DOTBot: Diffusion-Aligned Chain-of-Thought Bipedal Humanoid Robot for Interpretable Real-Time Autonomous Navigation, Hazard Mitigation, and Recovery in High-Risk Built Environments

Srikumar, Karthik (School: South Windsor High School)

Autonomous robotic systems and the vision-language models powering them share a critical failure: both produce actions through opaque latent embeddings with no causal justification, thereby preventing auditing. This creates a technical challenge by amplifying errors in obscured scenes, while introducing ethical risks of deploying unverifiable systems. To overcome these challenges, this study presents DOTBot, a plug-and-play framework enhancing any VLM backbone through three innovations: (1) Diffusion Perceptual Reconstructor: a DDPM-based module that recovers visual fidelity from smoke- and debris-corrupted frames by modeling the full pixel-level distribution rather than compressing to fixed embeddings; (2) Mutual Information Alignment Module, which formally couples perceptual latents to structured reasoning tokens via noise-contrastive MI maximization, enforcing causal coherence between observation and action; and (3) Structured Chain-of-Thought Head, an autoregressive decoder generating 3 to 6 typed intermediate reasoning steps before action commitment, eliminating opaque end-to-end prediction. DOTBot achieves SOTA across all evaluated benchmarks: 74.1% on EgoSchema (+7.7pp over RT-2), 75.8% on NExT-QA (+2.3pp over VideoTree), and 42.3% task success on ALFRED (+3.1pp over RT-2), with 87.4% CoT faithfulness, a capability absent from all baselines. The framework is validated on a physically engineered bipedal humanoid robot featuring a torque-stratified actuator hierarchy and a fault-tolerant distributed electrical architecture, deployed at low-cost. By introducing a formal interpretability mechanism into autonomous robotic decision-making, DOTBot builds operational trust that current disaster-response systems critically lack.

Awards Won: