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.

What a Frozen Foundation Model Already Knows about the Critical View of Safety: Toward a Real-Time Monitor

Soham Chaudhuri[1], Gerrit Jansen[1], Remco Hogerwerf[1], Joseph Krol[1], and Ana Manzano Rodríguez[2,3,4,5]

[1] University of Amsterdam
[2] Data Science Center HAVA-Lab, University of Amsterdam
[3] Video & Image Sense Lab, Informatics Institute, University of Amsterdam
[4] Amsterdam UMC Location University of Amsterdam, Surgery
[5] Amsterdam Public Health, Digital Health

Abstract. Bile duct injury during laparoscopic cholecystectomy is one of the most preventable surgical complications. It can be averted by confirming the Critical View of Safety (CVS), yet this is assessed subjectively and rarely monitored continuously. We investigate whether a single, frozen surgical foundation model already possesses the specialized representations required for automated CVS assessment. Benchmarking five foundation models, we select SurgeNetXL (DINOv3 ViT-L), which nearly doubles the performance of the strongest general domain model. We find that explicit depth and segmentation priors provide no meaningful uplift, consistent with the frozen backbone already encoding much of the relevant 3D geometry (RGB depth CKA = 0.76). Ablating this depth concept via null-space projection yields suggestive evidence that the model selectively uses this implicit geometry: erasing it degrades the 3D structural criterion (C3) while slightly improving the purely 2D semantic criterion (C1). Furthermore, parameter-efficient adaptation (mAP 0.48) outperforms full fine-tuning (0.43) by preventing overfitting. A sparse attention sweep reveals CVS temporal evidence is remarkably short-ranged (2-3 frame optimal causal window). Leveraging this, a confidence-aware causal label propagation framework turns sparse 18-frame anchors into dense targets. Our frozen-feature causal temporal bridge predicts each frame online at mAP 0.49±0.06 up from a 0.41 frozen-probe baseline, at 13.5 ms per frame with latency independent of clip length, outperforming fine-tuned variants and serving as a reliable rule-out signal in our evaluation (NPV ≈99%). As more complex methods do not exceed this bridge, we attribute this ceiling to the sparse-label budget rather than model capacity. Ultimately, strong frozen representations paired with efficient causal reasoning offer a principled path to a deployable CVS safety monitor.

Keywords: Critical View of Safety · Surgical foundation models · Frozen representations · Causal temporal modeling · Real-time surgical system.

1 Introduction

Laparoscopic cholecystectomy carries a persistent 0.4–1.5% risk of bile-duct injury (BDI) [4]. BDI is uncommon but disproportionately harmful: patients often need complex reconstruction and face elevated long-term disability and mortality [5,6], at a reported first-year cost above $60,000 per case.

Most BDIs arise not from faulty execution but from visual misperception of the anatomy [2]: because the harm originates in what the surgeon perceives rather than does, the most direct safeguard helps the team correctly interpret what they see.

The Critical View of Safety (CVS) [1], addresses exactly this. Before any structure is clipped, the surgeon confirms three binary conditions: (C1) only two tubular structures enter the gallbladder; (C2) the hepatocystic triangle is cleared of fat and fibrous tissue; and (C3) the lower third of the gallbladder is separated from the liver bed. When all three hold, the great majority of major BDIs are avoided [3]. Yet, CVS is documented as achieved in only an estimated 10-15% of procedures, largely because manual visual verification is subjective. A system that verifies CVS consistently and continuously would therefore add an intra-operative safety check.

Building such a verifier requires scarce and costly expert annotations, faces wide variation in operative appearance across institutions and equipment, and must meet a hard real-time constraint. Self-supervised foundation models help, transferring with far fewer labels than from-scratch models [10,11]. Existing automated-CVS systems, by contrast, are typically task-specific multimodal networks trained from scratch [9,22] –data-hungry, domain shift sensitive, and hard to trust in deployment [7].

Every extra component (a depth network, a segmentation stream, a fine-tuned backbone, or a long temporal memory) is another failure mode and a data and latency cost. We hence invert the usual design question; rather than what to add to a CVS model, we ask how much a single frozen surgical foundation model provides, so the deployed monitor stays minimal, real-time, and predictable. Equally important for safety, we ask what it fundamentally cannot do, since a monitor is only safe to field once its ceiling and failure modes are understood.

