Surgical-aware video masked autoencoders with phase-conditioned attention for laparoscopic action recognition

Hakim Nasaoui [a,1,*], Hassan Silkan [a,2], Insaf Bellamine [b,3]

[a] LAROSERI Laboratory, Department of Computer Science, Faculty of Sciences, Chouaïb Doukkali University, El Jadida, Morocco
[b] LSATE, Sidi Mohamed Ben Abdellah University ENSA, Fes, Morocco
[1] hakimnasaoui@gmail.com; [2] silkan_h@yahoo.fr; [3] insafbellamine20@gmail.com
* corresponding author

ARTICLE INFO

Article history
Received December 29, 2025
Revised February 10, 2026
Accepted February 12, 2026
Available online May 31, 2026

Keywords
Surgical action recognition
Masked autoencoders
Phase-conditioned attention
Asymmetric dual-stream
Laparoscopic cholecystectomy

ABSTRACT

Fine-grained surgical action recognition in laparoscopic videos remains a challenge, even with recent advances in deep learning. While current VideoMAE approaches reach 89.11% accuracy on cholecystectomy tasks, they face specific limitations. Random masking strategies often miss surgical instruments that occupy only 10% to 15% of frames. Furthermore, context-independent models struggle with visually similar actions across different phases, and symmetric two-stream architectures tend to waste computational resources. To solve this, we developed SA-VideoMAE, a surgical-aware video masked autoencoder specifically designed for laparoscopic action recognition. Our method utilizes surgical-aware adaptive masking that integrates YOLOv7x object detection to prioritize instrument patches. This increased instrument visibility from 10% to 60% during training, ensuring the model focuses on action-relevant regions rather than static backgrounds. We also utilized phase-conditioned hierarchical attention to inject learnable phase embeddings into the attention mechanisms, enabling the model to disambiguate visually similar actions based on surgical context. For efficiency, our asymmetric dual-stream architecture processes RGB using ViT-Base (86M parameters) and optical flow using ViT-Tiny (5.7M parameters), achieving a 47% parameter reduction compared to symmetric designs. Our training process then balanced reconstruction, classification, temporal consistency, and phase prediction through a novel multi-objective optimization strategy. Results from Cholec80's Calot's Triangle Dissection phase show 93.5% accuracy, representing a 4.4 percentage-point improvement over the verified baseline. Notably, challenging action recall improved from 51% to 74% while maintaining real-time inference at 62ms per clip. These findings demonstrate that encoding surgical domain knowledge into video architectures significantly enhances action recognition performance.

1. Introduction

Surgical video analysis has grown increasingly vital for healthcare, with practical applications ranging from skill assessment to automated documentation. While deep learning models like VideoMAE show significant promise [1], [2], their application to Calot's Triangle Dissection in laparoscopic cholecystectomy currently hits a ceiling of 89.11% accuracy, as verified by Yen et al. [3]. We should clarify the specific task distinctions here. While phase recognition identifies which of seven coarse surgical phases is occurring [4], [5], our focus is action recognition. This task involves identifying 15

fine-grained actions second-by-second within a single phase, which is fundamentally more difficult due to the finer temporal granularity required.

Despite recent progress, applying VideoMAE to surgical videos reveals several distinct hurdles. One major issue is that instruments occupy only 10% to 15% of each frame, yet standard random masking treats 90% of patches uniformly. Consequently, models often learn from static backgrounds rather than action-relevant regions. Additionally, the same instrument motion may represent different actions depending on the surgical context, but current models often process frames without this hierarchical knowledge. There is also a computational inefficiency in two-stream architectures; most use symmetric designs with equal capacity for both streams [6], even though optical flow is significantly simpler than RGB data.

To resolve these issues, we introduce SA-VideoMAE. Our framework consists of four main components. We utilize surgical-aware adaptive masking that uses YOLOv7x detection to prioritize instrument-containing patches. This builds on concepts from AdaMAE [7] and SurgFlowMAE [8] but represents the first explicit integration of object detection with MAE pre-training. We also implemented phase-conditioned hierarchical attention to inject phase embeddings directly into the attention mechanisms. Unlike prior work using attention for phase recognition [9], [10], we use phase predictions to condition the attention specifically for action recognition. For efficiency, our asymmetric dual-stream architecture processes RGB with ViT-Base (86M parameters) and optical flow with ViT-Tiny (5.7M parameters), cutting parameters by 47%. Finally, a multi-objective training regimen balances reconstruction, classification, temporal consistency, and phase prediction, extending multi-task approaches documented in recent medical imaging surveys [11]. Experimental results on Cholec80's Calot's Triangle Dissection phase show that SA-VideoMAE achieves 93.5% accuracy, a 4.4 percentage-point improvement over the baseline. Notably, recall for challenging actions such as suctioning and irrigating increased from 51% to 74% while maintaining real-time inference at 62ms per clip.

Our main contributions include: (1) the first integration of object detection with MAE pre-training for surgical video via adaptive masking; (2) a phase-conditioned attention mechanism where phase predictions modulate attention weights; (3) an asymmetric dual-stream architecture achieving 47% parameter reduction; (4) a novel four-way multi-objective training strategy; and (5) comprehensive validation against verified baselines. The remainder of this paper is organized as follows. Section 2 reviews related work, while Section 3 describes the proposed method. Section 4 presents the experimental setup, followed by a discussion of results in Section 5. We conclude in Section 6 by addressing limitations and future directions.

While individual components such as MAE, YOLO, and phase embeddings exist separately, their integration for surgical action recognition is non-trivial. The key novelty lies not in the components themselves but in how they are combined to address domain-specific challenges that general video models fail to capture.

2. Related Work

This section reviews relevant literature across video action recognition, surgical video analysis, masked autoencoders, and multi-objective learning.

