Endo-DyLat++: Semantic Readiness Monitoring for Endoscopic Vision-Language-Action Robots

Xiyue Xiao, Chi Kit Ng, and Hongliang Ren
Department of Electronic Engineering
The Chinese University of Hong Kong, Hong Kong SAR, China

Abstract—Healthcare robots increasingly rely on learned visual policies, language interfaces, and shared-autonomy loops, but a policy also needs to know when the current perceptual state is reliable enough for action. In endoscopic robotics, target loss, occlusion, blur, and tissue deformation can make a visually plausible frame unsafe for continued autonomy. We present Endo-DyLat++, a lightweight semantic readiness monitor for endoscopic vision-language-action robots. We define semantic readiness as whether the current visual state is reliable enough for a downstream system to proceed, given the anticipated risks of target loss, image degradation, and occlusion. The monitor separates relatively stable anatomical and camera context from time-varying risk cues. It predicts these near-future events from weak proxy labels and selects a validation operating point under a constraint on unsafe-ready acceptances, in which risky samples are incorrectly accepted as ready. On a Kvasir-SEG real-image pseudo-video benchmark, Endo-DyLat++ is comparable to a strong frozen-DINOv2 single-latent baseline within seed variance, reduces variance on several aggregate metrics, and exposes a clear safety/coverage tradeoff: higher readiness recall and coverage come with more unsafe-ready errors. These results should not be interpreted as clinical validation or evidence of a deployed surgical robotic system; rather, they provide a compact testbed for converting endoscopic visual risk into an auditable pre-action readiness signal.

I. Introduction

AI-enabled healthcare robots are moving toward shared autonomy, learned perception, and language-conditioned control. Robot foundation and vision-language-action (VLA) models such as SayCan, RT-1, RT-2, PaLM-E, and Open X-Embodiment show that semantic representations can connect language, visual context, and robot actions across tasks and embodiments [2, 4, 5, 9, 15]. In clinical and surgical settings, however, semantic action generation is only part of the problem. Before a learned policy moves an instrument, continues tracking, or accepts an autonomous subtask, the system should know whether the current visual state is ready for action.

This problem is acute in endoscopic robotics. The camera view is deformable, reflective, partially occluded, and highly dependent on operator motion. A current frame can appear usable while the near future contains target loss, image degradation, or occlusion. Existing surgical video benchmarks support phase recognition, tool detection, scene segmentation, and action triplet recognition [1, 7, 8, 12], while endoscopic foundation models study large-scale visual pretraining [11, 13]. These directions are complementary to our goal: a lightweight monitor that converts visual risk evidence into a pre-action readiness decision for a downstream healthcare robot.

We study semantic readiness monitoring as an intermediate layer between visual perception and VLA control. Here, semantic readiness means that the current visual state is sufficiently reliable for a downstream system to proceed, given anticipated target loss, image degradation, and occlusion. Given a short endoscopic clip, the monitor predicts these future visual risk states and a scalar readiness score. This visual criterion does not assess action feasibility, tissue interaction, or overall clinical risk. The safety-critical failure considered here is an unsafe-ready error: the monitor declares the state ready although at least one visual risk is present. Our contribution is threefold. First, we introduce Endo-DyLat++, a static/dynamic latent scaffold that separates anatomical context from dynamic failure evidence. Second, we build a weakly supervised real-image pseudo-video testbed from Kvasir-SEG polyp images and masks [6]. Third, we report a validation-selected operating-point interface that makes the safety/coverage tradeoff explicit rather than optimizing a single aggregate score.

II. Method

Problem formulation. For an input endoscopic clip $x_{1:T}$, the monitor predicts a future risk vector

$$r = [r_{\text{loss}}, r_{\text{degrade}}, r_{\text{occlude}}], \tag{1}$$

and a readiness score $q \in [0, 1]$. The components of r estimate whether the near future contains target loss, visual degradation, or occlusion. The readiness score estimates whether a downstream tracker, planner, shared autonomy module, or VLA policy should proceed. This formulation is deliberately policy-agnostic: it does not choose the robot action, but it can reject, pause, or defer before an action is executed.

