This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

Large-Small Model Collaboration for Zero-Shot Surgical Phase Recognition

Yiyi Zhang[1], Ying Zheng[2], Wenxin Fan[1], Yu Zhu[1], Yuchen Yuan[1(✉)], Litao Zhao[1(✉)], Zheng Li[3], and Pheng-Ann Heng[1,4]

[1] Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong, China
[2] Department of Electrical and Electronic Engineering, The Hong Kong Polytechnic University, Hong Kong, China
[3] Department of Surgery, The Chinese University of Hong Kong, Hong Kong, China
[4] Institute of Medical Intelligence and XR, The Chinese University of Hong Kong
ycyuan22@cse.cuhk.edu.hk, litaozhao@cuhk.edu.hk

Abstract. Task-specific lightweight models for surgical phase recognition excel at capturing temporal dynamics but generalize poorly under domain shift. Conversely, surgical foundation models (FMs) offer superior transferability via large-scale pretraining, yet their lack of explicit temporal modeling often yields temporally inconsistent predictions, leading to degraded performance. To exploit the complementary strengths of both paradigms, we propose **La**rge-**S**mall **T**emporal adaptation (**LaST**), a novel large-small collaborative framework that enables zero-shot adaptation to unseen clinical domains. In LaST, the FM initiates the pipeline by generating frame-level phase priors that serve as initial weak supervision. To effectively utilize these noisy phase priors, we introduce an iterative temporal refinement scheme that integrates dynamic quality control to filter reliable predictions and dual-model cross-learning to mitigate confirmation bias. Simultaneously, the lightweight model leverages its intrinsic temporal modeling ability to progressively correct inconsistent predictions and enhance overall accuracy across iterations. At the end, a cycle replay strategy is employed to close the loop: the refined, more accurate predictions are utilized as upgraded supervision signals for the subsequent iterations, fostering a self-reinforcing evolution of both label quality and model capability. Extensive experiments demonstrate that LaST achieves robust adaptation to unseen domains for zero-shot surgical phase recognition, outperforming the baseline (PeskaVLP) by 24.85%-43.17% in accuracy and even surpassing fully supervised linear probing and several state-of-the-art few-shot approaches. Codes will be released at https://github.com/YIYIZH/LaST.

Keywords: Surgical Phase Recognition · Large-Small Model Collaboration · Vision-Language Foundation Model · Zero-Shot Adaptation.

1 Introduction

Surgical phase recognition plays a pivotal role in computer-assisted intervention, offering real-time context for anomaly detection and clinical decision sup-

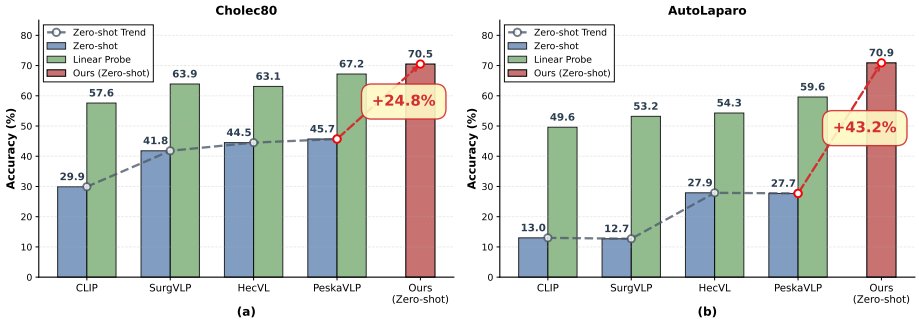

Fig. 1: Existing surgical vision-language FMs lack the precision required for clinical utility, and even supervised linear probing remains suboptimal due to the absence of temporal modeling. In contrast, our method addresses these limitations, outperforming PeskaVLP [21] by 24.8% and 43.2% without manual annotation.

port [10,25]. Recent progress has demonstrated that task-specific lightweight models can effectively capture complex temporal dependencies within surgical videos, achieving impressive performance when trained on specific procedures or datasets [6,19]. However, such models often struggle to generalize in real-world clinical environments, where workflows may vary significantly due to differences in operating room layouts, surgeon preferences, institutional protocols, and patient anatomies [8]. Addressing this domain shift typically requires extensive manual re-annotation and model retraining, which is costly, time-consuming, and impractical for widespread deployment. Although SPA [20] has been proposed to alleviate this issue, it still depends on few-shot annotations for adaptation, making it unsuitable for unseen clinical scenarios where labeled data are unavailable.