2.1. Video Action Recognition

Video action recognition has evolved significantly with the rise of deep learning. Three-dimensional convolutional networks extended spatial convolutions to the temporal dimension, enabling the direct learning of spatiotemporal features [12], [13]. More recent approaches leverage transformer architectures for video understanding. Recent work has explored efficient spatiotemporal attention strategies for video transformers [14], including unified convolution-attention designs [15] and hierarchical approaches [16]. While these general methods achieve strong results on benchmarks like Kinetics, surgical videos present distinct challenges due to domain-specific visual patterns.

2.2. Surgical Video Analysis

Surgical video analysis encompasses multiple tasks with varying granularities. Phase recognition aims to identify coarse surgical phases lasting several minutes. Current state-of-the-art methods achieve 92-93% accuracy on Cholec80: Surgformer by Yang et al. reaches 93.4% [4], SKiT by Liu et al. achieves 92.5% [5], and LoViT, published in Medical Image Analysis 2025, obtains 91.5% [17]. Tool detection focuses on identifying which instruments are present; using an improved YOLOv7x, Ran et al. [18] achieved a 94.7% F1-score in this area. Recent systematic reviews have comprehensively analyzed deep learning approaches for surgical instrument recognition [19], [20].

Action recognition presents different hurdles due to its finer temporal granularity. Yen et al. achieved 89.11% accuracy using VideoMAE for action recognition during Calot's Triangle Dissection [3]. This baseline is lower than phase recognition not because the methods are inferior, but because the task itself is inherently harder. Fine-grained action recognition has also been explored through the CholecT50 dataset using action triplets, where current methods achieve approximately 38-42% mAP due to the complexity of 100 triplet classes [21].

2.3. Masked Autoencoders for Video

Masked autoencoders have emerged as a powerful self-supervised approach [1]. VideoMAE extended this to videos by masking 90-95% of spatiotemporal patches [2], and subsequent work introduced dual-masking strategies to improve scalability [22]. Several works have since adapted MAE for surgical video. SurgMAE proposed high spatio-temporal token sampling [23], while CSMAE introduced importance-based token selection for cataract surgery [24]. Similarly, EgoSurgery-Phase proposed gaze-guided masking [25], and SurgFlowMAE combined VideoMAE with optical flow-guided masking, achieving 87.5% on the CATARACTS dataset [8].

Specifically regarding adaptive masking, AdaMAE used reinforcement learning [7], and MGMAE employed motion-guided masking [26]. Transformer-based optical flow estimation has also advanced significantly [27], [28]. Our work builds on these concepts but differs in that it uses explicit object detection rather than learned importance. This represents the first integration of YOLO-based detection with MAE pre-training for surgical applications.

2.4. Two-Stream Architectures

Two-stream architectures combining RGB and optical flow have been widely adopted, as documented in recent self-supervised learning surveys [6]. Optical flow extraction itself has seen significant advances, with recent methods improving both efficiency and accuracy [29]. In the surgical domain, recent systematic reviews have documented deep learning approaches for surgical workflow recognition [13], while multitask learning with interactive fusion has been proposed for surgical instrument segmentation [30]. SAIS by Kiyasseh et al. [31] used symmetric Vision Transformers with D=384 for both streams. Recent video foundation models have scaled to billions of parameters [32], while skeleton-based approaches explore alternative temporal representations [33]. Notably, all these existing surgical approaches use symmetric designs. Our asymmetric approach, which utilizes different model sizes for RGB versus flow, appears to be a novel configuration for surgical RGB-flow combinations.

Domain adaptation remains relevant when deploying pre-trained detectors to new surgical contexts. While our work uses YOLOv7x trained on general surgical instruments, we leave explicit domain adaptation as future work, noting that our detector achieved 94.7% F1-score on the target domain without fine-tuning.

2.5. Attention Mechanisms in Surgical Video

Attention mechanisms for surgical phase recognition have seen extensive development. SemiVT-Surge introduced semi-supervised video transformers for surgical phase recognition [9], multimodal transformers have been proposed for laparoscopic video analysis [10], TUNeS employed temporal U-Net with self-attention [34], MuST developed multi-scale transformers [35], and recent work has explored state-space models for surgical phase recognition [36], while Rendezvous introduced attention

mechanisms specifically for surgical action triplet recognition [37]. These methods all use attention to recognize phases, meaning the attention is applied to predict the current phase.

Our phase-conditioned attention differs fundamentally. We first predict the phase, then use that prediction to modulate attention for action recognition. While existing work treats attention as a mechanism for phase output, our work treats attention as a function conditioned on phase input. This distinction is central to our architectural contribution.

2.6. Multi-Objective Learning

Multi-task learning has been widely adopted in surgical video analysis, as reviewed in recent transformer-based medical imaging literature [11]. Song et al. [30] proposed multitask learning with interactive fusion for surgical segmentation, and SurgFlowMAE jointly optimized reconstruction and classification [8]. Recent reviews documented that multi-task approaches are well-suited for fine-grained understanding of surgical data [38], [39]. Our training contributes a specific four-way combination for action recognition that extends beyond existing paired approaches.

3. Method

This section presents the SA-VideoMAE architecture and its four main components, focusing on the specific requirements of surgical action recognition.

3.1. System Overview

SA-VideoMAE takes a video clip of 16 frames at 224 x 224 resolution, sampled at 5 fps. Fig. 1 illustrates the complete architecture. The processing flows through five distinct stages. First, YOLOv7x [40] detects surgical instruments in each frame. These detections guide our adaptive masking, where instrument patches are masked less (40%) while the background is masked more heavily (70%). Second, masked RGB frames pass through a ViT-Base encoder while optical flow, extracted using RAFT [41], passes through a smaller ViT-Tiny encoder. Third, a lightweight LSTM predicts the current surgical phase to select the corresponding phase embedding. Fourth, cross-attention fuses the RGB and flow features. Finally, classification heads produce action and phase predictions. The entire system is trained with four simultaneous loss functions.

