ICL-SURG: IN-CONTEXT LEARNING WITH LOGIC REFINEMENT FOR SURGICAL PHASE RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Few-shot surgical phase recognition aims to segment surgical videos into semantically meaningful phases using only a few annotated samples. However, the scarcity of labeled data limits existing methods in capturing the temporal structure and discriminative patterns of surgical workflows. To address these challenges, we present **ICL-Surg**, a novel framework that combines in-context learning with workflow-guided logic refinement. Given a single annotated demonstration video, ICL-Surg extracts complementary local temporal and global scene representations and adaptively fuses them via a Dynamic Confidence-aware Fusion (DCF) mechanism, which balances local and global similarities according to modality confidence. For each segment in a query video, phase labels are assigned by weighted cosine similarity to the demonstration. To enforce temporal coherence and medical plausibility, we present a Workflow-guided Logic Refinement (WLR) module that applies Viterbi decoding with dual-source priors: a local transition matrix derived from the demonstration, and empirical transition and positional statistics aggregated from a label pool, combined by linear interpolation. This module explicitly refines the predicted phase sequence against surgical logic. Extensive experiments across multiple public surgical video benchmarks demonstrate that ICL-Surg outperforms existing few-shot surgical phase recognition methods. **Code will be released upon acceptance.**

1 INTRODUCTION

Surgical phase recognition, which aims to segment surgical videos into clinically meaningful phase sequences (Garrow et al., 2021; Demir et al., 2023; Buyck et al., 2023; Guo et al., 2023), is a critical task in computer-assisted surgery. Since EndoNet (Twinanda et al., 2017) introduced a multi-task learning framework to jointly predict surgical phases and instrument usage, a series of deep learning methods have been developed for this task (Funke et al., 2025; Czempiel et al., 2020). Subsequent work incorporated temporal modeling modules to capture long-range dependencies, including LSTM-based multi-task recurrent convolutional networks (Jin et al., 2020) and segment-level semantic modeling for online recognition (Ding and Li, 2022). Despite their strong performance on benchmark datasets, these methods require frame-level annotations from expert surgeons, making the labeling process costly and difficult to scale, which limits their transferability to new procedures and institutions.

Few-shot learning offers a promising direction for alleviating the annotation bottleneck, aiming to enable rapid task adaptation with only a handful of labeled samples (Wang et al., 2020)—a setting that aligns naturally with the scarce-annotation nature of surgical phase recognition. Metric-based approaches, such as Prototypical Networks (Snell et al., 2017) and Matching Networks (Vinyals et al., 2016), have been widely adopted in image classification, while extensions to video have introduced temporal alignment modules to match the temporal structure between support and query sets (Cao et al., 2020; Zhang et al., 2021; Nguyen et al., 2022; Zhan et al., 2025). In the context of surgical video analysis, several recent efforts have begun to explore phase recognition under few-shot settings, including a cross-surgery transfer learning framework for surgical step recognition (Neimark et al., 2021), feasibility studies with limited annotations on single tasks (Bajraktari et al., 2026), and prototype-based learning with as few as a single annotated video (Yuan et al., 2025).

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.

However, directly applying these methods to surgical phase recognition presents two fundamental challenges: (i) surgical phase transitions follow clear temporal logic (Ding et al., 2024), where phases typically proceed in a fixed order and certain transitions occur far more frequently than others, yet conventional prototypical networks treat each video segment as an independent sample and fail to leverage such structural priors; and (ii) in the few-shot setting, the temporal patterns in the demonstration video may be incomplete due to variations in individual surgical practices, causing the model to produce predictions that blatantly violate medical logic, e.g., regressing to earlier phases or placing terminal phases at the beginning.

To address these challenges, we propose ICL-Surg, a novel framework for few-shot surgical phase recognition based on in-context learning with workflow-guided logic refinement. Our core idea is to leverage both local temporal features and global video representations for phase recognition. Specifically, we employ a GRU to model the frame sequence within each window, capturing local temporal dynamics, while simultaneously introducing a video-level global feature, which is obtained via element-wise max pooling over all frames of the entire video, to encode holistic scene information. These two types of features are used to compute local and global similarities, which are then adaptively fused through a Dynamic Confidence-aware Fusion (DCF) mechanism, yielding a more comprehensive representation for phase recognition.

While the above design achieves promising performance on standard datasets, we observe that in cases where phase transition patterns are complex or the demonstration video is short, the model still produces predictions that clearly violate surgical logic, suggesting that visual-feature-based few-shot methods alone cannot fully capture the underlying temporal structure of surgical videos. To this end, we introduce a Workflow-guided Logic Refinement (WLR) module that performs global sequence refinement via Viterbi decoding (Forney, 1973). Specifically, WLR estimates a local transition matrix and positional priors from the demonstration video, aggregating empirical transition and positional statistics from a label pool. These are fused via linear interpolation to serve as the transition probabilities and positional priors for Viterbi decoding. This enables global optimal path search over the predicted sequence of the target video, effectively mitigating the issue of incomplete temporal information in a single demonstration video.

Our main contributions are as follows:

- We propose ICL-Surg, a dual-branch temporal network based on in-context learning that achieves complementary modeling of local temporal dynamics and global scene semantics through joint optimization of GRU-based local modeling and global video representations, significantly improving few-shot surgical phase recognition accuracy.
- We propose a Dynamic Confidence-aware Fusion (DCF) module that adaptively fuses local and global similarities, allowing the more reliable modality to contribute more to the final prediction without requiring additional supervision.
- We present a Workflow-guided Logic Refinement (WLR) module, which applies Viterbi decoding with dual-source transition and positional priors to globally correct illogical phase transitions and temporal errors.
- Extensive experiments on multiple benchmarks demonstrate that our method consistently outperforms existing methods under few-shot settings, and ablation studies further validate the effectiveness of each key module.

