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.

Recognizing Surgical Phases Anywhere: Few-Shot Test-time Adaptation and Task-graph Guided Refinement

Kun Yuan[1,3], Tingxuan Chen[3], Shi Li[1], Joël L. Lavanchy[4], Christian Heiliger[5], Ege Özsoy[3], Yiming Huang[6], Long Bai[6], Nassir Navab[3], Vinkle Srivastav[1,2], Hongliang Ren[6], and Nicolas Padoy[1,2]

[1] University of Strasbourg, CNRS, INSERM, ICube, UMR7357, Strasbourg, France
[2] IHU Strasbourg, Strasbourg, France
[3] CAMP, Technische Universität München, Munich, Germany
[4] University Digestive Health Care Center – Clarunis, 4002 Basel, Switzerland
[5] Ludwig Maximilian University of Munich, Munich, Germany
[6] Chinese University of Hong Kong, Hong Kong SAR, China
kyuan@unistra.fr;npadoy@unistra.fr

Abstract. The complexity and diversity of surgical workflows, driven by heterogeneous operating room settings, institutional protocols, and anatomical variability, present a significant challenge in developing generalizable models for cross-institutional and cross-procedural surgical understanding. While recent surgical foundation models pretrained on large-scale vision-language data offer promising transferability, their zero-shot performance remains constrained by domain shifts, limiting their utility in unseen surgical environments. To address this, we introduce **S**urgical **P**hase **A**nywhere (SPA), a lightweight framework for versatile surgical workflow understanding that adapts foundation models to institutional settings with minimal annotation. SPA leverages few-shot spatial adaptation to align multi-modal embeddings with institution-specific surgical scenes and phases. It also ensures temporal consistency through diffusion modeling, which encodes task-graph priors derived from institutional procedure protocols. Finally, SPA employs dynamic test-time adaptation, exploiting the mutual agreement between multi-modal phase prediction streams to adapt the model to a given test video in a self-supervised manner, enhancing the reliability under test-time distribution shifts. SPA is a lightweight adaptation framework, allowing hospitals to rapidly customize phase recognition models by defining phases in natural language text, annotating a few images with the phase labels, and providing a task graph defining phase transitions. The experimental results show that the SPA framework achieves state-of-the-art performance in few-shot surgical phase recognition across multiple institutions and procedures, even outperforming full-shot models with 32-shot labeled data.

Keywords: Surgical data science · Vision-language · Few-shot learning.

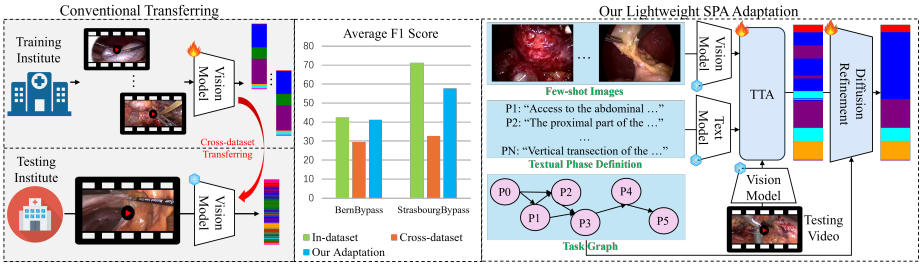

Fig. 1. Conventional adaptation struggles with performance drops across institutions due to spatial and procedural variations. Our institution-specific adaptation minimizes manual effort by requiring only textual phase definitions, a few labeled images, and a task graph from the target hospital, to adapt spatial understanding and enforce institution-specific temporal constraints.

1 Introduction

Deep learning models have made significant progress in the field of surgical data science [11, 27], particularly in surgical phase recognition [2, 23, 3], which is essential for improving patient safety [12], reducing errors, and optimizing communication in the operating room (OR) [14]. Surgical phase recognition can provide real-time insights into procedure progress, enabling early anomaly detection [13] and context-aware decision support. However, current phase recognition models face limitations when deployed in real clinical settings. Specifically, their performance often degrades in unseen institutions and procedures due to variability in surgical settings, such as different procedure protocols, OR setups [14], instrumentation [9, 8, 21], and variations in patient anatomies [4], as shown in Fig. 1. This variability necessitates costly re-annotation and retraining for each new clinical site, which hinders the scalability of phase recognition models. Therefore, there is an urgent need for generalizable and lightweight adaptation methods that can bridge these gaps without requiring extensive re-annotation or retraining, ensuring broader real-world applicability.

Few-shot transfer learning [17, 19, 28] with foundation models [15, 25] is a promising approach to address domain shifts in surgical tasks, particularly variations in instruments and anatomies. By leveraging pretrained representations from large-scale datasets such as SVL-Pretrain [27] and fine-tuning with a few labeled samples, it reduces reliance on extensive annotations. However, existing few-shot studies rely on fine-tuning with a significant fraction of data (e.g., 12.5% or 25% in [16]), requiring hundreds to thousands of labeled samples. This assumption is impractical in clinical applications. Standard few-shot learning [19, 28] with K-shot N-class settings often overfits on surgical video datasets, which are limited compared to radiology [22, 24, 10], histology [6, 7], or ophthalmology [18, 5]. Each surgical video is typically from one patient, resulting in sparse and unrepresentative training samples. Consequently, hyperparameter tuning on

small validation sets from a single surgical video fails to capture dataset diversity, resulting in poor generalization under test-time distribution shifts.

Current few-shot learning methods [17, 19, 28] focus on spatial, image-based adaptation and fail to capture temporal dependencies essential for surgical workflow modeling, which vary across institutions. For instance, different hospitals follow distinct procedural protocols for gastric bypass surgery [9]. Sparse, isolated frames lack the continuity needed to model such variations, making existing methods ineffective for cross-institutional generalization. To address this, we leverage institutional-specific task graphs as temporal priors to learn phase transition distributions using a generative approach, i.e., diffusion modeling.

We introduce Surgical Phase Anywhere (SPA), a lightweight adaptation framework enabling foundation models to generalize across institutions with minimal supervision by integrating spatiotemporal priors. SPA enhances spatial generalization via few-shot adaptation, training a lightweight linear layer on a frozen foundation model to align multi-modal embeddings using limited labeled images and textual phase descriptions. To address temporal variability, SPA employs diffusion adaptation module that encodes institutional task-graph priors, ensuring phase predictions align with the target institution's protocol. Additionally, SPA uses self-supervised test-time adaptation, leveraging multi-modal mutual agreement to refine predictions during inference, adapting to distribution shifts from procedural variations. SPA enables scalable, data-efficient surgical workflow modeling through spatial, temporal, and test-time adaptation.

Our key contributions are: (1) **S**urgical **P**hase **A**nywhere (SPA), an institution-specific adaptation framework that transfers surgical vision-language foundation models to phase recognition across procedures and institutions with minimal supervision, (2) a spatiotemporal adaptation strategy, integrating spatial priors via few-shot learning and temporal priors via diffusion modeling, (3) test-time adaptation, enhancing robustness against test-time distribution shifts through multi-modal mutual agreement, and (4) extensive evaluations, showing strong generalization, sometimes surpassing full-shot methods with 200× less data.

2 Method

SPA has three modules: spatial adaptation via few-shot learning (Sec.2.1), temporal adaptation via task-graph guided diffusion modeling (Sec.2.2), and test-time adaptation (Sec. 2.3). We follow the standard few-shot learning definition, where N-shot means using N × K labeled samples, with K samples per class.

