Prompt Engineering and Video Segmentation Strategies for VLM-Based Action Recognition in Prefabrication Workflows

Chien-Wen Chen[1], Ren-Jie Li[1], Liang-Ting Tsai[2], Cheng-Hsuan Yang[3], Meng-Han Tsai [1]

[1] National Taiwan University of Science and Technology, Taipei, Taiwan
[2] University of Alberta, Edmonton, Alberta, Canada
[3] RoBIM Technologies Inc., Edmonton, Alberta, Canada
menghan@mail.ntust.edu.tw

Abstract –

Automated productivity monitoring and robotic learning in prefabrication are hindered by the scarcity of structured demonstration data. To address this, this study investigates a Vision-Language Model (VLM) workflow for extracting engineering logs from wall panel assembly videos. We conducted a comparative empirical evaluation using a 2 × 3 factorial design to evaluate the interaction between video processing strategies (Continuous vs. Segmented) and prompt complexity. Results indicate that relying on internal attention mechanisms for long videos leads to significant temporal hallucinations. In contrast, the proposed "Segment-and-Caption" framework, combining physical temporal segmentation (visual constraint) with domain-specific prompting (semantic constraint), achieved the highest reliability, yielding a Duration-weighted Accuracy of 65.44% and a Macro-F1 score of 0.547 to account for class imbalance. Crucially, we identify a fundamental trade-off: while this "stateless" segmentation eliminates attention drift, it introduces an "Identity Gap" that precludes consistent worker tracking. This work establishes that decoupling visual context from temporal search is a critical strategy for reliable VLM-based monitoring, providing a validated baseline for converting unstructured footage into machine-readable data for downstream robotic planning.

Keywords –

Prefabrication; Prompt Engineering; Vision-Language Models; Action Recognition

1 Introduction

1.1 Industrial Background and Practical Motivation

Prefabrication relies heavily on synchronized assembly-line workflows to enhance efficiency [1, 2]. Despite this industrial shift, core tasks like wall panel assembly remain labor-intensive, with essential actions (e.g., fastening, alignment) performed manually due to automation limitations [3, 4]. Consequently, accurate monitoring is essential. However, current practices relying on manual supervision or passive CCTV are labor-intensive and fail to generate the structured, machine-readable data required for productivity profiling and human-robot collaboration [5].

1.2 Limitations of Existing Automation and Vision-Based Approaches

While sensor-based technologies like RFID track location effectively [6], they require intrusive instrumentation and lack the semantic context of specific assembly actions. Computer vision offers a non-intrusive alternative, yet conventional task-specific models require extensive training data and struggle to generalize across diverse site conditions [7]. Even emerging open-vocabulary models often underperform specialized detectors in identifying domain-specific elements [8]. Moreover, scaling these tools to real-world scenarios remains challenging due to robustness issues [9] and the difficulty of transforming raw visuals into digital twin-ready data [10].

1.3 Vision-Language Models for Construction Video Understanding: Opportunities and Challenges

Vision-Language Models (VLMs) offer a promising alternative by jointly reasoning over visual and natural language inputs. Unlike task-specific models, VLMs facilitate flexible, zero-shot activity understanding through text descriptions, eliminating the need for massive annotated datasets [11, 12]. Recent studies demonstrate that ontology-based prompting can guide VLMs to interpret construction images [13]. Theoretically, this allows for generating detailed activity logs directly from visual data, offering a pathway to automated productivity analysis.

1.4 Research Gap and Study Objectives

Despite their potential, applying VLMs to long, untrimmed prefabrication videos presents challenges. Models relying on internal attention often suffer from temporal hallucinations, inaccurately timestamping actions [14]. Furthermore, VLM performance is highly sensitive to prompt engineering [13]. While Video-LLMs are evolving, the specific interaction between prompt formulation and video segmentation strategies remains underexplored. To bridge this gap, this study provides initial empirical evidence by employing a 2 × 3 factorial design to evaluate how physical segmentation and domain-informed prompting jointly impact VLM performance and mitigate temporal hallucinations.

