A Vision Foundation Model for Cataract Surgery Using Joint-Embedding Predictive Architecture

Nisarg A. Shah[*1] SNISARG812@GMAIL.COM
[1] *Johns Hopkins University, Baltimore, USA*
Mingze Xia[*1] MXIA8@JHU.EDU
Subhasri Vijay[1] SVIJAY2@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 2025

Abstract

Vision foundation models can automate analysis of surgical videos and enable multiple applications that support patient care and surgical training. For cataract surgery, existing models are limited by reliance on small datasets, privacy concerns, and poor generalizability across surgical settings. In this paper, we introduce JHU-VPT(JEPA), a self-supervised vision foundation model leveraging Joint-Embedding Predictive Architecture (JEPA) to learn spatiotemporal representations via latent feature prediction on a large corpus of unlabeled cataract videos, without requiring extensive labeled datasets or pixel-level reconstruction. JHU-VPT(JEPA) is pretrained on 2591 videos from multiple sites that capture different surgical technique and style variations. Comprehensive evaluations on step recognition, surgical feedback, and skill assessment tasks demonstrate that JHU-VPT(JEPA) outperforms existing methods. JHU-VPT(JEPA)'s effectiveness is evident even when using attentive probing with a frozen encoder, highlighting the robustness of the learned features and addressing privacy concerns by not requiring access to raw videos during downstream tasks. Our approach offers a scalable, generalizable, and privacy-preserving solution for surgical video analysis, with significant potential to advance patient care and surgical education.

Keywords: Surgical Pretraining, Joint Embedding Predictive Network, Cataract Surgery

1. Introduction

Vision foundation models to analyze videos of the surgical field can have a substantial global impact on patient care. Intraoperative videos of the surgical field are a rich source of data for algorithms that can enable several critical applications such as activity recognition for situation awareness, skill and feedback prediction for supporting surgeons' learning and evaluation, among others (Maier-Hein et al., 2017; Yu et al., 2019; Padoy, 2019). The emergence of surgical data science has accelerated models for analyzing videos of the surgical field. However, the state-of-the-art models have several constraints including small datasets from convenience samples (Shah et al., 2023), limited evaluation on a few applications, and models that lack generalizability in new datasets (Lecuyer et al., 2020; Padoy, 2019; Funke

[*] Contributed equally

et al., 2019). While foundation models are rapidly being trained for applications in several domains (Kang et al., 2023; Lu et al., 2023), vision foundation models by pretraining on large surgical video datasets have not yet been developed.

Self-supervised learning (SSL) has emerged as a powerful paradigm to leverage large corpora of unlabeled video data and train vision foundation models. Traditional SSL methods for medical imaging often involve multimodal cues, e.g., textual radiology reports paired with X-ray images (Boecking et al., 2022; Moon et al., 2022). By contrast, surgical videos typically lack accompanying text annotations, necessitating visual self-supervised schemes. To address the lack of granular text annotations, we propose a new self-supervised approach tailored to the complexity of spatio-temporal information across the surgical videos. Building on the Joint-Embedding Predictive Architecture (JEPA) (Assran et al., 2023), our method focuses on *feature prediction* in latent space, a strategy that captures both spatio-temporal coherence and surgical scene semantics without requiring direct pixel-level reconstruction.

Unlike prior self-supervised strategies that primarily rely on contrastive learning or masked autoencoders (MAEs) (He et al., 2022; Tong et al., 2022), our JEPA-based approach operates in the latent feature domain, reducing the overhead of reconstructing pixel details that may be irrelevant for clinical applications. We develop our model, *JHU-VPT(JEPA):Cataract*, which we refer to as JHU-VPT(JEPA), by pretraining on a large corpus of cataract surgery videos including multiple sites and surgeons. The dataset diversity allows learning of domain-robust embeddings. The resultant representations can be shared more readily than raw videos (protecting patient privacy), and they excel in label-scarce scenarios, reducing the need for extensive manual annotations and data-hungry fine-tuning protocols. We comprehensively evaluate JHU-VPT(JEPA)'s learned embeddings on three key tasks: (1) **Step Recognition**, wherein the aim is to identify surgical steps or phases; (2) **Surgical Feedback**, to predict specific performance feedback for the surgeon; and (3) **Skill Assessment**, which is essential for both surgeon training and credentialing. By varying the size of the annotated subsets used for fine-tuning, we show that JHU-VPT(JEPA) achieves strong performance even with limited labels, highlighting its data efficiency. Furthermore, we validate cross-domain generalization by testing on previously unseen videos, demonstrating JHU-VPT(JEPA)'s capacity to adapt to new surgical styles or camera configurations.

