CONTROLLED TRANSFER AND SPARSE EXPERT ROUTING FOR SURGICAL FOUNDATION MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Adapting surgical workflow recognition to a new procedure requires choosing a visual representation with limited labeled videos. Existing comparisons often change the encoder, temporal model, and training recipe together, leaving it unclear which representations transfer well and when source training reduces annotation needs. We propose SurgFM-Control, a procedure-aware framework that combines timestamp-aligned frozen-feature caching, a common temporal learner, and adaptation to native procedure labels. This design evaluates representation transfer jointly across surgical procedures, label budgets, and visual perturbations while reusing cached features for downstream experiments. Across three datasets, six frozen encoders reveal complementary strengths: VideoMamba achieves 74.4% and 52.9% macro-F1 on Cholec80 and Microanastomosis, respectively, while DINOv2 reaches 68.8% on AutoLaparo. With one labeled target video, source-initialized temporal learning improves macro-F1 by 5.7% points on average over target only training. DINOv2 also retains 88.7% of clean macro-F1 across the evaluated pixel corruptions, while training sample thinning preserves performance more effectively for long phases than for rapid action transitions. These findings establish a practical basis for matching frozen representations and adaptation strategies to a procedure's annotation budget and temporal structure.

1 INTRODUCTION

Surgical video understanding spans procedures with different visual and temporal structure. These differences create a fundamental challenge for representation learning: features that are informative for one workflow or temporal scale may not transfer equally well to another. As pretrained image and video encoders become increasingly common in surgical workflow recognition, an important question is therefore not simply which representation achieves the highest overall performance, but *which representation transfers best under a given procedure, annotation budget, temporal scale, and visual condition*. This question is difficult to answer from existing results because representation quality is often entangled with downstream modeling choices. Encoder architecture, temporal head, feature dimensionality, and optimization strategy frequently change together, making it difficult to determine whether a performance difference arises from the representation itself or from the surrounding pipeline. The amount of supervision further changes this comparison. A controlled comparison must therefore separate representation learning from temporal prediction while accounting for procedure, temporal granularity, and annotation availability.

This need is increasingly important as the range of pretrained visual representations expands. General image foundation models provide strong appearance representations learned at scale Oquab et al. (2024); Siméoni et al. (2025). VideoMAE learns temporal features through masked video reconstruction Tong et al. (2022); V-JEPA2 predicts latent future representations and captures motion and anticipation Assran et al. (2025); and VideoMamba enables efficient long-context video modeling with linear complexity Li et al. (2024). Surgical pretraining introduces a complementary prior: Endo-FM learns from large endoscopic video collections Wang et al. (2023), procedure-aware video-language models connect visual observations with surgical semantics Yuan et al. (2024), and SurgVISTA scales self-supervised representation learning across surgical procedures Yang et al. (2026). These models differ not only in architecture and pretraining objective, but also in training data, temporal context, and feature structure, making results obtained under different downstream pipelines difficult to compare directly.

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.

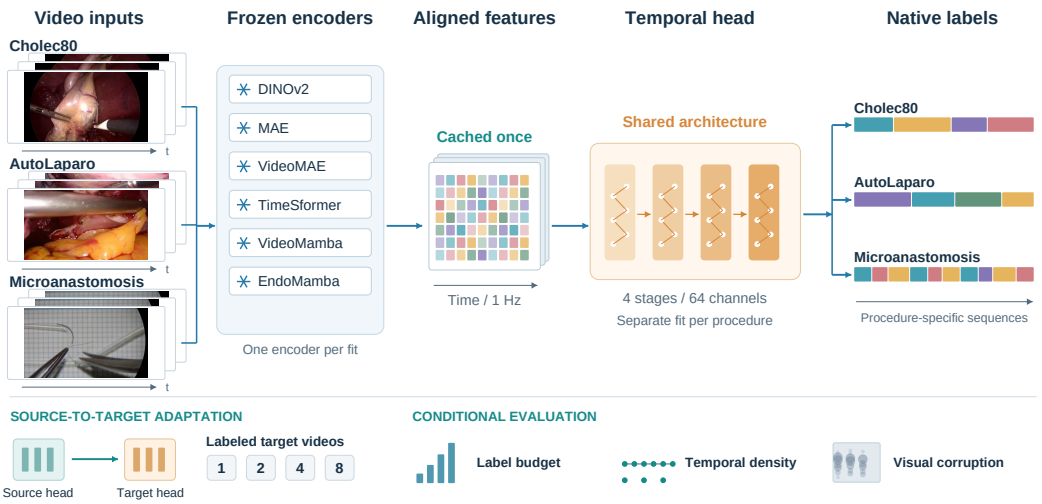

Figure 1: SurgFM-Control overview. Encoders are evaluated individually: timestamp-aligned features feed a separately trained temporal head with a common design and procedure-specific output labels. The lower branches show source-to-target initialization and the evaluation axes.

A controlled cataract study addressed part of this problem by evaluating cached image and video representations through a fixed MS-TCN++ temporal head Spencer et al. (2026). We extend this principle from controlled representation comparison to *conditional transfer*: rather than asking whether one encoder is universally superior, we examine how the relative usefulness of a representation changes with the target procedure, amount and density of supervision, temporal granularity, source initialization, visual corruption, and evaluation criterion.

We therefore present SurgFM-Control, a procedure-aware framework for controlled evaluation of frozen visual representations in surgical workflow recognition. The study makes three contributions:

- **A controlled framework for procedure-aware representation transfer.** We evaluate six frozen encoders across three surgical procedures using timestamp-aligned cached features, a common temporal learner, and procedure-native labels, with two additional procedure-adapted encoders included where cross-procedure comparison is appropriate.
- **A systematic analysis of label-efficient and cross-procedure adaptation.** Through case-budget, temporal-density, source-initialization, and corruption experiments, we distinguish the effects of representation choice, target supervision, temporal granularity, and transfer conditions across large range of saved experimental configurations.
- **A multidimensional characterization of representation utility.** Joint recognition, segmentation, and robustness evaluation identifies complementary image/video strengths and connects the effect of temporal subsampling to workflow granularity.

