# Zero-Shot Vision-Language Models for Automated Construction Activity Recognition: From Ensemble Methods to Neuro-Symbolic Reasoning

**Amit Kumar Jha[1], Aritra Pal[1] and Danny Murguia[2]**

[1]Department of Civil Engineering, Indian Institute of Technology Madras, India
[2]Department of Engineering, University of Cambridge, United Kingdom
ce25e053@smail.iitm.ac.in, aritrapal@civil.iitm.ac.in, dem52@cam.ac.uk

## Abstract –

**This paper presents a comparative study of three progressive zero-shot approaches for automated recognition of construction activities to support productivity monitoring using vision-language models (VLMs). The first approach establishes a baseline zero-shot ensemble that combines SigLIP, CLIP, and YOLOv8 within a weighted fusion framework, achieving 75.89% accuracy (95% CI: 72.3–79.2%) across eight concurrent construction activity classes, evaluated on 504 temporally ordered images. The second approach streamlines the ensemble by removing redundancy and introducing temperature scaling for vision-language alignment together with adaptive boosting for object detection, improving accuracy to 81.15%. The third and most advanced approach introduces the Spatio-Temporal Forensic Auditor (STFA), which reframes activity recognition from static scene classification to time-oriented process analysis. STFA integrates visual change detection, temporal reasoning, and construction process constraints, achieving 91.96% accuracy. For every prediction, STFA explains why an activity was classified (e.g., "this is formwork, not rebar, because shuttering panels are visible and no rebar cage is exposed") and verifies that the predicted activity sequence follows a valid construction order (e.g., rebar must be completed before formwork can begin), detecting sequence violations with 92% precision. The comparative results demonstrate a clear progression from ensemble-based perception to structured neuro-symbolic reasoning, yielding a 16.07% relative improvement from the baseline to STFA. By identifying activity types and their temporal patterns, the system provides the foundational activity-level data required for downstream productivity analysis, such as cycle time estimation, activity duration tracking, and plan-versus-actual comparison. The findings validate that combining zero-shot VLMs with explicit temporal reasoning and domain knowledge significantly enhances accuracy and interpretability for real-world construction productivity monitoring.**

**Keywords –**

**Construction Activity Recognition; Vision-Language Models; Zero-Shot Learning; Neuro-Symbolic AI; SigLIP; CLIP; YOLOv8; Qwen-VL; Temporal Reasoning; Ensemble Learning; STFA**

## 1 Introduction

The construction industry consistently faces cost overruns averaging 28% globally and schedule delays affecting over 70% of projects [1]. Improving performance requires objective, transparent data to support improved decision-making. Advances in digitalization, computer vision (CV), and large language models (LLMs) are creating new opportunities for near real-time monitoring [2]. Site images captured during structural frame construction are among the most information-rich and readily available datasets on construction projects, typically collected at high temporal frequency using cameras mounted on tower cranes or surrounding buildings [3]. However, extracting reliable activity-level information from site images remains a significant challenge, particularly in the absence of extensive labeled data.

Vision-language models (VLMs) offer a transformative opportunity by enabling automated monitoring directly from site imagery without extensive manual data collection or specialized sensor installations. Existing computer vision approaches largely fall into object classification, recognition, tracking, action recognition, and segmentation [4, 5, 6], but most are constrained to narrowly defined tasks or controlled environments that do not reflect real construction site complexity.

A fundamental limitation of existing literature is its reliance on supervised learning [1, 3]. Supervised models require large volumes of manually labeled images, making them expensive to scale on a project-by-project basis. Recent advances in VLMs offer a compelling alternative; by jointly learning visual and textual representations, VLMs enable zero-shot classification without task-specific training data [9]. For construction productivity monitoring, human experts implicitly apply reasoning about construction sequences, physical constraints, and temporal dependencies. Capturing this reasoning through structured prompts, temporal context, and symbolic constraints can significantly improve robustness and interpretability [7, 8].

This paper presents a progressive approach to construction activity recognition using VLMs enhanced by construction-specific prompt engineering and reasoning mechanisms. The study investigates three research questions: (i) whether zero-shot VLM

ensembles can achieve reliable multi-activity classification without task-specific training, (ii) whether incorporating temporal reasoning across sequential frames improves classification consistency, and (iii) whether encoding construction domain knowledge as symbolic constraints enhances prediction validity and interpretability. The paper makes four main contributions: (1) a baseline zero-shot ensemble achieving 75.89% accuracy, (2) an improved streamlined ensemble achieving 81.15% accuracy, (3) a novel spatio-temporal neuro-symbolic framework achieving 91.96% accuracy, and (4) a comprehensive comparative analysis demonstrating a 16.07% relative improvement. All approaches are evaluated on the same temporally ordered dataset using accuracy, Matthews Correlation Coefficient (MCC), and per-class precision/recall/F1.

