Shenghua Wan, Xiaohai Hu, Xunlan Zhou, Lei Yuan, Le Gan, De-Chuan Zhan

Nanjing University · University of Washington

ICML 2026

Outline

Latent Action Learning Suffers from View-Dependent Noise

Why Latent Actions?

- World models need **action signals**
- Annotating actions is expensive
- Self-supervised learning is scalable

Core Insight

While observations vary across viewpoints, the **physical action** remains **invariant**.

Key Limitation

- Conflate 2D displacement with dynamics
- Gaussian priors too simplistic
- VQ codes lack information density

Our Proposal

Multi-view consistency as a superior inductive bias for learning robust latent actions.

MuCoLA: Multi-view Consistent Latent Action Learning

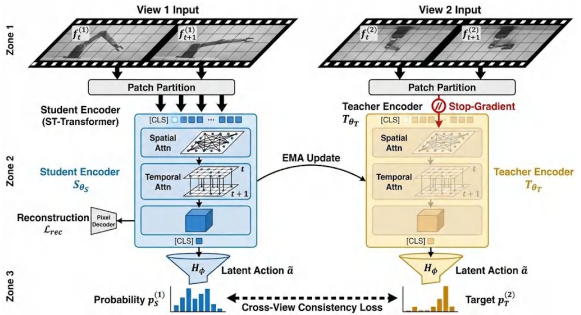

Figure 1. MuCoLA framework overview.

Stage 1: Latent Action Learning

- ST-Transformer with $[\texttt{CLS}]_{\text{act}}$ token
- Cross-view self-distillation (DINO)
- Reconstruction + consistency loss

Stage 2: World Model Training

- Freeze encoder, extract $\tilde{a}_{1:T}$
- Train iVideoGPT with $\tilde{a}$ conditioning

Cross-View Self-Distillation Filters View-Specific Noise

Student–Teacher Architecture

- **Student** S_{θ_S}: View 1 input
- **Teacher** T_{θ_T}: View 2 input
- EMA: $\theta_T \leftarrow \lambda\theta_T + (1-\lambda)\theta_S$
- Stop-gradient on Teacher

Cross-View Consistency Loss

$$\mathcal{L}_{\text{cons}} = -\sum_k \left[p_T^{(2)}(k) \log p_S^{(1)}(k) + p_T^{(1)}(k) \log p_S^{(2)}(k)\right]$$

- Cross-entropy vs. target
- Centering & sharpening

Total loss: $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{rec}} + \beta\,\mathcal{L}_{\text{cons}}\,, \quad \mathcal{L}_{\text{rec}} = \frac{1}{2}\|x - \hat{x}\|^2$

Multi-View Consistency Acts as a Spectral Noise Filter

Linear LAM Setup

For view v, the observation change is:

$$\Delta o^{(v)} = X^{(v)} a + \epsilon^{(v)}$$

- $a \sim \mathcal{N}(0, I_{d_a})$: shared action
- $X^{(v)}$: view-specific effect matrix
- $\epsilon^{(v)}$: view-specific noise

Proposition 3.1 (Noise Filtering)

Minimizing the cross-view consistency loss compels encoders $D^{(1)}, D^{(2)}$ to project onto the **shared action subspace** while remaining **orthogonal** to view-specific noise.

Implication

MuCoLA maximizes mutual information about the agent's intervention, discarding view-dependent nuisance factors.

Single-view (PCA): captures high-variance noise $\longrightarrow$ **MuCoLA:** isolates intrinsic action dynamics

MuCoLA Achieves Superior Reconstruction and Regression Accuracy

Table 1. Image Reconstruction Metrics

Method	PSNR↑	SSIM↑	LPIPS↓
AdaWorld (V1)	34.42	96.58	13.51
AdaWorld (V2)	32.53	91.45	15.72
FICC	30.28	90.07	18.65
LAPO	27.54	88.66	21.42
MuCoLA	**37.97**	**98.28**	**9.42**

Table 2. Action Regression (RMSE / MAE)

Method	Train	Test
AdaWorld	0.229	0.235
FICC	0.308	0.322
LAPO	0.264	0.289
MuCoLA	**0.202**	**0.214**
w/o consistency	0.220	0.225

MuCoLA consistently outperforms all baselines across views.

