# Learning Action-Conditioned World Models for Cataract Surgery from Unlabeled Videos

**Nisarg A. Shah**[*1] SNISARG812@GMAIL.COM
[1] *Johns Hopkins University, Baltimore, USA*
**Mingze Xia**[*1] MXIA8@JHU.EDU
**Shameema Sikder**[2,3] SSIKDER1@JHMI.EDU
[2] *Malone Center for Engineering in Healthcare, Baltimore, USA*
[3] *Wilmer Eye Institute, Johns Hopkins University, Baltimore, USA*
**S. Swaroop Vedula**[2] SWAROOP@JHU.EDU
**Vishal M. Patel**[1] VPATEL36@JHU.EDU

**Editors:** Under Review for MIDL 2026

## Abstract

Vision foundation models have enabled automated analysis of cataract surgery videos, but existing self-supervised approaches treat video as state-only sequences, limiting causal reasoning and sample efficiency in label-scarce settings. We present *SurgWorld*, an action-conditioned world model that learns surgical dynamics from unlabeled cataract videos by combining a Latent Action Tokenizer, which discretizes frame-to-frame motion into atomic action primitives, with a latent predictor trained on top of a frozen cataract foundation encoder. By modeling state transitions in feature space conditioned on inferred actions rather than generating pixels, *SurgWorld* separates tool motion from static anatomy and learns a latent control signal that is complementary to visual appearance. Pretrained on a multi-institutional corpus and evaluated on four cataract datasets, *SurgWorld* improves step recognition accuracy over state-only baselines, with gains of about 10 percentage points in low-data regimes, indicating that explicit dynamics provide a sample-efficient prior. Ablation studies show that action-only features are already discriminative, and that fusing actions with vision encoder features achieves state-of-the-art performance and consistent improvements in step anticipation. These results support the view that latent actions capture orthogonal temporal structure that describes how cataract procedures progress.

**Keywords:** Surgical Video Analysis, World Models, Latent Actions, SSL

## 1. Introduction

Automated surgical video analysis forms the foundation for modern intraoperative assistance, enabling applications such as step recognition, workflow optimization, and objective skill assessment (Maier-Hein et al., 2017; Yu et al., 2019; Padoy, 2019). To address the scarcity of annotated medical data, the field has increasingly adopted self-supervised learning (SSL) strategies on large unlabeled surgical videos (Shah et al., 2025b; Centeno López et al., 2025; Yang et al., 2025). Current state-of-the-art vision foundation models, including JHU-VPT (Shah et al., 2025b) and temporal transformers trained via masked autoencoding (Tong et al., 2022; Bandara et al., 2023; Shah et al., 2025a), learn video representations by reconstructing masked pixels or maximizing feature similarity. These architectures function

[*] Contributed equally

as state-only sequence models, approximating $p(x_{\text{masked}} \mid x_{\text{visible}})$ based on visual texture and temporal correlation. They act as passive observers that do not explicitly model the temporal dynamics, i.e., how instrument interactions ($a_t$) drive state transitions ($s_t \rightarrow s_{t+1}$) (Ding et al., 2025; Chen et al., 2025). By treating video as a sequence of textures rather than a physical process, these models struggle with tasks requiring temporal reasoning, such as long-horizon step anticipation (Damen et al., 2020), and exhibit poor sample efficiency in label-scarce regimes where memorizing visual patterns is insufficient (Lecuyer et al., 2020; Padoy, 2019; Funke et al., 2019).

To bridge the gap between passive observation and physical understanding, and to enable reasoning and planning, the field of embodied AI has developed world models (Ha and Schmidhuber, 2018; Hafner et al., 2019). A world model learns an internal simulation of the environment defined by the transition density $P(s_{t+1} \mid s_t, a_t)$ (LeCun, 2022). While world models have been applied to robotic-assisted surgery where kinematic logs provide ground-truth action labels (Agarwal et al., 2025; Assran et al., 2025; Yang et al., 2023; Zhai et al.), applying this paradigm to microscopic surgery performed without robotic assistance presents a data availability constraint. Human surgery lacks proprioceptive recording; millions of video frames exist without corresponding control signals. Without explicit action annotations, training action-conditioned dynamics models has historically been infeasible. Recent attempts to bridge this gap rely on pixel-space video generation (Chen et al., 2025; Koju et al., 2025). However, evaluating generative models in clinical settings indicates that high visual fidelity does not guarantee correct physical dynamics, and pixel hallucination poses safety and performance risks for downstream analysis (Assran et al., 2025; Tivnan et al., 2024; Kim et al., 2025a,b).

We address these limitations by proposing *SurgWorld*, a framework that learns action-conditioned dynamics from video without kinematic supervision. Our approach rests on the hypothesis that the control signal is implicitly encoded in the visual transformation between consecutive frames. We recover this signal and learn dynamics in two stages. First, we introduce a Latent Action Tokenizer (LAT), based on Vector-Quantized Variational Autoencoders (VQ-VAE) (Van Den Oord et al., 2017; Razavi et al., 2019; Ye et al., 2024), which discretizes the motion residual between frames into a compact vocabulary of atomic action primitives. This process separates instrument motion from static appearance, providing inferred action tokens ($a_t$) that contain information complementary to standard visual encoders. Second, we utilize these tokens to train a Latent World Model based on the V-JEPA architecture (Bardes et al., 2024; Shah et al., 2025b). Unlike generative baselines that operate in pixel space (Bruce et al., 2024; Koju et al., 2025), our model predicts future states in a high-level semantic feature space. This objective forces the model to capture the causal structure of the procedure: how specific primitives transform the surgical state, while abstracting away irrelevant pixel-level noise.

