Research Statement — Zhili Yang

Application to UCSD Hao AI Lab

I am currently a first year graduate student majoring in Computer Engineering at UCSD, with research interests centered around large language model efficient fine-tuning, model interpretability, and the emerging frontier of multimodal generation—particularly video diffusion models. Recently, I have been deeply drawn to how temporal coherence, spatial attention, and generative consistency can be optimized in video models, and I have read extensively on the latest advances such as DiT, Sparse VideoGen, Progressive Distillation & Consistency Models, and FastVideo. My goal is to contribute to bridging efficiency, fidelity, and interpretability in generative systems that extend beyond text into dynamic visual modalities.

My undergraduation project investigates improved LoRA variants across multi-domain benchmarks, incorporating visualization of parameter distributions and attention maps to better understand model adaptation behavior. This work has shaped the way I think about high-dimensional model dynamics and efficient adaptation — concepts that naturally extend to video generation, where computational cost and temporal consistency remain key bottlenecks. I have also implemented and reproduced various diffusion-based frameworks and re-engineered components from models like Stable Diffusion, which provided me a practical understanding of denoising processes and cross-attention mechanisms that underlie both image and video synthesis.

What attracts me most to the Hao AI Lab is its pragmatic and open-minded research philosophy. The lab's focus on scalable, efficient, and accessible machine learning systems — exemplified by works like FastVideo — resonates with my belief that the future of generative AI depends not only on better models, but on making those models faster, interpretable, and easier to deploy. I see a clear opportunity to extend my existing understanding of attention visualization and efficient model tuning to this context: analyzing how temporal attention shifts across frames, how model compression influences generative smoothness, and how LoRA-style adaptation or quantization might further reduce computational overhead without compromising perceptual quality.

If given the opportunity to join the lab, my initial month would focus on an more in-depth survey of recent multimodal and video diffusion research to identify where efficiency and quality can most effectively intersect. Meanwhile I would contribute to reproducing and profiling FastVideo's existing pipeline, followed by developing small-scale prototypes to explore temporal attention analysis and lightweight fine-tuning strategies. In the longer term(after this quarter), I aim to propose an independent sub-project on interpretable acceleration for multimodal generation, integrating both my visualization background and the lab's expertise in large-scale system design.

Joining the Hao AI Lab would allow me to deepen my expertise in multimodal generation and system-level optimization, and to work in a team that values both theoretical insight and real-world impact. I aspire to become a researcher who builds models that others can understand, use, and extend.