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.

# LaSeD: Label-Semantic Self-Distillation for Visual-Only Surgical Phase Recognition

Ye Tao[1], Claudia Scherl[2,3], and Sara Homscheid[1⋆]

[1] Mannheim Institute for Intelligent Systems in Medicine (MIISM), Medical Faculty Mannheim, Heidelberg University, Mannheim, Germany
[2] Department of Otorhinolaryngology, Head and Neck Surgery, University Medical Center Mannheim, Medical Faculty Mannheim, Heidelberg University, Mannheim, Germany
[3] Department of Otolaryngology, Head and Neck Surgery, Campus Klinikum Bielefeld Mitte, University Hospital OWL of Bielefeld University, Bielefeld, Germany

**Abstract.** Surgical phase recognition maps each video frame to a clinically meaningful workflow phase, supporting context-aware assistance, documentation, and postoperative analysis. Most methods treat phase annotations only as class IDs, whereas recent surgical vision-language models often require additional video–text data, captions, or instruction tuning. We propose *LaSeD*, a label-semantic self-distillation framework that uses phase names as privileged training-time context while retaining visual-only deployment without a ground-truth phase-name hint. LaSeD initializes a frozen teacher and a student from the same pretrained VLM checkpoint. The teacher receives the frame, a fixed task prompt, and the ground-truth phase-name hint; the student receives the same frame and prompt without the hint, and only its visual encoder is optimized. Training combines hard phase-token supervision with feature-level distillation from cached teacher representations. At inference, the teacher and hint are removed, and the student predicts one of the seven Cholec80 phases through constrained digit-token logits without an additional classifier head. On the Cholec80 evaluation split, LaSeD achieves 86.20% accuracy, 77.75% macro recall, 78.13% macro precision, and 64.15% macro Jaccard. Under the identical protocol, it improves a visual-only Qwen3-VL-4B baseline by 9.45, 9.42, 11.93, and 12.22 percentage points, respectively. Visual-only means that the image is the only sample-specific inference input, while all frames share the same fixed task prompt. These results support the benefit of phase-name-conditioned teacher supervision. Code: GitHub repository.

**Keywords:** Surgical phase recognition · Label-semantic distillation · Vision-language models · Self-distillation · Surgical workflow analysis

## 1 Introduction and Related Work

**Surgical phase recognition.** Surgical phase recognition (SPR) assigns each time step in a surgical video to a clinically meaningful workflow state and

⋆ Corresponding author: sara.Homscheid@medma.uni-heidelberg.de

supports intraoperative assistance, documentation, training, and postoperative review [2,1,5]. Cholec80 is a widely used laparoscopic cholecystectomy benchmark with seven phases, from preparation to gallbladder retraction [9]. SPR remains challenging because different phases may share instruments, tissue appearance, and viewpoints, while the same phase varies across patients and surgeons. Temporal models reduce this ambiguity using video context [17,6,11,10], but phase names are typically treated only as class indices. This leaves open whether ground-truth phase-name information can improve a deployable model that receives no sample-specific label text at inference.

**Language supervision for surgical understanding.** Vision-language models (VLMs) show that natural-language supervision can shape visual representations [8,12]. In surgery, this has motivated video-language pretraining and multimodal instruction tuning, including SurgVLP [15], procedure-aware pretraining [14], HecVL [13], and SurgVLM [16]. These methods demonstrate the value of language, but often require large-scale video–text data, generated captions, transcripts, or instruction data. LaSeD instead uses only the phase names already available in standard annotations.

**Knowledge distillation and label-guided teachers.** Knowledge distillation transfers information from a teacher to a student through soft predictions or intermediate features [4]. In SPR, self-distillation has been used to transfer richer representations from a teacher to a student [18]. Guan et al. [3] introduced a label-guided teacher that receives both video features and ground-truth class labels during training. Their teacher learns label-aware temporal representations through label–feature cross-attention and supervised contrastive learning, which are then distilled into a frame-only temporal student. These studies show that privileged label information can provide a richer training signal while preserving label-free inference.

