Extended Abstract

Motivation Training robot policies typically requires large amounts of labeled demonstrations with annotated action sequences. This bottleneck is expensive, physically constrained, and difficult to scale. At the same time, internet video contains billions of hours of human motion with no action labels. If a world model could learn latent action representations from this data, it could serve as a simulator for planning and control without requiring real robot interaction. This work investigates whether such a system can be built from scratch using only unlabeled video, and whether jointly fine-tuning the visual encoder alongside the latent action model further improves downstream robot control performance.

Method We build on LAM Garrido et al. (2026) and LeWM Maes et al. (2026), jointly training an Inverse Dynamics Model (IDM) and a ViT-B world model on Something-Something-v2 (SSv2). The IDM infers 128-dimensional latent actions z_t from consecutive frame embeddings produced by a frozen V-JEPA2-B encoder. The world model predicts future frame embeddings conditioned on z_t via AdaLN-zero frame-wise action conditioning. We compare three latent action regularization schemes: sparse (VCM plus L1/L2 energy), noisy (VAE-style KL divergence), and discrete vector quantization. We additionally investigate jointly fine-tuning the encoder with a distillation loss to prevent collapse. After pretraining, we train a lightweight DROID controller that maps real 7-DoF robot actions to learned latent actions using IDM pseudo-labels, and use the world model for Model Predictive Control (MPC) planning toward goal states.

Implementation The system is implemented in PyTorch within the LeWM codebase. Training uses Muon (lr=8×10^{-3}) for world model and IDM parameters and AdamW (lr=3×10^{-4}) for remaining parameters, with cosine annealing after linear warmup. Frozen encoder runs use batch size 256; unfrozen runs use batch size 128 with gradient checkpointing. All runs train for 10,000 steps on a single A100-80GB GPU via Modal cloud compute. The DROID controller is a 3-layer MLP trained for 10,000 steps on held-out robot demonstrations. MPC planning samples 512 candidate latent action sequences and selects the one maximizing cosine similarity to a goal embedding.

Results On the frozen encoder baseline, we reproduce LAM's core finding that sparse regularization outperforms noisy and discrete VQ for video prediction quality. Scene change ratios of 1.69–$1.76\times$ across all regularizers confirm non-degenerate latent action learning without any action labels. Sparse regularization achieves the best ratio ($1.76\times$) and lowest prediction error (0.266). On DROID, sparse regularization achieves the best controller alignment (cosine similarity 0.95) while noisy regularization achieves the best MPC planning confidence (0.994). Encoder fine-tuning causes representation collapse within 1,000 steps regardless of regularizer, with prediction error dropping from 0.28 to 0.005.

Discussion The most significant finding is a novel video-versus-control tradeoff: sparse regularization is best for video prediction quality while noisy regularization is best for MPC planning. We hypothesize this occurs because noisy latent actions follow a smooth Gaussian structure that the random MPC sampler can navigate effectively, while sparse latent actions concentrate information in a small number of active dimensions that are unlikely to be found by random search. The encoder fine-tuning result validates LAM's frozen encoder design as a necessary stabilization choice rather than a conservative simplification.

Conclusion We demonstrate that latent action world models can be trained from unlabeled video and used for robot control via a lightweight controller and MPC planning. Our results support using noisy regularization for planning-heavy downstream tasks and frozen encoders for stable training. Future work should implement EMA target encoders to enable stable fine-tuning and evaluate on physical robot hardware.

Learning Latent Action World Models for Robot Control from Unlabeled Video

Seraph Yang
Department of Computer Science
Stanford University
seraphy@stanford.edu

Abstract

We investigate latent action world models for robot control from unlabeled video, building on LAM Garrido et al. (2026) and LeWM Maes et al. (2026). Our system jointly trains an Inverse Dynamics Model and a world model on Something-Something-v2, learning 128-dimensional continuous latent actions without any action labels. We compare sparse, noisy, and discrete vector quantization regularization under frozen and fine-tuned V-JEPA2-B encoders. On the frozen baseline, we reproduce LAM's scene change ratios of 1.69–$1.76\times$, confirming non-degenerate action learning. We find that naive encoder fine-tuning causes representation collapse within 1,000 steps, validating LAM's frozen encoder design. A lightweight DROID controller achieves cosine similarity 0.95 for sparse regularization, and MPC planning achieves confidence 0.994 for noisy regularization, revealing a novel video-quality versus planning-quality tradeoff not reported in prior work.