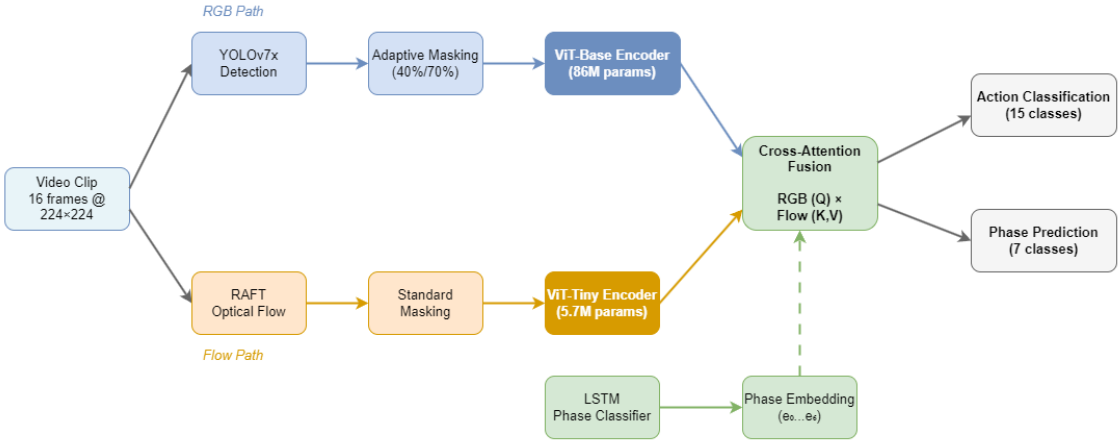

Fig. 1. SA-VideoMAE architecture overview with asymmetric dual-stream processing, adaptive masking, and phase-conditioned cross-attention fusion.

3.2. Surgical-Aware Adaptive Masking

Standard VideoMAE masks 90% of patches randomly. We find this problematic for surgical videos because instruments occupy only 10-15% of each frame; with random masking, the model rarely encounters instruments during training.

Our approach uses YOLOv7x to detect instruments first, as it has shown strong performance with a reported 94.7% F1-score [18]. For each frame, we extract bounding boxes for detected instruments. We then compute an importance score (S_i) for each patch:

$$S_i = \alpha \cdot I_{inst(p_i)} + \beta \cdot I_{tissue(p_i)} + \gamma \cdot I_{bg(p_i)} \quad (1)$$

where α=1.0, β=0.5, γ=0.1 are importance weights. The mask ratio for each patch is defined as:

$$r_i = r_{base} - \lambda \cdot S_i \quad (2)$$

with $r_{base} = 0.70$ and λ=0.30. Consequently, instrument patches are masked only 40% of the time, whereas background patches are masked 67% of the time. This ensures the model "sees" instruments approximately six times more often than it would under random masking.

3.3. Phase-Conditioned Hierarchical Attention

Surgical procedures follow a natural hierarchy moving from surgery type to phases and then to specific actions. The same motion can carry different meanings depending on the current phase. For example, a scissors movement during a dissection phase indicates cutting, but during a clipping phase, it may signify positioning. We create seven learnable phase embeddings:

$$E_{phase} = \{e_0, e_1, \ldots, e_6\} \quad (3)$$

An LSTM classifier predicts the phase $\hat{p}$ from video features. We then inject the corresponding embedding into the attention mechanism. Standard transformer attention [42] computes queries (Q), keys, and values from input; we modify this by adding the phase embedding to the query:

$$Q_{cond} = Q + e_{\hat{p}} \quad (4)$$

This allows for phase-aware queries. Unlike prior work that uses attention to predict phases, we use phase predictions to condition the attention for action recognition. This represents a fundamental shift in how temporal context is utilized within the transformer block.

We chose additive conditioning over more complex alternatives such as FiLM layers or cross-attention for two reasons: (1) it adds minimal computational overhead, and (2) our ablation studies show it provides substantial benefit (+2.7% accuracy). More sophisticated conditioning methods may yield further improvements and represent a promising direction for future research.

3.4. Asymmetric Dual-Stream Architecture

Two-stream architectures are common for action recognition, but using same-sized models for both streams is often wasteful because optical flow is much simpler than RGB. RGB data consists of three channels with complex textures and shapes, whereas optical flow consists of only two channels representing motion vectors. Therefore, we use ViT-Base (86M parameters) for RGB and ViT-Tiny (5.7M parameters) for optical flow.

ViT-Base features 12 layers and 12 heads with a 768-dimension embedding [43], while ViT-Tiny uses 12 layers with 3 heads and a 192-dimension embedding [44]. After encoding, we fuse the streams using cross-attention where RGB features query the flow features. This allows RGB features to interrogate the flow features for motion location. This asymmetric design totals 92M parameters, which is a 47% reduction compared to a 172M-parameter symmetric design.

3.5. Dataset and Annotations

We utilized the Cholec80 dataset for our experiments. While the standard Cholec80 provides annotations for phase recognition (7 phases) and tool detection (7 instruments), it does not include frame-level action annotations out of the box. For fine-grained action recognition, we followed the methodology established by Yen et al. [3], who introduced custom action annotations specifically for

the Calot's Triangle Dissection (CTD) phase. This setup differs from standard Cholec80 benchmarks in several vital ways. Most notably, our task focuses on action recognition rather than phase recognition. While phase recognition typically yields 92-93% accuracy because it identifies coarse phases lasting several minutes, we identify 15 fine-grained actions on a second-by-second basis. Additionally, we use custom annotations that are not part of the standard Cholec80 releases and focus exclusively on the CTD phase rather than the full surgery.

