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.

Annotation-Efficient Critical View of Safety Assessment with Vision Foundation Models

Sergio Cañar[0009−0000−5818−7259](✉), Javier Santiago Vera Rincón[0009−0007−8949−1270], Isabel Sofía Tovar Sánchez[0009−0008−8578−1972], and Pablo Arbeláez[0000−0001−5244−2407]

Universidad de los Andes, Bogotá, Colombia
{s.canar, j.verar, i.tovars, pa.arbelaez}@uniandes.edu.co

Abstract. Automated Critical View of Safety (CVS) assessment during laparoscopic cholecystectomy could reduce bile duct injuries, but existing methods require additional supervision (segmentation masks, bounding boxes, or language prompts), increasing annotation effort and limiting large-scale deployment. We ask whether a general-domain vision foundation model (FM) can remove this dependence on extra supervision. To answer this, we benchmark five pretrained encoders under identical conditions on two CVS benchmarks, Endoscapes 2023 and SAGES 2024, and find EVA-02 L the strongest. Fine-tuned per frame without additional supervision, this FM alone matches or surpasses prior methods. On top of the frozen FM, **PercEVA-CVS** adds a keyframe-anchored Perceiver aggregator that contributes a smaller, dataset-dependent gain and lets us characterise the temporal context CVS assessment requires. A sufficiently strong FM removes the need for additional supervision, and the frozen PercEVA-CVS pipeline builds on this to set a new state of the art on both benchmarks. Results are reported as mean ± std over three seeds.

Keywords: Critical View of Safety · Laparoscopic cholecystectomy · Foundation models · Surgical video analysis.

1 Introduction

Bile duct injury is one of the most serious complications of laparoscopic cholecystectomy, often arising from the misidentification of the biliary anatomy at the time of clipping the cystic duct [26]. The Critical View of Safety (CVS) protocol was introduced to prevent these injuries, requiring the surgeon to confirm three criteria before clipping: two tubular structures connected to the gallbladder (C1), a cleared hepatocystic triangle (C2), and an exposed cystic plate (C3) [24, 4]. Yet despite broad endorsement, surgeons show low inter-rater agreement when assessing the criteria intraoperatively [13]. This variability motivates automated CVS assessment, cast as multi-label binary classification of whether each criterion is satisfied at a given moment, as a form of intraoperative decision support [12, 15].

Most prior methods rely on some form of external supervision. DeepCVS [16] concatenates frames with segmentation masks as auxiliary input. More recently, graph-based methods use a pretrained detector to localize anatomical structures and build a latent representation over them: LG-CVS [17] does so per frame and SV2LSTG [18] extends it to the spatiotemporal domain. These methods depend on the quality of the detected structures, making them sensitive to detection errors. Language-supervised approaches instead condition on text: KnoCo-CVS [9] uses guideline prompts to modulate attention, while CVS-AdaptNet [3] aligns a surgical vision-language model via criterion text. Both still depend on textual descriptions of the criteria, reintroducing the same expert interpretation behind the low inter-rater agreement. SwinCVS [19] completely forgoes external supervision, combining a SwinV2 encoder [11] with an LSTM. Recurrent aggregators of this kind are, in principle, susceptible to a fixed-size state bottleneck over longer temporal contexts [25, 10], which motivates exploring a non-recurrent alternative. Foundation models (FMs), large-scale encoders pretrained on broad visual data [2], transfer well to surgical tasks such as phase recognition [27] and depth estimation [5]. Yet even a surgery-specific self-supervised FM still falls short of specialized CVS architectures [8], leaving open whether a general-domain FM can close this gap without any additional supervision.

We answer this affirmatively and find that the FM alone is decisive, benchmarking five general-domain encoders spanning four pretraining paradigms under identical conditions to isolate the encoder's contribution. EVA-02 L [6] is consistently the strongest; fine-tuned per frame without additional supervision, it already matches or surpasses prior methods on Endoscapes 2023 [14] and SAGES 2024 [1]. We then complete the pipeline with a keyframe-anchored Perceiver aggregator in which learnable latents seeded by the current frame attend in parallel over a short causal window of embeddings, avoiding the fixed-state bottleneck of recurrent alternatives while keeping the temporal context explicit and tunable. The resulting two-stage pipeline, ***PercEVA-CVS*** (Fig. 1), improves over the per-frame FM and lets us characterise how much temporal context CVS assessment requires.