2 Methodology

2.1 Proposed Approach

We propose the "Segment-and-Caption" framework to strictly constrain VLM inference (**Figure 1**). As illustrated, the **baseline (top path)** applies temporal prompting directly to the raw, continuous video, relying on internal attention mechanisms, a process often leading to contextual hallucinations due to attention drift. In contrast, our **proposed method (bottom path)** introduces a modular decomposition to isolate temporal constraints. Specifically, the workflow enforces two primary mechanisms:

Physical Temporal Segmentation slices the footage into discrete clips based on ground truth timestamps. This "hard constraint" effectively renders the input stateless, limiting the model's focus to visual evidence within a specific time window. Complementing this visual isolation, Domain-Informed Prompting acts as a "semantic constraint." By injecting a construction knowledge into the prompt logic, we prevent generic descriptions and ensure outputs remain structurally consistent with engineering standards.

2.2 Experimental Design

To evaluate reliability, we adopted a 2 × 3 factorial design intersecting Input Strategy and Prompt Complexity. The Input Strategy compares the full-length Continuous baseline against the Segmented Input strategy, where action events are physically isolated to remove the burden of temporal search. Superimposed on these are three prompt levels: (1) P-Natural (naive instructions), (2) P-Syntactic (employing structural delimiters like XML to optimize parsing and enforce CSV outputs), and (3) P-Domain (integrating the construction knowledge). This progression allows us to determine whether improvements stem from formatting rules or domain-specific knowledge.

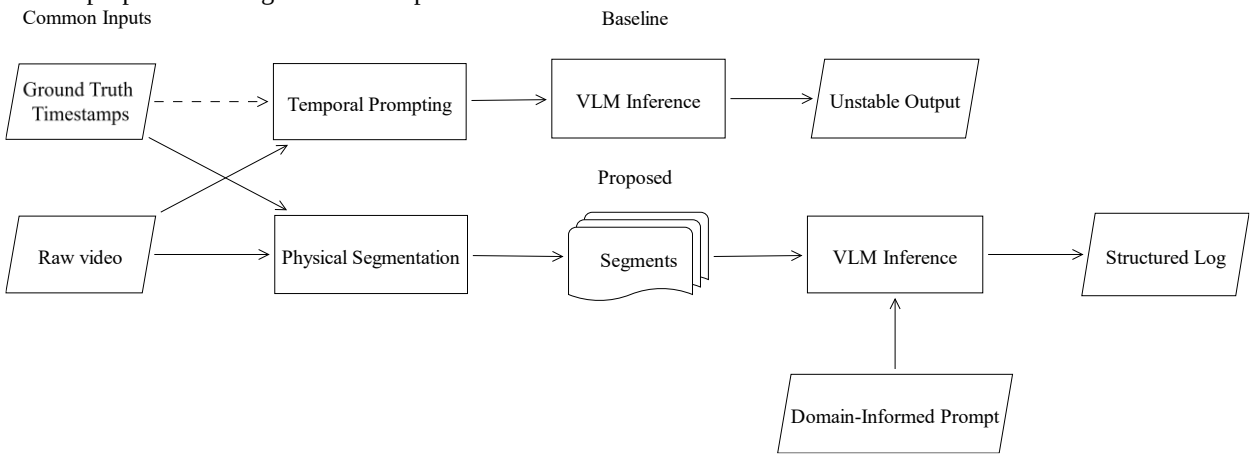

Figure 1. Workflow comparison between the Continuous Baseline and the Proposed Framework

3 Experimental Results and Analysis

3.1 Experimental Setup

The experiment utilized a high-density collaboration sequence (Duration: 6:57, 720p, 60fps) captured from a Canadian prefabrication factory, characterized by frequent occlusion and complex multi-worker interactions. The selection of this specific duration was governed by the native technical constraints of the Gemini 3 Pro web-based interface, which imposes a 2GB file upload limit. At approximately 1.87GB, this sequence represents the maximum continuous temporal context currently supportable by the platform's native multimodal pipeline, providing rigorous evaluation context for temporal stability without exceeding ingestion boundaries.

To simulate an open-world industrial monitoring scenario and ensure dataset transparency, the VLM was prompted with a construction knowledge comprising 8 task categories: *prepping*, *waiting*, *handling*, *fastening*, *process*, *measuring*, *fixing*, and *unsure*. While the ground-truth (GT) sequence contained 27 instances distributed across 4 active classes (*fastening*, *measuring*, *prepping*, *waiting*), the remaining 4 categories were treated as distractor classes to evaluate the model's robustness against over-classification.

