Research Statement

Xin Wang

Humans learn to perceive the world through multiple modalities such as vision, sound, and touch. Language is invented for communication and documentation, which enables knowledge reasoning and distinguishes humans from other animals. Therefore, language and perception lay foundations for artificial intelligence, and how to ground natural language onto real-world perception is a fundamental challenge to advance research from recognition to cognition, empowering various practical applications that require human-machine communication.

By bridging language, perception, and knowledge, my research goal is to build intelligent agents that understand, interact with, and reason about the multimodal world via natural language. I am interested in designing scalable inference and learning algorithms to **connect language, perception, and actions for knowledge acquisition**. More specifically, to close the loop between language and perception, my doctoral work mainly investigates the following questions:

- How can machines describe the world using natural language?
- How can machines interact with the world via natural language?
- How can visual grounding bridge different languages?

My research is also highly interdisciplinary—it draws methodologies from machine learning, natural language processing, computer vision, and robotics, along with inspirations from cognitive science and neuroscience. I enjoy collaborating with scientists and domain experts of different backgrounds for interdisciplinary research in these areas.

Below I describe my research experience as initial steps to answer the aforementioned questions, and present my future plans on this emerging research area.

1. Learning to Describe the Multimodal World

The world is fulfilled with various signals such as visuals, audios, text. Natural language serves as an essential means to go beyond visual recognition and describe the unstructured, multimodal world. Towards producing coherent, relevant, and generalizable descriptions of the activities in a given video (see Figure 1), we systematically study reinforcement learning, multimodal learning, knowledge integration, and transfer learning for multimodal grounded language generation. we propose the first *hierarchical reinforcement learning solution for video captioning* [7], employing a master agent that generates the latent semantic goals and a worker agent that renders the lexical items. By imposing structural constraints, we achieve the state-of-the-art performance of video captioning on the benchmark MSR-VTT dataset. We further consider a cross-modal attention based fusion approach "Watch, Listen, and Describe" [10] to improve the performance and account for the audio signals, as audio cues also play a crucial role in understanding the world in addition to visual signals.

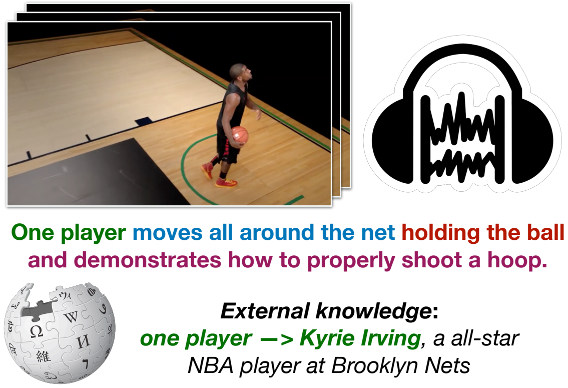

One player moves all around the net holding the ball
and demonstrates how to properly shoot a hoop.
External knowledge:
one player —> Kyrie Irving, a all-star
NBA player at Brooklyn Nets

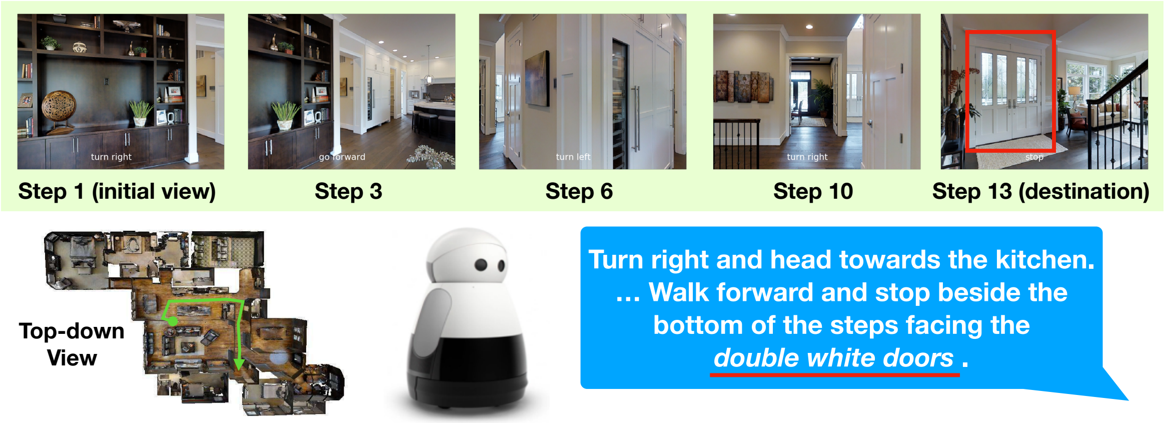
Figure 1. Learning to describe activities in a video.