In summary, this work makes the following contributions:

- **A state of the art without additional supervision, driven by the FM.** A general-domain FM (EVA-02 L) matches or surpasses prior supervised methods in mean Average Precision (mAP) on both Endoscapes 2023 and SAGES 2024. PercEVA-CVS adds a Perceiver temporal stage on top, whose contribution we report separately from the FM's.
- **A reproducible encoder benchmark for CVS.** Under a single controlled protocol, we identify EVA-02 L as the strongest encoder for CVS assessment and provide the first systematic evaluation of prior methods on SAGES 2024, a recent benchmark absent from prior CVS work.
- **An analysis of temporal context for CVS.** We quantify how aggregation strategy, context length, and readout affect performance, which saturates at a short causal context and a simple keyframe readout.

Our code is available at https://github.com/BCV-Uniandes/PercEVA-CVS.

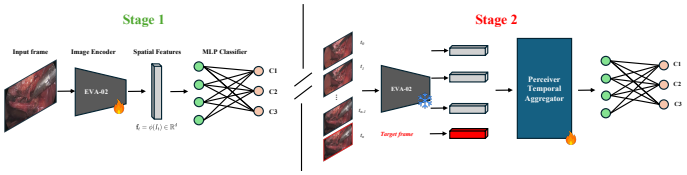


Fig. 1. Overview of the PercEVA-CVS pipeline. Stage 1 fine-tunes the FM (EVA-02 L) frame-by-frame to produce per-frame embeddings. Stage 2 freezes the FM and trains a lightweight Perceiver-based temporal aggregator that reads a short causal window of embeddings via keyframe-anchored latent attention, producing per-criterion CVS predictions.

2 Method

2.1 Overview

PercEVA-CVS divides CVS assessment into two stages (Fig. 1): an FM in Stage 1 and a keyframe-anchored Perceiver aggregator in Stage 2. In Stage 1 (Sect. 2.2), we fine-tune EVA-02 L [6] to produce discriminative per-frame embeddings for CVS criterion scoring. In Stage 2 (Sect. 2.3), the FM is frozen and the Perceiver aggregator is trained on top, reading a short causal window of FM embeddings to issue per-criterion predictions. This factorisation decouples encoder selection from temporal modelling, letting each stage be optimised independently.

2.2 Stage 1: Vision Foundation Model

Based on the encoder benchmark (Sect. 3.2), we adopt EVA-02 L [6] as our FM ϕ, fine-tuned end-to-end (E2E) for multi-label binary classification of the three CVS criteria with a per-criterion weighted binary cross-entropy, where w_c addresses class imbalance for each criterion $c \in \{c_1, c_2, c_3\}$. For each frame I_t, the FM produces an embedding $\mathbf{f}_t = \phi(I_t) \in \mathbb{R}^d$ ($d = 1024$) from its pre-logit layer which serves as the input to Stage 2.

2.3 Stage 2: Temporal Aggregator

Architecture. Inspired by the Perceiver [7], the temporal aggregator (Fig. 2) compresses temporal context into K latent vectors via parallel cross-attention over all T frames. Frames are sampled online at 1 fps and arranged causally, oldest to most recent, into a window $\mathbf{X} = [\mathbf{f}_1, \ldots, \mathbf{f}_T] \in \mathbb{R}^{T \times d}$, with $\mathbf{f}_T$ the *keyframe* (the frame under evaluation and the most recent in the window). Each embedding is linearly projected to model dimension D, layer-normalised, and augmented with learned positional encodings. The projected keyframe $\mathbf{k} \in \mathbb{R}^D$ is prepended to K learnable latents $\mathbf{Q} \in \mathbb{R}^{K \times D}$ to form the initial query sequence $\mathbf{L}_0 = [\mathbf{k}; \mathbf{Q}] \in \mathbb{R}^{(K+1) \times D}$. Anchoring the latents to the keyframe integrates temporal context relative to the frame under evaluation; since $\mathbf{f}_T$ also remains in $\mathbf{X}$, the keyframe token attends to itself alongside the rest of the window during cross-attention.