Concretely, we study five questions, each a barrier between a frozen backbone and a deployable CVS safeguard. **(RQ1)** Which frozen foundation model is better suited for learning CVS understanding without task-specific training? **(RQ2)** Does a safe monitor need explicit 3D or segmentation priors, or does the backbone already encode the geometry the CVS criteria depend on? **(RQ3)** How much adaptation is necessary and whether heavy adaptation is harmful or helpful? **(RQ4)** Over what temporal horizon must an online monitor integrate evidence, that is, how much memory and latency must it carry? **(RQ5)**

Do these compose into a trustworthy real-time causal monitor, and what is its clinical characterization and fundamental limit? Throughout, mAP is our primary metric for its threshold-free nature and robustness to the extreme class imbalance of CVS (roughly 2% positive frames).

2 Related Work

Vision has shifted from task-specific models toward foundation models pretrained at scale. CLIP [12] showed that such representations transfer zero-shot, and biomedical variants like BioMedCLIP [13] followed; these are trained largely on static modalities (pathology, radiographs) and transfer poorly to the dynamic surgical scene. Surgery-specific self-supervision has since emerged, from methodological studies [11] to large-scale in-domain backbones such as the SurgeNet family [10], with vision-language [18,16] and video-native [17] models adding multimodal and temporal pretraining.

Automated CVS assessment is established, and its trajectory motivates our design. Early systems are two-stage and segmentation-driven, which segment the gallbladder-region anatomy, then decide each criterion [22]. This is accurate but needs dense pixel annotations and runs offline. Recent work instead adapts foundation models, e.g., aligning image features with per-criterion text prompts [9], and the SAGES 2024 challenge provides a multi-centric benchmark [8]. Prior methods thus depend on segmentation or fine-tune the backbone and are evaluated offline; none ask what a single frozen backbone already provides in a real-time, causal label-scarce regime, a deployable monitor needs. That is our setting.

3 Methodology

3.1 Data and evaluation protocol

We use the official SAGES 2024 CVS Challenge dataset [8]: 1,000 laparoscopic cholecystectomy videos (700 train / 300 test) from 54 institutions in 24 countries. Each is a 90 s window before the first cystic clip, sampled at one frame per five seconds, giving exactly 18 anchor frames; three annotators rate each anchor with a confidence, and their majority vote is the evaluation label. Labels exist only on anchors, where all metrics are computed. We use two protocols and no cross-protocol comparisons are made: backbone selection uses a 5-fold CV linear probe, all other experiments the official single split (each table states its protocol). mAP is primary, Macro-F1 balances the criteria, and per-criterion accuracy is shown only alongside mAP given the $\sim$2% prevalence. The causal bridge and two ceiling models use three seeds (mean$\pm$std); other results are single runs. The training details can be found in the supplementary material.

3.2 Foundation model benchmarking (RQ1)

To measure intrinsic surgical knowledge before adaptation, we benchmark five backbones spanning general-purpose (EVA-02 [14], C-RADIO-v4 [15]), surgical (SurgeNetXL [10]), vision–language (SurgCLIP [16]), and video-pretrained (UniSurg [17]) families. We extract per-frame features from the 18 anchors and train a linear probe on frozen features; SurgCLIP, being natively vision–language aligned, is evaluated zero-shot via image–prompt cosine similarity.

