Instruction-Guided Video Representation Learning for Complex Procedure Understanding

Elliot Reed
School of Information Technology, University of Cincinnati, Cincinnati, OH, USA.
ereed@uc.edu

Abstract

The comprehension of complex, multi-step procedures from video data represents a critical frontier in artificial intelligence, with profound implications for autonomous systems, surgical robotics, industrial automation, and instructional technology. Traditional video representation learning has largely focused on action recognition or short-term temporal dynamics, yet the understanding of long-horizon, hierarchically structured procedures demands a fundamentally different representational paradigm. This paper introduces a framework for instruction-guided video representation learning that leverages natural language instructions as a structured supervisory signal to organize and interpret temporal sequences of procedural actions. We argue that the integration of linguistic instruction streams with visual perception enables the formation of hierarchical, goal-oriented representations that are essential for robust procedure understanding. The paper examines the architectural trade-offs between end-to-end learned embeddings and modular, instruction-conditioned feature spaces, analyzing how these choices impact generalizability, computational efficiency, and interpretability. We explore the governance and infrastructure implications of deploying such systems in high-stakes environments, including the need for auditability, fairness in instructional content, and robustness to distributional shifts. Sustainability considerations are addressed through the lens of computational cost versus representational fidelity. Cross-domain comparisons between surgical video understanding, cooking procedure recognition, and industrial assembly verification illustrate the structural invariants of procedural knowledge. The paper further discusses the policy and regulatory challenges that arise when instruction-guided systems are embedded in socio-technical infrastructures, particularly regarding accountability for procedural errors. By synthesizing insights from computer vision, natural language processing, cognitive science, and systems engineering, this research provides a comprehensive analytical framework for the next generation of video understanding systems that must operate reliably in complex, real-world procedural environments.

Keywords

instruction-guided learning, video representation, procedure understanding, hierarchical temporal modeling, socio-technical systems.

1. Introduction

The proliferation of video data across domains such as medical education, manufacturing, and autonomous navigation has created an urgent demand for artificial intelligence systems capable of understanding not merely what is happening in a video, but how a sequence of actions unfolds toward a defined goal. While substantial progress has been made in video action recognition and temporal action localization, these approaches typically treat actions as isolated events or short-range temporal patterns [1]. Complex procedures, by contrast, involve long-range dependencies, hierarchical task decompositions, and conditional branching that

cannot be captured by models optimized for discriminative classification of atomic actions. The challenge is compounded by the fact that procedures are inherently goal-directed and often accompanied by explicit or implicit instructional guidance, whether in the form of written manuals, spoken commands, or demonstrated sequences [2].

Instruction-guided video representation learning emerges as a response to these limitations, proposing that language instructions can serve as an organizing principle for learning visual representations that are both temporally structured and semantically grounded. This approach draws on the insight that human procedural learning is heavily mediated by verbal or textual instructions, which provide a scaffold for segmenting continuous experience into discrete, meaningful steps [3]. By aligning video representations with instructional text, models can learn to attend to task-relevant visual features, ignore irrelevant variations, and generalize across different environments that share the same procedural logic. The resulting representations are not merely descriptive but prescriptive, encoding the normative structure of how a procedure should be executed [4].

This paper investigates the systems-level implications of adopting instruction-guided video representation learning for complex procedure understanding. Rather than focusing on specific model architectures or benchmark evaluations, we examine the broader structural trade-offs that arise when such systems are designed, deployed, and governed. These trade-offs span the technical dimensions of representation granularity and temporal abstraction, the infrastructural requirements for data collection and annotation at scale, and the socio-technical challenges of ensuring fairness, accountability, and robustness in procedural AI systems. The analysis is informed by case studies from surgical video analysis, where procedural precision is paramount; cooking and instructional video understanding, where variability in execution is high; and industrial assembly verification, where consistency and error detection are critical [5][6].

A central argument of this paper is that the effectiveness of instruction-guided representations depends critically on the alignment between the level of abstraction in the instructional text and the temporal granularity of the video signal. Mismatches between instruction and visual dynamics can lead to brittle representations that fail to generalize or, worse, produce misleading procedural interpretations. We further contend that the governance of such systems must account for the fact that instructional content itself can embed biases, errors, or cultural assumptions that propagate through the learned representations into downstream decisions [7]. The paper concludes by outlining a research agenda for building instruction-guided video systems that are not only technically capable but also socially responsible and infrastructurally sustainable.