Despite the success of the supervised captioning methods, human activities are far beyond the fixed inventory of activities represented in the training corpus. We hence introduce a new novel task, *zero-shot video captioning* [12], which aims at describing out-of-domain videos of unseen activities. Accordingly, we propose to utilize external knowledge (*e.g.*, Wikipedia and WikiHow) and develop a topic-aware meta-learning model [12] for novel activity captioning. In addition to domain generalization, we also collect a large-scale, high-quality multilingual dataset for video-and-language research [11], which goes beyond English and enables multilingual caption generation.

In our paper "No Metrics Are Perfect" [6], we observe that most evaluation metrics for text generation are based solely on text matching and thus could be easily gamed, and hence we propose a novel solution of *adversarial reward learning (AREL)* for visual storytelling, which aims at learning a reward function from human-written stories and optimizing the policy in an alternating fashion. Moreover, natural language is inherently ambiguous, and even human-written references may not fully cover the image content. Therefore, most recently, we introduce a new automatic evaluation metric TIGEr [3], which evaluates caption quality not only based on how well machine-generated captions match reference captions, but also on how well a caption represents the image content.

2. Language Guided Real-world Interactions

Current natural language understanding focuses on learning statistical language models from text-only corpora. Humans, however, acquire language by communicating and interacting in the real world, rather than learning the meaning of a word purely based on its relationship to other words. To simulate human behaviors, it is a fundamental capability of intelligent robots to navigate in visual environments by following natural language guidance, because humans can easily reason about the language guidance and efficiently interact with the visual environments. In the past two years, we have explored *grounding natural language instructions and visual inputs to actions* in real-world robot navigation tasks [13, 8, 9, 2] (*e.g.*, the vision-and-language navigation task as presented in Figure 2).



Turn right and head towards the kitchen. … Walk forward and stop beside the bottom of the steps facing the double white doors .

Figure 2. Given a natural language instruction, the robot learns to navigate through a 3D indoor environment to a destination without access to the top-down view.

We view the vision-and-language navigation task as a Markov decision process and tackle it via reinforcement learning. First, we combine *model-based and model-free reinforcement learning*, and tightly integrates a look-ahead policy model for planning [13]. Then we identify a critical issue in VLN that the success signal is too coarse to guide the instruction following navigation well. Hence, in our CVPR 2019 paper [8], we propose a *reinforced cross-modal matching* model that introduces a cycle-reconstruction intrinsic reward to encourage instruction-trajectory alignment along with the extrinsic reward. Meanwhile, we design a *self-supervised imitation learning method* to explore unseen environments without any supervision, greatly reducing the performance gap between seen and unseen environments. **This paper was selected as the Best Student Paper at the top computer vision conference CVPR 2019**.

We also aim at resolving the generalization and data scarcity issues that commonly exist in robotics tasks. Most recently, we develop a *generalized and scalable multitask navigation agent* [9] that cannot only follow natural language instructions but also interact with humans through dialog. We build a distributed navigation learning framework and simultaneously train the model on multiple navigation tasks. In addition, we propose a solution of *environment-agnostic learning* to learn representations that are environment-invariant but still effective in navigation [9]. Furthermore, we adopt the concept of *counterfactual thinking in psychology* and propose a *adversarial path sampler* [2] to sample increasingly challenging paths and augment them with a back-translated speaker model, which shows its effectiveness on various navigation models regardless of their model architecture.

Aside from the language grounded navigation tasks, we present that the natural language query can also be used to retrieve and localize a short video clip from an untrimmed, long video. We present a novel framework that unifies the candidate moment encoding and temporal structural reasoning in a single-shot feed-forward network [18], and achieves the new state of the art results on two challenging benchmarks DiDeMo and Charades-STA.

