SURGJEPA: STRUCTURED CONTEXT FOR MULTI-HORIZON SURGICALWORKFLOW ANTICIPATION

Yan Meng, Daniel A. Donoho Department of Neurosourgery
Children's National Hospital
Washington, DC 20010, USA
{ymeng,ddonoho}@childrensnational.org

ABSTRACT

Anticipating surgical workflow requires integrating visual observations with procedural context while accounting for the substantial nuisance variation present in future frames. Rather than predicting future pixels directly, we model future surgical scenes in a learned latent representation space. We introduce SurgJEPA, a tool-conditioned joint embedding predictive architecture that forecasts distributions over latent surgical states at 5, 30, 60, and 300 seconds. A causal context encoder integrates historical visual features, phase information, and temporally aligned tool observations, together with explicit missingness indicators that enable robust inference when tool data are incomplete. For each prediction horizon, a horizon-conditioned predictor generates four weighted latent hypotheses, while an exponential-moving-average target encoder provides self-supervised representations of future observations. In a five datasets study comprising 301 cases with case disjoint splits, tool-conditioned training improves dataset macro future phase macro-F1 over a matched tool free configuration by 10.2%, 9.4%, 9.1%, and 5.0% at 5, 30, 60, and 300 seconds, respectively, while reducing phase negative log-likelihood at every horizon. Representation level analyses further show that the learned contextual state supports future-phase decoding accuracies of 83.5%, 74.3%, 72.7%, and 53.5% across the four horizons, and that the predicted mixture representation retains complementary information about future workflow state. A controlled pilot additionally demonstrates that latent temperature substantially affects the geometry of the predicted representation, increasing pooled effective rank from 2.80 to 17.54. Together, these findings demonstrate the value of structured, tool conditioning multi-horizon surgical workflow anticipation and clarify how future phase information is distributed across the learned latent representations.

1 INTRODUCTION

Surgical workflow anticipation asks what is likely to occur after the current observation. Unlike phase recognition, which identifies the activity currently being performed, anticipation reason about both how long the present activity is likely to continue and which procedural transition may follow. This distinction becomes increasingly important across prediction horizons. At short horizons, recent visual activity and instrument use may provide strong evidence about local continuation or an imminent transition. At longer horizons, accurate forecasting depends more heavily on accumulated procedural context, phase duration, case progress, optional workflow branches, and procedure specific chronology. A useful anticipation model should therefore operate across multiple horizons and present uncertainty rather than commit to a single deterministic future.

Two properties make this task challenging. First, visual appearance varies with camera motion, lighting, tissue deformation, and instrument visibility. Predicting every future pixel is therefore a demanding intermediate objective for a task concerned with workflow. Second, the available context is heterogeneous. Video features, phase observations, and tool annotations have different semantics and may be missing at different times. A representation should combine the information that is available without treating an unobserved instrument as a confirmed absence. Joint-embedding predictive approaches offer a natural alternative by learning to predict representations of future observations rather than reconstructing their exact pixel level appearance. The value of jointly modeling instru-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

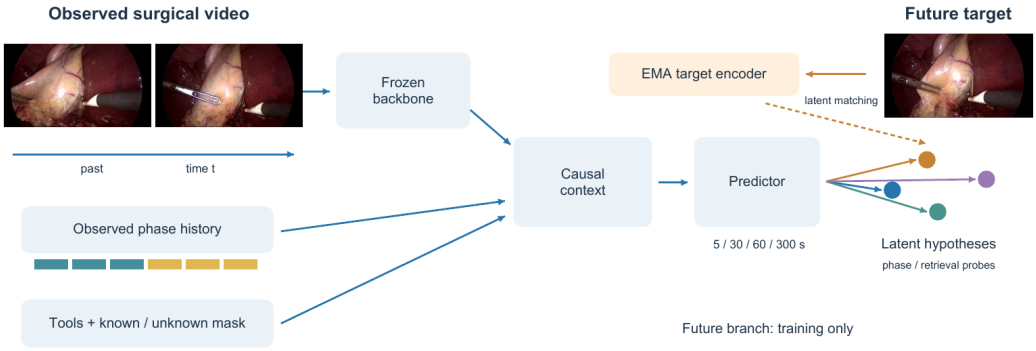


Figure 1: **SurgJEPA overview.** Past visual features, observed phase history, and tool observations form a causal context. The predictor produces weighted latent hypotheses for each forecast horizon. A future target branch provides representation-level supervision during training; a phase readout evaluates the resulting predictions. The target branch uses the same frozen visual backbone, omitted on that branch for clarity. Real Cholec80 case-01 frames illustrate the data flow. Their spacing, the colored history bars, and the latent symbols are schematic; the frames are neither generated predictions nor verified examples of the four forecast horizons.

