent understanding of the world by exploring, perceiving, and imagining through their many senses. They efficiently integrate information from multiple sources that provide complementary information about their environment. My long-term goal is to build computational models that can do the same, fusing different modalities (images, text, audio, videos, and more) to achieve robust and reliable performance on generation and grounding tasks. Ultimately, I want to develop systems that can generate diverse and accurate representations of the real world consistent with visual, semantic, physical, and geometrical phenomena. As first steps, in my existing research as a Research Software Engineer at Google, I have achieved state-of-the-art results in text-to-image synthesis, proposed the first world model capable of consistent high resolution generation in 3D indoor environments, and demonstrated that multi-modal information can improve semantic segmentation. My work has led to three first-author publications at top conferences (ICCV 2021 [1], CVPR 2021 [2], ECCV 2020 [3]), a first-author publication at WACV 2021 [4], and several other collaborations. My goal is to build on the flexibility, utility, and speed of such models to discover and identify new phenomena about cognition and the real world.

Selected past and current research: As an undergraduate, I was keen to research how multi-view and multi-modal data might improve performance for computer vision tasks. I explored fusing overhead satellite imagery with street-level photographs for semantic segmentation, and developed an end-to-end trainable network with conditional spatial distribution [3]. As the lead author, I proposed the initial idea for the model, ran experiments, and wrote most of the paper. A key novelty of our approach is its ability to use user attention: domain experts can influence model outputs at inference time using mouse trace annotations. This is essential for tasks with high failure costs (such as the medical domain).

After joining Google in October 2019, I initiated a project to apply similar ideas to improve text-to-image synthesis by grounding image generation with user attention. Here, users describe a scene by simultaneously speaking and moving their mouse, intuitively dictating the type of objects to be created alongside their corresponding spatial locations. The difficulty of this problem lies in generating realistic scenes that are coherent with language descriptions and spatial configuration. To approach this previously untackled problem, I tested several hierarchical and end-to-end generative models, converging on the final model TReCS [4] through a series of extensive experiments. TReCS both showed viability on this new problem and outperformed existing text-to-image synthesis models.

I was motivated by these strong results to further examine how multi-modal models might be improved through stronger inter-modality grounding. I experimented with adding inter-modal (text-image) and intra-modal (image-image) contrastive learning to text-to-image generation models, and discovered that they benefited immensely from such a training regime due to improved consistency between the language and vision modalities. This enabled us to greatly simplify text-to-image generation models while retaining strong performance (Fig. 1). Our approach, XMC-GAN [2], achieves state-of-the-art, improving results on the MS-COCO dataset by 62% over prior work. Human raters also much prefer the images generated, rating XMC-GAN as more realistic than three other models 77% of the time.

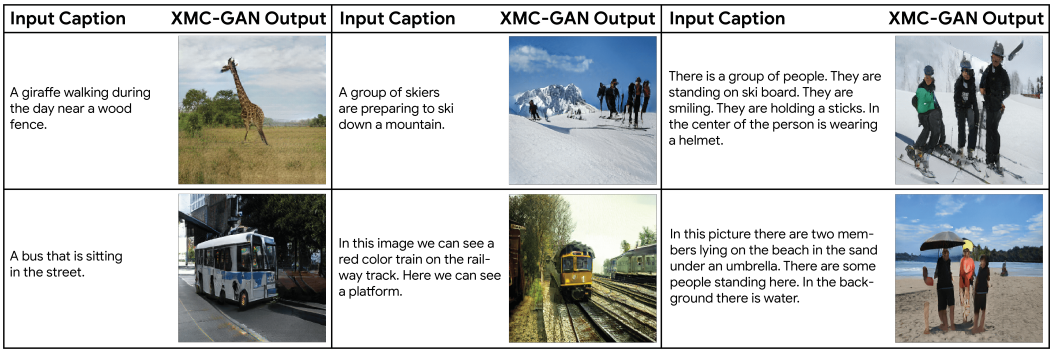

Figure 1: XMC-GAN [2] is capable of generating realistic scenes when conditioned on diverse input captions.

These exciting results encouraged me to push further on generative multi-modal research. Multi-modal models are more interpretable: we can easily change inputs and immediately observe understandable outputs. Throughout our development process, this was instrumental in allowing us to easily analyze correlations, biases, and examine the nuances of language and vision learnt by the model. Model robustness and quality is also enhanced through leveraging information from disparate modalities to improve the others.

The TReCS and XMC-GAN work was the catalyst for a new large cross-team collaboration on text-to-image generation within Google Research. My role in this has involved engineering to scale data and models to the multi-billion scale, and research in new geometry-aware generative models. This research has already resulted in a number of advances to generation quality [5] and our understanding of responsible AI for the text-to-image generation task, as well as a journal

paper on this work currently in preparation.