2 Related Work

Surgical workflow recognition. Surgical workflow recognition has progressed from framewise and recurrent approaches toward models that explicitly capture long-range spatiotemporal structure. Cholec80 established a widely used benchmark for joint phase and instrument recognition in laparoscopic cholecystectomy Twinanda et al. (2017), while AutoLaparo extended workflow analysis to laparoscopic hysterectomy Wang et al. (2022). Early temporal approaches modeled sequential dependencies using recurrent networks and multi-stage temporal convolutions Abu Farha & Gall (2019); Li et al. (2023); Czempiel et al. (2020). Subsequent methods increasingly integrated spatial and temporal representation learning Gao et al. (2022); Liu et al. (2023). More recent architectures further emphasize multi-scale and long-range temporal reasoning. Vido transformer models jointly models spatial and temporal information using divided attention and hierarchical temporal attention, aggregating local and global context from sparsely sampled frames Yang et al. (2024);

Meng et al. (2025). State-space models Mamba provide an alternative to attention-based sequence modeling Cao et al. (2024); Tian et al. (2025). However, they also introduce substantial variation in encoder architecture, sampling strategy, and optimization. This motivates our controlled setting, in which frozen representations are evaluated using a shared feature caching protocol and a common temporal learner.

Foundation models. Recent foundation models provide a diverse set of representations for surgical video analysis. DINO learns general purpose image features at scale while preserving strong dense visual representations Oquab et al. (2024); Siméoni et al. (2025). Video-based models incorporate temporal information through different learning objectives: VideoMAE uses high-ratio masked video reconstruction Tong et al. (2022), V-JEPA2 learns predictive latent representations from large-scale video Assran et al. (2025), and VideoMamba targets efficient modeling of long video sequences Li et al. (2024). Domain-specific pretraining introduces complementary surgical priors Wang et al. (2023); Yang et al. (2026), while procedure-aware vision-language pretraining incorporates textual supervision and hierarchical surgical knowledge Yuan et al. (2024). SurgFM-Control can evaluate representation quality together with practical properties such as cache size and robustness to visual corruption, rather than considering recognition accuracy in isolation.

Controlled comparison and temporal segmentation. Controlled evaluation is essential for separating representation quality from the effects of downstream temporal modeling. A surgical data efficiency study addressed this problem in cataract surgery by holding an MS-TCN++ temporal head fixed while comparing different pretrained encoders Spencer et al. (2026). We build on this principle and extend the evaluation to three procedures, cross-procedure transfer, label efficiency, visual corruption, and sparse representation routing. Other approaches target complementary aspects of temporal segmentation: FACT exchanges information between frame-level and action-level representations Lu & Elhamifar (2024), ASRF explicitly models action boundaries to reduce fragmented predictions Ishikawa et al. (2021), and recent boundary- and segment-aware objectives improve segmental F1 and edit scores with comparatively smaller changes in framewise accuracy Mitsuoka & Hotta (2026). These findings show that different metrics capture distinct properties of temporal predictions and motivate evaluating representations beyond frame level accuracy alone.

Weak labels, robustness, and calibration. The choice of representation also depends on how it behaves under limited supervision and degraded observations. Timestamp-supervised temporal segmentation has shown that sparse annotations can provide sufficient supervision for dense temporal predictions Li et al. (2021). Reliability is equally important: neural networks can produce poor confidence estimates even when their predictive accuracy is high Guo et al. (2017), and calibration can deteriorate further under distribution shift Ovadia et al. (2019). Robustness evaluation therefore provides a complementary view of representation quality. Common corruption benchmarks introduced systematic, severity-controlled perturbations for measuring visual robustness Hendrycks & Dietterich (2019). These perturbations affect spatial appearance and temporal evidence in different ways, making them useful for assessing not only whether a representation remains accurate under corruption, but also which forms of information it relies on.

3 METHOD

SurgFM-Control separates visual representation extraction from temporal workflow learning to enable a controlled comparison of pretrained encoders. The framework reuses existing pretrained encoders together with a multi-stage temporal model and therefore does not require the design of a new backbone. We first introduce the shared temporal learner, followed by the feature-caching strategy, procedure-specific adaptation, and evaluation.

3.1 UNIT OF COMPARISON

Let p denote a surgical type with K_p workflow labels, v a case, and e an encoder. At each sampled timestamp t, the frozen encoder produces a feature $f_t^{e,v} \in \mathbb{R}^{d_e}$, giving the sequence

$$F^{e,v} = (f_1^{e,v}, \ldots, f_{T_v}^{e,v}). \tag{1}$$

Table 1: Frozen visual encoders evaluated in SurgFM-Control. Features are extracted at one-second intervals using 224×224-pixel input frames. Image encoders process a single frame at each target timestamp; video encoders process 8 or 16 frames sampled across a two-second window centered on that timestamp, as specified in the Input column.