The emergence of surgical vision-language FMs [21,22,23] offers a promising direction toward improved generalization. By aligning visual features with semantic text embeddings, these models demonstrate zero-shot transferability to novel scenarios. However, direct deployment of FMs in clinical settings reveals two fundamental limitations (see Fig. 1). First, although FMs exhibit strong generalization across domains, their zero-shot predictions often fall short of clinical requirements, lacking the reliability and precision necessary for real-world surgical decision support. Second, fine-tuning approaches, such as linear probe adaptation, are often insufficient to close this gap. This is because most surgical FMs adopt CLIP-like architectures [14,26], which were originally designed for static image-text alignment. As a result, they generally fail to exploit the temporal dynamics inherent in surgical phases [10], leading to suboptimal performance even after adaptation [15].

These observations lead to a central question: *Can we design a complementary large-small collaboration in which a foundation model provides broad semantic generalization while a lightweight temporal model captures procedure-specific dynamics, enabling zero-shot adaptation to unseen clinical scenarios?*

An intuitive solution is to leverage an FM to generate initial phase predictions and treat them as weak supervision to train a specialized, temporally aware small model for improved adaptation in surgical phase recognition. However, effective large-small collaboration is non-trivial because FM's zero-shot outputs can be highly noisy and uncertain. Directly distilling these predictions risks causing the small model to overfit unreliable labels, triggering error accumulation and limiting adaptation [2]. Moreover, FM-based label filtering is unreliable across thresholds because FM confidence is not a faithful indicator of prediction quality and can be spuriously high for incorrect predictions. We therefore rely on the small model to absorb noisy supervision, limit error accumulation, and leverage temporal modeling to progressively enhance prediction quality.

To address these challenges, we propose **La**rge-**S**mall **T**emporal adaptation (**LaST**), a novel large-small collaborative framework that enables zero-shot adaptation to unseen clinical domains for surgical phase recognition. LaST operates through a self-reinforcing strategy comprising three key stages. (1) Zero-Shot Initialization: We utilize the FM to generate zero-shot predictions, serving as initial pseudo-labels for the target dataset. (2) Iterative Temporal Refinement: We first warm up two identical lightweight temporal models using the FM's coarse labels. After that, we fit a Gaussian Mixture Model (GMM) [13] to the per-sample loss distribution to distinguish clean from noisy samples in a class-wise manner at each iteration. The two temporal models are then cross-trained on peer-selected "clean" subsets, reducing confirmation bias and improving robustness to label noise. (3) Cycle Replay: After several iterations of temporal refinement, the refined predictions are fed back as higher-quality pseudo-labels for the next cycle of iterations. Repeating this loop progressively enhances pseudo-label quality and enables the small temporal models to leverage intrinsic temporal modeling capabilities to specialize to the target domain, yielding reliable phase recognition under distribution shift.

Our main contributions are threefold. (1) We reveal the inherent complementarity between large FMs and lightweight temporal models for surgical phase recognition, and propose a novel large-small collaborative framework (LaST) for zero-shot adaptation. (2) To enable effective large-small collaboration, we design an iterative temporal refinement scheme that integrates a dynamic quality control mechanism and a dual-model cross-learning protocol, together with a cycle replay strategy that forms a self-reinforcing loop. (3) Extensive experiments demonstrate the superior zero-shot adaptation capability of LaST. Specifically, it improves the FM's F1-score by 23.48% and 37.54% on Cholec80 [17] and AutoLaparo [18], respectively, and even surpasses fully supervised linear probing by 1.49% and 14.61%. LaST also outperforms the SOTA few-shot adapter, SPA, with 32 shots.

