Research Statement

Generally, my research interests lie at the intersection of **generative models, language modeling, and exploration-driven machine learning algorithms.** Recently, large-scale pre-trained generative models (e.g., LLMs) have demonstrated exceptional versatility and effectiveness across diverse applications, exhibiting superior generalization capabilities than small-scale task-specific models. However, their training lacks a systematic approach to fostering problem-solving skills through exploration and practice, as humans do, which hinders their ability to apply learned knowledge effectively. This deficiency limits the performance of **complex reasoning tasks** and results in the **generation of hallucinated** information. Furthermore, the growing scale of pre-trained models makes their application **on resource-constrained devices less feasible**. Motivated by these observations, I aim to focus my Ph.D. research on (but not limited to) the following themes:

- Developing scalable algorithms for training generative models to explore problem-solving skills (e.g., through **search, preference optimization, and reinforcement learning**).
- Optimizing large generative models for **computational efficiency** in both training and inference.
- Improving the **factuality, trustworthiness, and reliability** of (multimodal) LLMs.

The following sections provide an overview of my research journey, reflections on my experiences, and future plans, including the transition of my research interest from task-specific multimodal understanding to large-scale generative models (LLMs), my motivation for training models through exploration, and my vision for creating efficient and trustworthy LLMs.

1 Task-Specific Multimodal Understanding

Text-video Alignment. My research journey began early in my master's studies at Tsinghua University, where I focused on text-video alignment tasks under the guidance of Prof. **Shao-Lun Huang**. At that time, diffusion models had shown remarkable effectiveness in visual generative tasks due to their iterative denoising process. Motivated by the potential of this multi-step refinement characteristic in understanding tasks, I developed an innovative approach for video grounding and reformulated it as a conditional generative task, using diffusion models to sample segment locations that align with corresponding text input [1]. Benefiting from the iterative sampling during the denoising stage, the decoder progressively refines the location precision, achieving superior performance over existing DETR-like models.

Advised by Prof. **Yansong Tang**, Prof. **Wenbo Ding**, and my supervisor, I also co-led a project to enhance video-text alignment across both spatial and temporal dimensions, proposing an entangled spatio-temporal framework for encoding [2] and decoding multimodal features. Our model effectively mitigated inconsistencies in object predictions from the same event across adjacent video frames.

Multimodal Feature Fusion. For multi-modal understanding tasks, it is essential to align the features from different modalities. Contrastive learning is a common approach for multimodal alignment, which brings the feature representations of positive pairs closer while distancing those of negative pairs. Building on this, we proposed a model-agnostic framework for aligning multimodal features from video, speech, and text and experimented on emotional recognition tasks [3]. Additionally, we introduced an innovative approach of projecting discrete labels into dense embeddings for alignment with the input modalities. Our framework successfully demonstrated compatibility with various base models and improved task performance.

Rethinking the Specialized Understanding Models. Although my earlier work on multimodal understanding showed satisfactory results on specific tasks, these small-scale models lacked adaptability, generalizability, and the capacity for complex reasoning, making them prone to overfitting to training distribution. To pursue general and robust intelligence for real-world applications, I shifted my research focus to large-scale pre-trained (multimodal) language models.

2 Adapting General LLMs for Domain-Specific Tasks

Large language models (LLMs) have achieved significant performance improvements in various applications. However, for domain-specific tasks, general pre-trained LLMs often exhibit subpar

performance. During my internship at **Microsoft**, I focused on how to adapt general pre-trained LLMs to specific target domains, advised by Dr. **Yeyun Gong**. Drawing inspiration from the way humans learn — **acquiring foundational knowledge and applying it through exercises** — we developed a two-stage continual learning approach to enhance LLM adaptation. To support the two-stage training, we curated two distinct training corpora: unlabeled in-domain data for knowledge acquisition and task-oriented passages aimed at problem-solving. These corpora were combined in varying proportions for two-stage continual training [4]. In the first stage, the model was trained on in-domain data to inject domain knowledge, while the second stage focused on task-oriented passages to develop problem-solving skills in the target domain. Through this progressive adaptation, the models achieved improved performance in both few-shot and fine-tuned settings for domain-specific tasks. Additionally, the ablation study demonstrated that both knowledge injection and problem-solving training are essential for effective adaptation, complementing each other in enhancing performance.

Rethinking the Enhancement of Problem-Solving Capabilities. While learning through knowledge injection and problem-solving imitation helped the model address tasks in specific domains, it is usual for us humans to strengthen our problem-solving skills through iterative practice (e.g., doing exercises for preparing an exam). Similarly, for generative models like LLMs, I believe **it is crucial for them to improve their problem-solving skills by exploring their own generations**, receiving feedback from the environment, and refining their responses through extensive exploration and practice. In [5], we initially explored how to enhance LLMs to produce summaries better aligned with relevant queries through feedback from downstream retrieval task performance. In my PhD journey, I am eager to delve deeper into advancing LLM problem-solving capabilities through exploration and interaction with external feedback.