ments and surgical phases has been recognized since EndoNet Twinanda et al. (2017). V-JEPA demonstrated the effectiveness of predicting masked target representations instead of pixels Assran et al. (2024), while V-JEPA 2 extended predictive representation learning to action anticipation Assran et al. (2025). These findings motivate a surgical anticipation framework that focuses on the predictable semantic state of the procedure rather than its precise future visual realization.

We address these challenges with SurgJEPA, a tool-conditioned joint-embedding predictive architecture for multi-horizon surgical workflow anticipation (Fig. **??**). The model predicts future latent representations rather than future images. A causal context encoder integrates visual features from the observed clip sequence with past phase information and temporally aligned tool observations. Our study focuses on three questions. First, does tool-conditioned training improve future workflow anticipation under otherwise matched training and evaluation conditions? Second, where is futur -phase information retained within the learned representation: in the contextual state, the predicted mixture representation, or the individual latent hypotheses? Third, how does the latent temperature affect the concentration and geometry of the predictive representation? We address these questions through comparison between tool-conditioned and tool-free training, a representation specific linear probe benchmark, and a matched temperature pilot. These analyses are complementary: the tool comparison evaluates the effect of conditioning during training, the probe benchmark measures how readily future phase information can be decoded from different internal representations, and the temperature pilot examines how prediction geometry changes under controlled variation of the latent distribution.

Our contributions are fourfold:

- We formulate a causal, multi-horizon joint-embedding predictive framework for surgical workflow anticipation that forecasts multiple weighted latent hypotheses at 5, 30, 60, and 300 seconds rather than a single deterministic future.
- We introduce structured, missingness-aware conditioning that integrates visual context, observed phase history, and temporally aligned tool information while distinguishing an observed empty tool set from an unknown tool observation.
- We conduct a five datasets, 301 case study with case-disjoint splits and a matched three-seed comparison between tool-conditioned and tool-free configurations, showing consistent improvements in future phase anticipation with tool-conditioned training.
- We characterize the learned representation using matched linear probes of the contextual state, predicted mixture representation, and individual hypotheses, and further examine how latent temperature alters representation geometry in a controlled pilot.

2 Related Work

2.1 Surgical workflow modeling

Surgical workflow analysis has largely focused on recognizing the current procedural state from video. EndoNet demonstrated the benefit of jointly modeling surgical phases and instrument presence in laparoscopic procedures Twinanda et al. (2017). Subsequent approaches have increasingly emphasized temporal context. TeCNO applies multi-stage temporal convolutions for causal phase recognition Czempiel et al. (2020); Trans-SVNet combines short- and long-range temporal dependencies Gao et al. (2022); SKiT is designed to capture long-range context efficiently for online recognition Liu et al. (2023); and Surgformer employs hierarchical temporal attention to model surgical workflow dynamics Yang et al. (2024). Related work in temporal action segmentation has introduced multi-stage temporal convolutional models and transformer-based architectures together with edit and segmental overlap metrics Abu Farha & Gall (2019); Yi et al. (2021); Meng et al. (2026), while boundary-aware approaches seek to reduce over-segmentation around activity transitions Ishikawa et al. (2021).

These methods establish the importance of temporal structure for surgical video understanding, but their primary objective is to identify the current procedural state. We want to focus on anticipating future workflow states at explicit prediction horizons. This distinction requires the representation to encode not only the current phase but also information about likely continuation, transition, and longer-term procedural progression. Because phase definitions and workflow structure vary substantially across procedures, our evaluation preserves the native label space of each dataset rather than imposing a common phase taxonomy.

The growing diversity of surgical workflow datasets further motivates evaluation across multiple procedure families. AutoLaparo provides workflow annotations for laparoscopic hysterectomy Wang et al. (2022); Cataract-101 and SICS-105 cover different forms of cataract surgery Schoeffmann et al. (2018); Mueller et al. (2025); HeiChole captures multicenter variability in laparoscopic cholecystectomy Wagner et al. (2023); M2CAI16-Workflow provides an established benchmark for cholecystectomy phase recognition Stauder et al. (2016); and PmLR50 characterizes workflow surrounding the Pringle maneuver in liver resection Guo et al. (2025). Evaluating predictive representations across such heterogeneous procedures provides a stronger test of whether the learned temporal structure transfers beyond a single operation or dataset.