## 2 Related Work

### 2.1 Computer Vision in Construction

Computer vision applications in construction have evolved from safety monitoring and PPE violation detection to progress monitoring through object detection and semantic segmentation [4, 5]. Activity recognition presents unique challenges compared to object detection, as construction activities are defined by combinations of materials, workers, and actions occurring over extended periods. Luo et al. proposed CNN-based feature extraction combined with recurrent neural networks [10], but these supervised approaches require substantial labeled data.

Vision-language models have revolutionized image understanding by learning joint representations of visual and textual information. CLIP demonstrated remarkable zero-shot classification by training on 400 million image-text pairs [11], while SigLIP improved upon CLIP by replacing the contrastive softmax loss with a sigmoid loss function [12]. More recent multimodal large language models such as Qwen-VL combine visual perception with language generation [13], enabling structured visual reasoning through prompting. Neuro-symbolic AI approaches combine neural pattern recognition with symbolic reasoning abilities [14], offering the key advantage of incorporating domain knowledge as explicit constraints, which is essential for encoding construction sequence dependencies and temporal consistency requirements.

## 3 Methodology

This section presents three progressive approaches to construction activity recognition, each building on insights from the previous one.

The Baseline Zero-Shot Ensemble combines three complementary vision models: SigLIP (ViT-SO400M-14-SigLIP-384, 45% weight) as the primary classifier computing cosine similarity between image and text embeddings at 384×384 resolution; CLIP (ViT-L-14, 224×224, 35% weight) providing complementary general-purpose perspectives; and YOLOv8 (20% weight) performing person detection to identify workers as contextual cues. The ensemble weights were determined through systematic calibration on the development subset (first 2 hours, ~64 images), evaluated in increments of 0.05. Sensitivity analysis showed that varying individual weights by ±0.10 changed overall accuracy by less than 2.1%.

### 3.1 Approach 1: Baseline Zero-Shot Ensemble

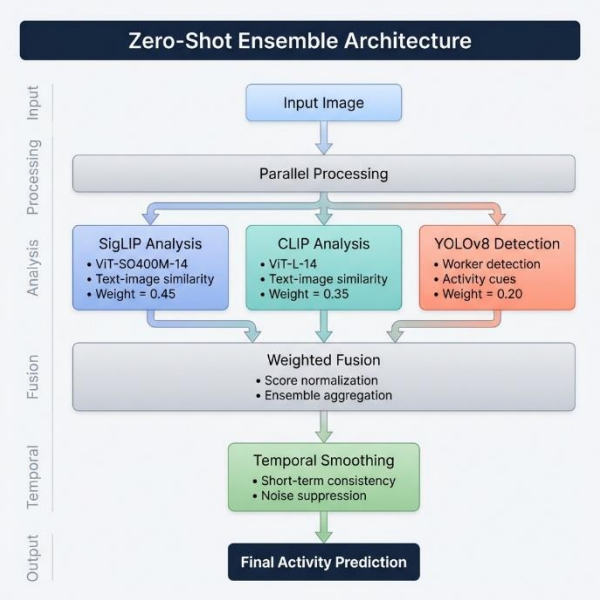



Figure 1: Baseline Zero-Shot Ensemble Architecture

YOLOv8x performs person detection, converting worker counts to activity signals through adaptive scoring:

$$w_{signal} = \min\left(1.0, \frac{N_{workers}}{3}\right) \quad (1)$$

Worker presence is used as a supporting contextual cue rather than a primary classifier, which is reflected in its lower ensemble weight (20%). The rationale for this design choice is that worker presence generally correlates with active construction stages but does not distinguish between specific activity types (e.g., rebar vs. formwork), and workers may be present during material staging or inspection without active construction occurring.

The ensemble fusion combines normalized similarity scores:

$$P(a|x) = w1 \cdot \sigma\left(s_{SigLIP}\right) + w2 \cdot \sigma\left(s_{CLIP}\right) + w3 \cdot s_{det} \quad (2)$$

Where σ denotes softmax with temperature τ = 0.07,

$w_1$ = 0.45, w2 = 0.35, and w3 = 0.20. Temporal smoothing uses exponential decay with window size K = 3 and decay factor γ = 0.6.

