Gi-Cheon Kang

Research Statement

My goal is to build **interactive AI systems** that can help people with real-world tasks, such as enabling robots to perform household tasks based on language instruction "Wash my clothes," or allowing digital assistants to help blind people with real visual challenges by talking with them. To build systems like these, I perform **interdisciplinary research** spanning the intersection of computer vision, natural language processing, and robotics. My research focuses on **grounding**: connecting language with perception (mostly vision) and action, enabling machines to understand the semantics of the physical world. By integrating insights from these diverse disciplines, I seek to advance the development of **AI agents that can see, talk, and act**, thereby contributing to solutions for pressing societal needs and pushing the boundaries of AI capabilities.

Vision is one of the most essential modalities of human intelligence. To bridge the gap between vision and language, I have developed **visually grounded interactive systems** that can continuously communicate with humans about images [1, 2, 3]. A major bottleneck in training these systems is **the difficulty of scaling visually grounded conversation data**. To tackle this challenge, I have introduced a new approach that automatically generates synthetic dialogue data regarding millions of images obtained from the Web. By leveraging synthetic data to train visually grounded dialogue systems, I have found that they produce accurate and robust responses about images when talking with humans. I have extended the image-based systems to **video-based interactive systems** [4]. We proposed a methodology for effectively fusing temporal and spatial information grounded in language, considering the unique properties of video data.

Young children learn about the semantics of the physical world not only through perception but also by manipulating their perception through interaction with the environment [5]. This view has helped me to extend visually grounded systems to **embodied AI systems** [6, 7, 8] that perform real-world tasks through language interaction with humans. My work has focused on **language-guided robotic manipulation**, where a physical robot arm should manipulate objects based on natural language instruction from human users. I have studied a new scenario in which the initial instruction is ambiguous without mentioning the target object. The embodied system should disambiguate the target object by seeing and talking with users. My work successfully executes real-world tasks with minimal interaction with humans.

1. Visually-grounded interaction with language

I have worked towards modeling visually grounded dialogue systems that can continuously communicate with humans about images. In addition to the challenge of grounding vision and language, these systems should hold a meaningful dialog history with humans in natural language about visual content. One of the critical problems in training visual dialog systems is the difficulty of scaling human-to-human visual dialog data. To this end, I introduced a semi-supervised learning approach, called Generative Self-Training (GST) [1], to scale data without human annotation. The key idea of GST is to generate synthetic dialog data for unlabeled Web images and train models on the data. I have shown that synthetic data makes significant performance gains on visual dialog [9]. Moreover, the use of synthetic data improves robustness against adversarial attacks.

I have also addressed the problem of visual reference resolution, where visual dialog systems should resolve ambiguous expressions in human utterances (*e.g.,* "What color is *it*?") and ground them to a given image. I have proposed attention-based methods [3, 2] that effectively retrieve relevant dialog history to clarify ambiguous expressions. They have demonstrated strong performance improvements over the prior art.

2. Embodied AI

Another big theme of my research is embodied interaction in real-world environments. I have worked on language-conditioned robotic manipulation, where embodied agents (*i.e.,* robots) manipulate objects based on natural language instructions from humans. A typical scenario of this problem specifies the category of the target object in instructions. Inspired by the fact that humans often convey their *intended* meanings by relying on context, I have introduced a new task and corresponding dataset to study the task [6]. The goal of the task is for robots to pick up the desired object in the given scene, but the language instructions are ambiguous (*e.g.,*"I'm thirsty."). Therefore, the agents should interact with humans by asking questions to disambiguate the target object. Based on the task setup, we propose a new embodied system that effectively interprets the user's intention and picks up the target object. The key mechanism of our system is determining one target object among candidates based on how well each object candidate explains the human-AI conversation, which we call *pragmatic inference.* We showcase that pragmatic inference helps identify the target object correctly with minimal human-AI interaction.