Encoder	Checkpoint / initialization	Pretraining signal	Input	Dim
DINOv2 ViT-B/14 Oquab et al. (2024)	vit_base_patch14_dinov2.lvd142m	image self-distillation	frame	768
MAE ViT-B He et al. (2022)	mae_pretrain_vit_base.pth	masked image reconstruction	frame	768
VideoMAE-B Tong et al. (2022)	videomae-base	masked video reconstruction	16 frames / 2 s	768
TimeSformer Bertasius et al. (2021)	TimeSformer_divST_8x32_224_K400.pyth	supervised video classification	8 frames / 2 s	768
VideoMamba-M Li et al. (2024)	VideoMamba-M16, K400	selective state-space video	16 frames / 2 s	576
EndoMamba Tian et al. (2025)	phase-recognition checkpoint	supervised endoscopic phases	16 frames / 2 s	384

where T_v is the number of sampled timestamps and d_e is the encoder's output feature dimension. The temporal head maps this sequence to probabilities over the K_p native labels:

$$q_t^{e,v} = \mathrm{softmax}(H_{\theta,e,p}(F^{e,v})_t) \in [0,1]^{K_p}, \tag{2}$$

where θ denotes the trainable head parameters and the entries of $q_t^{e,v}$ sum to one.

For each encoder–procedure pair, we train a separate head using the training cases and evaluate it on held-out cases. The head's first layer projects the native feature dimension d_e to a common width of 64 channels. Subsequent temporal layers have the same architecture across comparisons, although the input-projection parameter count varies with d_e.

The temporal head follows an MS-TCN style refinement architecture Abu Farha & Gall (2019). It contains four stages, each with six residual temporal-convolution layers, dilation factors $1, 2, 4, 8, 16, 32$, and 64 channels.

Within each comparison, we fix the feature sampling grid, input resolution, temporal-head architecture, optimizer, maximum epoch budget, and validation rule. Early stopping may produce different numbers of updates. All visual encoders remain frozen during downstream training.

3.2 Encoders

The encoder set covers several architectures and pretraining objectives. DINOv2 is a stable image baseline and MAE is its masked-reconstruction counterpart. VideoMAE tests masked video reconstruction, and TimeSformer tests supervised video classification. VideoMamba tests a selective state-space video architecture, and EndoMamba is a supervised endoscopic phase-recognition checkpoint. Two additional encoders,VideoMamba and SVFormer fine-tuned on microanastomosis phase labels, are admitted only in cross-procedure comparisons whose target is not microanastomosis, because fine-tuning on a procedure disqualifies an encoder from being evaluated on that procedure's target split. Table 1 lists checkpoints, inputs, dimensions, and admissible targets.

3.3 Feature caching and traceability

All procedures are first converted to a shared frame store sampled at 10 frames per second. Encoders receive 224×224 resized frames, and their outputs are aligned to a common 1-Hz target grid. Image encoders process a single frame at each target timestamp, whereas video encoders sample clips of 8

or 16 frames from a two-second window centered on the target time. Because these centered clips may include frames occurring after the target timestamp, this setup corresponds to offline rather than strictly causal recognition.

For each encoder, we define a cache contract that records the encoder identifier, native feature dimensionality, sampling rate, input resolution, clip configuration, preprocessing pipeline, and corruption settings. Each case is stored as aligned arrays containing visual features, workflow labels, timestamps, and available instrument indicators. Run manifests additionally record the corresponding input paths and experimental configurations. Together, these records provide traceability from each cached feature sequence back to the data and settings used to generate it.

Caching separates computationally expensive visual feature extraction from downstream temporal learning. Once features have been extracted, experiments involving different label budgets, target procedures, or transfer settings can reuse the same visual representations without repeatedly running the encoder. This reduces computational cost and ensures that downstream comparisons are performed on identical cached observations.

3.4 Evaluation methods and statistics

In-domain. In-domain experiments train and test within each procedure. We report accuracy, macro-F1 (mF1), edit score, segmental F1@$\{10, 25, 50\}$, boundary F1 within ± 1 s, and 15-bin expected calibration error (ECE). Metrics are first computed per case and then averaged without weighting by case duration. Scores are percentages except ECE, which is on a 0–1 scale.

Cross-procedure transfer. Cross-procedure experiments initialize a target temporal head from a source-trained head, then train on $k \in \{1, 2, 4, 8\}$ target videos. The baseline trains a new head on the same target cases ("scratch"). The implementation copies all parameters with matching names and shapes. The procedures each have class-dependent layers despite different label meanings. We refer to this operation as *source-initialized temporal learning*: the transferred head is adapted to the target labels during supervised training. Target training uses native labels throughout. Procedure-adapted encoders are excluded when microanastomosis is the target.

Label efficiency. To evaluate label efficiency, we train a separate temporal head using 10%, 25%, 50%, or 100% of the available training videos, while keeping the visual encoder frozen. Each selected video contributes its full labeled sequence on the 1-Hz feature grid. The number of selected videos is rounded to the nearest integer, with a minimum of one video. Validation and test sets remain fixed across training budgets.

We summarize performance across these four budgets using the normalized area under the label efficiency curve, $\mathrm{AULC}_{\log}$. Let f_j denote the jth training video fraction in increasing order, and let m_j denote the corresponding test macro-F1, averaged across test videos. We connect these four measurements with straight lines on a logarithmic training-fraction axis and integrate using the trapezoidal rule:

$$x_j = \log_{10} f_j, \qquad j = 1, \ldots, 4, \tag{3}$$

$$\mathrm{AULC}_{\log} = \frac{1}{x_4 - x_1} \sum_{j=1}^{3} \frac{m_j + m_{j+1}}{2} (x_{j+1} - x_j). \tag{4}$$

The logarithmic axis gives equal weight to equal proportional increases in training data. Dividing by the axis width produces a summary on the same scale as the underlying performance metric. Higher values indicate better performance across the evaluated training budgets.

Temporal sampling strategies. We investigate how the density and location of training samples affect workflow recognition while retaining all training videos. Uniform subsampling selects every second, fifth, or tenth sample from the 1 Hz training sequences, allowing us to assess sensitivity to reduced temporal sampling density.

