PHASE AS A STRONG PRIOR: SURGICAL PHASE FEATURE MODULATION FOR VESSEL RECOGNITION IN COMPLEX SURGICAL VIDEOS

Anonymous authors
Paper under double-blind review

ABSTRACT

Surgical procedures unfold in stereotypical phases, and the anatomy exposed at any moment is strongly conditioned on the current surgical phase. This work investigates whether the surgical phase constitutes a sufficiently informative prior for fine-grained anatomical segmentation and, if so, how such a prior should be incorporated into a segmentation foundation model. We study this problem on key vessel segmentation in laparoscopic gastrectomy, using the LapGC-KVAD-30 dataset (30 procedures, 8 surgical phases, 15 vessel classes, case-level splits). A diagnostic experiment shows that the phase label alone predicts 15-class vessel presence with a macro-F1 of 0.810, whereas image-only prediction with a compact convolutional network remains near chance (macro-F1 0.136), indicating that the phase prior dominates appearance for vessel presence. Motivated by this observation, we inject the phase prior as feature-level modulation: the proposed Phase-FiLM maps the phase embedding to channel-wise scale and shift parameters that modulate the SAM2 image features before the mask decoder. On binary vessel segmentation, Phase-FiLM improves the test mean Dice from 0.5892 ± 0.0190 to 0.6245 ± 0.0103 over three seeds (mean gain $+3.5$ points), outperforming the no-phase baseline on every seed, and yields consistent gains across two backbones (SAM2-L $+3.1$ points; MedSAM2-T $+7.9$ points). The improvements concentrate in the data-scarce dissection phases, where single-phase training collapses (S4: 0.320 vs. 0.428). Robustness analysis shows the method remains beneficial down to a break-even phase-recognition accuracy of 82.6%, within the 85–92% range reported for current surgical phase recognizers. Substituting an incorrect phase at test time collapses performance to 0.4212, indicating that the improvement arises from genuine exploitation of phase information rather than from increased model capacity. These results suggest that, in phase-structured procedures, workflow priors are most effectively incorporated as multiplicative feature modulation.

1 INTRODUCTION

Deep learning models for surgical scene understanding are typically trained to map pixels directly to labels, ignoring a property that distinguishes surgery from most natural-image domains: procedures unfold in stereotypical phases, and the anatomy visible at any moment is a strong function of the current phase—certain vessels appear only during specific dissection phases, while others remain exposed throughout. Surgeons exploit this structure routinely; segmentation networks largely do not.

We study this problem in laparoscopic radical gastrectomy for gastric cancer, a representative complex laparoscopic procedure. Gastric cancer caused more than 968,000 new cases and nearly 660,000 deaths worldwide in 2022, ranking fifth in both incidence and mortality among all cancers (Bray et al., 2024). Radical gastrectomy remains the primary curative treatment and is therefore performed in large volumes; the procedure is also technically demanding, involving extensive D2 lymphadenectomy around major abdominal vessels and typically lasting three to four hours even in experienced hands (Liao et al., 2019; Kim et al., 2016).

This gap matters because key vessel injury is a leading cause of intraoperative bleeding, and vessel segmentation is hardest precisely where phase information is most informative: vessels are thin,

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.

low-contrast, and frequently occluded, so appearance alone is weak evidence. If phase priors could be effectively injected into strong segmentation backbones, they could disambiguate exactly the ambiguous cases where pure appearance fails.

Foundation models such as SAM (Kirillov et al., 2023) and SAM2 (Ravi et al., 2025) offer powerful, general-purpose segmentation features, but they are prompt-driven and class-agnostic—they have no mechanism to accept a categorical workflow prior such as "the current phase is S3."

This paper makes four main contributions:

1. **A quantified diagnosis.** We show that the surgical phase is a *dominant* prior for vessel presence: phase-only multi-label classification reaches macro-F1 0.810 while image-only prediction by a compact CNN is near chance (0.136). This measurement both motivates the method and exposes an evaluation pitfall: models can achieve high presence accuracy while learning little about localization.
2. **A method: Phase-FiLM injection into SAM2.** We adapt FiLM (Perez et al., 2018) to inject an 8-way phase embedding as channel-wise scale/shift modulation of SAM2 image features before the mask decoder. The mechanism is parameter-light (approximately 17K parameters), leaves the pretrained architecture intact, and trains end-to-end.
3. **A robust improvement on binary vessel segmentation.** Phase-FiLM yields a clear and robust gain on binary vessel segmentation (test mDice improving from 0.5892 to 0.6245, a gain of $+3.5$ points, mean of three seeds), consistently outperforming concatenation-based injection and the autonomous-research system Camyla. Ablations—frozen encoder, concatenation injection, wrong-phase control—attribute the gain to end-to-end phase-conditioned adaptation, not parameter count or architecture artifacts.
4. **A deployment analysis.** We quantify how much upstream phase-recognition error the method tolerates (break-even at 82.6% phase-recognition accuracy, within the range of current recognizers), and demonstrate that gains concentrate in data-scarce phases where single-phase training collapses—evidence that phase conditioning transfers knowledge across phases.

We frame the contribution as a principle for conditional medical segmentation: in phase-structured procedures, workflow context is best injected as multiplicative feature modulation close to the representation, not as constraints on outputs.

2 Related Work

Surgical phase recognition and phase-aware scene understanding. Phase recognition from video is well studied in laparoscopic surgery (Twinanda et al., 2017), with temporal models such as TeCNO (Czempiel et al., 2020) and Trans-SVNet (Gao et al., 2021) reaching 85–92% frame accuracy. These works treat phase as an *output*. A complementary line of work instead uses phase as *input context* to constrain downstream tasks such as instrument detection or action recognition, typically via multi-task learning or temporal context (Maier-Hein et al., 2022). Our work follows the latter direction but differs in both the prior and the injection mechanism: to our knowledge we are the first to quantify the phase prior for *vessel presence* and to inject it into a segmentation foundation model via feature modulation. The two paradigms are complementary and could be cascaded (Section 5).

Anatomical structure segmentation in laparoscopy. Semantic segmentation of surgical anatomy has been most extensively studied in laparoscopic cholecystectomy, where models identify the liver, gallbladder, and hepatocystic structures to support the critical view of safety (Madani et al., 2022), and in multi-organ datasets such as the Dresden Surgical Anatomy Dataset (Carstens et al., 2023). Explicit *vessel* segmentation, by contrast, remains comparatively scarce: existing efforts are largely confined to a small number of vascular structures in specific procedures, e.g. portal and mesenteric veins in pancreatic surgery (Shi et al., 2025). For gastric cancer surgery in particular, prior segmentation work is limited to a few structures (pancreas, transverse mesocolon, right gastroepiploic vein) in the robotic setting, and the recently released LapGC-KVAD-30 dataset (Gou et al., 2025) provides the first multi-class key-vessel annotations, spanning 15 vascular categories, for D2 lymph node dissection. A concurrent autonomous-research system that searches over thousands of model variants

on this dataset reports a best Dice of 0.570 (Gao et al., 2026), indicating that fine vessels in this complex, phase-structured setting remain a bottleneck. Our work targets this regime and shows that conditioning on the surgical phase yields a clear improvement over appearance-only networks.

Segmentation foundation models and conditional modulation. Promptable foundation models such as SAM (Kirillov et al., 2023) and SAM2 (Ravi et al., 2025), together with their medical adaptations MedSAM (Ma et al., 2024) and MedSAM2 (Ma et al., 2025), transfer strongly but are prompt-driven and unconditional: they expose no mechanism for injecting categorical workflow context. Parameter-efficient fine-tuning methods such as LoRA (Hu et al., 2022) add capacity to these models but likewise provide no semantic conditioning. A separate line of work conditions features through scale-and-shift modulation, as in FiLM (Perez et al., 2018), AdaIN (Huang & Belongie, 2017), and the conditioning layers of diffusion transformers (Peebles & Xie, 2023). Our Phase-FiLM bridges these two lines: it applies feature-wise modulation to condition a segmentation foundation model on the surgical phase with negligible parameters, and provides, to our knowledge, the first head-to-head evidence in this domain that feature-level modulation outperforms concatenation-based injection.

