From Video Understanding to Clinical Insight: Applying Hierarchical Interleaved Motion Encoding for Surgical Workflow Analysis

Vikram Mahajan
Department of Computer Science, University of Houston, Houston, TX, USA.
contactvikram@uh.edu

Aditya L. Gokhale
Department of Computer Science, University of New Hampshire, Durham, NH, USA.
aditya1999@unh.edu

Abstract

The translation of raw video data into clinically actionable insight represents a central challenge in modern surgical informatics. This paper examines the application of hierarchical interleaved motion encoding for surgical workflow analysis, a paradigm that combines multi-scale temporal abstraction with interleaved spatial-motion representations to capture the complex, non-linear dynamics of surgical procedures. Unlike conventional frame-level or single-stream approaches, hierarchical interleaved motion encoding decomposes video streams into multiple complementary motion cues, such as optical flow, temporal differences, and long-range feature correlations, and then interleaves them across hierarchical levels to preserve both fine-grained instrument interactions and global procedural context. We argue that this architecture offers significant structural advantages for surgical workflow analysis: it naturally handles long temporal dependencies, reduces redundant computation through scale-specific feature reuse, and enables robust performance across varied surgical settings. However, deploying such models in clinical infrastructures introduces trade-offs among computational efficiency, interpretability, data governance, and fairness. This paper provides a system-level analysis of these trade-offs, addressing the architectural choices, deployment strategies, data privacy considerations, and regulatory implications. We situate the hierarchical interleaved motion encoding approach within the broader landscape of video understanding and surgical AI, drawing comparisons to transformer-based and graph-based alternatives. We also discuss sustainability, robustness to domain shift, and the need for equitable model performance across diverse patient populations. The paper concludes with forward-looking recommendations for integrating such systems into clinical decision support frameworks while maintaining alignment with ethical and policy standards.

Keywords

surgical workflow analysis, video understanding, hierarchical motion encoding, clinical AI, infrastructure, governance, fairness.

1. Introduction

The operating room generates an enormous volume of video data through endoscopic cameras, room cameras, and wearable devices. This data, if systematically analyzed, can provide real-time feedback to surgeons, support post-operative debriefing, and enable large-scale studies of surgical technique variability. Yet the automatic interpretation of surgical video remains a

formidable problem because of the long duration of procedures, the high degree of intra-operative variability, and the need to recognize subtle transitions between phases that are often defined by multiple simultaneous cues: instrument presence, tissue deformation, and surgeon motion. Early approaches relied on frame-wise convolutional neural networks trained to classify individual still images into surgical phases, but these models lacked temporal context and frequently misclassified brief transitions. Recurrent architectures improved temporal modeling but struggled with the very long sequences typical of procedures such as laparoscopic cholecystectomy or robotic prostatectomy. More recently, transformer-based video models have shown promise, yet their quadratic complexity in input length poses severe computational challenges for deployment in real-time or near-real-time clinical environments.

The emerging paradigm of hierarchical interleaved motion encoding addresses these limitations by explicitly constructing multi-scale temporal representations that are interleaved across complementary motion modalities. This approach, as introduced in recent technical reports [15], builds on the insight that surgical video contains distinct motion patterns at different time scales: rapid instrument actions (on the order of milliseconds) and slower phase transitions (on the order of minutes). By encoding these scales separately and then interleaving the resulting features, the model can capture both fine-grained dynamics and long-range dependencies without the computational overhead of processing every frame at the highest resolution. This paper does not aim to present a specific algorithmic implementation; rather, it provides a systems-level analysis of the structural trade-offs, infrastructure requirements, and socio-technical implications of applying hierarchical interleaved motion encoding to surgical workflow analysis. We examine how this architecture interacts with clinical data governance frameworks, deployment constraints in resource-limited settings, and ethical requirements for fairness and robustness. We also draw comparisons to alternative architectures [1, 2, 3, 4] and discuss how the hierarchical interleaved design can be adapted to comply with emerging regulatory guidelines for AI-assisted medical devices.