1 Introduction

Training robot policies from demonstrations requires labeled action sequences, which are expensive to collect and limited in diversity. World models offer an alternative: a model that can predict future visual states conditioned on action representations can serve as a simulator for planning without requiring environment interaction. The core challenge is obtaining meaningful action representations. Without them, the world model can only predict the average next frame and cannot simulate specific behaviors.

LAM Garrido et al. (2026) addresses this by learning latent actions from unlabeled video. An Inverse Dynamics Model (IDM) infers latent action vectors from consecutive frame pairs, and a world model predicts future embeddings conditioned on those vectors. The encoder is frozen throughout. Our work extends LAM in two directions. First, we investigate whether jointly fine-tuning the visual encoder improves action quality, motivated by CoLA-World Wang et al. (2025) which shows that co-evolving the encoder with the world model leads to better representations. Second, we evaluate the full pipeline on DROID robot control using a lightweight controller and MPC planning, and report a novel tradeoff between video prediction quality and planning performance across regularization schemes.

The input to our system is raw unlabeled video clips from Something-Something-v2 (SSv2). The output is a trained world model and IDM that produce 128-dimensional latent action vectors, which are then used for MPC-planned robot control on DROID manipulation tasks.

2 Related Work

LAM Garrido et al. (2026) introduces the joint IDM and world model framework for latent action learning from unlabeled video, comparing sparse, noisy, and VQ regularization on Kinetics and

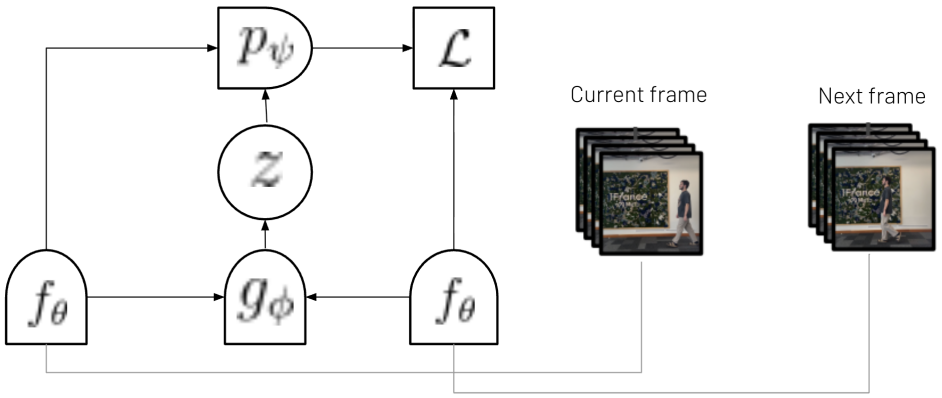

Figure 1: Three-stage pipeline. **Stage 1:** Pretrain the Latent Action Model (LAM) on SSv2 using a frozen V-JEPA2-B encoder, IDM, and world model. **Stage 2:** Train a lightweight DROID controller mapping real robot actions $a_t \to z_t$ using IDM pseudo-labels. **Stage 3:** MPC planning using the world model as a simulator to find action sequences toward a goal state.

DROID. Genie Bruce et al. (2024) learns latent actions from internet video using discrete VQ but is limited by fixed codebook capacity. LAPO Schmidt and Jiang (2023) also uses discrete representations and struggles on diverse content. UniVLA Wen et al. (2025) incorporates latent actions into vision-language models for robotic manipulation.