All evaluations were performed using the Gemini 3 Pro (Web Version) with system-level decoding parameters (e.g., Temperature, Top-K, and Top-P) maintained at their default platform settings to ensure a standardized baseline and eliminate manual tuning bias. For the Continuous Input baseline, we fed the maximum possible context length directly into the Gemini 3 Pro web-based interface. Unlike API-based approaches that allow for programmatic chunking, this setup meant we were bounded by the platform's native file upload limits. Conversely, the Segmented Input strategy involved slicing this sequence into 27 discrete clips based on ground truth timestamps, isolating individual action events to test the efficacy of our "hard constraint" approach.

3.2 Quantitative Results

To provide a comprehensive evaluation and clarify the performance metrics against inherent class imbalance, we report three indicators in Table 1: Duration-weighted Accuracy, which reflects the proportion of total video time correctly classified; Clip-level Accuracy (Top-1), representing the standard discrete instance-based classification rate; and the Macro-F1 score, calculated exclusively over the active ground-truth categories to account for the recall of minority tasks.

As detailed in Table 1, the experimental data confirms that the proposed Segmented P-Domain strategy yielded the most reliable performance across all metrics, achieving a Duration-weighted Accuracy of 65.44%, a Clip-level Accuracy of 55.56%, and a Macro-F1 score of 0.547. This result validates the hypothesis that layering physical segmentation as a visual constraint serves as an effective method for stabilizing VLM outputs.

It is worth noting, however, that these metrics require careful interpretation, particularly regarding the Segmented P-Syntactic model. While this configuration achieved a seemingly competitive accuracy of 58.17%, a deeper inspection and the resulting lower Macro-F1 score (0.419) reveals this score was inflated by frequency bias (class imbalance). Given that the dataset is naturally dominated by repetitive "Fastening" tasks, the P-Syntactic model adopted a heuristic strategy of predicting the majority class for nearly every segment. Consequently, it failed to register minority but critical tasks such as "Checking" or "Alignment". In contrast, the P-Domain strategy demonstrated a genuine sensitivity to action variance as evidenced by its superior Macro-F1 performance, making it far more practically viable for industrial monitoring despite the modest numerical gap in raw accuracy.

Table 1. Comparative Performance of VLM Workflows

Input Strategy	Prompt Level	Duration-weighted Acc	Clip-level Acc (Top-1)	Macro-F1
Segmented	P-Domain (Proposed)	65.44%	55.56%	0.547
Segmented	P-Syntactic	58.17%	48.15%	0.419
Continuous	P-Syntactic	47.54%	37.04%	0.383
Continuous	P-Domain	47.09%	33.33%	0.257
Continuous	P-Natural	42.62%	37.04%	0.246
Segmented	P-Natural	42.06%	33.33%	0.167