2.1 Spatial adaptation via few-shot learning

As shown in Fig. 2 (a), we first build a few-shot classifier using the target institution-specific priors, i.e., textual phase class descriptions from each center and a few-shot set of labeled images. For each image $\boldsymbol{x}_i$, its vision embedding is computed as: $\boldsymbol{f}_i = \boldsymbol{\theta}_v(\boldsymbol{x}_i)$ using the the frozen pre-trained visual encoder θ_v. For each phase class $k \in \{1, \ldots, K\}$, we encode its textual description $\boldsymbol{d}_k$ (e.g.,

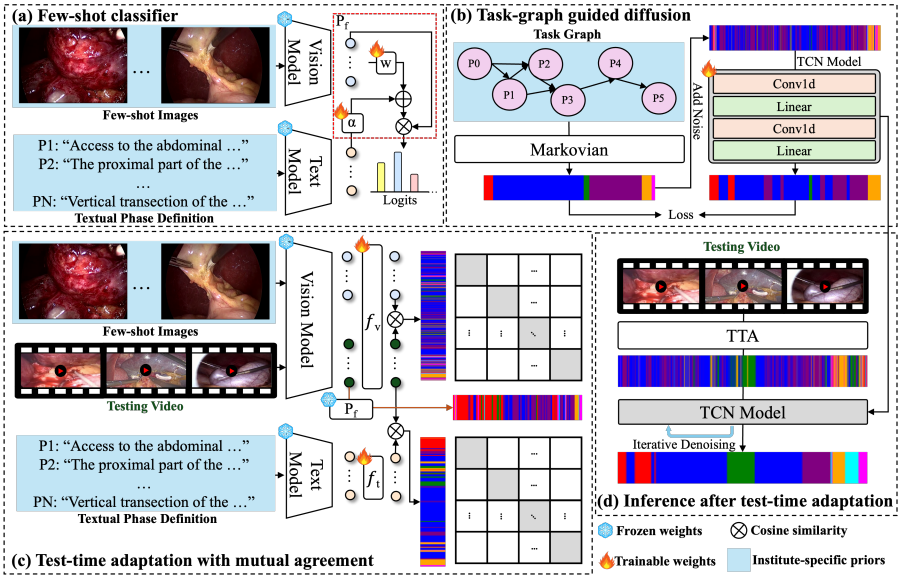

Fig. 2. Pipeline of SPA. During training, we first train (a) few-shot classifier for spatial adaptation, and (b) task-graph guided diffusion model for temporal adaptation. During inference, TTA updates model parameters using self-supervision, followed by the diffusion model to ensure temporally coherent phase transitions.

"Preparation: we introduce trocar and ...") using a frozen pretrained text encoder: $\boldsymbol{t}_k = \boldsymbol{\theta}_t(\boldsymbol{d}_k)$. Following the work of Shakeri et al. [17], we then learn a text-driven linear classifier P_f consisting of vision class prototypes $\mathbf{w} = (\boldsymbol{w}_k)_{1 \le k \le K}$ blended with text embeddings via learnable multipliers $\boldsymbol{\alpha} = (\alpha_k)_{1 \le k \le K}$:

$$p_{ik}(\mathbf{w}, \boldsymbol{\alpha}) = \frac{\exp\left(\boldsymbol{f}_i^t(\boldsymbol{w}_k + \alpha_k \boldsymbol{t}_k)\right)}{\sum_{j=1}^{K} \exp\left(\boldsymbol{f}_i^t(\boldsymbol{w}_j + \alpha_j \boldsymbol{t}_j)\right)}. \tag{1}$$

During training, only the weights of the class prototypes in linear layer $\mathbf{w}$ and the learnable multiplier $\boldsymbol{\alpha}$ are optimized via gradient descent, while text embeddings $\boldsymbol{t}_k$ remain fixed. The learning objective is cross-entropy loss to learn to classify each few-shot image into one of the phase classes.

2.2 Temporal adaptation via task-graph guided diffusion

The initial non-temporal predictions can be noisy and inaccurate. To enforce temporal coherence and improve prediction accuracy, we introduce a task-graph guided diffusion model. The institution-specific task graph generates synthetic phase sequences, which are then used to train the diffusion model, enabling it to capture the contextual relationships between different phases. During inference, the trained diffusion model refines the noisy coarse phase predictions to align with the specific procedural protocols of the institution.