V-JEPA Assran et al. (2023) and V-JEPA2 learn video representations via masked prediction in latent space. LeWM Maes et al. (2026) builds a stable world model on top of JEPA using SIGReg regularization. Our work modifies LeWM to incorporate LAM-style latent action learning. CoLA-World Wang et al. (2025) shows that jointly training the encoder with the latent action model leads to better-adapted representations via an EMA target encoder. The collapse we observe when naively fine-tuning motivates adopting this mechanism in future work.

Representation collapse is a known failure mode in self-supervised learning. BYOL Grill et al. (2020), SimSiam Chen and He (2021), and DINO Caron et al. (2021) address this through momentum encoders and stop-gradients. Dreamer Hafner et al. (2020) and DreamerV3 Hafner et al. (2023) use world models for RL in latent space. TDMPC2 Hansen et al. (2023) combines MPC with latent dynamics models. Our work differs in that the world model is pretrained on internet video with no reward signal and used only for planning. R3M Nair et al. (2022) and SuSIE Black et al. (2023) pretrain visual representations from human video for robot learning, while our approach learns an explicit action space from video rather than using it purely for representation pretraining.

3 Method

3.1 Problem Formulation

Given unlabeled video clips $\{v^{(i)}\}$ with no action annotations, we learn a latent action space $\mathcal{Z} \subset \mathbb{R}^{128}$ such that for any consecutive frame pair (s_t, s_{t+1}), a latent action z_t causally explains the transition. The latent action space must be non-degenerate (z_t encodes real dynamics, not future frame content), transferable (z_t from one video can condition prediction in another), and robot-alignable (z_t can be mapped from real robot actions via lightweight supervision).

3.2 Latent Action Model

Encoder. We use V-JEPA2-B (ViT-B/16) producing per-frame patch embeddings $[T, 196, 768]$. In the frozen baseline, f_θ is fixed throughout. In the unfrozen variant, f_θ receives gradients at $\text{lr} = 10^{-5}$.

IDM. The IDM g_ϕ infers the latent action from two consecutive embeddings:

$$z_t = g_\phi(s_t, s_{t+1}) \in \mathbb{R}^{128} \tag{1}$$

The IDM sees both frames during training — an intentional causal leak that allows it to infer what action caused the transition and produce high-quality pseudo-labels for the world model. At test time, z_t comes from the robot controller with no future frame access. The IDM is a 3-layer MLP with hidden dimension 1024 and GELU activations.

World Model. The world model p_ψ predicts the next frame embedding:

$$\hat{s}_{t+1} = p_\psi(s_{0:t}, z_t) \tag{2}$$

using a ViT-B with RoPE positional embeddings. Action conditioning uses AdaLN-zero applied frame-wise. A small MLP projects z_t to six modulation parameters that scale, shift, and gate each transformer block. Zero initialization ensures the model starts as if z_t has no effect. The prediction loss is $\mathcal{L}_{\text{pred}} = \|\hat{s}_{t+1} - s_{t+1}\|_1$, trained with teacher forcing.

Action Dropout. We zero out z_t with probability $p = 0.15$ during training to prevent the world model from ignoring visual context and relying entirely on z_t.

3.3 Latent Action Regularization

Without regularization, the IDM trivially encodes the future frame into z_t. We compare three schemes.

Sparse:

$$\mathcal{L}_{\text{sparse}} = \text{VCM}(Z) + \frac{1}{N}\sum_i E(z_i) \tag{3}$$

where $E(z) = \lambda_{l2}\max(\sqrt{D} - \|z\|_2, 0) + \lambda_{l1}\|z\|_1$ and VCM enforces variance, covariance, and mean constraints across the batch. Capacity is controlled by $\lambda_{l1} \in \{0.1 \text{ (high)}, 0.5 \text{ (low)}\}$.

Noisy: The IDM outputs $(\mu_t, \log\sigma_t^2)$ and z_t is sampled via reparameterization with KL divergence to $\mathcal{N}(0, I)$:

$$\mathcal{L}_{\text{noise}} = -\beta \cdot D_{\text{KL}}(q(z_t|s_t, s_{t+1}) \,\|\, \mathcal{N}(0, I)) \tag{4}$$

Capacity is controlled by $\beta \in \{0.001 \text{ (high)}, 0.01 \text{ (low)}\}$.