These observations raise a question: can phase-name conditioning improve a visual-only VLM for surgical phase recognition? Here, *visual-only* means that the image is the only sample-specific inference input, while all frames share a fixed prompt defining the candidate phases and output format. The deployed model requires neither an external teacher nor captions, reports, or ground-truth phase-name text. We address this question with *LaSeD*, a label-semantic self-distillation framework. During training, a frozen teacher receives the frame, fixed prompt, and ground-truth phase-name hint, while a student initialized from the same checkpoint receives the frame and prompt without the hint. The student predicts the phase and aligns its representation with the phase-name-conditioned teacher. Our contributions are threefold: we extend privileged-label distillation to pretrained VLMs using human-readable phase names as natural-language context; introduce prompt-controlled self-distillation in which only the student visual encoder is optimized; and evaluate LaSeD on Cholec80 with a constrained digit-token interface, showing clear gains over a hard-label visual-only baseline under the same protocol. We do not claim overall SPR state of the art; our goal is to assess whether phase-name conditioning improves visual-only VLM adaptation.

## 2 Method

**Problem formulation.** We consider surgical phase recognition from sampled laparoscopic video frames. Let $\mathcal{V} = \{I_t\}_{t=1}^{T}$ denote a video sequence, where $I_t$ is a preprocessed RGB frame and $y_t \in \mathcal{Y} = \{1, \ldots, C\}$ its phase label. For Cholec80, $C = 7$, and each class $c$ has a human-readable phase name $s_c$, such as *Preparation* or *Calot Triangle Dissection*. A fixed prompt lists the seven candidate phases and specifies a digit-only answer format. LaSeD provides only the correct phase name $s_{y_t}$ as privileged, sample-specific text to the teacher during training; no external captions, reports, transcripts, or additional annotations are used.

**Constrained phase-token interface.** Rather than adding a task-specific classifier, we formulate phase recognition as constrained next-token prediction. Each class $c$ is mapped to a candidate digit token $\tau_c$, using the same mapping for the visual-only baseline, teacher, and student. The implementation verifies that each candidate digit is a single vocabulary token and obtains its ID from the VLM tokenizer. Given a frame $I_t$ and complete chat prompt $q$, the autoregressive VLM $f_\theta$ produces vocabulary logits at each input position. We use the logits at the final prompt position $a(q)$, which predict the first answer token; the answer digit is not included in the input prefix. Retaining only the $C$ candidate-token logits gives

$$\mathbf{z}_t(q) = [z_{t,1}(q), \ldots, z_{t,C}(q)], \qquad z_{t,c}(q) = [f_\theta(I_t, q)]_{a(q),\tau_c}, \tag{1}$$

with the restricted distribution

$$p(y = c \mid I_t, q) = \frac{\exp(z_{t,c}(q))}{\sum_{j=1}^{C} \exp(z_{t,j}(q))}. \tag{2}$$

The prediction is $\hat{y}_t = \arg\max_c p(y = c \mid I_t, q)$. This shared interface avoids randomly initialized classifier parameters.

**Label-visible teacher and visual-only student.** Inspired by the privileged-label teacher–student principle of Guan et al. [3], LaSeD adapts this idea to pretrained VLMs through prompt-controlled information visibility rather than training a dedicated label-aware temporal teacher. As shown in Figure 1, both branches are initialized from the same pretrained VLM checkpoint and use the same fixed system prompt, which lists the seven phases and requests one digit. The teacher $f_{\theta_T}$ is entirely frozen, whereas the student $f_{\theta_S}$ retains frozen language-model and output components and updates only its visual encoder.