2. Background and Related Work

Surgical workflow analysis has evolved from simple rule-based phase detection to deep learning models that leverage temporal and spatial features. Early work by Twinanda et al. [5] introduced a convolutional neural network combined with a hidden Markov model for phase recognition on the Cholec80 dataset, establishing a baseline that relied on frame-level features and temporal smoothing. Subsequent efforts incorporated long short-term memory networks [6] to capture longer-range dependencies, but these models often required careful tuning of sequence lengths and suffered from vanishing gradients when applied to videos spanning several hours. The introduction of two-stream architectures [1], which separately process RGB frames and optical flow, demonstrated that motion information is critical for recognizing actions. Later, three-dimensional convolutions [2] attempted to jointly model spatial and temporal dimensions, albeit at high computational cost. The advent of transformer models [3] for video understanding offered a way to model global temporal attention, but the self-attention mechanism’s quadratic cost in sequence length made it impractical for hour-long videos without aggressive downsampling or chunking.

Hierarchical approaches have been explored in various video domains. Feichtenhofer et al. [7] proposed a slow-fast network that processes video at two frame rates, thereby capturing both fine and coarse temporal cues. This concept was extended by the multi-scale transformer [8], which uses a pyramid of temporal resolutions. The hierarchical interleaved motion encoding framework [15] differs from these prior works by explicitly interleaving multiple motion

streams at each hierarchical level, rather than simply concatenating or aggregating features from different scales. This design choice allows the model to retain complementary motion information—such as flow magnitude, temporal gradients, and long-range feature correlations—throughout the network, rather than collapsing them at an early stage. The interleaving mechanism also facilitates gradient flow during training, enabling the model to learn from both short-term and long-term patterns without the need for auxiliary losses or curriculum learning strategies.

In the surgical domain, several benchmark datasets have been established, including Cholec80 [5], M2CAI [9], and EndoVis [10]. These datasets have enabled systematic comparisons across models. However, most existing work treats surgical video analysis as a supervised phase classification task, ignoring the rich structure of finer-grained activities such as tool usage, suture placement, and tissue handling. Recent research has begun to address action segmentation and instrument detection simultaneously [11], but these multitask models often require complex decoders and suffer from label imbalance. The hierarchical interleaved motion encoding approach offers a unified architecture that can be trained for both phase recognition and action segmentation without requiring multiple output heads, because the hierarchical representations naturally contain information at different granularities. This structural property has implications for system design, as it reduces the number of separate models that must be deployed and maintained in a clinical setting.

3. Hierarchical Interleaved Motion Encoding Framework

The core architectural principle of hierarchical interleaved motion encoding is the decomposition of video into multiple motion streams at each temporal scale, followed by the interleaved fusion of these streams across scales. At the lowest level, the model computes dense optical flow, frame differences, and feature-tracked motion vectors from short temporal windows. These streams are processed by lightweight convolutional encoders that produce multi-channel motion features. At the next hierarchical level, the temporal resolution is reduced by a factor, and the same streams are computed over longer windows using pooled or subsampled representations. Critically, the streams from adjacent levels are interleaved: the low-level motion features are combined with the mid-level aggregated features through learned gating or attention mechanisms, allowing the model to selectively propagate fine-grained information into coarser scales and vice versa. This interleaving is repeated across multiple levels, building a hierarchy where each level contains a fused representation of motion at its native resolution as well as contributions from the level below and above.

This architecture offers several structural advantages for surgical workflow analysis. First, it naturally handles the variable duration of surgical phases. A phase such as “dissection” may last several minutes, while a specific instrument motion within that phase may last only milliseconds. By preserving information at multiple scales, the model can simultaneously recognize the overall phase and the detailed action. Second, the interleaving mechanism reduces the curse of dimensionality: instead of processing all frames at full resolution, the model allocates computational resources according to the temporal scale. This makes real-time deployment feasible on existing hospital GPU servers or edge devices. Third, the design is inherently modular; individual motion streams can be replaced or augmented with domain-specific streams, such as depth motion or instrument segmentation masks, without retraining the entire hierarchy [12]. This flexibility is important for adapting to different surgical modalities, including laparoscopy, endoscopy, and robotic surgery, each of which exhibits distinct motion characteristics.

From an infrastructure perspective, the hierarchical architecture imposes specific requirements on data flow and memory management. Because each level depends on features computed at adjacent levels, the system must maintain a temporal buffer that stores intermediate representations. The size of this buffer is a critical design parameter: a larger buffer allows longer-range dependencies but increases memory footprint and latency. In a clinical deployment where video is streamed live from an endoscopic camera, the model must operate in an online fashion, processing frames as they arrive. The hierarchical interleaved design can be adapted to online processing by caching hierarchical features from previous time steps and updating them incrementally. However, this introduces trade-offs between latency and accuracy: if the model waits for sufficient temporal context at the coarsest level, it may introduce a delay of several seconds, which is acceptable for post-operative analysis but problematic for real-time decision support. Practical implementations often use a sliding window approach at each level, with the window size doubling at each coarser scale [13].