Static and dynamic risk latents. Endo-DyLat++ encodes each $T = 8$ frame clip with a frozen DINOv2 ViT-S/14 frame encoder [10]. Frame features are pooled by concatenating temporal mean, final-frame feature, first-to-last feature delta, and temporal maximum. A compact adapter maps this pooled feature into two 48-dimensional latents:

$$z_s = f_s(x_{1:T}), \qquad z_d = f_d(x_{1:T}). \tag{2}$$

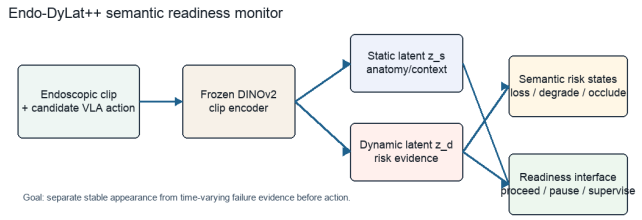

Fig. 1. Endo-DyLat++ semantic readiness monitor. A short endoscopic clip is encoded into static anatomy and dynamic risk latents. Future visual risk and a readiness score are converted into a policy-facing proceed, pause, or defer signal.

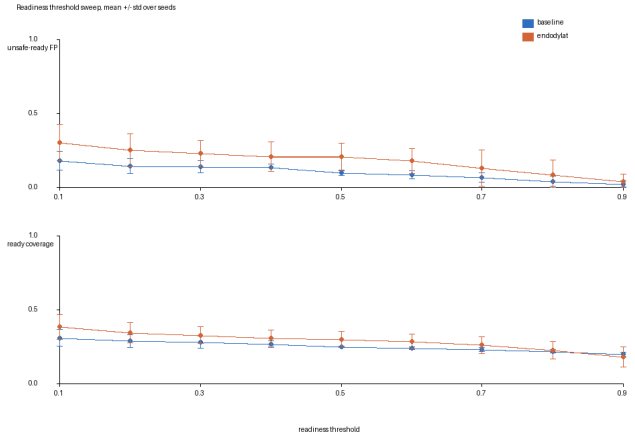

Fig. 2. Readiness operating-point sweep for DINOv2 with temporal mean/last/delta/max pooling. The selected threshold maximizes ready coverage under a validation unsafe-ready FP-rate constraint, then is evaluated on held-out test data.

The static latent z_s is encouraged to capture slowly varying anatomical and camera context, following temporal invariance ideas from slow feature analysis [14]. The dynamic latent z_d is encouraged to carry time-varying risk evidence, following the intuition of video disentanglement [3]. The two latents are concatenated for the risk and readiness heads.

Training objective and readiness interface. The training loss combines binary cross-entropy for future risk and readiness with static consistency, dynamic contrast, and static/dynamic orthogonality regularizers:

$$\mathcal{L} = \lambda_r \mathcal{L}_{risk} + \lambda_q \mathcal{L}_{ready} + \lambda_s \mathcal{L}_{static} + \lambda_d \mathcal{L}_{dynamic} + \lambda_o \mathcal{L}_{orth}. \tag{3}$$

Same-source pseudo-clips are regularized to keep z_s stable, while clips with different weak any-risk states are separated in z_d. The orthogonality term penalizes static/dynamic covariance and squared cosine similarity. We use a risk-first setting with $\lambda_r = \lambda_q = 2.0$, $\lambda_s = \lambda_d = 0.05$, and $\lambda_o = 0.01$.

At evaluation time, 0.5 is used as a conventional threshold for converting each risk output into a binary prediction. It is not a clinically validated safety boundary and does not determine whether the downstream system proceeds; that decision uses a separate readiness threshold. We select the readiness threshold on the validation split by maximizing ready coverage subject to an unsafe-ready FP-rate constraint of 0.06, then apply it once to the held-out test split. This interface exposes a safety/utility operating point rather than choosing a favorable threshold on the test set. Both thresholds would require application-specific calibration before deployment.