2.2 Predictive representations and surgical representation learning

Representation learning for video has increasingly moved beyond direct pixel-level supervision. VideoMAE learns transferable video representations through masked spatiotemporal reconstruction Tong et al. (2022). Joint-embedding approaches instead predict learned target representations, thereby reducing the need to reconstruct visually complex and potentially task-irrelevant pixel details. V-JEPA predicts target-encoder representations for masked video regions without relying on pixel reconstruction, negative pairs, or text supervision Assran et al. (2024), while V-JEPA 2 extends predictive representation learning to larger-scale video understanding and action anticipation Assran et al. (2025).

Related efforts in surgical representation learning have explored domain-specific pretraining and procedure-aware multimodal supervision. Endoscopic pretraining has been used to learn representations adapted to surgical imagery Ender et al. (2024), while video-language approaches incorporate semantic and procedural information through paired supervision Yuan et al. (2024). We build on the broader idea of representation-level prediction but differs in several respects. First, the prediction targets correspond to explicit future clips at multiple horizons rather than only masked regions within the observed video. Second, prediction is conditioned on structured surgical context, including phase history and tool observations. Third, the learned representation is evaluated according to how much future-phase information it retains across multiple procedures.

2.3 Forecasting and uncertainty

Forecasting future surgical activity has been studied from several complementary perspectives. Future frame prediction seeks to extrapolate the visual evolution of a surgical scene Gao et al. (2021),

whereas goal-conditioned procedure planning predicts actions required to reach a supplied target state or goal image Zhao et al. (2024). Our setting differs from both. Rather than reconstructing future images or conditioning on an externally specified goal, we predict latent representations of future surgical states directly from the observed past.

An important challenge in this setting is that surgical futures may be ambiguous. Multi-hypothesis prediction provides a mechanism for representing several plausible latent outcomes, but diversity among hypotheses should not be interpreted as calibrated uncertainty by itself. Multiple embeddings may differ geometrically without corresponding to distinct procedural futures, and a diverse latent set may still yield poorly calibrated probabilities. We therefore separate representation diversity from probabilistic evaluation and assess future phase decisions independently from latent matching and representation geometry.

This distinction is consistent with the broader uncertainty estimation literature. Deep ensembles provide a practical approach for capturing predictive variability Lakshminarayanan et al. (2017), while expected calibration error measures mismatch between predictive confidence and empirical correctness Guo et al. (2017). Proper probabilistic scores such as negative log likelihood further evaluate whether predicted distributions assign appropriate probability mass to observed outcomes. In our evaluation, these diagnostics complement representation-level analyses and help distinguish improved future-state encoding from improved confidence calibration.

3 METHOD

3.1 OBSERVED CONTEXT AND FORECASTING TARGETS

Each case is represented on a one-second temporal grid. Let t denote the observation time, $v_t \in \mathbb{R}^{768}$ the frozen visual feature at that time, y_t the observed native phase label, u_t the tool observations, and m_t a mask indicating whether the corresponding tool information is known. The model receives the most recent $L = 60$ time steps and predicts future states at horizons in the set $\mathcal{H} = \{5, 30, 60, 300\}$ seconds. The observed context is summarized as:

$$c_t = C_\theta(v_{t-L+1:t}, y_{t-L+1:t}, u_{t-L+1:t}, m_{t-L+1:t}) \,, \tag{1}$$

where C_θ is the context encoder with trainable parameters θ, and a subscript interval denotes the sequence of one-second observations between its endpoints. The resulting contextual state $c_t \in \mathbb{R}^{768}$ summarizes all information available through time t. All evaluated SurgJEPA configurations use annotated past phase history. Future phase labels are excluded from the context encoder and are used only for downstream training and evaluation.

For each horizon $h \in \mathcal{H}$, the future target is defined as

$$z_{t,h} = \mathrm{norm}(E_{\bar{\theta}}\left(v_{t+h:t+h+W_h-1}\right)) \,, \tag{2}$$

where $E_{\bar{\theta}}$ is the target encoder, whose parameters $\bar{\theta}$ are updated as an exponential moving average (EMA) of the corresponding online encoder parameters, and W_h denotes the number of feature tokens in the future target window. We use $W_h = 2$ for the 5- and 30-second horizons and $W_h = 4$ for the 60- and 300-second horizons. The operator norm performs ℓ_2 normalization, yielding a unit length target representation $z_{t,h} \in \mathbb{R}^{512}$. An anchor contributes to a given horizon only when its complete future target window is available; future sequences are not padded beyond the end of a case.