4. System Architecture and Infrastructure Considerations

Deploying a hierarchical interleaved motion encoding model for surgical workflow analysis requires a carefully designed system architecture that spans data acquisition, preprocessing, inference, and interpretation. The data acquisition pipeline must ingest high-resolution video streams from multiple cameras, synchronize them, and optionally compress them to reduce bandwidth. The preprocessing stage typically involves frame sampling, motion estimation (e.g., optical flow computation), and temporal chunking. Each of these steps introduces latency and computational overhead. Optical flow computation, for instance, is a dense per-pixel operation that can dominate the inference time if implemented naively. In practice, many deployments use lightweight flow approximations [14] or learned flow networks that can be executed on the same GPU as the main model. The hierarchical interleaved architecture can exploit the fact that optical flow is computed only at the finest temporal scale; at coarser scales, it can be approximated by temporally pooling the fine-scale flow, reducing the overall flow computation cost by an order of magnitude.

Another critical infrastructure consideration is data governance. Surgical video contains highly sensitive patient information, including faces, body parts, and identifiable anatomical features. Recording and transmitting such video off-premises for processing raises privacy and legal concerns under regulations like HIPAA in the United States and GDPR in Europe. Therefore, the inference pipeline must be deployable on local hospital infrastructure, often in a hybrid edge-cloud configuration. The hierarchical model's modularity facilitates on-premises deployment of the low-level stream processing (which is data-intensive and privacy-sensitive) with cloud-based aggregation and analysis of cross-institutional de-identified feature vectors. This federated approach reduces the risk of data leakage while still enabling large-scale learning across hospitals [15]. The interleaving mechanism must be designed to support such distributed processing by allowing each level to communicate only aggregated, non-reversible features across network boundaries.

Scalability is another challenge. A single hospital might record hundreds of surgical videos per week. Storing and processing all video at full resolution is prohibitive. The hierarchical interleaved architecture naturally enables a tiered storage strategy: only the highest-level aggregated representations are stored permanently, while low-level motion streams are discarded after inference. This not only reduces storage costs but also aligns with data minimization principles. Moreover, the model's computational footprint can be scaled by adjusting the number of hierarchy levels or the resolution of motion streams. In a resource-

constrained setting, a two-level hierarchy might be used, trading some accuracy for significantly lower latency and memory usage. This adaptability is essential for equity in global surgical AI, where hospitals in low-resource regions may lack high-end GPU clusters.

5. Deployment and Governance in Clinical Settings

The integration of a surgical workflow analysis system into clinical workflows demands careful governance to ensure that the system supports rather than disrupts surgical practice. One major governance issue is model validation and continuous monitoring. The hierarchical interleaved model, like all deep learning systems, is sensitive to domain shift: a model trained on laparoscopic cholecystectomy videos from a Western hospital may perform poorly on robotic sleeve gastrectomy in an Asian hospital due to differences in instrumentation, lighting, and surgeon technique. Robustness to such shifts can be improved by incorporating domain-invariant motion representations, but no model can eliminate the need for retraining or fine-tuning on local data. Therefore, a clinical deployment must include a governance structure that mandates periodic model evaluation on held-out local data and defines procedures for model updating. The hierarchical interleaved architecture supports incremental fine-tuning because individual stream encoders can be updated while preserving the interleaving weights, reducing the amount of retraining data required [16].

Interpretability is a related governance concern. Surgeons and hospital administrators require explanations for why a model predicted a certain phase transition or flagged an anomalous motion. The hierarchical interleaved approach offers inherent interpretability avenues: because each level is associated with a specific temporal scale, a clinician can examine which motion stream contributed most to a prediction at a given time. For example, if the model detects a transition from “clip application” to “cutting,” a surgeon can inspect the fine-scale optical flow stream to see the specific instrument movement that triggered the change. This level of granularity is more informative than a saliency map over RGB pixels. However, implementing such interpretability mechanisms requires additional visualization tools and user interfaces, which themselves must be validated against user comprehension and potential cognitive biases.

