VTAM: Video-Tactile-Action Models for Complex Physical Interaction Beyond VLAs

Haoran Yuan[1,*], Weigang Yi[1,*], Zhenyu Zhang[2,*], Wendi Chen[3,*] Yuchen Mo[1], Jiashi Yin[1], Xinzhuo Li[1], Xiangyu Zeng[1], Chuan Wen[3], Cewu Lu[3], Katherine Driggs-Campbell[1], Ismini Lourentzou[1]

PLAN LAB Perception and LANguage Lab

[1] University of Illinois Urbana-Champaign, [2] Stanford University, [3] Shanghai Jiao Tong University

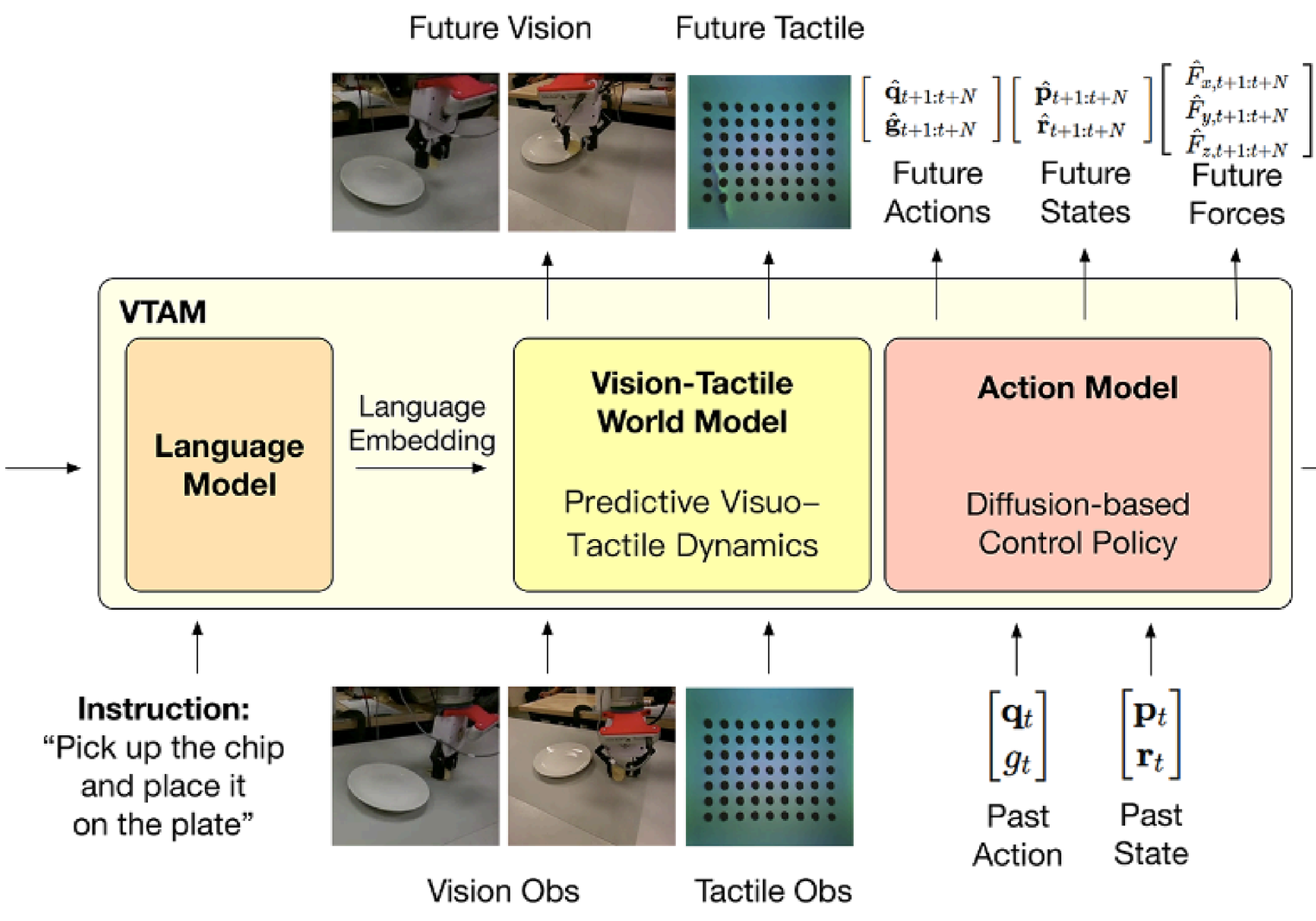
plan-lab.github.io/vtam

VTAM

We propose a new video-tactile world action model for contact-rich robotic manipulation. Given multi-view visual observations, tactile signals, and robot state/action context, VTAM predicts interaction dynamics and generates physically grounded actions for complex real-world tasks.

