2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

EXOM: An Excavator Operation Monitoring Framework with Onboard Vision and Sensor Data

Seok-Kyu Kang, Seong-Gye Lee, Gye-Bong Jang

Abstract— **Reliable monitoring of excavator operations in real-world environments requires accurate excavation counting to ensure productivity, efficient computation for real-time inference, and cost-effective on-board sensing—a combination that most prior systems fail to achieve. We present EXOM (EXcavator Operation Monitoring), a lightweight and deployable framework that relies solely on a factory-installed cabin camera and built-in hydraulic sensors. EXOM integrates two embedded-friendly modules: a Video data Processing Module (VPM), where an ECSE algorithm leverages bucket detection to estimate excavation sections and counts from state transitions, and a Sensor data Processing Module (SPM), where an Adaptive Window (AW) process sparsifies time-series signals and drives a segmentation model through a learnable sparse tensor. To capture deployability, we introduce EXOM-I, a unified index that combines section-level F1 and normalized excavation counting accuracy. Experiments with real-world data demonstrate that EXOM consistently outperforms previous approaches, achieving state-of-the-art performance with real-time latency on resource-limited embedded excavator hardware.**

I. INTRODUCTION

The construction industry, valued at over 10 trillion USD globally, is undergoing rapid transformation through robotics and automation technologies [1]–[4] aimed at improving productivity, reducing costs, and enhancing safety. Among heavy machinery, excavators are central not only to earthmoving and material handling but also as key platforms for deploying intelligent monitoring systems in the field.

Accurate real-time monitoring of excavator operations is technically challenging yet crucial [5]–[9], as tasks vary by terrain, operator behavior, and context. Precise understanding reduces idle time, fuel consumption, and maintenance costs. Furthermore, the rise of autonomous systems [10]–[13] heightens the need for reliable, fine-grained data to ensure robust deployment across diverse environments.

Despite growing interest, existing approaches face key limitations in accuracy and deployability. Sensor-based methods (RFID, UWB, IMU, GNSS) [14]–[22] require costly infrastructure and lack scalability. Multi-camera vision systems [23]–[30] suffer from occlusion and installation overhead. Multimodal frameworks [31]–[33] integrate vision and sensors, but their high computational demand prevents real-time deployment on embedded excavator hardware. Crucially, these methods often fail to accurately delineate excavation segments and counts, which leads to an unreliable interpretation of other operational phases and limits their overall utility for productivity analysis and automation.

The authors are with the HD Hyundai AI Center, Seongnam, Republic of Korea (e-mail: {seokkyu.kang, seonggye.lee, gyebong.jang}@hd.com).

To address these challenges, we present **EXOM (EXcavator Operation Monitoring)**, a lightweight and deployable framework tailored for embedded excavator platforms. EXOM operates solely with a factory-installed cabin camera and built-in hydraulic sensors, avoiding additional infrastructure. As shown in Fig. 1, EXOM integrates two embedded-friendly modules: the Video data Processing Module (VPM) and the Sensor data Processing Module (SPM).

VPM, driven by the Excavation Counting and Section Estimation (ECSE) algorithm, applies object detection to classify bucket states from monocular video and tracks their transitions to infer excavation counts and sections. This design enables efficient use of visual information, ensuring real-time performance while preserving high accuracy in excavation section and count estimation. SPM complements VPM by targeting non-excavation operations captured by hydraulic sensors. Its Adaptive Window (AW) process sparsifies time-series signals and provides a learnable sparse tensor for deep learning–based segmentation, allowing fine-grained discrimination of excavation-like but distinct operations such as leveling or other transitions.

EXOM achieves state-of-the-art performance, delivering accurate and real-time operation monitoring with end-to-end latency suitable for resource-limited excavators. To demonstrate practical feasibility, we further validate EXOM on full-scale 30–36t excavators (HX300 and DX340LC-5), confirming its effectiveness under real-world field conditions.

Our contributions are as follows:

- **EXOM**: an on-board framework for real-time excavator operation monitoring that relies only on a factory-installed cabin camera and built-in hydraulic sensors, enabling lightweight and deployable use in the field.
- **VPM with ECSE**: a vision data processing module that applies an object (bucket) detection and state-transition tracking algorithm to provide accurate excavation counting and section estimation from monocular video.
- **SPM with AW**: a sensor data processing module that sparsifies uninformative sequences through an Adaptive Window (AW) process and supplies a learnable sparse tensor to a segmentation model, enabling fine-grained discrimination of non-excavation operations.
- **System validation**: extensive real-world experiments demonstrating that EXOM outperforms previous approaches, achieving state-of-the-art performance with end-to-end latency of ≤ 30 ms on an NVIDIA Jetson Orin NX 8 GB, ensuring real-time deployability on resource-limited excavators.