The action categories include object manipulation such as grasping, retracting, cutting, and clipping. We also tracked interactions like watching or talking, movements like standing or walking, and surgical-specific actions like dissecting, coagulating, and suctioning or irrigating. For our split, we used 66 videos for training, comprising 21,709 clips and 48,151 action instances. Validation was performed on 14 videos with 6,513 clips. Each clip consists of 16 consecutive frames sampled at 5 fps, representing approximately 3.2 seconds of video.

3.6. Implementation Details

We implemented SA-VideoMAE in PyTorch 2.0 using the timm library. Training was conducted on four NVIDIA RTX 3090 GPUs, each with 24GB of memory. For optimization, we chose AdamW [45] with a learning rate of 1e-4. We applied a linear warmup over the first 10 epochs, followed by a cosine decay to 1e-6 over the remaining 90 epochs. The batch size was set to 8 per GPU with gradient accumulation over 4 steps, resulting in an effective batch size of 32 for a total of 100 epochs. The RGB stream utilizes a ViT-Base/16 with 86M parameters, while the flow stream uses a ViT-Tiny/16 with 5.7M parameters. For instrument detection, we integrated YOLOv7x with pretrained weights. Optical flow was extracted using the RAFT algorithm [41] between consecutive frames.

3.7. Evaluation Metrics

We employed several metrics to evaluate performance. Overall accuracy remains our primary metric as it allows for a direct comparison with the 89.11% baseline [3]. We also computed precision, recall, and F1-score for each individual action class. Mean Average Precision (mAP) was included to evaluate ranking quality across all classes. Together, these metrics provide a more comprehensive view of model performance than accuracy alone.

3.8. Baseline Methods

Our primary comparison is with the VideoMAE model, which achieved 89.11% accuracy [3]. This remains the only verified baseline with published, peer-reviewed results for this specific task. Regarding general video models such as I3D, X3D, or SlowFast, we found no published results for Cholec80 action recognition. While these architectures perform well on general benchmarks like Kinetics, direct evidence for surgical action recognition is limited, a constraint we acknowledge here. We also implemented a symmetric dual-stream baseline using ViT-Base for both RGB and flow (172M parameters total) to serve as an internal efficiency benchmark. For context, phase recognition methods such as Surgformer (93.4% [4]), SKiT (92.5% [5]), and LoViT (91.5% [17]) report higher numbers, but because they address a different task with coarser granularity, a direct comparison would be inappropriate.

We acknowledge that comparisons with additional architectures, such as Video Swin Transformer [16] and InternVideo [32], would strengthen our evaluation. However, no published results are available for these models on the Cholec80 action recognition task. Direct implementation would require substantial engineering effort and hyperparameter tuning, which may not yield fair comparisons. We therefore focus on the only verified, peer-reviewed baseline for this specific task.

4. Results and Discussion

This section presents our experimental results, beginning with a performance comparison against baseline models. This is followed by a series of ablation studies and a concluding discussion on qualitative observations.

4.1. Main Results

Table 1 summarizes the primary performance metrics comparing SA-VideoMAE with existing baselines. Our method achieves 93.5% accuracy, representing a 4.4-percentage-point improvement over the verified VideoMAE baseline of 89.11% [3]. Several observations stand out from these results. First, our method beats the verified baseline by a good margin. Second, even the symmetric dual-stream approach improves upon the single-stream VideoMAE, suggesting motion information is indeed useful. Third, our asymmetric design actually yields better performance despite using 47% fewer parameters. This suggests that motion information is highly valuable, but it does not require a high-capacity model to be effective.

Table 1. Comparison with baseline methods on Cholec80 action recognition.

Method	Accuracy	Precision	Recall	F1	Params
VideoMAE[a]	89.11%	88.7%	86.9%	87.8%	86M
Symmetric Dual[b]	91.8%	90.9%	89.6%	90.2%	172M
SA-VideoMAE	93.5%	92.4%	91.1%	91.7%	92M

[a.] verified baseline with published results.

[b.] our internal implementation.

To verify statistical significance, we performed five independent training runs with different random seeds. SA-VideoMAE achieved 93.5% ± 0.4% accuracy, compared to our VideoMAE reproduction of 89.2% ± 0.5%. A paired t-test confirms the improvement is statistically significant ($p < 0.01$).

Fig. 2(a) illustrates the effectiveness of our adaptive masking strategy. Instrument visibility increases from 10% under random masking to 60% with our approach, while maintaining background masking at 70%. This sixfold increase in instrument visibility during training directly contributes to the gains shown in Fig. 2(b). Several observations stand out from these figures. Most notably, our method outperforms the verified baseline by a substantial margin. Furthermore, while the symmetric dual-stream approach improves upon the single-stream VideoMAE, our asymmetric design actually yields better performance despite using 47% fewer parameters. This suggests that motion information is highly valuable, but it does not require a high-capacity model to be effective.