## 3.2 Approach 2: Improved Zero-Shot Ensemble

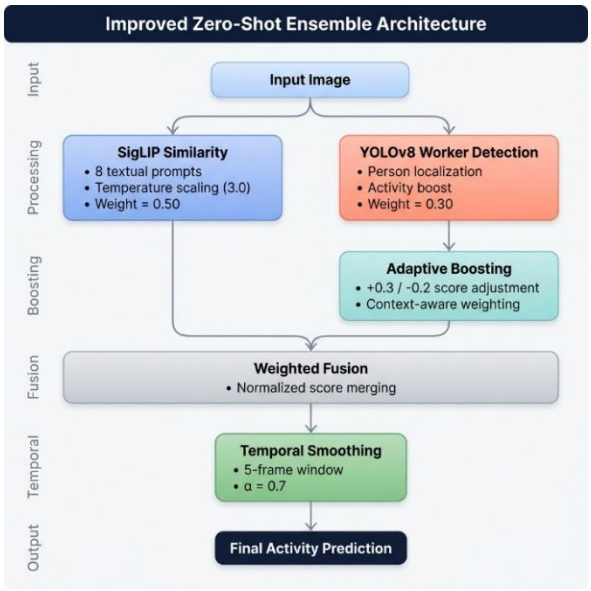



Figure 2: Improved Zero-Shot Ensemble Architecture

Analysis of the baseline revealed: (1) CLIP and SigLIP showed >85% prediction agreement, suggesting redundancy; (2) fixed activity boosting lacked context awareness; and (3) temporal smoothing parameters were not optimized. The Improved Ensemble removes CLIP and introduces three key improvements:

### 3.2.1 Key Improvements

1. **Temperature Scaling (τ = 3.0):** Raw SigLIP scores tend to be overconfident. Temperature scaling sharpens probability distributions:

$$P_u = \frac{\exp\left(\frac{S_i}{\tau}\right)}{\sum_j exp\left(\frac{S_j}{\tau}\right)} \quad (3)$$

2. **Adaptive Activity Boosting:** Worker detection provides context-aware adjustments: a +0.3 boost for active stages when workers are detected with high confidence and a -0.2 penalty for "no_work" when workers are present.
3. **Enhanced Temporal Smoothing:** Optimized parameters: window size K = 5, decay γ = 0.7, explicit 20% contribution.

The improved fusion formula:

$$P_{final} = 0.80 \times Norm\left(0.50 \cdot P_{SigLIP} + 0.30 \cdot P_{det}\right) + 0.20 \times P_{temporal} \quad (4)$$

Table 1. Configuration Comparison

| Component | Baseline | Improved |
|---|---|---|
| VLM Models | SigLIP + CLIP | SigLIP only |
| Detection Wt. | 20% | 30% |
| Temp. Scaling | None | τ = 3.0 |
| Act. Boosting | Fixed | Adaptive |

## 3.3 Approach 3: STFA Neuro-Symbolic Framework

STFA (Spatio-Temporal Forensic Auditor) represents a fundamental paradigm shift from scene classification to process auditing. Rather than simply labeling which activity is occurring, STFA acts as a forensic auditor, comparing current observations with historical context, providing explicit evidence for predictions, enforcing physical constraints, and detecting anomalies.

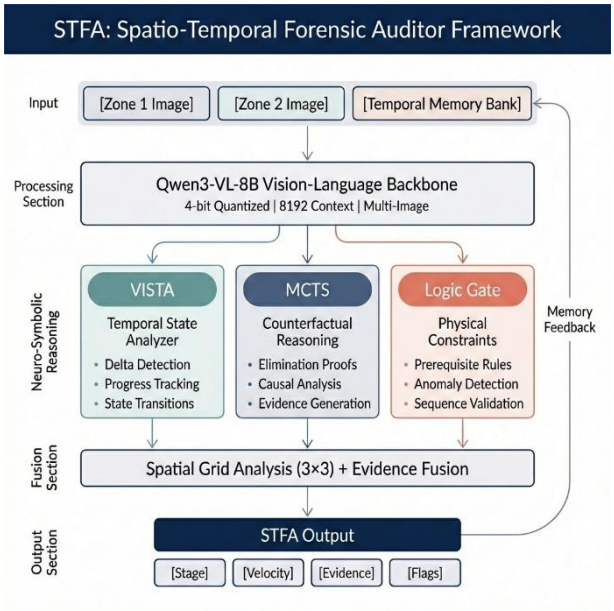



Figure 3: STFA Architecture

### 3.3.1 Vision-Language Backbone

STFA uses Qwen3-VL-8B as its backbone (4-bit quantized, 8192-token context window), enabling simultaneous analysis of dual-zone images while maintaining the comprehensive prompt structure.