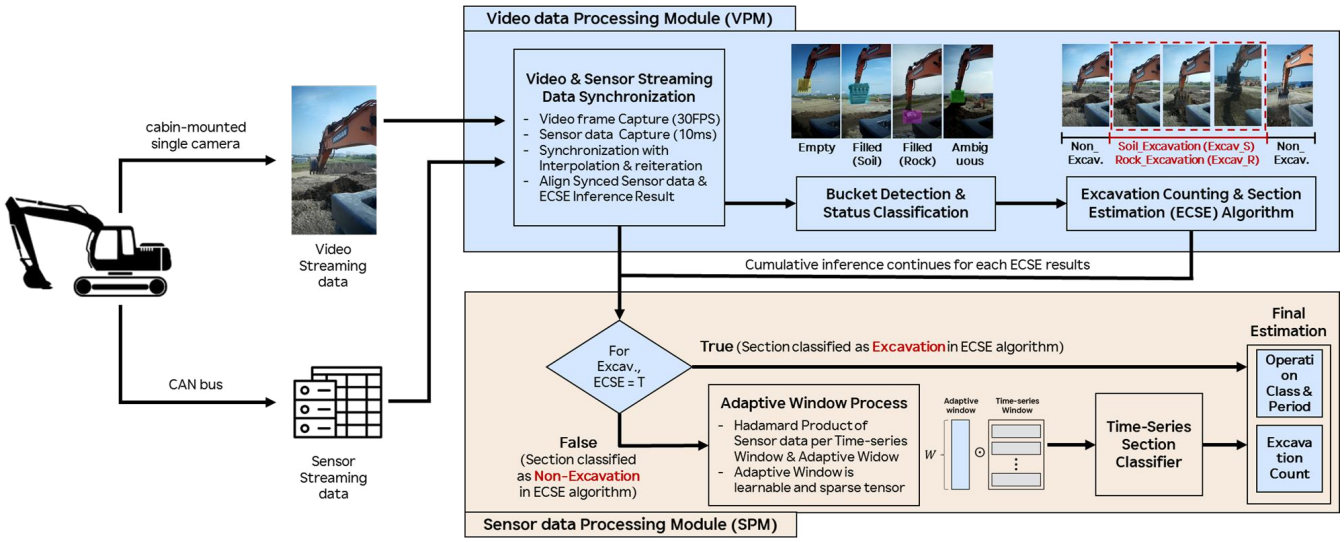


Fig. 1: Overview of the proposed **EXOM** framework. A factory-installed cabin camera feeds the Video data Processing Module (VPM), where the ECSE algorithm estimates *excavation sections and counts* from bucket-state transitions. Built-in hydraulic sensors feed the Sensor data Processing Module (SPM), where the Adaptive Window (AW) process sparsifies time-series signals and a deep learning–based model estimates *non-excavation sections*. Together, these modules provide accurate and lightweight operation monitoring deployable on embedded excavator hardware.

II. Related Work

Early attempts to monitor excavator operations relied on external sensors such as RFID, UWB, IMU, or GNSS [14]–[22]. While these methods can provide direct motion or location signals, they require costly infrastructure for installation and calibration and are often tied to specific machine configurations, making retrofitting onto rental or legacy excavators infeasible and limiting portability across job sites [34].

To support real-time inference under constrained conditions, researchers have applied various machine learning and lightweight deep learning models to sensor time-series data [29]. Early work explored classical classifiers such as KNN, SVM, and MLPs [35], [36], while later studies adopted ANN, LSTM, CNN-LSTM, and attention-based architectures [21], [22], [37]–[39]. These methods demonstrate the potential of data-driven modeling, but sensor-only modalities struggle to distinguish fine-grained operation modes without visual context [40]. Moreover, many are validated only on curated datasets or laboratory-scale settings, raising concerns about generalizability to real-world excavator environments.

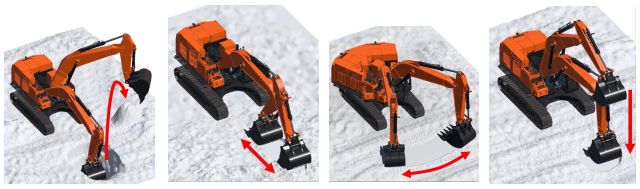
Vision-based methods have also been widely studied. Earlier approaches used fixed external cameras with classical classifiers such as Bayesian networks, SVMs, or Random Forests [24]–[27], whereas recent work employs CNN-LSTM pipelines or multi-stage detectors [23], [28], [30], [38], reporting accuracies up to 90%. However, these systems often depend on multi-camera setups, suffer from occlusion and overlapping machinery [7], [27], [28], [38], and require large annotated datasets and high computational budgets, making real-time embedded deployment infeasible [23], [29]. In addition, these methods estimate excavation counts only indirectly from activity sequences, whereas direct bucket observation, as adopted by EXOM, is essential for reliable accuracy in practice.

Recent multimodal frameworks attempt to combine vision and sensors [31]–[33]. While they report improved recognition accuracy, their computational complexity prevents real-time performance, and the need for additional infrastructure undermines practicality for resource-constrained excavators.

Taken together, existing systems either achieve accuracy at the cost of heavy infrastructure and computation or remain lightweight but fail at fine-grained discrimination, particularly for reliable excavation counting. This gap motivates the development of a framework that *balances accuracy, efficiency, and deployability* under real-world constraints—the focus of our proposed EXOM system.

III. Proposed Framework

EXOM is a modular framework for real-time excavator operation monitoring using only onboard resources. As illustrated in Fig. 1, it integrates a factory-installed cabin camera and built-in hydraulic sensors into two complementary modules, each tailored to a distinct activity type. Excavation activities are strongly reflected in bucket state transitions and can be efficiently captured through vision, whereas non-excavation activities manifest as subtler temporal variations better represented in hydraulic signals. VPM first distinguishes excavation from non-excavation intervals; excavation segments are then analyzed within VPM to estimate counts and section types, while non-excavation intervals are passed to SPM for further segmentation and classification. This cooperative division of labor enables robust monitoring across diverse conditions while preserving low latency and computational efficiency for embedded deployment.



(a) Excavation (b) Lv-FB (c) Lv-H (d) Lv-V

Fig. 2: Representative excavator operation modes. Excavation denotes scooping and discharging material, counted as a single cycle. Leveling tasks are subdivided into forward-backward (Lv-FB), horizontal (Lv-H), and vertical (Lv-V) leveling. Additional modes such as traveling and idle are also considered in our dataset but omitted here for brevity.