For each training pair $(I_t, y_t)$, the student input contains the frame and fixed prompt $q^{\text{stu}}$, with no ground-truth phase name or answer token. The teacher receives the same frame and task instruction together with the sample-specific statement "This phase belongs to {Label}," where {Label}=$s_{y_t}$, denoted by $q^{\text{tea}}(s_{y_t})$. Thus,

$$x_t^{\text{tea}} = \big(I_t, q^{\text{tea}}(s_{y_t})\big), \qquad x_t^{\text{stu}} = \big(I_t, q^{\text{stu}}\big). \tag{3}$$

The branches therefore differ only in access to the ground-truth phase-name hint, not in model size or architecture. Because the teacher receives the correct

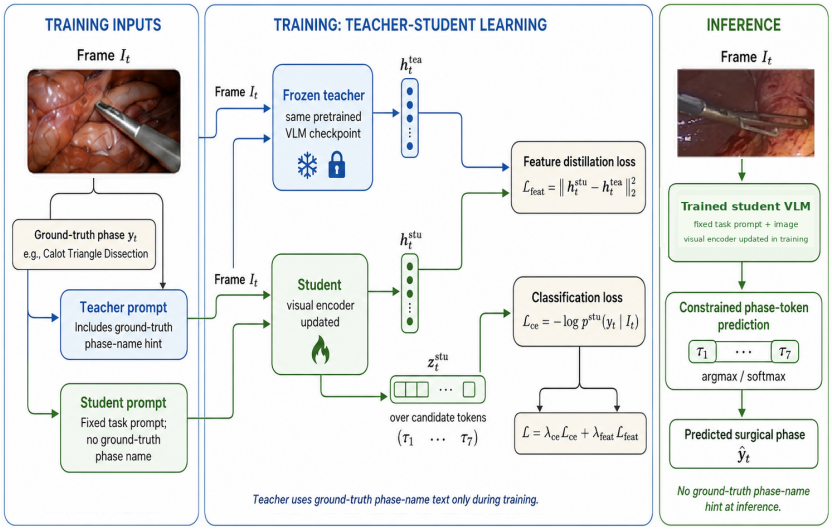



**Fig. 1.** Overview of LaSeD. The same frame and fixed task prompt are provided to both branches; only the teacher additionally receives the ground-truth phase-name hint, while the student receives no label text. The frozen teacher provides representation targets to the student, which is trained with hard phase-token supervision and feature distillation. At inference, only the student and fixed prompt are retained.

phase name, it serves only as a privileged-information branch that provides a label-conditioned representation target. The student learns to approximate this target from the frame and fixed prompt alone.

**Prompt/message structure.** Both branches use the same fixed system prompt, which lists the seven phase names with their digit mapping and instructs the model to return one digit. The student user message contains only the image. The teacher user message contains the same image plus the sample-specific text "This phase belongs to {Label}." Code: GitHub repository.

**Self-distillation objective.** Let $\mathbf{h}_{t,\ell}^{\text{tea}}, \mathbf{h}_{t,\ell}^{\text{stu}} \in \mathbb{R}^d$ denote the language-model representations extracted at decoder layer $\ell$ from $x_t^{\text{tea}}$ and $x_t^{\text{stu}}$, respectively, where $d$ is the hidden dimension. Each representation is taken at its branch-specific final prompt position, whose logits predict the first answer token. Because the prompts differ in length, these positions are determined independently. Teacher representations are precomputed, cached, and detached during student optimization. For mini-batch $\mathcal{B}$ and distillation layers $\mathcal{S}$, the hard-label loss over the restricted phase-token distribution is

$$\mathcal{L}_{\text{ce}} = -\frac{1}{|\mathcal{B}|} \sum_{t \in \mathcal{B}} \log p_{\text{stu}}(y = y_t \mid I_t, q^{\text{stu}}), \tag{4}$$

and the feature-distillation loss is

$$\mathcal{L}_{\text{feat}} = \frac{1}{|\mathcal{B}|\,|\mathcal{S}|\,d} \sum_{t\in\mathcal{B}} \sum_{\ell\in\mathcal{S}} \left\| \mathbf{h}_{t,\ell}^{\text{stu}} - \text{sg}\big(\mathbf{h}_{t,\ell}^{\text{tea}}\big) \right\|_2^2, \tag{5}$$