We also evaluate boundary-focused selection to investigate whether samples surrounding workflow transitions provide sufficient context for recognition. A boundary is a change between consecutive phase or action labels. We retain samples within five time steps of either sample adjacent to each

boundary and concatenate the retained samples in chronological order. Videos with fewer than ten retained samples keep their full sequence. Validation and testing use complete sequences on the 1 Hz grid for both strategies.

Boundary-focused selection uses existing training annotations to locate transitions, so it evaluates sample selection rather than reduced annotation effort. Concatenation also alters temporal context by making previously separated portions adjacent.

Robustness and calibration. We evaluate robustness by measuring how recognition performance changes when visual inputs or cached temporal sequences are perturbed. Visual corruptions, such as smoke (blur plus contrast loss), specularity, and resolution downsampling, are re-extracted on the test split at severities $\{1, 3, 5\}$. Sequence corruptions, such as independent frame removal, integer label shifts, and feature-space Gaussian noise, are evaluated at levels $\{1, 2, 3\}$ and applied to the clean test caches. Feature noise represents a generic disturbance to the visual representation rather than a calibrated model of detector errors. The trained models are evaluated under these conditions without retraining.

We quantify robustness by the fraction of clean condition macro-F1 retained under corruption. For each encoder–procedure pair, let F_{clean} denote macro-F1 on the unperturbed test sequences and $F_{c,s}$ denote macro-F1 under corruption type c at severity s. Both scores are computed separately for each test video and then averaged across videos. Using the same trained model for clean and corrupted evaluation, we define percentage retention as:

$$\mathrm{R}(c) = \frac{100}{|\mathcal{S}_c|} \sum_{s \in \mathcal{S}_c} \frac{F_{c,s}}{F_{\text{clean}}}, \tag{5}$$

where $\mathcal{S}_c$ contains the evaluated severity levels and $F_{\text{clean}} > 0$. Thus, we first express performance at each severity as a fraction of clean performance, average these fractions across severities, and multiply by 100. A retention of 100% indicates unchanged performance on average; lower values indicate degradation, and values above 100% indicate improved performance under the evaluated perturbations.

4 Experiments

4.1 Datasets and experimental setup

The executed study covers the three timing-verified procedures. Cholec80 supplies 80 laparoscopic cholecystectomy videos with seven phases and a seven-instrument stream. AutoLaparo supplies 21 hysterectomy videos with seven phases. Microanastomosis64, including 64 microanastomosis cases, are fine-grained repeated-state streams: their action annotation is event-like, with a 17.5s median segment, in contrast to the 138s–380s medians of the other two procedures.

The segment density is 556.8 boundaries/hour in microanastomosis, 9.1 in Cholec80, and 6.2 in AutoLaparo, so boundary scores depend strongly on temporal scale: a ± 5 s illustrative window exceeds the median microanastomosis segment but is only 1.3% of the median AutoLaparo segment. The reported boundary metric uses the narrower ± 1 s tolerance. This structural contrast also predicts ranking instability. A spatial encoder can obtain many correct frames in a 380s appearance-defined phase even if it reacts slowly, whereas on microanastomosis the same delay spans multiple labels and severely harms edit and segmental F1. Pooling the procedures without an interaction term would make encoder conclusions depend primarily on whether frames or datasets receive equal weight.

The evaluation contains 100 configurations, including the auxiliary fusion study in Sec. D.1: 18 in-domain, 18 label-efficiency, 44 cross-procedure, 18 robustness, and two fusion configurations. These configurations organize the experimental comparisons rather than count individual model-training runs. Consequently, the total number of training runs exceeds the number of saved configurations.

Cases are disjoint across training, validation, and test sets. The splits are Cholec80 40/10/30, AutoLaparo 12/3/6, and microanastomosis 50/8/6. All modes use Adam with learning rate 10^{-4}, weight decay 10^{-4}, gradient clipping at 5, at most 60 epochs, patience 10, and are run on devices with single NVIDIA Blackwell RTX Pro 6000 GPU. Model selection uses pooled validation-frame accuracy.

Table 2: In-domain results. Accuracy and macro-F1 in %; edit, F1@10/25/50, and boundary F1; ECE in [0,1]. Best macro-F1 per dataset in bold.

Dataset	Encoder	Acc.	mF1	Edit	F1@10	F1@25	F1@50	B-F1	ECE
Cholec80	DINOv2 ViT-B/14	81.2	73.9	66.8	77.5	74.5	63.3	5.9	0.167
Cholec80	MAE ViT-B	69.3	55.3	50.7	60.9	53.3	42.4	3.5	0.246
Cholec80	VideoMAE-B	66.0	53.1	54.5	63.9	57.2	41.3	2.1	0.245
Cholec80	TimeSformer	82.0	73.8	69.7	78.3	76.3	64.9	7.1	0.154
Cholec80	VideoMamba-M	83.8	**74.4**	70.4	80.9	78.5	68.3	5.3	0.139
Cholec80	EndoMamba	53.5	37.1	49.4	54.0	46.8	30.1	0.8	0.389
AutoLaparo	DINOv2 ViT-B/14	79.4	**68.8**	53.8	64.5	61.5	48.1	9.3	0.180
AutoLaparo	MAE ViT-B	62.1	49.6	38.7	46.1	38.6	29.9	0.0	0.302
AutoLaparo	VideoMAE-B	53.5	37.7	32.3	37.6	29.6	13.9	0.0	0.405
AutoLaparo	TimeSformer	77.7	66.0	53.9	64.5	61.8	50.8	7.6	0.212
AutoLaparo	VideoMamba-M	73.9	57.6	46.7	52.5	47.1	36.8	4.6	0.229
AutoLaparo	EndoMamba	28.6	12.7	27.8	27.7	22.1	9.4	1.3	0.230
Microanastomosis	DINOv2 ViT-B/14	68.9	52.8	47.5	52.0	43.6	29.4	37.6	0.132
Microanastomosis	MAE ViT-B	62.0	27.1	24.8	29.2	24.1	13.7	22.7	0.136
Microanastomosis	VideoMAE-B	59.8	28.4	25.9	31.3	26.5	15.1	18.6	0.155
Microanastomosis	TimeSformer	74.2	52.6	44.3	52.4	46.5	32.2	37.2	0.080
Microanastomosis	VideoMamba-M	70.3	**52.9**	44.6	51.3	44.1	27.8	36.4	0.116
Microanastomosis	EndoMamba	63.7	28.0	27.7	30.5	23.0	11.6	18.7	0.067