Synthetic data generation from task graph: First, we synthesize temporally coherent surgical phase sequences $X \in \mathbb{R}^L$ by encoding phase transition graph, i.e., Task Graph, $\mathcal{G} = (V, E)$. The task graph represents the valid transitions between surgical phases. Here, V is the set of graph nodes v_j to represent surgical phases, and E is the set of directed edges e_{ij} indicating permissible phase transitions. The edges encode institution-specific phase dependencies, ensuring that generated sequences follow target procedure protocols. For each surgical phase node, we also have an estimated time of their temporal bounds ($[L^i_{\min}, L^i_{\max}]$). Therefore, we can synthesize surgical phase sequence using a Markovian process using the following:

$$P(X_{l+1}|X_l) = \begin{cases} \mathcal{U}(v_j \in V \mid e_{ij} \in E) & \text{if } \Delta l^i < L^i_{\max} \\ 0 & \text{otherwise} \end{cases}. \tag{2}$$

This equation dictates that the next phase X_{l+1} is uniformly sampled from the set of valid successor phases in the task graph ($\mathcal{U}(v_j \in V \mid e_{ij} \in E)$), and the current phase duration has not exceeded its upper bound ($L^i_{\max}$). The synthetic sequences generated using Eq. 2 are representative of the target institution's surgical phase transition distribution. Also, the Markovian generation process enables the creation of diverse yet valid sequences, effectively modeling variations in real-world procedures.

Diffusion model in hidden state space: We train a diffusion model on the generated synthetic phase sequences X to learn the phase transition distribution specific to the target institution as shown in Fig. 2(b). The model operates in a *hidden state space* $H \subseteq \mathbb{R}^C$, where each hidden state $h_l \in H$ represents the phase label at timestamp l. These hidden states are mapped to phase probabilities: $H = \{h_1, \ldots, h_L\}\, \sigma(h_l) \in \mathbb{R}^C$, where $\sigma(h_l)$ is the one-hot feature vector.

Forward process in hidden space. The forward process gradually introduces noise into the phase sequences, disrupting their structure. This helps the model learn to recover meaningful phase transitions from noisy data. Formally, the forward process is defined as:

$$q(H_t|H_{t-1}) = \mathcal{N}\left(H_t; \sqrt{1-\beta_t}H_{t-1}, \beta_t I\right), \tag{3}$$

β_t is the noise variance at step t. Here, $\mathcal{N}(\mu, \Sigma)$ represents a multivariate normal (Gaussian) distribution with mean μ and covariance Σ. This means that at each step t, the hidden state H_t is sampled from a Gaussian distribution.

Reverse process with surgical priors. The reverse process learns to denoise sequences to learn the temporal priors from the synthetic data:

$$p_\theta(H_{t-1}|H_t) = \mathcal{N}\left(H_{t-1}; \boldsymbol{\mu}_\theta(H_t, t), \boldsymbol{\Sigma}_\theta(H_t, t)\right), \tag{4}$$

where $\boldsymbol{\mu}_\theta$ and $\boldsymbol{\Sigma}_\theta$ are learned parameters. During training, noise is progressively added to phase sequences, and the reverse process learns to denoise them, refining noisy predictions into surgically plausible sequences. During inference, a coarse phase prediction and noise step are input, and the reverse process refines it iteratively to align with learned temporal priors, i.e., procedure protocols.

2.3 Test-time adaptation (TTA) with mutual agreement

Surgical foundation models generate multi-modal embeddings, enabling image-text comparisons via cosine similarity. In the SPA framework, each test image produces three prediction streams: (1) **Reference prediction**, which compares the test image to few-shot reference images (Sec. 2.1), given by $S_{ref} = \sigma(V \cdot R^\top)C$, where V is the test image embedding, R the reference visual embeddings, and C is their class associations; (2) **Vision-language prediction**, which matches the image to phase descriptions using $S_{vl} = \sigma(V \cdot T^\top)$; and (3) **Few-shot prediction**, obtained from a previously trained few-shot classifier (Sec. 2.1), given by S_{fs}.

