Statement of Purpose: **Yifei Li** http://realliyifei.github.io/

I am open to any research topic related to AI, as long as I can develop my own research skills and taste. Based on my prior experience, my current **research agenda** is motivated by two overarching, complementary goals: to advance intelligence in NLP and multimodal to understand concepts in a more composite, consistent, and hierarchical way; and to align these models to collaborate with humans more robust and fairer. They can be seen as four sweet marriages:

Machine Meets Human. The omnipresence of AI foundation models in human life brings efficiency but also *fairness and safety* issues. It is critical to *align* the behaviors and values between humans and AI; and let them *collaborate* better.

*Past.* Both natural language processing (NLP) and computer vision (CV) models encode and amplify the bias inherited from the training data. As the first author of conceptor debiasing, we use conceptor--a soft projection--to catch bias subspace from large language models (LLMs), e.g. BERT and GPT, then project out the bias by conceptor NOT operation as post-processing. We also propose a new architecture to incorporate bias in all layers through continued training. Besides weights and output, it is more fundamental to debias the input. In my recent research about self-supervised learning, we found that even CV models pretrained on non-human images would lead to gender bias after fine-tuning, due to the implicit confounders. These culprits are then found by tracing the gradients between pretraining and finetuning following the TracIn method. In this process, I notice that current debiasing methods largely suffer from unstable metrics (e.g. inconsistent debiasing benchmarks) and data (e.g. bad attribute wordlists lead to unstable debiasing). Further, the safety issue of LLMs such as discrimination and hallucination is hard to root out by these traditional machine-based methods due to their opaque. Human cooperation and alignment are promising to be the next step to tackle them.

*Future.* I aim to construct new metrics and data for ***stable and general debiasing***. The semantic-level or cultural bias is far less addressed: for example, the red color symbolizes luck in Chinese culture more than in the west, where it might indicate danger. Further, the event-level bias is rarely seen in discussion: for instance, in children's books, it is often that a male hero rescues a female beauty but not the other way. These biases are more implicit and thus harder to quantify and mitigate, and I hope to utilize semantic-, event-, and document-level understanding techniques to tackle them. Beyond fairness, I hope to dive into the ***human-AI alignment*** to inject desired behavior and values into the machine and let them ***collaborate*** *more productively*. For instance, we shall align and calibrate the machine's behavior with human-in-the-loop and ChatGPT unleashes the potential of RLHF. But human involvement is often expensive and may bring undesired phenomena like bias. Also, the alignment in one domain is not guaranteed to generalize to other domains. Therefore, scaling up human oversight more efficiently and robustly is significant for AI safety, and more theoretical and empirical work is needed. Another indispensable part is ***the analysis of LLMs***: e.g. what contributes to in-context learning, chain-of-thought, and model scaling and what are their limitations; a follow-up question is how to make these tricks and models more robust to let general users (not only the researchers) incorporate with them without any catastrophic output.

Concept Meets Hierarchy. We live in a world of concepts; to understand this world, models must dive into the ocean of concepts and understand how they exist and relate in the hierarchy. How are concepts captured and represented? How are they composed to emerge a new concept, disentangled into smaller granularity, or distilled as a meta-concept?

*Past.* Concepts can be captured and manipulated in a logical form. Within my conceptor paper, we capture and negate the bias concept by conceptor NOT operator. Alternatively, two bias concepts (e.g. gender and race) can be conjoined by conceptor AND operator and then removed intersectionally. This project inspires my interest in capturing and (de-)composing concepts logically. For mapping the concepts, one traditional way is to segment concepts by coarse graphs--such as CLVER, Visual Genome, and ConceptNet--as objects, attributes, and relations. However, this discrete and static paradigm cannot perfectly describe the real world, where concepts exist in a continuous, hierarchical, and implicit way; and the concepts are changing constantly.

*Future.* I hope to capture and represent more ***natural and semantic concepts*** along with their relations to construct graphs that are fancier than their traditional counterparts, using the internal knowledge of foundation models in NLP and multimodality (e.g. by templated prompting, or retrieved embeddings). These concepts can be further (de-)composed to challenge and improve the current datasets and models (see the next two sections). Challenges come with great opportunities. On one hand, concepts and their relations are not immutable. It varies temporally (e.g. the specific American president in the knowledge graph changes in different years), procedurally (e.g. the microwave status is different before and after being placed food inside), and culturally (e.g. the consensus about politeness is different among the nations). It is essential to equip AI with a ***dynamic concept*** graph that is adaptive to different z-axises on-the-fly, rather than relying solely on fixed knowledge graphs and pretrained representations. On the other hand, it is fun to play with the geometry of the concept graphs: the distance between concepts can enable zero-shot reasoning and debias; symmetry can

help to improve robustness, consistency, and analysis of embeddings. I am excited to explore these topics using LLM and multimodal representations, perhaps combining neuro-symbolic approaches.