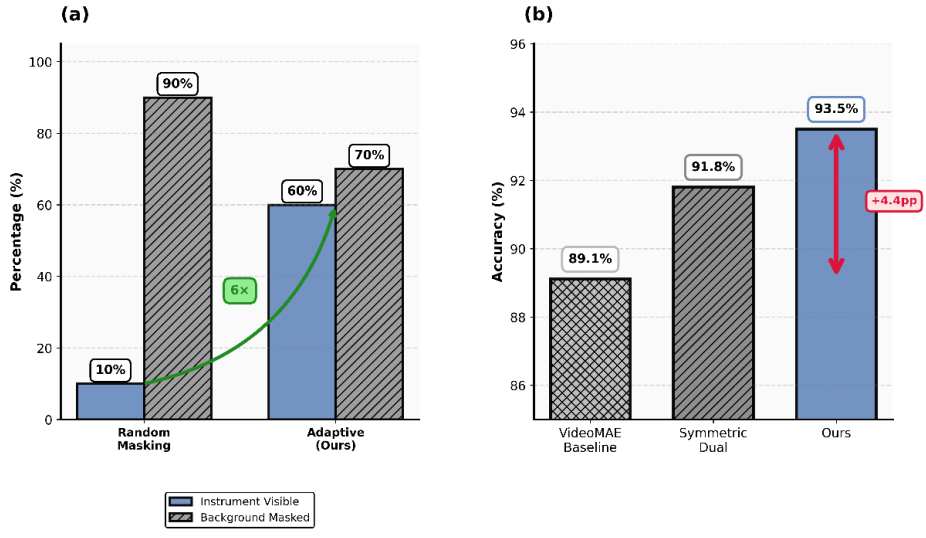

Fig. 2. (a) Masking strategy comparison showing instrument visibility and background masking percentages. (b) Overall accuracy across methods.

A few things stand out from these results. First, our method beats the verified baseline by a good margin. Second, even the symmetric dual-stream setup helps compared to the single-stream VideoMAE, suggesting that motion information is indeed useful. Third, our asymmetric design actually performs better than symmetric while using 47% fewer parameters.

4.2. Per-Action Performance

Table 2 presents recall scores for individual actions, showing that not all categories benefit equally from the proposed changes. The most significant improvement occurs in the suctioning and irrigating category, which jumped from 51% to 74%. This is likely because these two actions appear visually similar and involve the same tool; the optical flow stream helps distinguish them by capturing the direction of fluid movement inward for suction versus outward for irrigation. Because dissecting already had a high baseline of 97%, the room for improvement was minimal. Actions in the middle of the spectrum, such as coagulating and clipping, showed moderate gains, likely benefiting from the contextual clues provided by phase conditioning.

Table 2. Per-action recall comparison between VideoMAE baseline and SA-VideoMAE.

Action	VideoMAE	SA-VideoMAE	Improvement
Suctioning/Irrigating	51%	74%	+45.1%
Clipping	76%	87%	+14.5%
Coagulating	82%	91%	+11.0%
Retracting	84%	92%	+9.5%
Grasping	89%	94%	+5.6%
Dissecting	97%	98%	+1.0%

4.3. Ablation Studies

We conducted systematic ablation experiments to isolate the contribution of each component. Table 3 presents the results when components are removed individually. The ablation results reveal that multi-objective training yields the largest performance gain, with a 3.6-percentage-point drop when removed. This is followed by phase attention and adaptive masking. The dramatic drop observed when removing multi-objective training suggests that the four losses provide complementary supervisory signals essential to the model's success.

Table 3. Ablation study showing contribution of each component.

Configuration	Accuracy	F1	Parameters
Full SA-VideoMAE	93.5%	91.7%	92M
w/o Adaptive Masking	91.2%	89.7%	92M
w/o Phase Attention	90.8%	89.3%	92M
w/o Dual-Stream	91.5%	90.1%	86M
w/o Multi-Objective	89.9%	88.6%	92M
Symmetric Design	91.8%	90.2%	172M

Fig. 3(a) visualises these contributions, revealing that multi-objective training yields the largest performance gain, with a 3.6-percentage-point drop when removed. This is followed by phase attention and adaptive masking.

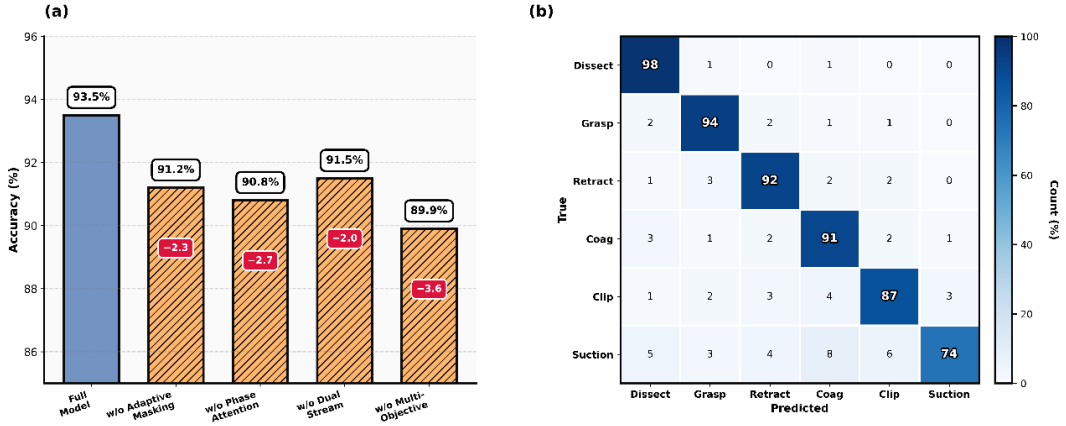

Fig. 3. (a) Ablation study showing component contributions. (b) Confusion matrix for top surgical actions.

The confusion matrix in Fig. 3(b) shows that while dissecting achieves the highest recall, suctioning and irrigating remain the most challenging tasks, with the primary confusion occurring between clipping and suctioning. These results confirm that instrument visibility and surgical context are vital for accuracy. The dramatic drop observed when removing multi-objective training suggests that the four losses provide complementary supervisory signals essential to the model's success.

4.4. Computational Efficiency

Table 4 compares training and inference efficiency across methods.