3.3 Geometric priors and multi-stream fusion (RQ2)

Because confirming isolated tubular structures (C1) and gallbladder separation (C3) depends on 3D geometry that is not directly observable in laparoscopic video, we analyze the contribution of geometric priors in SurgeNetXL. First, a Siamese model fuses monocular depth maps with RGB images: RGB frames are encoded with the DINOv3 ViT-Large (1024-d) backbone and depth maps with the SurgeNetXL CaFormer-S18 backbone (512-d), and the fused representations are fed to an MLP fusion head, with the RGB modality subjected to 20% dropout in training to avoid over-reliance on the dominant modality. To assess the independence of complementary modality streams, we use three frozen encoders per frame SurgeNetXL CaFormer-S18 (512-d), DepthAnythingV2 Base [19] (768-d), and SAM 2 Hiera-L [20] (256-d) and explore three fusion techniques: (i) simple concatenation followed by a two-hidden-layer MLP; (ii) fusion with learned stream-specific weights via a softmax gating mechanism to reduce the importance of low-quality streams; and (iii) a cross-attention mechanism based on a compact transformer encoder. Additionally, we analyze the representations with a linear Centered Kernel Alignment (CKA) on the RGB and depth streams (Eq. 1) and a dense spatial probe, bypassing the global [CLS] token, to reconstruct ground-truth depth maps from the frozen DINOv3 spatial tokens.

$$\mathrm{CKA}(X, Y) = \frac{\langle XX^\top, YY^\top \rangle_F}{\|XX^\top\|_F \, \|YY^\top\|_F} \tag{1}$$

Finally, to test whether the model actively relies on this implicit geometry, we project the 1024-d features onto the null-space of the trained depth probe, erasing geometric variance while preserving semantic texture, and measure the effect on each CVS criterion.

3.4 Adaptation strategies (RQ3)

We compare three adaptation levels for SurgeNetXL: a frozen backbone with a trainable MLP head; full fine-tuning of the final transformer block; and LoRA adapters on the attention and projection layers of the final blocks, keeping the rest of the backbone frozen.

3.5 Sparse temporal attention (RQ4)

Sparse attention [21] is used to describe the temporal evolution of the CVS model instead of dense attention to ensure interpretability and computational efficiency. Three types of models were screened: learnable spatial, temporal-local, and hybrid type, and the temporal-local model was adopted. The temporal-local model uses a causal local mask that allows the model to attend to only a few recent predecessors before the current frame. Following this, varying the causal window size helps determine how much temporal information is required by the model. Since the attention weights are exposed, it leads to an interpretable model, it enables per-video attention maps, lag-conditioned summaries, and targeted frame-pair

inspections without sacrificing an end-to-end trainable architecture. The model is trained for 25 epochs with evaluations after each epoch to identify the optimal operating point before overfitting.

3.6 Causal label propagation and the streaming bridge (RQ5)

To enable the applicability in real-time, dense and causal supervision based on sparse labels is needed. For an anchor frame with label l_a at time t_a, we propagate a strictly causal target, which decays to prior of 0.5 (Eq. 2); where τ is the persistence factor multiplied by the annotation confidence and consensus, while $H(t)$ is the cumulative hazard due to brightness instabilities. By decaying toward 0.5, rather than toward zero, we convey epistemic uncertainty—a stale label means unknown state, not negative class. Every frame is assigned a weight, which is calculated as confidence $2|\hat{l} - 0.5|$. More detailed description of the procedure follows in the supplementary material.

$$\hat{l}(t) = 0.5 + (l_a - 0.5)\,\exp\left(-\frac{t - t_a}{\tau} - \lambda H(t)\right) \tag{2}$$

For real-time application, the deployed model freezes SurgeNetXL and calculates 1024-d features in every frame once, followed by the operation of the small causal bridge—a transformer which attends only to the past ring buffer of features and predicts the newest frame. To see whether the bridge could be improved further, we create two models which should surpass it—an anchor-conditioned causal memory and the proof-faithful propagation, implementing the theory exactly with asymmetric persistence and cumulative hazard.

4 Results and Discussion

4.1 A strong CVS prior exists off-the-shelf (RQ1)

Backbones differ widely (Table 1a): SurgeNetXL reaches mAP 0.534, nearly double the best general-domain model (C-RADIO-v4, 0.283) and far above EVA-02 (0.156), which we attribute to in-domain self-supervision on ∼4.7M surgical frames. SurgCLIP (0.171) is zero-shot and not directly comparable. C1 is consistently the hardest criterion across models, which we read as a representational limit rather than a missing training signal. We adopt SurgeNetXL.