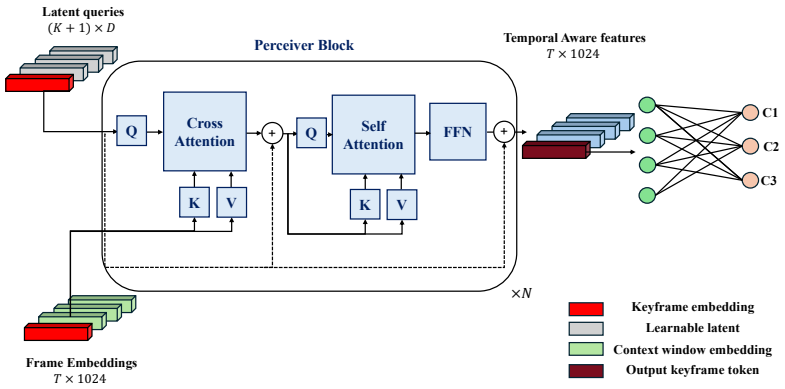


Fig. 2. Stage 2 temporal aggregator. The keyframe embedding (red) is prepended to K learnable latents (grey) and refined by N Perceiver blocks, each cross-attending over the T context-window embeddings (green) followed by a latent self-attention + FFN. The refined keyframe token (dark red) is read out to a two-layer MLP producing per-criterion logits $[\hat{c}_1, \hat{c}_2, \hat{c}_3]$.

Gated-residual Perceiver blocks. $\mathbf{L}_0$ is refined by a stack of N identical blocks (Fig. 2), each revising the latent array in three steps. First, cross-attention lets the queries read the temporal window $\mathbf{X}$, drawing context from across the T frames into the latents. Then, self-attention mixes this information among the $K+1$ tokens and a position-wise feed-forward network (FFN) refines each token independently. Rather than a standard residual, every sublayer uses a learned gated residual inspired by Highway Networks [23],

$$\mathbf{L} \leftarrow \sigma\big(W_g\,\mathrm{LN}(\mathbf{L})\big) \odot s\big(\mathrm{LN}(\mathbf{L})\big) + \mathbf{L}, \tag{1}$$

where LN is layer normalisation and $s(\cdot)$ is the sublayer: cross-attention (with $\mathbf{X}$ as keys and values), self-attention, or the FFN. $W_g \in \mathbb{R}^{D\times D}$ is a learned projection, σ the sigmoid, and $\odot$ the element-wise product. The gate acts per token and per dimension, allowing each of the $K+1$ tokens to control how much sublayer output it admits, while its previous state always flows through untouched. Gate biases are initialised to zero, so all gates start half-open. After N blocks, the refined keyframe token, the slot seeded by $\mathbf{k}$, is read out as $\mathbf{z} = \mathrm{LN}(\mathbf{L}_{\mathrm{kf}}^{(N)}) \in \mathbb{R}^D$. Reading out the keyframe token rather than pooling all K latents preserves the frame-level grounding established at the input. A two-layer MLP with hidden dimension 128 and GELU activations then maps $\mathbf{z}$ to per-criterion logits $[\hat{c}_1, \hat{c}_2, \hat{c}_3] \in \mathbb{R}^3$.

E2E variant. The two-stage protocol freezes the Stage 1 FM during Stage 2 training. As an alternative, we explore joint optimisation: we load the Stage 1 weights and unfreeze the last four transformer blocks of EVA-02 L, targeting the highest-level task-relevant layers while keeping lower-level features fixed to

limit overfitting and compute. The aggregator is randomly initialised and trained jointly with these blocks (Sect. 2.4). This trades the modularity and efficiency of the frozen regime for a potentially richer gradient signal at substantially higher compute.

2.4 Implementation Details

Frames are resized to 448×448 and normalised with ImageNet statistics, without augmentation, and sampled at 1 fps. Per-criterion soft labels in $[0, 1]$ are used directly as training targets and evaluated against hard labels, with a criterion considered positive when marked satisfied by the majority of annotators. Positive weights w_c are computed as the inverse positive frequency. Stage 1 is initialized from the public ImageNet-1k fine-tuned EVA-02 L checkpoint [6] and fine-tuned independently on each dataset (no weight sharing). Stage 2 uses $N = 2$ blocks, model dimension $D = 1024$, 16 attention heads, FFN width 512, dropout 0.01, $K = 64$ latent tokens, and a $T = 15$ causal window, trained with batch size 64; shorter windows are zero-padded and masked during cross-attention. Both stages use AdamW, StepLR, and early stopping on validation mAP (patience 5, up to 15 epochs). Learning rate, weight decay, step size, and decay factor are selected by grid search on the validation set, with final settings released alongside the code. Experiments are conducted on an NVIDIA Quadro RTX 8000 GPU.