A. *Vision Processing Module (VPM)*

Distinguishing excavation from other activities using sensor signals alone is inherently difficult. Operations such as leveling often exhibit motion patterns that closely resemble excavation, leading to ambiguity and reduced accuracy in sensor-only approaches. Vision provides a clearer signal, enabling both accurate excavation detection and material identification. However, prior vision-based systems typically relied on multiple fixed cameras around the job site [23]–[30], which are computationally expensive, prone to occlusions, and difficult to deploy at scale. In contrast, the factory-installed cabin camera offers a first-person view of the bucket and work area, inherently avoiding occlusion from other machinery and ensuring consistent visibility. This eliminates the need for additional infrastructure while guaranteeing coverage of the excavation process.

A key insight of EXOM is that excavation can be identified by tracking bucket states rather than analyzing entire frames. When an empty bucket becomes filled with soil or rock and is later emptied, an excavation event can be counted directly; if the same motion occurs with the bucket remaining empty, it is classified as non-excavation. This bucket-centric perspective enables both excavation section segmentation and counting with minimal computation, making real-time deployment feasible even on low-power embedded platforms.

*1) **Bucket Detection and Status Classification**:* To realize this design, we isolate the bucket region and train a dedicated detection model for status classification. Bounding boxes for the bucket were manually annotated, and each frame is labeled as one of four states: *Empty*, *Filled with Soil*, *Filled with Rock*, or *Ambiguous*. We adopt YOLOX-s [41] for bucket detection, as its anchor-free design and decoupled head strike an effective balance between accuracy and speed, making it suitable for embedded deployment in excavators.

*2) **Excavation Counting and Section Estimation (ECSE)**:* The ECSE algorithm tracks state transitions over time to estimate both excavation counts and section types in real time. A fixed-length queue stores recent predictions and applies majority filtering to suppress short-term noise. A valid excavation cycle is detected when a transition pattern `Empty → Filled → Empty` occurs, and the history of filled states determines whether the segment corresponds to soil excavation, rock excavation, or a non-excavation action.

Algorithm 1 Excavation Counting and Section Estimation (ECSE) - Compact Version

Initialize queue length k, threshold p, history list
for each incoming frame **do**
  $b \leftarrow$ DetectBucketStatus(frame)
  Update bucket_queue with b
  $b_{est} \leftarrow$ majority **if** freq(b) $> k \times p$ **else** b
  if $b_{prev} = 0$ **and** $b_{est} \in \{1,2\}$ **then**
    Append b_{est} to history
  else if $b_{prev} \in \{1,2\}$ **and** $b_{est} = 0$ **and** history$[-1] \in \{1,2\}$ **then**
    count $\leftarrow$ count $+1$
    section $\leftarrow$ Majority(history): Soil_Excavation, Rock_Excavation, or Non_Excavation
    Reset history
  $b_{prev} \leftarrow b_{est}$

Key components of ECSE:

- **Status Detection:** Each frame is classified into one of four bucket states.
- **Queue-Based Filtering:** A fixed queue smooths predictions and removes short-term fluctuations.
- **Transition Detection:** Excavation cycles are defined by `Empty → Filled → Empty`.
- **Section Estimation:** The dominant filled label (soil vs. rock) defines the section type; otherwise non-excavation.
- **Output:** Real-time excavation counts and section labels for downstream monitoring.

This efficient, bucket-centric algorithm avoids full-frame video modeling and heavy temporal networks, enabling accurate excavation monitoring with end-to-end latency under 30 ms on embedded hardware.

B. *Sensor Data Processing Module (SPM)*

Sensor-only approaches have historically struggled to achieve high accuracy in excavator operation recognition, largely because non-excavation activities often resemble excavation motions. As shown in Fig. 2, leveling operations exhibit trajectories similar to excavation, yet correspond to fundamentally different tasks. The problem is further compounded by operator variability: preparatory gestures, motion intensity, and individual working styles introduce additional noise, making reliable classification difficult when relying solely on sensor signals.

In EXOM, this challenge is alleviated because excavation segments are already separated by VPM. SPM therefore focuses exclusively on non-excavation intervals, where it can more precisely distinguish operations such as forward-backward leveling (Lv-FB), horizontal leveling (Lv-H), vertical leveling (Lv-V), driving, idle, and undefined. This division of labor not only simplifies the classification task but also improves robustness by narrowing the decision space.

The core principle of SPM is efficient computation. Directly feeding raw sensor streams into deep models is wasteful, as much of the signal reflects operator-specific noise rather than task-relevant information. To address this, we introduce the Adaptive Window (AW) process, which conceptually acts as a temporal filter that preserves only informative subsequences. Unlike static preprocessing, AW is learned jointly with the model: it assigns higher weights to salient time steps and suppresses irrelevant ones, effectively producing a sparse input representation. Originally designed to reduce computation by lightweight vector operations, AW unexpectedly also contributed to accuracy improvements by emphasizing discriminative patterns. This makes SPM both faster and more accurate, enabling efficient operation recognition within non-excavation segments.

*1) **Time-series Section Classifier**:* SPM adopts a Bi-LSTM as the backbone classifier to capture temporal dependencies in sensor signals. This choice is standard but effective: it allows the model to exploit both past and future context, improving robustness for sequences where current behavior depends on preceding or subsequent motions. The novelty of SPM lies not in the backbone itself but in how inputs are preprocessed by the Adaptive Window mechanism.