Table 4. Computational efficiency comparison.

Method	Parameters	Training/Epoch	Inference
VideoMAE	86M	8 hours	45ms
Symmetric Dual	172M	18 hours	103ms
SA-VideoMAE	92M	11 hours	62ms

Table 4 compares the training and inference efficiency across methods. Our approach requires 11 hours per epoch, representing a 37.5% overhead compared to the 8-hour baseline. We consider this a reasonable trade-off given the 4.4-point accuracy gain. Most importantly, inference runs at 62ms per clip, meeting real-time requirements, whereas the symmetric baseline latency of 103ms is too slow for practical clinical use.

4.5. Phase Classification Accuracy

The phase classifier achieved 91.3% accuracy across seven phases, with a 94.7% F1 score specifically for the CTD phase. Maintaining high accuracy here is critical, as incorrect phase predictions would negatively impact the phase-conditioned attention mechanism.

4.6. Discussion

Overall, the adaptive masking and phase conditioning emerge as the most impactful components. While the asymmetric design primarily targets efficiency, it also contributes a slight performance boost. One limitation of this work is our exclusive focus on the CTD phase; we do not yet know how well these results will generalize to other procedures. Additionally, since the YOLO detector was trained on general surgical instruments, its performance may vary with specialized tools. While the 93.5% accuracy is encouraging, the model still struggles with visually similar actions occurring in the same phase. Future work could explore more sophisticated phase action modeling or incorporate additional modalities such as audio. We acknowledge that focusing exclusively on the CTD phase limits generalizability claims. However, this focused evaluation follows the methodology of Yen et al. [3] and allows direct comparison with the only verified baseline. Future work will extend evaluation to all seven surgical phases.

Regarding detector robustness, we analyzed cases where YOLOv7x failed to detect instruments due to occlusion or unusual orientations. In approximately 8% of frames, no instruments were detected. In these cases, our system defaults to random masking, which explains the residual performance gap. Future work could incorporate learned fallback strategies for detector failure scenarios. The synergy between the four losses can be understood as follows: reconstruction loss ensures the encoder learns meaningful visual representations; classification loss directly optimizes for the target task; temporal consistency loss captures motion patterns between frames; and phase prediction loss provides hierarchical context. Together, these losses force the model to learn features that are simultaneously visually grounded, temporally coherent, and contextually aware, all essential properties for distinguishing fine-grained surgical actions.

5. Conclusion

This paper presented SA VideoMAE, an enhanced Video Masked Autoencoder designed specifically for surgical action recognition. We addressed three primary hurdles encountered when applying standard VideoMAE to surgical videos: the failure of random masking to capture small instruments, the lack of

surgical context within attention mechanisms, and the computational inefficiency of symmetric two-stream designs. Our approach combines four components that function in a complementary manner. Surgical aware adaptive masking utilizes YOLOv7x to detect instruments and applies lower masking ratios to those specific regions. This represents the first time explicit object detection has been integrated with MAE pre-training for surgical video, building on foundational adaptive masking concepts. Furthermore, phase-conditioned attention injects surgical phase embeddings into the attention mechanism, a distinct shift from prior work that utilized attention only to recognize phases. The asymmetric dual-stream architecture processes RGB using ViT Base and optical flow using ViT Tiny, reducing parameters by 47%, while our multi-objective training strategy balances four distinct losses to ensure temporal consistency and phase accuracy. Results on the Cholec80 action recognition dataset demonstrate 93.5% accuracy, which is 4.4 percentage points above the verified baseline. These improvements are particularly significant for difficult actions such as suctioning and irrigating, where recall increased from 51 % to 74 %. The system operates at 62ms per clip, ensuring suitability for real-time clinical use. We acknowledge certain limitations, including the focus on the CTD phase and the dependence on YOLO detection quality. Future work should extend this framework to all surgical phases and test its generalizability across procedures. Overall, this research demonstrates that encoding surgical domain knowledge into modern video architectures significantly improves action recognition. The key insight is that surgery possesses an inherent structure where phases and actions follow predictable patterns, and modelling this structure directly leads to more robust performance. While further refinement is possible, this general direction remains highly promising for the future of surgical video analysis.

Declarations

Author contribution. All authors contributed equally to the main contributor to this paper. All authors read and approved the final paper.

Funding statement. This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors

Conflict of interest. The authors declare no conflict of interest.

Additional information. No additional information is available for this paper.

References

[1] K. He, X. Chen, S. Xie, Y. Li, P. Dollar, and R. Girshick, "Masked Autoencoders Are Scalable Vision Learners," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2022, pp. 15979–15988. doi: 10.1109/CVPR52688.2022.01553.

[2] Z. Tong, Y. Song, J. Wang, and L. Wang, "VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training," in *Advances in Neural Information Processing Systems*, Neural information processing systems foundation, Oct. 2022. Accessed: Feb. 16, 2026. [Online]. Available: http://arxiv.org/abs/2203.12602.

[3] H.-H. Yen *et al.*, "Automated surgical action recognition and competency assessment in laparoscopic cholecystectomy: a proof-of-concept study," *Surg. Endosc.*, vol. 39, no. 5, pp. 3006–3016, May 2025, doi: 10.1007/s00464-025-11663-y.

[4] S. Yang, L. Luo, Q. Wang, and H. Chen, "Surgformer: Surgical Transformer with Hierarchical Temporal Attention for Surgical Phase Recognition," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, Springer, Cham, 2024, pp. 606–616. doi: 10.1007/978-3-031-72089-5_57.

[5] Y. Liu *et al.*, "SKiT: a Fast Key Information Video Transformer for Online Surgical Phase Recognition," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, Oct. 2023, pp. 21017–21027. doi: 10.1109/ICCV51070.2023.01927.