3.2 ENCODING VISUAL AND STRUCTURED CONTEXT

The model integrates visual, phase, procedure, and tool information within a common temporal representation. Each visual feature vector is mapped through a learned linear projection to match the temporal model's hidden dimension. Phase identities are represented using dataset specific embeddings, while a procedure embedding provides the known procedure context.

Tool observations encode category, presence, entry, exit, confidence or availability, and observation age. Because multiple tools may occur within the same one-second interval, their token representations are combined using permutation invariant pooling and then processed by a causal temporal

encoder to summarize tool history. Unknown observations receive a learned missingness representation, distinct from an observed empty tool set. The model therefore distinguishes "no tool observed" from "tool status unknown."

The visual, phase, tool, and procedure embeddings are combined and passed through a six-block causal Transformer with eight attention heads. Causal attention restricts each position to information available at the same or earlier time steps. The final temporal output is used as the contextual state c_t in Eq. equation 1.

To reduce dependence on complete tool annotations, training includes several forms of tool stream perturbation. The entire tool stream is removed with probability 0.25, individual tool observations are dropped with probability 0.15, and tool timestamps are perturbed with probability 0.10. These augmentations expose the model to incomplete and temporally imperfect tool information while preserving the distinction between missing and observed empty states.

3.3 Horizon-conditioned latent prediction

A learned embedding e_h specifies the prediction horizon. Given the contextual state and horizon embedding, the predictor P_ϕ produces $K = 4$ candidate future representations and their corresponding mixture weights:

$$P_\phi(c_t, e_h) = \{(\hat{z}_{t,h,k}, \pi_{t,h,k})\}_{k=1}^{K}, \tag{3}$$

where ϕ denotes the predictor parameters. Each $\hat{z}_{t,h,k} \in \mathbb{R}^{512}$ is an ℓ_2-normalized latent prediction, and the mixture weights are obtained by applying a softmax to the corresponding logits:

$$\pi_{t,h,k} \geq 0, \qquad \sum_{k=1}^{K} \pi_{t,h,k} = 1. \tag{4}$$

The multiple hypotheses allow the predictor to represent alternative outcomes in latent space without forcing all plausible futures into a single embedding. We do not assume, however, that individual hypotheses correspond to distinct clinical trajectories.

For retrieval and linear probing, we also form a single summary,

$$q_{t,h} = \mathrm{norm}\left(\sum_{k=1}^{K} \pi_{t,h,k} \hat{z}_{t,h,k}\right), \tag{5}$$

where $q_{t,h}$ is the normalized mixture-mean prediction. Studying c_t, $q_{t,h}$, and each individual hypothesis allows us to distinguish information available in the observed context from information retained after predictive projection.

3.4 Learning from future representations

For unit vectors a and b, define cosine distance as $\delta(a, b) = 1 - a^\top b$, where $a^\top b$ is their inner product. Latent prediction minimizes the mixture matching energy:

$$\ell_{t,h} = -\log\left[\sum_{k=1}^{K} \pi_{t,h,k} \exp\left(-\frac{\delta(\hat{z}_{t,h,k}, z_{t,h})}{\sigma_h}\right)\right], \tag{6}$$

where $\sigma_h > 0$ is a horizon-specific latent temperature. Smaller values of σ_h place greater emphasis on hypotheses close to the target, whereas larger values distribute matching mass more broadly across hypotheses. The main experiments learn a separate temperature for each horizon within $[0.02, 0.5]$. A controlled pilot additionally compares this setting with a fixed temperature of 0.1.

Equation 6 operates entirely in latent representation space and should therefore not be interpreted as a categorical phase likelihood. Future-phase negative log likelihood and calibration are computed separately from downstream phase probabilities.

The training objective augments the mixture matching energy with four additional terms. First, a best-of-four cosine-distance loss encourages at least one hypothesis to approach the future target

Table 1: Case-disjoint visual-study split. Tool annotations are available for bypass and Cholec80. All five sources retain their native phase definitions.