4.2 The backbone already encodes the geometry CVS needs (RQ2)

Explicit geometric priors don't improve performance meaningfully. The Siamese depth-fusion model matches the RGB-only baseline.Three findings explain why: first, the RGB-depth Branchwise Centered Kernel Alignment (CKA) is 0.76, suggesting strong correlation between the two branches. Second, dense spatial

Code: `https://github.com/S-Chaudhuri/Critical-View-of-Safety`

probe allows to reconstruct DepthAnythingV2 depth maps from the frozen DINOv3 spatial tokens (Figure 1), meaning that geometry is already encoded in the RGB features. Third, [CLS] token attends to the 3D anatomy without depth supervision. Consistently, adding depth and SAM 2 yields only a 1.2 pp mAP gain, almost all on C3, with SAM 2 negligible; learned fusion underperforms simple concatenation, likely for a lack of data to converge the fusion weights. Hence, in-domain self-supervision, the geometry of the operative field is already encoded.

In order to determine if the model only embeds 3D knowledge or uses it in some way, we used null-space concept erasure. Deletion of the concept of depth caused an insignificant drop in performance overall (ΔmAP $= -0.0022$), suggesting that multi-modal explicit architectures are unnecessary. On the other hand, the criterion-level effects are consistent with clinical intuition: deletion of depth hurts C3 (gallbladder separation) more than any other criterion ($\Delta \mathrm{AP}_{\mathrm{C3}} = -0.0137$), while C1 (tubular structures) actually gets better ($+0.0129$). Because C1 is a 2D problem semantically speaking, deletion of the 3D variance helps to denoise the representation of this criterion, which depends geometrically on 3D priors.

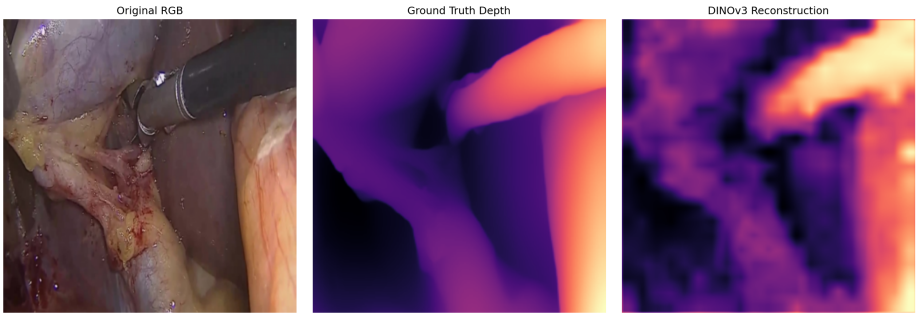


Fig. 1: Reconstruction of the depth map from the frozen DINOv3 spatial tokens is close to the DepthAnythingV2 ground truth, which means that the backbone encodes three-dimensional geometry without explicit depth supervision.

4.3 Light adaptation beats full fine-tuning (RQ3)

Consistent with the geometry experiment results, light adaptation is shown to perform better. LoRA adapter, which trains only 0.36% of the weights, increases frozen MLP baseline of SurgeNetXL from mAP 0.41 to 0.48, while full fine-tuning achieves only 0.43 mAP and tends to overfit to the 700 videos in the training set. The overall message here is that on a small dataset, adding additional modalities and unfreezing the backbone could actually hurt the performance; light adaptation yields the best results.

Table 1: Frozen linear probes (5-fold CV; anchor mAP). (a) SurgeNetXL (DINOv3 ViT-L) nearly doubles the best general backbone; †SurgCLIP is zero-shot, not directly comparable. (b) Adding depth/SAM 2 gives little uplift and gating underperforms concatenation.

(a) Backbone benchmark.

Model	C1	C2	C3	mAP
sSurgeNetXL	**0.413**	**0.635**	**0.554**	**0.534**
C-RADIO-v4	0.216	0.354	0.278	0.283
UniSurg	0.158	0.264	0.197	0.206
SurgCLIP†	0.124	0.240	0.149	0.171
EVA-02	0.112	0.243	0.112	0.156

(b) Multi-stream fusion.