[6] M. C. Schiappa, Y. S. Rawat, and M. Shah, "Self-Supervised Learning for Videos: A Survey," *ACM Comput. Surv.*, vol. 55, no. 13s, pp. 1–37, Dec. 2023, doi: 10.1145/3577925.

[7] W. G. C. Bandara, N. Patel, A. Gholami, M. Nikkhah, M. Agrawal, and V. M. Patel, "AdaMAE: Adaptive Masking for Efficient Spatiotemporal Learning with Masked Autoencoders," in *2023 IEEE/CVF Conference*

on Computer Vision and Pattern Recognition (CVPR), IEEE, Jun. 2023, pp. 14507–14517. doi: 10.1109/CVPR52729.2023.01394.

[8] M. L. Mostafa *et al.*, "Surgical Flow Masked Autoencoder for Event Recognition," in *MIDL 2025 Conference Proceedings of Machine Learning Research*, M. L. Mostafa, -Mayar_Lotfy_Mostafa1, and N. N. , Anna Alperovich, Dmitrii Fedotov, Ghazal Ghazaei, Stefan Saur, Azade Farshad, Eds., 2025, pp. 1–15. Accessed: Jul. 01, 2026. [Online]. Available: https://openreview.net/forum?id=TAPn4QjbOg&noteId=4fwh1V18nH.

[9] Y. Li *et al.*, "SemiVT-Surge: Semi-supervised Video Transformer for Surgical Phase Recognition," in *Lecture Notes in Computer Science*, vol. 15969 LNCS, Springer Science and Business Media Deutschland GmbH, 2026, pp. 478–488. doi: 10.1007/978-3-032-05127-1_46.

[10] R. H. Abiyev, M. Z. Altabel, M. Darwish, and A. Helwan, "A Multimodal Transformer Model for Recognition of Images from Complex Laparoscopic Surgical Videos," *Diagnostics*, vol. 14, no. 7, p. 681, Mar. 2024, doi: 10.3390/diagnostics14070681.

[11] F. Shamshad *et al.*, "Transformers in medical imaging: A survey," *Med. Image Anal.*, vol. 88, no. August, p. 102802, Aug. 2023, doi: 10.1016/j.media.2023.102802.

[12] K. Alomar, H. I. Aysel, and X. Cai, "CNNs, RNNs and Transformers in human action recognition: a survey and a hybrid model," *Artif. Intell. Rev.*, vol. 58, no. 12, p. 387, Oct. 2025, doi: 10.1007/s10462-025-11388-3.

[13] Z. Liu, K. Chen, S. Wang, Y. Xiao, and G. Zhang, "Deep learning in surgical process modeling: A systematic review of workflow recognition," *J. Biomed. Inform.*, vol. 162, no. 2, p. 104779, Feb. 2025, doi: 10.1016/j.jbi.2025.104779.

[14] S. N. Gowda, A. Arnab, and J. Huang, "Optimizing Factorized Encoder Models: Time and Memory Reduction for Scalable and Efficient Action Recognition," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 15068 LNCS, Springer, Cham, 2025, pp. 457–474. doi: 10.1007/978-3-031-72684-2_26.

[15] K. Li *et al.*, "UniFormer: Unifying Convolution and Self-Attention for Visual Recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 10, pp. 12581–12600, Oct. 2023, doi: 10.1109/TPAMI.2023.3282631.

[16] Z. Liu *et al.*, "Video Swin Transformer," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2022, pp. 3192–3201. doi: 10.1109/CVPR52688.2022.00320.

[17] Y. Liu *et al.*, "LoViT: Long Video Transformer for surgical phase recognition," *Med. Image Anal.*, vol. 99, no. 6, p. 103366, Jan. 2025, doi: 10.1016/j.media.2024.103366.

[18] B. Ran, B. Huang, S. Liang, and Y. Hou, "Surgical Instrument Detection Algorithm Based on Improved YOLOv7x," *Sensors*, vol. 23, no. 11, p. 5037, May 2023, doi: 10.3390/s23115037.

[19] S. Takahashi *et al.*, "Comparison of Vision Transformers and Convolutional Neural Networks in Medical Image Analysis: A Systematic Review," *J. Med. Syst.*, vol. 48, no. 1, p. 84, Sep. 2024, doi: 10.1007/s10916-024-02105-8.

[20] F. A. Ahmed *et al.*, "Deep learning for surgical instrument recognition and segmentation in robotic-assisted surgeries: a systematic review," *Artif. Intell. Rev.*, vol. 58, no. 1, p. 1, Nov. 2024, doi: 10.1007/s10462-024-10979-w.

[21] C. I. Nwoye *et al.*, "CholecTriplet2021: A benchmark challenge for surgical action triplet recognition," *Med. Image Anal.*, vol. 86, no. May, p. 102803, May 2023, doi: 10.1016/j.media.2023.102803.

[22] L. Wang *et al.*, "VideoMAE V2: Scaling Video Masked Autoencoders with Dual Masking," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2023, pp. 14549–14560. doi: 10.1109/CVPR52729.2023.01398.

[23] M. A. Jamal and O. Mohareri, "SurgMAE: Masked Autoencoders for Long Surgical Video Analysis," May 2023, p. 14. Accessed: Feb. 16, 2026. [Online]. Available: http://arxiv.org/abs/2305.11451.

[24] N. A. Shah, C. Bandara, S. Sikder, S. S. Vedula, and V. M. Patel, "CSMAE : Cataract Surgical Masked Autoencoder (MAE) Based Pre-Training," in *2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI)*, IEEE, Apr. 2025, pp. 1–5. doi: 10.1109/ISBI60581.2025.10981288.