These predictions may individually generalize poorly due to test-time distribution shifts. However, ensembling them provides complementary information that if they co-agree on a prediction, it is likely correct. To leverage this mutual agreement, we introduce a self-supervised learning strategy for test-time adaptation. We add two linear layers f_v and f_t on top of the vision and text encoders to refine embeddings, thus $S_{vl} = \sigma(f_v(V) \cdot f_t(T)^\top/\tau)$ and $S_{ref} = \sigma(f_v(V) \cdot f_v(R)^\top)C$.

To align predictions across modalities, we define a contrastive loss:

$$\mathcal{L} = \mathcal{L}_{mutual}(S_{ref}, S_{vl}) + \mathcal{L}_{mutual}(S_{fs}, S_{ref}),$$

where $\mathcal{L}_{mutual}(A, B) = -\frac{1}{L}\sum_{l=1}^{L}\sum_{k=1}^{K} A_{l,k} \log B_{l,k}$ penalizes frame-wise mismatches by aligning the diagonal elements of similarity matrices, enforcing cross-modal consistency at each timestamp. Overall, our SPA inference has two steps: first, we adapt the vision and text linear layers using $\mathcal{L}$ to refine predictions for the test video. Then, we fuse the three prediction streams and further refine them using the previously trained diffusion model (Sec. 2.2).

3 Experiment setup

Target procedures and institutions: Evaluation includes diverse surgical phase recognition datasets spanning various surgical procedures. **Cholecystectomy:** We utilize Cholec80 [20] as it represents a well-established and routinely performed surgical procedure. This dataset requires anatomical understanding, including structures such as the gallbladder and liver; **Gastric Bypass:** To evaluate cross-dataset generalizability, we consider both StrasbourgBypass [9] and BernBypass [9]. Despite both datasets featuring gastric bypass surgeries, differences in phase definitions and procedural protocols create a significant domain gap, requiring adaptation of surgical foundation models. **Hysterectomy**: We also evaluate laparoscopic hysterectomy phase recognition using Autolaparo [23].

Implementation details: This work adapts a generalist foundation model for institution-specific surgical phase recognition. We use PeskaVLP [26], an open-access surgical vision-language foundation model pretrained on surgical video lectures [27]. To retain pretraining knowledge and minimize computational

Table 1. Comparison of few-shot methods using average F1 (%) across four benchmarks. For each N-shot K-class setting, results are averaged over three subsets. The RN50 row is populated with values from SOTA papers, Cholec80 [1], Bypass [9].

Methods	Shots	Cholec80	StrasBypass	BernBypass	Autolaparo
CLIP [15]	0	13.1	5.5	4.1	18.3
PeskaVLP [26]	0	34.2	28.6	22.6	28.6
RN50	Full	80.3	71.1	42.4	NA
Tip-Adapter-F [28]	1	$29.29_{\pm 0.53}$	$14.88_{\pm 0.03}$	$14.29_{\pm 0.00}$	$26.58_{\pm 0.00}$
Linear probe (LP) [17]	1	$26.19_{\pm 6.75}$	$24.70_{\pm 3.64}$	$14.95_{\pm 1.35}$	$33.90_{\pm 5.19}$
LP+text [17]	1	$32.21_{\pm 1.88}$	$23.53_{\pm 3.59}$	$14.12_{\pm 2.36}$	$32.80_{\pm 5.81}$
SPA	1	$44.53_{\pm 5.77}$	$30.68_{\pm 5.24}$	$24.31_{\pm 4.29}$	$51.48_{\pm 6.26}$
Tip-Adapter-F [28]	16	$40.75_{\pm 2.15}$	$34.29_{\pm 2.12}$	$24.54_{\pm 1.97}$	$37.51_{\pm 1.31}$
Linear probe (LP) [17]	16	$39.06_{\pm 1.04}$	$38.48_{\pm 1.88}$	$27.36_{\pm 0.45}$	$42.60_{\pm 2.78}$
LP+text [17]	16	$38.24_{\pm 2.33}$	$35.42_{\pm 1.35}$	$25.01_{\pm 0.50}$	$39.43_{\pm 4.83}$
SPA	16	$55.05_{\pm 4.80}$	$52.39_{\pm 1.88}$	$38.73_{\pm 1.58}$	$58.91_{\pm 4.72}$
Tip-Adapter-F [28]	32	$42.05_{\pm 1.04}$	$35.83_{\pm 4.08}$	$26.98_{\pm 1.68}$	$39.75_{\pm 2.11}$
Linear probe (LP) [17]	32	$38.37_{\pm 1.37}$	$42.79_{\pm 0.61}$	$29.74_{\pm 0.85}$	$44.72_{\pm 1.25}$
LP+text [17]	32	$38.36_{\pm 1.81}$	$40.35_{\pm 1.25}$	$26.01_{\pm 0.79}$	$39.40_{\pm 3.35}$
SPA	32	$55.69_{\pm 3.99}$	$55.80_{\pm 1.32}$	$43.54_{\pm 1.96}$	$60.36_{\pm 3.14}$