Streams	mAP	C3
SurgeNetXL	0.386	0.316
+ Depth	0.395	0.374
+ Depth + SAM 2	**0.398**	**0.374**
+ Depth + SAM 2 (gated)	0.366	0.338

4.4 CVS temporal evidence is short-range (RQ4)

Screening 3 sparse designs against the SurgeNetXL linear-probe baseline (Acc/F1/mAP 0.848/0.506/0.534), only temporal-local avoids an F1 regression (0.862/0.561/0.576), vs hybrid-sparse (0.847/0.464/0.459) and ViT-learnable (0.812/0.418/0.392). The sweep of its causal window shows that CVS temporal evidence is short-range (maximal accuracy with a 2-3-frame window, reaching mAP = 0.576; details are available in the supplementary material) and decreases with larger windows. Following the criteria, temporal context helps C3 the most, due to progressive separation of the gallbladder, and has almost no effect on C1, implying C1's problem is spatial rather than temporal.

4.5 A real-time causal monitor and its ceiling (RQ5)

Integrating these findings yields a real-time monitor. With the frozen backbone, per-frame features are computed in 13.5 ms latency, and causal attention operates over a short window, so the latency does not depend on clip length. Frozen causal bridge achieves mAP 0.49 ± 0.06, outperforming frozen probe (0.41) and fine-tuned causal variant (0.40); as in the adaptation experiment (RQ3), unfreezing and fine-tuning the backbone for this causal task degrades the self-supervised representation, whereas fine-tuning benefited only the offline sparse model (Table 2a). A distilled student achieves mAP 0.44 with 4.3× lower latency. As a reference, upper bounds of mAP 0.534 (linear probe) and 0.576 (sparse attention) are provided under a different protocol and are not directly challenged. For deployment purposes, we demonstrate the performance of the CVS monitor at the high-specificity operating point, justified by the safety-critical aspect of the task. With approximately 2% prevalence, the positive predictive value is inherently low (at most 0.19), while negative predictive value is close to 0.99, and the bridge demonstrates the best discrimination (AUROC 0.85). Hence, the monitor serves as rule-out instrument—a reliable signal that dissection should continue rather than a green light for clipping.

Table 2: Real-time monitor (single split; anchor mAP, 3 seeds where shown, best seed 0.581). (a) Frozen streaming bridge is the deployable system and beats the fine-tuned variant. (b) Methods built to exceed do not, evidence for label-information ceiling.

(a) Real-time models.

Model	**mAP**	**Streaming**
Frozen probe	0.41	per-frame
Fine-tuned causal	0.40	yes
Distilled student	0.44	yes (4.3×)
Frozen causal bridge	**0.49**	yes

(b) Pushing past the bridge.

Model	**mAP**
Frozen causal bridge	**0.49±0.06**
Anchor-cond. memory	0.46±0.03
Proof-faithful prop.	0.46±0.03

Finally, we examine whether the causal bridge can be surpassed. We develop two models to surpass the causal bridge—an anchor-conditioned causal memory and proof-faithful propagation, but both of them achieve mAP 0.46 ± 0.03, still underperforming the bridge (Table 2b). This result can be interpreted as label information ceiling: pseudo-labels, generated from 18 anchors per video, cannot contain more information, and re-labeling with the same frozen backbone adds no additional information. Thus, the constraint is imposed by label information, not model capacity. The residual weakness on C1 is representational a single pooled feature cannot assert that *exactly two* tubular structures are present, which is itself safety-relevant, since C1 concerns the tubular-structure identification at the heart of the injury mechanism. The levers are spatial patch-token heads for C1 and temporal self-supervision on unlabeled frames, not a larger temporal model.