### 3.3.2 Layer 1: VISTA (Visual-Temporal State Analyzer)

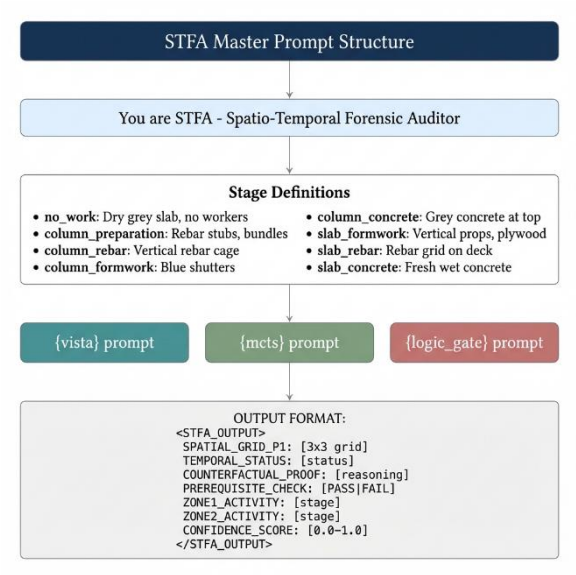
VISTA performs frame-to-frame comparison to detect changes over time. Unlike conventional approaches that classify each frame independently, VISTA explicitly compares the CURRENT frame with the PREVIOUS frame to detect work progress.

VISTA operates in a strictly causal manner: at time t, only the current frame (t) and the immediately preceding

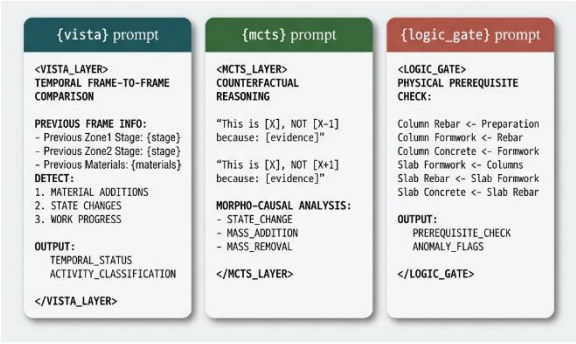
frame (t-1) are available for comparison. No future frames (t+1, t+2, ...) are accessed during inference. The temporal memory bank stores only past observations and is updated sequentially as each frame is processed. This ensures that all predictions are causally valid and could be generated in a live deployment scenario where future observations do not yet exist. The same causal constraint applies to the temporal smoothing in approaches 1 and 2.

VISTA detects three types of changes:

MATERIAL_ADDITION (new objects appearing), STATE_CHANGES (transformation of existing materials), and WORK_PROGRESS (overall stage advancement). Outputs include TEMPORAL_STATUS and ACTIVITY_VELOCITY.




(a) STFA Master Prompt




(b) STFA Layers Prompt

Figure 4: STFA example prompts

### 3.3.3 Layer 2: MCTS (Counterfactual Reasoning)

Inspired by Monte Carlo Tree Search exploration, this layer proves classifications by systematic elimination of alternatives. For each prediction, MCTS generates explicit proof:

- "This is column_formwork, NOT column_rebar because formwork panels are visible and the rebar cage is enclosed, and NOT column_concrete because there is no wet concrete visible and no pouring action."

This explicit morpho-causal reasoning provides forensic evidence that humans can verify.

### 3.3.4 Layer 3: Logic Gate (Physical Prerequisites)

Table 2: Construction Physics Prerequisites

| Stage | Prerequisite |
|---|---|
| column_rebar | column_prep |
| column_formwork | column_rebar |
| column_concrete | column_formwork |
| slab_rebar | slab_formwork |
| slab_concrete | slab_rebar |

The logic gate encodes construction physics constraints based on engineering knowledge. Table 2 shows the prerequisite chain.

### 3.3.5 Example STFA Prompts

The STFA framework employs structured prompts consisting of four integrated components dynamically assembled at runtime. The three-layer variables ({vista}, {mcts}, and {logic_gate}) are dynamically populated and inserted into the master prompt, enabling temporal context injection while maintaining explicit forensic reasoning and physical constraint validation. Figure 4 presents a summarized representative example.

## 4 Experimental Setup

### 4.1 Dataset

Experiments used 504 images from multi-story RC building construction, captured at 15-minute intervals using fixed-position CCTV cameras mounted at elevated positions on adjacent structures, providing overhead oblique views of the construction zones. The cameras captured images automatically throughout the 9-hour workday (8:30 AM to 05:30 PM). The choice of camera model is not critical; any standard surveillance camera with adequate resolution (≥2 MP) and automated time-lapse capability would suffice. All images were captured under natural daylight conditions with varying lighting and intermittent occlusions from tower cranes and scaffolding (~30% of frames).