2 Method

2.1 Overview of the LaST Framework

To effectively adapt a frozen surgical Foundation Model (FM) to unseen target domains without manual annotations, we propose the **La**rge-**S**mall **T**emporal

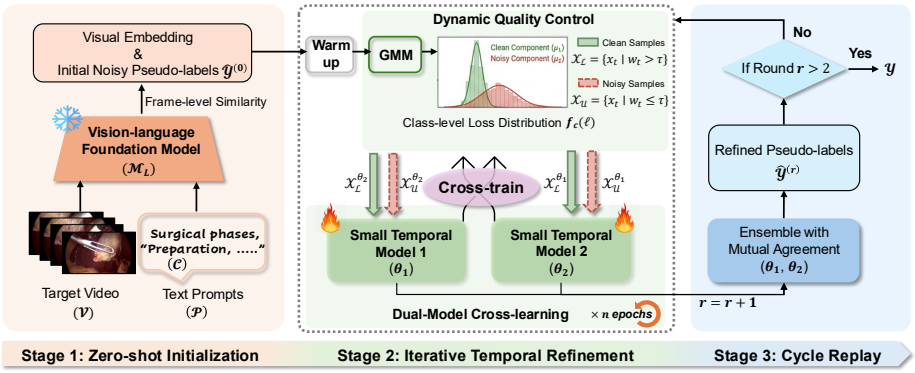

Fig. 2: Overview of the LaST framework consisting of three stages. (1) Zero-Shot Initialization: the FM generates an initial sequence of pseudo-labels for the target videos; (2) Iterative Temporal Refinement: after a warm-up step for dual small models, a dynamic quality control mechanism is used to filter high-quality samples and a cross-learning protocol is adopted to train both models; (3) Cycle Replay: the ensemble predictions from the current cycle serve as improved supervision for the next cycle. We take the predictions from the final cycle as the output.

adaptation (**LaST**) framework, as illustrated in Fig. 2. LaST establishes a self-reinforcing loop that synergizes the generalization capabilities of large FMs with the temporal modeling efficiency of lightweight networks. Let $\mathcal{V} = \{x_1, x_2, \ldots, x_T\}$ denote an unannotated surgical video sequence of length T. Our objective is to predict the corresponding sequence of phase labels $Y = \{y_1, y_2, \ldots, y_T\}$, where $y_t \in \mathcal{C}$ and $C = |\mathcal{C}|$ represents the number of surgical phase categories.

2.2 Zero-Shot Initialization

We employ a pre-trained surgical vision-language FM, $\mathcal{M}_L$, to generate initial pseudo-labels for the target video $\mathcal{V}$. For each frame $x_t \in \mathcal{V}$, $\mathcal{M}_L$ computes the similarity between the visual embedding of x_t and the text embeddings of a set of predefined prompts $\mathcal{P}$ describing the surgical phases, denoted as:

$$\hat{y}_t^{(0)} = \arg\max_{c \in \mathcal{C}} \mathrm{Sim}(\mathcal{M}_L(x_t, \mathcal{P}_c)). \tag{1}$$

Let $\hat{\mathcal{Y}}^{(0)} = \{\hat{y}_1^{(0)}, \ldots, \hat{y}_T^{(0)}\}$ denote the pseudo-label sequence for the entire video. Though $\hat{\mathcal{Y}}^{(0)}$ provides useful global semantic priors, it lacks temporal consistency and contains inherent noise, therefore can be regarded as noisy supervision signals.

2.3 Iterative Temporal Refinement

To avoid overfitting to noisy FM pseudo-labels $\hat{\mathcal{Y}}^{(0)}$, we introduce an iterative temporal refinement scheme. We first warm up two identical temporal models

with different initializations using $\hat{\mathcal{Y}}^{(0)}$, then divide samples into "clean" and "noisy" sets by a dynamic quality control mechanism. Each model is updated using the peer model's filtered clean set through dual-model cross-learning.