We evaluate *SurgWorld* on four cataract surgery datasets, including Cataract-1k (Ghamsarian et al., 2024) for step recognition. To analyse the benefit of action-conditioned modeling, we compare to the state-only JHU-VPT baseline. In 10-25% label regimes, accuracy improves by about 10 percentage points over JHU-VPT (e.g., from 63.8% to 73.9% on Cataract-1k-JHU at the 10% split). We also observe consistent gains in step anticipation, indicating that the learned world model encodes useful information about future surgical progression.

**Contributions.**

- **Unsupervised discovery of action primitives.** We propose a Latent Action Tokenizer that infers discrete surgical interactions from raw video pixels and provides latent action tokens that contain information complementary to visual features.
- **Action-conditioned feature prediction.** We introduce a surgical world model based on V-JEPA that integrates explicit action conditioning in latent space, enabling dynamics modeling without pixel-space generation.
- **Data-efficient cataract representations.** We demonstrate that SurgWorld outperforms state-of-the-art vision baselines (JHU-VPT, VideoMAE) on step recognition, particularly in low-data settings, and achieves consistent improvements in step anticipation.

## 2. Method

We formulate learning surgical dynamics from unlabeled cataract videos as a partially observable Markov decision process (POMDP) without explicit control signals. Let $\mathcal{D} = \{\mathbf{x}^{(i)}\}_{i=1}^{N}$ be a collection of monocular microscope videos. Each video $\mathbf{x}^{(i)} = (x_1^{(i)}, \ldots, x_{T_i}^{(i)})$ is sampled at 1 fps, where $x_t \in \mathbb{R}^{H \times W \times C}$ denotes the surgical field of view at time $t$. Conceptually, each frame $x_t$ is generated from a latent surgical state $s_t$ (tool configuration and ocular structures) and an action $a_t$ that summarizes the surgeon's tool motion applied between $t$ and $t+1$.

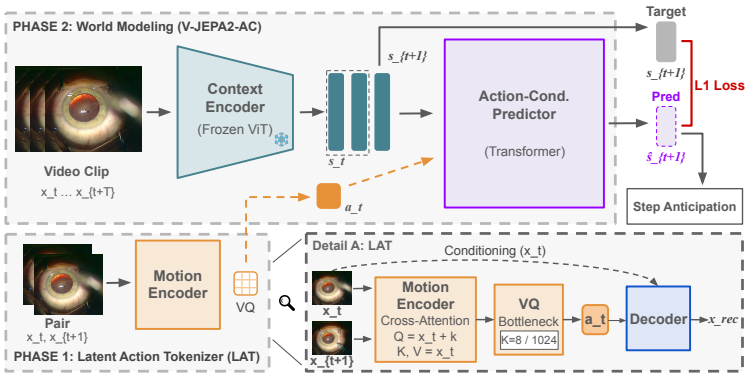


Figure 1: SurgWorld consists of two stages. Phase 1 (LAT): consecutive frames $(x_t, x_{t+1})$ are encoded with a motion encoder and vector-quantized into a discrete action primitive $a_t$, with a decoder conditioned on $x_t$ enforcing reconstruction of $x_{t+1}$. Phase 2 (Latent World Model): a frozen cataract encoder maps frames to latent states $s_t$, and an action-conditioned transformer predicts next states $\hat{s}_{t+1}$ from $(s_t, a_t)$ using a feature-prediction loss, enabling downstream step recognition and anticipation.

In robot-assisted surgery, $a_t$ is observed via kinematic logs and world models can be trained directly on $(s_t, a_t, s_{t+1})$ tuples (Agarwal et al., 2025; Assran et al., 2025; Ye et al., 2024). In microscopic cataract surgery, $a_t$ is unobserved and only the video sequence $\mathbf{x}_{1:T}$ is available. Our goal is to move from state-only sequence modeling to an action-conditioned world model that approximates one-step dynamics $p(s_{t+1} \mid s_t, a_t)$ using video alone. We assume that the missing action signal is implicitly encoded in the visual transformation between consecutive frames. *SurgWorld* recovers this signal and learns the dynamics in two stages. First, a **Latent Action Tokenizer** (LAT) approximates an inverse dynamics map $a_t \approx f_{\text{LAT}}(x_t, x_{t+1})$ and discretizes the motion residual between frames into a sequence of

atomic action primitives. Second, a **Latent World Model** uses these primitives to predict future states in a semantic feature space derived from a surgical video encoder. This section describes the LAT; the world model is presented in Section 2.2.

### 2.1. Latent Action Tokenizer (LAT)

The core hypothesis behind LAT is that the missing action $a_t$ is encoded in the visual residual that maps the current frame $x_t$ to the next frame $x_{t+1}$. In cataract surgery, this residual corresponds to changes such as tool motion within the anterior chamber, lens material removal, or capsule deformation. LAT is designed to extract this residual as a discrete latent variable, approximating an inverse dynamics map $a_t \approx f(x_t, x_{t+1})$.

**Motion encoding and pre-quantized action features.** LAT operates on consecutive frame pairs $(x_t, x_{t+1})$ sampled at 1 fps from the same video. Both frames are first passed through a shared patch embedding and spatial transformer, producing token sequences for $x_t$ and $x_{t+1}$ in a latent space (Villegas et al., 2022). To focus on the transformation rather than static appearance, we use a cross-attention module in which tokens of $x_{t+1}$ act as queries and tokens of $x_t$ act as keys and values, following the latent action quantization setup in GENIE (Bruce et al., 2024).