Inference latency. We measure per-keyframe latency at batch size 1 in fp16. The frozen pipeline runs in 37.5 ms (34.4 ms for the EVA-02 L encoder and ∼3 ms for the Perceiver aggregator), encoding each frame once as it arrives, whereas the jointly-trained E2E variant re-encodes its context window at every step and takes ∼400 ms. Both fall within the 1 s interval of the 1 fps assessment rate. The two configurations trade off accuracy against deployment form: the frozen pipeline is more accurate but separates encoding from a later temporal read-out over stored embeddings, which suits offline evaluation more than intraoperative use, while the E2E variant has the single-pass form better suited to the operating room but is an order of magnitude slower. We report this comparison to characterise the trade-off rather than to claim real-time deployment.

3 Experiments

3.1 Datasets

Datasets. We evaluate on Endoscapes 2023 [14] and SAGES 2024 [1], to our knowledge the only public CVS benchmarks with frame-level criterion labels. Endoscapes 2023 comprises 201 laparoscopic cholecystectomy recordings totalling 11,090 frames, each labelled for the three CVS criteria and annotated with segmentation masks and anatomical bounding boxes. We use the official train/validation/test splits. SAGES 2024 is a multi-centre dataset of 1,000 clips (90 s each) with 18 labelled frames per clip, without segmentation masks or bounding boxes. We follow the official split of 700 training and 300 test clips, holding

out 200 training clips for validation. Both datasets are publicly released under their respective terms of use, with ethical approval obtained by the original authors [14, 1].

Evaluation metrics. Performance is evaluated using mAP over the three CVS criteria, the most robust metric for this multi-label, class-imbalanced setting and one that ranks methods independently of any decision threshold. Improvements are reported as absolute percentage-point (pp) differences. Statistical significance is assessed with two-sided paired t-tests over three random seeds against the reference method indicated in each table. We also report the total parameter count (M) as a measure of model size.

3.2 Encoder Benchmark

We benchmark five general-domain image encoders under identical conditions that isolate the encoder's contribution to CVS assessment. These span four pretraining paradigms: supervised (Swin-L [11]), language-contrastive (ViT-L/CLIP [21]), self-distilled (DINOv2 [20], DINOv3 [22]), and masked image modelling with language alignment (EVA-02 L [6]). All encoders are fine-tuned E2E on CVS data, with each output projected to $d = 1024$ and classified by a shared two-layer MLP head with per-criterion positive weights addressing class imbalance; hyperparameters are grid-searched independently per encoder and dataset, selecting the best configuration on the validation split.

Table 1. Encoder comparison on Endoscapes 2023 and SAGES 2024 test sets. Best results are in bold, second best underlined.

Encoder	Params (M)	C1	C2	C3	mAP
Endoscapes 2023					
CLIP [21]	304.0	53.16 ± 2.8	50.66 ± 3.1	68.78 ± 3.9	57.53 ± 3.2
Swin-L [11]	195.2	60.95 ± 4.5	$\underline{56.75 \pm 1.4}$	66.82 ± 6.4	61.50 ± 3.1
DINOv2 [20]	304.0	64.53 ± 0.9	50.52 ± 5.3	$\underline{70.64 \pm 1.2}$	61.90 ± 1.6
DINOv3 [22]	304.0	$\underline{67.10 \pm 1.0}$	52.20 ± 1.7	67.50 ± 1.9	$\underline{62.30 \pm 0.7}$
EVA-02 L [6]	304.1	$\mathbf{72.45 \pm 1.3}$	$\mathbf{59.93 \pm 0.6}$	$\mathbf{70.67 \pm 1.3}$	$\mathbf{67.68 \pm 0.3}$
SAGES 2024					
CLIP [21]	304.0	43.79 ± 1.4	75.21 ± 0.0	48.87 ± 1.3	55.96 ± 0.9
Swin-L [11]	195.2	48.20 ± 0.0	75.07 ± 0.7	49.55 ± 0.8	57.61 ± 0.5
DINOv2 [20]	304.0	48.65 ± 0.9	77.34 ± 0.5	52.26 ± 0.7	59.42 ± 0.2
DINOv3 [22]	304.0	$\underline{50.30 \pm 1.1}$	$\underline{78.90 \pm 0.7}$	$\underline{53.40 \pm 1.7}$	$\underline{60.90 \pm 0.4}$
EVA-02 L [6]	304.1	$\mathbf{52.96 \pm 1.5}$	$\mathbf{80.57 \pm 0.9}$	$\mathbf{56.41 \pm 1.4}$	$\mathbf{63.32 \pm 1.1}$

Table 1 reports test-set results on both benchmarks. EVA-02 L leads on mAP on both: 67.68 on Endoscapes 2023 (+5.38 pp over DINOv3) and 63.32 on SAGES 2024 (+2.42 pp). It is the only encoder ranking first on mAP across both benchmarks, indicating more consistent transfer to CVS assessment; these results motivate its use in all subsequent experiments.