2 Related Work

2.1 Surgical Phase Recognition and Few-Shot Learning

Recent architectures jointly model spatial and temporal information for surgical phase recognition. Transformer-based approaches have shown particular promise: SKiT (Liu et al., 2023), a fast Key information Transformer that achieves 92.5% accuracy on Cholec80 with O(1) temporal complexity, and LoViT (Liu et al., 2025a), a two-stage Long Video Transformer with a temporally-rich spatial feature extractor and phase transition map for modeling complex temporal dependencies. Beyond Transformers, other architectures have also advanced the field: TeCNO (Czempiel et al., 2020), a multi-stage temporal convolutional network that refines predictions through iterative temporal aggregation, and TUNeS (Funke et al., 2025), a temporal U-Net with self-attention for surgical phase

recognition. Despite these advances, existing methods predominantly rely on fully supervised learning with dense frame-level annotations, making the labeling process costly and difficult to scale, which limits their transferability to new procedures and institutions (Demir et al., 2023).

To address the annotation bottleneck, few-shot learning has emerged as a promising direction for surgical phase recognition. A cross-surgery transfer learning framework for surgical step recognition (Neimark et al., 2021) demonstrated that models can be trained on one procedure type and adapted to another with minimal additional annotation. Surgical Phase Anywhere (SPA) (Yuan et al., 2025) adapts surgical phase recognition models across institutions with minimal supervision through few-shot spatial adaptation and task-graph guided refinement. A systematic feasibility study (Bajraktari et al., 2026) demonstrated that few-shot models are viable under low-label and transfer settings. However, directly applying conventional few-shot methods to surgical phase recognition presents fundamental limitations. These methods treat each video segment as an independent sample, failing to leverage the structural temporal priors inherent in surgical workflows—phases typically follow a fixed order, with certain transitions occurring frequently and others rarely (Ding et al., 2024). Moreover, the limited demonstrations available in the few-shot setting may not fully capture the diversity of temporal patterns in surgical workflows, as individual surgical practices can vary significantly, leading to incomplete or biased transition statistics that cause the model to produce predictions violating medical logic.

2.2 In-Context Learning

In-Context Learning (ICL) refers to the paradigm where a model performs a task from input-output examples provided in the context without updating its parameters. ICL originated in language modeling, with seminal works such as Brown et al. (2020) and Min et al. (2022b). It has since been adapted to computer vision through visual prompting (Bar et al., 2022) and SegGPT (Wang et al., 2023). In medical imaging, Tyche (Rakic et al., 2024) was proposed for stochastic in-context segmentation, and EICSeg (Xie et al., 2026) introduced exemplar-instructed universal segmentation. Despite these advances, ICL for surgical video understanding remains unexplored. Surgical videos present unique challenges: they exhibit strong temporal dependencies, specialized semantic structures, and significant inter-surgeon variability, requiring reasoning over long temporal horizons and understanding procedural logic.

2.3 Meta-Learning

Meta-learning provides a framework for rapid task adaptation by training across a distribution of tasks (Finn et al., 2017). MetaICL (Min et al., 2022a) demonstrated that meta-training can significantly improve in-context learning performance, and further investigated what makes demonstrations effective, finding that the format of demonstrations plays a more critical role than the specific input-label mapping (Min et al., 2022b). In surgical phase recognition, Meta-SurDiff (Liu et al., 2025b) introduced a meta-learning-optimized classification diffusion model that achieves precise frame-level distribution estimation for online recognition. While meta-learning and ICL share the goal of few-shot adaptation, they differ fundamentally: meta-learning relies on parameter updates during adaptation (Finn et al., 2017), whereas ICL performs task adaptation without any gradient updates (Brown et al., 2020). Meta-learning typically requires gradient computation at test time, which can be computationally expensive, while ICL performs adaptation purely through forward computation conditioned on demonstrations. Our work adopts the ICL paradigm while drawing inspiration from meta-learning's episodic training strategy (Min et al., 2022a) to prepare the model for in-context learning during training.

2.4 Temporal Modeling and Structured Prediction

Temporal coherence in phase predictions has been addressed through structured prediction techniques. Neural Finite-State Machines (Ding et al., 2024) explicitly model phase transitions as state transitions in a finite automaton. CoStoDet-DDPM (Yang et al., 2025) proposed a collaborative training framework for surgical workflow anticipation and recognition.

Viterbi decoding (Forney, 1973) has been applied to refine surgical phase predictions by incorporating transition priors (Lalys et al., 2010; 2011). However, these approaches typically estimate

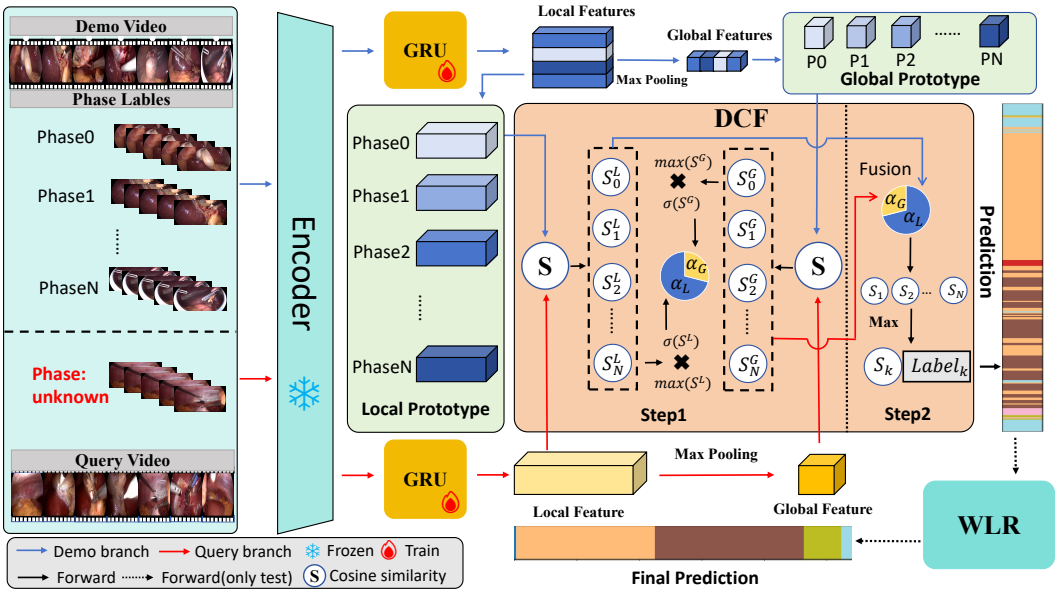


Figure 1: **Overview of ICL-Surg**. Demo and query videos are encoded by PeskaVLP and GRU into local and global features. Demo prototypes are computed per phase class, and query-to-prototype similarities are fused by DCF using an adaptive weight α to yield initial predictions, which WLR refines into the final phase sequence.