Given the exciting results from our text-to-image synthesis work, I was keen to further explore how additional real-world constraints, such as those imposed by 3D environments and physical actions, might be used to develop models with richer representations of their simulated world. I saw an opportunity to address the even more challenging grounding problem of novel view synthesis in 3D environments. When tasked to navigate in a foreign environment, humans plan actions by imagining vivid scenes of what we expect to see. Developing a predictive world model is one way of equipping computational agents with planning and pragmatic reasoning abilities. To facilitate research towards this goal, I led the design and development of Pathdreamer [1], a high resolution generative world model for indoor environments. Pathdreamer unifies 2D image representations with 3D geometry to generate consistent long-term predictions from a single image of a previously unseen environment (Fig. 2). Pathdreamer is useful as a predictive world model for downstream tasks, enabling low-cost simulations of multiple potential futures that ease planning and reasoning for navigation agents. Notably, Pathdreamer is the first known model-based approach for the vision-and-language navigation problem, and improves success rate by 24% relative to baselines.

Figure 2: Outputs from the Pathdreamer [1] model. The model generates observations for 3 new viewpoints traversing a corridor, generating scenes that demonstrate semantic knowledge of indoor environments.

My current work extends Pathdreamer to continuous video sequences (such as house tour videos [6]). Video data contains rich temporal dependencies which encode informative structure about the environment. The primary drawback is that imagery from a video is often much sparser compared to richly annotated 3D environments. To equip Pathdreamer with the ability to handle large regions of missing information, I massively simplified and improved the original two-stage model. Pathdreamer++ is a single-stage, end-to-end trainable model which is more efficient and greatly improves generation quality (24% over prior work). By training on diverse video sequences, Pathdreamer++ is capable of synthesizing higher quality 3D environments from single images. This unlocks generative data augmentation for indoor environments, which complements instruction augmentation, a standard procedure for achieving high performance in vision-and-language navigation. Augmentation is essential due to the paucity of diverse training environments, and our approach enables training of stronger agents, achieving a state-of-the-art success rate of 68%. These exciting results further fueled my interests in creating models grounded in real world phenomena: the geometrical constraints of the 3D world, the temporal dependencies of videos, and the expressivity of natural language are rich structures for training stronger models. Analogous to how humans learn through multiple senses, these grounded generative models fuse vision, language, environment, and action to achieve stronger and more robust performance.

Future research: I believe that gaining a better understanding of multi-modal models and generative models will provide fundamental insights on computational intelligence, paving the way for more robust, responsible, and creative models. Having seen and developed models for various multi-modal problems, I wish to further advance our understanding of how different modalities interact, and develop improved multi-modal models for vision-and-language reasoning, video and audio understanding, and 3D scene representation. I envision further research on grounded generative models for image synthesis, video generation, and language and audio generation. My research goals are twofold: (1) to develop models which are capable of producing representations of the world faithful to visual, semantic, and physical constraints, and (2) to extract new fundamental insights about intelligence and the world from computational models trained on curated data and simulations.

I would be thrilled to pursue a PhD at CMU. The work of several faculty members is specifically well-matched to my own research interests. Prof. Daniel Fried's research on grounded language understanding and embodied tasks is well-matched with my interests on grounded multi-modal learning. Prof. Yonatan Bisk's work on grounded language understanding is well-matched with my interests on multi-modal learning and language grounding for video prediction. Prof. Katerina Fragkiadaki's research on video understanding and scene representation is very closely aligned with my interests in multi-modal learning and developing robust world models. Prof. Jun-Yan Zhu's research on image and video synthesis is closely aligned with my research goals of developing stronger generative models and exploring alternative generative modeling paradigms. The broader CMU faculty and centers will provide an ideal research community as I pursue these questions. I believe that CMU will provide the best environment for me to learn, grow, and develop the necessary skills to succeed as a researcher and academic.

References

[1] **J. Y. Koh**, H. Lee, Y. Yang, J. Baldridge, and P. Anderson, "Pathdreamer: A world model for indoor navigation," *International Conference on Computer Vision (ICCV)*, October 2021. [PDF].

[2] H. Zhang*, **J. Y. Koh***, J. Baldridge, H. Lee, and Y. Yang, "Cross-modal contrastive learning for text-to-image generation," *Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021. (* denotes equal contribution) [PDF].

[3] **J. Y. Koh**, D. T. Nguyen, Q.-T. Truong, S.-K. Yeung, and A. Binder, "SideInfNet: A deep neural network for semi-automatic semantic segmentation with side information," *The European Conference on Computer Vision (ECCV)*, August 2020. [PDF].

[4] **J. Y. Koh**, J. Baldridge, H. Lee, and Y. Yang, "Text-to-image generation grounded by fine-grained user attention," *The IEEE Winter Conference on Applications of Computer Vision (WACV)*, January 2021. [PDF].

[5] J. Yu, X. Li, **J. Y. Koh**, H. Zhang, R. Pang, J. Qin, A. Ku, Y. Xu, J. Baldridge, and Y. Wu, "Vector-quantized image modeling with improved VQGAN," *In submission.*, October 2021. [PDF].

[6] T. Zhou, R. Tucker, J. Flynn, G. Fyffe, and N. Snavely, "Stereo magnification: learning view synthesis using multiplane images," *ACM Transactions on Graphics (TOG)*, vol. 37, no. 4, pp. 1–12, 2018.