Dynamic Quality Control Deep networks typically learn clean samples before memorizing noise [1,3]. We therefore use the per-sample loss as a proxy for label cleanliness and estimate it in a class-wise manner to reduce class-imbalance bias, where samples from rare classes are more likely to be misclassified as noise. For a specific class c, let ℓ_t denote the loss of a sample x_t assigned the pseudo-label c. We fit a two-component Gaussian Mixture Model (GMM) [13] to the class-specific loss distribution $f_c(\ell)$:

$$f_c(\ell) = \sum_{k=1}^{2} \pi_k \mathcal{N}(\ell \mid \mu_k, \sigma_k^2), \tag{2}$$

where π_k, μ_k, and σ_k^2 denote the mixture weight, mean, and variance of component k, respectively. We assume the component with the smaller mean corresponds to the clean distribution. For each sample x_t, we compute the posterior probability [9] $w_t = P(\text{clean} \mid \ell_t)$ and partition the data into a labeled set $\mathcal{X}_L = \{x_t \mid w_t > \tau\}$ and an unlabeled set $\mathcal{X}_U = \{x_t \mid w_t \leq \tau\}$, based on a confidence threshold τ.

Dual-Model Cross-Learning To mitigate self-training confirmation bias, we jointly train two temporal networks θ_1 and θ_2 via DivideMix-inspired cross-learning [11]. Unlike DivideMix, which targets static images with the costly MixMatch strategy, we omit these operations and instead leverage the sequential nature of surgical videos. Specifically, model θ_1 is trained using the data partition $(\mathcal{X}_L^{\theta_2}, \mathcal{X}_U^{\theta_2})$ determined by the loss distribution of θ_2, and vice versa. For a given model θ_i $(i \in \{1, 2\})$, the total loss $\mathcal{L}_{total}^{\theta_i}$ comprises three terms:

1) Sparse Supervision. We apply cross-entropy loss only on the clean samples $\mathcal{X}_L^{\theta_j}$ identified by the peer network θ_j, where $\hat{y}_t$ denotes the pseudo-label of sample x_t and $p_{\theta_i}(\hat{y}_t \mid x_t)$ represents the predicted probability assigned by θ_i:

$$\mathcal{L}_{\text{sparse}}^{\theta_i} = -\frac{1}{|\mathcal{X}_L^{\theta_j}|} \sum_{x_t \in \mathcal{X}_L^{\theta_j}} \log p_{\theta_i}(\hat{y}_t \mid x_t), \quad j \neq i. \tag{3}$$

2) Temporal Smoothing. To encourage temporal continuity and reduce abrupt phase transitions, we enforce prediction consistency between adjacent frames in $\mathcal{V}$. We use a stop-gradient operator (sg) [4] to stabilize the target, ensuring the prediction of t-th frame x_t is smoothed based on its previous frame x_{t-1}:

$$\mathcal{L}_{temp}^{\theta_i} = \frac{1}{T-1} \sum_{t=2}^{T} \| \log p_{\theta_i}(\cdot \mid x_t) - \text{sg}[\log p_{\theta_i}(\cdot \mid x_{t-1})] \|_2^2 . \tag{4}$$

3). Class Balance Regularization. Under severe label noise, models may degenerate to a trivial solution (e.g., predicting a single dominant class) in

order to minimize the temporal loss. To prevent such collapse, we impose a prior distribution constraint that encourages balanced predictions across classes. Specifically, we introduce a uniform prior distribution $\phi \in \mathbb{R}^C$ with $\phi_c = \frac{1}{C}$ for all c. The regularization term is defined as the Kullback–Leibler divergence (D_{KL}) [7,16] between the prior ϕ and the empirical prediction $\bar{p}_{\theta_i}$:

$$\mathcal{L}_{reg}^{\theta_i} = D_{KL}(\phi \parallel \bar{p}_{\theta_i}) = \sum_{c=1}^{C} \phi_c \log \frac{\phi_c}{\bar{p}_{\theta_i,c}}, \quad \text{where } \bar{p}_{\theta_i} = \frac{1}{|\mathcal{V}|} \sum_{x \in \mathcal{V}} p_{\theta_i}(\cdot \mid x). \tag{5}$$