Contributions. In summary, our main contributions are:

- **JEPA-based approach for cataract videos.** We introduce JHU-VPT(JEPA) with a novel architecture for surgical video analysis that employs feature prediction in latent space to learn rich spatio-temporal representations.These representations are validated via attentive probing with a frozen encoder, confirming their high transferability and effectiveness in downstream tasks without fine-tuning.

- **Large-scale video pretraining using a large dataset.** We use an extensive dataset of unlabeled surgical videos from multiple institutions and surgeons, ensuring robust, domain-generalizable embeddings.

- **Comprehensive downstream evaluation.** We test JHU-VPT(JEPA) on three important tasks—step recognition, surgical feedback, and skill assessment—and show notable gains under varying amounts of labeled data, underscoring its potential clinical utility.

2. JHU-VPT(JEPA): Cataract model

In this section, we describe our proposed *Vision Foundation Model for Cataract Surgery* **JHU-VPT(JEPA)**, which builds upon the JEPA principle (Garrido et al., 2024; Bardes et al., 2024) for learning rich, robust visual representations from cataract surgery videos. Our goal is to exploit *feature prediction* as a stand-alone objective, enabling the model to learn meaningful spatio-temporal embeddings without extra supervision. A high-level overview of JHU-VPT(JEPA) is shown in Figure 1.

2.1. Overview

At the core of feature prediction as a stand-alone objective, the model learns by predicting the representation of a target input $\mathbf{y}$ from the representation of a context input $\mathbf{x}$. Specifically, an encoder $E_\psi(\cdot)$ projects $\mathbf{x}$ into latent space, while a predictor $P_\phi(\cdot)$ attempts to recover the embedding of $\mathbf{y}$ given $\mathbf{x}$. A conditioning variable δ, indicating the transformation or corruption that links $\mathbf{x}$ and $\mathbf{y}$, guides the predictor to generate distinct outputs for different transformations. In our setting, $\mathbf{x}$ and $\mathbf{y}$ are disjoint spatio-temporal patches from a surgical clip, and δ encodes the masking pattern (or offset) between these two regions.

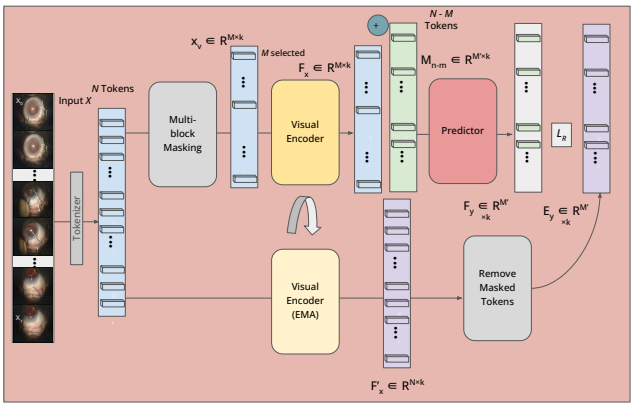

Figure 1: Overview of the JHU-VPT(JEPA) architecture. The framework consists of Block Masking, an Encoder, a Predictor, and an EMA-updated Target Encoder. The Encoder processes the non-masked tokens, predicting their feature representations. The Predictor combines these representations with learnable mask tokens and a conditioning variable to predict the embeddings of masked regions. The Target Encoder encodes all tokens, generating target embeddings for the feature-prediction loss.

2.2. Training Objective

To learn robust representations, we train the visual encoder $E_\psi(\cdot)$ and the predictor $P_\phi(\cdot)$ via a feature-prediction loss. Let *context* region $\mathbf{x}$ and *target* region $\mathbf{y}$ be two non-overlapping subsets of video tokens from a video $\mathbf{X}$, selected according to a masking scheme (see Section 2.3). We define the loss function to encourage the predicted representation of $\mathbf{y}$ to

match the actual representation of $\mathbf{y}$, generated by a *target encoder* $E_{\bar{\psi}}(\cdot)$. Concretely, we minimize:

$$\min_{\psi,\phi} \left\| P_\phi\big(E_\psi(\mathbf{x}),\, \delta\big) \;-\; \mathrm{sg}\big(E_{\bar{\psi}}(\mathbf{y})\big) \right\|_1, \tag{1}$$

where $\mathrm{sg}(\cdot)$ is a stop-gradient blocking updates to $E_{\bar{\psi}}(\cdot)$. In practice, $\bar{\psi}$ is maintained as an exponential moving average (EMA) of ψ, consistent with prior work that mitigates representation collapse (Garrido et al., 2024). Using L1 loss and a stop-gradient on the target encoder prevents trivial solutions(i.e. feature collapse) by forcing the encoder and predictor to capture meaningful spatio-temporal information in the surgical video.

Collapse Prevention. Combining an EMA target encoder, a stop-gradient, and a predictor prevents representation collapse in various self-supervised contexts (Grill et al., 2020; Assran et al., 2023). Intuitively, $\bar{\psi}$ changes more slowly than ψ, compelling $E_\psi(\mathbf{x})$ to capture detailed information needed by $P_\phi(\cdot)$ to match the slowly evolving target representation. This strategy drives the encoder to encode distinct semantic cues (e.g., instruments, ocular structures, movements) rather than collapsing to constant outputs.

2.3. Prediction Task and Masking Strategy

We implement the feature-prediction objective using a masked modeling approach. Each video clip is partitioned into 3D tokens, and large continuous blocks are sampled to form the masked regions $\mathbf{y}$; the remaining tokens constitute the visible regions $\mathbf{x}$. Applying large or continuous masks across time creates a challenging prediction task, encouraging the model to capture dynamic interactions between surgical instruments and ocular tissue.

To achieve this, we use multi-block masking (Bardes et al., 2024). First, short-range masks involve sampling several small blocks (e.g., 8) that cover about 15% of each frame, applied consistently across all frames. This forces the model to rely on temporal cues to infer fine-grained details and quick instrument movements. Second, long-range masks involve sampling fewer, larger blocks (e.g., 2) covering approximately 70% of each frame and extending over time, forcing the model to understand broader surgical phases and slower eye changes from limited visible areas. This multi-block masking strategy challenges the predictor to reconstruct features of large masked regions from small visible segments, enhancing the model's understanding of actions and anatomy in surgery videos.

2.4. Implementation Details

JHU-VPT(JEPA) comprises three learnable modules and an EMA-updated target encoder.
Tokenizer: The tokenizer converts the raw video $\mathbf{X} \in \mathbb{R}^{T\times C\times H\times W}$ into non-overlapping 3D tokens representing spatio-temporal volumes. We apply a 3D convolutional layer with kernel and stride (t, h, w), producing $N = \frac{T}{t} \times \frac{H}{h} \times \frac{W}{w}$ tokens, each of dimension k. Fixed 3D positional encodings (He et al., 2022; Tong et al., 2022) are added to retain spatio-temporal information.
Encoder $E_\psi(\cdot)$: The encoder is a Vision Transformer (ViT) backbone (Dosovitskiy et al.; Arnab et al., 2021) that processes the visible tokens $\mathbf{x}$, producing an embedding $\mathbf{F}_x \in \mathbb{R}^{|\mathbf{x}|\times d}$, where d is the embedding dimension. This embedding is passed to the predictor.
Predictor $P_\phi(\cdot)$: The predictor is a lightweight transformer that maps $\mathbf{F}_x$ to a predicted embedding $\widetilde{\mathbf{F}}_y$. It also receives learnable mask tokens $\mathbf{M}$ (one per masked patch) with

positional encodings and the conditioning variable δ, which encodes positional offsets or transformations between $\mathbf{x}$ and $\mathbf{y}$. Formally,

$$\widetilde{\mathbf{F}}_y = P_\phi\big(\mathbf{F}_x,\, \mathbf{M},\, \delta\big). \tag{2}$$

Target Encoder $E_{\bar{\psi}}(\cdot)$: The target encoder is an EMA copy of the encoder, updated at each training iteration by

$$\bar{\psi} \leftarrow \alpha\, \bar{\psi} + (1-\alpha)\, \psi, \tag{3}$$

where $\alpha \in [0, 1)$ is a momentum coefficient. It processes the masked tokens $\mathbf{y}$, generating $\mathbf{F}_y$ for the loss in Eq. (1).