2. Architectural Foundations of Instruction-Guided Video Representation

The design of an instruction-guided video representation system involves a set of interrelated architectural choices that fundamentally shape its capacity for procedure understanding. At the highest level, these choices concern the modality fusion strategy, the temporal modeling architecture, and the representation learning objective. Early approaches to video-language understanding employed late fusion techniques, where visual features and text features were independently extracted and combined only at the decision layer [8]. While computationally simple, such architectures fail to capture the fine-grained alignment between instructional phrases and specific visual segments, which is essential for procedure understanding where a single instruction may correspond to a temporally extended action sequence.

More recent architectures adopt cross-modal attention mechanisms that allow the model to dynamically align video frames or segments with instruction tokens. These models often employ a transformer-based encoder that processes both visual and textual inputs in a shared representational space, enabling bidirectional information flow between modalities [9]. The hierarchical nature of procedures, however, demands additional architectural considerations. A single procedural step, such as "suture the incision," may encompass multiple sub-actions that are visually distinct but semantically unified under the instruction. To capture this hierarchy, some architectures introduce multi-stream motion encoding that operates at different temporal resolutions, allowing the model to represent both fine-grained motion dynamics and coarse-grained procedural phases [10]. This hierarchical interleaving of motion streams is particularly effective for long video sequences where actions unfold over minutes or hours, as it mitigates the vanishing gradient problem inherent in recurrent architectures while preserving temporal context.

Another critical architectural dimension is the choice between disentangled and entangled representation learning. In a disentangled approach, the visual representation is decomposed into components that are explicitly conditioned on instructional content and components that capture environment-specific variations. This separation facilitates transfer learning, as the instruction-conditioned component can be reused across different video domains that share the same procedural instructions [11]. Conversely, entangled representations, where instruction and visual information are fused into a single embedding, may achieve higher accuracy on specific tasks but suffer from poor generalization when the visual environment or instructional phrasing changes. The trade-off between specificity and generality must be carefully calibrated based on the deployment context. In surgical settings, where the visual environment is relatively controlled and the instructions are standardized, entangled representations may offer superior performance. In industrial assembly, where lighting conditions, camera angles, and component variations are unpredictable, disentangled representations provide greater robustness [12].

The representation learning objective itself is a subject of active investigation. Contrastive learning objectives, which pull together visual and textual representations of the same procedure while pushing apart representations of different procedures, have shown promise for learning semantically meaningful embeddings [13]. However, contrastive methods require careful negative sampling strategies to avoid trivial solutions, particularly when procedures share sub-steps. Alternative objectives based on temporal ordering prediction or masked instruction modeling encourage the model to learn the sequential logic of procedures, which is essential for understanding not only what actions occur but also when they should occur relative to one another [14]. The selection of objective function has direct implications for the types of procedural errors the system can detect. A contrastive objective may be sufficient for identifying whether a procedure has been executed correctly in a global sense, but a temporal ordering objective is necessary for detecting out-of-order steps, a common source of procedural failure.

3. Structural Trade-offs in Representation Granularity and Temporal Abstraction

The efficacy of instruction-guided video representation is heavily influenced by the granularity at which both instructions and video are segmented and aligned. This granularity is not a free parameter but is constrained by the nature of the procedure, the availability of annotated data, and the computational budget. At one extreme, fine-grained alignment at the level of individual video frames and instruction words offers high resolution but is

computationally prohibitive for long videos and often leads to overfitting to spurious correlations between specific visual features and lexical items [15]. At the other extreme, coarse-grained alignment at the level of entire procedure steps and video segments may miss critical sub-actions that are essential for error detection or skill assessment.

A structural trade-off exists between representational fidelity and computational tractability. Hierarchical models that operate at multiple temporal scales attempt to navigate this trade-off by learning representations at the level of atomic actions, sub-steps, and full procedures, with each level conditioned on the instructional context appropriate to that scale [16]. For example, the instruction "prepare the patient for surgery" may be associated with a high-level visual segment spanning several minutes, while the sub-instruction "apply antiseptic" aligns with a shorter, more specific visual segment. The challenge lies in learning the hierarchical decomposition automatically, as manual annotation of sub-step boundaries is labor-intensive and domain-specific. Weakly supervised methods that leverage instructional text as a source of implicit structure have been proposed, but they often rely on assumptions about the consistency of instruction-video alignment that may not hold across diverse procedural domains [17].