III. Experimental Setup

Data and weak labels. We use Kvasir-SEG, a real colonoscopy image dataset with polyp segmentation masks [6]. Because it is image-based rather than a robot video dataset, we construct pseudo-video clips by applying controlled temporal transformations to each image and mask. The transformations simulate stable observation, boundary drift, target loss, degradation, occlusion, and proximity-like mask changes. Weak labels are derived from mask visibility, target area, target center, future visibility loss, and occlusion proxies. Samples are split by image identifier into train, validation, and test sets using a 70/15/15 split for each random seed. These clips retain real endoscopic appearance but do not reproduce the full temporal, anatomical, or procedural complexity of endoscopic video. The transformations were not evaluated by clinicians in this study. We therefore use this dataset only as a controlled proof-of-concept testbed, not as evidence of clinical realism or real-world safety.

Baselines and metrics. The baseline uses the same frozen DINOv2 encoder, temporal pooling, and prediction heads but maps the pooled clip feature into one 64-dimensional latent instead of a static/dynamic split. Excluding the shared frozen backbone, the baseline has 0.239M trainable adapter/head parameters and Endo-DyLat++ has 0.280M. We report AUROC/AUPRC for target loss, degradation, occlusion, aggregated any-risk prediction, and readiness. Any-risk is positive when at least one of the three visual risks is present. Ready coverage is the fraction of all samples accepted as ready, while readiness precision and recall use the weak readiness label as ground truth. We define the unsafe-ready FP rate as the number of accepted samples with any ground-truth visual risk divided by the total number accepted as ready. Thus, despite the "FP" label, it measures contamination of the accepted set rather than the conventional class-conditional false-positive rate. We also report absolute ready and unsafe-ready accepted counts.

IV. Results

Table I shows that the split monitor is not a large-margin winner over a strong frozen-DINOv2 baseline. Aggregate ranking metrics are comparable within seed variance, and per-event behavior is mixed: target-loss recall increases slightly, degradation AUROC increases

TABLE I

Five-seed Kvasir-SEG pseudo-video results with frozen DINOv2 features. Values are mean±std; Δ is Endo-DyLat++ minus baseline. Arrows indicate whether larger (↑) or smaller (↓) is better.

Metric	Baseline	Endo-DyLat++	Δ
Target-loss recall ↑	0.748±0.102	0.761±0.149	+0.012
Degradation AUROC ↑	0.959±0.009	0.963±0.005	+0.004
Occlusion AUROC ↑	0.999±0.001	0.997±0.005	-0.003
Any-risk AUROC ↑	0.980±0.012	0.985±0.005	+0.004
Any-risk AUPRC ↑	0.993±0.004	0.994±0.002	+0.001
Readiness AUROC ↑	0.982±0.009	0.985±0.007	+0.003
Readiness AUPRC ↑	0.956±0.022	0.964±0.015	+0.008

TABLE II

Five-seed readiness operating points. Thresholds are selected on validation data under an unsafe-ready FP-rate constraint of 0.06 and evaluated on held-out test data. Counts are out of 150 test samples. Arrows indicate whether larger (↑) or smaller (↓) is better; selected threshold is an operating parameter.

Metric	Baseline	Endo-DyLat++
Selected threshold	0.72±0.16	0.60±0.16
Ready accepted ↑	31.6±4.2	34.8±3.1
Unsafe accepted ↓	1.8±2.2	2.4±2.1
Ready coverage ↑	0.211±0.028	0.232±0.021
Unsafe-ready FP ↓	0.051±0.060	0.066±0.056
Readiness recall ↑	0.742±0.046	0.808±0.060

slightly, and occlusion AUROC is near saturated for both models. The most consistent positive pattern is variance reduction on aggregate metrics. Standard deviation decreases for any-risk AUROC (0.0117 to 0.0052), readiness AUROC (0.0092 to 0.0066), and degradation AUROC (0.0088 to 0.0048), which is useful for a safety monitor even when mean differences are small.