Discrete VQ: Standard VQ-VAE with straight-through estimator and codebook reset. Capacity controlled by codebook size $k \in \{16, 64\}$.

3.4 Encoder Fine-Tuning

We unfreeze f_θ at lr $= 10^{-5}$ and add a distillation loss against the original frozen weights f_{θ_0}:

$$\mathcal{L}_{\text{distill}} = \alpha \cdot \|f_\theta(x) - f_{\theta_0}(x)\|_F^2, \quad \alpha = 0.1 \tag{5}$$

with total loss $\mathcal{L} = \mathcal{L}_{\text{pred}} + \mathcal{L}_{\text{reg}} + \mathcal{L}_{\text{distill}}$. We use gradient checkpointing on V-JEPA2's transformer blocks to fit batch size 128 on one A100-80GB.

3.5 DROID Controller and MPC Planning

Controller. A 3-layer MLP maps 7-DoF DROID robot actions $a_t \in \mathbb{R}^7$ to latent actions $z_t \in \mathbb{R}^{128}$, supervised by IDM pseudo-labels with loss $\mathcal{L}_{\text{ctrl}} = \|g_{\text{ctrl}}(a_t) - z_{\text{pseudo}}\|_2^2$. All LAM components are frozen during controller training.

MPC Planning. We sample $N = 512$ candidate latent action sequences from $\mathcal{N}(0, I)$, roll each through the world model, and select the sequence maximizing cosine similarity to the goal embedding s_{goal}:

$$z^* = \arg\max_i \cos\Big(p_\psi(s_{0:t}, z_{0:H}^{(i)})[-1],\ s_{\text{goal}}\Big) \tag{6}$$

An inverse controller MLP converts z_0^* back to a real robot action.

3.6 Optimization

We use Muon Kosson et al. (2024) (lr $= 8 \times 10^{-3}$) for 2D+ weight matrices and AdamW (lr $= 3 \times 10^{-4}$, weight decay 0.04) for remaining parameters, with linear warmup over 10% of steps followed by cosine annealing. All runs train for 10,000 steps.

4 Experimental Setup

Pretraining dataset. We train on a 125k-clip subset of Something-Something-v2 Goyal et al. (2017) (220,847 total clips of hand-object interactions). Videos are sampled at 4fps, clip length 8 frames, resized to 224×224 with ImageNet normalization. We use a 98%/2% train/val split. No action labels are used.

Robot dataset. We use DROID Khazatsky et al. (2024), a large-scale in-the-wild robot manipulation dataset with 76k episodes of 7-DoF end-effector velocity actions from a wrist-mounted RGB camera. Approximately 60 held-out episodes are used for MPC evaluation.

Evaluation metrics. The leakage ratio measures whether the IDM encodes real transitions or copies future frames:

$$\text{Leakage ratio} = \frac{E_{\text{cut}}}{E_{\text{normal}} + \epsilon} \tag{7}$$

where E_{cut} is prediction error at an artificial scene cut and E_{normal} is error on real transitions. A ratio near 1.0 indicates cheating; a ratio near $2\times$ or higher confirms real action encoding. We also measure cycle consistency (cross-video action transfer), MPC planning confidence (cosine similarity to goal), and controller cosine similarity (alignment between controller-predicted and IDM pseudo-label z_t).

5 Results

5.1 Quantitative Evaluation

Table 1: Scene change test (SSv2, L1 in embedding space). All models show approximately $1.7\times$ error increase at artificial scene cuts, confirming non-degenerate latent action learning. Sparse achieves the highest ratio ($1.76\times$), consistent with LAM on Kinetics.