*2) **Adaptive Window Process**:* The Adaptive Window (AW) process is designed to filter out operator-specific noise and redundant patterns, ensuring that the classifier attends only to task-relevant signals. Conceptually, AW serves as a learnable temporal mask that selects informative subsequences and suppresses irrelevant ones before they are passed to the Bi-LSTM.

Formally, given an input segment $X \in \mathbb{R}^{W\times d}$ with window length W and feature dimension d, a mask vector $AW \in \mathbb{R}^{W}$ is learned and broadcast across features:

$$X' = AW^{(d)} \odot X, \quad AW^{(d)} \in \mathbb{R}^{W\times d}.$$

The mask is optimized jointly with the classifier, so salient time steps receive larger weights during training. After convergence, a binary mask is obtained by thresholding the top $c\%$ of values in AW:

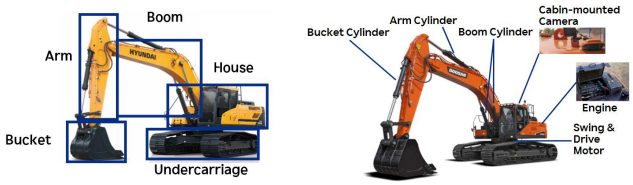
$$AW_{\text{bin},i} = \begin{cases} 1 & \text{if } |AW_i| \geq \theta_c, \\ 0 & \text{otherwise,} \end{cases}$$

where θ_c is the cut-off determined by sparsity ratio c. The masked input is then:

$$X'' = AW_{\text{bin}}^{(d)} \odot X,$$

which is passed to the Bi-LSTM while zero-masked steps are skipped.

This mechanism achieves two goals simultaneously. First, by discarding uninformative segments, it reduces computational overhead, allowing faster inference on embedded hardware. Second, it improves accuracy by emphasizing discriminative temporal cues while suppressing operator-dependent noise. In practice, we found that AW not only shortened inference time but also led to consistent performance gains, highlighting its effectiveness as both a computational and representational improvement.



(a) Excavator structure (b) Sensor configuration

Fig. 3: Excavator structure and sensor configuration used in our study. (a) HX300 (HD Hyundai Construction Equipment) illustrating the fundamental components of an excavator, including boom, arm, bucket, house, and undercarriage. (b) DX340LC-5 (DEVELON) with annotated factory-installed onboard sensors, namely the cabin-mounted camera, hydraulic cylinders (boom, arm, bucket), and drive/swing motors connected via the CAN-bus.

IV. Experiment Setup

A. Configuration and Dataset

We categorize excavator operations into eight classes: Soil Excavation, Rock Excavation, Forward-Backward Leveling (Lv-FB), Horizontal Leveling (Lv-H), Vertical Leveling (Lv-V), Drive, Undefined, and Idle. Experiments were conducted on two 30–36 t class excavators widely used in practice: the HD Hyundai Construction Equipment HX300 and the DEVELON DX340LC-5.

As shown in Fig. 3, the system relies solely on onboard resources: a factory-installed cabin-mounted camera and built-in hydraulic sensors accessed via the Controller Area Network (CAN) bus, a standard protocol in heavy machinery for robust sensor communication. Video was recorded at 30 FPS, while hydraulic signals were retrieved through the CAN-bus. Since individual sensor channels operate at slightly different sampling rates, all signals were aligned to the slowest update cycle and then synchronized with the video stream using timestamps. Sensor readings were further interpolated to the video frame rate, ensuring that each frame was paired with temporally consistent hydraulic data.

The dataset comprises 425 minutes of recorded operations (388 for training, 37 for testing), yielding a total of 765,180 synchronized frames. Operations were collected in balanced proportions across the eight classes and annotated at 0.1 s intervals by domain experts, reducing class bias and improving robustness across diverse operators and site conditions.

In addition, a dedicated bucket detection dataset was created to support VPM training. A total of 5,276 frames were annotated with bounding boxes and bucket status labels (Empty, Soil, Rock, Ambiguous), providing ground truth for the ECSE algorithm. This dataset enables reliable excavation section segmentation and counting without requiring full-frame visual analysis. All experiments were deployed on an embedded platform, with EXOM running on an NVIDIA Jetson Orin NX (8 GB) installed inside the excavator cabin. Detailed specifications are summarized in Table I.

TABLE I: Excavator, Dataset, and On-board H/W Overview

Excavators	DX340LC-5 (30 ton), HX300 (36 ton)
Operation Time	425 minutes of real operation data (train+val: 388 min, test: 37 min from a separate day; val = 10% sampled per class)
Video	1920×1080, 30 FPS, 765,180 frames
Hydraulic Sensors	Engine Speed [rpm], Engine Torque [%], Boom Pressure [bar], Arm Pressure [bar], Bucket Pressure [bar], Travel Pilot, Swing Pressure [bar]
Operation Classes	Soil Excavation, Rock Excavation, Forward- Backward Leveling (Lv-FB), Horizontal Leveling (Lv-H), Vertical Leveling (Lv-V), Drive, Undefined, Idle
Bucket Dataset	5,276 frames (88 minutes at 1 FPS), each labeled with bucket Bbox and class (Empty, Soil, Rock, Ambiguous)
Evaluation Granularity	Operation classes evaluated under frame-level alignment of video and sensor data at 0.1-second intervals
On-board Hardware	NVIDIA Jetson Orin NX (8 GB), installed in the excavator cabin and operated using the excavator's own power supply

B. Excavator Operation Monitoring Evaluation Index (EXOM-I)

To holistically assess monitoring performance, we propose **EXOM-I**, a unified metric that evaluates both operation classification and excavation counting. Section-level classification metrics such as accuracy or F1 score capture how well a system distinguishes between different activities, while counting metrics evaluate how precisely excavation cycles are identified. For deployment, both dimensions must be satisfied simultaneously—accurate recognition without reliable counting, or vice versa, is insufficient.