3.3 Comparison with State of the Art (SOTA)

Table 2 compares PercEVA-CVS against publicly available CVS methods: LG-CVS [17], SV2LSTG [18], CVS-AdaptNet [3], and SwinCVS [19]; all pipelines are re-trained and re-evaluated. On Endoscapes 2023, EVA-02 L fine-tuned at frame level matches or exceeds prior methods despite using no additional supervision, suggesting that a sufficiently strong vision FM can learn discriminative representations for CVS assessment directly from images. The frozen two-stage model achieves the highest mAP among non-E2E methods, improving over SwinCVS Frozen and, significantly, over the per-frame FM alone (+1.73 pp). Temporal aggregation improves all three criteria over the per-frame FM, most on C1 and C3 and more modestly on C2. The one clear gap remains C2, where the frozen model still trails SwinCVS Frozen. E2E variant closes this gap and achieves the best overall mAP, at the cost of higher training compute.

Table 2. Comparison with SOTA on Endoscapes 2023 and SAGES 2024 test sets. *Sup.*: anatomy supervision (*segm*/*box*/*text*/-). p: two-sided paired t-test ($n = 3$) of each method versus PercEVA-CVS (—: reference). ‡ On SAGES, graph based methods use detectors pre-trained on Endoscapes. Best results are in bold, second best underlined.

Method	Sup.	Params (M)	C1	C2	C3	mAP	p
Endoscapes 2023							
CVS-AdaptNet [3]	*text*	133.4	52.9 ± 2.5	50.0 ± 1.0	56.8 ± 1.0	53.2 ± 1.5	0.005
LG-CVS‡ [17]	*box*	112.1	67.2 ± 0.6	56.8 ± 1.8	70.2 ± 2.0	64.7 ± 1.2	0.039
LG-CVS‡ [17]	*segm*	114.8	66.9 ± 2.2	53.6 ± 2.1	67.7 ± 2.2	62.8 ± 1.8	0.043
SV2LSTG‡ [18]	*box*	277.5	69.5 ± 1.6	58.6 ± 1.9	69.4 ± 1.2	65.8 ± 1.5	0.071
SV2LSTG‡ [18]	*segm*	280.1	67.3 ± 0.8	55.1 ± 3.2	67.3 ± 1.3	63.3 ± 1.2	0.025
SwinCVS Frozen [19]	-	88.73	66.6 ± 1.0	$\underline{64.5 \pm 0.8}$	65.5 ± 1.8	65.5 ± 1.0	0.035
SwinCVS E2E [19]	-	88.73	63.8 ± 1.6	61.1 ± 0.4	65.8 ± 2.1	63.6 ± 1.3	0.028
EVA-02 L (frame)	-	304.1	$\underline{72.5 \pm 1.3}$	59.9 ± 0.6	70.7 ± 1.3	67.7 ± 0.3	0.019
PercEVA-CVS	-	331.6	$\mathbf{74.8 \pm 0.5}$	61.0 ± 0.2	$\underline{72.5 \pm 0.3}$	$\underline{69.4 \pm 0.2}$	—
PercEVA-CVS E2E	-	331.6	71.0 ± 2.8	$\mathbf{65.6 \pm 4.1}$	$\mathbf{74.8 \pm 3.5}$	$\mathbf{70.5 \pm 3.1}$	
SAGES 2024							
CVS-AdaptNet [3]	*text*	133.4	35.2 ± 1.3	59.1 ± 1.1	29.6 ± 0.2	41.3 ± 0.7	0.003
LG-CVS‡ [17]	*box*	112.1	47.3 ± 0.6	73.7 ± 0.4	51.9 ± 0.8	57.6 ± 0.4	0.008
LG-CVS‡ [17]	*segm*	114.8	47.1 ± 0.3	72.7 ± 0.3	50.3 ± 1.5	56.7 ± 0.6	0.024
SV2LSTG‡ [18]	*box*	277.5	48.3 ± 2.4	73.7 ± 1.2	56.1 ± 1.2	59.4 ± 1.2	0.008
SV2LSTG‡ [18]	*segm*	280.1	46.9 ± 1.9	72.4 ± 0.5	50.1 ± 2.7	56.5 ± 0.6	0.023
SwinCVS Frozen [19]	-	88.7	40.1 ± 0.2	70.0 ± 0.3	44.7 ± 2.2	51.6 ± 0.8	0.001
SwinCVS E2E [19]	-	88.7	38.4 ± 1.1	64.8 ± 2.1	38.9 ± 2.6	47.4 ± 1.5	0.012
EVA-02 L (frame)	-	304.1	53.0 ± 1.5	80.6 ± 0.9	56.4 ± 1.4	63.3 ± 1.1	0.091
PercEVA-CVS	-	331.6	$\mathbf{54.1 \pm 1.1}$	$\underline{80.9 \pm 1.0}$	$\mathbf{60.4 \pm 2.1}$	$\mathbf{65.1 \pm 1.3}$	—
PercEVA-CVS E2E	-	331.6	$\underline{53.8 \pm 1.1}$	$\mathbf{81.0 \pm 0.9}$	$\underline{59.4 \pm 0.9}$	$\underline{64.7 \pm 0.9}$	