The final objective for each network is:

$$\mathcal{L}_{total}^{\theta_i} = \mathcal{L}_{sparse}^{\theta_i} + \mathcal{L}_{temp}^{\theta_i} + \mathcal{L}_{reg}^{\theta_i}. \tag{6}$$

2.4 Cycle Replay

Building upon the intra-cycle refinement, we introduce an outer-loop cycle replay strategy to progressively elevate pseudo-label quality. At the end of each cycle, we obtain refined pseudo-labels $\hat{\mathcal{Y}}^{(r)}$ by ensembling the predictions of θ_1 and θ_2. The refined sequence $\hat{\mathcal{Y}}^{(r)}$ serves as the supervision for the next cycle, replacing the previous $\hat{\mathcal{Y}}^{(r-1)}$. This creates a self-reinforcing loop wherein the models become more accurate across cycles, enabling the quality control mechanism to identify a larger and more reliable set of clean samples, which in turn leads to better models. After three rounds of cycles ($r = 3$), we take the ensemble predictions from the final cycle as the final phase output $\mathcal{Y}$ for the target video.

3 Experimental Setup

Evaluation Benchmarks: We evaluate our framework on two public benchmarks: *Cholec80* [17] and *AutoLaparo* [18]. Cholec80 comprises 80 videos of cholecystectomy surgeries, while AutoLaparo contains 21 videos of laparoscopic hysterectomy. Both datasets are annotated with 7 distinct surgical phases. We adhere to the official data split protocols for both benchmarks [17,18]. We utilize only the testing set videos for both adaptation and final evaluation, ensuring no supervision from any human annotation is accessed.

Implementation Details: The proposed LaST framework is implemented using a single RTX 3090. Following SPA [20], we utilize the frozen visual encoder from PeskaVLP [21] as the FM $\mathcal{M}_L$ for our zero-shot initialization; PeskaVLP was pretrained on surgical video lectures (SVL) [23] from open e-learning platforms. The dual small models (θ_1, θ_2) are instantiated as MS-TCN [5]. All models are optimized using the Adam optimizer with a learning rate of 5×10^{-4}. The dynamic quality control mechanism performs more reliably with relatively high confidence thresholds (τ>0.7), where $\tau = 0.9$ achieves the best performance based on sensitivity analysis. We train the framework for 50, 30, and 20 epochs across $r = 3$ rounds of cycles, respectively. The training process is lightweight, taking only 0.71 and 0.22 GPU hours on Cholec80 and AutoLaparo.

Table 1: Surgical phase recognition results on Cholec80 and AutoLaparo datasets. We report Accuracy (Acc) and average F1-Score (F1). The values in subscripts indicate the performance gap compared to LaST. **Best** and Second Best.

			Cholec80		AutoLaparo	
Model	**Shots**	**Temporal**	**Acc (%)**	**F1 (%)**	**Acc (%)**	**F1 (%)**
Fully Supervised						
TransVNet [6]	Full	✔	89.10	-	78.30	-
Linear Probe [15]	Full	✘	$67.18_{(+3.34)}$	$\underline{56.83}_{(+1.49)}$	$59.64_{(+11.22)}$	$46.41_{(+14.61)}$
Zero-Shot FMs						
SurgVLP [23]	0	✘	$41.78_{(+28.74)}$	$28.14_{(+30.18)}$	$12.74_{(+58.12)}$	$9.43_{(+51.59)}$
HecVL [22]	0	✘	$44.50_{(+26.02)}$	$25.71_{(+32.61)}$	$27.90_{(+42.96)}$	$23.33_{(+37.69)}$
PeskaVLP [21]	0	✘	$45.67_{(+24.85)}$	$34.84_{(+23.48)}$	$27.69_{(+43.17)}$	$23.48_{(+37.54)}$
Few-Shot Adapters						
Linear Probe [15]	32	✘	-	$38.37_{(+19.95)}$	-	$44.72_{(+16.30)}$
LP+Text [15]	32	✘	-	$38.36_{(+19.96)}$	-	$39.40_{(+21.62)}$
Tip-Adapter-F [24]	32	✘	-	$42.05_{(+16.27)}$	-	$39.75_{(+21.27)}$
SPA [20]	1	✔	-	$44.53_{(+13.79)}$	-	$51.48_{(+9.54)}$
SPA [20]	32	✔	-	$55.69_{(+2.63)}$	-	$\underline{60.36}_{(+0.66)}$
Ours (LaST)	**0**	✔	**70.52**	**58.32**	**70.86**	**61.02**