transition probabilities either from the training set or from the test video. In the few-shot setting, local statistics are sparse and biased, while empirical statistics from a label pool may not reflect the specific workflow of the target procedure. Our work addresses this by combining local transition statistics from the demonstration video with empirical statistics from a label pool, along with positional priors that capture the typical temporal location of each phase.

3 Method

3.1 Overview and Problem Formulation

The overall framework of ICL-Surg is illustrated in Fig. 1. A demo video and a query video are encoded by a frozen PeskaVLP encoder and a GRU to obtain local and global features. For the demonstration video, local and global prototypes are constructed by averaging features within each phase class. For the query video, local and global cosine similarities are computed against these prototypes. A Dynamic Confidence-aware Fusion (DCF) module fuses the two similarities in two steps: it first computes an adaptive fusion weight from modality confidence, and then combines the similarities to produce class scores and initial predictions. Subsequently, a Workflow-guided Logic Refinement (WLR) module refines these predictions into the final phase sequence.

To formalize the framework described above, we define the time window as the basic processing unit. Each surgical video is partitioned into windows of L consecutive frames. Each window is fed into a pre-trained visual encoder (PeskaVLP) to extract frame-level features, producing a window feature sequence $\mathbf{X}_t \in \mathbb{R}^{L\times D}$, where t indexes the windows, T_v is the total number of windows in the video, and D is the feature dimension. A video is therefore represented by

$$\mathcal{X}_v = \{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_{T_v}\}, \tag{1}$$

with the corresponding phase label sequence:

$$\mathcal{Y}_v = \{y_1, y_2, \ldots, y_{T_v}\}, \quad y_t \in \mathcal{C}_v, \tag{2}$$

where $\mathcal{C}_v$ is the set of phase classes present in the video.

These windows serve as the basic processing units for both training and testing: during training, a subset of windows is sampled to construct episodes, while at test time all windows of the target video are processed to produce dense predictions, which are subsequently refined as a whole sequence.

3.2 Dual-Branch Temporal Representation

We represent each temporal window with two complementary branches. The local branch uses a unidirectional GRU to capture short-term temporal dynamics. The global branch uses cumulative max pooling to encode holistic scene semantics. The two branches are complementary: the local branch preserves transition information, while the global branch summarizes instrument, tissue, and illumination context.

Local Feature Extraction. Given an input window feature sequence $\mathbf{X} \in \mathbb{R}^{L\times D}$, we employ a uni-directional Gated Recurrent Unit (GRU) (Cho et al., 2014)to model the frame sequence, obtaining $\mathbf{H} = \text{GRU}(\mathbf{X})$, and take the hidden state at the final timestep as the local temporal representation $\mathbf{h} = \mathbf{H}[-1, :] \in \mathbb{R}^D$. The GRU's recurrent structure integrates dependencies across frames, preserving dynamic transition information that would be lost by simple temporal averaging.

Global Context Representation. We compute a global scene representation $\mathbf{g} \in \mathbb{R}^D$ for the entire video via cumulative max pooling. Starting from a zero-initialized vector, we update it after each window by taking the element-wise maximum between the current global vector and the local feature of the newly processed window:

$$\mathbf{g}_t = \max(\mathbf{g}_{t-1}, \mathbf{h}_t), \quad \mathbf{g}_0 = \mathbf{0}, \tag{3}$$

where $\mathbf{h}_t$ is the GRU output of the t-th window. The final global representation $\mathbf{g}_{T_v}$ is obtained after processing all windows. This global representation encodes holistic scene semantics such as instrument types, tissue appearance, and illumination conditions. We note that element-wise max pooling is a standard practice widely adopted in surgical video analysis (Liu et al., 2023; 2025a); it is not a core contribution of this work.

3.3 Dynamic Confidence-aware Fusion

Existing approaches to fusing multiple similarity measures typically rely on fixed weights or simple averaging, which assume that each modality contributes equally across all samples. In surgical phase recognition, however, the reliability of local and global modalities can vary significantly depending on the query sample. A fixed weighting scheme may thus fail to emphasize the more reliable modality when it matters most. We therefore propose the DCF module that adaptively adjusts the fusion weight for each query sample based on the confidence of each modality.

Given the local feature $\mathbf{h}_j$ and global feature $\mathbf{g}_j$ for query sample j, and their counterparts $\mathbf{h}_i$ and $\mathbf{g}_i$ for the i-th support sample, we compute their cosine similarities as

$$\begin{cases} s_{j,i}^{\text{loc}} = \cos(\mathbf{h}_j, \mathbf{h}_i), \\ s_{j,i}^{\text{glo}} = \cos(\mathbf{g}_j, \mathbf{g}_i). \end{cases} \tag{4}$$

We then define the confidence of the local modality for query sample j as follows:

$$c_j^{\text{loc}} = \sigma(\{s_{j,i}^{\text{loc}}\}_{i\in\mathcal{I}_c}) \cdot \max_{i\in\mathcal{I}_c} s_{j,i}^{\text{loc}}, \tag{5}$$

where $\sigma(\cdot)$ denotes the standard deviation and $\mathcal{I}_c$ is the set of support sample indices for the c-class. Then, the confidence of the global modality c_j^{glo} is defined analogously.