For classification, we adopt the **F1 Score**, which balances precision and recall. For counting, we define the **Normalized Excavation Counting Accuracy (NECA)**, which scales the Mean Absolute Counting Error (MACE) by the true excavation count:

$$\text{NECA} = 1 - \frac{\text{MACE}}{\text{True Excavation Count}}$$

$$\text{MACE} = \frac{1}{N}\sum_{i=1}^{N} |\text{True}_i - \text{Pred}_i|$$

where N is the number of samples. NECA penalizes deviations in predicted counts proportionally to task size, ensuring that both underestimation and overestimation are fairly captured.

The final EXOM-I score is the simple average of the two components:

$$\text{EXOM-I} = \tfrac{1}{2}\left(\text{F1 Score} + \text{NECA}\right).$$

A higher EXOM-I indicates that a method achieves strong performance in both section classification and excavation counting under a single, deployment-oriented evaluation. By unifying these complementary measures, EXOM-I provides a clearer benchmark for assessing excavator monitoring systems beyond isolated accuracy or error metrics.

V. Experiment Results

We evaluate EXOM against four representative baselines frequently adopted in excavator operation recognition: an ANN-based model [21], an LSTM-based model [22], a CNN-LSTM hybrid [38], and a Dual Attention network [39]. These methods span from classical machine learning to more advanced deep models, providing a fair spectrum of comparison. All baselines were implemented with their original configurations and optimized for real-time inference.

In contrast, prior vision- and multimodal-based approaches [23]–[33] rely on multiple external cameras and sensors or heavy visual pipelines. While they report promising accuracy, such methods require hundreds of milliseconds per frame on GPU servers, far exceeding the 30 ms real-time threshold for embedded excavator deployment. Consequently, they could not be embedded on excavators for practical onboard use and were therefore excluded from direct comparison. All experiments were executed on an NVIDIA Jetson Orin NX (8 GB), demonstrating feasibility on a resource-limited embedded excavator platform.

A. Exp. 1: Evaluation on Excavator Operation Section and Excavation Counting Estimation

TABLE II: Final performance comparison with baseline models (Exp. 1). EXOM achieves the best results across all metrics, with a particularly large margin in NECA, showing its strength in accurate excavation count estimation.

Framework	Acc	Recall	Precision	F1	NECA	EXOM-I
ANN [21]	0.8639	0.8319	0.8416	0.8337	0.6034	0.7186
LSTM [22]	0.8982	0.8791	0.8991	0.8873	0.5862	0.7367
CNN-LSTM [38]	0.9070	0.8829	0.8956	0.8867	0.7069	0.7968
Dual Attn [39]	0.8697	0.8569	0.8741	0.8603	0.5690	0.7146
EXOM (ours)	**0.9382**	**0.9183**	**0.9330**	**0.9242**	**0.9828**	**0.9535**

Table II presents the overall comparison across all metrics. EXOM outperforms all baselines on section-level metrics and further raises the composite index EXOM-I by nearly 20% over the strongest baseline. The largest margin appears in NECA, where baselines fall below 0.71 due to indirect counting, while EXOM directly observes bucket status transitions and achieves 0.98. This performance is particularly notable given that video and sensor data were aligned at 0.1-second intervals, making the task substantially more demanding than benchmarks with coarser temporal labels.

Accurate excavation counting is critical in excavator monitoring, as even small miscounts can mislead operators or autonomous controllers regarding productivity and schedule adherence. By sustaining strong section-level performance and near-perfect NECA under such strict alignment, EXOM demonstrates that it can satisfy the stringent requirements of real-world deployment. These results establish our core contribution: reliable operation monitoring under realistic constraints. In the following sections, we further analyze these gains through ablation studies of VPM (Exp. 2) and SPM (Exp. 3), isolating the role of ECSE and AW.

(a) Status transitions during continuous excavation.

(b) Accurate section recognition even after relocating (drive) and resuming excavation.

Fig. 4: Live test footage from real excavators with EXOM installed. (a) The framework tracks bucket status transitions (Empty → Ambiguous → Filled (Soil) → Ambiguous → Empty) and updates excavation counts in real time. (b) Even after relocation, EXOM continues to identify excavation sections accurately, demonstrating robustness to task interruptions.

B. *Exp. 2: Ablation on VPM with the ECSE Algorithm*

Table III serves as an ablation study to isolate the contribution of the Video Processing Module (VPM) with the ECSE algorithm. Each baseline model is evaluated both in its original sensor-only configuration and in an augmented version with VPM attached, thereby holding the backbone constant while toggling bucket-centric vision on or off. This design explicitly tests whether modeling bucket transitions improves recognition beyond temporal sensor signals alone. Across all architectures, EXOM-I improves consistently, with relative gains between 12.5% and 20.7%. The strongest gains appear in NECA, where even shallow ANN and LSTM baselines—which otherwise miscount excavations—achieve sharp accuracy improvements once VPM supplies stable excavation cues. This confirms that direct vision of bucket state provides critical information that sensor dynamics cannot fully capture, particularly for distinguishing excavation from visually similar but non-productive motions such as leveling.

TABLE III: Effect of incorporating VPM with ECSE into baseline models (Exp. 2). VPM consistently improves both F1 and NECA, leading to large EXOM-I gains.

