Statement of Purpose

Wanjiang Weng

From video generation to world models. Video models can now learn interactive environments from visual experience alone, as Genie demonstrates [1]. Yet a model that renders plausible frames is not yet a world model: it must represent the structure of how things move, respond faithfully to what it is asked to do, and produce futures an agent can act on. My doctoral research asks how generative video models can acquire these three properties: *structure*, *alignment*, and *actionability*. I have approached this question through human motion, where structure and intent can be examined explicitly, and through video generation in industry. I completed my master's degree in Computer Technology at Southeast University under Prof. Hongsong Wang.

Representation learning for human motion. My research began with a basic question: what should a representation of motion capture? Existing self-supervised methods tended to learn either global, sequence-level features or local, frame-level ones, which limited what the representation could support. In USDRL [2], I proposed decorrelating features at multiple granularities so that local and global motion structure are learned in one dense representation; its extension into a skeleton foundation model [3] tested this idea across a wide range of action-understanding tasks. This work convinced me that explicit structure is what allows a learned model of dynamics to transfer.

Aligning motion generation with instructions. I then turned from understanding motion to generating it, and found that realistic motion is often unfaithful to its instruction. Correcting this with rewards raised two questions. At which denoising steps should a reward intervene? ReAlign [4] answered with a step-aware reward that guides sampling without retraining the generator. Why is reward fine-tuning so costly? In EasyTune [5], we traced the cost to recursive gradient dependence across denoising steps and removed it with step-wise optimization. MotionRFT [6] extended these ideas to rewards shared across heterogeneous motion representations. Together, these projects taught me that alignment is fundamentally a problem of credit assignment along the generative process.

Unifying motion and language, and moving to video. Generation and understanding of motion are usually treated as separate problems. In BiMoGen [7], I asked whether one model could translate in both directions between motion and language, using a unified masked discrete diffusion model in which captioning and generation share a single cross-modal correspondence. This suggested that structured motion can serve as a common interface between language, video, and action. At Shopee, I now work on a video generation foundation model, where I lead its enhancement stage; my current research there, MendVE, asks how much text conditioning a video enhancer actually needs. Working in pixel space has shown me that controllability and consistency do not emerge from scale alone.

Proposed Research. Building on these experiences, I plan to study three connected questions.

(1) *Video and motion models as mutual priors.* Video generation models implicitly encode how people move, interact with objects, and appear from different viewpoints, whereas motion models offer explicit, physically interpretable structure but are limited by scarce motion-capture data [8]. Can each serve as a prior for the other? From video to motion, I would distill human dynamics from video models into motion models, combining 4D human recovery with score distillation in the spirit of DreamFusion. From motion to video, I would extend pose- and camera-conditioned generation so that full-body motion becomes an explicit control signal for consistency across frames and views.

(2) *Language models as motion planners.* Large language models (LLMs) such as GPT and Claude can write motion directly from a description, as keyframes, joint trajectories, or code, much as they write robot programs and reward functions; motion itself can be modeled as a language. Yet LLM-written motion is often physically implausible and temporally coarse. I would treat it as a coarse plan that a motion model refines and a video model renders, and conversely use motion and video models as critics whose feedback improves the motion that LLMs write.

(3) *Credit assignment from denoising to action.* Rewards for motion and video are typically one score per sequence, while errors arise in particular frames and denoising steps. Extending ReAlign and EasyTune alongside reward fine-tuning and preference optimization for diffusion models [9, 10], I would study how to distribute semantic, geometric, and task rewards across time and the denoising trajectory, using held-out evaluators and human judgments to guard against reward exploitation. Ultimately, I would test whether futures aligned in this way improve policy learning, building on video-based policies such as UniPi [11] and beginning in simulation.

Research fit. I hope to join a group where generative modeling meets vision, graphics, language, or robot learning, and where progress is judged by careful evaluation as much as by benchmark gains. I would bring experience in motion representation, reward-based post-training, and large-scale video generation, and hope to learn from advisors whose expertise in geometry, language, or embodied learning complements mine.

Long-term goal. I aim to lead research on generative world models in an industrial AI lab.

References

[1] J. Bruce et al. Genie: Generative interactive environments. In *International Conference on Machine Learning (ICML)*, 2024.

[2] W. Weng et al. USDRL: Unified skeleton-based dense representation learning with multi-grained feature decorrelation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2025.

[3] H. Wang, W. Weng, et al. Foundation model for skeleton-based human action understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 48(1), 2026.

[4] W. Weng et al. ReAlign: Text-to-motion generation via step-aware reward-guided alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2026.

[5] X. Tan, W. Weng, H. Lei, and H. Wang. EasyTune: Efficient step-aware fine-tuning for diffusion-based motion generation. In *International Conference on Learning Representations (ICLR)*, 2026.

[6] X. Tan, W. Weng, et al. MotionRFT: Unified reinforcement fine-tuning for text-to-motion generation. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2026. To appear.

[7] W. Weng, Y. Wu, et al. BiMoGen: Bidirectional motion-text generation via unified masked discrete diffusion. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2026. To appear.

[8] G. Tevet et al. Human motion diffusion model. In *International Conference on Learning Representations (ICLR)*, 2023.

[9] K. Black et al. Training diffusion models with reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2024.

[10] K. Clark, P. Vicol, K. Swersky, and D. J. Fleet. Directly fine-tuning diffusion models on differentiable rewards. In *International Conference on Learning Representations (ICLR)*, 2024.

[11] Y. Du et al. Learning universal policies via text-guided video generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.