The dynamic fusion weight is then defined by

$$\alpha_j = \frac{c_j^{\text{loc}}}{c_j^{\text{loc}} + c_j^{\text{glo}} + \epsilon_1}, \tag{6}$$

where ϵ_1 is a small constant for numerical stability. The final fused similarity is then computed by

$$s_{j,i} = \alpha_j \cdot s_{j,i}^{\text{loc}} + (1 - \alpha_j) \cdot s_{j,i}^{\text{glo}}. \tag{7}$$

The score of query sample j for class c is the maximum over all support samples of that class:

$$s_{j,c} = \max_{i \in \mathcal{I}_c} s_{j,i}. \tag{8}$$

3.4 Episodic Meta-Training and In-Context Inference

Episode Construction. At each training iteration, we randomly draw two distinct videos: one serves as the demonstration video and the other as the query video. We select at least five phase categories from the demonstration video and sample N_S windows per category to form the support set $\mathcal{S}$. From the query video, we retain only windows belonging to categories that appear in the demonstration video, and sample N_Q windows per category to form the query set $\mathcal{Q}$; if a category has fewer than N_Q windows, all of them are used. This yields a meta-task $\mathcal{T} = \{\mathcal{S}, \mathcal{Q}\}$.

This cross-video construction is essential. It requires the model to extract contextual information from the demonstration video and use it to classify windows from a different video, which mirrors the test-time setting where only one demonstration video is available.

Loss Function. The training objective minimizes the negative log-likelihood loss over all query samples:

$$\mathcal{L} = -\frac{1}{|\mathcal{Q}|} \sum_{j \in \mathcal{Q}} \log \frac{\exp(s_{j,y_j}/\tau)}{\sum_{c=1}^{C} \exp(s_{j,c}/\tau)}, \tag{9}$$

where τ is a temperature parameter, y_j is the ground-truth phase of query window j, and $s_{j,c}$ is its score for class c. Only the GRU parameters are updated during meta-training; the PeskaVLP encoder remains frozen.

In-Context Inference. After meta-training, phase recognition is performed by in-context inference without any parameter updates. Given a target video and a single demonstration video, all windows of the demonstration video are encoded into local and global features, which together form the context. For each window in the target video, the model computes its fused similarity to every context window through DCF and assigns the phase label of the most similar context window.

3.5 Workflow-guided Logic Refinement

DCF produces per-window class scores independently and does not enforce temporal consistency across the video. Surgical workflows, however, follow a largely ordered phase sequence, and a single illogical transition can substantially degrade recognition accuracy. The WLR module addresses this by refining the initial predictions with three sources of workflow knowledge: transition statistics, positional priors, and observation probabilities from DCF. An overview of the proposed WLR module is presented in Fig. 2.

Transition Prior Modeling. We first estimate a local transition matrix $\mathbf{M}^{\text{demo}} \in \mathbb{R}^{C \times C}$ from the demonstration video:

$$M_{i,j}^{\text{demo}} = \frac{N_{i \to j}^{\text{demo}} + \epsilon_2}{\sum_k (N_{i \to k}^{\text{demo}} + \epsilon_2)}, \tag{10}$$

where $N_{i \to j}$ is the frequency of transitions from phase i to phase j, and ϵ_2 is a small smoothing constant.

However, relying solely on a single demonstration video may yield unreliable estimates due to individual surgical variations. To improve robustness, we pre-compute an empirical transition matrix $\mathbf{M}^{\text{emp}}$ from the entire training set. The final transition matrix is obtained via linear interpolation:

$$\mathbf{M} = \lambda \cdot \mathbf{M}^{\text{demo}} + (1 - \lambda) \cdot \mathbf{M}^{\text{emp}}, \tag{11}$$

where $\lambda \in [0, 1]$ balances the influence of local and empirical statistics. This combination retains the procedure-specific ordering of the demonstration while borrowing statistical strength from the label pool.

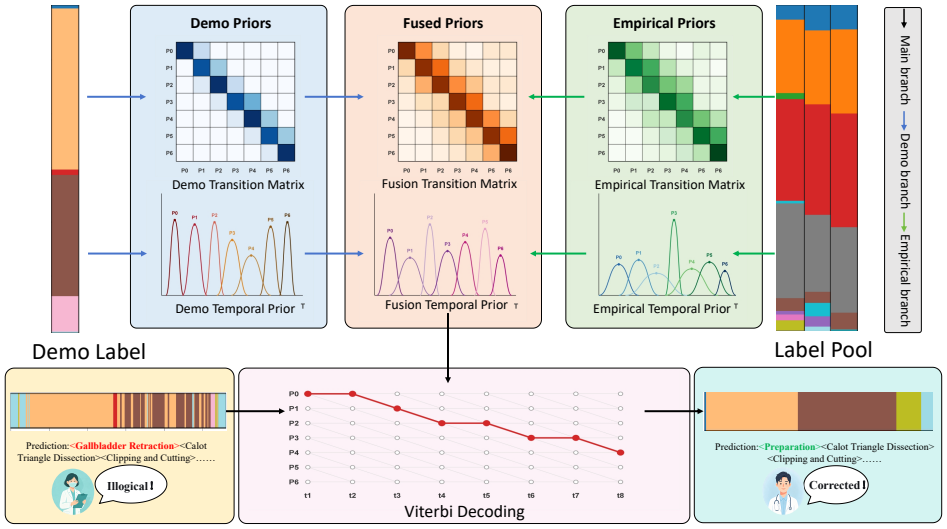


Figure 2: Overview of the proposed workflow-guided logic refinement module. The module estimates transition matrices and positional priors from the labels of both the demonstration video and a label pool that aggregates phase label sequences from all training videos, and fuses the corresponding estimates via linear interpolation. The fused transition matrix, fused positional priors, and observation probabilities from DCF are then jointly fed into Viterbi decoding to search for a globally optimal phase sequence that conforms to surgical workflow logic.