Another important trade-off concerns the abstraction level of the instructional text. Procedural instructions can range from highly specific, step-by-step commands to abstract, goal-oriented descriptions. Representations learned from specific instructions tend to be more precise but less transferable, as they encode fine-grained visual details that may not generalize to different tools, environments, or user skill levels. Representations learned from abstract instructions are more robust to visual variation but may lack the granularity needed for detailed error analysis [18]. In practice, a multi-level instruction representation that combines both specific and abstract descriptions can provide a more complete procedural understanding, but this requires additional annotation effort and more complex fusion mechanisms.

The temporal abstraction problem is further complicated by the presence of procedural variation. Different individuals may perform the same procedure in slightly different orders, with different pacing, or with different tool usage. An instruction-guided representation must be invariant to such permissible variations while remaining sensitive to impermissible deviations that constitute errors. Achieving this balance requires the representation to capture the normative structure of the procedure, which is often implicitly encoded in the instructional text rather than explicitly stated [19]. For instance, the instruction "close the incision" implies a sequence of sub-actions that are standard but not always enumerated. The representation must therefore learn to infer the expected sub-steps from the instructional context and the visual evidence, a capability that remains an open challenge in the field.

4. Infrastructure, Deployment, and Sustainability Considerations

Deploying instruction-guided video representation systems at scale requires substantial infrastructural investment across the data pipeline, model serving, and monitoring layers. The data infrastructure must support the collection, annotation, and storage of paired video-instruction datasets that capture the diversity of procedural execution. Unlike standard video classification datasets, procedural datasets require temporal annotations that align instructional steps with video timestamps, a process that is time-consuming and subject to annotator disagreement [20]. Moreover, procedures in domains such as surgery or manufacturing are often protected by privacy or proprietary concerns, necessitating secure data sharing frameworks and de-identification protocols. The governance of these datasets must address questions of consent, data sovereignty, and representational fairness, particularly

when procedures are performed by diverse populations whose variations must be accurately captured to avoid algorithmic bias [21].

The computational infrastructure for training and inference presents significant sustainability challenges. Large-scale transformer models that process both video and text modalities require substantial energy consumption, and the trend toward larger models for improved accuracy exacerbates this issue. The hierarchical multi-stream architectures discussed earlier, while effective for long video understanding, further increase computational demands due to the parallel processing of multiple temporal resolutions [10]. Sustainable deployment strategies include model distillation, where a smaller student model is trained to mimic the behavior of a larger teacher model, and quantization techniques that reduce the precision of model weights without significant performance degradation. Additionally, edge deployment for real-time procedure monitoring, such as in surgical operating rooms or assembly lines, requires model compression and hardware acceleration to meet latency constraints while maintaining representational quality [22].

Monitoring and maintenance of deployed systems introduce another layer of infrastructural complexity. Procedural environments evolve over time due to changes in equipment, protocols, or user practices. An instruction-guided representation that was trained on data from one generation of surgical instruments may fail to generalize to a new instrument that has different visual appearance but identical functional role. Continuous learning mechanisms that update the representation based on new data without catastrophic forgetting are essential for long-term robustness [23]. However, such mechanisms raise governance challenges regarding version control, validation, and certification, particularly in regulated domains where changes to an AI system may require regulatory re-approval. The trade-off between adaptability and stability must be managed through careful system design and institutional oversight.

5. Governance, Fairness, and Policy Implications

The integration of instruction-guided video representation into socio-technical systems raises profound governance questions that extend beyond technical performance. One central concern is the accountability for procedural errors when the AI system is used to monitor, guide, or assess human performance. If a surgical assistant system based on instruction-guided representations fails to detect a critical procedural deviation, who is responsible: the system developer, the hospital, the surgeon, or the instruction author? Existing liability frameworks are ill-equipped to handle the distributed agency inherent in such systems, where the instructional content, the learned representation, and the human operator all contribute to the final outcome [24]. A governance framework must establish clear lines of responsibility, including mechanisms for auditing the representation's behavior and for contesting its outputs.

Fairness considerations are particularly salient in instruction-guided systems because the instructional content itself can encode cultural, linguistic, or professional biases. Instructions written for a Western surgical context may assume specific tools, anatomical terminology, or procedural norms that do not hold in other settings, leading to systematic misalignment between the learned representation and the actual procedure [25]. Furthermore, if the training data over-represents certain demographic groups or skill levels, the resulting representation may perform poorly for underrepresented groups, exacerbating disparities in access to quality procedural guidance or assessment. Mitigating these biases requires not only diverse training data but also participatory design processes that involve stakeholders from diverse procedural communities in the development of instructional content and the validation of system outputs.