The output is aggregated into a continuous *pre-quantized action feature*

$$h_t = E_{\text{mot}}(x_t, x_{t+1}) \in \mathbb{R}^{d_a},$$

where $E_{\text{mot}}$ is the motion encoder and $d_a$ is the action feature dimension. The vector $h_t$ encodes the motion residual between $x_t$ and $x_{t+1}$ in a coordinate system aligned with surgical dynamics rather than pixel space. In our ablations, we also use $h_t$ directly as a continuous action representation to assess how much information is lost during quantization.

**Vector quantization and action primitives.** To obtain a compact vocabulary of reusable surgical interactions, we discretize $h_t$ using a vector quantizer (**???**). We maintain a learnable codebook $\mathcal{C} = \{e_k\}_{k=1}^{K} \subset \mathbb{R}^{d_a}$ and assign each $h_t$ to its nearest codebook vector:

$$a_t = \arg\min_k \|h_t - e_k\|_2^2, \qquad \tilde{h}_t = e_{a_t}. \tag{1}$$

Here $a_t \in \{1, \ldots, K\}$ is the *latent action primitive index* and $\tilde{h}_t$ is the corresponding *post-quantized action feature*. The sequence $(a_1, \ldots, a_T)$ defines the latent control signal that will later condition the world model.

We train the codebook with a standard VQ-VAE objective (Razavi et al., 2019) and adopt Noise Substitution Vector Quantization (NSVQ) (Vali and Bäckström, 2022) to avoid codebook collapse (Ye et al., 2024). NSVQ perturbs the quantized vector by noise scaled with the quantization error, which improves exploration of the codebook early in training and leads to more uniform code usage.

**Conditional reconstruction objective.** To ensure that $a_t$ and $\tilde{h}_t$ capture the control information needed to drive the procedure, LAT includes a conditional decoder $D_{\text{mot}}$ that reconstructs the next frame $\hat{x}_{t+1}$ from the current frame $x_t$ and the quantized action feature:

$$\hat{x}_{t+1} = D_{\text{mot}}(x_t, \tilde{h}_t).$$

The decoder operates in a spatial transformer architecture conditioned on $x_t$, so the action bottleneck is not required to store static anatomical content. The training objective combines a reconstruction term with standard codebook and commitment losses:

$$\mathcal{L}_{\text{LAT}} = \|x_{t+1} - \hat{x}_{t+1}\|_2^2 + \mathcal{L}_{\text{VQ}}(h_t, \tilde{h}_t), \tag{2}$$

where $\mathcal{L}_{\text{VQ}}$ follows the formulation in VQ-VAE (Razavi et al., 2019) with stop-gradient on the appropriate paths. This objective encourages the codebook to represent the surgical action that transforms the state at time $t$ into the state at time $t+1$.

**Macro- and micro-action regimes.** The codebook size $K$ governs the resolution of the discretized action space. We investigate two regimes representing distinct levels of abstraction in surgical dynamics. First, **Macro-Action Primitives ($K = 8$)** constrain the model to cluster continuous motion into a small set of prototypical dynamic states. In this regime, tokens represent broad categories of motion magnitude and direction rather than precise trajectories. While this yields a highly compressed bottleneck, it quantizes away the fine-grained kinematic variance required to distinguish subtle maneuvers within a single step. Second, **Micro-Action Primitives ($K = 1024$)** provide a high-capacity vocabulary designed to approximate the continuous control signal. This granularity allows the tokenizer to assign distinct codes to variations in instrument velocity, orientation, and local tissue deformation. For SurgWorld, we adopt the $K = 1024$ configuration to maximize the information content of the dynamics model, ensuring the predictor learns from the most expressive possible control signal.

### 2.2. Latent World Model

The Latent World Model in *SurgWorld* is trained on top of a fixed surgical vision encoder and uses the latent action primitives from LAT to model cataract surgery dynamics. We reuse the JHU-VPT (JEPA) encoder $E(\cdot)$ (Shah et al., 2025b), pretrained on unlabeled cataract videos, as a frozen state encoder. Given a clip $(x_t)_{t=1}^{T}$ sampled at 1 fps, each frame is mapped to a latent state representation $s_t = E(x_t)$. These embeddings summarize the surgical scene at time $t$, including tool configuration and ocular structures, but do not contain an explicit control variable.

**Action-conditioned state prediction.** The goal of the world model is to approximate one-step dynamics in this latent space conditioned on the inferred actions from LAT. For each time step $t$, we associate $s_t$ with the corresponding latent action primitive index $a_t$ and its embedding $\tilde{h}_t$ from Section 2.1. Before entering the predictor, we project the action embedding into the same dimensionality as the state via a learnable linear map $\text{Proj}(\cdot)$. The action-conditioned predictor $g_\phi$, implemented as a Transformer (Assran et al., 2025), receives the pair $(s_t, \text{Proj}(\tilde{h}_t))$ and outputs a prediction of the next latent state:

$$\hat{s}_{t+1} = g_\phi\big(s_t, \text{Proj}(\tilde{h}_t)\big).$$

The encoder $E(\cdot)$ is kept frozen; only the predictor parameters $\phi$ are optimized. Conditioning on the action embedding allows the model to distinguish between different future latent states that share similar appearance at time $t$ but correspond to different tool motions, and thus approximates $p(s_{t+1} \mid s_t, a_t)$ in feature space.