3. Multilingual Language Grounding

When we talk about natural language processing, we should keep in mind that there are thousands of languages on this planet though currently, English studies dominate this inclusive research area. To promote the inclusion of different languages and serve speakers with different linguistic backgrounds, we are dedicated to working on the multilingual study grounded on multiple modalities and domains [11, 17, 1, 15].

Figure 3. Vision can be the bridge between languages.

We consider vision as a bridge between languages, as humans first perceive the world through their eyes and then develop languages for communication. For instance, when presented an image in Figure 3, a person can immediately acquire the meaning of those words in different languages regardless of what languages she or he can speak. We introduce the *first large-scale multilingual video-and-language dataset* VaTeX [11] to enable multilingual study (*e.g.*, English and Chinese) in various downstream tasks such as video captioning, video-guided machine translation, and video-text retrieval. Meanwhile, as an initial step to move beyond monolingual interactions with robots, we propose a *cross-lingual vision-and-language navigation framework* [17], which meta-learns visually grounded cross-lingual representations for executing a command in multiple languages in zero-shot and low-resource settings.

In addition to language and vision, we have also made fundamental breakthroughs in the multilingual study of other domains. We draw inspiration from comparable corpora mining and develop an unsupervised machine translation approach [15] to extract and edit real sentences from the target monolingual corpora, which bypasses the overreliance of machine translation systems on large parallel corpora and avoids the error accumulation issue in back-translation-based approaches. Moreover, we make use of the machine translation systems and propose to perform dialogue state tracking of foreign dialogues (*e.g.*, German and Italian) without the needs of annotations [1].

4. Future Research Agenda

My past research has outlined the importance of integrating language, perception, and actions to acquire knowledge for real-world applications. Moving beyond, my long-term research goal is to build an AI platform that can learn from enormous unlabeled data of all kinds of modalities, that reasons about the world with the acquired commonsense knowledge, that communicates with people of various backgrounds, and that can perform complex tasks for humans in this physical world. With the philosophy in mind, I identify the following four research topics that I am thrilled to pursue next.

Self-supervised Learning Despite the incredible progress supervised learning methods have achieved on many human-annotated tasks, tremendous data are created without annotations every day by billions of internet users. So going beyond human supervision, I believe self-supervised learning has great potential to learn meaningful representations from raw data by utilizing learning signals in data itself. I will do self-supervised learning for more real-world tasks from two perspectives. First, we can utilize the underlying structured information in raw data to learn general-purpose representations that can benefit a variety of downstream tasks. For example, we can utilize the noisy alignment between different modalities in raw web data (*e.g.*, text and visuals) to train a generalizable representation model for language and vision tasks such as visual captioning and visual question answering. Second, we can use the task-specific, self-supervised learning signals to boost the task performance. For instance, we employed the order consistence of the dialog as the self-supervised signal for dialog tasks [16] and sentence context for text summarization tasks [14].

Knowledge-based Commonsense Reasoning Today's multimodal learning algorithms have mostly focused on task-specific domain learning from raw data with no usage of external knowledge that humans have built. Humans, instead, not only acquire new knowledge from observations of the world, but also inherit knowledge from the past and others. So I think the key missing ingredients of multimodal learning are external knowledge utilization and never-end learning for commonsense reasoning. We have proved that external knowledge can be a reliable source for zero-shot learning [12, 5]. I will then work on how to use the structured external data for commonsense reasoning and how to continually harvest new knowledge and incorporate it into the existing repertoire.

Connecting Language and Perception for Robotics One of the long-term challenges of robotics is to enable robots to interact with humans in the visual world via natural language, as humans are visual animals that communicate through language. Overcoming this challenge requires the ability to perform a wide variety of complex tasks in response to multifarious instructions from humans. To connect language and perception for large-scale robot learning, I plan to close the simulation-deployment loop. In addition to collecting more practical and more challenging data for language grounded robotics tasks [4], I will continue building a scalable and generalizable multitask simulation platform [9] to enable large-scale training across tasks. Meanwhile, I will work on physical robotics and transfer the knowledge learned from photo-realistic simulation environments to the real world.