Dataset	Train	Validation	Test
AutoLaparo	11	5	5
Bypass	44	10	10
Cholec80	44	19	17
PmLR50	32	10	8
Pituitary	47	20	19
Total	178	64	59

and is weighted by 0.5. Second, a variance–covariance regularizer on normalized online target representations, weighted by 0.1, discourages collapse of the latent space. Third, a hypothesis diversity term with coefficient 0.05 encourages separation among candidate predictions; its margin is adapted smoothly according to the running within-phase dispersion of target representations. Finally, an auxiliary future-tool binary cross entropy loss with coefficient 0.2 encourages the predicted representation to retain information relevant to future tool usage. The tool loss is applied only when the corresponding annotations are available.

Because the number of valid future targets differs across horizons, the mixture-matching and best-of-four losses are first averaged over the valid anchors for each horizon and then combined using inverse valid-anchor-count weighting within each minibatch. Let N_h denote the number of valid anchors for horizon h, and let $\mathcal{H}_{\text{valid}}$ denote the set of horizons with at least one valid target in the minibatch. The normalized horizon weight is:

$$w_h = \frac{N_h^{-1}}{\sum_{h' \in \mathcal{H}_{\text{valid}}} N_{h'}^{-1}} \left| \mathcal{H}_{\text{valid}} \right|. \tag{7}$$

Thus, the weights sum to the number of evaluated horizons, preserving the overall scale of the predictive objective, while horizons with fewer valid anchors receive greater relative weight. For example, before the common normalization factor is applied, a horizon with 20 valid anchors receives twice the weight of a horizon with 40. This weighting prevents horizons with more available future targets from dominating optimization.

The target branch receives no gradient. Instead, its parameters are updated by an EMA of the corresponding online components, including the visual projection, temporal trunk, and the first hypothesis slice of the latent projection. The EMA momentum is scheduled from 0.996 toward 0.9999 during training.

For the matched tool-free configuration, tool conditioning and the auxiliary future-tool loss are both disabled, while the remaining architecture and optimization procedure are unchanged. Consequently, the tool-conditioned versus tool-free comparison measures the combined effect of tool-conditioned training rather than isolating the contribution of tool inputs alone.

4 Experiments and Results

4.1 Datasets and evaluation protocol

The visual study includes 301 cases from AutoLaparo, Bypass, Cholec80, PmLR50, and Pituitary Tumore Surgery (PTS) from established procedure-specific workflow benchmarks Wang et al. (2022); Twinanda et al. (2017); Guo et al. (2025); Meng et al. (2025). We use 178 training, 64 validation, and 59 test cases (Table 1). The split is case-disjoint and shared across training seeds. Aligned tool annotations are available for Bypass and Cholec80; missing tool streams in other sources are represented explicitly.

A source-to-cache audit covers all 301 cases and reports zero mismatches on labeled positions, complete phase vocabulary mappings, and media coverage of the cached spans. PTS annotations

Table 2: Future-phase anticipation on the shared test split. Each cell gives accuracy / macro-F1 in percentage. Metrics average cases within each dataset and then datasets equally; SurgJEPA rows additionally average three training seeds. Persistence and both SurgJEPA configurations use annotated current or past phase information.

Method	5 s	30 s	60 s	300 s
Current-phase persistence	89.4 / 86.1	77.0 / 68.2	72.7 / 59.8	54.3 / 41.1
Per-dataset majority	49.8 / 17.4	49.2 / 19.7	50.8 / 21.9	42.3 / 18.9
SurgJEPA, without tools	15.6 / 7.4	11.6 / 6.3	9.0 / 5.2	15.7 / 8.0
SurgJEPA, with tools	27.7 / 17.5	25.5 / 15.7	24.9 / 14.3	23.8 / 13.0

Table 3: Tool-conditioned minus tool-free configuration. Accuracy and macro-F1 differences are percentage points; lower phase NLL is better. The contrast includes both tool inputs and auxiliary tool supervision.

Horizon (s)	ΔAcc.	ΔF1	ΔNLL
5	+12.1	+10.2	-0.171
30	+13.9	+9.4	-0.142
60	+15.9	+9.1	-0.146
300	+8.1	+5.0	-0.059

contain unlabelled intervals, represented as unknown labels and excluded from phase scoring. Three pituitary training or validation cases have shorter cached spans than their manifest durations; anchor eligibility uses the actual cache length. These checks support consistent alignment without treating missing annotations as negative labels.