3 METHOD

3.1 DATASET AND TASKS

We conduct experiments on the LapGC-KVAD-30 dataset (Gou et al., 2025), which comprises 5303 images drawn from the D2 lymph node dissection procedure and annotated with 15 key vessel classes. Each image is further labeled with one of eight surgical phases, denoted S1 through S8. Following the case-level split, the data are partitioned into training, validation, and test sets of 3544, 880, and 879 images, respectively. Full vessel-class and phase definitions are provided in Section S1 of the supplementary material. We consider two main tasks, one diagnostic and one primary:

- **Vessel presence classification** (diagnostic): given the image and the phase label, predict the 15-dimensional multi-hot presence vector indicating which vessel classes are visible. This isolates the information content of the phase prior from localization. Metrics: macro-F1 and macro average precision.
- **Multi-vessel segmentation** (primary task): segment the key vessels in each frame. We frame this as binary segmentation of the union mask over the 15 classes (metric: mean Dice).

3.2 DIAGNOSTIC PRESENCE CLASSIFICATION

We treat a phase prior as the conditional probability that a given vessel is visible under a given phase, estimated from the training split; the phase label alone is thus already informative for predicting presence. We compare four conditioning modes on a compact CNN: image-only, phase-only (a learnable phase embedding), image-phase concatenation, and phase FiLM. The purpose is twofold: to quantify the information content of the phase prior independent of localization, and to select the fusion mechanism to port to segmentation. Implementation details are deferred to Section S2 of the supplementary material.

3.3 PHASE-FILM: FEATURE-LEVEL PHASE INJECTION INTO SAM2

Figure 1 illustrates the overall architecture. We build on SAM2: the image encoder (Hiera-Large) produces a top-level FPN feature f, which a prompt-free mask decoder turns into binary vessel logits. To inject surgical-phase information, the phase label $s \in \{1, \ldots, 8\}$ is first mapped to a learned phase embedding. This embedding is then routed through two parallel lightweight heads, a scale head and a shift head, each consisting of a fully connected layer, a GELU nonlinearity, and a second fully connected layer. The two heads produce channel-wise scale and shift parameters γ and β, respectively, which modulate the image feature FiLM-style,

$$\tilde{f} = f \odot (1 + \gamma) + \beta, \tag{1}$$

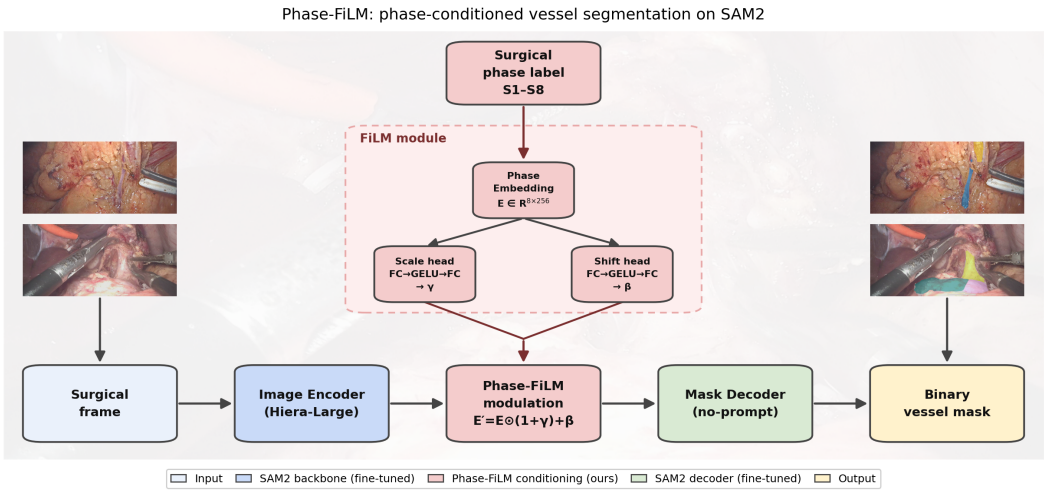


Figure 1: Overview of Phase-FiLM.

and the modulated feature $\tilde{f}$ is what the mask decoder sees. In the figure, the phase-conditioning pathway is highlighted in red, and the SAM2 backbone and decoder are shown in blue and green. This pathway is the only architectural addition to an otherwise standard SAM2 backbone–decoder pipeline; it adds only a negligible number of parameters and is trained jointly with full end-to-end fine-tuning. Detailed implementation and training hyperparameters are deferred to Section S2 of the supplementary material.