Positional Prior Modeling. Beyond transition order, surgical phases also exhibit strong positional priors — certain phases are highly likely to appear at specific positions along the video timeline. We model the positional distribution of each phase on the normalized time axis using a Gaussian distribution. For the c-class in the demonstration video, we have

$$\mu_c^{\text{demo}} = \frac{1}{|\mathcal{I}_c^{\text{demo}}|} \sum_{t \in \mathcal{I}_c^{\text{demo}}} \frac{t}{T_{\text{demo}}}, \tag{12}$$

$$\sigma_c^{\text{demo}} = \sqrt{\frac{1}{|\mathcal{I}_c^{\text{demo}}|} \sum_{t \in \mathcal{I}_c^{\text{demo}}} \left(\frac{t}{T_{\text{demo}}} - \mu_c^{\text{demo}} \right)^2} + \delta, \tag{13}$$

where $\mathcal{I}_c^{\text{demo}}$ is the set of window indices for class c, T_{demo} is the total number of windows, and δ prevents zero variance.

Empirical positional priors $(\mu_c^{\text{emp}}, \sigma_c^{\text{emp}})$ are similarly computed from the training set. The two are fused via interpolation:

$$\rho(c,t) = \lambda_{\text{pos}} \cdot \mathcal{N}\left(\frac{t}{T}; \mu_c^{\text{demo}}, \sigma_c^{\text{demo}} \right) + (1 - \lambda_{\text{pos}}) \cdot \mathcal{N}\left(\frac{t}{T}; \mu_c^{\text{emp}}, \sigma_c^{\text{emp}} \right), \tag{14}$$

where t denotes the temporal position of the current window, T is the target video length, and λ_{pos} is the interpolation coefficient.

Viterbi Decoding with Logic Constraints. Given the observation probability matrix $\mathbf{O} \in \mathbb{R}^{T \times C}$ from the initial predictions, the fused transition matrix $\mathbf{M}$, and the positional prior ρ, we seek the optimal hidden state sequence $\hat{\mathbf{y}} = \{\hat{y}_1, \ldots, \hat{y}_T\}$ that maximizes the joint probability:

$$\hat{\mathbf{y}} = \arg\max_{\mathbf{y}} \left[\sum_{t=1}^{T} \log O_t(y_t) + \sum_{t=1}^{T} \log \rho(y_t, t) + \sum_{t=2}^{T} \log M_{y_{t-1}, y_t} \right]. \tag{15}$$

The optimal sequence is obtained by solving the above maximization problem via the Viterbi algorithm in $O(T \cdot C^2)$ time, yielding a globally refined prediction that balances visual evidence, transition plausibility, and positional priors.

Table 1: Performance comparison on four surgical datasets. Best results are in bold, and second-best are underlined.

Method	Cholec80	AutoLaparo	BernBypass	StrasBypass
D^{2}ST-Adapter (Pei et al., 2025)	37.55	<u>52.83</u>	<u>58.84</u>	42.83
TEAM (Lee et al., 2025)	<u>49.61</u>	52.38	45.06	<u>52.90</u>
FSL (Bajraktari et al., 2026)	45.76	40.53	47.92	43.94
ICL-Surg (Ours)	**68.14**	**58.72**	**62.77**	**70.84**

Table 2: Comparison with additional baselines on four surgical datasets. Best results are in bold, and second-best are underlined.

Method	Shot	Cholec80	AutoLaparo	BernBypass	StrasBypass
ImageNet (He et al., 2016)	10	57.4	44.9	53.3	53.3
Moco (He et al., 2020)	10	57.6	49.9	59.1	63.1
CLIP (Radford et al., 2021)	10	57.5	46.2	52.8	54.3
SurgVLP (Yuan et al., 2025)	10	55.0	48.5	57.7	57.0
PeskaVLP (Yuan et al., 2024)	10	61.9	53.1	<u>62.9</u>	63.8
HyperVLP (Hu et al., 2026)	10	<u>64.2</u>	**60.1**	**72.3**	<u>66.9</u>
ICL-Surg (Ours)	**1 video ($\leq$5)**	**68.14**	<u>58.72</u>	62.77	**70.84**

4 Experiments

4.1 Experimental Setup

Datasets and Evaluation Metrics. We use seven publicly available surgical video datasets covering different surgical specialties: Cholec80 (Twinanda et al., 2017), Cataract-101 (Schoeffmann et al., 2018), AutoLaparo (Wang et al., 2022), BernBypass, StrasBypass(Lavanchy et al., 2024), ArthroPhase (Malayeri et al., 2025), and TMVP-SurVideo (Lan et al., 2025). In each experiment, one dataset is held out as the test set and the remaining six are used for meta-training. This ensures that the model is evaluated on procedure types unseen during training. Four datasets are evaluated as test targets in separate runs: Cholec80, AutoLaparo, BernBypass, and StrasBypass. We evaluate all methods using frame-level accuracy.

Implementation Details. The PeskaVLP visual encoder is frozen during training. Episodes are configured with $N = 5$ classes, $K = 5$ support samples, and 5 queries per class. We train for 30 epochs with 100 episodes per epoch using Adam and a learning rate of 1×10^{-4} with temperature $\tau = 0.1$. Besides, L and D are set to 16 and 768, respectively. The interpolation coefficients for refinement are set to $\lambda = 0.6$ for the transition matrix and $\lambda_{\text{pos}} = 0.5$ for the positional prior.All results are averaged over 5 runs with different random seeds. Experiments are conducted on a single NVIDIA RTX 3090 GPU.

4.2 Comparison with State-of-the-art Methods

To validate the effectiveness of the proposed model, we compare ICL-Surg with three representative few-shot learning methods on the four test datasets, including FSL (Bajraktari et al., 2026), TEAM (Lee et al., 2025), and D^{2}ST-Adapter (Pei et al., 2025). As shown in Table 1, ICL-Surg achieves the highest accuracy on all four datasets, substantially outperforming all baselines. Notably, whereas the baselines exhibit considerable variance across datasets, ICL-Surg maintains consistently strong performance across all four procedure types. The comparison results are visualized in Fig. 3. As can be observed, our model achieves more accurate surgical phase prediction than the other comparison methods.