2.5. Pretraining Architecture Analysis

By predicting the representations of large, masked video regions from limited visible cues, JHU-VPT(JEPA) captures both fine-grained details and long-range context inherent to cataract surgery workflows. The joint-embedding mechanism directs the encoder to focus on discriminative aspects such as surgical instruments, subtle eye movements, and relevant clinical steps. The combined effect of the EMA target encoder, stop-gradient mechanism, and predictor network prevents representation collapse, enabling the learning of temporally coherent and anatomically relevant features. This design is scalable to various downstream tasks, including surgical phase recognition and skill assessment, and demonstrates strong generalization with minimal labeled data. In Section 3.3, we demonstrate JHU-VPT(JEPA)'s effectiveness in capturing the complexities of real-world surgical workflows while maintaining low annotation requirements

2.6. Downstream Task Evaluation

After pretraining JHU-VPT(JEPA), we evaluate its representations on downstream tasks using two approaches: *fine-tuning* and *attentive probing*. In fine-tuning, we initialize the encoder $E_\psi(\cdot)$ with the pretrained weights and attach a linear classification head. The entire model, including the encoder and the classification head, is then optimized jointly on the downstream dataset.

In contrast, attentive probing keeps the pretrained encoder $E_{\bar{\psi}}(\cdot)$ fixed to assess the quality of the learned features without updating them. We introduce a learnable cross-attention layer with a query token that attends to the output features of the frozen encoder. The output of the cross-attention layer is added to the query token via a residual connection and passed through a two-layer multilayer perceptron (MLP) for prediction:

$$\mathbf{h} = \mathrm{MLP}\big(\mathbf{q} + \mathrm{CrossAttn}(\mathbf{q},\, E_{\bar{\psi}}(\mathbf{x}))\big), \tag{4}$$

where $\mathbf{q}$ is the learnable query token, and $\mathbf{h}$ is the output used for classification or regression tasks. Attentive probing evaluates the robustness of the pretrained features while keeping the feature extractor unchanged, ensuring that the representation quality is not influenced by further training. This approach is useful when labeled data is limited or when data privacy restrictions prevent sharing raw videos, as it allows training downstream models on new tasks using shared features without accessing the raw video data.

2.7. Datasets

For **Pretraining** JHU-VPT(JEPA), we assembled a multi-institutional dataset of 2,591 unlabeled cataract surgery videos. This dataset comprises 1,838 internal videos averaging 30 minutes at 59 *fps*, and 753 videos from Cataract-1k (Ghamsarian et al., 2024) averaging 8 minutes, with total of 2591 unique videos. We did not pretrain on the Cataract-1k videos for which step recognition annotations were provided. All videos were subsampled to 1 *fps* and resized to 250×250 pixels for pretraining, following prior protocols (Gao et al., 2021; Twinanda et al., 2016). CSMAE (Shah et al., 2025) was pretrained on the D-450 dataset (an extension of D99 videos), following methodology for MAE-based pretraining (Bandara et al., 2023).

We evaluated JHU-VPT(JEPA) on three downstream tasks: step recognition, surgical feedback, and skill assessment. For **step recognition**, experiments were conducted under both low-data (10%, 25%, 50%) and full-data settings using four cataract surgery datasets: Cataract-101 (Schoeffmann et al., 2018), D99 (Yu et al., 2019), Cataract-1k (subset for which annotations were provided with the original dataset) (Ghamsarian et al., 2024), and a larger subset of Cataract-1k which we internally annotated (referred to as Cataract-1k-JHU and includes the annotated videos in the original dataset). Cataract-101 contains 101 videos at 25 *fps* with 10 annotated steps and a resolution of 720×540 pixels, split into 50 training, 10 validation, and 40 testing videos (Shah et al., 2023). D99 comprises 99 videos at 59 *fps* with 12 annotated steps and resolution 640×480 pixels, partitioned into 60 training, 20 validation, and 19 testing videos (Shah et al., 2023). For Cataract-1k, we used 25 training, 7 validation, and 24 testing videos. For Cataract-1k-JHU, we employed 181 training, 31 validation, and 91 testing videos. All evaluation videos were subsampled to 1 *fps* and resized to 250×250 pixels for consistency.

In the **surgical feedback** task, we evaluated JHU-VPT(JEPA) on feedback items (Xia et al., 2025) during the capsulorhexis step using the D99 dataset (Hira et al., 2022) of 99 videos. Frames were resized to 224×224 pixels, applying data augmentations like rotation and color jitter. Data was split into training (60%), validation (20%), and testing (20%) sets, we repeated experiments with three random splits and averaged the results.