Learned Latent Space is Semantically Organized and View-Invariant

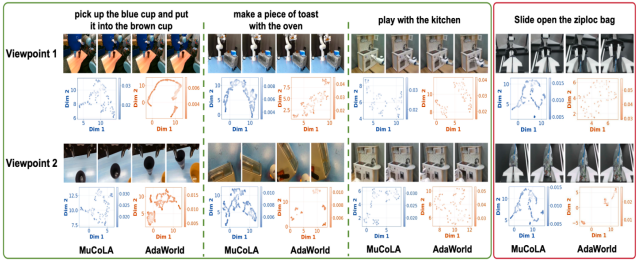

Figure 2. UMAP visualization: MuCoLA latent actions show consistent semantic clusters across views.

MuCoLA

- Recovers **shared action manifold**
- Well-defined semantic clusters
- High cross-view congruency

Gaussian-Prior Baselines

- Diffuse distributions
- No semantic partitioning
- Fail on multi-modal actions

Cross-View Generalization Enables Robust Unseen-View Prediction

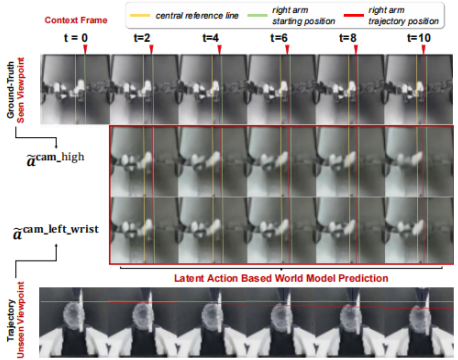

Figure 3. World model prediction using latent actions from an *unseen* left-wrist viewpoint.

Key Observations

- Correctly anticipates leftward motion
- Predictions match ground-truth
- Captures **physical semantics**

Implication

Train on multi-view data, **deploy with single view**, retaining semantic robustness.

MuCoLA Latent Actions Improve Visual Planning, RL, and Control

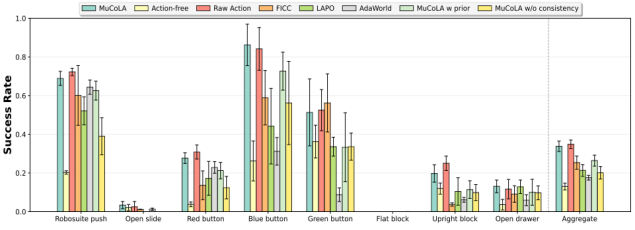

Figure 4. MPC success rates on VP2 benchmark across 8 tasks.

Table 3. Behavior Cloning on LIBERO (%)

Method	Long	Goal	Obj	Spat
AdaWorld	53.1	74.7	49.2	55.8
LAPO	59.1	61.6	39.5	41.0
MuCoLA	**67.1**	**71.0**	**55.3**	**61.2**

MuCoLA also achieves **state-of-the-art MBRL** on Meta-World across two distinct camera views.

Ablation Confirms Design Choices; MuCoLA Scales Favorably

Ablation Findings

- **Gaussian prior hurts**: cannot model multi-modal actions
- **Consistency loss is critical**: removal degrades all metrics

Label Efficiency

- Robust with only **5K** labels
- Non-trivial at 1K samples
- Low sample complexity

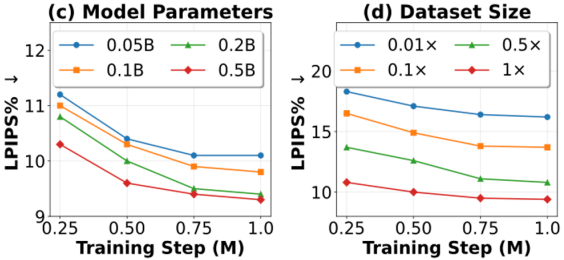

Figure 5. Scaling: larger models and more data yield monotonic gains.

MuCoLA: View-Invariant Actions Enable Scalable World Models

Key Contributions

1. Identified limitations of Gaussian priors and single-view reconstruction
2. Proposed MuCoLA: cross-view self-distillation for latent action learning
3. Proved multi-view consistency as **spectral noise filter**
4. SOTA on reconstruction, planning, MBRL, and behavior cloning

Limitations

- Requires synchronized multi-view data
- Long-horizon tasks remain challenging

Future Work

- Large-scale multi-view datasets
- View augmentation & pseudo-view synthesis
- Internet-scale video pretraining

Thank You!

Questions Welcome

Shenghua Wan · De-Chuan Zhan

zhandc@nju.edu.cn