4.2 In-domain results

Table 2 reports the 18 in-domain comparisons. VideoMamba has the highest mean macro-F1 on Cholec80 (74.4%; 95% case-bootstrap CI 69.9–78.3) and microanastomosis (52.9%; 45.5–60.4). DINOv2 leads AutoLaparo at 68.8% (62.7–74.6). The top encoders are closely grouped on Cholec80 and microanastomosis. EndoMamba has the lowest macro-F1 in each procedure, including 12.7% on AutoLaparo. The Cholec80 ECE is 0.389, compared with 0.116–0.212 for the leading encoders. Microanastomosis has a much denser transition structure in the audited subset. Its boundary F1 should therefore be interpreted within that dataset, rather than compared directly with boundary F1 on long surgical phases.

4.3 Label efficiency analysis

Table 3 reports case fraction and training sequence subsampling results. Case-level subsampling is far more damaging than frame-level subsampling: at 10% of training cases, encoder-mean macro-F1 falls to 36.6 (Cholec80), 8.6 (AutoLaparo), and 28.0 (microanastomosis) versus 61.0/52.5/37.8 at 100%. AutoLaparo is the most fragile: 10% means a single training case, and several encoders fall to the 5.7-point majority-class floor. Frame strides are comparatively benign on coarse phase datasets (encoder-mean mF1 at stride 10 vs. stride 1: 0.85 Cholec80, 0.95 AutoLaparo) but severe on microanastomosis (0.56). TimeSformer is the most fraction robust encoder, with the best 10%-case macro-F1 on Cholec80 (55.0) and AutoLaparo (15.6). Boundary-focused sequence selection performs poorly on coarse phases (20.3 for DINOv2 on Cholec80) yet matches or exceeds full supervision on microanastomosis (56.4 vs. 50.4 for TimeSformer): Dense action transitions leave a larger fraction of the sequence near a boundary, so this setting removes less training context.

4.4 Cross-procedure transfer

We vary the target supervision budget using $k \in \{1, 2, 4, 8\}$ labeled training videos from the target procedure. At each budget, the source-initialized head and the randomly initialized baseline use the same selected target videos. Table 4 aggregates the six directions over admissible encoders; Table 5 breaks the k=8 column out by encoder. The largest average transfer gain occurs with the fewest target cases: at k=1 the source-initialized head beats scratch training by $+5.7$ macro-F1 points on average (median $+3.6$), but the advantage decays to -1.2 (median -1.5) by k=8, and transfer wins only 17 of 44 encoder-direction cells (38.6%). Only one direction shows

Table 3: Label efficiency: macro-F1 (%) under case fractions $\{10, 25, 50, 100\}\%$, area under the label-efficiency curve over $\log_{10}$ fraction (AULC, Sec. 3.5), frame strides at full case budget, and boundary-focused sequence selection. The 100% cell is retrained inside the label-efficiency RNG stream and may differ slightly from Table 2.

		Case fraction					Frame stride			
Dataset	Encoder	10%	25%	50%	100%	$\mathrm{AULC}_{\log}$	2	5	10	Bnd.
Cholec80	DINOv2 ViT-B/14	50.9	57.2	75.2	77.2	64.4	74.0	72.2	65.5	20.3
Cholec80	MAE ViT-B	36.5	48.8	52.8	63.1	49.7	58.0	51.2	51.7	16.7
Cholec80	VideoMAE-B	25.2	14.3	15.9	47.2	21.9	47.1	42.7	28.6	14.4
Cholec80	TimeSformer	55.0	69.0	75.9	77.4	69.5	75.6	69.5	65.4	20.2
Cholec80	VideoMamba-M	43.2	68.8	67.6	74.8	64.2	72.7	68.7	58.2	19.5
Cholec80	EndoMamba	8.6	17.7	21.0	26.4	18.2	30.7	23.9	21.8	17.2
AutoLaparo	DINOv2 ViT-B/14	5.7	58.1	63.6	66.0	50.5	55.6	64.6	52.9	22.0
AutoLaparo	MAE ViT-B	5.7	40.6	42.6	47.2	35.2	39.7	48.8	45.1	20.9
AutoLaparo	VideoMAE-B	5.7	5.7	37.7	41.3	20.7	22.7	36.0	41.7	13.6
AutoLaparo	TimeSformer	15.6	61.0	65.2	66.0	54.0	61.8	63.8	58.0	25.5
AutoLaparo	VideoMamba-M	13.4	51.8	58.9	61.5	47.8	53.6	62.9	54.3	20.1
AutoLaparo	EndoMamba	5.7	5.7	22.4	32.9	14.8	7.4	5.8	11.1	18.4
Microanastomosis	DINOv2 ViT-B/14	41.1	50.3	53.0	56.3	50.2	54.8	48.7	35.0	54.0
Microanastomosis	MAE ViT-B	23.6	30.7	22.3	30.2	26.7	28.3	31.3	22.1	39.1
Microanastomosis	VideoMAE-B	10.9	15.0	10.9	24.7	14.4	19.4	15.7	12.0	27.4
Microanastomosis	TimeSformer	40.2	45.1	46.2	50.4	45.3	49.8	41.6	30.9	56.4
Microanastomosis	VideoMamba-M	41.2	49.0	45.6	44.2	45.7	48.4	41.7	29.9	48.1
Microanastomosis	EndoMamba	10.9	29.8	17.2	21.1	20.9	23.2	16.2	10.9	30.1

Table 4: Cross-procedure macro-F1 (%), averaged over admissible encoders. Each cell reports macro-F1 for the source-initialized head (first value) and the randomly initialized head (second value). k is the number of labeled target training videos. Microanastomosis-adapted encoders are excluded when microanastomosis is the target.