To isolate what drives the gain, we compare against two controls that share the same backbone, decoder, and training protocol. *Phase-Concat* replaces multiplicative FiLM modulation with concatenation-based injection: the phase embedding is spatially broadcast, concatenated to the image feature, and projected back to the original channel dimension by a 1×1 convolution. This tests whether multiplicative modulation, specifically, is responsible for any improvement over a generic feature-fusion baseline. *Frozen-encoder* keeps the same Phase-FiLM architecture but freezes the image encoder, training only the FiLM layer and decoder; this tests whether the gain requires co-adaptation of the encoder features with the phase conditioning.

4 Experiments

4.1 Experimental Protocol

We adhere to a strict evaluation protocol to prevent information leakage and ensure unbiased estimates of generalization. All experiments use the official case-level splits of LapGC-KVAD-30, thereby preventing the model from exploiting patient-specific appearance cues between the train and test sets. The test set is reserved exclusively for final evaluation; no test data is used during model selection, hyperparameter tuning, or the estimation of phase priors. All checkpoints are selected based on validation set performance.

No data augmentation is applied during training. Horizontal flipping is deliberately disabled: unlike generic visual recognition tasks, laparoscopic vessel segmentation requires distinguishing left and right anatomical structures whose identities are not invariant under mirroring. No photometric or geometric transformations are used.

All reported metrics are extracted from per-experiment JSON artifacts generated automatically during evaluation, eliminating manual transcription errors. Implementation details including network architectures, optimization hyperparameters, and training schedules are provided in Section S2 of the supplementary material.

4.2 DIAGNOSTIC: PHASE IS A DOMINANT PRIOR FOR VESSEL PRESENCE

Table 1 summarizes test-set 15-class vessel presence classification across the four conditioning modes (879 frames). The phase-only mode attains a macro-F1 of 0.810, substantially exceeding the image-only baseline of 0.136. Phase FiLM achieves 0.804, outperforming image-phase concatenation (0.800); this ordering informs the fusion mechanism adopted for segmentation. The near-chance performance of the image-only mode indicates that appearance features alone are largely uninformative for vessel presence, whereas the phase label provides a dominant source of prior information.

Table 1: Test-set 15-class vessel presence classification (879 frames).

Mode	macro-F1
image-only	0.1361
image-phase concatenation	0.8000
phase-only	0.8097
phase FiLM	0.8040

4.3 MAIN RESULT: PHASE-FILM IMPROVES BINARY VESSEL SEGMENTATION

Phase-FiLM substantially improves binary vessel segmentation over the no-phase baseline. Table 2 reports test mDice over three seeds: Phase-FiLM attains a mean test mDice of 0.6245 ± 0.0103 against 0.5892 ± 0.0190 for the baseline, outperforming it on all three seeds (gains of $+0.029$ to $+0.046$) with a mean gain of $+3.5$ points, confirming that the improvement is statistically robust rather than a single-seed artifact. Table 3 further compares Phase-FiLM against alternative injection schemes and external systems. Concatenation-based injection yields no improvement over the no-phase baseline, and the autonomous-research system Camyla (Gao et al., 2026), which searches over thousands of model variants on this dataset, reports a best test mDice of 0.5698; Phase-FiLM outperforms both.

Table 2: Multi-seed comparison of no-phase baseline and Phase-FiLM (test mDice).

Model	seed 1	seed 2	seed 3	mean $\pm$ std
No-phase baseline	0.5925	0.6064	0.5687	0.5892 ± 0.0190
Phase-FiLM	**0.6239**	**0.6350**	**0.6145**	$\mathbf{0.6245 \pm 0.0103}$
Improvement	$+0.031$	$+0.029$	$+0.046$	$+0.035 \pm 0.0094$

Table 3: Injection-method comparison on binary vessel segmentation (SAM2-L backbone).

Method	test mDice	Improvement over baseline
No-phase baseline	0.5892	—
Phase-Concat (concat injection)	0.5888	-0.0004
Camyla (Gao et al., 2026)	0.5698	-0.0194
Phase-FiLM (ours)	**0.6245**	$+$**0.0353**

4.4 TRAINING-MECHANISM ABLATIONS

Table 4 examines the training configuration of Phase-FiLM. Three controls are decisive. First, freezing the image encoder and training only the FiLM layer and the decoder attains a test mDice of 0.4425, far below the fully fine-tuned 0.6239, indicating that the gain requires end-to-end co-adaptation with the encoder features rather than merely additional capacity. Second, replacing the multiplicative FiLM injection with concatenation-based injection yields 0.5888, statistically indistinguishable from the no-phase baseline (0.5925) and well below Phase-FiLM, despite the concatenation head carrying

more than four times as many parameters as the FiLM head; multiplicative feature modulation—rather than mere access to phase information or additional capacity—is therefore responsible. Third, feeding an incorrect phase at every test frame (100% phase error) collapses the test mDice to 0.4212, confirming that the network genuinely reads vessel plausibility from the phase rather than exploiting an architectural artifact.