For **skill assessment** in the main incision and capsulorhexis steps, we used 56 videos from D99 and an additional 37 videos captured under consistent conditions. Expert surgeons evaluated the videos using ICO-OSCAR:Phacoemulsification (Puri et al., 2017). Skill was categorized as *novice* (scores 2–4) and *expert* (score 5) for main incision and *novice* and *expert* for capsulorhexis following (Hira et al., 2022; Kim et al., 2019).

3. Experiments and Results

3.1. Evaluation Metrics

We evaluate JHU-VPT (JEPA) using Accuracy, Precision, Recall, and Jaccard Index for step recognition (Shah et al., 2023; Kim et al., 2019), and Accuracy, Sensitivity, Specificity, and AUC for surgical feedback and skill assessment.

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

We compare JHU-VPT(JEPA) with existing pretraining models on the Cataract-101, D99, Cataract-1k, and Cataract-1k-JHU datasets. As shown in Table 1, in the attentive probing setup, where the encoder remains frozen during downstream evaluation, JHU-VPT(JEPA) consistently outperforms VideoMAE (Tong et al., 2022)(pretrained on same pretraining set) across all data splits and all datasets. For example, on Cataract-1k with 10% training data, JHU-VPT(JEPA) attains an accuracy of 35.12%, surpassing VideoMAE's 20.70% by 14.42 percentage points—a relative improvement of approximately 70%. These results indicate that the features learned by JHU-VPT(JEPA) are more robust and generalizable, effectively capturing important surgical patterns without updating the feature extractor during downstream tasks.

Table 1: Comparison of Step Recognition Accuracy across different dataset splits. These results are based on Attentive Probe experiments, showing that our model, JHU-VPT(JEPA) consistently outperforms VideoMAE (Tong et al., 2022), JHU-VPT(MAE) across all settings when pretrained on the same pretraining set (D-2591).

Dataset	10% Split		25% Split		50% Split		100% Split	
	VideoMAE	Ours	VideoMAE	Ours	VideoMAE	Ours	VideoMAE	Ours
Cataract-101	29.87	**56.95**	43.14	**79.73**	57.13	**84.79**	65.49	**89.82**
Cataract-1k	20.70	**35.12**	32.12	**45.09**	46.04	**58.80**	52.26	**79.58**
D99	20.14	**45.56**	28.71	**63.21**	32.21	**71.51**	40.60	**77.20**
Cataract-1k-JHU	43.77	**63.81**	52.69	**74.55**	55.24	**80.71**	58.76	**83.65**

In full fine-tuning experiments (Tab. 2, Figure 2), JHU-VPT(JEPA) shows substantial improvements over CSMAE (Shah et al., 2025), despite CSMAE employing advanced sampling strategies (Bandara et al., 2023). Pretraining on a larger and more diverse dataset enhances JHU-VPT(JEPA)'s performance, emphasizing the critical role of data diversity in self-supervised learning for surgical video analysis. While JHU-VPT(MAE), VideoMAE model pretrained on our pretraining dataset, often achieves higher accuracy under full fine-tuning (Feichtenhofer et al., 2022), JHU-VPT(JEPA) is pretrained with a predictive objective that emphasizes the learning of abstract, high-level representations. These robust, generalizable features perform better when evaluated using attention probes without updating all network parameters showcasing robustness of pretraining features. In contrast, full fine-tuning adjusts every parameter, which can perturb the delicate representations and lower the performance compared to models optimized for end-to-end updates.

Overall, JHU-VPT(JEPA)'s feature prediction approach, enabled by large and diverse pretraining data, yields significant performance gains in cataract surgery analysis. Its robust features allow surgical video analysis in privacy-constrained scenarios with minimal fine-tuning (Garrido et al., 2024).

3.3. Comparison on Feedback and Skill Performance

Table 3(a) shows that our method improves feedback prediction by 10% in AUC. Compared to other methods, JHU-VPT(JEPA) improves specificity, i.e., reduces false positives, indi-

Table 2: Quantitative results of step recognition from different methods on the Cataract-101 and D99 datasets.