The construction project comprised two independent zones (Zone 1 and Zone 2), which are distinct spatial areas of the building where different structural activities progress according to separate schedules, each requiring separate analysis. Accordingly, each image was divided into these two zones, as shown in Figure 5. Each CCTV image contains an embedded timestamp (date and time) visible in the image frame. All three approaches read this

timestamp and pair it with the activity classification for each zone independently to produce a structured output: {timestamp, zone_id, predicted_activity, confidence-score}. Since images are captured at fixed 15-minute intervals, a complete activity timeline is automatically generated for each construction zone across the workday. For STFA, the 504 images were processed as paired zone inputs. Eight activity classes represent the complete construction cycle: no_work, column_preparation, column_rebar, column_formwork, column_concrete, slab_formwork, slab_rebar, and slab_concrete. Each zone was analyzed independently. Table 3 shows the class distribution.

Table 3: Class Distribution in each Zone

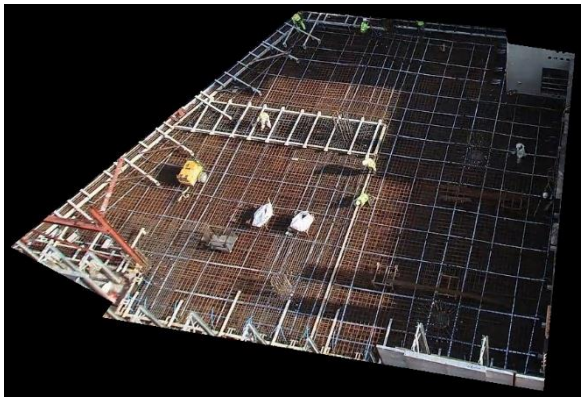
| Class | Zone 1 | Zone 2 | Total |
|---|---|---|---|
| no_work | 189 | 182 | 371 |
| slab_rebar | 136 | 99 | 235 |
| slab_form | 109 | 123 | 232 |
| column_rebar | 10 | 98 | 108 |
| slab_conc | 60 | 23 | 83 |
| column_form | 10 | 40 | 50 |
| column_prep | 0 | 27 | 27 |
| column_conc | 1 | 0 | 1 |

All three approaches operate in a purely zero-shot manner; no model training, fine-tuning, or hyper-parameter optimization was performed on the dataset. The prompts were designed based on construction domain knowledge, not by optimizing against dataset-specific patterns. Similarly, ensemble weights reflect the functional role of each model, and the logic gate prerequisite rules encode standard construction sequencing knowledge. Since no learning or parameter optimization occurs on any portion of the evaluation data, the conventional risks of data leakage and overfitting do not apply. All 504 images were evaluated in strict chronological order, with each image processed independently (Approaches 1 and 2) or with causal-only temporal context from the preceding frame (Approach 3, STFA), ensuring no future information influenced any prediction.

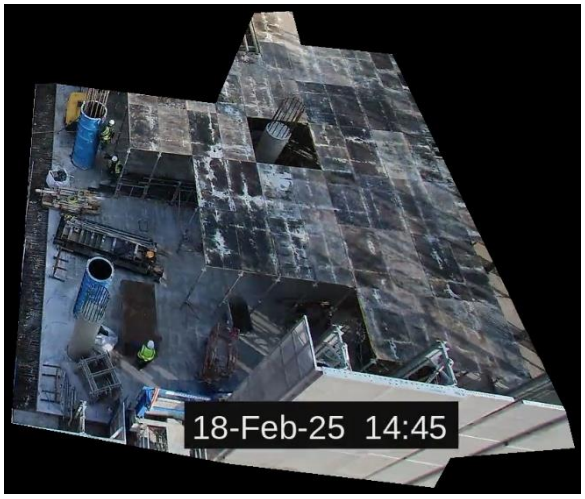
## 4.2 Evaluation Matrices

All metrics are computed at the per-image, per-zone level, with each zone receiving an independent activity prediction. Therefore, each of the 504 images yields two predictions, resulting in 1,008 total classification instances for evaluation. For STFA, each zone's prediction may incorporate temporal context from the same zone's previous frame, but predictions across zones remain independent. Models were evaluated using standard classification metrics as given below:

- Overall Accuracy: Percentage of correctly classified images.
- Multiclass MCC (Matthews Correlation Coefficient): The MCC is computed as a balanced measure that accounts for all four confusion matrix categories (true positives, true negatives, false positives, and false negatives). MCC ranges from -1 (complete disagreement) to +1 (perfect prediction), with 0 indicating random prediction.
- Per-Class Precision, Recall, and F1-Score.