Training setup. The main comparison contains three seeds, for each of the tool-conditioned and tool-free configurations. The model uses AdamW, learning rate warm-up, gradient clipping, and validation-based early stopping. Early stopping can produce different completed update counts. All visual features remain frozen. The primary phase readout is trained separately for each dataset and horizon with class balanced loss, using only training cases. It maps predicted hypotheses to phase probabilities and combines them using their mixture weights.

Metrics and aggregation. Accuracy measures the proportion of correctly predicted future phases. Macro-F1 averages phase-specific F1 scores, giving phases equal importance within the scoring protocol. In the primary evaluation, metrics are computed per case, averaged within each dataset, and then averaged equally across datasets. Configuration summaries additionally average the three training seeds. The saved per-run evaluation provides stratified case-bootstrap intervals. ECE measures the discrepancy between confidence and empirical correctness.

4.2 Contribution of tool-conditioned training

Tool-conditioned training improves future-phase recognition across all four horizons. At 5 seconds, mean accuracy increases from 15.6% to 27.7%, and macro-F1 increases from 7.4% to 17.5%. At 60 seconds, the corresponding gains are from 9.0% to 24.9% accuracy and from 5.2% to 14.3% macro-F1. The macro-F1 differences are 10.2, 9.4, 9.1, and 5.0 percentage points at 5, 30, 60, and 300 seconds, respectively (Table 3). Each of the three matched seed comparisons favors the tool-conditioned configuration in macro-F1 at every horizon.

The improvement extends to phase NLL, with reductions of approximately 0.171, 0.142, 0.146, and 0.059 across the four horizons. This indicates that the configuration changes both phase decisions and assigned probabilities. The mean ECE over horizons is nevertheless higher with tools, approximately 0.250 versus 0.216, so the NLL gain does not establish uniformly improved calibration.

Current-phase persistence provides an essential reference because the model receives annotated phase history. Its macro-F1 remains higher than that of either SurgJEPA configuration, including at 300 seconds (Table 2). The supported result is therefore a consistent benefit of tool-conditioned

Table 4: Representation-specific linear-probe accuracy (%). This secondary benchmark pools anchors within datasets before averaging datasets. SurgJEPA uses annotated phase history and tools where available; the frozen video reference uses neither.

Representation	5 s	30 s	60 s	300 s
Frozen video context	61.9	53.3	50.1	51.0
SurgJEPA context	83.5	74.3	72.7	53.5
Predicted mixture mean	67.0	59.6	62.6	50.7
Individual hypotheses	66.3–70.2	52.1–62.6	57.8–60.2	49.5–51.0

Table 5: Centered effective rank of predicted mixture means at 60 s in the matched six-epoch pilot. Higher rank denotes a less concentrated covariance spectrum; it is not an anticipation score.

Dataset	Learned temp.	Fixed temp.
Pooled	2.80	17.54
AutoLaparo	22.55	18.26
Bypass	2.94	7.22
Cholec80	1.84	5.10
PmLR50	17.11	13.97
Pituitary	20.19	23.56

training within the proposed architecture, while improved anticipation over phase-history controls remains a further objective.

4.3 Future-phase information within the representation

We next examine which internal representations expose future-phase information to a simple classifier. The secondary benchmark fits standardized logistic probes separately for each dataset and horizon. It compares mean frozen visual features over the observed context, the SurgJEPA contextual state, the predicted mixture mean, and each individual hypothesis. Standardization is fitted on training data. Probe optimization settings and training-label budgets are shared across representations within each benchmark; eligible anchor pools are shared, although capped training samples can differ between probe fits.

This benchmark averages anchor-level accuracy within a dataset and then averages datasets equally. It therefore differs from the primary case-averaged evaluation and is interpreted within its own table. Accuracy characterizes linear decodability in this analysis; the primary evaluation provides the complementary task-level macro-F1 results.

The contextual state supports 83.5%, 74.3%, 72.7%, and 53.5% future-phase accuracy across the four horizons. The predicted mixture mean supports 67.0%, 59.6%, 62.6%, and 50.7%. These results show that the context retains information useful at multiple horizons, while the predictive projection is less effective at preserving linearly accessible phase information in this setting. Individual hypotheses perform in a similar range to the mixture mean. At 5 seconds, their accuracy ranges from 66.3% to 70.2%, compared with 67.0% for the mixture mean.