where $\text{sg}(\cdot)$ stops gradients through the teacher features. Teacher logits are not distilled. We use $\mathcal{S} = \{L\}$, where $L$ denotes the final language-model decoder layer. The complete objective is

$$\mathcal{L} = \lambda_{\text{ce}}\mathcal{L}_{\text{ce}} + \lambda_{\text{feat}}\mathcal{L}_{\text{feat}}, \tag{6}$$

with $\lambda_{\text{ce}} = \lambda_{\text{feat}} = 1.0$.

**Visual-only test-time prediction.** At inference, the teacher, ground-truth phase-name hint, and distillation loss are removed. The student receives $I_t$ with the same fixed prompt used for every frame, which defines the candidate phases and digit-only output format but contains no sample-specific label information. The predicted phase is selected by argmax over the seven candidate-token probabilities. Thus, the frame is the only sample-specific inference input, and LaSeD retains the same deployment interface as the visual-only baseline.

## 3 Experiments and Results

**Dataset and protocol.** We evaluate on Cholec80 [9], using the first 40 videos for training and the remaining 40 as an evaluation split. Sampling at 1 fps yields 86,304 training and 98,193 evaluation frames. During training, the 40-video evaluation split was monitored for checkpoint selection; consequently, it is not an untouched test set and the reported absolute performance may be optimistic. The controlled LaSeD–baseline comparison uses the same split, evaluation schedule, and checkpoint-selection rule for both methods. We denote the seven phases as P1–P7: Preparation, Calot Triangle Dissection, Clipping and Cutting, Gallbladder Dissection, Gallbladder Packaging, Cleaning and Coagulation, and Gallbladder Retraction. We report accuracy, macro recall, macro precision, and macro Jaccard, emphasizing macro metrics because of phase imbalance.

**Implementation details.** All experiments use Qwen3-VL-4B-Instruct [7], with teacher, student, and baseline initialized from the same checkpoint. The fixed prompt maps the seven phase names to digits 1–7 and requests a one-digit answer. The student user input contains only the image, whereas the teacher additionally receives the ground-truth phase-name hint. All compared Qwen3-VL models use the same data division, sampling, candidate tokens, preprocessing, inference prompt, and checkpoint-selection rule. Models were evaluated every 500 optimization steps, with checkpoint selection based on evaluation accuracy. With eight GPUs, per-device batch size 1, and gradient accumulation over eight steps, the effective global batch size was 64. Images are resized to shortest edge 224 while preserving aspect ratio. Neither LoRA nor QLoRA is used; the language model and output components remain frozen, and only the student visual encoder is updated. The frozen teacher caches the final-decoder representation at

**Table 1.** Controlled Qwen3-VL-4B comparison on Cholec80 under the same data division, preprocessing, constrained digit-token interface, prompt, and checkpoint-selection rule. Values are percentages.

| Method | Accuracy | Recall | Precision | Jaccard |
|---|---|---|---|---|
| Visual-only baseline | 76.75 | 68.33 | 66.20 | 51.93 |
| LaSeD | **86.20** | **77.75** | **78.13** | **64.15** |
| *Absolute gain* | +9.45 | +9.42 | +11.93 | +12.22 |
| *Relative gain* | +12.31% | +13.79% | +18.02% | +23.53% |

**Table 2.** Video-level performance over the 40 evaluation videos, reported as mean ± standard deviation across videos. Each model was trained once; these deviations quantify video-to-video variability, not variability across random seeds.