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., 2025)	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

cating that the model has meaningful discrimination between positive and negative labels. Table 3(b,c) demonstrates our model performance on skill assessment for main incision and capsulorhexis. We observe steady improvement of 10%-20% for both phases, which highlights the robustness of its learned feature representations across various phases.

Table 3: Model evaluation for predicting feedback items, skill assessment in main incision, and skill assessment in capsulorhexis.

Feedback Prediction (Table 3a)				
Model	**Accuracy**	**Sensitivity**	**Specificity**	**AUC**
CNN-LSTM (Wan et al., 2024)	76.3 ± 1.6	$\mathbf{94.3 \pm 1.5}$	15.3 ± 1.7	0.659 ± 0.049
CNN-LSTM-GNN (Xia et al., 2025)	75.0 ± 1.1	85.6 ± 2.6	34.5 ± 6.7	0.559 ± 0.048
JHU-VPT(MAE) (D-2591)	80.4 ± 2.9	93.1 ± 8.0	35.9 ± 8.2	0.817 ± 0.032
TimeSformer (Bertasius et al., 2021)	77.2 ± 1.1	85.7 ± 5.9	40.2 ± 12.4	0.710 ± 0.066
JHU-VPT(JEPA) (Ours)	$\mathbf{82.3 \pm 1.4}$	92.6 ± 6.9	$\mathbf{40.8 \pm 16.5}$	$\mathbf{0.842 \pm 0.045}$
Main Incision Skill Assessment (Table 3b)				
Model	**Accuracy**	**Sensitivity**	**Specificity**	**AUC**
CNN-LSTM (Hira et al., 2022)	63.0	92.0	36.0	0.64
ViT (Dosovitskiy et al.)	62.0	10.0	100.0	0.55
JHU-VPT(JEPA) (Ours)	73.0	60.0	63.0	0.72
Capsulorhexis Skill Assessment (Table 3c)				
Model	**Accuracy**	**Sensitivity**	**Specificity**	**AUC**
ResNet-101 (He et al., 2016)	62.0	76.0	80.0	0.45
STMAE (Feichtenhofer et al., 2022)	66.0	68.0	80.0	0.55
JHU-VPT(MAE) (D-2591)	71.0	85.0	90.0	0.55
JHU-VPT(JEPA) (D-2591)	80.0	70.0	56.25	0.80

4. Conclusion

We introduced JHU-VPT(JEPA) for cataract surgery video analysis. By leveraging feature prediction in the latent space, JHU-VPT(JEPA) captures rich spatio-temporal representations without dependence on pixel-level reconstruction or large amounts of labeled data. It allows clinical use of the model while preserving patient privacy. While JHU-VPT (JEPA) shows strong performance, further improvements can be achieved by increasing temporal resolution and incorporating finer-grained motion features for feedback prediction and skill assessment. Future work may explore domain adaptation techniques to improve generalization across different surgical environments.

Acknowledgments

This research was supported by a grant from the National Institutes of Health, USA; R01EY033065. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6836–6846, 2021.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15619–15629, 2023.

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.

Benedikt Boecking, Naoto Usuyama, et al. Making the most of text semantics to improve biomedical vision–language processing. In *ECCV*, pages 1–21. Springer, 2022.

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.

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.

Quentin Garrido, Mahmoud Assran, Nicolas Ballas, Adrien Bardes, Laurent Najman, and Yann LeCun. Learning and leveraging world models in visual representation learning. *arXiv preprint arXiv:2403.00504*, 2024.

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.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

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.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

Sanchit Hira, Digvijay Singh, Tae Soo Kim, Shobhit Gupta, Gregory Hager, Shameema Sikder, and S Swaroop Vedula. Video-based assessment of intraoperative surgical skill. *International journal of computer assisted radiology and surgery*, 17(10):1801–1811, 2022.

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.

Qingbo Kang, Jun Gao, Kang Li, and Qicheng Lao. Deblurring masked autoencoder is better recipe for ultrasound image recognition. *arXiv preprint arXiv:2306.08249*, 2023.

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.

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

Mengkang Lu, Tianyi Wang, and Yong Xia. Multi-modal pathological pre-training via masked autoencoders for breast cancer diagnosis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 457–466. Springer, 2023.

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.

Jong Hak Moon, Hyungyung Lee, Woncheol Shin, Young-Hak Kim, and Edward Choi. Multimodal understanding and generation for medical images and text via vision-language pre-training. *IEEE Journal of Biomedical and Health Informatics*, 26(12):6070–6080, 2022.