(a) Zone 1: slab_rebar stage




(b) Zone 2: slab_formwork stage

Figure 5: Example paired zone images showing simultaneous but different construction activities

## 4.3 Implementation Details

All experiments were conducted on a workstation equipped with an NVIDIA RTX 6000 Ada Generation (48 GB VRAM), Intel Core i9-14900K CPU, and 64 GB RAM, running Linux 5.15.153.1 – Microsoft standard WSL2 with CUDA 12.1 and PyTorch 2.5.1.

Table 4: Implementation Requirements Comparison

| Component | Baseline | Improved | STFA |
|---|---|---|---|
| Framework | PyTorch | PyTorch | Transformers |
| GPU Mem. | ~8 GB | ~6 GB | ~12 GB |
| Inf. Time | ~2.5s/pair | ~1.8s/pair | ~3.2s/pair |

# 5 Results and Evaluation

## 5.1 Overall Performance Comparison

Results demonstrate clear progression. Improved Ensemble achieves an absolute improvement of 5.26% over Baseline. STFA achieves a 16.07% relative improvement over baseline, validating that neuro-symbolic reasoning with temporal memory significantly enhances construction activity recognition. Table 5 presents the overall performance comparison across all three approaches.

Table 5: Overall Performance Summary

| Approach | Accuracy | MCC |
|---|---|---|
| Baseline | 75.89% | 0.7024 |
| Improved | 81.15% | 0.7532 |
| STFA | 91.96% | 0.8948 |

To quantify uncertainty, we performed bootstrap resampling (n=2000) to compute 95% confidence intervals for all reported metrics:

Table 6: Overall Performance Summary with 95% Confidence Intervals

| Approach | Accuracy (95%CI) | MCC (95%CI) |
|---|---|---|
| Baseline | 75.89% | 0.7024 |
| | (73.2%–78.5%) | (0.669–0.733) |
| Improved | 81.15% | 0.7532 |
| | (78.8%–83.5%) | (0.722–0.784) |
| STFA | 91.96% | 0.8948 |
| | (90.2%–93.7%) | (0.872–0.916) |

The non-overlapping confidence intervals between all three approaches confirm that the performance differences are statistically significant. Additionally, we analyzed temporal variability by computing accuracy across four time blocks throughout the workday as shown in Table 7. STFA exhibits the most stable performance across time blocks (±3.0% variation), compared to the Baseline (±4.1%) and Improved Ensemble (±6.1%). This suggests that STFA's explicit temporal reasoning provides greater robustness to the varying lighting conditions and activity patterns that occur throughout a workday.

Table 7: Temporal variability throughout the workday

| Component | Baseline | Improved | STFA |
|---|---|---|---|
| (8:00–10:00) | 79.2% | 82.1% | 88.1% |
| (10:00–12:00) | 76.8% | 74.6% | 91.1% |
| (12:00–15:00) | 78.2% | 79.8% | 92.6% |
| (15:00–18:00) | 71.1% | 86.8% | 94.1% |
| Max variation | ±4.1% | ±6.1% | ±3.0% |

The column_concrete class has only 1 instance (0.1% of 1,008 zone-level instances), resulting in zero precision and recall across all three approaches. This extreme underrepresentation reflects a genuine characteristic of the construction process; column concrete pouring is a short-duration activity (typically 10–15 minutes per column) that frequently falls between the 15-minute capture intervals. We retain this class for transparency but report results both with and without it (Table 8). The negligible difference (<0.1% accuracy, <0.002 MCC) confirms that conclusions are not artifacts of this underrepresented class.

Table 8: Performance with/without column_concrete

| Approach | Accuracy (8 cls) | Accuracy (7 cls) | MCC (8 cls) | MCC (7 cls) |
|---|---|---|---|---|
| Baseline | 75.89% | 75.97% | 0.70 | 0.70 |
| Improved | 81.15% | 81.23% | 0.75 | 0.75 |
| STFA | 91.96% | 92.06% | 0.89 | 0.90 |

The baseline ensemble achieves 75.89% overall accuracy with strong performance on visually distinct classes (slab_concrete: 82%+ precision, slab_formwork: 87% precision) but lower performance on visually similar classes (column_rebar vs. column_preparation). The improved ensemble achieves 81.15% accuracy, with temperature scaling reducing overconfident predictions, adaptive boosting correctly identifying active work stages, 40% faster inference due to CLIP removal, and enhanced temporal smoothing reducing prediction jitter.