Fairness in Multimodal Machine Learning I intend to conduct inclusive research to serve people of diverse backgrounds, *e.g.*, genders, races, and nationalities. Our work on multilingual language grounding [11, 17, 1] has made me realize that people of different linguistic backgrounds have different focuses on describing the surroundings. Recent study [19] also suggests that severe gender bias issues exist in various tasks. So I will promote and study the fairness in multimodal machine learning, in order to build fair models to remove potential biases and accommodate differences among people.

References

[1] Wenhu Chen, Jianshu Chen, Yu Su, **Xin Wang**, Dong Yu, Xifeng Yan, and William Yang Wang. XL-NBT: A cross-lingual neural belief tracking framework. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

[2] Tsu-Jui Fu, **Xin Wang**, Matthew Peterson, Scott Grafton, Miguel Eckstein, and William Yang Wang. Conterfactual vision-and-language navigation via adversarial path sampling. *Submitted to The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[3] Ming Jiang, Qiuyuan Huang, Lei Zhang, **Xin Wang**, Pengchuan Zhang, Zhe Gan, Jana Diesner, and Jianfeng Gao. TIGEr: Text-to-image grounding for image caption evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019.

[4] Yuankai Qi, Qi Wu, Peter Anderson, **Xin Wang**, William Yang Wang, Chunhua Shen, and Anton van den Hengel. Reverie: Remote embodied visual referring expression in real indoor environments. *Submitted to the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[5] Pengda Qin, **Xin Wang**, Wenhu Chen, Chunyun Zhang, Weiran Xu, and William Yang Wang. Generative adversarial zero-shot relational learning for knowledge graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2020.

[6] **Xin Wang***, Wenhu Chen*, Yuan-Fang Wang, and William Yang Wang. No metrics are perfect: Adversarial reward learning for visual storytelling. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2018.

[7] **Xin Wang**, Wenhu Chen, Jiawei Wu, Yuan-Fang Wang, and William Yang Wang. Video captioning via hierarchical reinforcement learning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

[8] **Xin Wang**, Qiuyuan Huang, Asli Celikyilmaz, Jianfeng Gao, Dinghan Shen, Yuan-Fang Wang, William Yang Wang, and Lei Zhang. Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[9] **Xin Wang**, Vihan Jain, Eugene Ie, William Yang Wang, Zornitsa Kozareva, and Sujith Ravi. Generalized natural language grounded navigation via environment-agnostic multitask learning. *Submitted to International Conference on Learning Representations (ICLR)*, 2020.

[10] **Xin Wang**, Yuan-Fang Wang, and William Yang Wang. Watch, listen, and describe: Globally and locally aligned cross-modal attentions for video captioning. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT)*, 2018.

[11] **Xin Wang***, Jiawei Wu*, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *The IEEE International Conference on Computer Vision (ICCV)*, 2019.

[12] **Xin Wang**, Jiawei Wu, Da Zhang, Yu Su, and William Yang Wang. Learning to compose topic-aware mixture of experts for zero-shot video captioning. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2019.

[13] **Xin Wang***, Wenhan Xiong*, Hongmin Wang, and William Yang Wang. Look before you leap: Bridging model-free and model-based reinforcement learning for planned-ahead vision-and-language navigation. In *The European Conference on Computer Vision (ECCV)*, 2018.

[14] Hong Wang, **Xin Wang**, Wenhan Xiong, Mo Yu, Xiaoxiao Guo, Shiyu Chang, and William Yang Wang. Self-supervised learning for contextualized extractive summarization. 2019.

[15] Jiawei Wu, **Xin Wang**, and William Yang Wang. Extract and edit: An alternative to back-translation for unsupervised neural machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT)*, 2019.

[16] Jiawei Wu, **Xin Wang**, and William Yang Wang. Self-supervised dialogue learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2019.

[17] An Yan*, **Xin Wang***, Jiangtao Feng, Lei Li, and William Yang Wang. Cross-lingual vision-language navigation. *Submitted to the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020.

[18] Da Zhang, Xiyang Dai, **Xin Wang**, Yuan-Fang Wang, and Larry S Davis. Man: Moment alignment network for natural language moment retrieval via iterative graph adjustment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[19] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2017.