The frozen-video row is an input-restricted reference: SurgJEPA additionally receives annotated phase history and tool information where available. Its stronger contextual scores cannot therefore be attributed solely to improved visual features or to the predictive objective. The comparison characterizes the complete structured-context representation under the stated supervision.

4.4 Temperature control and representation geometry

The matched pilot uses 37 of the 178 training cases, with validation and test cases unchanged. Both arms use the same training subset, and a six-epoch budget. The intervention fixes the latent temperature at 0.1; the comparison arm retains a learned temperature. This design tests whether the scale in Eq. equation 6 influences the structure of the predicted mixture means.

Table 6: Pilot mixture-mean linear-probe accuracy (%) under the representation-benchmark protocol. Both arms use the same 37-case training subset, seed, and six-epoch budget.

Temperature	5 s	30 s	60 s	300 s
Learned	85.1	73.4	68.4	56.6
Fixed	83.7	74.8	68.0	53.7

We measure geometry using the centered covariance of 60-second predictions. Effective rank is the exponential of the entropy of the covariance eigenvalues after normalizing them to sum to one. It measures how broadly variation is distributed across directions. The leading eigenvalue fraction measures the proportion of total variance explained by the dominant direction. These are representation descriptors, rather than measures of forecast correctness.

With fixed temperature, pooled effective rank increases from 2.80 to 17.54, and the leading eigenvalue fraction decreases from 0.82 to 0.30. The effect varies by source: rank increases for bypass, Cholec80, and pituitary, and decreases for AutoLaparo and PmLR50 (Table 5). The pooled calculation uses the sampled anchors and is not a dataset-macro statistic. The intervention therefore changes covariance concentration, with heterogeneous effects across procedures.

The matched linear-probe results in Table 6 show that greater pooled rank does not produce uniform anticipation gains. Mixture-mean accuracy changes from 73.4% to 74.8% at 30 seconds, but from 56.6% to 53.7% at 300 seconds. Under the primary mixture-probe evaluation, 60 second macro-F1 is 2.9% for the learned-temperature pilot and 2.6% for the fixed-temperature pilot. These findings motivate evaluating representation geometry and phase information together.

4.5 Scope and practical interpretation

The experiments shows that structured tool conditioning improves the proposed architecture relative to its tool-free configuration, and temperature control can materially change the geometry of its predicted representations. The strongest phase information is accessible in the contextual state, suggesting that future designs should preserve that information through the predictive stage.

The model receives manually annotated past surgical phases, whereas a deployed system would typically obtain this information from an automatic phase recognition model. Further evaluation is therefore needed to determine how recognition errors affect anticipation performance. Because the current test set has also informed exploratory analyses, future confirmation should use previously unexamined cases, with the evaluation protocol specified in advance. Furthermore, a visible instrument is not necessarily being used, and changing the model's tool inputs reveals its sensitivity to those inputs rather than the clinical effects of instrument use. The results therefore support further investigation of surgical workflow anticipation from recorded procedures; they do not yet establish suitability for clinical decision support.

5 Conclusion

We presented SurgJEPA, a predictive representation learning framework for multi-horizon surgical workflow anticipation. By integrating visual observations with phase history and missingness-aware tool information, SurgJEPA learns a structured temporal representation of the observed procedure and predicts multiple latent hypotheses for future surgical states without reconstructing future images. Across five surgical data sources, tool-conditioned training consistently improved future phase anticipation, while representation level analyses showed that the learned contextual state retains substantial information about subsequent workflow. Our findings further demonstrate that the geometry of the predicted latent representation can be modulated through temperature control, highlighting an additional mechanism for shaping predictive representations. Together, these results demonstrate the value of combining visual evidence with structured procedural context for surgical workflow anticipation. SurgJEPA provides a foundation for future work on predictive surgical representations, particularly toward modeling longer-term workflow uncertainty, incomplete contextual observations, and the complementary contributions of visual, temporal, and instrument information to understanding how surgical procedures evolve.

REFERENCES

Yazan Abu Farha and Juergen Gall. Ms-tcn: Multi-stage temporal convolutional network for action segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3575–3584, 2019.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *Transactions on Machine Learning Research*, 2024.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 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 *Medical Image Computing and Computer Assisted Intervention*, pp. 343–352, 2020.

Mert Ender et al. Endovit: Pretraining vision transformers on a large collection of endoscopic images. *International Journal of Computer Assisted Radiology and Surgery*, 19:1407–1416, 2024.

