Research Statement

Weijun Wang

AI has scaled from DNN to foundation models to its next frontier: *embodied intelligence*, physical agents that perceive, reason, and act in the physical world. Model capability is only half the story. The system stack underneath must catch up, or capability gains never reach a robot.

As a AI systems researcher, my goal is to build the training and inference systems that let models scale effectively and runs efficiently. My research tracks AI's evolution at the system level: efficient DNN inference system at the edge in the *pre-LLM era*; effective LLM serving systems on cloud for the *LLM era*; and training and inference frameworks for physical agents in the *embodied era*.

Impact. Owing to my intrigue for *real-world* systems, my research has received attention widely. Some of my research has been used in production of ByteDance, Huawei, AsiaInfo, etc., serving millions of users. I'm also honored to be selected for the National Postdoctoral Talent Introduction Program, Outstanding Postdoctoral Fellow Runner-Up (top 20/3500 at Tsinghua), the Shuimu Tsinghua Scholar, and several paper awards[1] from the community.

AI Systems for the Pre-LLM Era

Before LLMs, the dominant AI workload was DNN-based perception, video analytics, real-time monitoring, etc. The biggest challenge was the mismatch between *lantency and accuracy* in SLOs.

Models shold pay more attention on accuracy-aware data. Conventional visual applications shipped opaque pixel streams to the edge or cloud. However, *inference accuracy gain is non-uniform across frames and within a frame*. Only a few frame, even the partial content within frame, matter accuracy. AccDecoder[2] and RegenHance[3] reframe video as a non-uniform information signal, reconstructing regions of interest at higher fidelity at the decoder. BiSwift[4] further extends this non-uniform structure to multiple streams. These work has been received well in the community[5] and adopted by AsiaInfo in production[6].

AI Systems for the LLM Era

At earlier of LLM era, systems inherit a serving stack tuned for compute, such as FlashAttention. But along the model architecture tend to converge, the dominant cost shifts from compute to *storage and network*. Serving system must manage the state of LLM, e.g., LoRA adapters or experts in MoE (model parameters), request prefixes (user inputs), KV caches (inference intermediate data), etc. Tranditional work hide these into inference engines, invisible to the scheduler. I make them first-class and build systems explicitly manage them.

Model management. LLMs are powerful but domain-blind. LoRA adapters offers good choices to ingest domain knowledge into LLMs; the mixture-of-experts (MoE) architecture offers a way to scale model capacity without increasing inference cost. I designed systems to efficient manage these model states explicitly. ValoRA[7] treats LoRA adapters as *batched and orchestrated entities*, enabling heterogeneous adapters concurrency; SwapMoE[8] makes experts swappable, locality-aware entities for storage-constrained GPU on edge, while SMoE[9] further turns expert importance into a scheduling signal.

[1] ACM MobiSys 2026 Best Paper Award Runner-Up; IEEE ICPADS 2025 Best Student Paper Award; IEEE CBD 2022 Best Paper Award.

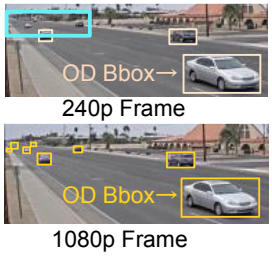

Figure 1: Inference accuracy is non-uniform within one frame; e.g., the blue boxes part needs pay more attention. This non-uniform data distribution also occurs across streams and frames.

[2] T. Yuan, L. Mi, **W. Wang**, H. Dai, X. Fu. *AccDecoder: Accelerated Decoding for Neural-enhanced Video Analytics*. IEEE INFOCOM 2023.

[3] **W. Wang**, L. Mi, S. Cen, H. Dai, et al. *Region-based Content Enhancement for Efficient Video Analytics at the Edge*. USENIX NSDI 2025.

[4] L. Sun†, **W. Wang**†, T. Yuan, L. Mi, et al. *BiSwift: Bandwidth Orchestrator for Multi-Stream Video Analytics on Edge*. IEEE INFOCOM 2024.

[5] Invited CCF Talk, "DNN and LMM Empowering Video Analytics Systems", online, 03.2025, with 4,000+ simultaneous watchers.

[6] RegenHance has been adopted by AsiaInfo's Edge AI product line with 3M RMB order in 9 months.

[7] L. Mi†, **W. Wang**†*, W. Tu, et al. *Empower Vision Applications with LoRA LMM*. ACM EuroSys 2025.

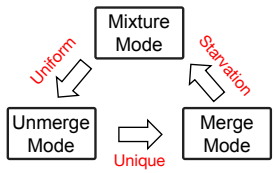

Figure 2: ValoRA explicitly manages LoRA adapters and switches among three inference modes. It reduces latency by 20–89% over Punica and S-lora and has been deployed in AsiaInfo's Edge FM product with 5.8M RMB order.

[8] R. Kong, Y. Li, Q. Feng, **W. Wang**, et al. *SwapMoE: Serving Off-the-shelf MoE-based Large Language Models with Tunable Memory Budget.*. ACL 2024.