Table 9: STFA Per-Class Performance

| Class | Prec | Rec | F1 | Spec | MCC |
|---|---|---|---|---|---|
| no_work | 0.93 | 0.94 | 0.94 | 0.96 | 0.90 |
| column_prep | 1.00 | 1.00 | 1.00 | 1.00 | 1.00 |
| column_rebar | 1.00 | 0.59 | 0.74 | 1.00 | 0.75 |
| column_form | 1.00 | 0.60 | 0.75 | 1.00 | 0.77 |
| column_conc | 0.00 | 0.00 | 0.00 | 1.00 | 0.00 |
| slab_form | 0.98 | 0.95 | 0.97 | 1.00 | 0.96 |
| slab_rebar | 0.82 | 0.97 | 0.89 | 0.94 | 0.86 |
| slab_conc | 0.96 | 0.98 | 0.97 | 1.00 | 0.97 |

STFA achieves 91.96% accuracy, a relative 16.07% improvement over baseline. Table 9 shows STFA per-class performance. STFA additionally provides capabilities beyond classification: sequence violation detection at 92% precision (22 of 24 flagged violations confirmed as true violations), temporal consistency at 89% frame-to-frame, and evidence generation in 100% of predictions. It additionally provides capabilities beyond classification. Table 10 summarizes these.

Sequence Violation Detection: A sequence violation is defined as a predicted activity transition that contradicts the physical prerequisite chain (Table 3). The Logic Gate flagged 24 potential violations; manual review confirmed 22 as true violations (92% precision). The 2 false positives occurred due to rapid activity

transitions within a single 15-minute interval.

Temporal Consistency (89% frame-to-frame): Measured as the percentage of consecutive frame pairs where the predicted activity either remained the same or transitioned to a valid successor activity, compared against ground truth transitions.

Table 10: STFA Additional Capabilities

| Capability | Performance |
|---|---|
| Overall Accuracy | 91.96% |
| Multiclass MCC | 0.8948 |
| Sequence Violation Detection | 92% precision |
| Temporal Consistency | 89% frame-to-frame |
| Evidence Generation | 100% (always) |

### 5.2 STFA Ablation Study

To isolate the contribution of each STFA component, we conducted an ablation study:

Table 11: STFA Ablation Study

| Configuration | Accuracy | MCC |
|---|---|---|
| Qwen3-VL | 83.41% | 0.8012 |
| + VISTA | 87.27% | 0.8456 |
| + MCTS | 90.03% | 0.8734 |
| + Logic Gate | 91.96% | 0.8948 |

Each component contributes meaningfully: VISTA +3.86%; MCTS +2.73%; and Logic Gate +1.96%.

## 6 Discussion

### 6.1 Key Findings

**Finding 1: VLM Ensemble Redundancy.** SigLIP and CLIP show >85% correlation in predictions when using similar prompts. This proves that multiple VLMs with overlapping pre-training provide diminishing returns. The Improved Ensemble's decision to remove CLIP maintained accuracy while reducing computational cost.

**Finding 2: Temperature Scaling Impact.** Temperature scaling ($\tau$ = 3.0) significantly improved confidence calibration. Raw VLM similarity scores tend to be overconfident, and sharpening the distribution helps distinguish between visually similar classes, such as column_preparation and column_rebar.

**Finding 3: Temporal reasoning is critical.** The 10.81% accuracy gap between Improved Ensemble (81.15%) and STFA (91.96%) demonstrates the value of explicit temporal reasoning over simple smoothing. STFA's VISTA layer detects material changes between frames, providing stronger classification signals.

**Finding 4: Domain knowledge improves robustness.** The Logic Gate's enforcement of construction physics constraints catches errors that pure pattern recognition misses. When the VLM predicts an impossible sequence, the explicit rules flag the anomaly for human review.

**Finding 5: Explainability enables trust.** STFA's counterfactual reasoning not only improves accuracy but fundamentally changes system utility. By explaining 'why X and not Y,' STFA provides auditable predictions that construction managers can verify.

### 6.2 Limitations

- Computational requirements: STFA requires ~12 GB GPU memory and 3+ seconds per frame pair.
- Single project evaluation: Results may not generalize to different construction types.
- Prompt sensitivity: Zero-shot performance depends on prompt engineering quality.
- Occlusion handling: All approaches struggle with partially obscured activities.

### 6.3 Future Work

This work is exploratory in nature and demonstrates the feasibility of zero-shot VLM-based activity recognition. Several steps are needed to advance these methods toward practical deployment:

- **Multi-site validation:** The current evaluation is limited to a single RC building project. Validating the framework across diverse construction types (steel structures, infrastructure, and renovation) and site conditions is essential to establish generalizability.