Policy frameworks for instruction-guided video systems must also address transparency and explainability. In high-stakes procedural environments, users need to understand why a system flagged a particular action as an error or why it provided a specific instructional recommendation. The hierarchical and multimodal nature of instruction-guided representations makes explainability particularly challenging, as the reasoning may involve interactions between visual features and instructional text across multiple temporal scales [26]. Regulatory requirements for explainability, such as those emerging in the European Union's AI Act, will necessitate the development of interpretability tools that can trace system outputs back to specific instructional phrases and visual segments. These tools must be designed in consultation with domain experts to ensure that the explanations are meaningful and actionable for end-users.

6. Cross-Domain Analysis and Structural Invariants

Comparing instruction-guided video representation across different procedural domains reveals both domain-specific challenges and structural invariants that can inform general-purpose system design. In surgical video analysis, procedures are highly standardized, with well-defined steps and strict protocols for error avoidance. The instructional text is typically detailed and unambiguous, and the visual environment is controlled, albeit with variations due to patient anatomy and surgeon technique. The primary challenge in this domain is achieving the high temporal precision needed to detect subtle deviations that could lead to adverse outcomes [5]. The hierarchical interleaved motion encoding approach is particularly well-suited here, as it can capture both the macroscopic phases of a surgery and the microscopic hand movements that constitute individual sutures [10].

In contrast, cooking procedure understanding involves high variability in ingredients, equipment, and execution style. The same recipe may be performed differently by different cooks, and the instructional text often leaves room for interpretation. Instruction-guided representations in this domain must be robust to visual variation while capturing the functional equivalence of different actions, such as "chopping" with a knife versus a food processor [6]. The structural invariant across these domains is the presence of a hierarchical task decomposition, where high-level goals decompose into sub-goals and atomic actions. This decomposition is reflected in the instructional text, whether explicitly structured as numbered steps or implicitly embedded in procedural language. A general-purpose instruction-guided representation system must therefore be capable of learning this hierarchical structure from the instructional signal, regardless of the domain.

Industrial assembly verification presents yet another set of constraints, including the need for real-time feedback, the presence of repetitive actions, and the requirement for high precision in part placement and fastening. The instructional text in this domain is often generated automatically from computer-aided design models, providing a structured but potentially noisy signal [12]. The key challenge here is aligning the instruction-guided representation with the physical constraints of the assembly process, such as the order-dependent nature of part insertion. The structural invariant across all three domains is the critical role of temporal ordering, which distinguishes procedural understanding from mere action recognition. Instruction-guided representations that explicitly encode temporal dependencies, whether through attention mechanisms or recurrent architectures, are better positioned to generalize across domains than those that treat time as a passive dimension.

7. Conclusion

This paper has presented a comprehensive analysis of instruction-guided video representation learning for complex procedure understanding, emphasizing the systems-level trade-offs, infrastructural requirements, and governance challenges that accompany the deployment of such technology. We have argued that the alignment between instructional text and visual dynamics is the central design problem, with implications for architectural choices, representation granularity, and generalization capability. The hierarchical interleaving of motion streams at multiple temporal resolutions represents a promising architectural direction, particularly for long and complex procedures where both fine-grained actions and coarse-grained phases must be captured [10]. However, the effectiveness of this approach depends on the availability of high-quality instructional data, the computational resources for training and inference, and the governance frameworks that ensure accountability and fairness.

The cross-domain analysis revealed that while specific challenges vary across surgical, culinary, and industrial contexts, the underlying structural invariants of hierarchical task decomposition and temporal ordering provide a foundation for general-purpose system design. Future research should focus on developing self-supervised methods for learning hierarchical procedure structure from unlabeled video and instructional text, reducing the dependence on costly manual annotations. Additionally, the integration of causal reasoning into instruction-guided representations could enhance their ability to distinguish between permissible procedural variation and genuine errors, a capability that is essential for high-stakes applications.