Intermediates management. Long-context inference carries a hidden tax: prefill recomputation across users that share contexts. KVCodec[10] makes a counter-intuitive observation: *KV caches look like videos*, high-dimensional, spatially correlated, redundant. It leverage idel GPU-native video codec chips to compress/de- and migrate KV caches among hosts. Tokens, especially visual tokens of high-defined images in VLM, can also be managed and compressed. I propose a RL-based agent to do so during VLM inference (under review by NeurIPS).

AI Systems for the Embodied Era — Future Work

Efficient RL training framework

In the embodied era, RL post-training fuses simulation/real-machine operations, generation (inference), and policy update (training) three stages into one closed loop. My future research targets on building efficient training and inference systems for embodied AI.

Asynchronous Embodied RL. Current embodied RL systems synchronize simulation, generation, and policy update at coarse batch boundaries, leaving CPUs and GPUs idle when any stage stalls. EmbodiRL[11] addresses this inefficiency with asynchronous stages execution. It pools CPU cores, GPU SMs, and memory, then schedules rollout, generation, and training without strict stage-level barriers. This design keeps heterogeneous resources busy, substantially improves resource utilization.

Disaggregated Embodied RL. Beyond asynchronous execution, I will explore disaggregated embodied RL systems that separate simulation from policy generation. CPU-side simulation, environment stepping, and rendering often cannot keep pace with GPU-side generation, making rollout throughput simulation-bound and leaving GPU capacity underutilized. One design runs closed-source simulators on CPU clusters and generators on GPU clusters, linked by efficient observation transport. The other targets open-source stacks by splitting CPU physics from GPU rendering and synchronizing compact physical state. Together, these designs improve rollout throughput across closed- and open-source simulation infrastructures.

Efficient embodied model inference runtime

Embodied models will become a key workload on edge and IoT devices, where agents must perceive, reason, reflect, and act under tight latency and energy budgets. Running them efficiently is difficult for two reasons. First, their architectures have not converged. Visual language action models, world action models, and openclaw-based embodied agents expose different operator mixes, state layouts, and control cycles, so mature high-performance kernels and runtime frameworks are still missing for NPUs and embedded GPUs. Second, their dataflow is not unified. Many models reason, reflect, and predict over language, vision, even latent representations, making inference stateful, multimodal, and hard to batch, cache, or schedule. I have built an unified KV cache management system for GPU memory constrained VLA inference[12], and a skill-aware self-reflection scheme for self-improving embodied agents[13] . My future work will build efficient embodied model inference runtime for diverse workloads and hardware platforms.

[9] G. Zhu, M. Li, H. Dai, X. Liu, **W. Wang**, et al. *SMoE: An Algorithm-System Co-Design for Pushing MoE to the Edge via Expert Substitution*. IEEE/ACM ISCA 2026.

[10] L. Mi†, **W. Wang**†, J. Chen, T. Cao, H. Dai, Y. Liu. *Efficient Remote KV Cache Reuse with GPU-native Video Codec*. ACM SIGCOMM 2026.

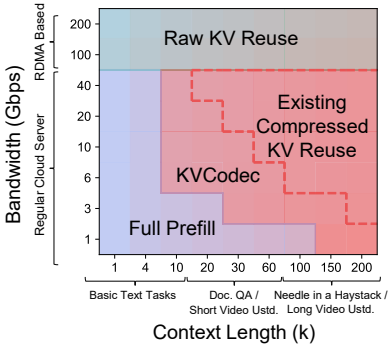

Figure 3: KVCodec significantly enlarge the scope of prefix cache application, by viewing KV cache as video and encoding them with GPU-native video codec ASICs. KVCodec is under deployment by ByteDance HPC platform and Huawei MindSpore product.

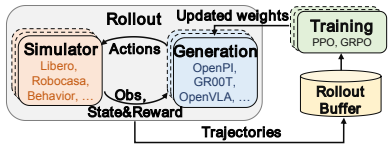

Figure 4: Embodied RL training interleaves simulation, generation, and training with highly diverse resource requirements.

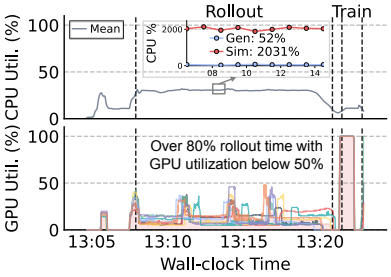

Figure 5: Strict synchronization between rollout and training, and also the simulating step and generation step significantly limits the efficiency.

[11] **W. Wang**. *Asynchronous Embodied RL Training via Fine-Grained Pooling*. In submission to ASPLOS 2026.

[12] X. Li, H. Tang, X. Ding, **W. Wang**, et al. *OxyGen: Unified KV Cache Management for VLA Inference*. In submission to NeurIPS 2026.

[13] R. Ju, X. Wang, X. Ding, **W. Wang**, et al. *EmbodiSkill: Skill-Aware Reflection for Self-Evolving Embodied Agents*. In submission to NeurIPS 2026