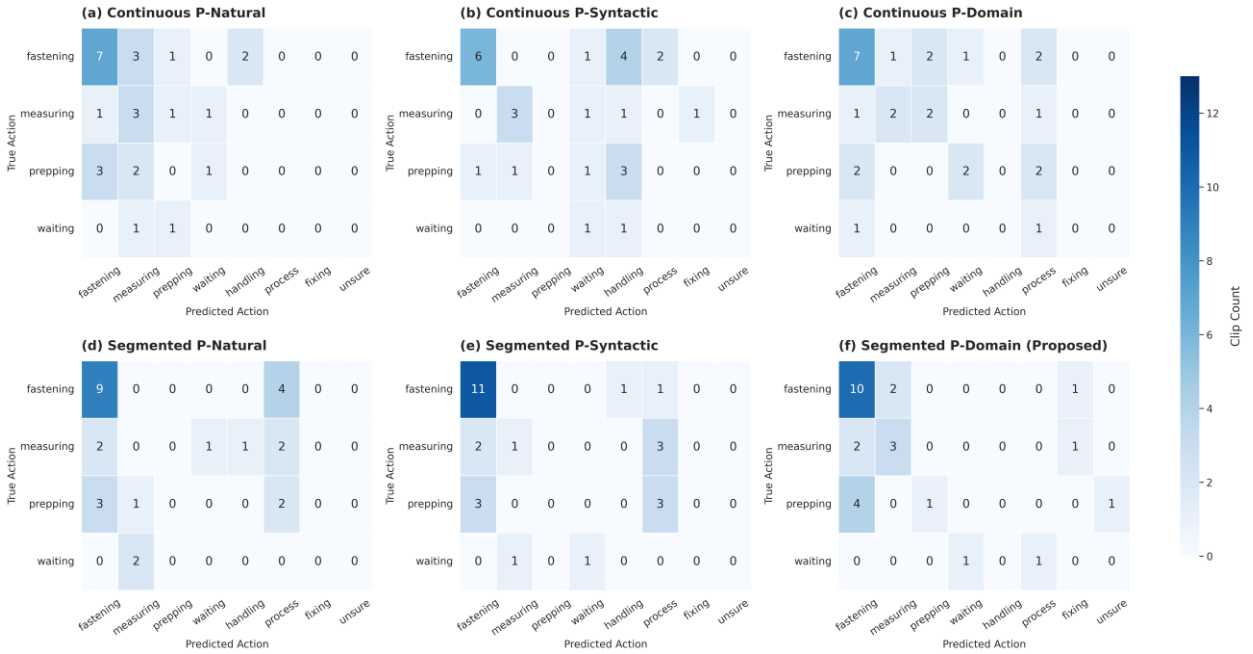

Figure 2. Comparison of action recognition performance across six experimental configurations.

3.3 Qualitative Analysis

While the quantitative metrics demonstrate the stability of the proposed method, a granular analysis utilizing the comprehensive 6-panel Confusion Matrix (Figure 2) reveals distinct operational characteristics and vulnerabilities of VLMs. As illustrated in Figure 2, the Proposed Method (f) achieves the highest concentration of correct predictions along the diagonal. This visually demonstrates the superior reliability of combining physical temporal segmentation with domain-informed prompting, effectively mitigating the attention drift and temporal inconsistency observed in continuous baselines.

The most prominent limitation is Small Object Blindness. Across all testing configurations, the models consistently struggled to resolve small tools, such as chalk lines or spray cans. The continuous baseline often compensated for this visual deficit by leveraging the full temporal sequence to infer actions (e.g., deducing "Chalking" from walking motions). However, the segmented approach, lacking this sequential context, frequently misclassified these same motions as generic "Measuring."

This visual limitation was mirrored by a semantic rigidity in the domain-informed models. By enforcing a strict construction knowledge, the VLM exhibited a tendency toward Taxonomy Over-fitting. As illustrated in Figure 2, by evaluating the model against 8 predefined options, we observed that it frequently misclassified ambiguous manual segments into distractor categories like *'fixing'* or *'process'* that were absent in the physical sequence. In its attempt to satisfy the "cognitive constraint," the model often forced nuanced or ambiguous behaviors into the nearest predefined category. This semantic hallucination explains the performance gap between raw accuracy and Macro-F1, as misclassifications into zero-support distractor classes directly penalized the recall of actual minority tasks like *prepping*. This indicates that while domain constraints successfully standardize output formatting, providing an overly broad knowledge without dynamic context-awareness can inadvertently degrade precision.

4 Discussion and Conclusion

4.1 Limitation and Future Work

Our results confirm that physical segmentation effectively eliminates the temporal hallucinations observed in continuous baselines; however, the relatively small sample size derived from this initial high-complexity sequence precludes a rigorous inferential statistical analysis at this stage. Consequently, these findings provide initial empirical evidence supporting the efficacy of the 'Segment-and-Caption' framework. While the current sample size precludes broad statistical claims of universal superiority, the mechanistic advantages of decoupling visual context from temporal search are clearly demonstrated.

Furthermore, this architectural choice creates a fundamental trade-off: by gaining temporal stability, we sacrifice narrative continuity. This results in an "Identity

Gap," where the "stateless" processing prevents the system from maintaining consistent worker identities across discrete clips (e.g., labeling the same individual as "Worker A" then "Worker B"). Consequently, while action recognition is accurate, individual productivity profiling remains challenging. Additionally, the current workflow relies on ground-truth timestamps. Future iterations will address these dual constraints by: (1) developing an Automated Event Detection module to dynamically slice video streams, removing manual dependency; and (2) fusing the VLM workflow with traditional computer vision-based tracking (Re-ID) to link worker identities across disjointed segments, thereby restoring continuity without sacrificing the stability gains of segmentation.