We further compare with additional baselines under the 10-shot setting as reported in HyperVLP (Hu et al., 2026), as shown in Table 2. Notably, our method uses only a single demonstration video per dataset ($\leq$5 shots per class), a stricter and more practical setting than all baselines. Despite using far fewer annotations, ICL-Surg achieves the highest accuracy on Cholec80 and StrasBypass, and

Table 3: Ablation study on each key component in the proposed model. Results are reported as mean ± std over 5 runs.

Dataset	Local	Global	LSF	DCF	Full
Cholec80	50.30 ± 1.60	54.50 ± 5.23	55.62 ± 2.09	59.49 ± 2.30	**68.14 ± 1.52**
AutoLaparo	43.79 ± 5.32	28.60 ± 7.38	46.94 ± 3.30	49.12 ± 4.19	**58.72 ± 2.04**
BernBypass	50.04 ± 4.35	33.11 ± 3.83	49.67 ± 2.22	52.31 ± 1.67	**62.77 ± 2.33**
StrasBypass	56.31 ± 1.66	25.55 ± 2.91	58.48 ± 1.72	57.99 ± 1.59	**70.84 ± 1.30**

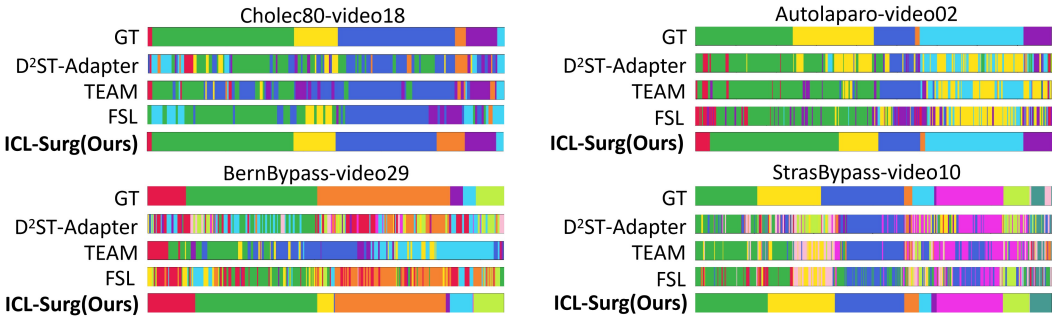


Figure 3: Qualitative comparison between ICL-Surg and other comparison methods.

remains competitive on AutoLaparo and BernBypass. This demonstrates the strong applicability of our method, as it delivers superior or comparable performance with substantially less supervision. This comparison is therefore conservative in favor of the baselines.

4.3 Ablation Study

To systematically evaluate the contribution of each component in ICL-Surg, we conduct ablation experiments by progressively adding modules to the local-only baseline. We consider five configurations: (i) **Local**, which uses only local features for classification, serving as the baseline; (ii) **Global**, which uses only global features; (iii) **LSF**, which combines local and global similarities via a learnable scalar fusion parameter; (iv) **DCF**, which replaces the scalar fusion with the proposed Dynamic Confidence-aware Fusion; and (v) **Full**, which adds WLR on top of DCF. As shown in Table 3, DCF outperforms LSF on most datasets, with the largest gain on Cholec80 (+3.87%), indicating that per-sample adaptive fusion is more effective than a fixed scalar weighting scheme. The only exception is StrasBypass (-0.49%), where one modality remains consistently more reliable. Adding WLR on top of DCF yields substantial improvements across all datasets, ranging from +8.65% on Cholec80 to +12.85% on StrasBypass, with the largest gains on the two bypass datasets, where phase transitions are more complex and variable. These corrections are particularly important for procedures with strict phase ordering, such as gastric bypass, where even a single illogical transition can substantially degrade accuracy. These results suggest that DCF and WLR address two different failure modes: unreliable modality weighting and illogical phase transitions. Correcting both is necessary for robust phase recognition.

5 Conclusion

We proposed ICL-Surg, a framework for few-shot surgical phase recognition that integrates in-context learning with workflow-guided logic refinement. Leveraging a single demonstration video as context, ICL-Surg adaptively fuses local and global similarities via DCF and refines predictions via WLR using dual-source priors. Experiments on four surgical datasets demonstrate that ICL-Surg achieves strong performance, outperforming baselines under stricter annotation settings. Overall, ICL-Surg offers a practical solution for few-shot surgical phase recognition.

AI USE STATEMENT

In this work, we used generative AI tools only for language polishing and improving readability. We did not use them for any tasks requiring disclosure (e.g., generating synthetic data, developing theoretical models, formulating or proving mathematical claims, proposing hypotheses, designing experiments, implementing methods, translation, data cleaning, qualitative analysis, or interpreting results); all other listed tasks are not applicable. We have reviewed all AI-assisted text and take responsibility for the final content, including text and claims.

ETHICS STATEMENT

This work uses only publicly available clinical video datasets that have undergone the corresponding ethics review and were released for research purposes. We did not collect any new human-subject data, and no additional ethics approval was required for this study. We use these datasets strictly in accordance with their original licenses and terms of use and do not redistribute any raw data. We are not aware of any conflicts of interest or privacy, security, or fairness concerns related to this work.

REPRODUCIBILITY STATEMENT

We have made efforts to ensure reproducibility. The method is described in detail in Sec. 3. The appendix provides dataset splits, preprocessing, network architecture, training hyperparameters, evaluation protocol, and Viterbi decoding details. Code will be released upon acceptance. All datasets are publicly available and used under their original licenses; we do not redistribute raw data.

REFERENCES

Bajraktari, F., Asmußen, R., Giacoppo, G. A., and Pott, P. P. Few-shot learning for surgical phase recognition: performance and generalization in cholecystectomy. *Computer Methods and Programs in Biomedicine*, 282:109386, 2026.

