STATEMENT OF PURPOSE

Youzhe Song
Homepage
Computer Science PhD Applicant, Fall 2026

I. INTRODUCTION: MY RESEARCH INTERESTS AND PERSPECTIVE

Current AI models have achieved remarkable success across tasks, yet their internal mechanisms remain largely a "black box." We know what models can do, and to some extent how they do it in specific details, but we lack an intuitively human-understandable view of their cognitive and reasoning processes. I believe that a key step toward more general and reliable AI lies in understanding and designing the model's internal information-processing mechanisms. My research interests center on two core questions: First, how can we enable the model to **understand and align information from different modals** (e.g., images and text) in a human-like manner? Second, can we construct within the model a **human-intuitive, stepwise, and organic reasoning trajectory**?

My goal is to explore how to realize the fusion of these two abilities in a high-dimensional latent space, enabling models not only to "see" and "read" the world, but also to reason about it in a transparent and trustworthy way.

II. ROBUST FEATURE REPRESENTATIONS (COREFACE) [1]

My research journey began with a deep fascination for self-supervised contrastive learning. It does not rely on expensive human annotations; instead, it learns from the structure inherent in data. Its elegance and potential captivated me. However, the realities of **my lab with very limited computing resource and diverse student directions** pushed me to give it up and pursue a more inexpensive alternative.

I found face recognition to be an ideal "sandbox." I observed its deep kinship with contrastive learning: **both carefully shape the feature distributions in high-dimensional space** and share the same goal. Building on this, I proposed the CoReFace framework, which introduces contrastive learning into classification to achieve robust feature representations.

In fine-grained recognition tasks such as face recognition, traditional image enhancement methods tend to destroy critical identity information. Since addressing the issue at the image level was not feasible, **I shifted my focus to exploring feature-level processing approaches.** I employed Dropout, which is essentially a random perturbation at the feature level; it can construct feature pairs for contrastive learning without undermining the semantic information of images.

Furthermore, I identified and addressed compatibility issues in the joint training process of classification and contrastive learning. These issues included two key aspects: **the supervision signal in contrastive learning being too weak to generate effective optimization, and the problem of semantic duplicate calculation in some sample pairs.** To tackle these, I conducted in-depth optimization of the contrastive learning loss function by introducing a dynamic similarity constraint, which ensured the maintenance of effective training gradients throughout the entire training process. For the issue of semantic duplication in sample pairs, I opted to restructure the scheme for constructing sample pairs used in contrastive learning.

By introducing a regularization term guided by contrastive learning, I proactively "arranged" the geometric structure of the feature space. Ultimately, this method successfully **increased the similarity margin between positive and negative pairs by 15%**. In this work, I took the lead in overseeing the entire lifecycle, spanning from background research and scheme design to independent verification and paper writing. This experience enabled me to achieve a cognitive transition in my identity, moving from a "student" to a "researcher".

III. A UNIFIED FRAMEWORK FOR HETEROGENEOUS DATA (QGFACE) [2]

After completing CoReFace, I was eager to conduct research that would establish a more direct connection with the real world. **This idea was inspired while I was browsing through my phone's photo album**: the built-in AI face clustering feature didn't perform well when processing my family photos, often missing pictures of certain individuals. I realized that **low-quality data** in real-world scenarios, resulting from factors like shooting distance, composition, and lighting, **differs significantly from the datasets commonly used in general face recognition**.

Hoping to achieve unified processing of data with varying quality **through an elegant and efficient method**, I initiated the QGFace project. It was developed based on a single-encoder architecture, eliminating redundant components such as super-resolution modules and teacher-student networks used in existing solutions. **Clear images facilitate accurate person identification, while extremely blurry or occluded images make it nearly impossible to determine the corresponding identity**. Based on this assumption, I applied different supervision signals to encoded features of varying quality: classification loss for high-quality data and contrastive loss for low-quality data.

During the process, I discovered that contrastive learning underperformed due to insufficient sample pairs, prompting me to design a real-time encoding queue. **Unlike the momentum encoding queue commonly used in contrastive learning**[3], its features do not come from momentum encoders but from the training encoder. Meanwhile, these encoding results are updated using the differences between classification vectors at different update steps.

To meet the submission deadline, I immersed myself fully in the work. During that period, **my day was divided into multiple work sessions, and GPU processing time became my only breaks**. This experience was not a heavy burden for me, nor did I fall into a state of depression, because **I knew I was creating something valuable**: it convinced me that my motivation stemmed not from external expectations, but from an intrinsic drive and joy in solving difficult problems.

Ultimately, QGFace achieved SOTA performance on low-quality datasets while barely sacrificing performance on high-quality data (with only a minimal 0.3%/95.5% performance trade-off), **demonstrating that my research philosophy can guide the development of practical solutions that are both elegant and robust**. Additionally, this work represents **an exploration of "unimodal" data content**. I have always believed that current methods still have room for improvement in exploring the modality itself. For instance, models tend to prioritize textures over shapes, unlike human perception which relies more on shapes. Exploring internal alignment within unimodal data will be a focus of my future work.