The advantage is larger on SAGES 2024, where frozen PercEVA-CVS outperforms SwinCVS Frozen by 13.53 pp in mAP. Since both methods use the same two-stage frozen-encoder pipeline, this improvement is primarily attributable to the stronger representations learned by EVA-02 L. The gap is even larger relative to anatomy-supervised baselines because SAGES 2024 provides no anatomy annotations, forcing those methods to rely on detectors pretrained on Endoscapes

that degrade under the cross-centre distribution shift. In contrast, our approach avoids this dependence on anatomy supervision. Qualitatively, Fig. 3 shows that PercEVA-CVS localises the anatomy relevant to each CVS criterion more precisely than SwinCVS. This is most evident for C1, where our model sharply highlights the two tubular structures connected to the gallbladder, whereas SwinCVS produces diffuse, off-target activations.

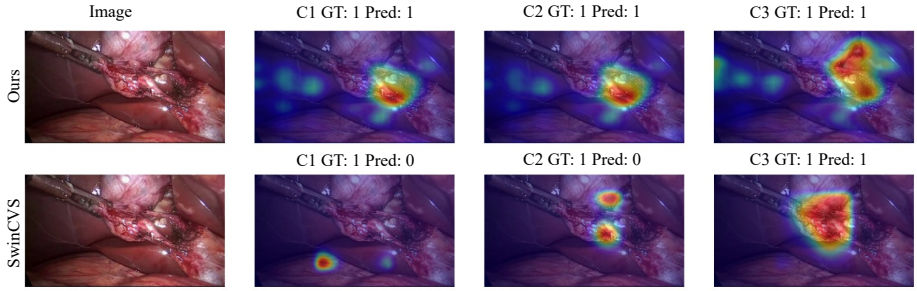


Fig. 3. Grad-CAM comparison between PercEVA-CVS and SwinCVS, computed on each method's encoder (EVA-02 L and SwinV2, respectively)

3.4 Ablations

All ablations use the 2 stage variant on SAGES 2024. **(a)** The keyframe token significantly outperforms both mean pooling (+2.01 pp) and a CLS token (+4.81 pp), showing that anchoring the latent representation to the evaluated frame is more effective than global pooling. Highway gating provides a further significant improvement of +0.67 pp. **(b)** The Perceiver and LSTM achieve statistically indistinguishable performance, indicating that both are effective temporal aggregators for this task. We nevertheless adopt the Perceiver for its parallel, keyframe-anchored formulation over a fixed temporal window (Sect. 2.3). Increasing the number of Perceiver blocks does not significantly affect performance; we therefore use N=2, which achieves the highest mean mAP. (c) A 15 s temporal window achieves the highest mean mAP. However, differences across windows are not statistically significant, including the single-frame setting, indicating that CVS assessment requires only short-term temporal context.

4 Conclusion

We presented PercEVA-CVS, a two-stage pipeline for CVS assessment. A controlled benchmark identifies EVA-02 L as the strongest vision FM, and a lightweight keyframe-anchored Perceiver provides a modest additional gain by aggregating short-term temporal context. The resulting pipeline surpasses prior

Table 3. Ablation studies on SAGES 2024. Adopted configuration in bold. p: two-sided paired t-test (n=3 seeds) versus the adopted configuration (—: reference).