**Training objective.** We train the predictor with a feature prediction loss in the JHU-VPT (JEPA) latent space (Bardes et al., 2024). For each frame pair $(x_t, x_{t+1})$ in an unlabeled cataract video, we compute the target state $s_{t+1} = E(x_{t+1})$ and minimize the $L_1$ distance between $\hat{s}_{t+1}$ and $s_{t+1}$ with a stop-gradient on the target: $\mathcal{L}_{\text{world}} = \sum_t \left\| \hat{s}_{t+1} - \text{sg}\left(E(x_{t+1})\right) \right\|_1$. This teacher-forcing objective is analogous to the post-training procedure in V-JEPA2-AC (Assran et al., 2025), but uses latent action primitives instead of continuous robot kinematics. To improve robustness under rollout, we also include a short-horizon rollout term: starting from $s_t$ and a sequence of future actions from LAT, we unroll the predictor for two steps in latent space and penalize the deviation between the rolled-out state and $E(x_{t+2})$. The total loss $\mathcal{L}_{\text{WM}}$ is the sum of the teacher-forcing and rollout terms, and is minimized only with respect to $\phi$.

**Application to cataract surgery.** By predicting future encoder features instead of pixels, the world model learns how latent surgical states evolve under the inferred actions without modeling image formation (Bruce et al., 2024; Koju et al., 2025). In cataract surgery, this forces the representation to capture how tool motion and changes in intraocular anatomy (e.g., progression of capsulorhexis, nucleus fragmentation, cortical cleanup) follow from the sequence of latent action primitives, while ignoring nuisance variation such as illumination changes or fluid artifacts (Assran et al., 2025; Tivnan et al., 2024). At inference time, we use the sequence $(s_t)$ and the action-conditioned predictions $(\hat{s}_{t+1})$ as inputs to simple downstream heads for step recognition and step anticipation. This aligns the learned representation with the temporal structure of the procedure while preserving the benefits of the pretrained cataract foundation model.

### 2.3. Downstream Task Evaluation

To evaluate the quality of the learned representations, we freeze all *SurgWorld* components (the JHU-VPT (JEPA) encoder $E(\cdot)$, the LATr, and the world model predictor) and train a lightweight probe. This setup isolates the contribution of the self-supervised representations from the capacity of the downstream head. We compare three feature configurations to disentangle the contributions of visual context and dynamic control signals.

**State-only features.** As a baseline, we use the latent state sequence $s_{1:T}$ from the frozen JHU-VPT (JEPA) encoder, where $s_t = E(x_t)$. This represents a state-only model that captures static appearance and coarse temporal structure but does not use the inferred control signal from LAT, matching prior cataract foundation models (Shah et al., 2025b).

**Action-only features.** To quantify the information content of the discovered action primitives independent of visual context, we construct an action-only representation. For each time step $t$, we take the post-quantized action feature $\tilde{h}_t$ from LAT (Section 2.1) and project it to the model dimension via a linear map $\text{Proj}(\cdot)$. The resulting sequence $u_t^{\text{act}} = \text{Proj}(\tilde{h}_t)$ encodes dynamics without direct image features. Performance in this setting tests whether motion residuals alone contain sufficient semantic structure for step discrimination.

**State + action features.** To test whether LAT provides information that is complementary to the visual encoder, we form a joint representation. At each time step, we concatenate the state embedding and the projected action feature and fuse them via a learnable linear

layer $W_f$:

$$u_t^{\text{joint}} = W_f \left[ s_t \, ; \, \text{Proj}(\tilde{h}_t) \right], \tag{3}$$

where $[\cdot\,;\,\cdot]$ denotes concatenation. This yields a sequence $u_{1:T}^{\text{joint}}$ that combines appearance (from $s_t$) and inferred control (from LAT) in a unified feature stream. The gain of this configuration over state-only and action-only baselines quantifies the degree to which dynamics contribute non-redundant information.

**Attentive probing.** For all configurations (state-only, action-only, state+action), we use the same attentive probing head (Bardes et al., 2024). Given an input sequence $u_{1:T}$, we introduce a learnable query token $q$ and apply a cross-attention layer followed by a multilayer perceptron (MLP):

$$h = \text{MLP}\left(q + \text{CrossAttn}(q, u_{1:T})\right). \tag{4}$$

The vector $h$ is fed to a linear classifier trained with cross-entropy loss for step recognition. Only the probe parameters (cross-attention, MLP, classifier, and $W_f$) are updated. Comparing performance across the three feature configurations allows us to quantify (i) the intrinsic semantic value of the latent actions and (ii) the additional benefit of explicitly modeling dynamics alongside static appearance.

## 3. Experiments and Results

### 3.1. Datasets

**Pretraining.** For pretraining *SurgWorld* (LAT and the latent world model), we reuse the unlabeled cataract video corpus assembled in prior work (Shah et al., 2025b). This dataset comprises 1,838 internal microscope videos (average length ∼30 minutes at 59 *fps*) and 753 videos from Cataract-1k (Ghamsarian et al., 2024) (average length ∼8 minutes), for a total of 2,591 unique procedures. All videos are temporally subsampled to 1 *fps* and resized to $224 \times 224$ pixels. We do not use any Cataract-1k videos that carry step annotations for pretraining, to avoid label leakage into downstream evaluation. The JHU-VPT (JEPA) encoder (Shah et al., 2025b) is kept frozen during *SurgWorld* training; only LAT and the world model predictor are updated.