costs, PeskaVLP's encoders remain frozen during our adaptation. Few-shot adaptation (Sec.2.1) is conducted on an RTX 3090 with a learning rate of 0.01. The task-graph-guided diffusion model (Sec.2.2) is trained on a V100 GPU for 20 epochs with a batch size of 128 and a learning rate of 0.0001. Test-time adaptation (Sec. 2.3) runs on an RTX 3090 for 15 epochs with a 0.0001 learning rate, processing a 30-minute video in 22 seconds. Inference estimates noise levels and applies the diffusion model, requiring just 0.26 seconds for a 30-minute video.

4 Results and discussions

Tab. 1 shows that SPA effectively adapts across diverse surgical procedures (e.g., cholecystectomy vs. gastric bypass) and institutions (e.g., StrasBypass vs. BernBypass) by leveraging general knowledge from pretrained vision-language model [26]. By leveraging institutional-specific spatiotemporal priors, SPA surpasses the SOTA few-shot methods, e.g., improving F1 by +17.33% on Cholec80 (32-shot). This scalability highlights its ability to refine domain-agnostic pretrained features with minimal institution-specific data. SPA even achieves a 43.54% F1 on BernBypass (32-shot), surpassing the full-shot method (42.4%) with 300× less labeled data, demonstrating its viability for real-world settings where labeled data is scarce. As shown in Fig. 3, SPA reduces fragmentation and errors in phase predictions. Both zero-shot and few-shot LP predictions exhibit high fragmentation, with disordered phase transitions and weak temporal consistency. TTA improves phase prediction accuracy and reduces fragmentation,

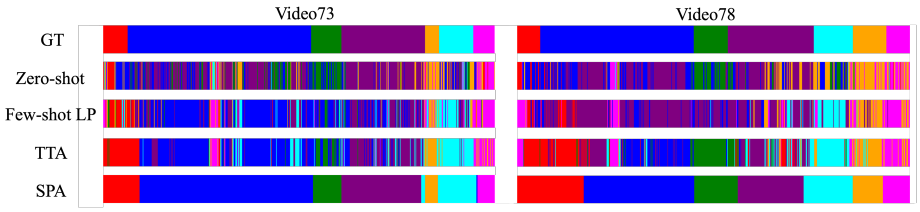

Fig. 3. 32-shot phase recognition shows SPA reduces noise and fragmentation.

Table 2. Ablation study. We conduct surgical phase recognition and report F1 Score. TTA: test-time adaptation. TG: Task-graph guided diffusion modeling.