Xiaojie Gao, Yueming Jin, Zixu Zhao, Qi Dou, and Pheng-Ann Heng. Future frame prediction for robot-assisted surgery. In *Information Processing in Medical Imaging*, pp. 533–544, 2021.

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. *IEEE Transactions on Medical Imaging*, 41(5):1057–1068, 2022.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *International Conference on Machine Learning*, pp. 1321–1330, 2017.

Diandian Guo, Weixin Si, Zhixi Li, Jialun Pei, and Pheng-Ann Heng. Surgical workflow recognition and blocking effectiveness detection in laparoscopic liver resection with pringle maneuver. In *AAAI Conference on Artificial Intelligence*, volume 39, pp. 3220–3228, 2025. doi: 10.1609/aaai.v39i3.32332.

Yuchi Ishikawa, Seito Kasai, Yoshimitsu Aoki, and Hirokatsu Kataoka. Alleviating over-segmentation errors by detecting action boundaries. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 2322–2331, 2021.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, 2017.

Yang Liu, Jiayu Huo, Jingjing Peng, Rachel Sparks, Prokar Dasgupta, Alejandro Granados, and Sebastien Ourselin. Skit: A fast key information video transformer for online surgical phase recognition. In *IEEE/CVF International Conference on Computer Vision*, pp. 21074–21084, 2023.

Yan Meng, Daniel A Donoho, Marcelle Altshuler, and Omar Arnaout. Ai-driven evaluation of surgical skill via action recognition. *arXiv preprint arXiv:2512.24411*, 2025.

Yan Meng, Jack Cook, XY Han, Kaan Duman, Shauna Otto, Dhiraj Pangal, Jonathan Chainey, Ruth Lau, Margaux Masson-Forsythe, Daniel A Donoho, et al. Surgphase: Time efficient pituitary tumor surgery phase recognition via an interactive web platform. *arXiv preprint arXiv:2603.24897*, 2026.

Simon Mueller, Bhuvan Sachdeva, Singri Niharika Prasad, Raphael Lechtenboehmer, Frank G. Holz, Robert P. Finger, et al. Phase recognition in manual small-incision cataract surgery with ms-tcn++ on the novel sics-105 dataset. *Scientific Reports*, 15:16886, 2025. doi: 10.1038/s41598-025-00303-z.

Klaus Schoeffmann, Mario Taschwer, Stephanie Sarny, Bernd Muenzer, Manfred Juergen Primus, and Doris Putzgruber. Cataract-101: Video dataset of 101 cataract surgeries. In *ACM Multimedia Systems Conference*, pp. 421–425, 2018. doi: 10.1145/3204949.3208137.

Ralf Stauder, Daniel Ostler, Michael Kranzfelder, Sebastian Koller, Hubertus Feussner, and Nassir Navab. The tum lapchole dataset for the m2cai 2016 workflow challenge. *arXiv preprint arXiv:1610.09278*, 2016.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems*, 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, 2017. doi: 10.1109/TMI.2016.2593957.

Martin Wagner et al. Comparative validation of machine learning algorithms for surgical workflow and skill analysis with the heichole benchmark. *Medical Image Analysis*, 86:102770, 2023. doi: 10.1016/j.media.2023.102770.

Ziyi Wang, Bo Lu, Yonghao Long, Fangxun Zhong, Tak-Hong Cheung, Qi Dou, and Yunhui Liu. Autolaparo: A new dataset of integrated multi-tasks for image-guided surgical automation in laparoscopic hysterectomy. In *Medical Image Computing and Computer Assisted Intervention*, pp. 486–496, 2022.

Shu Yang, Luyang Luo, Qiong Wang, and Hao Chen. Surgformer: Surgical transformer with hierarchical temporal attention for surgical phase recognition. In *Medical Image Computing and Computer Assisted Intervention*, pp. 606–616, 2024.

Fangqiu Yi, Hongyu Wen, and Tingting Jiang. Asformer: Transformer for action segmentation. In *British Machine Vision Conference*, 2021.

Kun Yuan, Vinkle Srivastav, Nassir Navab, and Nicolas Padoy. Procedure-aware surgical video-language pretraining with hierarchical knowledge augmentation. In *Advances in Neural Information Processing Systems*, 2024.

Zixu Zhao et al. See, predict, plan: Diffusion for procedure planning in surgical videos. In *Medical Image Computing and Computer Assisted Intervention*, 2024.