Direction	$k = 1$	$k = 2$	$k = 4$	$k = 8$
C→A	13.8 / 8.6	29.0 / 25.7	38.6 / 43.8	45.2 / 48.6
C→M	12.0 / 10.9	18.2 / 17.8	22.7 / 25.4	30.9 / 33.2
A→C	20.5 / 12.9	26.7 / 22.6	31.3 / 37.6	43.0 / 45.1
A→M	13.3 / 10.9	19.4 / 14.0	23.8 / 28.5	30.5 / 32.5
M→C	19.8 / 10.8	31.7 / 29.9	34.9 / 38.3	46.5 / 41.5
M→A	15.0 / 8.0	24.2 / 21.3	39.4 / 39.0	45.0 / 48.0

C: Cholec80; A: AutoLaparo; M: Microanastomosis. Arrows indicate source → target.

Table 5: Cross-procedure performance by encoder using $k = 8$ labeled target training videos. Dirs. counts admissible directions; transfer and scratch are mean macro-F1 (%); Δ is the mean paired difference in % points.

Encoder	Dirs.	Transfer	Scratch	Δ
DINOv2 ViT-B/14	6	57.8	61.1	−3.3
MAE ViT-B	6	34.7	33.6	+1.1
VideoMAE-B	6	26.8	25.1	+1.7
TimeSformer	6	52.8	60.3	−7.6
VideoMamba-M	6	53.8	54.1	−0.3
EndoMamba	6	15.9	18.5	−2.6
VideoMamba†	4	34.6	28.3	+6.2
SurgFormer†	4	54.3	57.5	−3.2

†Microanastomosis-adapted checkpoint.

a mean k=8 gain, Microanastomosis→Cholec80 (46.5 vs. 41.5). The best transfer sources are also the best in-domain encoders: DINOv2 (57.8 mean k=8 macro-F1 over its six directions), VideoMamba (53.8), and TimeSformer (52.8). The two microanastomosis-adapted sources do not systematically beat generalist encoders: SurgFormer (micro.-adapted) averages 54.3 over its four admissible directions and VideoMamba (micro.-adapted) only 34.6, both below DINOv2, and on Microanastomosis→AutoLaparo the adapted sources (61.4 and 19.6) trail TimeSformer (67.1).

4.5 Robustness

Table 6 reports clean macro-F1 from the robustness mode heads and relative retention per corruption family. Smoke dominates: encoder-mean macro-F1 retention across severities is 55.7% for smoke versus 92.2% (specularity) and 90.6% (resolution) for the other pixel cor-

Table 6: Measured robustness. Clean macro-F1 (%) from the robustness-mode head; R is mean relative macro-F1 retention (%) over severities $\{1, 3, 5\}$ (pixel, re-extracted) or levels $\{1, 2, 3\}$ (sequence, applied to clean caches). "Noise" is the feature-space Gaussian implementation of the detector-error corruption. Retention above 100% indicates a higher score on the perturbed inputs; its cause is not established.

			Pixel R (%)			Sequence R (%)		
Dataset	Encoder	Clean mF1	Smoke	Spec.	Res.	Drop	Jitter	Noise
Cholec80	DINOv2 ViT-B/14	74.7	69.1	97.6	96.7	97.0	98.9	99.5
Cholec80	MAE ViT-B	50.6	32.4	91.2	85.1	96.9	98.5	63.7
Cholec80	VideoMAE-B	46.6	84.7	100.6	107.0	100.1	98.3	77.0
Cholec80	TimeSformer	76.4	65.4	84.5	98.0	98.2	99.0	99.3
Cholec80	VideoMamba-M	76.1	69.0	91.0	94.4	96.9	99.0	97.4
Cholec80	EndoMamba	35.3	17.0	100.2	83.8	97.8	99.0	67.6
AutoLaparo	DINOv2 ViT-B/14	68.8	85.0	96.7	98.8	100.0	99.9	99.6
AutoLaparo	MAE ViT-B	49.6	42.0	93.7	76.6	99.3	100.1	41.4
AutoLaparo	VideoMAE-B	37.7	50.7	92.5	83.2	97.3	98.5	45.9
AutoLaparo	TimeSformer	66.0	70.1	83.7	95.9	101.6	99.8	101.6
AutoLaparo	VideoMamba-M	57.6	55.4	77.2	78.4	99.3	101.1	90.7
AutoLaparo	EndoMamba	12.7	51.0	100.1	99.5	97.0	100.3	99.7
Microanastomosis	DINOv2 ViT-B/14	54.1	57.6	98.8	97.6	96.7	76.7	98.5
Microanastomosis	MAE ViT-B	34.5	39.4	90.7	82.0	90.6	88.0	57.8
Microanastomosis	VideoMAE-B	28.7	80.8	88.8	92.5	92.4	87.9	69.1
Microanastomosis	TimeSformer	54.0	46.1	87.5	85.5	97.8	71.9	96.0
Microanastomosis	VideoMamba-M	49.3	51.7	79.5	91.0	93.6	85.6	79.7
Microanastomosis	EndoMamba	25.8	35.6	104.8	85.3	90.6	82.4	59.8

ruptions, and at severity 5 the absolute encoder-mean macro-F1 drops to 20.2/21.9/13.3 on Cholec80/AutoLaparo/microanastomosis from clean 60.0/48.7/41.1. Sequence corruptions are gentler: frame drop retains 96.8%, temporal jitter 93.6%, and feature noise 80.2%. DINOv2 is the most pixel-robust encoder (88.7% mean retention over the three families) and MAE the least (70.3%). A few weak-model cells exceed 100% retention (e.g., VideoMAE under Cholec80 specularity, EndoMamba under microanastomosis specularity): a perturbed test input can occasionally yield a higher score than the clean input.

