Statement of Purpose

Ziteng Wang MS/PhD in Computer Science

I stepped into deep learning 7 years ago by building neural networks for hand-written figures classification, following Andrew Ng's videos on YouTube. My excitement about how deep learning connects computation and diverse aspects of our world continues to grow as I witness the potential of large language and vision models to serve as a unified engine powering machine intelligence in our complex and heterogeneous world. My research interest aims to fully realize the potential that Multimodal Large Language Models see more clearly and think more precisely. This target might be realized by offering MLLMs (1)**a pair of penetrating eyes** and (2) **a clear mind**.

Fine-grained visual features extraction Vision Transformer (ViT) is the eye of MLLM, as it converts images into visual patch tokens before the language model can take them as input. How models can compress billions of pixels into input tokens seems immensely beautiful to me. However, this process is not always perfectly done. I am drawn to uncover why CLIP fails in distinguishing between some similarly looking pairs [1]. This can be attributed to (1) limitted text input length to describe small details in an image during contrastive learning and (2) limitted source of very similar image pairs in CLIP training. In CLIP-IN [2], I solved these problems by introducing a symmetric hard negative loss, as well as instruction based image editing data (they are by default very similar!) into basic CLIP training.

Learning and thinking with what MLLMs have seen Even if we have solved the problem for visual representation (probably only a fantasy) - we have a perfect vision encoder that can capture every details of the input image. Now the problem is, the MLLM that takes visual and text tokens as inputs may ignore our perfect visual tokens and output the answer solely rely on the knowledge they have learned during pretraining, or only the input text token. This makes our ViT becomes useless. In Thinking-Pilot [3], I used a reward-driven post-training pipeline to guide the MLLM focusing more on visual information real-time input knowledge. This work only focused on spatial reasoning, a highly vision-centric domain. I believe this idea can be expanded generally.

Besides research experience, I have also exposed to industrial level projects.

Mixture of Data through Training I started investigating the impact of different combination of mixture of data when I was an intern at **Meituan** and engaged in the SFT stage of LongCat-Flash-Omni [4], an industry level full-modality large language model that trained with 2000+ NVIDIA H100s for more than three days. My goal is to find out the optimal data ratio from different sources (i.e charts, documents, OCRs .etc) and different modalities (i.e images and audio)for SFT. I started with the most intuitive approach: run the training with different data mixtures for couple of hours and determined the validation according to the loss curves and log output. However, the enormous scale of today's data prevents abundant experiments. I was excited to find recent works of clever methods for quickly finding good data mixtures such as averaging model weights and truncating training runs. These observations led me to identify the core component of each work and build a unified benchmark of data mixture estimators. This process involves large-scale model training with many datasets and model configurations. I iterated through multiple systems for creating and managing experiments and learned valuable skills in building versatile yet easy-to-use model training infrastructures.

In conclusion, my previous experience focused on multimodal learning and vision language alignment. My long-term research interests (not limited) are listed below.

Further explorations in vision-language alignment I believe that there still exists some topics under explored in the field of vision-language alignment. Vision signals are continuous while text tokens are discrete, which might be the high level abstraction of why alignment is difficult. Can we direactly align the type of signals in representation learning?

Unified framework for multimodal understanding and generation Traditionally, an understanding model uses a ViT as vision encoder while a generation model uses DiT/VAE for visual representations. I am interested in exploring whether these two tasks can share a same latent space and benefit from each other. For example, some reasoning tasks (i.e geometry math reasoning) in modern MLLMs benefit from generating additional images as part of CoT. These images are usually generated by an external agent, which has no correspondence to the understanding model itself. I believe the reasoning ability can be enhanced if we use an affinal model in this process.

Vision language actions While I think that MLLM understanding should not be restricted to fragmented images and video, instead, it should be able to perceive and interact with the physical world. This can be a huge topic, as moving from perception to action requires more than recognition - MLLMs need to understand spatial structure and leverage past experience. We can handled this by studying spatial intelligence and memory-driven action planning.

References

[1] Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9568–9578, 2024. URL https://api.semanticscholar.org/CorpusID:266976992.

[2] Ziteng Wang, Siqi Yang, Limeng Qiao, and Lin Ma. Vitrix-clipin: Enhancing fine-grained visual understanding in clip via instruction editing data and long captions. In *NeurIPS*, 2025. URL https://api.semanticscholar.org/CorpusID:280422347.

[3] Ziteng Wang, Jiancheng Huang, Yu Gao, and Liang Xiang. From what to where: Reward-driven post-training for spatially-aware mllms. In *Submitted to CVPR*, 2026.

[4] Meituan LongCat Team. Longcat-flash-omni technical report. *ArXiv*, abs/2511.00279, 2025. URL https://api.semanticscholar.org/CorpusID:282740018.

* denotes equal contribution.