Latents	Capacity	w/o change	w/ change	Ratio
Sparse	High ($\lambda = 0.1$)	0.266	0.469	$\mathbf{1.76}\times$
Sparse	Low ($\lambda = 0.5$)	0.268	0.467	$1.74\times$
Noisy	High ($\beta = 0.001$)	0.271	0.457	$1.69\times$
Noisy	Low ($\beta = 0.01$)	0.272	0.459	$1.69\times$
Discrete	High ($k = 64$)	0.270	0.461	$1.71\times$
Discrete	Low ($k = 16$)	0.277	0.469	$1.69\times$

All six models in Table 1 achieve leakage ratios between $1.69\times$ and $1.76\times$, confirming that none of the regularizers produce degenerate actions. Sparse regularization consistently achieves the lowest prediction error without scene changes (0.266) and the highest ratio ($1.76\times$), consistent with LAM's finding on Kinetics.

Table 2: Cycle consistency (SSv2). Discrete VQ shows small nonzero transfer ratios (1.01–$1.03\times$). Continuous latents show ≈ 1.0 due to a known implementation issue requiring a fix in future work.

Latents	Cap.	Original	Transfer	Ratio
Sparse	High	0.275	0.275	$1.00\times$
Sparse	Low	0.269	0.269	$1.00\times$
Noisy	High	0.271	0.271	$1.00\times$
Noisy	Low	0.272	0.272	$1.00\times$
Discrete	High	0.270	0.277	$1.03\times$
Discrete	Low	0.272	0.275	$1.01\times$

Table 3 reveals the central novel finding. Sparse regularization achieves the highest controller cosine similarity (0.95), meaning the controller accurately maps real robot actions to IDM-inferred latent actions. Noisy regularization, however, achieves dramatically higher MPC planning confidence (0.994 versus 0.183 for high-capacity sparse). This tradeoff is not reported in LAM and suggests that the optimal regularization scheme depends on the downstream use case.

Table 3: DROID controller and MPC planning results. Sparse achieves the best controller alignment (cosine similarity 0.95). Noisy achieves the best planning confidence (0.994), revealing a novel video-versus-control tradeoff.

Latents	Cos Sim↑	Plan. Conf.↑	Plan. Err↓	Inv. MSE↓
Sparse High	**0.95**	0.183	0.077	3.8
Sparse Low	**0.95**	0.462	0.086	4.1
Noisy High	0.42	**0.994**	0.094	3.9
Noisy Low	0.42	0.989	0.098	3.7