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

Sidharth Puri, Divya Srikumaran, Christina Prescott, Jing Tian, and Shameema Sikder. Assessment of resident training and preparedness for cataract surgery. *Journal of Cataract & Refractive Surgery*, 43(3):364–368, 2017.

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)*, 2025.

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.

Andru P Twinanda, Sherif Shehata, Didier Mutter, Jacques Marescaux, Michel De Mathelin, and Nicolas Padoy. Endonet: a deep architecture for recognition tasks on laparoscopic videos. *IEEE transactions on medical imaging*, 36(1):86–97, 2016.

Bohua Wan, Michael Peven, Gregory Hager, Shameema Sikder, and S Swaroop Vedula. Spatial-temporal attention for video-based assessment of intraoperative surgical skill. *Scientific Reports*, 14(1):26912, 2024.

Chen Wei, Haoqi Fan, Saining Xie, Chao-Yuan Wu, Alan Yuille, and Christoph Feichtenhofer. Masked feature prediction for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14668–14678, 2022.

Mingze Xia, Nisarg A. Shah, Vishal M. Patel, S. Swaroop Vedula, and Shameema Sikder. Leveraging graph attention networks for targeted feedback from operating room surgery videos. In *Proceedings of the International Symposium on Biomedical Imaging (ISBI)*, TBD, 2025. IEEE.

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.

Appendix A. More results of comparison on Pretraining dataset and Masking methods

Table 4 presents a detailed comparison of JHU-VPT(JEPA) with several state-of-the-art methods for D99 Step Recognition under various data-regime settings (10

Notably, JHU-VPT(JEPA), which is pretrained on the large and diverse D-2591 dataset using a Multi-block masking approach, achieves the highest performance in low-data regimes (62.2 at 10%, 65.86 at 25%, and 70.42 at 50%). These results underscore JHU-VPT(JEPA)'s ability to learn robust representations that are particularly effective when labeled data is scarce. While some methods, such as GLSFormer, surpass JHU-VPT(JEPA) at the full data regime (100%), our approach offers a compelling advantage in scenarios where extensive labeled data is unavailable.

Overall, these findings highlight the effectiveness of combining extensive pretraining with tailored masking strategies, positioning JHU-VPT(JEPA) as a strong candidate for applications with privacy constraints and limited annotation resources.

Table 4: Comparison of JHU-VPT(JEPA) with other state-of-the-art methods on D99 Step Recognition, under different data-regime settings and pretraining datasets.

Methods	Pre-training Dataset	Masking	Data Regime (%)			
			10	25	50	100
MaskFeat (Wei et al., 2022)	Kinetics-400	Random	47.28	59.32	60.47	72.85
GLSFormer (Shah et al., 2023)	Kinetics-400	-	47.19	61.54	63.76	**80.24**
VideoMAE (Tong et al., 2022)	D-450	Frame	48.62	58.73	60.84	70.91
STMAE (Feichtenhofer et al., 2022)	D-450	Random	52.37	60.42	63.58	74.16
VideoMAE (Tong et al., 2022)	Kinetics-400	Tube	46.16	59.76	60.99	73.35
VideoMAE (Tong et al., 2022)	D-450	Random	50.24	60.89	62.34	72.98
VideoMAE (Tong et al., 2022)	D-450	Tube	52.11	61.59	63.72	74.39
CSMAE (Shah et al., 2025)	D-450	Token Selection	54.75	63.12	65.83	78.14
JHU-VPT(JEPA)	D-2591	Multi-block	**62.2**	**65.86**	**70.42**	75.61

Figure 2 shows the step recognition accuracy across different dataset splits after full fine-tuning. Both our JHU-VPT(JEPA) and the D-2591 models (VideoMAE pretraining) consistently outperform CSMAE on the Cataract-1k and Cataract-1k-JHU datasets, reinforcing the robustness of our pretraining with a large dataset.

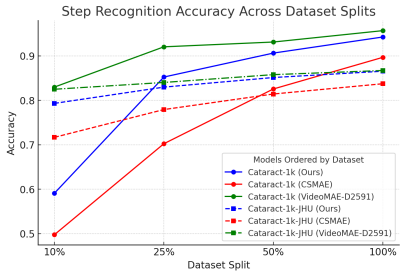

Figure 2: Step Recognition Accuracy across different dataset splits after complete fine-tuning.