Table II makes the monitor's main value explicit. Endo-DyLat++ accepts more samples and improves readiness recall, but it also accepts more unsafe-ready samples. Its validation-selected threshold does not fully generalize to the test split: the mean test unsafe-ready FP rate is 0.066, above the 0.06 budget, while the baseline remains at 0.051. We therefore do not claim that Endo-DyLat++ is safer at the reported operating point. Instead, the readiness interface exposes the tradeoff that a clinical or robot-system designer would need to tune.

Two diagnostic analyses help interpret these results. A three-seed backbone-strength sweep showed that the target-loss recall gain is larger with a weak smoke-test encoder and contracts toward zero with DINOv2, consistent with the interpretation that stronger pretrained features already encode some dynamic evidence. Three-seed ablations showed mixed but informative effects: removing the dynamic loss increased ready coverage and target-loss recall but also increased the unsafe-ready FP rate, while removing the orthogonality term increased covariance leakage and worsened target-loss recall. These diagnostics support a cautious conclusion: the split helps make monitor tradeoffs inspectable, but the present loss weighting is not an optimized safety objective.

V. Discussion and Conclusion

Endo-DyLat++ is intended as a small monitoring layer for healthcare robotics, not as a complete surgical autonomy stack. The current benchmark uses real endoscopic appearance and real segmentation masks, but its temporal dynamics are procedural and its labels are weak proxies. It does not include robot actions, force, depth, tool contact, or clinical intervention labels. The evaluation also uses a single image dataset, so it does not establish temporal or cross-dataset generalization. Its synthetic transformations were not validated by clinicians and cannot establish that the pseudo-videos represent real endoscopic failure modes. Moreover, improvements over the baseline are modest and generally within seed variance, and the reported operating point does not reduce unsafe-ready errors. The results therefore support a limited claim: static/dynamic risk latents and a validation-selected readiness threshold can turn weakly labeled endoscopic visual evidence into an auditable pre-action safety/coverage interface. They do not demonstrate improved clinical safety or readiness for deployment.

This limitation also defines the next step. In a full healthcare robot system, readiness should be conditioned on candidate actions and evaluated on real surgical or endoscopic video with tool, anatomy, and robot-state annotations. The monitor should answer not only whether the scene is visually ready, but whether a proposed action is likely to preserve target visibility, avoid occlusion, and keep the system ready for the next step. Endo-DyLat++ provides a compact experimental scaffold for that direction: a semantic readiness signal that can decide when a learned endoscopic policy should proceed, pause, or defer to human supervision.

References

[1] Max Allan, Satoshi Kondo, Sebastian Bodenstedt, Stefan Leger, Rahim Kadkhodamohammadi, Imanol Luengo, Felix Fuentes, Evangello Flouty, Ahmed Mohammed, Marius Pedersen, et al. 2018 robotic scene segmentation challenge. arXiv preprint arXiv:2001.11190, 2020.

[2] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, et al. RT-1: Robotics transformer for real-world control at scale. In Robotics: Science and Systems, 2023.

[3] Emily L. Denton and Vighnesh Birodkar. Unsupervised learning of disentangled representations from video. In Advances in Neural Information Processing Systems, 2017.

[4] Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe

Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. PaLM-E: An embodied multimodal language model. In International Conference on Machine Learning, volume 202 of Proceedings of Machine Learning Research, 2023.

[5] Brian Ichter, Anthony Brohan, Yevgen Chebotar, Chelsea Finn, Karol Hausman, Alexander Herzog, Daniel Ho, Julian Ibarz, Alex Irpan, Eric Jang, Ryan Julian, Dmitry Kalashnikov, Sergey Levine, Yao Lu, Carolina Parada, Kanishka Rao, Pierre Sermanet, Alexander T. Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Mengyuan Yan, Noah Brown, Michael Ahn, Omar Cortes, Nicolas Sievers, Clayton Tan, Sichun Xu, Diego Reyes, Jarek Rettinghouse, Jornell Quiambao, Peter Pastor, Linda Luu, Kuang-Huei Lee, Yuheng Kuang, Sally Jesmonth, Nikhil J. Joshi, Kyle Jeffrey, Rosario Jauregui Ruano, Jasmine Hsu, Keerthana Gopalakrishnan, Byron David, Andy Zeng, and Chuyuan Kelly Fu. Do as i can, not as i say: Grounding language in robotic affordances. In Proceedings of The 6th Conference on Robot Learning, volume 205 of Proceedings of Machine Learning Research, pages 287–318. PMLR, 2023.