4 Experimental Results

We compare LaST against three categories of state-of-the-art (SOTA) methods, with results summarized in Table 1. **VS. Few-Shot Adapters:** Despite utilizing zero manual annotations, LaST remains highly competitive against SOTA few-shot methods. Notably, our zero-shot approach surpasses the 32-shot SPA in F1-score on both datasets. Furthermore, LaST significantly outperforms the 1-shot SPA variant, achieving improvements of +13.79% and +9.54%. This suggests that our self-reinforcing framework effectively mines latent supervision from the unlabeled video structure, offering a robust alternative to few-shot annotation. **VS. Zero-Shot FMs:** On the Cholec80 benchmark, LaST surpasses the strongest FM (PeskaVLP) by a substantial margin of +24.85%. The gains are even more pronounced on the challenging AutoLaparo dataset, where LaST improves upon PeskaVLP by +43.17%. These results underscore the critical limitation of applying foundation models in a naive frame-wise manner and validate the efficacy of our temporal refinement module in rectifying noisy zero-shot predictions. **VS. Fully Supervised Approaches:** While a performance gap exists between zero-shot and fully supervised upper bound (e.g., TransVNet), LaST consistently demonstrates superior efficacy against spatially supervised baselines, such as Linear Probe. This highlights the importance of temporal modeling in our method to capture the necessary procedural dynamics for surgical phase recognition.

Ablation Study: (1).We first assess the importance of **temporal modeling** in the zero-shot pipeline by comparing two strategies for exploiting temporal information: temporal-window feature averaging and MS-TCN. As illustrated in Fig. 3(a–b), simple averaging over adjacent-frame features consistently improves upon the frame-level baseline (window size 0), indicating that temporal context

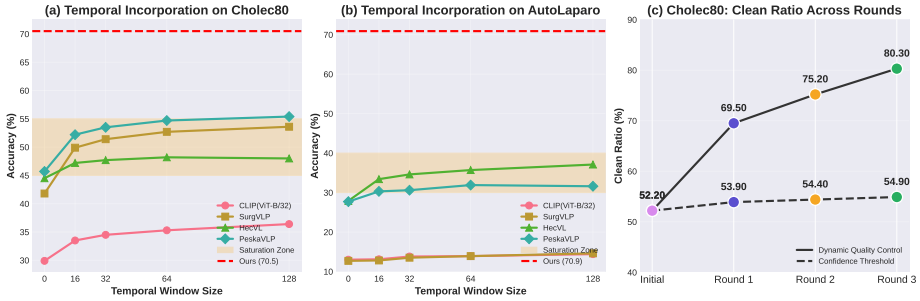

Fig. 3: Ablation study on different components of LaST.

Table 2: F1 (%) across different rounds on AutoLaparo and Cholec80 datasets.

	AutoLaparo			Cholec80		
	Round 1	Round 2	Round 3	Round 1	Round 2	Round 3
Ours w. single-model	50.81	56.60 (+5.79 ↑)	59.72 (+8.91 ↑)	46.89	51.43(+4.54 ↑)	53.80 (+6.91 ↑)
Ours w. dual-model	51.27	57.91 (+6.64 ↑)	**61.02** (+9.75 ↑)	50.58	57.84 (+7.26 ↑)	**58.32** (+7.74 ↑)

offers complementary information for surgical phase recognition. Nevertheless, the gains from this naive smoothing scheme quickly plateau, since larger windows may introduce ambiguity from neighboring phases. By contrast, our proposed method effectively overcomes this limitation through adaptive temporal modeling at the video level, yielding a clear advantage over static window-based smoothing.