4.6 Limitations

SurgFM-Control relies on frozen visual representations, which make repeated downstream training efficient but restrict adaptation to new surgical domains. The framework also requires labeled examples from each target procedure. Cross-procedure initialization may consequently be less effective when source and target workflows differ substantially. Finally, the robustness evaluation characterizes responses to predefined synthetic perturbations. These perturbations cannot represent the full variability of clinical practice, including changes in acquisition systems, surgical technique, and unexpected events.

5 Conclusion

We presented SurgFM-Control, a procedure-aware framework for evaluating and transferring frozen visual representations in surgical workflow recognition. Aligned feature caching and a common temporal learner make it possible to examine encoder choice jointly with label availability, temporal granularity, and visual robustness. Across three procedures, the experiments identify complementary image/video strengths, demonstrate the benefit from source initialization with few labeled target videos, and show why sampling density should reflect workflow duration. These findings provide a concrete basis for selecting representations and adaptation strategies when extending workflow recognition to a new procedure.

REFERENCES

Yazan Abu Farha and Jürgen Gall. Ms-tcn: Multi-stage temporal convolutional network for action segmentation. In *CVPR*, pp. 3575–3584, 2019.

Mido Assran et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.

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

Rui Cao, Jiangliu Wang, and Yun-Hui Liu. Sr-mamba: Effective surgical phase recognition with state space model. *arXiv preprint arXiv:2407.08333*, 2024.

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*, pp. 343–352, 2020.

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 *ICML*, pp. 1321–1330, 2017.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, pp. 16000–16009, 2022.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *ICLR*, 2019.

Yuchi Ishikawa, Seito Kasai, Yoshimitsu Aoki, and Hirokatsu Kataoka. Alleviating over-segmentation errors by detecting action boundaries. In *WACV*, pp. 2322–2331, 2021.

Maurice G Kendall et al. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93, 1938.

Kunchang Li, Xinhao Li, Yi Wang, Yinan He, Yali Wang, Limin Wang, and Yu Qiao. Videomamba: State space model for efficient video understanding. In *ECCV*, 2024.

Shi-Jie Li, Yazan Abu Farha, Yun Liu, Ming-Ming Cheng, and Juergen Gall. Ms-tcn++: Multi-stage temporal convolutional network for action segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):6647–6658, 2023.

Zhe Li, Yazan Abu Farha, and Juergen Gall. Temporal action segmentation from timestamp supervision. In *CVPR*, pp. 8365–8374, 2021.

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 *ICCV*, pp. 21074–21084, 2023.

Zijia Lu and Ehsan Elhamifar. Fact: Frame-action cross-attention temporal modeling for efficient action segmentation. In *CVPR*, pp. 18175–18185, 2024.

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.

Hinako Mitsuoka and Kazuhiro Hotta. Combining boundary supervision and segment-level regularization for fine-grained action segmentation. In *CVPR Workshops*, pp. 8865–8873, 2026.

Maxime Oquab et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Yaniv Ovadia et al. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *NeurIPS*, 2019.

Oriane Siméoni et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Lincoln Spencer, Song Wang, and Chen Chen. Data-efficient surgical phase segmentation in small-incision cataract surgery: A controlled study of vision foundation models. In *CVPR Workshops*, pp. 8805–8811, 2026.

Qingyao Tian, Huai Liao, Xinyan Huang, Bingyu Yang, Dongdong Lei, Sebastien Ourselin, and Hongbin Liu. Endomamba: An efficient foundation model for endoscopic videos via hierarchical pre-training. *arXiv preprint arXiv:2502.19090*, 2025.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *NeurIPS*, 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.

Zhao Wang, Chang Liu, Shaoting Zhang, and Qi Dou. Foundation model for endoscopy video analysis via large-scale self-supervised pre-train. In *MICCAI*, 2023.

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 *MICCAI*, 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 *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 606–616. Springer, 2024.

Shu Yang, Fengtao Zhou, Leon Mayer, et al. Large-scale self-supervised video foundation model for intelligent surgery. *npj Digital Medicine*, 9:220, 2026.

Kun Yuan, Vinkle Srivastav, Nassir Navab, and Nicolas Padoy. Procedure-aware surgical video-language pretraining with hierarchical knowledge augmentation. In *NeurIPS*, 2024.

A Appendix - General Metric Definitions

Frame accuracy is the fraction of correctly labeled samples. Macro-F1 averages class-wise F1 over all K_p dataset classes within each case; a class absent from both labels and predictions receives zero. This convention also applies to macro Jaccard and can lower scores on cases missing some phases. Dataset summaries are unweighted case means.

Edit score collapses consecutive repeated labels and reports $100(1 - d_{\text{edit}}/\max(N_{\text{gt}}, N_{\text{pred}}))$, with a denominator floor of one. Segmental F1@α greedily assigns each predicted segment to the unmatched same-label segment of largest temporal intersection-over-union, provided it is at least $\alpha \in \{0.10, 0.25, 0.50\}$. Segments use sample-index intervals.