Finally, the governance and policy dimensions of instruction-guided video systems demand urgent attention from the research community, industry, and regulators. As these systems become embedded in critical infrastructure, from operating rooms to factory floors, the need for transparent, fair, and accountable AI becomes paramount. The development of audit trails, bias mitigation protocols, and explainability tools must proceed in parallel with technical advances, ensuring that the benefits of instruction-guided video representation are realized without compromising human autonomy or safety. The path forward requires interdisciplinary collaboration that bridges computer vision, natural language processing, cognitive science, law, and ethics, building systems that not only understand procedures but also respect the human contexts in which those procedures are performed.

References

1. 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.

2. Shao, J., Wang, J., Chang, K. W., & Lim, J. J. (2020). Fine-grained procedural understanding from video and text. Proceedings of the European Conference on Computer Vision, 123-139.

3. Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. Behavioral and Brain Sciences, 40, e253.

4. Miech, A., Zhukov, D., Alayrac, J. B., Tapaswi, M., Laptev, I., & Sivic, J. (2019). HowTo100M: Learning a text-video embedding by watching hundred million narrated video clips. Proceedings of the IEEE International Conference on Computer Vision, 2630-2640.

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. Malmaud, J., Huang, J., Rathod, V., Johnston, N., Rabinovich, A., & Murphy, K. (2015). What's cookin'? Interpreting cooking videos using text, speech and vision. Proceedings of the North American Chapter of the Association for Computational Linguistics, 143-152.

7. Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. Proceedings of the Conference on Fairness, Accountability, and Transparency, 77-91.

8. Xu, J., Mei, T., Yao, T., & Rui, Y. (2016). MSR-VTT: A large video description dataset for bridging video and language. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 5288-5296.

9. 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, 5998-6008.

10. 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.

11. Denton, E., & Birodkar, V. (2017). Unsupervised learning of disentangled representations from video. Advances in Neural Information Processing Systems, 30, 4414-4423.

12. Paul, G., & Newman, P. (2010). FAB-MAP: Probabilistic localization and mapping in the space of appearance. The International Journal of Robotics Research, 29(6), 647-665.

13. Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. Proceedings of the International Conference on Machine Learning, 1597-1607.

14. Xu, H., Ghosh, G., Huang, P. Y., Okhonko, D., Aghajanyan, A., Metze, F., ... & Zettlemoyer, L. (2021). VideoCLIP: Contrastive pre-training for zero-shot video-text understanding. Proceedings of the Conference on Empirical Methods in Natural Language Processing, 6787-6800.

15. Arandjelovic, R., & Zisserman, A. (2017). Look, listen and learn. Proceedings of the IEEE International Conference on Computer Vision, 609-617.

16. Todorovic, S., & Nechyba, M. C. (2005). A vision system for intelligent mission profiles of micro air vehicles. IEEE Transactions on Vehicular Technology, 54(5), 1713-1726.

17. Alayrac, J. B., Recasens, A., Schneider, R., Arandjelovic, R., Ramapuram, J., De Fauw, J., ... & Zisserman, A. (2020). Self-supervised multimodal versatile networks. Advances in Neural Information Processing Systems, 33, 25-37.

18. Goyal, Y., Khot, T., Summers-Stay, D., Batra, D., & Parikh, D. (2017). Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 6904-6913.

19. Zacks, J. M., & Tversky, B. (2001). Event structure in perception and conception. Psychological Bulletin, 127(1), 3-21.

20. Sigurdsson, G. A., Varol, G., Wang, X., Farhadi, A., Laptev, I., & Gupta, A. (2016). Hollywood in homes: Crowdsourcing data collection for activity understanding. Proceedings of the European Conference on Computer Vision, 510-526.

21. Raji, I. D., Gebru, T., Mitchell, M., Buolamwini, J., Lee, J., & Denton, E. (2020). Saving face: Investigating the ethical concerns of facial recognition auditing. Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 145-151.

22. Han, S., Mao, H., & Dally, W. J. (2016). Deep compression: Compressing deep neural networks with pruning, trained quantization and Huffman coding. International Conference on Learning Representations.

23. Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., ... & Hadsell, R. (2017). Overcoming catastrophic forgetting in neural networks. Proceedings of the National Academy of Sciences, 114(13), 3521-3526.

24. Selbst, A. D., Boyd, D., Friedler, S. A., Venkatasubramanian, S., & Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. Proceedings of the Conference on Fairness, Accountability, and Transparency, 59-68.

25. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? Proceedings of the ACM Conference on Fairness, Accountability, and Transparency, 610-623.

26. Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.