**Step recognition.** For step recognition, we evaluate on four cataract surgery datasets: Cataract-101 (Schoeffmann et al., 2018), D99 (Yu et al., 2019), Cataract-1k (annotated subset) (Ghamsarian et al., 2024), and an extended Cataract-1k split with additional internal annotations (Cataract-1k-JHU). Cataract-101 contains 101 videos with 10 annotated steps, and D99 contains 99 videos with 12 steps; for both we follow the standard train/validation/test splits used in prior work (Shah et al., 2023). For Cataract-1k and Cataract-1k-JHU, we adopt the splits described in (Shah et al., 2025b). All evaluation videos are subsampled to 1 *fps* and resized to $224 \times 224$ pixels for consistency with pretraining.

**Step anticipation.** For step anticipation, we derive future labels from the same datasets by shifting the step annotations forward in time. For each frame at time $t$ with step label $y_t$, we construct targets $y_{t+\Delta}$ for anticipation horizons $\Delta \in \{1, 3, 5, 10\}$ seconds, discarding frames near video boundaries where $t + \Delta$ is undefined. This yields matched recognition and

Table 1: **Step Recognition Accuracy** across varying labeled data regimes. We compare ***SurgWorld*** (Ours) against the state-only baseline **JHU-VPT** (Shah et al., 2025b) and **VideoMAE** (Tong et al., 2022), organized from label-scarce (10%) to full-data (100%) settings. Our approach consistently outperforms baselines, with particularly strong gains in low-data regimes (e.g., **+13.1%** on D99 and **+10.1%** on Cataract-1k-JHU at 10% split), demonstrating the sample efficiency of dynamic priors.

| Dataset | 10% Split | | | 25% Split | | | 50% Split | | | 100% Split | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | MAE | JEPA | Ours | MAE | JEPA | Ours | MAE | JEPA | Ours | MAE | JEPA | Ours |
| Cataract-1k | 36.61 | 35.12 | **38.21** | 46.91 | 45.09 | **50.59** | 59.65 | 58.80 | **70.65** | 63.75 | 79.58 | **89.56** |
| Cataract-1k-JHU | 58.36 | 63.81 | **73.90** | 63.93 | 74.55 | **80.76** | 66.18 | 80.71 | **84.51** | 70.03 | 83.65 | **85.39** |
| Cataract-101 | 58.60 | 56.95 | **61.12** | 70.64 | 79.73 | **82.01** | 72.41 | 84.79 | **88.21** | 79.31 | **89.82** | 89.75 |
| D99 | 42.10 | 45.56 | **58.70** | 47.56 | **63.21** | 61.91 | 51.10 | **71.51** | 69.24 | 66.13 | **77.20** | 76.85 |

anticipation splits, allowing us to study how explicit dynamics modeling affects both current-step classification and short-horizon forecasting. We report frame-level step recognition accuracy as our primary metric, and additionally provide Jaccard index, precision, and recall for comparison with prior work (Shah et al., 2023; Kim et al., 2019). For step anticipation, we measure accuracy at different prediction horizons.

### 3.2. Comparison to State-of-the-Art Cataract Pretraining Models

We compare *SurgWorld* to state-only pretraining baselines on step recognition across four cataract datasets and multiple label budgets using the same attentive probing protocol (Section 2.3), where the encoder and LAT remain frozen and only a lightweight probe is trained. As shown in Table 1, *SurgWorld* consistently outperforms VideoMAE and JHU-VPT (JEPA) across all datasets. In the low-data regime (10% labels), accuracy on D99 increases from 45.56% (JEPA) to 58.70% (Ours), and on Cataract-1k-JHU from 63.81% to 73.90%, corresponding to absolute gains of about 10%. Similar trends hold at 25% and 50% labels, indicating that latent action primitives provide dynamic information that is not captured by texture-based pretraining alone and help resolve ambiguities between visually similar frames corresponding to different steps.

Table 2 reports full fine-tuning results on Cataract-101 and D99. On Cataract-101, SurgWorld (Post-Quant) attains the highest accuracy (92.09%) and improves precision over both JHU-VPT (MAE) and JHU-VPT (JEPA). On D99, *SurgWorld* (Pre-Quant) achieves the best recall (66.15%) among all methods, suggesting that the pre-quantized action features retain fine-grained dynamics useful for detecting step transitions. Across attentive probing and full fine-tuning, these results support the main hypothesis of this work: explicitly modeling latent control through action primitives yields features that complement visual encoders and improve cataract step recognition in both label-scarce and fully supervised settings.

### 3.3. Step Anticipation

On Cataract-1k, we evaluate step anticipation at horizons of 1-10 seconds (Table 3) by training the same attentive probe to predict the step label $y_{t+\Delta}$ from features at time $t$. As the horizon increases, the state-only baseline JHU-VPT (JEPA) shows a gradual

Table 2: Quantitative results of step recognition on Cataract-101 and D99 datasets. We compare standard baselines against our state-only foundation model **JHU-VPT (JEPA)** and our proposed ***SurgWorld*** framework. By integrating latent action primitives, full fine-tuning *SurgWorld* consistently outperforms state-only baselines, demonstrating that explicitly modeling dynamics provides critical semantic information for surgical workflow analysis.