(a) Readout			(b) Aggregator			(c) Window		
Condition	mAP	p	Condition	mAP	p	T	mAP	p
Readout			*Aggregator*					
Keyframe	$\mathbf{65.12 \pm 1.32}$	—	LSTM	64.88 ± 0.95	0.720	1	63.98 ± 0.95	0.151
Mean	63.11 ± 1.29	<0.001	**Perceiver**	$\mathbf{65.12 \pm 1.32}$	—	3	65.11 ± 0.95	0.976
CLS	60.31 ± 2.08	0.038	*Blocks*			5	64.95 ± 1.29	0.498
Highway gate			$N = 1$	64.84 ± 0.85	0.645	10	65.04 ± 1.32	0.551
On	$\mathbf{65.12 \pm 1.32}$	—	$\boldsymbol{N = 2}$	$\mathbf{65.12 \pm 1.32}$	—	**15**	$\mathbf{65.12 \pm 1.32}$	—
Off	64.45 ± 1.15	0.042	$N = 3$	64.87 ± 1.10	0.532	20	65.09 ± 1.31	0.109

supervised methods in mAP on both Endoscapes 2023 and SAGES 2024. Our analysis shows that nearly all performance gains arise from the FM, suggesting that a sufficiently strong FM can compensate for the absence of additional anatomical supervision. CVS assessment further requires little temporal context, remaining robust even from a single frame. Future work will explore video foundation models, surgical-domain pre-training, uncertainty-aware predictions, and single-pass architectures suited to the latency constraints of intraoperative deployment.

Acknowledgments. This work was supported by Azure sponsorship credits granted by Microsoft's AI for Good Research Lab.

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

References

1. Alapatt, D., Eckhoff, J., Lyu, Z., Ban, Y., Mazellier, J.P., Choksi, S., et al.: The SAGES critical view of safety challenge: A global benchmark for AI-assisted surgical quality assessment. arXiv preprint arXiv:2509.17100 (2026)
2. Awais, M., Naseer, M., Khan, S., Anwer, R.M., Cholakkal, H., Shah, M., Yang, M.H., et al.: Foundational models defining a new era in vision: A survey and outlook. arXiv preprint arXiv:2307.13721 (2023). https://doi.org/10.48550/arXiv.2307.13721
3. Baby, B., Srivastav, V., Jain, P.P., Yuan, K., Mascagni, P., et al.: Multi-modal representations for fine-grained multi-label critical view of safety recognition. arXiv preprint arXiv:2507.05007 (2025). https://doi.org/10.48550/arXiv.2507.05007
4. Brunt, L.M., Deziel, D.J., Telem, D.A., Strasberg, S.M., Aggarwal, R., Asbun, H., et al.: Safe cholecystectomy multi-society practice guideline and state-of-the-art consensus conference on prevention of bile duct injury during cholecystectomy. Surgical Endoscopy **34**(7), 2827–2855 (2020)
5. Cui, B., Islam, M., Bai, L., Ren, H.: Surgical-dino: Adapter learning of foundation models for depth estimation in endoscopic surgery. International Journal of Computer Assisted Radiology and Surgery (2024)