Framework	Without VPM			With VPM (ECSE)			Gain
	F1	NECA	EXOM-I	F1	NECA	EXOM-I	
ANN [21]	0.8337	0.6034	0.7186	0.8409	0.8621	0.8515	+18.5%
LSTM [22]	0.8873	0.5862	0.7367	0.9011	0.7931	0.8471	+15.0%
CNN-LSTM [38]	0.8867	0.7069	0.7968	0.8953	0.8966	0.8959	+12.5%
Dual Attn [39]	0.8603	0.5690	0.7146	0.8797	0.8448	0.8632	+20.7%
EXOM (ours)	—	—	—	**0.9242**	**0.9828**	**0.9535**	—

From a methodological standpoint, ECSE embodies a key principle: excavation can be identified through state transitions rather than full-frame analysis. By encoding the canonical Empty → Filled → Empty cycle into compact signals, VPM introduces a structured representation of excavation events that is both lightweight and domain-relevant. This explains why the benefit is universal across backbones—VPM provides complementary supervision that regularizes ambiguous temporal patterns, enabling models to achieve reliable counting with minimal overhead.

Finally, the results should be interpreted as controlled evidence of VPM's effectiveness rather than as a new set of competing benchmarks. The ablation clearly demonstrates that, without VPM, models often conflate excavation with leveling due to shared motion signatures. With VPM, however, bucket-state transitions act as explicit discriminative features, resolving these ambiguities and substantially boosting both section-level accuracy and excavation counting. This establishes VPM with ECSE not only as the main driver of EXOM's performance but also as a transferable plug-in that can enhance monitoring pipelines across architectures.

C. *Exp. 3: Ablation on SPM with the AW Process*

TABLE IV: Performance with and without the Adaptive Window (AW) process (Exp. 3). AW consistently improves F1 while reducing inference time, confirming its role as an efficient temporal filter.

Framework	Without AW		With AW	
	F1	Time (ms)	F1	Time (ms)
ANN [21]	0.8267	1.92	**0.8346**	**1.71**
LSTM [22]	0.8817	2.35	**0.8938**	**2.04**
CNN-LSTM [38]	0.8736	2.86	**0.8777**	**2.79**
Dual Attn [39]	0.8603	3.01	**0.8845**	**2.69**
EXOM (w/o VPM)	0.8850	1.98	**0.9090**	**1.84**

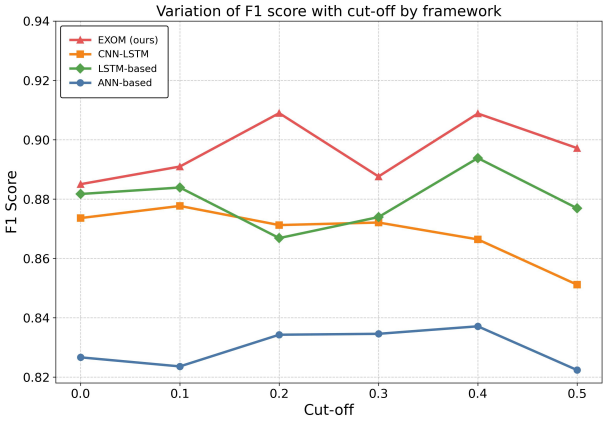
Table IV presents an ablation study on the Sensor Processing Module (SPM), isolating the contribution of the Adaptive Window (AW) process. Across all baselines, AW improves F1 scores by up to 0.02 while also reducing inference latency by 0.07–0.32 ms, showing that it is both accurate and efficient. The effect is most pronounced in EXOM without VPM, where the F1 score rises from 0.8850 to 0.9090, confirming that AW is particularly valuable when sensor signals carry most of the discriminative load.

AW achieves these gains by pruning uninformative time steps and retaining only salient temporal patterns. This selective masking reduces redundancy, shortens the effective input length, and prevents idle or noisy sensor readings from degrading classification. In practice, this enables the Bi-LSTM to focus on meaningful transitions in hydraulic signals while skipping irrelevant fluctuations tied to operator variability. The result is not just a marginal boost in accuracy, but a measurable improvement in efficiency that is critical for real-time embedded deployment.

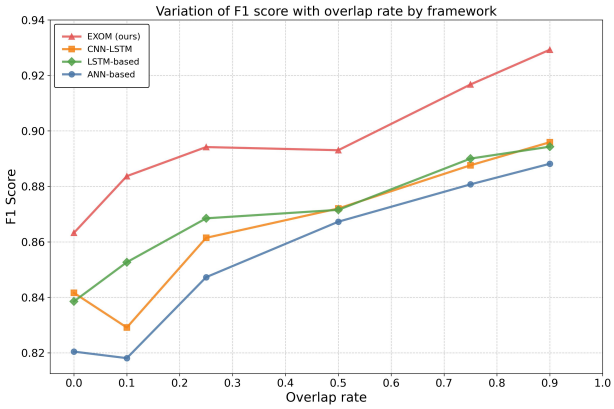
To further examine robustness, we varied the cut-off threshold and the temporal overlap rate between input windows (Fig. 5). EXOM consistently outperforms all baselines across settings, with its best F1 reaching 0.9292 at 0.9 overlap. These results confirm that AW generalizes beyond a single configuration, offering a stable and transferable mechanism for improving sensor-based activity recognition.