| Method | Cataract-101 | | | | D99 | | | |
|---|---|---|---|---|---|---|---|---|
| | Jaccard | Precision | Recall | Accuracy | Jaccard | Precision | Recall | Accuracy |
| ResNet (He et al., 2016) | 62.58 | 76.68 | 74.73 | 82.64 | 37.98 | 54.76 | 52.28 | 72.06 |
| SV-RCNet (Jin et al., 2017) | 66.51 | 84.96 | 76.61 | 86.13 | 39.15 | 58.18 | 54.25 | 73.39 |
| OHFM (Yi and Jiang, 2019) | 69.01 | 85.37 | 78.29 | 87.82 | 40.01 | 59.12 | 55.49 | 73.82 |
| TeCNO (Czempiel et al., 2020) | 70.18 | 86.03 | 79.52 | 88.26 | 41.31 | 61.56 | 55.81 | 74.07 |
| TMRNet (Jin et al., 2021) | 71.83 | 85.09 | 82.44 | 89.68 | 41.42 | 61.37 | 56.02 | 75.11 |
| Trans-SVNet (Gao et al., 2021) | 72.32 | 86.72 | 81.12 | 89.45 | 42.06 | 60.12 | 56.36 | 74.89 |
| ViT (Dosovitskiy et al.) | 64.77 | 78.51 | 75.62 | 84.56 | 38.18 | 55.15 | 53.60 | 72.45 |
| TimeSformer (Bertasius et al., 2021) | 75.97 | 85.38 | 84.47 | 90.76 | 42.69 | 64.24 | 55.17 | 77.83 |
| STMAE (Feichtenhofer et al., 2022) | 70.54 | 81.47 | 78.67 | 85.29 | 41.67 | 59.38 | 53.22 | 74.16 |
| VideoMAE (Tong et al., 2022) | 71.39 | 82.13 | 80.16 | 86.47 | 42.58 | 61.24 | 56.35 | 74.39 |
| CSMAE (Shah et al., 2025a) | 76.82 | 84.26 | 86.73 | 89.83 | 43.51 | *64.32* | 52.45 | 78.14 |
| JHU-VPT (MAE) | *79.95* | 87.80 | **89.10** | *92.00* | **49.95** | **64.78** | *64.46* | **78.69** |
| JHU-VPT (JEPA) | 79.58 | 87.88 | *88.89* | 91.52 | 43.63 | 55.39 | 62.19 | 75.61 |
| ***SurgWorld* (Pre-Quant)** | 79.10 | *88.13* | 87.89 | 91.17 | 45.48 | 57.38 | **66.15** | *76.31* |
| ***SurgWorld* (Post-Quant)** | **80.09** | **89.76** | 88.02 | **92.09** | *47.06* | 64.11 | 60.19 | 76.18 |

Table 3: **Step Anticipation Accuracy** on Cataract-1k. We evaluate the model's ability to forecast the surgical step $\Delta$ seconds into the future. While the baseline JHU-VPT (state-only) degrades as the horizon increases, ***SurgWorld*** maintains higher predictive accuracy, indicating that the learned world model captures additional temporal structure.

| **Method** | **Anticipation Horizon ($\Delta$)** | | | |
|---|---|---|---|---|
| | **1 sec** | **3 sec** | **5 sec** | **10 sec** |
| JHU-VPT (JEPA) (Shah et al., 2025b) | 67.06 | 64.42 | 61.88 | 59.15 |
| ***SurgWorld* (Ours)** | **69.50** | **66.08** | **64.73** | **62.05** |

drop in accuracy, reflecting the difficulty of forecasting future steps from appearance alone. *SurgWorld* provides consistent improvements of roughly 2–3% absolute across all horizons, indicating that incorporating latent action primitives adds useful temporal structure for near-future prediction. These anticipation results complement the larger gains observed in low-data step recognition, and together suggest that action-conditioned representations support both current-step classification and short-horizon forecasting.

### 3.4. Ablation: Action Granularity and Quantization

We study how the action vocabulary size $K$ and the choice of pre- versus post-quantized features affect step recognition. Table 4 reports results for LAT alone (action-only) and for SurgWorld (state + action) on Cataract-1k and Cataract-1k-JHU.

**Action-only representations.** For the LAT in isolation, increasing the vocabulary from Macro-Action Primitives ($K = 8$) to Micro-Action Primitives ($K = 1024$) yields substantial gains. On Cataract-1k, action-only accuracy with post-quantized features increases from

Table 4: Ablation of feature components and vocabulary size ($K$) on Cataract-1k datasets. Action-only performance indicates that motion primitives capture discriminative step information independent of visual context. The improvement of SurgWorld over state-only baselines demonstrates that dynamic control signals provide complementary information to static visual embeddings.

| Method | Config | Cataract-1k | | Cataract-1k-JHU | |
|---|---|---|---|---|---|
| | | Post-Q | Pre-Q | Post-Q | Pre-Q |
| ***Baselines (State-Only)*** | | | | | |
| VideoMAE (Tong et al., 2022) | - | 63.75 | | 70.03 | |
| JHU-VPT (JEPA) (Shah et al., 2025b) | - | 79.58 | | 83.65 | |
| ***Latent Action Tokenizer (Action-Only)*** | | | | | |
| LAT (Ours) | K=8 | 17.60 | 38.78 | 21.29 | 50.46 |
| | K=1024 | **40.69** | **46.40** | **44.60** | **54.62** |
| ***SurgWorld (State + Action)*** | | | | | |
| ***SurgWorld*** | K=8 | 90.03 | **90.13** | 84.47 | 84.63 |
| | K=1024 | **90.30** | 89.56 | **84.85** | **85.39** |

roughly 18% to 41%, with a similar improvement for pre-quantized features; Cataract-1k-JHU shows the same trend. These results indicate that a coarse codebook collapses distinct motion patterns into a few dynamic modes and loses step-discriminative detail, whereas a larger codebook can encode finer differences in tool velocity, orientation, and local tissue response. Pre-quantized features systematically outperform post-quantized ones, showing that discretization trades some dynamic fidelity for a compact representation, although the discrete tokens still support non-trivial step recognition without any visual input.