(2). We further analyze the effectiveness of our **dynamic quality control mechanism** based on loss distribution modeling, in comparison with a standard label filtering baseline based on prediction confidence [12]. To measure the reliability of selected samples, we report the *Clean Ratio*, defined as the percentage of true clean samples among the set of selected "clean" pseudo-labels. Results in Fig. 3(c) highlight a clear divergence between the two strategies. On Cholec80, the baseline quickly saturates, while our method progressively recovers a larger and more accurate clean set over rounds. We observe a similar trend on AutoLaparo, although its curve is omitted for clarity. These results indicate that label filtering based on prediction confidence is unreliable under domain shift, whereas our adaptive loss modeling efficiently separates clean samples from noisy ones.

(3). We compare the **dual-model cross-learning** protocol against standard single-model training and assess the effect of **cycle replay**. The single-model variant removes peer teaching and ensembling, training a single MS-TCN on its own selected labels. As reported in Table 2, the dual-model variant achieves superior results against single-model (61.02% VS. 59.72%, 58.32% VS. 53.80%) due to cross-supervision and mutual agreement. In addition, progressively retraining on refined pseudo-labels consistently improves performance, with Round 3 yielding improvements of 9.75% and 7.74% over Round 1 with dual-model. These results demonstrate that our proposed iterative temporal refinement scheme and cycle replay strategy are robust to both single and dual modes, where model evolution and pseudo-label quality can mutually reinforce each other across rounds.

5 Conclusion

In this work, we presented **LaST**, a novel framework for zero-shot surgical phase recognition that combines the transferability of surgical FM with the temporal modeling strength of lightweight models to produce accurate predictions under domain shift. By nesting dynamic quality control and dual-model cross-learning within an Iterative Temporal Refinement loop, and further reinforcing this via a global Cycle Replay strategy, our framework enables the system to iteratively self-correct and refine noisy pseudo-labels without human supervision. Experiments show that LaST substantially outperforms zero-shot baselines and even SOTA few-shot methods that require manual annotations, highlighting large-small collaboration as a practical route for deployment in annotation-scarce settings.

Acknowledgments. This work was supported by the Research Grants Council of the Hong Kong Special Administrative Region (T45-401/22-N, C4042-23GF, 14214322, 14200623, 14203424, 14206325), the Hong Kong Innovation and Technology Fund (GHP/252/23SZ), the National Natural Science Foundation of China (62106236); and by the AIR@InnoHK Multiscale Medical Robotics Center, the Chow Yuk Ho Technology Center for Innovative Medicine, and the Li Ka Shing Institute of Health Sciences.

Disclosure of Interests. The authors have no competing interests.

References

1. Arazo, E., Ortego, D., Albert, P., O'Connor, N., McGuinness, K.: Unsupervised label noise modeling and loss correction. In: International conference on machine learning. pp. 312–321. PMLR (2019)
2. Chen, B., Jiang, J., Wang, X., Wan, P., Wang, J., Long, M.: Debiased self-training for semi-supervised learning. Advances in Neural Information Processing Systems **35**, 32424–32437 (2022)
3. Chen, P., Liao, B.B., Chen, G., Zhang, S.: Understanding and utilizing deep neural networks trained with noisy labels. In: International conference on machine learning. pp. 1062–1070. PMLR (2019)
4. Chen, X., He, K.: Exploring simple siamese representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15750–15758 (2021)
5. Farha, Y.A., Gall, J.: Ms-tcn: Multi-stage temporal convolutional network for action segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3575–3584 (2019)
6. Gao, X., Jin, Y., Long, Y., Dou, Q., Heng, P.A.: Trans-svnet: Accurate phase recognition from surgical videos via hybrid embedding aggregation transformer. In: International conference on medical image computing and computer-assisted intervention. pp. 593–603. Springer (2021)
7. Hershey, J.R., Olsen, P.A.: Approximating the kullback leibler divergence between gaussian mixture models. In: 2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07. vol. 4, pp. IV–317. IEEE (2007)