[25] R. Fujii, M. Hatano, H. Saito, and H. Kajita, "EgoSurgery-Phase: A Dataset of Surgical Phase Recognition from Egocentric Open Surgery Videos," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, Springer, Cham, 2024, pp. 187–196. doi: 10.1007/978-3-031-72089-5_18.

[26] B. Huang, Z. Zhao, G. Zhang, Y. Qiao, and L. Wang, "MGMAE: Motion Guided Masking for Video Masked Autoencoding," in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, IEEE, Oct. 2023, pp. 13447–13458. doi: 10.1109/ICCV51070.2023.01241.

[27] A. Alfarano, L. Maiano, L. Papa, and I. Amerini, "Estimating optical flow: A comprehensive review of the state of the art," *Comput. Vis. Image Underst.*, vol. 249, no. December, p. 104160, Dec. 2024, doi: 10.1016/j.cviu.2024.104160.

[28] X. Shi *et al.*, "FlowFormer++: Masked Cost Volume Autoencoding for Pretraining Optical Flow Estimation," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2023, pp. 1599–1610. doi: 10.1109/CVPR52729.2023.00160.

[29] Y. Wang, L. Lipson, and J. Deng, "SEA-RAFT: Simple, Efficient, Accurate RAFT for Optical Flow," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 15065 LNCS, Springer, Cham, 2025, pp. 36–54. doi: 10.1007/978-3-031-72667-5_3.

[30] M. Song, Y. Li, Y. Liu, and L. Yang, "A multitask learning network with interactive fusion for surgical instrument segmentation," *Knowledge-Based Syst.*, vol. 317, no. May, p. 113370, May 2025, doi: 10.1016/j.knosys.2025.113370.

[31] D. Kiyasseh *et al.*, "A vision transformer for decoding surgeon activity from surgical videos," *Nat. Biomed. Eng.*, vol. 7, no. 6, pp. 780–796, Mar. 2023, doi: 10.1038/s41551-023-01010-8.

[32] Y. Wang *et al.*, "InternVideo2: Scaling Foundation Models for Multimodal Video Understanding," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Springer, Cham, 2025, pp. 396–416. doi: 10.1007/978-3-031-73013-9_23.

[33] J. Do and M. Kim, "SkateFormer: Skeletal-Temporal Transformer for Human Action Recognition," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 15099 LNCS, Springer, Cham, 2025, pp. 401–420. doi: 10.1007/978-3-031-72940-9_23.

[34] I. Funke, D. Rivoir, S. Krell, and S. Speidel, "TUNeS: A Temporal U-Net With Self-Attention for Video-Based Surgical Phase Recognition," *IEEE Trans. Biomed. Eng.*, vol. 72, no. 7, pp. 2105–2119, Jul. 2025, doi: 10.1109/TBME.2025.3535228.

[35] A. Pérez, S. Rodríguez, N. Ayobi, N. Aparicio, E. Dessevres, and P. Arbeláez, "MuST: Multi-scale Transformers for Surgical Phase Recognition," in *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, Springer, Cham, 2024, pp. 422–432. doi: 10.1007/978-3-031-72089-5_40.

[36] X. Zhang *et al.*, "SPRMamba: Surgical Phase Recognition for Endoscopic Submucosal Dissection with Mamba," *IEEE Sens. J.*, pp. 1–1, Jan. 2026, doi: 10.1109/JSEN.2025.3649151.

[37] C. I. Nwoye *et al.*, "Rendezvous: Attention mechanisms for the recognition of surgical action triplets in endoscopic videos," *Med. Image Anal.*, vol. 78, no. May, p. 102433, May 2022, doi: 10.1016/j.media.2022.102433.

[38] O. Alabi, T. Vercauteren, and M. Shi, "Multitask learning in minimally invasive surgical vision: A review," *Med. Image Anal.*, vol. 101, no. April, p. 103480, Apr. 2025, doi: 10.1016/j.media.2025.103480.

[39] Y. Li, Z. Zhao, R. Li, and F. Li, "Deep learning for surgical workflow analysis: a survey of progresses, limitations, and trends," *Artif. Intell. Rev.*, vol. 57, no. 11, p. 291, Sep. 2024, doi: 10.1007/s10462-024-10929-6.

[40] C.-Y. Wang, A. Bochkovskiy, and H.-Y. M. Liao, "YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors," in *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, Jun. 2023, pp. 7464–7475. doi: 10.1109/CVPR52729.2023.00721.

[41] Z. Teed and J. Deng, "RAFT: Recurrent All-Pairs Field Transforms for Optical Flow," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12347 LNCS, Springer, Cham, 2020, pp. 402–419. doi: 10.1007/978-3-030-58536-5_24.

[42] A. Vaswani *et al.*, "Attention is All you Need," in *Advances in Neural Information Processing Systems*, 2017, p. 11. Accessed: Feb. 17, 2026. [Online]. Available: https://arxiv.org/abs/1706.03762.

[43] A. Dosovitskiy *et al.*, "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale," in *ICLR 2021 - 9th International Conference on Learning Representations*, International Conference on Learning Representations, ICLR, Jun. 2021, p. 22. Accessed: Feb. 17, 2026. [Online]. Available: http://arxiv.org/abs/2010.11929.

[44] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *Proceedings of Machine Learning Research*, ML Research Press, Jan. 2021, pp. 10347–10357. Accessed: Feb. 17, 2026. [Online]. Available: http://arxiv.org/abs/2012.12877.

[45] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," in *7th International Conference on Learning Representations, ICLR 2019*, International Conference on Learning Representations, ICLR, Jan. 2019, p. 19. Accessed: Feb. 17, 2026. [Online]. Available: http://arxiv.org/abs/1711.05101.