**State + action fusion.** When we fuse latent actions with JHU-VPT (JEPA) state embeddings in *SurgWorld*, the effect of $K$ becomes much smaller. On Cataract-1k, *SurgWorld* attains around 90% accuracy across all ($K$, Pre/Post) settings, compared to 79.58% for state-only JEPA; on Cataract-1k-JHU, the joint model improves JEPA from 83.65% to approximately 85%. Both Macro-Action and Micro-Action configurations perform similarly once combined with a strong visual encoder, suggesting that latent actions provide information that is largely complementary in the action-only regime but partially redundant with the state trajectory when fused with high-capacity visual features. In all SurgWorld experiments, we adopt $K = 1024$ Micro-Action Primitives and use the continuous pre-quantized features to condition the world model.

## 4. Conclusion

We introduced *SurgWorld*, an action-conditioned world model for cataract surgery that learns surgical dynamics from unlabeled video using latent action tokens and latent-state prediction on a frozen foundation encoder. Across four datasets, *SurgWorld* consistently improves step recognition, with the largest gains in label-scarce settings, demonstrating strong sample efficiency. Action-only features are discriminative, and their fusion with JEPA features yields the best performance, confirming the complementary nature of latent dynamics and static appearance. Step anticipation gains over JHU-VPT further indicate that *SurgWorld* captures meaningful future surgical progression.

## Acknowledgments

The authors are supported by grants from the National Institutes of Health, U.S.A.; NIH 1R01EY033065 and NIH 1R01EB038734. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. Also, we would like to thank the Johns Hopkins Research IT team in IT@JH for their support and infrastructure resources, where some of these analyses were conducted, especially DISCOVERY HPC. Their commitment to advancing research has been invaluable in the successful completion of this study.

## References

Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai, Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, et al. Cosmos world foundation model platform for physical ai. *arXiv preprint arXiv:2501.03575*, 2025.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

Wele Gedara Chaminda Bandara, Naman Patel, Ali Gholami, Mehdi Nikkhah, Motilal Agrawal, and Vishal M Patel. Adamae: Adaptive masking for efficient spatiotemporal learning with masked autoencoders. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14507–14517, 2023.

Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mahmoud Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *arXiv preprint arXiv:2404.08471*, 2024.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, volume 2, page 4, 2021.

Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *Forty-first International Conference on Machine Learning*, 2024.

Alba Centeno López, Ángela González-Cebrián, Igor Paredes, Alfonso Lagares, and Paula de Toledo. Improving surgical phase recognition using self-supervised deep learning. *Scientific Reports*, 15(1):39087, 2025.

Zhen Chen, Qing Xu, Jinlin Wu, Biao Yang, Yuhao Zhai, Geng Guo, Jing Zhang, Yinlu Ding, Nassir Navab, and Jiebo Luo. How far are surgeons from surgical world models? a pilot study on zero-shot surgical video generation with expert assessment. *arXiv preprint arXiv:2511.01775*, 2025.

Tobias Czempiel, Magdalini Paschali, Matthias Keicher, Walter Simson, Hubertus Feussner, Seong Tae Kim, and Nassir Navab. Tecno: Surgical phase recognition with multi-stage temporal convolutional networks. In *MICCAI 2020*, pages 343–352. Springer, 2020.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. The epic-kitchens dataset: Collection, challenges and baselines. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(11):4125–4141, 2020.

Jingtao Ding, Yunke Zhang, Yu Shang, Yuheng Zhang, Zefang Zong, Jie Feng, Yuan Yuan, Hongyuan Su, Nian Li, Nicholas Sukiennik, et al. Understanding world or predicting future? a comprehensive survey of world models. *ACM Computing Surveys*, 58(3):1–38, 2025.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Christoph Feichtenhofer, Yanghao Li, Kaiming He, et al. Masked autoencoders as spatiotemporal learners. *Advances in neural information processing systems*, 35:35946–35958, 2022.

Isabel Funke, Sören Torge Mees, Jürgen Weitz, and Stefanie Speidel. Video-based surgical skill assessment using 3d convolutional neural networks. *IJCARS*, 2019.

Xiaojie Gao, Yueming Jin, Yonghao Long, Qi Dou, and Pheng-Ann Heng. Trans-svnet: Accurate phase recognition from surgical videos via hybrid embedding aggregation transformer. In *MICCAI 2021*, pages 593–603. Springer, 2021.

Negin Ghamsarian, Yosuf El-Shabrawi, Sahar Nasirihaghighi, Doris Putzgruber-Adamitsch, Martin Zinkernagel, Sebastian Wolf, Klaus Schoeffmann, and Raphael Sznitman. Cataract-1k dataset for deep-learning-assisted analysis of cataract surgery videos. *Scientific data*, 11(1):373, 2024.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2(3), 2018.

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*, 2019.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Yueming Jin, Qi Dou, Hao Chen, Lequan Yu, Jing Qin, Chi-Wing Fu, and Pheng-Ann Heng. Sv-rcnet: workflow recognition from surgical videos using recurrent convolutional network. *IEEE transactions on medical imaging*, 37(5):1114–1126, 2017.

Yueming Jin, Yonghao Long, Cheng Chen, Zixu Zhao, Qi Dou, and Pheng-Ann Heng. Temporal memory relation network for workflow recognition from surgical video. *IEEE Transactions on Medical Imaging*, 40(7):1911–1923, 2021.

Seunghoi Kim, Henry FJ Tregidgo, Matteo Figini, Chen Jin, Sarang Joshi, and Daniel C Alexander. Tackling hallucination from conditional models for medical image reconstruction with dynamicdps. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 593–603. Springer, 2025a.