D. *Discussion and Summary*

The results across Exp. 2 and Exp. 3 highlight the complementary contributions of EXOM's two modules. The Video Processing Module (VPM), through the ECSE algorithm, provides bucket-centric visual cues that are highly discriminative for excavation, enabling accurate section segmentation and reliable counting. The Sensor Processing Module (SPM) enhances robustness during non-excavation periods by applying the Adaptive Window (AW) process, which filters noise and reduces overhead in sensor-based time-series modeling. Without VPM, excavation is often confused with visually similar motions such as leveling, while without AW, sensor classifiers become inefficient and error-prone under operator variability. By combining discriminative vision with efficient temporal filtering, EXOM achieves a balanced architecture that not only outperforms baselines but also meets the low-latency constraints of embedded platforms. This balance enhances deployment relevance, ensuring that the framework translates into tangible productivity and safety benefits in real construction environments.



(a) F1 score variation with changes in cut-off threshold.



(b) F1 score variation with changes in overlap rate.

Fig. 5: Ablation study on Adaptive Window (AW). EXOM consistently maintains higher F1 across thresholds and overlap rates, demonstrating robust accuracy and adaptability.

VI. Conclusion

This paper introduced **EXOM**, a deployable framework for excavator operation monitoring that relies only on factory-installed sensors. By unifying bucket-centric vision (VPM with ECSE) and adaptive sensor modeling (SPM with AW), EXOM achieves both accurate excavation counting and robust classification of non-excavation activities. Extensive experiments on real excavator data show clear and consistent improvements over strong baselines, underscoring not only technical validity but also readiness for integration into construction workflows. We believe EXOM marks a concrete step toward practical, real-time monitoring systems that can enhance productivity, support autonomous excavation, and strengthen safety in construction robotics.

REFERENCES

[1] B. Xiao, C. Chen, and X. Yin, "Recent advancements of robotics in construction," *Automation in Construction*, vol. 144, p. 104591, 2022.

[2] G. M. Daniel Eriksson, Reza Ghabcheloo, "Automatic loading of unknown material with a wheel loader using reinforcement learning," *2024 IEEE International Conference on Robotics and Automation (ICRA2024)*, 2024.

[3] M. F. Joseph Rowell, Lintong Zhang, "Lista: Geometric object-based change detection in cluttered environments," *2024 IEEE International Conference on Robotics and Automation (ICRA2024)*, 2024.

[4] D. S. Ejup Hoxha, Jinglun Feng and J. Xiao, "Robotic inspection and subsurface defect mapping using impact-echo and ground penetrating radar," *2024 IEEE International Conference on Robotics and Automation (ICRA2024)*, 2024.

[5] J. Yang, Z. Shi, and Z. Wu, "Vision-based action recognition of construction workers using dense trajectories," *Advanced Engineering Informatics*, vol. 30, no. 3, pp. 327–336, 2016.

[6] J. Yang, M.-W. Park, P. A. Vela, and M. Golparvar-Fard, "Construction performance monitoring via still images, time-lapse photos, and video streams: Now, tomorrow, and the future," *Advanced Engineering Informatics*, vol. 29, no. 2, pp. 211–224, 2015. Infrastructure Computer Vision.

[7] Z. Zhu, X. Ren, and Z. Chen, "Integrated detection and tracking of workforce and equipment from construction jobsite videos," *Automation in Construction*, vol. 81, pp. 161–171, 2017.

[8] L. Z. Sibo Zhang, "Construction site safety monitoring and excavator activity analysis system," *Construction Robotics*, 2022.

[9] A. B. K. Rabbi and I. Jeelani, "Ai integration in construction safety: Current state, challenges, and future opportunities in text, vision, and audio based applications," *Automation in Construction*, vol. 164, p. 105443, 2024.

[10] L. Chen, S. Jin, H. Wang, and L. Zhang, "Exact: An end-to-end autonomous excavator system using action chunking with transformers," *ICRA Workshop 2024: 3rd Workshop on Future of Construction: Lifelong Learning Robots in Changing Construction Sites*, 2024.

[11] S. Parascho, "Construction robotics: From automation to collaboration," *Annual Review of Control, Robotics, and Autonomous Systems*, vol. 6, no. 1, pp. 183–204, 2023.

[12] A. S. Rao, M. Radanovic, Y. Liu, S. Hu, Y. Fang, K. Khoshelham, M. Palaniswami, and T. Ngo, "Real-time monitoring of construction sites: Sensors, methods, and applications," *Automation in Construction*, vol. 136, p. 104099, 2022.

[13] T. Fu, T. Zhang, Y. Lv, X. Song, G. Li, and H. Yue, "Digital twin-based excavation trajectory generation of uncrewed excavators for autonomous mining," *Automation in Construction*, vol. 151, p. 104855, 2023.

[14] C. Zhang, A. Hammad, and S. Rodriguez, "Crane pose estimation using uwb real-time location system," *Journal of Computing in Civil Engineering*, vol. 26, no. 5, pp. 625–637, 2012.

[15] N. Pradhananga and J. Teizer, "Automatic spatio-temporal analysis of construction site equipment operations using gps data," *Automation in Construction*, vol. 29, pp. 107–122, 2013.

[16] S. El-Omari and O. Moselhi, "Integrating automated data acquisition technologies for progress reporting of construction projects," *Proceedings of the 2009 International Symposium on Automation and Robotics in Construction (ISARC 2009)*, pp. 86–94, June 2009.

[17] E. Ergen, B. Akinci, B. East, and J. Kirby, "Tracking components and maintenance history within a facility utilizing radio frequency identification technology," *Journal of Computing in Civil Engineering - J COMPUT CIVIL ENG*, vol. 21, 01 2007.

[18] A. Oloufa, M. Ikeda, and H. Oda, "Gps-based wireless collision detection of construction equipment," *Proceedings of the 19th International Symposium on Automation and Robotics in Construction (ISARC)*, pp. 461–466, September 2002.