Bar, A., Gandelsman, Y., Darrell, T., Globerson, A., and Efros, A. A. Visual prompting via image inpainting. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Brown, T. B., Mann, B., Ryder, N., et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems (NeurIPS)*, 33: 1877–1901, 2020.

Buyck, F., Vandemeulebroucke, J., Ceranka, J., Van Gestel, F., Cornelius, J. F., Duerinck, J., and Bruneau, M. Computer-vision based analysis of the neurosurgical scene – a systematic review. *Brain and Spine*, 3:102706, 2023.

Cao, K., Ji, J., Cao, Z., et al. Few-shot video classification via temporal alignment. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10618–10627, 2020.

Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., and Bengio, Y. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1724–1734, 2014.

Czempiel, T., Paschali, M., Keicher, M., Simson, W., Feussner, H., Kim, S. T., and Navab, N. TeCNO: surgical phase recognition with multi-stage temporal convolutional networks. *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 343–352, 2020.

Demir, K. C., Schieber, H., Weise, T., May, M., Maier, A., and Yang, S. H. Deep learning in surgical workflow analysis: a review of phase and step recognition. *IEEE Journal of Biomedical and Health Informatics*, 27(11):5405–5417, 2023.

Ding, X., and Li, X. Exploring segment-level semantics for online phase recognition from surgical videos. *IEEE Transactions on Medical Imaging*, 41(11):3309–3319, 2022.

Ding, H., Gao, Z., Planche, B., Luan, T., Sharma, A., Zheng, M., Lou, A., Chen, T., Unberath, M., and Wu, Z. Neural finite-state machines for surgical phase recognition. *arXiv*, 2024. arXiv: 2411.18018.

Finn, C., Abbeel, P., and Levine, S. Model-agnostic meta-learning for fast adaptation of deep networks. *International Conference on Machine Learning (ICML)*, pp. 1126–1135, 2017.

Forney, G. D. The Viterbi algorithm. *Proceedings of the IEEE*, 61(3):268–278, 1973.

Funke, I., Rivoir, D., Krell, S., and Speidel, S. TUNeS: a temporal U-Net with self-attention for video-based surgical phase recognition. *IEEE Transactions on Biomedical Engineering*, 72(7):2105–2119, 2025.

Garrow, C. R., Kowalewski, K.-F., Li, L., Wagner, M., Schmidt, M. W., Engelhardt, S., Hashimoto, D. A., Kenngott, H. G., Bodenstedt, S., Speidel, S., et al. Machine learning for surgical phase recognition: a systematic review. *Annals of Surgery*, 273(4):684–693, 2021.

Guo, K., Tao, H., Zhu, Y., Li, B., Fang, C., Qian, Y., and Yang, J. Current applications of artificial intelligence-based computer vision in laparoscopic surgery. *Laparoscopic, Endoscopic and Robotic Surgery*, 6(3):91–96, 2023.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

He, K., Fan, H., Wu, Y., Xie, S., and Girshick, R. Momentum contrast for unsupervised visual representation learning. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9729–9738, 2020.

Hu, Y., Yuan, K., Navab, N., Ying, H., Wu, J., and Padoy, N. HyperVLP: Enhancing Hierarchical Surgical Video-Language Pre-training in Hyperbolic Space. *arXiv preprint arXiv:2606.31245*, 2026.

Jin, Y., Li, H., Dou, Q., et al. Multi-task recurrent convolutional network with correlation loss for surgical video analysis. *Medical Image Analysis*, 59:101572, 2020.

Lalys, F., Riffaud, L., Morandi, X., and Jannin, P. Automatic phases recognition in pituitary surgeries by microscope images classification. *International Conference on Information Processing in Computer-Assisted Interventions (IPCAI)*, 2010.

Lalys, F., Riffaud, L., Morandi, X., and Jannin, P. Surgical phases detection from microscope videos by combining SVM and HMM. *International Workshop on Machine Learning in Medical Imaging (MLMI)*, 2011.

Lan, M., Yan, X., Si, W., and Li, X. A new dataset and versatile multi-task surgical workflow analysis framework for thoracoscopic mitral valvuloplasty. *Medical Image Analysis*, 2025.

Lavanchy, J. L., Ramesh, S., Dall'Alba, D., Gonzalez, C., Fiorini, P., Müller-Stich, B. P., Nett, P. C., Marescaux, J., Mutter, D., and Padoy, N. Challenges in multi-centric generalization: phase and step recognition in Roux-en-Y gastric bypass surgery. *International Journal of Computer Assisted Radiology and Surgery*, 19(11):2249–2257, 2024.

Lee, S., Moon, W., Seong, H. S., and Heo, J.-P. Temporal alignment-free video matching for few-shot action recognition. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5412–5421, 2025.

Liu, Y., Huo, J., Peng, J., et al. SKiT: a fast key information video transformer for online surgical phase recognition. *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21017–21027, 2023.

Liu, Y., Boels, M., Garcia-Peraza-Herrera, L. C., et al. LoViT: long video transformer for surgical phase recognition. *Medical Image Analysis*, 99: 103366, 2025.

Liu, Y., Boels, M., and Grammatikopoulou, M. Meta-SurDiff: meta-learning optimized surgical workflow recognition via classification diffusion models. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Malayeri, A. B., Seibold, M., Cavalcanti, N. A., et al. ArthroPhase: a novel dataset and method for phase recognition in arthroscopic video. *Computer Assisted Surgery*, 30(1): 2508144, 2025.

Min, S., Lewis, M., Zettlemoyer, L., and Hajishirzi, H. MetaICL: learning to learn in context. *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pp. 2791–2809, 2022.

Min, S., Lyu, X., Holtzman, A., et al. Rethinking the role of demonstrations: what makes in-context learning work? *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 11048–11064, 2022.

Neimark, D., Bar, O., Zohar, M., Hager, G. D., and Asselmann, D. "Train one, classify one, teach one" – cross-surgery transfer learning for surgical step recognition. *Proceedings of Machine Learning Research (MIDL)*, 143:532–544, 2021.