Language Meets Vision. Humans learn the world lying at the intersection of modalities; so should the machine. NLP and CV are born to be complementary: language can ground tremendous details not shown in vision from its rich encoded knowledge; vision can provide the details missed in language due to obviousness or indescribability. I am eager to bridge their gap by fusing the linguistic and visual concepts further and thus ground the downstream tasks better.

*Past.* ***Fine-grained concepts*** are often challenging to multimodal AI. During my independent study of probing CLIP, I figure out that despite its strong zero-shot ability, CLIP performs poorly on the iNaturalist dataset, where CLIP is struggling to map the small details of different creatures' features from descriptions to images. I then try to build a language utility to help users ask the right questions to CLIP: leverage SBERT to cluster different types of descriptions from the web, then compare and rank their CLIP similarity scores. ***The order of information*** is another challenge. It is well-known that CLIP and Dall-E are bad at consistency and composition as they treat prompts as bags-of-wolds (e.g. they cannot distinguish the spatial relation), partly due to the lack of penalizing disordering when training. Worse, diffusion models not only fail to consider all information in prompts in inference but also might mistranslate one word in several ways simultaneously (e.g. `bat` is treated as `stick` and `animal` together), causing concept leakage.

*Future.* All of these phenomena urge us to (1) construct the training datasets and rethink the finetuning methods that take this information into account; (2) redesign the model architecture to glue the order information across modalities more concretely. My target is to create benchmarks like compositional visual question answering to quantify the multimodal performance systematically and to push multimodal AI to understand relations by fine-tuning it on the concepts graphs from the last section with the help of counterfactual data augmentation. Besides, language and vision can help each other. ***Language grounding***, for instance, can improve CLIP zero-shot accuracy on classification. One idea is to concatenate GPT and CLIP alike imagine-then-verbalize: e.g. prompting GPT given keywords by labels and their knowledge graph query results to generate detailed descriptions and then feed them to CLIP; or even co-finetuning them jointly. This can be further boosted via semantic enrichment techniques like (contextualized) retrofitting, and adapted GPT to other tasks involving CLIP and diffusion generation. Vice versa, *vision can help language tasks* by, say, inputting the texts to text-to-image diffusion models and analyzing the distance of the image embeddings.

Connectionism Meets Symbolicism. Neural systems are universal approximators but lack interpretability and robustness, while symbolic systems are explicit and trustworthy but traditionally depend on domain constraints and hand-written rules. Neuro-symbolic systems aim to integrate these two paradigms to gain both benefits.

*Past.* In my research course project, I propose an idea that leverages neuro-symbolic dual system to improve the consistency and coherence in the task-oriented dialog: our team builds user belief states to ground human knowledge and domain-specific constraints, then fine-tune GPT3 to generate utterances and verify the consistency via belief states by a symbolic parser; repeat this process until it is consistent then update the belief states. However, along with many neuro-symbolic implementations, they are heavily limited by handcrafted rules and predefined symbols.

*Future.* My ambition is to ***scale neuro-symbolic approaches***--parsing neural models to symbols, reasoning among them, then updating the result back to its neural representations. We can overcome the limits of fixed symbols and rules by treating language as symbols (e.g. by weak unification) and implementing tasks in generative forms (e.g. prompting). In detail, we shall elicit the symbolic knowledge graph from the neural LLMs by prompting to exploit their encyclopedic knowledge, atomize them by weak unification, then propagate back to LLMs after the graph representations are upgraded by downstream tasks or external knowledge. This paradigm can be extended to multimodality and language groundings, where we can parse and ground other tasks by neural language representations which can be later processed symbolically and iteratively, sometimes incorporating the information from other modalities and knowledge databases.

Career Plan. My long-term career goal is to conduct research as a professor. For teaching, I always love imparting knowledge. As a teaching assistant in courses (ML, DL, and NLP), I advised 15+ AI course projects and led dozens of students in academic discussions. For research, besides exploring the above ideas, my lifelong desideratum is to bring the dream of Artificial General Intelligence (AGI) closer to reality and leverage it to empower humans in application tasks.

Other Projects. Besides the above research, I have been involved in other AI projects in NLP (dialogues, interactive textual games, summarization), CV (product match, medical image segmentation), and ML (hotel prediction).