	Shots	LP+text	+TTA	+TTA +TG-In	+TTA +TG-Cross
Cholec80	16	$38.24_{\pm 2.33}$	$51.89_{\pm 4.39}$	$55.05_{\pm 4.80}$	NA
	32	$38.36_{\pm 1.81}$	$51.23_{\pm 2.93}$	$55.69_{\pm 3.99}$	NA
StrasBypass	16	$35.42_{\pm 1.35}$	$41.70_{\pm 1.60}$	$52.39_{\pm 1.88}$	$39.09_{\pm 3.01}$
	32	$40.35_{\pm 1.25}$	$43.32_{\pm 0.38}$	$55.80_{\pm 1.32}$	$42.86_{\pm 1.35}$
BernBypass	16	$25.01_{\pm 0.50}$	$28.72_{\pm 3.66}$	$38.73_{\pm 1.58}$	$41.64_{\pm 1.05}$
	32	$26.01_{\pm 0.79}$	$29.09_{\pm 0.62}$	$43.54_{\pm 1.96}$	$45.38_{\pm 1.78}$
Autolaparo	16	$39.43_{\pm 4.83}$	$47.59_{\pm 3.42}$	$58.91_{\pm 4.72}$	NA
	32	$39.40_{\pm 3.35}$	$50.68_{\pm 1.07}$	$60.36_{\pm 3.14}$	NA

but the predictions remain noisy. SPA further enhances recognition, yielding predictions closest to the ground truth with clear phase boundaries.

4.1 Ablation study

We conduct an ablation study in Tab. 2 to evaluate the effect of our proposed modules, i.e., test-time adaptation (TTA), task-graph diffusion modeling (TG). **Test-time adaptation:** TTA significantly enhances model generalization for unseen test videos. Across all datasets, TTA improves F1 scores by notable margins, e.g., +13.65% on Cholec80 (16-shot) and +6.28% on StrasBypass (16-shot). These results highlight the effectiveness of self-supervised adaptation in mitigating distribution shifts, enabling more robust phase recognition in varied surgical environments. **Task-graph guided diffusion:** We assess the transferability of task graphs using institution-specific (TG-In) and cross-institutional (TG-Cross) priors. As shown in Tab. 2, TG-In significantly boosts performance, improving StrasBypass by +16.97% (16-shot) and BernBypass by +13.72%. However, TG-Cross shows mixed results: applying StrasBypass's task graph enhances BernBypass, while BernBypass's task graph reduces StrasBypass performance. This shows that procedural differences impact effectiveness, highlighting the importance of task graph quality and compatibility for cross-institutional adaptation.

5 Conclusion

This work presents **S**urgical **P**hase **A**nywhere (SPA), a lightweight framework that adapts foundation models for cross-institutional surgical workflow understanding with minimal annotation. By combining few-shot spatial adaptation, temporal adaptation, and dynamic test-time adaptation, SPA effectively handles domain shifts and ensures reliable phase recognition. SPA enables scalable, rapid deployment of customized phase recognition models across diverse institutions. This approach advances data-efficient surgical AI, offering a practical solution for real-world clinical applications.

Acknowledgments. This work has received funding from the European Union (ERC, CompSURG, 101088553). Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them. This work was also partially supported by French state funds managed by the ANR under Grant ANR-10-IAHU-02. The authors would like to acknowledge the High Performance Computing Center of the University of Strasbourg for supporting this work by providing scientific support and access to computing resources. Part of the computing resources were funded by the Equipex Equip@Meso project (Programme Investissements d'Avenir) and the CPER Alsacalcul/Big Data.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Alapatt, D., Murali, A., Srivastav, V., Consortium, A., Mascagni, P., Padoy, N.: Jumpstarting surgical computer vision. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 328–338. Springer (2024)
2. Czempiel, T., Paschali, M., Keicher, M., Simson, W., Feussner, H., Kim, S.T., Navab, N.: Tecno: Surgical phase recognition with multi-stage temporal convolutional networks. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2020: 23rd International Conference, Lima, Peru, October 4–8, 2020, Proceedings, Part III 23. pp. 343–352. Springer (2020)
3. 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)
4. Honarmand, M., Jamal, M.A., Mohareri, O.: Vidlpro: A video-language pretraining framework for robotic and laparoscopic surgery. In: Advancements In Medical Foundation Models: Explainability, Robustness, Security, and Beyond (2024)
5. Hu, M., Yuan, K., Shen, Y., Tang, F., Xu, X., Zhou, L., Li, W., Chen, Y., Xu, Z., Peng, Z., et al.: Ophclip: Hierarchical retrieval-augmented learning for ophthalmic surgical video-language pretraining. arXiv preprint arXiv:2411.15421 (2024)
6. Huang, Z., Bianchi, F., Yuksekgonul, M., Montine, T.J., Zou, J.: A visual–language foundation model for pathology image analysis using medical twitter. Nature medicine **29**(9), 2307–2316 (2023)