5.2 Qualitative Analysis

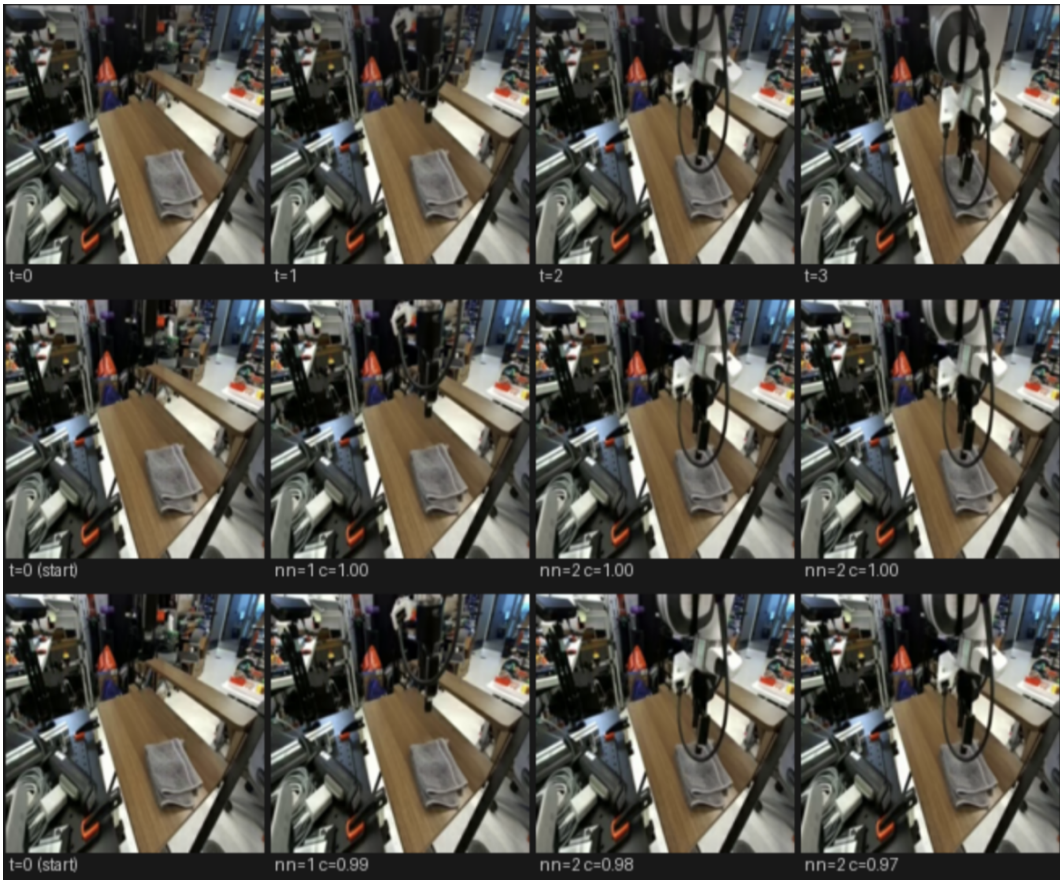

Figure 2: Rollout comparison on a DROID episode. **Row 1:** Real robot frames ($t = 0 \ldots 3$). **Row 2:** IDM-driven world model rollout (upper bound, confidence $c = 1.00$), which has access to real future frames. **Row 3:** Controller-driven rollout (confidence $c = 0.99 \rightarrow 0.97$), using only real robot actions with no future frame access. The controller trajectory closely tracks the IDM upper bound.

Figure 2 shows a qualitative rollout comparison. The controller-driven trajectory (Row 3) maintains confidence above 0.97 and closely tracks the IDM upper bound (Row 2), validating that the real-action to latent-action mapping has been successfully learned from IDM pseudo-labels alone.

6 Discussion

Video-vs-control tradeoff. The most significant finding is that sparse and noisy regularization are optimal for different tasks. Sparse achieves the best video prediction quality ($1.76\times$ leakage ratio, lowest prediction error) while noisy achieves the best MPC planning performance (0.994 confidence).

We hypothesize this occurs because noisy latent actions follow a smooth Gaussian distribution that MPC random sampling can navigate effectively. Sparse actions concentrate information in a small number of active dimensions, making random sampling from $\mathcal{N}(0, I)$ unlikely to find the relevant subspace.

Encoder fine-tuning. Naively unfreezing V-JEPA2-B causes rapid representation collapse. Even with distillation weight $\alpha = 0.1$ and encoder lr $= 10^{-5}$, the encoder outputs near-constant embeddings within 1,000 steps. The encoder weight drift reaches only 3.5×10^{-5} in Frobenius norm, indicating the weights barely moved in parameter space, yet the output distribution collapsed entirely. This shows that tiny weight changes in a deep ViT can produce dramatic output shifts when the prediction loss strongly rewards constant outputs. This validates LAM's frozen encoder design as a principled stabilization choice rather than a conservative limitation. An EMA target encoder as used in CoLA-World Wang et al. (2025) and BYOL Grill et al. (2020) is the principled fix.