Data governance also intersects with model governance when considering patient consent and data reuse. Most institutional review boards approve retrospective video studies only if any patient-identifiable information is removed. The hierarchical interleaved model can be trained entirely on de-identified motion features derived from optical flow, which do not contain any anatomical detail. This makes it possible to share motion encodings across institutions for collaborative model training without sharing raw video. However, the question of whether such motion features are truly non-reversible remains open; adversarial reconstruction attacks could potentially recover coarse spatial information from aggregated motion vectors. Therefore, governance policies must also include technical safeguards such as noise injection or differential privacy applied to the motion streams [17].

6. Robustness, Fairness, and Sustainability

Robustness in surgical video analysis refers to the model’s ability to maintain performance under varying conditions, including camera occlusions, smoke, blurring, and instrument glare. The hierarchical interleaved design is inherently more robust than single-stream models because it relies on multiple motion cues. If one stream, such as optical flow, becomes unreliable due to motion blur, other streams like temporal differences or long-range feature correlations may still provide useful information. Furthermore, the interleaving mechanism

allows the model to down-weight corrupted streams through learned gating. However, robustness does not automatically translate to fairness. Fairness concerns arise when model performance degrades for certain patient groups due to differences in anatomy, surgical procedure variants, or equipment. For instance, a model trained primarily on videos of normal-weight patients may misclassify phases in obese patients because of altered tissue motion patterns [18]. To address this, training datasets must be carefully stratified, and the hierarchical architecture should be evaluated across demographic subgroups. The interleaving mechanism can be extended to include domain-adaptation modules that adjust stream weights based on known covariates like body mass index or procedure type.

Sustainability is another dimension that receives insufficient attention in surgical AI research. Training large video models consumes substantial energy, contributing to carbon emissions. The hierarchical interleaved architecture, by design, is more efficient than monolithic transformers because it processes coarser scales at lower resolutions and compute. Yet the training still requires hours of GPU usage for datasets of hundreds of videos. The sustainability trade-off must be considered alongside performance gains. One approach is to pre-train the hierarchical streams on large, publicly available action recognition datasets and then fine-tune only on surgical data, reducing the total training compute. Additionally, model quantization and pruning can reduce the inference energy footprint, making deployment in low-power edge devices feasible. However, these efficiency improvements must not compromise the model's accuracy, particularly for safety-critical phase detection. A careful cost-benefit analysis is needed for each deployment context.

7. Policy Implications and Regulatory Challenges

The clinical deployment of hierarchical interleaved motion encoding for surgical workflow analysis raises several regulatory and policy issues. In the United States, the Food and Drug Administration has issued guidance for software as a medical device, including AI-based decision support tools. A system that provides real-time phase identification or alerts for critical events would likely be classified as a moderate-risk device, requiring premarket submission. The hierarchical model's innate modularity could facilitate regulatory approval by allowing each stream to be validated independently; for example, the optical flow stream could be certified as a general motion capture module, and the overall phase classification module could be certified as a separate device. However, the interleaving mechanism creates dependencies between modules, complicating the regulatory pathway. Policy frameworks must evolve to accommodate these new architectures, possibly through a "split-validation" approach where each hierarchical level is validated separately and then the combined system is tested for end-to-end accuracy.

Another policy consideration is the liability for incorrect predictions. If a surgical workflow analysis model fails to detect an intra-operative complication or mislabels a critical phase, who is responsible? The surgeon, the hospital, the model developer, or the data provider? The hierarchical interleaved model's interpretability features could help allocate responsibility by identifying which stream contributed to the error, but legal frameworks are not yet equipped to handle such granular attribution. International standards for AI in healthcare, such as those being developed by the International Organization for Standardization and the Institute of Electrical and Electronics Engineers, should incorporate hierarchical model testing protocols. Furthermore, policies governing data sharing across institutions must be harmonized to enable the collaborative training and continuous improvement of these models without violating privacy regulations. The European Health Data Space initiative provides a promising model

for such data governance, but its implementation requires alignment with the architectural requirements of hierarchical systems. Finally, the sustainability implications of large-scale video processing should be accounted for in hospital green IT policies, encouraging the adoption of efficient architectures like hierarchical interleaving over less efficient alternatives.