Seunghoi Kim, Henry FJ Tregidgo, Chen Jin, Matteo Figini, and Daniel C Alexander. Hallugen: Synthesizing realistic and controllable hallucinations for evaluating image restoration. *arXiv preprint arXiv:2512.03345*, 2025b.

Tae Soo Kim, Molly O'Brien, Sidra Zafar, Gregory D Hager, Shameema Sikder, and S Swaroop Vedula. Objective assessment of intraoperative technical skill in capsulorhexis using videos of cataract surgery. *International journal of computer assisted radiology and surgery*, 14(6):1097–1105, 2019.

Saurabh Koju, Saurav Bastola, Prashant Shrestha, Sanskar Amgain, Yash Raj Shrestha, Rudra PK Poudel, and Binod Bhattarai. Surgical vision world model. In *MICCAI Workshop on Data Engineering in Medical Imaging*, pages 1–10. Springer, 2025.

Yann LeCun. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1):1–62, 2022.

Gurvan Lecuyer, Martin Ragot, Nicolas Martin, Laurent Launay, and Pierre Jannin. Assisted phase and step annotation for surgical videos. *IJCARS*, 15:673–680, 2020.

Lena Maier-Hein, Swaroop S Vedula, Stefanie Speidel, Nassir Navab, Ron Kikinis, Adrian Park, Matthias Eisenmann, Hubertus Feussner, Germain Forestier, Stamatia Giannarou, et al. Surgical data science for next-generation interventions. *Nature Biomedical Engineering*, 1(9):691–696, 2017.

Nicolas Padoy. Machine and deep learning for workflow recognition during surgery. *Minimally Invasive Therapy & Allied Technologies*, 28(2):82–90, 2019.

Ali Razavi, Aaron Van den Oord, and Oriol Vinyals. Generating diverse high-fidelity images with vq-vae-2. *Advances in neural information processing systems*, 32, 2019.

Klaus Schoeffmann, Mario Taschwer, Stephanie Sarny, Bernd Münzer, Manfred Jürgen Primus, and Doris Putzgruber. Cataract-101: video dataset of 101 cataract surgeries. In *Proceedings of the 9th ACM multimedia systems conference*, pages 421–425, 2018.

Nisarg A Shah, Shameema Sikder, S Swaroop Vedula, and Vishal M Patel. Glsformer: Gated-long, short sequence transformer for step recognition in surgical videos. In *MICCAI*, 2023.

Nisarg A. Shah, Chaminda Bandara, Shameema Skider, S. Swaroop Vedula, and Vishal M. Patel. CSMAE: Cataract surgical masked autoencoder (MAE) based pre-training. In *Proceedings of the International Symposium on Biomedical Imaging (ISBI)*, 2025a.

Nisarg A Shah, Mingze Xia, Subhasri Vijay, Shameema Sikder, S Swaroop Vedula, and Vishal M Patel. A vision foundation model for cataract surgery using joint-embedding predictive architecture. In *Medical Imaging with Deep Learning*, 2025b.

Matthew Tivnan, Siyeop Yoon, Zhennong Chen, Xiang Li, Dufan Wu, and Quanzheng Li. Hallucination index: An image quality metric for generative reconstruction models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 449–458. Springer, 2024.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022.

Mohammad Hassan Vali and Tom Bäckström. Nsvq: Noise substitution in vector quantization for machine learning. *IEEE Access*, 10:13598–13610, 2022.

Aaron Van Den Oord, Oriol Vinyals, et al. Neural discrete representation learning. *Advances in neural information processing systems*, 30, 2017.

Ruben Villegas, Mohammad Babaeizadeh, Pieter-Jan Kindermans, Hernan Moraldo, Han Zhang, Mohammad Taghi Saffar, Santiago Castro, Julius Kunze, and Dumitru Erhan. Phenaki: Variable length video generation from open domain textual description. *arXiv preprint arXiv:2210.02399*, 2022.

Mengjiao Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. *arXiv preprint arXiv:2310.06114*, 1(2):6, 2023.

Shu Yang, Fengtao Zhou, Leon Mayer, Fuxiang Huang, Yiliang Chen, Yihui Wang, Sunan He, Yuxiang Nie, Xi Wang, Ömer Sümer, et al. Large-scale self-supervised video foundation model for intelligent surgery. *arXiv preprint arXiv:2506.02692*, 2025.

Seonghyeon Ye, Joel Jang, Byeongguk Jeon, Sejune Joo, Jianwei Yang, Baolin Peng, Ajay Mandlekar, Reuben Tan, Yu-Wei Chao, Bill Yuchen Lin, et al. Latent action pretraining from videos. *arXiv preprint arXiv:2410.11758*, 2024.

Fangqiu Yi and Tingting Jiang. Hard frame detection and online mapping for surgical phase recognition. In *MICCAI 2019*, pages 449–457. Springer, 2019.

Felix Yu, Gianluca Silva Croso, Tae Soo Kim, Ziang Song, Felix Parker, Gregory D Hager, Austin Reiter, S Swaroop Vedula, Haider Ali, and Shameema Sikder. Assessment of automated identification of phases in videos of cataract surgery using machine learning and deep learning techniques. *JAMA network open*, 2(4):e191860–e191860, 2019.

Guangyao Zhai, Xingyuan Zhang, and Nassir Navab. Recurrent world model with tokenized latent states. In *ICLR 2025 Workshop on World Models: Understanding, Modelling and Scaling*.