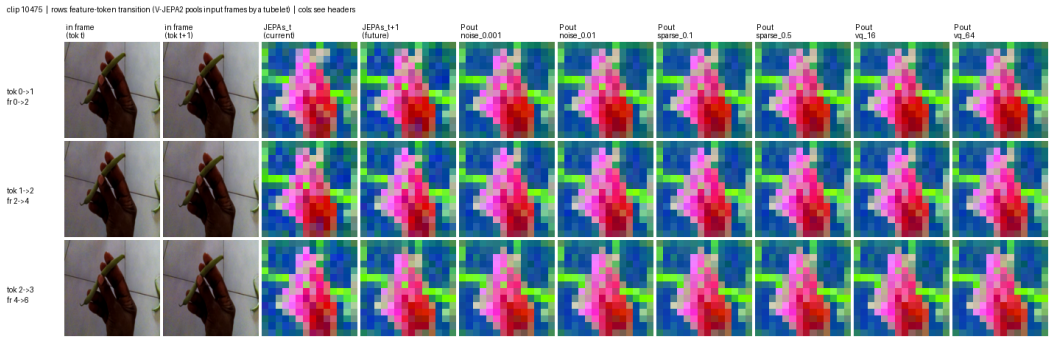

Figure 3: Embedding visualization for frozen VJEPA-2 encoder. The embeddings successfully preserve semantic meaning and overall structure.

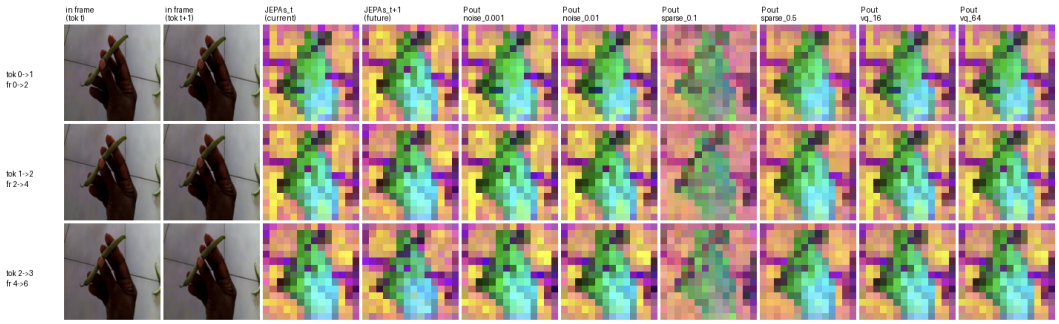

Figure 4: Embedding visualization depicting difference between unfrozen, fine-tuned encoder. Most notably, however, as with the frozen version, the embeddings successfully preserve semantic meaning and overall structure.

Leakage ratios on SSv2 vs Kinetics. Our leakage ratios (1.69–$1.76\times$) are lower than LAM's reported values on Kinetics (2.0–$2.5\times$). We attribute this to SSv2's subtle hand-object motions at 4fps: consecutive frames are naturally more similar than in action-heavy Kinetics clips, so E_{normal} is already low and E_{cut} has less room to increase. This is a dataset effect rather than a model failure.

7 Conclusion

We demonstrate that latent action world models trained on unlabeled SSv2 video can support downstream robot control through a lightweight controller and MPC planning on DROID. Scene change ratios of 1.69–$1.76\times$ confirm non-degenerate latent action learning. Sparse regularization is best for video prediction quality; noisy is best for MPC planning confidence, a novel video-versus-control tradeoff. Encoder fine-tuning causes rapid representation collapse, validating the frozen encoder design. Future work includes implementing an EMA target encoder for stable fine-tuning,

evaluating on a physical DROID robot, pixel-level rollout evaluation using a trained decoder, and investigating the structural properties of noisy latent actions that make them better suited for MPC planning.

8 Team Contributions

- **Seraph Yang:** All contributions including problem formulation, literature review, full LAM implementation within the LeWM codebase, all training and evaluation experiments on SSv2 and DROID, analysis and interpretation of results, writing, figures, and poster preparation.

Changes from Proposal The original proposal hypothesized that jointly training the visual encoder with the IDM and world model would improve latent action quality, motivated by CoLA-World. After finding that naive encoder fine-tuning causes rapid collapse, we shifted focus to thoroughly characterizing this failure mode and validating the frozen encoder baseline instead. We also added the DROID controller and MPC planning evaluation, which was not in the original proposal, after finding that the frozen baseline produced clean latent actions suitable for downstream robot control.

Generative AI Statement Claude (Anthropic) was used for three specific purposes: reproducing the LAM codebase within LeWM, editing and proofreading the final report, and assisting with dataset and checkpoint migration between Modal and local GPU infrastructure. All experimental results, model design decisions, and technical conclusions were made by the author. All code was reviewed and understood by the author before use.