Table 4: Training-mechanism ablations for Phase-FiLM (test mDice).

Configuration	Test mDice	Improvement over baseline
No-phase baseline	0.5925	—
Phase-FiLM, fully fine-tuned (full model)	**0.6239**	**+0.0314**
Frozen encoder (FiLM + decoder only)	0.4425	−0.1500
Concat injection instead of FiLM	0.5888	−0.0037
Wrong phase at test time (100% phase error)	0.4212	−0.1713

4.5 Cross-backbone generality

We next examine whether the benefit of phase conditioning is tied to a particular pretrained backbone. Figure 2 reports the test mDice of the no-phase baseline and Phase-FiLM on two backbones of different scale and pretraining: SAM2-L (Hiera-Large, self-supervised video pretraining) and MedSAM2-T (Ma et al., 2025) (Hiera-Tiny, medical-image pretraining). Phase-FiLM improves over the no-phase baseline on both backbones: on SAM2-L the test mDice rises from 0.5925 to 0.6239 (a gain of $+3.1 points), and on MedSAM2-T from 0.4956 to 0.5742 (a gain of +7.9$ points). The improvement is larger on the smaller backbone, suggesting that feature modulation is especially helpful when the backbone's own features are weaker. SAM2-L outperforms MedSAM2-T in absolute terms under both settings, so we use it as the default backbone elsewhere. That the gain transfers across architectures and pretraining regimes indicates that Phase-FiLM exploits the phase prior rather than a backbone-specific artifact. Both backbones follow the same implementation protocol (architecture, training hyperparameters, and evaluation) detailed in Section S2 of the supplementary material.

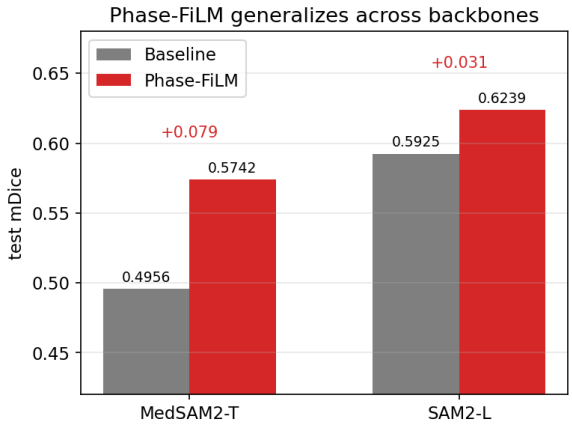


Figure 2: Cross-backbone comparison.

4.6 Robustness to phase-recognition errors

The results above assume an oracle phase. To assess deployment realism, we corrupt the test-time phase by replacing the true phase with a uniformly random incorrect phase with probability p (a fixed seed makes the corruption reproducible); at 100% error, every frame receives a wrong phase, yielding a test mDice of 0.4212. Figure 3 shows that the test mDice of Phase-FiLM degrades gracefully as the error rate grows, crossing the no-phase baseline (0.5925) at a break-even point of 82.6%

phase-recognition accuracy (approximately 17% error). The shaded band marks the 85–92% frame accuracy reported for current surgical phase recognizers (Czempiel et al., 2020; Gao et al., 2021), which lies to the right of the break-even point; Phase-FiLM therefore remains beneficial across the accuracy range of existing phase recognizers.

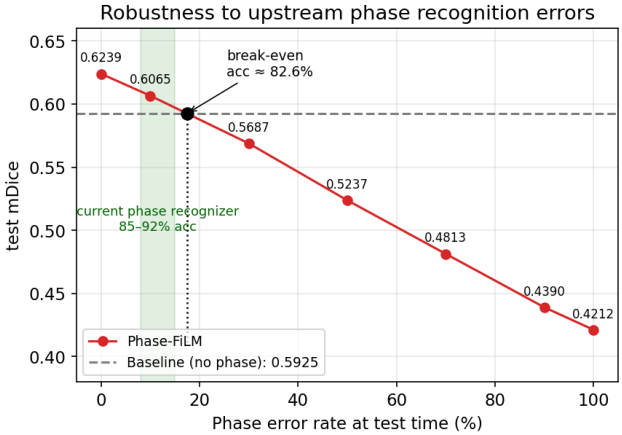


Figure 3: Robustness to upstream phase-recognition errors.