When deploying real-world embodied systems, we often witness some components, like vision models, struggle to perform in real-world environments due to the distributional shift. However, collecting real-world data for domain adaptation is expensive. I have investigated a lifelong learning framework [7] that enables models to adapt to the real-world environment without human supervision. In my work, the visual grounding model was found to be vulnerable to the domain shift, so I have proposed a method that automatically generates language instructions to the given scene to adapt the model.

3. Other work

In a separate line of research, I have studied two computer vision projects. The first work [10] is about reducing the texture bias of ConvNets in image classification. We propose a shape-focused augmentation, demonstrating its efficacy in robust visual representation learning. In another work [11], we investigate a new approach for synthesizing high-quality images using only a few image samples. The key idea was to bridge the gap between the source and target domain via contrastive learning, and it significantly improved performance on few-shot image generation.

4. Future research plans

Embodied foundation models. My short-term research plan is to develop foundation models for embodied AI that can perform diverse real-world tasks, including robotic manipulation and navigation. To achieve this, combining vision, language, and low-level control is a fundamental challenge. Existing approaches address this challenge by training language-conditioned visuomotor policies. However, they require copious amounts of human-annotated data to learn simple robotic tasks (*e.g.,* pick & place objects). Accordingly, I am excited to study an approach that trains language-conditioned visuomotor policies in a sample-efficient manner.

Lifelong multimodal learning. Unlike current AI models, humans continuously learn their grounded models of the world. By taking inspiration from humans, my long-term research will explore lifelong learning models capable of continuous learning and adaptation in dynamic environments. These models should evolve their capabilities by circumventing catastrophic forgetting when exposed to new data or tasks. I am excited to study new paradigms for training lifelong multimodal models, enabling them to expand their grounded knowledge over time.

References

[1] **Gi-Cheon Kang**, Sungdong Kim, Jin-Hwa Kim, Donghyun Kwak, and Byoung-Tak Zhang. The dialog must go on: Improving visual dialog via generative self-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

[2] **Gi-Cheon Kang**, Junseok Park, Hwaran Lee, Byoung-Tak Zhang, and Jin-Hwa Kim. Reasoning visual dialog with sparse graph learning and knowledge transfer. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2021.

[3] **Gi-Cheon Kang**, Jaeseo Lim, and Byoung-Tak Zhang. Dual attention networks for visual reference resolution in visual dialog. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019.

[4] Ahjeong Seo, **Gi-Cheon Kang**, Joonhan Park, and Byoung-Tak Zhang. Attend what you need: Motion-appearance synergistic networks for video question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2021.

[5] Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, et al. Experience grounds language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

[6] **Gi-Cheon Kang**, Junghyun Kim, Jaein Kim, and Byoung-Tak Zhang. Prograsp: Pragmatic human-robot communication for object grasping. In *International Conference on Robotics and Automation (ICRA)*, 2024.

[7] Junghyun Kim, **Gi-Cheon Kang**, Jaein Kim, Suyeon Shin, and Byoung-Tak Zhang. Gvcci: Lifelong learning of visual grounding for language-guided robotic manipulation. In *2023 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2023.

[8] Junghyun Kim, **Gi-Cheon Kang**, Jaein Kim, Seoyun Yang, Minjoon Jung, and Byoung-Tak Zhang. Pga: Personalizing grasping agents with single human-robot interaction. *arXiv preprint arXiv:2310.12547 (Submitted to IROS 2024)*, 2024.

[9] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. Visual dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[10] Sangjun Lee, Inwoo Hwang, **Gi-Cheon Kang**, and Byoung-Tak Zhang. Improving robustness to texture bias via shape-focused augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[11] Hyuk-Gi Lee, **Gi-Cheon Kang**, Chang-Hoon Jeong, Han-Wool Sul, and Byoung-Tak Zhang. C3: Contrastive learning for cross-domain correspondence in few-shot image generation. In *Proceedings of Workshop on Controllable Generative Modeling in Language and Vision (Ctrl-Gen) at NeurIPS*, 2021.