Boundary F1 greedily matches predicted internal boundaries in temporal order to the closest unused true boundary within ± 1 s. Starts and ends are excluded. This is not a global optimal bipartite assignment. ECE uses 15 equal-width confidence bins and is averaged over cases. We report only macro-F1/accuracy retention for frame-removal corruptions: the archived evaluator does not subset the timestamp array with the same removal mask, so its post-removal boundary metrics require correction before use.

B Appendix - Corruption Parameters

Pixel severities are $a \in \{1, 3, 5\}$. Simulated smoke applies Gaussian blur with standard deviation a and multiplies intensities by $1 - 0.1a$. It is an appearance perturbation, not a physical smoke simulator. Specularity draws a saturated white ellipses at seeded locations, with axis lengths based on image dimensions and severity. Resolution corruption downsamples both dimensions by $1 - 0.12a$ and restores the original size by interpolation.

Sequence levels use independent frame-removal probabilities $\{0.1, 0.25, 0.5\}$; a constant, uniformly sampled integer label shift within $\pm\{2, 5, 10\}$ samples, with edge-label padding; or additive feature noise with standard deviation $\{0.1, 0.5, 1.0\}$. At the main 1-Hz grid, a sample offset corresponds to one second. These synthetic perturbations are reproducible from a seed, but have not been calibrated to measured clinical failure distributions.

C Appendix - Statistics and ranking stability

The saved macro-F1 intervals use 1,000 percentile bootstrap resamples of test cases within each procedure. They measure case-sampling uncertainty, not training-seed uncertainty. Encoders share test cases, but the archived summaries do not provide paired difference intervals for every comparison. We therefore interpret small ranking gaps descriptively rather than as significant or equivalent performance. We assess the consistency of encoder rankings across procedures, evaluation metrics, and corruption conditions using Kendall's rank correlation coefficient τ Kendall et al. (1938). It measures agreement in the relative ordering of encoder pairs: a pair is concordant when the same encoder ranks higher in both settings and discordant when their ordering is reversed. Values of $\tau = 1$ indicate identical rankings, $\tau = -1$ indicate completely reversed rankings, and values near zero indicate little overall agreement.

Kendall's τ between macro-F1 rankings is 0.73 for Cholec80 versus AutoLaparo, 0.73 for Cholec80 versus microanastomosis, and 0.47 for AutoLaparo versus microanastomosis. In the order Cholec80/AutoLaparo/microanastomosis, accuracy versus macro-F1 rankings give $\tau = 0.87/1.00/0.47$, clean versus severity-5 smoke gives $0.60/1.00/0.60$, and full versus 10% case budgets gives $0.60/0.27/0.67$. Thus the best encoder depends on the dataset and evaluation condition.

The framework reveals complementary strengths across recognition conditions. VideoMamba leads mean in-domain macro-F1 on Cholec80 and microanastomosis, whereas DINOv2 leads AutoLaparo, microanastomosis edit and boundary scores, and mean pixel-corruption retention. Source initialization is most useful in the one-video target setting, and coarse phase workflows tolerate training sample thinning better than the rapid microanastomosis action stream. Together, these results connect encoder choice to procedure, annotation budget, and temporal scale rather than to a single aggregate score.

D Appendix - Auxiliary Multi-Encoder Fusion

D.1 Fusion architecture

We design SurgRouter to combine six encoder-specific towers, each with a trainable input projection and a four-layer, 64-channel causal TCN. We evaluate SurgRouter on Cholec80 and microanastomosis, for which instrument-presence signals are available and aligned with the visual features. These signals form part of the routing input. AutoLaparo is excluded because the prepared dataset lacks an equivalent aligned instrument stream. Extending this evaluation to AutoLaparo would require constructing this additional input or adapting the gate to operate on visual features alone.

A two-layer gate receives cumulative means of the cached features and tool bits (seven annotated instruments for Cholec80; five coordinate-derived presence signals for microanastomosis). Temperature-1 softmax weights are restricted to the top two experts and renormalized before mixing logits. Every expert is evaluated, so sparse weights do not reduce encoder or tower computation. The archived objective adds a positive entropy term with weight 0.01 to the classification loss; minimizing it favors concentrated usage rather than load balancing. Causality applies to the gate and tower operations on cached features, not to the full pipeline, because video features use centered clips.

The uniform ensemble averages six towers of the same design, and each single-expert baseline trains one tower alone. The gate adds parameters and tool inputs; total capacity is therefore not identical across variants. We compare the router with both controls on Cholec80 and microanastomosis. AutoLaparo is excluded because it has no aligned tool stream.

Table 7: Measured router results (macro-F1 / accuracy, %). All variants share the same trainer and expert towers; the router adds tool-bit gate input and top-2 sparsification.

	Cholec80		Microanastomosis	
Variant	mF1	Acc.	mF1	Acc.
SurgRouter (top-2 MoE)	68.3	77.6	53.6	73.1
Uniform ensemble	71.7	81.6	51.8	72.1
Expert alone: DINOv2 ViT-B/14	60.4	71.9	53.6	70.9
Expert alone: MAE ViT-B	51.0	63.6	22.9	65.2
Expert alone: VideoMAE-B	42.4	56.3	14.8	60.3
Expert alone: TimeSformer	68.8	78.0	51.9	71.7
Expert alone: VideoMamba-M	67.6	78.0	45.4	69.1
Expert alone: EndoMamba	31.2	44.8	18.6	60.9

D.2 Fusion results

Table 7 shows the outcome. On Cholec80 the uniform ensemble is best (71.7 macro-F1), ahead of the best single expert (TimeSformer, 68.8) and the learned router (68.3). On microanastomosis the router ties the best single expert (53.6, DINOv2) and beats the uniform ensemble (51.8). Across both datasets the learned gate never dominates both static controls: with six towers with the same temporal design over frozen features, the observed gate provides no clear advantage over the strongest static control. This does not establish a performance ceiling for other routers.