8. Conclusion

The application of hierarchical interleaved motion encoding to surgical workflow analysis represents a significant step forward in bridging video understanding and clinical insight. This architecture, by decomposing motion into complementary streams at multiple temporal scales and interleaving them, addresses the fundamental challenge of capturing both fine-grained actions and long-range procedural context. The system-level analysis presented in this paper reveals that this approach offers structural advantages in computational efficiency, modularity, and interpretability that are well-suited for clinical deployment. However, these advantages come with trade-offs in memory management, latency, and the need for robust governance frameworks. The successful integration of such models into real surgical environments will depend not only on algorithmic improvements but also on careful infrastructure design, data privacy safeguards, fairness validation, and alignment with evolving regulatory standards. As surgical video datasets grow in size and diversity, the hierarchical interleaved paradigm provides a scalable and adaptable foundation for building intelligent systems that can augment surgical decision-making, improve training, and ultimately enhance patient outcomes. Future research should focus on extending the framework to multi-modal inputs, such as audio and physiological signals, and on developing standardized evaluation benchmarks that account for the socio-technical dimensions discussed here.

References

1. Simonyan, K., & Zisserman, A. (2014). Two-stream convolutional networks for action recognition in videos. Advances in Neural Information Processing Systems, 27.

2. Tran, D., Bourdev, L., Fergus, R., Torresani, L., & Paluri, M. (2015). Learning spatiotemporal features with 3D convolutional networks. Proceedings of the IEEE International Conference on Computer Vision, 4489–4497.

3. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30.

4. Carreira, J., & Zisserman, A. (2017). Quo vadis, action recognition? A new model and the Kinetics dataset. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 6299–6308.

5. Twinanda, A. P., Shehata, S., Mutter, D., Marescaux, J., de Mathelin, M., & Padoy, N. (2017). EndoNet: A deep architecture for recognition tasks on laparoscopic videos. IEEE Transactions on Medical Imaging, 36(1), 86–97.

6. Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural Computation, 9(8), 1735–1780.

7. Feichtenhofer, C., Fan, H., Malik, J., & He, K. (2019). SlowFast networks for video recognition. Proceedings of the IEEE International Conference on Computer Vision, 6202–6211.

8. Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., & Schmid, C. (2021). ViViT: A video vision transformer. Proceedings of the IEEE International Conference on Computer Vision, 6836–6846.

9. Yengera, G., Mutter, D., Marescaux, J., & Padoy, N. (2018). Less is more: Surgical phase recognition with minimal annotations using temporal knowledge distillation. International Conference on Medical Image Computing and Computer-Assisted Intervention, 371–379.

10. Bodenstedt, S., Wagner, M., Katic, D., & Dillmann, R. (2017). EndoVis: A common platform for surgical training and evaluation. International Journal of Computer Assisted Radiology and Surgery, 12(1), 1–10.

11. Jin, Y., Li, Q., & Dou, Q. (2020). Multi-task learning for surgical instrument segmentation and phase recognition. Medical Image Computing and Computer Assisted Intervention, 12263, 148–158.

12. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 770–778.

13. Wu, C., & Krähenbühl, P. (2021). Towards long-form video understanding. arXiv preprint arXiv:2106.08986.

14. Dosovitskiy, A., Fischer, P., Ilg, E., Hausser, P., Hazirbas, C., Golkov, V., ... & Brox, T. (2015). FlowNet: Learning optical flow with convolutional networks. Proceedings of the IEEE International Conference on Computer Vision, 2758–2766.

15. Jin, H., Yi, H., Zhao, W., Luo, J., Ye, S., Guan, Z., ... & Yu, T. (2026). HY-Himmel Technical Report: Hierarchical Interleaved Multi-stream Motion Encoding for Long Video Understanding. arXiv preprint arXiv:2605.08158.

16. Wang, M., & Deng, W. (2018). Deep visual domain adaptation: A survey. Neurocomputing, 312, 135–153.

17. Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, 308–318.

18. Hashimoto, D. A., Rosman, G., Rus, D., & Meireles, O. R. (2018). Artificial intelligence in surgery: Promises and perils. Annals of Surgery, 268(1), 70–76.

19. Topol, E. J. (2019). High-performance medicine: The convergence of human and artificial intelligence. Nature Medicine, 25(1), 44–56.

20. Vokinger, K. N., & Gasser, U. (2021). Regulating AI in medicine in the United States and Europe. Nature Medicine, 27(1), 35–37.