4.2 Conclusion

This study addresses the critical bottleneck of converting unstructured construction videos into machine-readable engineering logs. By validating the "Segment-and-Caption" framework, we demonstrated that decoupling visual context from temporal search is a necessary evolution for industrial monitoring. Our results confirm that applying physical segmentation as a hard constraint significantly outperforms continuous baseline methods in metric stability, achieving an Action Type Accuracy of 65.44%. While trade-offs in semantic depth and identity continuity remain, this framework establishes a viable technical pathway for automating granular productivity tracking in the next generation of prefabrication factories.

References

[1] Y. Fu, J. Chen, and W. Lu. Human-robot collaboration for modular construction manufacturing: Review of academic research. *Automation in Construction*, 158:105196, 2024. Doi: 10.1016/j.autcon.2023.105196

[2] A. Abu Nokta. *A Lean-based Approach to Increase Process Efficiency in Precast Concrete Panel Manufacturing*. Master's thesis, University of Alberta, Edmonton, Canada, 2025. Doi:10.7939/83076

[3] C. P. Chea, Y. Bai, X. Pan, M. Arashpour, and Y. Xie. An integrated review of automation and robotic technologies for structural prefabrication and construction. *Transportation Safety and Environment*, 2(2):81-96, 2020. Doi: 10.1093/tse/tdaa007

[4] X. Yang, F. Amtsberg, M. Sedlmair, and A. Menges. Challenges and potential for human-robot collaboration in timber prefabrication. *Automation in Construction*, 160:105333, 2024. Doi: 10.1016/j.autcon.2024.105333

[5] Z. Ren and J.I. Kim. The role of AI in on-site construction robotics: A state-of-the-art review using the sense-think-act framework. *Buildings*, 15(13):2374, 2025. Doi: 10.3390/buildings15132374

[6] I. Awolusi, E. Marks, and M. Hallowell. Wearable technology for personalized construction safety monitoring and trending: Review of applicable devices. Automation in Construction, 85:96-106, 2018. Doi: 10.1016/j.autcon.2017.10.010

[7] R. Xiong, Y. Wang, J. Cai, K. Liu, Y. Zhu, P. Tang, and N. El-Gohary. OpenConstruction: A systematic synthesis of open visual datasets for data-centric artificial intelligence in construction monitoring. *arXiv*, 2025. Doi: 10.48550/arXiv.2508.11482

[8] A. Abdalwhab, A. Imran, S. Heydarian, I. Iordanova, and D. St-Onge. Are open-vocabulary models ready for detection of MEP elements on construction sites. *arXiv*, 2025. Doi: 10.48550/arXiv.2501.09267

[9] M. Wang, J. Cai, D. Hu, Y. Hu, Z. Han, and S. Li. AI-based robots in industrialized building manufacturing. *Frontiers of Engineering Management*, 12(1):59-85, 2025. Doi: 10.1007/s42524-025-4099-x

[10] A. Pal, J. J. Lin, S. H. Hsieh, and M. Golparvar-Fard. Automated vision-based construction progress monitoring in built environment through digital twin. *Developments in the Built Environment*, 16:100247, 2023. Doi: 10.1016/j.dibe.2023.100247

[11] Y. Tang, J. Bi, S. Xu, L. Song, S. Liang, T. Wang, D. Zhang, J. An, J. Lin, R. Zhu, A. Vosoughi, C. Huang, Z. Zhang, P. Liu, M. Feng, F. Zheng, J. Zhang, P. Luo, J. Luo, and C. Xu. Video understanding with large language models: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. Doi: 10.48550/arXiv.2312.17432

[12] S. Yin, C. Fu, S. Zhao, K. Li, X. Sun, T. Xu, and E. Chen. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403, 2024. Doi: 10.1093/nsr/nwae403

[13] C. Zeng, T. Hartmann, and L. Ma. Ontology-based prompting with large language models for inferring construction activities from construction images. *Advanced Engineering Informatics*, 69:103869, 2026. Doi: 10.1016/j.aei.2025.103869

[14] Y. Wang, Y. Wang, P. Wu, J. Liang, D. Zhao, Y. Liu, and Z. Zheng. Efficient temporal extrapolation of multimodal large language models with temporal grounding bridge. *arXiv*, 2024. Doi: 10.48550/arXiv.2402.16050