8. Jaspers, T.J., de Jong, R.L., Li, Y., Kusters, C.H., Bakker, F.H., van Jaarsveld, R.C., Kuiper, G.M., van Hillegersberg, R., Ruurda, J.P., Brinkman, W.M., et al.: Scaling up self-supervised learning for improved surgical foundation models. arXiv preprint arXiv:2501.09436 (2025)
9. Jiang, Q., Huang, B., Yan, X.: Gmm and optimal principal components-based bayesian method for multimode fault diagnosis. Computers & Chemical Engineering **84**, 338–349 (2016)
10. Jin, Y., Long, Y., Chen, C., Zhao, Z., Dou, Q., Heng, P.A.: Temporal memory relation network for workflow recognition from surgical video. IEEE Transactions on Medical Imaging **40**(7), 1911–1923 (2021)
11. Li, J., Socher, R., Hoi, S.C.: Dividemix: Learning with noisy labels as semi-supervised learning. In: International Conference on Learning Representations (2020)
12. Liu, P., Liu, J.: When confidence fails: Revisiting pseudo-label selection in semi-supervised semantic segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21874–21884 (2025)
13. Permuter, H., Francos, J., Jermyn, I.: A study of gaussian mixture models of color and texture features for image classification and segmentation. Pattern recognition **39**(4), 695–706 (2006)
14. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
15. Shakeri, F., Huang, Y., Silva-Rodríguez, J., Bahig, H., Tang, A., Dolz, J., Ben Ayed, I.: Few-shot adaptation of medical vision-language models. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 553–563. Springer (2024)
16. Tanaka, D., Ikami, D., Yamasaki, T., Aizawa, K.: Joint optimization framework for learning with noisy labels. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5552–5560 (2018)
17. Twinanda, A.P., Shehata, S., Mutter, D., Marescaux, J., De Mathelin, M., Padoy, N.: Endonet: a deep architecture for recognition tasks on laparoscopic videos. IEEE transactions on medical imaging **36**(1), 86–97 (2016)
18. Wang, Z., Lu, B., Long, Y., Zhong, F., Cheung, T.H., Dou, Q., Liu, Y.: Autolaparo: A new dataset of integrated multi-tasks for image-guided surgical automation in laparoscopic hysterectomy. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 486–496. Springer (2022)
19. Yang, S., Luo, L., Wang, Q., Chen, H.: 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)
20. Yuan, K., Chen, T., Li, S., Lavanchy, J.L., Heiliger, C., Özsoy, E., Huang, Y., Bai, L., Navab, N., Srivastav, V., et al.: Recognizing surgical phases anywhere: Few-shot test-time adaptation and task-graph guided refinement. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 467–477. Springer (2025)
21. Yuan, K., Navab, N., Padoy, N., et al.: Procedure-aware surgical video-language pretraining with hierarchical knowledge augmentation. Advances in Neural Information Processing Systems **37**, 122952–122983 (2024)
22. Yuan, K., Srivastav, V., Navab, N., Padoy, N.: Hecvl: Hierarchical video-language pretraining for zero-shot surgical phase recognition. In: International Conference

on Medical Image Computing and Computer-Assisted Intervention. pp. 306–316. Springer (2024)
23. Yuan, K., Srivastav, V., Yu, T., Lavanchy, J.L., Marescaux, J., Mascagni, P., Navab, N., Padoy, N.: Learning multi-modal representations by watching hundreds of surgical video lectures. Medical Image Analysis p. 103644 (2025)
24. Zhang, R., Fang, R., Zhang, W., Gao, P., Li, K., Dai, J., Qiao, Y., Li, H.: Tip-adapter: Training-free clip-adapter for better vision-language modeling. arXiv preprint arXiv:2111.03930 (2021)
25. Zhang, Y., Yuan, Y., Zheng, Y., Pei, J., Li, J., Li, Z., Heng, P.A.: Mejo: Mllm-engaged surgical triplet recognition via inter-and intra-task joint optimization. arXiv preprint arXiv:2509.12893 (2025)
26. Zhao, Z., Liu, Y., Wu, H., Wang, M., Li, Y., Wang, S., Teng, L., Liu, D., Cui, Z., Wang, Q., et al.: Clip in medical imaging: A comprehensive survey. arXiv preprint arXiv:2312.07353 (2023)