[19] A. Montaser and O. Moselhi, "Rfid+ for tracking earthmoving operations," *Construction Research Congress*, pp. 1011–1020, 05 2012.

[20] W. Lu, G. Q. Huang, and H. Li, "Scenarios for applying rfid technology in construction project management," *Automation in Construction*, vol. 20, no. 2, pp. 101–106, 2011. Building Information Modeling and Changing Construction Practices.

[21] R. Akhavian and A. H. Behzadan, "Construction equipment activity recognition for simulation input modeling using mobile sensors and machine learning classifiers," *Advanced Engineering Informatics*, vol. 29, no. 4, pp. 867–877, 2015.

[22] K. M. Rashid and J. Louis, "Times-series data augmentation and deep learning for construction equipment activity recognition," *Advanced Engineering Informatics*, vol. 42, p. 100944, 2019.

[23] J. Kim, "Visual analytics for operation-level construction monitoring and documentation: State-of-the-art technologies, research challenges, and future directions," in *Frontiers in Built Environment*, 2020.

[24] C. Chen, B. Xiao, Y. Zhang, and Z. Zhu, "Automatic vision-based calculation of excavator earthmoving productivity using zero-shot learning activity recognition," *Automation in Construction*, vol. 146, p. 104702, 2023.

[25] J. Gong, C. H. Caldas, and C. Gordon, "Learning and classifying actions of construction workers and equipment using bag-of-video-feature-words and bayesian network models," *Advanced Engineering Informatics*, vol. 25, no. 4, pp. 771–782, 2011. Special Section: Advances and Challenges in Computing in Civil and Building Engineering.

[26] M. Golparvar-Fard, A. Heydarian, and J. C. Niebles, "Vision-based action recognition of earthmoving equipment using spatio-temporal features and support vector machine classifiers," *Advanced Engineering Informatics*, vol. 27, no. 4, pp. 652–663, 2013.

[27] A. K. Langroodi, F. Vahdatikhaki, and A. Doree, "Activity recognition of construction equipment using fractional random forest," *Automation in Construction*, vol. 122, p. 103465, 2021.

[28] I.-S. Kim, K. Latif, J. Kim, A. Sharafat, D.-E. Lee, and J. Seo, "Vision-based activity classification of excavators by bidirectional lstm," *Applied Sciences*, vol. 13, no. 1, 2023.

[29] B. Sherafat, C. Ahn, R. Akhavian, A. Behzadan, M. Golparvar-Fard, H. Kim, Y. Lee, A. Rashidi, and E. Azar, "Automated methods for activity recognition of construction workers and equipment: State-of-the-art review," *Journal of Construction Engineering and Management*, vol. 146, June 2020. Publisher Copyright: © 2020 American Society of Civil Engineers.

[30] J. Kim and S. Chi, "Action recognition of earthmoving excavators based on sequential pattern analysis of visual features and operation cycles," *Automation in Construction*, vol. 104, pp. 255–264, 2019.

[31] J.-Y. Kim and S.-B. Cho, "A deep neural network ensemble of multimodal signals for classifying excavator operations," *Neurocomputing*, vol. 470, pp. 290–299, 2022.

[32] J.-Y. Kim and S.-B. Cho, "Classifying excavator operations with fusion network of multi-modal deep learning models," in *14th International Conference on Soft Computing Models in Industrial and Environmental Applications (SOCO 2019)*, (Cham), pp. 25–34, Springer International Publishing, 2020.

[33] H. S. Cho, K. Latif, A. Sharafat, and J. Seo, "Multi-modal excavator activity recognition using two-stream cnn-lstm with rgb and point cloud inputs," *Applied Sciences*, vol. 15, no. 15, 2025.

[34] E. R. Azar, S. Dickinson, and B. McCabe, "Server-customer interaction tracker: Computer vision–based system to estimate dirt-loading cycles," *Journal of Construction Engineering and Management*, vol. 139, no. 7, pp. 785–794, 2013.

[35] J. Ryu, J. Seo, H. Jebelli, and S. Lee, "Automated action recognition using an accelerometer-embedded wristband-type activity tracker," *Journal of Construction Engineering and Management*, vol. 145, p. 04018114, 10 2018.

[36] H. Lee, C. R. Ahn, N. Choi, T. Kim, and H. Lee, "The effects of housing environments on the performance of activity-recognition systems using wi-fi channel state information: An exploratory study," *Sensors*, vol. 19, no. 5, 2019.

[37] T.-K. Lim, S.-M. Park, H.-C. Lee, and D.-E. Lee, "Artificial neural network–based slip-trip classifier using smart sensor for construction workplace," *Journal of Construction Engineering and Management*, vol. 142, no. 2, p. 04015065, 2016.

[38] T. Slaton, C. Hernandez, and R. Akhavian, "Construction activity recognition with convolutional recurrent networks," *Automation in Construction*, vol. 113, p. 103138, 2020.

[39] Y. Shen, J. Wang, C. Feng, and Q. Wang, "Dual attention-based deep learning for construction equipment activity recognition considering transition activities and imbalanced dataset," *Automation in Construction*, vol. 160, p. 105300, 2024.

[40] C. Chen, Z. Zhu, and A. Hammad, "Automated excavators activity recognition and productivity analysis from construction site surveillance videos," *Automation in Construction*, vol. 110, p. 103045, 2020.

[41] Z. Ge, S. Liu, F. Wang, Z. Li, and J. Sun, "Yolox: Exceeding yolo series in 2021," *ArXiv*, vol. abs/2107.08430, 2021.