4.7 All-data Phase-FiLM versus per-phase specialist models

We compare Phase-FiLM, trained on all data with phase conditioning, against a separate single-phase baseline trained independently on each phase's data. Figure 4 reports the per-phase test mDice. Phase-FiLM matches or exceeds the single-phase specialist in 7 of 8 phases, and the improvements concentrate in the data-scarce phases: S4 improves by +0.108, where single-phase training collapses (0.320 versus 0.428), and S1 improves by +0.057. This indicates that phase conditioning transfers knowledge from data-abundant phases to data-scarce phases—a core benefit of training a single model on all data with phase conditioning rather than a separate specialist per phase. S3 is the only phase where the single-phase baseline is slightly higher (by 0.013), because S3 is itself data-abundant and anatomically simple, so additional cross-phase data offers little benefit.

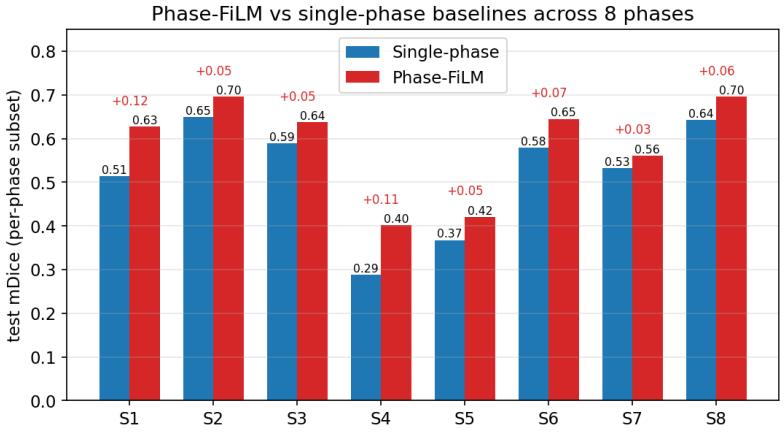


Figure 4: Per-phase comparison with single-phase baselines.

5 DISCUSSION

Results and model discussion. The experiments establish two points that we summarize briefly here. First, the phase prior is necessary: the wrong-phase collapse shows that the network reads vessel plausibility from phase rather than from the conditioning pathway alone. Second, the per-phase breakdown (Figure 4) shows that gains concentrate where vessel visibility is most phase-dependent and where data is scarcest. Together these results suggest that a workflow prior helps when it modulates *what features are extracted* multiplicatively, close to the representation. Since our contribution is the conditioning mechanism rather than any particular backbone, we build on SAM2 rather than the more recent SAM3: SAM2 is the established video-segmentation foundation model and the standard backbone in recent surgical segmentation, which keeps the comparison with existing baselines fair, and Phase-FiLM is architecture-agnostic—the gain already transfers across backbones (SAM2-L and MedSAM2-T), so using SAM2 isolates the effect of phase injection rather than confounding it with a backbone upgrade. The same mechanism is domain-agnostic and could inject categorical context into foundation backbones in other phase-structured video domains at a cost of approximately 0.1M parameters.

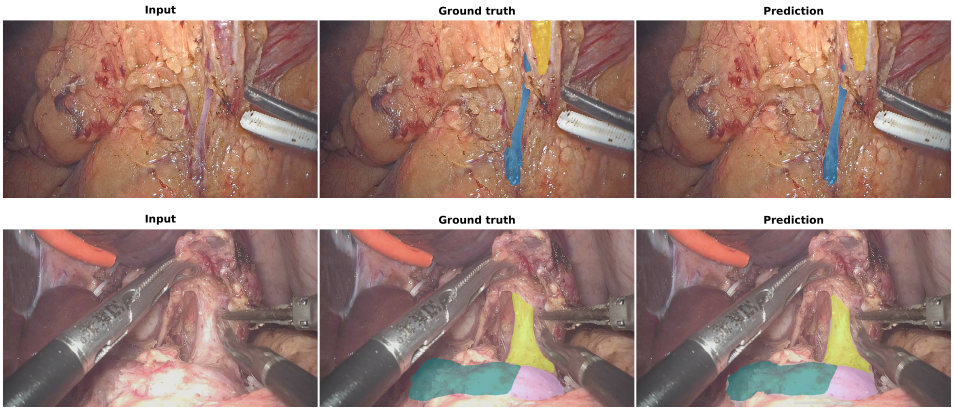


Figure 5: Qualitative multi-vessel segmentation results.