Vision-Tactile-Action Multimodal Data

- ✅ 10 Minute Teleoperation Per Task
- ✅ Zero Tactile Pretraining on Real Data
- ✅ Video-Derived Dynamics Priors
- ✅ No Additional Wrist Force Sensor Needed



Contact-Rich Tasks

Contributions

- ✓ Predictive visuo-tactile world model
- ✓ Lightweight modality-transfer finetuning
- ✓ Tactile regularization for balanced fusion
- ✓ Action + state + virtual-force prediction

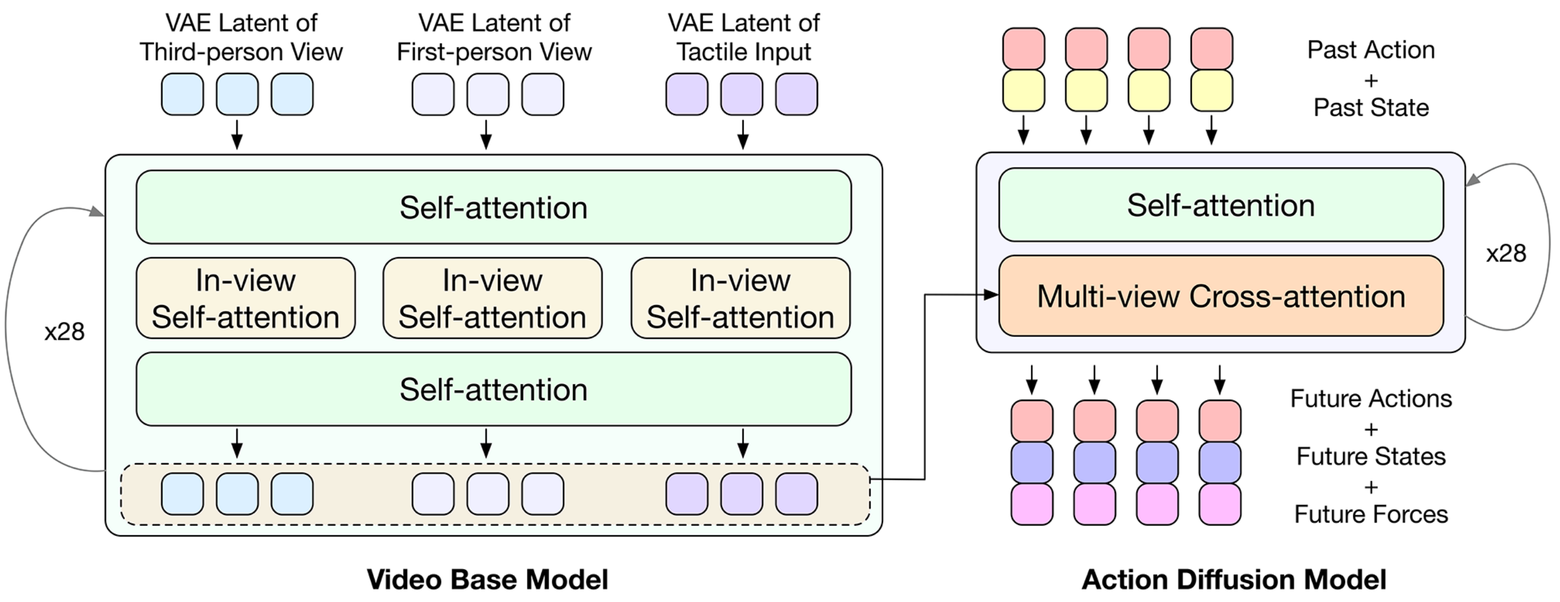
90% **average success** across contact-rich tasks

10 min **teleoperation / task** efficient real-world data

Why tactile?

- Vision misses subtle contact, slip, deformation, and force transitions.
- Contact-rich manipulation needs precise force modulation in addition to semantic understanding.
- Simply adding tactile input can still fail when visual features dominate multimodal attention.

Method



Stage I: joint visuo-tactile predictive dynamics

Stage II: conditional diffusion predicts future actions, states, and virtual forces

Teleoperation data collection

Experiments

80 trials/model · 20 trials/task setting

Model	Chip	Peel	Wipe
Genie Envisioner	0%	0%	2.5%
$\pi_{0.5}$ (Vision)	10%	0%	0%
$\pi_{0.5}$ + Tactile	5%	0%	0%
VTAM (Ours)	**90%**	**85%**	**95%**

Chip pick-and-place