Nguyen, K. D., Tran, Q.-H., Nguyen, K., Hua, B.-S., and Nguyen, R. Inductive and transductive few-shot video classification via appearance and temporal alignments. *European Conference on Computer Vision (ECCV)*, 2022.

Pei, W., Tan, Q., Lu, G., Tian, J., and Yu, J. D^2ST-Adapter: Disentangled-and-deformable spatio-temporal adapter for few-shot action recognition. *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11317–11326, 2025.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. Learning transferable visual models from natural language supervision. *International Conference on Machine Learning (ICML)*, pp. 8748–8763, 2021.

Rakic, M., Wong, H. E., Gonzalez Ortiz, J. J., Cimini, B. A., Guttag, J. V., and Dalca, A. V. Tyche: stochastic in-context learning for medical image segmentation. *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2024.

Schoeffmann, K., Taschwer, M., Sarny, S., Münzer, B., Primus, M. J., and Putzgruber, D. Cataract-101: video dataset of 101 cataract surgeries. *Proceedings of the 9th ACM Multimedia Systems Conference (MMSys)*, pp. 421–425, 2018.

Snell, J., Swersky, K., and Zemel, R. Prototypical networks for few-shot learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Twinanda, A. P., Shehata, S., Mutter, D., et al. EndoNet: a deep architecture for recognition tasks on laparoscopic videos. *IEEE Transactions on Medical Imaging*, 36(1):86–97, 2017.

Vinyals, O., Blundell, C., Lillicrap, T., et al. Matching networks for one shot learning. *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

Wang, Y., Yao, Q., Kwok, J. T., and Ni, L. M. Generalizing from a few examples: a survey on few-shot learning. *ACM Computing Surveys*, 53(3):Article 63, 2020.

Wang, Z., Lu, B., Long, Y., Zhong, F., Cheung, T. H., Dou, Q., and Liu, Y. AutoLaparo: a new dataset of integrated multi-tasks for image-guided surgical automation in laparoscopic hysterectomy. *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 486–496, 2022.

Wang, X., Zhang, X., Cao, Y., Wang, W., Shen, C., and Huang, T. SegGPT: towards segmenting everything in context. *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1130–1140, 2023.

Xie, S., Zhang, L., Niu, Z., Ye, F., Zhong, Q., Xie, D., Chen, Y.-W., and Lin, L. EICSeg: Universal Medical Image Segmentation via Explicit In-Context Learning. *IEEE Transactions on Medical Imaging*, 45(1):53–69, 2026.

Yang, K., Li, X., Li, Q., and Wang, Z. CoStoDet-DDPM: collaborative training of stochastic and deterministic models improves surgical workflow anticipation and recognition. *arXiv*, 2025.

Yuan, K., Srivastav, V., Navab, N., and Padoy, N. Procedure-aware surgical video-language pretraining with hierarchical knowledge augmentation. *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Yuan, K., Srivastav, V., Yu, T., Lavanchy, J. L., Marescaux, J., Mascagni, P., Navab, N., and Padoy, N. Learning multi-modal representations by watching hundreds of surgical video lectures. *Medical Image Analysis*, 2025.

Yuan, K., Chen, T., Li, S., Lavanchy, J. L., et al. Recognizing surgical phases anywhere: few-shot test-time adaptation and task-graph guided refinement. *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2025.

Zhan, Y., Pan, Y., Wu, S., Zhang, Y., and Sun, X. Hierarchical task-aware temporal modeling and matching for few-shot action recognition. *Neurocomputing*, 624:129467, 2025.

Zhang, S., Zhou, J., and He, X. Learning implicit temporal alignment for few-shot video classification. *International Joint Conference on Artificial Intelligence (IJCAI)*, 2021.

A Dataset Details

We use seven publicly available surgical video datasets covering different surgical specialties. Table 4 summarizes the number of videos, number of phase classes, and the role of each dataset. We adopt a leave-one-dataset-out protocol:in each experiment, one dataset is held out as the test set, and the remaining six datasets are used for meta-training.We report test results on four datasets:Cholec80, AutoLaparo, BernBypass, and StrasBypass. The other three datasets (Cataract-101, ArthroPhase, and TMVP-SurVideo) are used exclusively for meta-training and are never used as test targets.

Table 4: Statistics of the seven surgical video datasets.

Dataset	Videos	Phases	Role
Cholec80 (Twinanda et al., 2017)	80	7	Test target
Cataract-101 (Schoeffmann et al., 2018)	101	10	Training only
AutoLaparo (Wang et al., 2022)	21	7	Test target
BernBypass (Lavanchy et al., 2024)	70	12	Test target
StrasBypass (Lavanchy et al., 2024)	70	12	Test target
ArthroPhase (Malayeri et al., 2025)	27	5	Training only
TMVP-SurVideo (Lan et al., 2025)	23	11	Training only

All datasets are split at the video level, ensuring no overlap between training and test videos. BernBypass and StrasBypass are two subsets of the MultiBypass140 dataset, corresponding to the Bern and Strasbourg medical centers, respectively. For TMVP-SurVideo, we use a subset of 23 videos with 11 phases for meta-training only.

B Additional Implementation Details

We provide additional details not covered in the main text. Each surgical video is partitioned into windows of $L = 16$ consecutive frames with a stride of 8. The DCF module uses $\epsilon_1 = 10^{-8}$ for numerical stability. The WLR module uses $\epsilon_2 = 10^{-3}$ for Laplacian smoothing and $\delta = 0.01$ to prevent zero variance. The GRU contains approximately 3.5M trainable parameters.

C Reproducibility Statement

We have made efforts to ensure the reproducibility of our work. The code will be released upon acceptance. All datasets used in this paper are publicly available and can be obtained from their original sources. Detailed hyperparameters and implementation settings are provided in Section 4.1 and Appendix B. We use fixed random seeds for all experiments and report the mean and standard

deviation over 5 runs. The model is implemented in PyTorch. All experiments are conducted on a single NVIDIA RTX 3090 GPU.