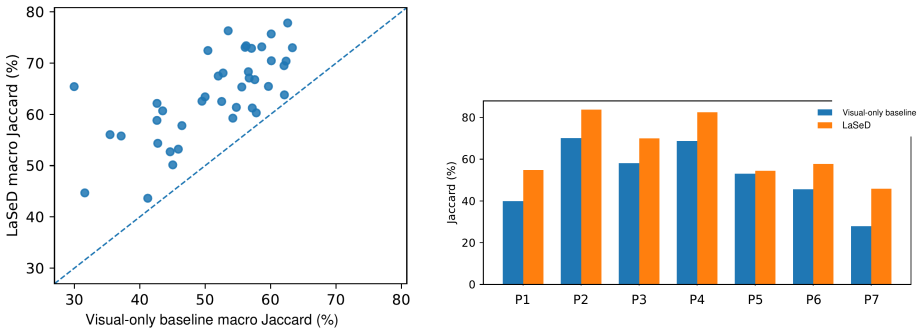
| Method | Accuracy | Recall | Precision | Jaccard |
|---|---|---|---|---|
| Visual-only baseline | $76.91 \pm 10.04$ | $68.40 \pm 7.75$ | $65.87 \pm 8.02$ | $51.51 \pm 8.82$ |
| LaSeD | $\mathbf{86.55 \pm 5.39}$ | $\mathbf{78.18 \pm 7.26}$ | $\mathbf{76.82 \pm 7.52}$ | $\mathbf{63.92 \pm 8.37}$ |

the answer-prediction position and is then removed from memory. The baseline uses only hard phase-token supervision, whereas LaSeD adds feature distillation with $\lambda_{\text{ce}} = \lambda_{\text{feat}} = 1.0$. We use AdamW with cosine scheduling, linear warmup, learning rate $1 \times 10^{-5}$, weight decay 0.01, two epochs, per-device batch size 1, and gradient accumulation over eight steps. Training took approximately 19.5 hours on 8 NVIDIA H20 GPUs; inference required approximately 3.5 hours on one NVIDIA RTX 5090, or 0.13 seconds per frame.

**Quantitative comparison.** To avoid implying direct comparability across mismatched evaluation protocols, Table 1 is restricted to the controlled comparison between the visual-only baseline and LaSeD. Both use the same Qwen3-VL-4B initialization, data division, preprocessing, digit-token interface, inference prompt, and checkpoint-selection rule. LaSeD achieves 86.20% accuracy, 77.75% macro recall, 78.13% macro precision, and 64.15% macro Jaccard, improving the visual-only baseline by 9.45, 9.42, 11.93, and 12.22 percentage points, respectively. The largest relative improvement is observed for macro Jaccard (+23.53%).

**Video-level and phase-wise analysis.** Table 2 reports variability across the 40 evaluation videos. Per-video values differ slightly from the global frame-level metrics in Table 1. LaSeD improves the mean of every metric, reduces accuracy variability, and outperforms the baseline in accuracy and macro Jaccard on all 40 videos. These statistics characterize variation across videos, not training seeds. Figure 2 shows that LaSeD also improves Jaccard for all seven phases, with the largest gains for P7, P1, P4, and P2. The smallest gain occurs for P5, where higher recall but lower precision indicates additional false positives for this short, visually similar late phase.

**Confusion patterns and temporal behavior.** Figure 3 shows stronger diagonal responses for LaSeD, indicating improved phase separation. Remaining errors mainly involve the temporally adjacent phases P5–P7, which can appear




(a) Per-video macro Jaccard (b) Per-phase Jaccard

**Fig. 2.** Additional analysis on the 40 evaluation videos: (a) per-video macro Jaccard and (b) per-phase Jaccard.

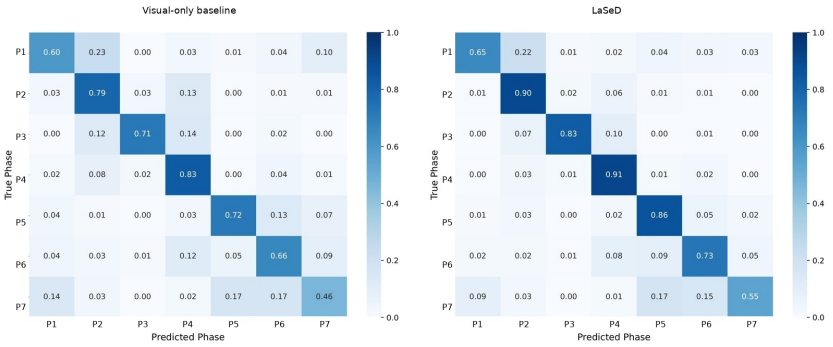



**Fig. 3.** Normalized confusion matrices. LaSeD strengthens most diagonal entries, while visually similar late phases remain challenging.