6. Fang, Y., Sun, Q., Wang, X., Huang, T., Wang, X., Cao, Y.: Eva-02: A visual representation for neon genesis. Image and Vision Computing **149**, 105171 (Sep 2024). https://doi.org/10.1016/j.imavis.2024.105171
7. Jaegle, A., Gimeno, F., Brock, A., Zisserman, A., Vinyals, O., et al.: Perceiver: General perception with iterative attention. In: Proceedings of the 38th International Conference on Machine Learning (ICML). Proceedings of Machine Learning Research, vol. 139 (2021), https://arxiv.org/abs/2103.03206
8. Jaspers, T.J., de Jong, R.L., Li, Y., Kusters, C.H., Bakker, F.H., et al.: Scaling up self-supervised learning for improved surgical foundation models. Medical Image Analysis **108**, 103873 (2026). https://doi.org/10.1016/j.media.2025.103873
9. Li, Y., Yang, K., Li, X., Li, Q., Wang, Z.: Knowledge-contrastive fine-grained attention for cvs recognition in cholecystectomy without dense labels. In: ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 7336–7340. Barcelona, Spain (2026). https://doi.org/10.1109/ICASSP55912.2026.11465051
10. Liu, Y., Boels, M., Garcia-Peraza-Herrera, L.C., Vercauteren, T., Dasgupta, P., Granados, A., Ourselin, S.: Lovit: Long video transformer for surgical phase recognition. Medical Image Analysis **99**, 103366 (2025). https://doi.org/10.1016/j.media.2024.103366
11. Liu, Z., Hu, H., Lin, Y., Yao, Z., Xie, Z., et al.: Swin Transformer V2: Scaling up capacity and resolution. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12009–12019 (2022)
12. Maier-Hein, L., Eisenmann, M., Sarikaya, D., März, K., et al.: Surgical data science – from concepts toward clinical translation. Medical Image Analysis **76**, 102306 (2022). https://doi.org/10.1016/j.media.2021.102306
13. Mascagni, P., Alapatt, D., Lapergola, A., Vardazaryan, A., Mazellier, J.P., et al.: Early-stage clinical evaluation of real-time artificial intelligence assistance for laparoscopic cholecystectomy. British Journal of Surgery **111**(1), znad353 (2024)
14. Mascagni, P., Alapatt, D., Murali, A., Vardazaryan, A., Garcia, A., Okamoto, N., Costamagna, G., et al.: Endoscapes, a critical view of safety and surgical scene segmentation dataset for laparoscopic cholecystectomy. Scientific Data **12**(1) (Feb 2025). https://doi.org/10.1038/s41597-025-04642-4
15. Mascagni, P., Alapatt, D., Sestini, L., Altieri, M.S., Madani, A., et al.: Computer vision in surgery: from potential to clinical value. npj Digital Medicine **5**(1), 163 (2022)
16. Mascagni, P., Vardazaryan, A., Alapatt, D., Urade, T., Emre, T., et al.: Artificial intelligence for surgical safety: automatic assessment of the critical view of safety in laparoscopic cholecystectomy using deep learning. Annals of Surgery **275**(5), 955–961 (2022)
17. Murali, A., Alapatt, D., Mascagni, P., Vardazaryan, A., Garcia, A., Okamoto, N., Mutter, D., et al.: Latent graph representations for critical view of safety assessment. IEEE Transactions on Medical Imaging **43**(3), 1247–1258 (2024). https://doi.org/10.1109/TMI.2023.3333034
18. Murali, A., Alapatt, D., Mascagni, P., Vardazaryan, A., Garcia, A., et al.: Encoding surgical videos as latent spatiotemporal graphs for object and anatomy-driven reasoning. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 647–657. Springer (2023)
19. Nowak, F.M., Duda, A., Błasiak, P., Urbanik, A.: Swincvs: A unified approach to classifying critical view of safety structures in laparoscopic cholecystectomy. International Journal of Computer Assisted Radiology and Surgery **20**(6), 1145–1152 (2025). https://doi.org/10.1007/s11548-025-03354-9

20. Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., Fernandez, P., Haziza, D., Massa, F., El-Nouby, A., et al.: DINOv2: Learning robust visual features without supervision. Transactions on Machine Learning Research (2024), arXiv:2304.07193.
21. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., et al.: Learning transferable visual models from natural language supervision. In: International Conference on Machine Learning (ICML). pp. 8748–8763 (2021)
22. Siméoni, O., Vo, H.V., Seitzer, M., Baldassarre, F., Oquab, M., Jose, C., Khalidov, V., Szafraniec, M., Yi, S., Ramamonjisoa, M., Massa, F., Haziza, D., Wehrstedt, L., Wang, J., Darcet, T., Moutakanni, T., Sentana, L., Roberts, C., Vedaldi, A., Tolan, J., Brandt, J., Couprie, C., Mairal, J., Jégou, H., Labatut, P., Bojanowski, P.: Dinov3. arXiv preprint arXiv:2508.10104 (2025). https://doi.org/10.48550/arXiv.2508.10104
23. Srivastava, R.K., Greff, K., Schmidhuber, J.: Highway networks. In: ICML Deep Learning Workshop (2015)
24. Strasberg, S.M., Hertl, M., Soper, N.J.: An analysis of the problem of biliary injury during laparoscopic cholecystectomy. Journal of the American College of Surgeons **180**(1), 101–125 (1995)
25. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., et al.: Attention is all you need. In: Advances in Neural Information Processing Systems. vol. 30, pp. 5998–6008 (2017)
26. Way, L.W., Stewart, L., Gantert, W., Liu, K., Lee, C.M., Whang, K., et al.: Causes and prevention of laparoscopic bile duct injuries: analysis of 252 cases from a human factors and cognitive psychology perspective. Annals of Surgery **237**(4), 460–469 (2003)
27. Yang, S., Cai, Z., Luo, L., Ma, N., Xu, S., et al.: Surgpetl: Parameter-efficient image-to-surgical-video transfer learning for surgical phase recognition. IEEE Transactions on Medical Imaging **45**(3), 1013–1023 (2026). https://doi.org/10.1109/TMI.2025.3615967