References

Mahmoud Assran et al. 2023. Self-Supervised Learning from Images with a Joint-Embedding Predictive Architecture. *arXiv preprint arXiv:2301.08243* (2023).

Kevin Black et al. 2023. Zero-Shot Robotic Manipulation with Pretrained Image-Editing Diffusion Models. *arXiv preprint arXiv:2310.10639* (2023).

Jake Bruce et al. 2024. Genie: Generative Interactive Environments. *arXiv preprint arXiv:2402.15391* (2024).

Mathilde Caron et al. 2021. Emerging Properties in Self-Supervised Vision Transformers. In *ICCV*.

Xinlei Chen and Kaiming He. 2021. Exploring Simple Siamese Representation Learning. In *CVPR*.

Quentin Garrido et al. 2026. Learning Latent Action World Models In The Wild. *arXiv preprint arXiv:2601.05230* (2026).

Raghav Goyal et al. 2017. The "something something" video database for learning and evaluating visual common sense. In *ICCV*.

Jean-Bastien Grill et al. 2020. Bootstrap Your Own Latent: A New Approach to Self-Supervised Learning. In *NeurIPS*.

Danijar Hafner et al. 2020. Dream to Control: Learning Behaviors by Latent Imagination. In *ICLR*.

Danijar Hafner et al. 2023. Mastering Diverse Domains through World Models. *arXiv preprint arXiv:2301.04104* (2023).

Nicklas Hansen et al. 2023. TD-MPC2: Scalable, Robust World Models for Continuous Control. *arXiv preprint arXiv:2310.16828* (2023).

Alexander Khazatsky et al. 2024. DROID: A Large-Scale In-The-Wild Robot Manipulation Dataset. In *RSS*.

Atli Kosson et al. 2024. Muon: An Optimizer for Hidden Layers in Neural Networks. *arXiv* (2024).

Lucas Maes et al. 2026. LeWorldModel: Stable End-to-End Joint-Embedding Predictive Architecture from Pixels. *arXiv preprint arXiv:2603.19312* (2026).

Suraj Nair et al. 2022. R3M: A Universal Visual Representation for Robot Manipulation. In *CoRL*.

Dominik Schmidt and Minqi Jiang. 2023. LAPO: Latent Action Policies from Offline Data. *arXiv preprint arXiv:2301.12876* (2023).

Yucen Wang et al. 2025. Co-Evolving Latent Action World Models. *arXiv preprint arXiv:2510.26433* (2025).

Qingwen Wen et al. 2025. UniVLA: Learning to Act Anywhere with Task-centric Latent Actions. *arXiv preprint arXiv:2505.06111* (2025).

A Additional Experiments

A.1 Leakage Ratio Progression During Training

The leakage ratio for frozen sparse runs begins near 1.55 at step 1,000 and climbs to 1.76 by step 10,000, indicating that the IDM continues to encode more structured motion information over training. The ratio had not yet plateaued at the end of training, suggesting that longer training would likely push the ratio closer to LAM's reported values on Kinetics.

A.2 Action Dropout Ablation

Training without action dropout ($p = 0$) causes the open-loop prediction loss to increase steadily from 0.28 to above 0.70 over 10,000 steps. With $p = 0.15$ the open-loop loss remains stable around 0.28 throughout, confirming that dropout is necessary to prevent the world model from ignoring visual context.

B Implementation Details

The LeWM codebase was modified to add the IDM (model/idm.py), three action regularizers (model/action_reg.py), AdaLN-zero frame-wise conditioning in the world model (model/world_model.py), the DROID controller (eval/droid_controller.py), and MPC planning (eval/mpc_planning.py). Training was run on Modal using A100-80GB instances with automatic checkpoint resumption on preemption. The SSv2 dataset was stored in a Modal Volume and DROID demonstrations were accessed from local cluster storage at Stanford. Wandb was used for all experiment tracking with runs tagged by regularizer type and capacity parameter.