similar in isolated frames. The phase bars in Figure 4 show more coherent predictions than the baseline, although residual fluctuations suggest that explicit temporal modeling could further improve workflow consistency.

## 4 Discussion and Conclusion

**Why phase-name conditioning helps.** The controlled comparison indicates that adding a phase-name-conditioned teacher target provides a richer training signal than hard phase-token supervision alone. The frozen teacher combines the surgical frame with the phase-name hint, producing a representation that reflects visual evidence and class-level context. By aligning with this representation, the student learns phase-discriminative visual features while operating without the hint at inference. This is especially useful when phases share instruments, anatomical appearance, or viewpoints, where hard class supervision

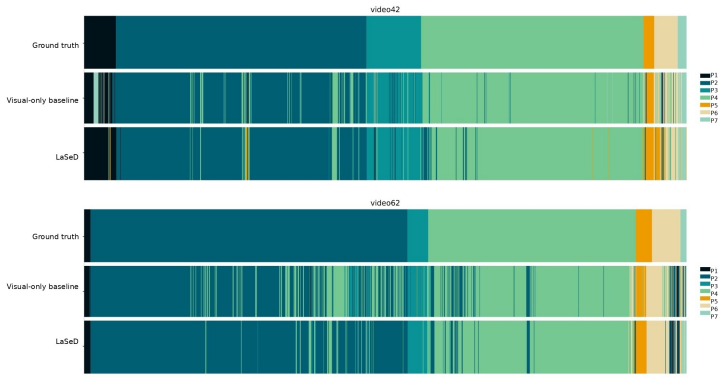



**Fig. 4.** Representative phase predictions for two evaluation videos. Rows show ground truth, the visual-only baseline, and LaSeD.

may be insufficient. LaSeD therefore exploits phase-name information already present in standard annotations without requiring captions, reports, text data, or a separately trained teacher.

**Consistency and practical implications.** LaSeD improves all four global metrics, with the largest relative gain in macro Jaccard, indicating better class-wise separation rather than improvement driven only by dominant phases. The gain is consistent across all 40 evaluation videos and all seven phases, showing that the benefit is broadly distributed. The phase-bar visualizations show more coherent predictions than the visual-only baseline, while the confusion matrices reveal stronger diagonal responses for most phases. Importantly, these improvements do not increase deployment-time input requirements: the teacher and ground-truth phase-name hint are removed, and inference uses only the trained student with the same fixed prompt as the baseline. The approach therefore adds training supervision without requiring a teacher model, external text, or sample-specific language at deployment.

**Remaining challenges and outlook.** Most residual errors occur among the temporally adjacent late phases P5–P7, which can be difficult to distinguish from isolated frames. Combining LaSeD with temporal sequence modeling or lightweight smoothing may further improve workflow consistency. The current experiments support the value of a phase-name-conditioned privileged teacher relative to the matched visual-only baseline, but do not determine whether human-readable linguistic semantics are responsible for the gain; arbitrary or shuffled identifiers might also help. Results are further limited to Cholec80, one surgical procedure, one VLM backbone, and one training run. Future work should therefore test numeric/shuffled label variants, additional procedures and VLMs, multiple random seeds, and a protocol with separate validation and untouched test sets.

## Disclosure of Interests

The authors have no competing interests to declare.