5 Conclusion

We examined how much of a CVS safety monitor a single scaled, frozen surgical foundation model already provides. Benchmarking five foundation models, we found that SurgeNetXL nearly doubles the best general-domain backbone (RQ1) and already encodes the 3D geometry a depth prior would add, using it selectively for the geometric criteria (RQ2); light adaptation beats full fine-tuning (RQ3), CVS temporal evidence is short-range (RQ4), and a confidence-aware causal propagation with a frozen streaming bridge gives an online monitor at mAP 0.49±0.06 and NPV≈99%, bounded by the 18-anchor label budget rather than model capacity (RQ5). These favour frozen representations over multi-modal fusion, light adaptation over unfreezing, and short causal windows over long context. Read as a safety instrument, our findings say as much about what is unreliable as what works: high NPV makes the bridge a dependable rule-out, but low precision at ∼2% prevalence means a positive reading cannot confirm CVS, and the residual C1 weakness sits on the criterion most tied to the injury mechanism. Such a monitor could support, not replace, a surgeon's verification;

establishing safety would need prospective, human-in-the-loop validation rather than the retrospective, anchor-only evaluation used here. Spatial patch-token heads for C1 and temporal self-supervision on unlabeled video are the most promising next directions.

Acknowledgments. We thank Luuk Jonker for their insights and feedback in an earlier draft and Junis Nagel for exploratory experiments. We also appreciate Tim Feenstra, Riëtte Meijer and Esther Barsom for their invaluable clinical insights.

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

References

1. Strasberg, S.M., Hertl, M., Soper, N.J.: An analysis of the problem of biliary injury during laparoscopic cholecystectomy. J. Am. Coll. Surg. **180**(1), 101-125 (1995)
2. Way, L.W., et al.: Causes and prevention of laparoscopic bile duct injuries. Ann. Surg. **237**(4), 460–469 (2003)
3. Buddingh, K.T., et al.: Intraoperative assessment of biliary anatomy for prevention of bile duct injury. Surg. Endosc. **25**(8), 2449-2461 (2011)
4. de'Angelis, N., et al.: 2020 WSES guidelines for detection and management of bile duct injury during cholecystectomy. World J. Emerg. Surg. **16**(1), 30 (2021)
5. Gawlik, C., Carneval, M.: A review of the management of bile leaks. Cureus **13**(5) (2021)
6. Booij, K.A., et al.: Long-term impact of bile duct injury on morbidity, mortality, quality of life, and work related limitations. Ann. Surg. **268**(1), 143-150 (2018)
7. Korndorffer Jr, J.R., et al.: Situating artificial intelligence in surgery. Ann. Surg. **272**(3), 523-528 (2020)
8. Alapatt, D., et al.: The SAGES critical view of safety challenge: a global benchmark for AI-assisted surgical quality assessment. arXiv:2509.17100 (2025)
9. Baby, B., et al.: Multi-modal representations for fine-grained multi-label critical view of safety recognition. In: MICCAI, pp. 423-432. Springer (2025)
10. Jaspers, T.J., et al.: Scaling up self-supervised learning for improved surgical foundation models. Med. Image Anal. **108**, 103873 (2026)
11. Ramesh, S., et al.: Dissecting self-supervised learning methods for surgical computer vision. arXiv:2207.00449 (2023)
12. Radford, A., et al.: Learning transferable visual models from natural language supervision. In: ICML, pp. 8748-8763 (2021)
13. Zhang, S., et al.: BiomedCLIP: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. arXiv:2303.00915 (2023)
14. Fang, Y., et al.: EVA-02: a visual representation for neon genesis. Image Vis. Comput. **149**, 105171 (2024)
15. Ranzinger, M., et al.: C-RADIO-v4 (tech report). arXiv:2601.17237 (2026)
16. Perez, A., et al.: SurgLAVI: large-scale hierarchical dataset for surgical vision-language representation learning. arXiv:2509.10555 (2026)
17. Wu, J., et al.: UniSurg: a video-native foundation model for universal understanding of surgical videos. arXiv:2602.05638 (2026)
18. Zeng, Z., et al.: SurgVLM: a large vision-language model and systematic evaluation benchmark for surgical intelligence. arXiv:2506.02555 (2025)

19. Yang, L., et al.: Depth Anything V2. arXiv:2406.09414 (2024)
20. Ravi, N., et al.: SAM 2: segment anything in images and videos. In: ICLR (2025)
21. Li, Y., et al.: SViTT: temporal learning of sparse video-text transformers. arXiv:2304.08809 (2023)
22. Mascagni, P., et al.: Artificial intelligence for surgical safety: automatic assessment of the critical view of safety. Ann. Surg. **275**(5), 955-961 (2022)