7. Ikezogwo, W., Seyfioglu, S., Ghezloo, F., Geva, D., Sheikh Mohammed, F., Anand, P.K., Krishna, R., Shapiro, L.: Quilt-1m: One million image-text pairs for histopathology. Advances in neural information processing systems **36** (2024)
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. Lavanchy, J.L., Ramesh, S., Dall'Alba, D., Gonzalez, C., Fiorini, P., Müller-Stich, B.P., Nett, P.C., Marescaux, J., Mutter, D., 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 pp. 1–9 (2024)
10. Ma, C., Jiang, H., Chen, W., Li, Y., Wu, Z., Yu, X., Liu, Z., Guo, L., Zhu, D., Zhang, T., et al.: Eye-gaze guided multi-modal alignment for medical representation learning. Advances in Neural Information Processing Systems **37**, 6126–6153 (2025)
11. Maier-Hein, L., Vedula, S.S., Speidel, S., Navab, N., Kikinis, R., Park, A., Eisenmann, M., Feussner, H., Forestier, G., Giannarou, S., et al.: Surgical data science for next-generation interventions. Nature Biomedical Engineering **1**(9), 691–696 (2017)
12. Mascagni, P., Alapatt, D., Sestini, L., Altieri, M.S., Madani, A., Watanabe, Y., Alseidi, A., Redan, J.A., Alfieri, S., Costamagna, G., et al.: Computer vision in surgery: from potential to clinical value. npj Digital Medicine **5**(1), 163 (2022)
13. Murali, A., Alapatt, D., Mascagni, P., Vardazaryan, A., Garcia, A., Okamoto, N., Mutter, D., Padoy, N.: Latent graph representations for critical view of safety assessment. arXiv preprint arXiv:2212.04155 (2022)
14. Özsoy, E., Pellegrini, C., Keicher, M., Navab, N.: Oracle: Large vision-language models for knowledge-guided holistic or domain modeling. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 455–465. Springer (2024)
15. 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)
16. Ramesh, S., Srivastav, V., Alapatt, D., Yu, T., Murali, A., Sestini, L., Nwoye, C.I., Hamoud, I., Sharma, S., Fleurentin, A., et al.: Dissecting self-supervised learning methods for surgical computer vision. Medical Image Analysis **88**, 102844 (2023)
17. 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)
18. Silva-Rodriguez, J., Chakor, H., Kobbi, R., Dolz, J., Ayed, I.B.: A foundation language-image model of the retina (flair): Encoding expert knowledge in text supervision. Medical Image Analysis **99**, 103357 (2025)
19. Silva-Rodriguez, J., Hajimiri, S., Ben Ayed, I., Dolz, J.: A closer look at the few-shot adaptation of large vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23681–23690 (2024)
20. 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)

21. Wang, H., Yang, G., Zhang, S., Qin, J., Guo, Y., Xu, B., Jin, Y., Zhu, L.: Video-instrument synergistic network for referring video instrument segmentation in robotic surgery. IEEE Transactions on Medical Imaging (2024)
22. Wang, Z., Wu, Z., Agarwal, D., Sun, J.: Medclip: Contrastive learning from unpaired medical images and text. arXiv preprint arXiv:2210.10163 (2022)
23. 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)
24. Wu, C., Zhang, X., Zhang, Y., Wang, Y., Xie, W.: Medklip: Medical knowledge enhanced language-image pre-training for x-ray diagnosis. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 21372–21383 (2023)
25. 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)
26. Yuan, K., Srivastav, V., Navab, N., Padoy, N., et al.: Procedure-aware surgical video-language pretraining with hierarchical knowledge augmentation. arXiv preprint arXiv:2410.00263 **2** (2024)
27. 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)
28. 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)