## References


1. Demir, K.C., Schieber, H., Weise, T., Roth, D., May, M., Maier, A., Yang, S.H.: Deep learning in surgical workflow analysis: A review of phase and step recognition. IEEE Journal of Biomedical and Health Informatics **27**(11), 5405–5417 (2023). https://doi.org/10.1109/JBHI.2023.3311628
2. 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., Müller-Stich, B.P., Nickel, F.: Machine learning for surgical phase recognition: A systematic review. Annals of Surgery **273**(4), 684–693 (2021). https://doi.org/10.1097/SLA.0000000000004425
3. Guan, J., Zou, X., Tao, R., Zheng, G.: Label-guided teacher for surgical phase recognition via knowledge distillation. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2024. pp. 349–358. Springer (2024). https://doi.org/10.1007/978-3-031-72089-5_33
4. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. In: NIPS Deep Learning and Representation Learning Workshop (2015)
5. Li, Y., Zhao, Z., Li, R., Li, F.: Deep learning for surgical workflow analysis: A survey of progresses, limitations and trends. Artificial Intelligence Review **57**, 291 (2024). https://doi.org/10.1007/s10462-024-10929-6
6. Pan, X., Gao, X., Wang, H., Zhang, W., Mu, Y., He, X.: Temporal-based swin transformer network for workflow recognition of surgical video. International Journal of Computer Assisted Radiology and Surgery **18**(1), 139–147 (2023). https://doi.org/10.1007/s11548-022-02785-y
7. Qwen Team: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
8. 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)
9. Twinanda, A.P., Shehata, S., Mutter, D., Marescaux, J., de Mathelin, M., Padoy, N.: Endonet: A deep architecture for recognition tasks on laparoscopic videos. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016. pp. 575–582. Springer (2016)
10. Wu, H., Wang, T.H., Lechner, M., Hasani, R., Eckhoff, J.A., Pak, P., Meireles, O.R., Rosman, G., Ban, Y., Rus, D.: Holistic surgical phase recognition with hierarchical input dependent state space models. arXiv preprint arXiv:2506.21330 (2025)
11. Yang, S., Luo, L., Wang, Q., Chen, H.: Surgformer: Surgical transformer with hierarchical temporal attention for surgical phase recognition. In: Medical Image Computing and Computer Assisted Intervention – MICCAI 2024. pp. 606–616. Springer (2024). https://doi.org/10.1007/978-3-031-72089-5_57
12. Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., Wu, Y.: Coca: Contrastive captioners are image-text foundation models. arXiv preprint arXiv:2205.01917 (2022)

13. Yuan, K., Srivastav, V., Navab, N., Padoy, N.: Hecvl: Hierarchical video-language pretraining for zero-shot surgical phase recognition. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2024. pp. 306–316. Springer (2024)
14. Yuan, K., Srivastav, V., Navab, N., Padoy, N.: Procedure-aware surgical video-language pretraining with hierarchical knowledge augmentation. In: Advances in Neural Information Processing Systems (2024)
15. 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 **105**, 103644 (2025). https://doi.org/10.1016/j.media.2025.103644
16. Zeng, Z., Zhuo, Z., Jia, X., Zhang, E., Wu, J., Zhang, J., Wang, Y., Low, C.H., Jiang, J., Zheng, Z., et al.: Surgvlm: A large vision-language model and systematic evaluation benchmark for surgical intelligence. arXiv preprint arXiv:2506.02555 (2025)
17. Zhang, B., Goel, B., Sarhan, M.H., Goel, V.K., Abukhalil, R., Kalesan, B., Stottler, N., Petculescu, S.: Surgical workflow recognition with temporal convolution and transformer for action segmentation. International Journal of Computer Assisted Radiology and Surgery **18**(4), 785–794 (2023). https://doi.org/10.1007/s11548-022-02811-z
18. Zhang, J., Barbarisi, S., Kadkhodamohammadi, A., Stoyanov, D., Luengo, I.: Self-knowledge distillation for surgical phase recognition. International Journal of Computer Assisted Radiology and Surgery (2024), arXiv:2306.08961