3 Towards Efficient Training and Inference for LLMs

The emergence of ChatGPT from OpenAI has attracted the attention of many researchers and internet users, sparking my interest in this field. However, due to the $O(n^2)$ complexity of the attention mechanism in standard Transformers, token limits remain a bottleneck in initial models. To address this, my colleagues and I focused on reducing the number of memorized tokens used during the decoding stage to help alleviate this limitation. Unlike most token selection strategies focusing on attention scores or feature representations, we evaluate token effectiveness based on its contribution to task performance, **employing reinforcement learning by training a token selector with a reward function that combines task performance with a length penalty for selected tokens** [6]. During inference, the trained token selector filters out irrelevant tokens, effectively reducing the computational burden during decoding.

Rethinking Efficiency Improvement for LLMs. While general pre-trained LLMs have demonstrated remarkable capabilities across various applications, their massive parameter size and computation cost hinder practical utilization. Recent work, such as DeepSeek-Math, has shown that smaller domain-specific models can match or surpass larger models in specialized tasks. From an architectural perspective, memory-efficient models like Mamba and its hybrid variant have also demonstrated promise in both efficiency and performance. In my PhD studies, I am delighted to explore the possibility of: (1) Developing **efficient and scalable generative foundation models**; (2) Designing memory- and computation-efficient pipelines for LLMs, including **quantization, parameter-efficient training, KV-cache optimization, and long-sequence modeling.**

4 Reducing Hallucinations and Building Trustworthy LLMs

LLMs sometimes inevitably produce content that deviates from the input or generated contexts, or contradicts established world knowledge—a phenomenon known as hallucination. To address this issue, my labmate and I proposed **a self-consistency-based decoding strategy.** This approach involves sampling multiple initial responses to a prompt, separately concatenating them with the prompt, and forming a batch. The final response is generated during the second decoding phase, where the next token is selected by aggregating predictions from all responses in the batch [7]. Our method not only enhanced the factuality of LLMs but also pushed the boundaries of self-consistency in open-ended generation tasks. **Beyond language modeling**, I also contributed to research on balancing hallucination and descriptiveness in multimodal large language models (MLLMs) by leveraging

contrastive decoding, which utilizes both descriptive intermediate hidden states and outputs from the last layer to generate the final response [8].

Rethinking on Building Reliable LLMs. Despite being pre-trained on vast amounts of web corpus and synthetic data to acquire world knowledge, it remains challenging for LLMs to ensure all their outputs align with factuality. In my previous research, we initially investigated how to better utilize their 'parameter knowledge' to mitigate hallucinations in generated answers. Taking a further step, I wish to explore the **root causes behind hallucinated generations from LLMs** and how they ensure the accuracy of their outputs. Building on these insights, I plan to design **more effective and targeted frameworks for evaluating factuality** in LLMs and mitigating hallucinations.

References

[1] **Xiao Liang***, Tao Shi*, Yaoyuan Liang, Te Tao, and Shao-Luo Huang. Exploring iterative refinement with diffusion models for video grounding. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2024.

[2] Yaoyuan Liang*, **Xiao Liang***, Yansong Tang, Zhao Yang, Ziran Li, Jingang Wang, Wenbo Ding, and Shao-Lun Huang. Costa: End-to-end comprehensive space-time entanglement for spatio-temporal video grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3324–3332, 2024.

[3] Tao Shi*, **Xiao Liang***, Yaoyuan Liang, Xinyi Tong, and Shao-Lun Huang. Sslcl: An efficient model-agnostic supervised contrastive learning framework for emotion recognition in conversations. *arXiv preprint arXiv:2310.16676*, 2023.

[4] **Xiao Liang***, Xinyu Hu*, Simiao Zuo, Yeyun Gong, Qiang Lou, Yi Liu, Shao-Lun Huang, and Jian Jiao. Task oriented in-domain data augmentation. *EMNLP*, 2024.

[5] **Xiao Liang**, Xinyu Hu, Simiao Zuo, Jimi He, Yu Wang, Victor Ye Dong, Yeyun Gong, Kushal S Dave, Yi Liu, Qiang Lou, et al. What you see is what you get: Entity-aware summarization for reliable sponsored search. In *Neurips Safe Generative AI Workshop 2024*.

[6] Jiawen Xie, Pengyu Cheng, **Xiao Liang**, Yong Dai, and Nan Du. Chunk, align, select: A simple long-sequence processing method for transformers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 13500–13519, 2024.

[7] Yi Cheng, **Xiao Liang**, Yeyun Gong, Wen Xiao, Song Wang, Yuji Zhang, Wenjun Hou, Kaishuai Xu, Wenge Liu, Wenjie Li, et al. Integrative decoding: Improve factuality via implicit self-consistency. *arXiv preprint arXiv:2410.01556*, 2024.

[8] Yaoyuan Liang, Zhuojun Cai, Jian Xu, Guanbo Huang, Yiran Wang, **Xiao Liang**, Jiahao Liu, Ziran Li, Jingang Wang, and Huang Shao-Lun. Unleashing region understanding in intermediate layers for mllm-based referring expression generation. *Advances in Neural Information Processing Systems*, 2024.