IV. EXPLORATION: CLARIFYING MY DIRECTION THROUGH WORKING PRACTICE

Despite achieving some academic results, repeated setbacks and uncertainties in the submission process plunged me into a profound period of confusion. I began to strongly doubt myself: **Was I truly suited for an academic career?** In search of answers, I made a decision: to proactively step into industry and conduct a thorough "exploration" of my true inner passions.

I first joined KeyoneAI, a startup led by Mr. Fang Jie, former Chief AI Architect of IBM China. There, **I worked on implementing cutting-edge generative AI technologies into products**, experiencing the impact of rapid iterations, collaborating with teams, engaging with potential users, and acting as a technology evangelist. However, I found it rather tormenting to be so close to technology yet caught in the trivil work of product refinement, while feeling distanced from intellectual innovation. Thus, I eventually left.

Seeking to understand how technology empowers various industries, I joined an organization further from the source of AI, the organizing committee of the Worldwide Educators Conference (WWEC)[4]. As the sole technical lead, I leveraged technology to enable **an international conference attended by tens of thousands of people**. Over nearly three months of working 80 hours a week, I ensured system functionality while promoting team digitalization, developing internal tools to support the generation of **over 1,000 complex posters, and connecting more than 3,000 attendees, 1,00 exhibitors, 10+ vendors**, and managing various ad-hoc meetings. I enjoyed the AI coding process to build more functionality for our team. I would like to build an academic automation system in the future.

While these experiences brought significant challenges and a sense of accomplishment afterward, I realized none could replace the pure, intellectual excitement and flow I felt during research: witnessing how researchers make breakthroughs in certain technical directions, sometimes several times a year or after years of effort, and being part of such breakthroughs myself. This period of "deliberate detachment" brought unprecedented clarity: **my deepest desire lies in questioning the underlying principles of things, refuting and reconstructing existing solutions**, and building innovations that draw on diverse sources and are truly effective.

More importantly, my work experiences helped transform what once troubled me. That is excessive sensitivity, self-examination, and "internal friction"into valuable life accumulations. Throughout my work, I constantly felt a certain void within; although my contributions were indispensable to the team, my inner creative passion remained unexpressed. To balance my mind, I turned to outdoor activities and read books on mind-body-spirit wellness. Among all this, thoughts about achieving intelligence kept emerging in my mind. I now firmly believe that **the hallmark of a truly intelligent system should not be flawless one-way reasoning, but precisely the ability to handle internal conflicts, engage in self-examination, and iteratively correct itself**. This is the essence of human "reflection" and "hesitation," and the foundation of complex reasoning and intelligence.

V. FUTURE: ALIGN SEMANTICS & REASON IN LATENT SPACE

Now, I am planning my doctoral research with a clearer and more determined goal. I aim to combine my past experience in **feature space optimization and heterogeneous data processing** with my reflections on human cognitive processes, focusing on exploring two core capabilities of large models:

1. **Constructing a Unified Semantic Space:** My primary objective is to investigate how to effectively map various types of information into a unified latent space with interpretable semantics. **This not only emphasizes cross-modal mapping but also involves the exploration of unimodal data content.** I believe that the semantic alignment is the first step in building a foundational model capable of comprehensively understanding the world. My experience in handling heterogeneous data during the QGFace project has provided me with a valuable practical basis for exploring ways to align and fuse heterogeneous information.
2. **Enabling Structured Reasoning Processes in the Latent Space:** In the process of achieving semantic alignment, my another goal is to design and implement the model's autonomous, structured reasoning paths within the latent space. **Existing models are similar to early computers that used paper tapes for input and output**, lacking a unified "brain" and a coherent thinking process. I hope the **model can not only generate answers but also demonstrate a decomposable and traceable reasoning process.** This will not only enhance the model's interpretability and reliability but may also offer new approaches to solving more complex, open-ended problems that require multi-step logic.

In summary, my research plan progresses in two parallel directions: on one hand, enabling the model to "**perceive**" a richer world through multi-modal alignment; on the other hand, allowing the model to "**acheive**" more clearly and logically through latent reasoning. Most published works start by exploring **how to extend existing models**; by contrast, my research begins with clarifying **what the model needs to achieve**.

VI. CONCLUSION

With hands-on experience designing internal mechanisms and a clear plan to integrate *semantic alignment* with *latent reasoning*, I am prepared for the challenges of a Ph.D. I look forward to contributing to the next generation of more capable and trustworthy AI systems in a creative and supportive environment.

REFERENCES

[1] **Youzhe Song** and Feng Wang. Coreface: Sample-guided contrastive regularization for deep face recognition. *Pattern Recognition*, 152:110483, 2024.

[2] **Youzhe Song** and Feng Wang. Quality-guided joint training for mixed-quality face recognition. In *2024 IEEE International Conference on Automatic Face and Gesture Recognition (FG)*. IEEE, 2024.

[3] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9726–9735, 2019.

[4] Worldwide educators conference (wwec). `https://www.wwec820.com/`.