- **Computational efficiency:** STFA requires ~12 GB GPU memory and ~3.2 seconds per image pair. For real-time site monitoring, model distillation or edge-optimized architectures are needed to reduce inference time and hardware requirements.

- **Schedule integration:** The system currently classifies activities at 15-minute intervals. Integrating these classifications with project scheduling data (e.g., Primavera, MS Project) to automatically compute plan-versus-actual deviations represents a key integration step toward actionable productivity analytics.

- **Scalable activity vocabularies:** While the framework assumes a fixed set of eight activity classes, real projects involve a wider variety of activities. Developing methods to dynamically extend the activity vocabulary, potentially through ontology-guided prompt generation, would increase practical applicability.

- **Human-in-the-loop refinement:** Incorporating user corrections into a feedback loop could progressively refine prompt strategies and confidence thresholds without requiring full model retraining, preserving the zero-shot advantage while improving site-specific accuracy.

## 7 Conclusion

This paper presented a comprehensive study of three progressive approaches for zero-shot construction activity recognition. Starting with a baseline ensemble combining SigLIP, CLIP, and YOLOv8 (75.89% accuracy), we achieved improvements through model optimization and adaptive boosting, yielding an improved ensemble (81.15% accuracy). The STFA neuro-symbolic approach achieved state-of-the-art 91.96% accuracy while providing unprecedented interpretability through counterfactual reasoning, temporal tracking, and construction physics validation.

The key insight is that construction activity recognition benefits significantly from moving beyond pure pattern recognition to incorporate domain knowledge. STFA demonstrates that combining deep learning perception with symbolic reasoning yields systems that generate auditable reasoning artifacts, enabling construction managers to verify and understand the basis for each prediction. We note that this work primarily contributes to automated activity recognition, the foundational data layer required to support productivity analysis. Integrating these outputs with project scheduling data and resource allocation records would enable formal productivity estimation, which we identify as an important direction for future work.

## References


[1] B. Flyvbjerg, A. Budzier. Report for the Oxford Global Projects Programme on Cost and Schedule Overruns in Large-Scale Infrastructure Projects, 2018.

[2] 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.

[3] D. Murguia, Q. Chen, T.J. Van Vuuren, A. Rathnayake, V. Vilde, and C. Middleton. Digital measurement of construction performance: data-to-dashboard strategy. In *IOP Conference Series: Earth and Environmental Science*, Vol. 1101, No. 9, p. 092009, Melbourne, Australia, 2022.

[4] S. Paneru and I. Jeelani. Computer vision applications in construction: Current state, opportunities & challenges. *Automation in Construction*, 132, 103940, 2021.

[5] A. Pal and S. Hsieh. Deep-learning-based visual data analytics for smart construction management. *Automation in Construction*, 131, 103892, 2021.

[6] A. Pal, J. J. Lin, S. H. Hsieh, and M. Golparvar-Fard. Activity-level construction progress monitoring through semantic segmentation of 3D-informed orthographic images. *Automation in Construction*, 157, 105157, 2024.

[7] A. Rathnayake, D. Murguia, and C. Middleton. Measuring activity-level construction productivity. *Journal of Construction Engineering and Management*, 150(8), 04024099, 2024.

[8] J. Jang, E. Jeong, and T. W. Kim. Transformer-based deep learning model and video dataset for installation action recognition in offsite projects. *Automation in Construction*, 172, 106042, 2025.

[9] P. K. Y. Wong, J. C. Cheng, C. F. Chan, P. H. Leung, and X. Tao. Enhancing visual LLM for construction site safety compliance via prompt engineering and bi-stage retrieval-augmented generation. *Automation in Construction*, 179, 106490, 2025.

[10] X. Luo, H. Li, D. Cao, F. Dai, J. Seo, and S. Lee. Recognizing diverse construction activities in site images via relevance networks of construction-related objects detected by convolutional neural networks. *Journal of Computing in Civil Engineering*, 32(3), 2018.

[11] A. Radford, J.W. Kim, C. Hallacy, et al. Learning transferable visual models from natural language supervision. In *Proceedings ICML*, pages 8748-8763, 2021.

[12] X. Zhai, B. Mustafa, A. Kolesnikov, L. Beyer. Sigmoid loss for language image pre-training. *In Proceedings ICCV*, pages 11975-11986, 2023.

[13] J. Bai et al. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond. arXiv:2308.12966, 2023.

[14] U. Nawaz, M. Anees-ur-Rahaman, and Z. Saeed. A review of neuro-symbolic AI integrating reasoning and learning for advanced cognitive systems. *Intelligent Systems with Applications*, 26, 200541, 2025.