[6] Debesh Jha, Pia H. Smedsrud, Michael A. Riegler, Pål Halvorsen, Thomas de Lange, Dag Johansen, and Håvard D. Johansen. Kvasir-seg: A segmented polyp dataset. In MultiMedia Modeling, pages 451–462. Springer, 2020.

[7] Chinedu Innocent Nwoye and Nicolas Padoy. Data splits and metrics for method benchmarking on surgical action triplet datasets. arXiv preprint arXiv:2204.05235, 2022.

[8] Chinedu Innocent Nwoye, Deepak Alapatt, Tong Yu, et al. Cholectriplet2021: A benchmark challenge for surgical action triplet recognition. Medical Image Analysis, 86:102803, 2023.

[9] Open X-Embodiment Collaboration, Abby O'Neill, Abdul Rehman, Abhinav Gupta, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, et al. Open x-embodiment: Robotic learning datasets and RT-X models. arXiv preprint arXiv:2310.08864, 2024.

[10] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. Transactions on Machine Learning Research, 2024. URL https://openreview.net/forum?id=a68SUt6zFt.

[11] Qingyao Tian, Huai Liao, Xinyan Huang, Bingyu Yang, Dongdong Lei, Sebastien Ourselin, and Hongbin Liu. Endomamba: An efficient foundation model for endoscopic videos via hierarchical pre-training. In Medical Image Computing and Computer Assisted Intervention – MICCAI 2025, volume 15968 of Lecture Notes in Computer Science, pages 224–234. Springer, 2026. doi: 10.1007/978-3-032-05114-1_22.

[12] Andru P. 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.

[13] Zhao Wang, Chang Liu, Shaoting Zhang, and Qi Dou. Foundation model for endoscopy video analysis via large-scale self-supervised pre-train. In Medical Image Computing and Computer Assisted Intervention – MICCAI 2023, pages 101–111. Springer, 2023.

[14] Laurenz Wiskott and Terrence J. Sejnowski. Slow feature analysis: Unsupervised learning of invariances. Neural Computation, 14(4):715–770, 2002.

[15] Brianna Zitkovich, Tianhe Yu, Sichun Xu, Peng Xu, Ted Xiao, Fei Xia, Jialin Wu, Paul Wohlhart, Stefan Welker, Ayzaan Wahid, Quan Vuong, Vincent Vanhoucke, Huong Tran, Radu Soricut, Anikait Singh, Jaspiar Singh, Pierre Sermanet, Pannag Sanketi, Grecia Salazar, Michael Ryoo, Krista Reymann, Kanishka Rao, Karl Pertsch, Igor Mordatch, Henryk Michalewski, Yao Lu, Sergey Levine, Tsang-Wei Edward Lee, Lisa Lee, Isabel Leal, Yuheng Kuang, Dmitry Kalashnikov, Ryan Julian, Nikhil Joshi, Alex Irpan, Brian Ichter, Jasmine Hsu, Alexander Herzog, Karol Hausman, Keerthana Gopalakrishnan, Chuyuan Fu, Pete Florence, Chelsea Finn, Avinava Dubey, Danny Driess, Tianli Ding, Krzysztof Choromanski, Xi Chen, Yevgen Chebotar, Justice Carbajal, Noah Brown, Anthony Brohan, Montserrat Gonzalez Arenas, and Kehang Han. RT-2: Vision-language-action models transfer web knowledge to robotic control. In Proceedings of The 7th Conference on Robot Learning, volume 229 of Proceedings of Machine Learning Research, pages 2165–2183, 2023.