Clinical significance. Figure 5 shows qualitative multi-vessel segmentation examples (input, ground truth, and prediction) in which Phase-FiLM accurately delineates multiple vessel classes against complex anatomical backgrounds (S3, Dice 0.911; S6, Dice 0.926). Accurate intraoperative vessel recognition can assist surgeons in localizing critical vessels during D2 lymphadenectomy, which may improve operative efficiency and reduce the risk of inadvertent vessel injury and bleeding. Because Phase-FiLM conditions recognition on the surgical phase, its predictions are constrained by the operative context: vessels are highlighted preferentially in the phases in which they are anatomically expected, rather than wherever vessel-like appearance happens to occur. This phase consistency makes the recognition output more trustworthy as decision support, since spurious detections outside the relevant phase are suppressed.

Deployment. The method assumes access to a phase label at inference time, which can be supplied by an off-the-shelf phase recognizer in a cascaded setup. Two properties make this deployment practical. First, the phase prior is injected at virtually no extra computational cost: on a single NVIDIA A100-80GB GPU, Phase-FiLM SAM2 runs at 30.3 ms per frame (approx. 33 FPS) at 1024×1024 resolution, with the image encoder dominating the cost (28.5 ms, approx. 94%) and phase conditioning adding only 5.4 ms per frame, of which the FiLM layer itself accounts for less than 1 ms. Second, the gain is robust to realistic phase-recognition errors: our error-robustness analysis (Section 4.6) shows that the improvement persists down to a phase-recognition accuracy of 82.6%, which is below the 85–92% frame accuracy reported for current surgical phase recognizers (Czempiel et al., 2020; Gao et al., 2021); per-frame smoothing would further reduce the effective error rate.

Limitations. (i) Although the dataset comprises 5303 images, they are drawn from only 30 procedures at a single institution; broader multi-center validation is needed. (ii) The break-even analysis uses random phase errors; correlated errors from a real phase recognizer may behave differently and warrant a full cascaded evaluation.

6 CONCLUSION

We quantified the surgical phase as a dominant prior for vessel presence and demonstrated that feature-level FiLM injection into SAM2 yields a clear and robust gain (test mDice improving from 0.5892 to 0.6245, a gain of $+3.5$ points, mean of three seeds) with negligible parameters. The gain is consistent across backbones, is supported by wrong-phase, frozen-encoder, and concatenation-injection controls, remains beneficial down to a break-even phase-recognition accuracy of 82.6%, within the accuracy range of current recognizers, and concentrates in data-scarce phases where single-phase training collapses. Phase-structured workflow context should be injected as multiplicative feature modulation—a principle we expect to transfer across surgical and other phase-structured video domains.

AI USE STATEMENT

In this work, we used generative AI tools to assist with language editing and polishing of the manuscript. We have not used generative AI tools for research ideation, experimental design, data analysis, or the generation of scientific claims. We have reviewed all AI-assisted work and take responsibility for the final content of this work, including text produced with the aid of generative AI.

ETHICS STATEMENT

The study uses retrospective anonymized surgical data and is framed as decision support, not autonomous action. No personally identifiable information is contained in the dataset, and the study protocol was approved by the institutional review board.

REPRODUCIBILITY STATEMENT

All experiments are fully specified by the dataset definitions in Section S1 of the supplementary material and the architectural and training details in Section S2 of the supplementary material. We report results over three random seeds for the main comparison, and all evaluation metrics are extracted automatically from per-experiment artifacts to eliminate transcription errors.

REFERENCES

Freddie Bray, Mathieu Laversanne, Hyuna Sung, Jacques Ferlay, Rebecca L. Siegel, Isabelle Soerjomataram, and Ahmedin Jemal. Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians*, 74 (3):229–263, 2024.

Matthias Carstens, Franziska M. Rinner, Sebastian Bodenstedt, Alexander C. Jenke, Jürgen Weitz, Marius Distler, Stefanie Speidel, and Fiona R. Kolbinger. The Dresden surgical anatomy dataset for abdominal organ segmentation in surgical data science. *Scientific Data*, 10:3, 2023.

Tobias Czempiel, Magdalini Paschali, Matthias Keicher, Walter Simson, Hubertus Feussner, Seong Tae Kim, and Nassir Navab. TeCNO: Surgical phase recognition with multi-stage temporal convolutional networks. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 343–352, 2020.

Xiaojie Gao, Yueming Jin, Yonghao Long, Qi Dou, and Pheng-Ann Heng. Trans-SVNet: Accurate phase recognition from surgical videos via hybrid embedding aggregation transformer. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 593–603, 2021.

Yifan Gao, Haoyue Li, Feng Yuan, Xin Gao, Weiran Huang, and Xiaosong Wang. Camyla: Scaling autonomous research in medical image segmentation. *arXiv preprint arXiv:2604.10696*, 2026.

Longfei Gou, Haolin Wu, Chang Chen, Jiayu Lai, Hua Yang, Yuqing Qiu, Boer Su, Hongyu Wang, Bingyu Zhao, Xin Ye, Jinming Li, Xiaobing Bao, Guoxin Li, Jiang Yu, Yanfeng Hu, Qi Dou, and Hao Chen. Dynamic key vascular anatomy dataset for D2 lymph node dissection during laparoscopic gastric cancer surgery. *Scientific Data*, 12:903, 2025.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022.

Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 1501–1510, 2017.

Hyoung-Il Kim, Sang-Uk Han, Han-Kwang Yang, Young-Woo Kim, Hyuk-Joon Lee, Keun Won Ryu, Joong-Min Park, Ji Yeong An, Min-Chan Kim, Sungsoo Park, et al. Multicenter prospective comparative study of robotic versus laparoscopic gastrectomy for gastric adenocarcinoma. *Annals of Surgery*, 263(1):103–109, 2016.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4015–4026, 2023.

Gang Liao, Ziwei Wang, Hui Li, Jiang Min, Jinkun Zhong, Sandrie Mariella, Kun Qian, and Wei Zhang. Learning curve and short-term outcomes of modularized laparoscopy-assisted distal gastrectomy for advanced gastric cancer: A retrospective study. *Medicine*, 98(10):e14670, 2019.

Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024.

Jun Ma, Zongxin Yang, Sumin Kim, Bihui Chen, Mohammed Baharoon, Adibvafa Fallahpour, Reza Asakereh, Hongwei Lyu, and Bo Wang. MedSAM2: Segment anything in 3D medical images and videos. *arXiv preprint arXiv:2504.03600*, 2025.

Amin Madani, Babak Namazi, Maria S. Altieri, Daniel A. Hashimoto, Amanda M. Rivera, Philip H. Pucher, Antonia Navarrete-Welton, Ganesh Sankaranarayanan, L. Michael Brunt, Allan Okrainec, et al. Artificial intelligence for intraoperative guidance: Using semantic segmentation to identify surgical anatomy during laparoscopic cholecystectomy. *Annals of Surgery*, 276(2):363–369, 2022.

Lena Maier-Hein, Matthias Eisenmann, Duygu Sarikaya, Keno März, Toby Collins, Anand Malpani, Johannes Fallert, Hubertus Feussner, Stamatia Giannarou, Pietro Mascagni, et al. Surgical data science—from concepts toward clinical translation. *Medical Image Analysis*, 76:102306, 2022.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4172–4182, 2023.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. FiLM: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. In *International Conference on Learning Representations (ICLR)*, 2025.

Jile Shi, Ruohan Cui, Hu Zhou, Lu Ping, Xianlin Han, Wenming Wu, et al. Deep learning HRNet-FCN for blood vessel identification in laparoscopic pancreatic surgery. *npj Digital Medicine*, 8, 2025.

Andru Putra Twinanda, Sherif Shehata, Didier Mutter, Jacques Marescaux, Michel de Mathelin, and Nicolas Padoy. EndoNet: A deep architecture for recognition tasks on laparoscopic videos. *IEEE Transactions on Medical Imaging*, 36(1):86–97, 2017.