ISSN 1392-3730/eISSN 1822-3605

VILNIUS TECH
Vilnius Gediminas Technical University

JOURNAL of CIVIL ENGINEERING and MANAGEMENT

2026
Volume 32
Issue 3
Pages 374–393
https://doi.org/10.3846/jcem.2026.26154

AN INTELLIGENT CONSTRUCTION SYSTEM BASED ON DIGITAL TWIN AND FOUNDATION MODEL OPTIMIZATION

Fengyi GUO[1], Cynthia Changxin WANG[2], Jun SUN[1], Kaixin HUANG[1], Xin WANG[3]✉

[1]*Department of Civil and Hydraulic Engineering, Huazhong University of Science and Technology, 430074 Wuhan, China*
[2]*School of Built Environment, University of New South Wales, 2052 Sydney, Australia*
[3]*China State Construction International Investments (Hubei) Limited, 430014 Wuhan, Hubei, China*

Article History:
- received 18 July 2025
- accepted 14 November 2025

Abstract. Construction sites routinely face multi-trade concurrency, spatiotemporal coupling, and high safety risk; relying solely on manual inspection and heuristic scheduling often leads to lagging detection and inconsistent execution. In response, recent practice has introduced digital twins (DT) to fuse video, sensors, and BIM and thus improve site visibility; however, most implementations remain at monitoring/visualization, lacking a mechanism to convert cognition into executable, verifiable decisions. Meanwhile, Transformer foundation models show strong capabilities in multimodal perception and representation learning, yet they are rarely closed-looped with engineering constraints and on-site execution.

Against this backdrop, taking high-rise self-climbing platform (SCP) operations as a representative scenario, we build a DT×Transformer closed-loop system. We align video/sensor/BIM/text at the component level via "Component-ID + Timestamp", train a multimodal Transformer for operation-state recognition and short-horizon risk prediction, and then explicitly encode safety, resource, and spatial precedence constraints in a policy module to generate feasible task sequences, which are delivered to crews via AR with acknowledgments to close the loop. The system integrates multi-source perception, digital twin, foundation-model reasoning, and AR-assisted execution, and was validated on a high-rise self-climbing platform project for its overall improvement of construction performance. The evaluation covered four key aspects – safety management, operational efficiency, communication and execution, and information transparency. Results show that the system significantly extends the lead time of risk warnings, reduces violation rates, stabilizes construction rhythm, shortens decision latency, and markedly improves the consistency between instruction delivery and on-site feedback.

Keywords: intelligent construction, digital twin, foundation model, high-rise self-climbing platform, AR interaction.

✉Corresponding author. E-mail: *wangxin@cohl.com*

1. Introduction

The construction industry, a cornerstone of global economic development, has long struggled with inefficiencies, frequent accidents, and substantial resource waste. Over the past two decades, productivity growth in construction has significantly lagged behind other sectors, with global labor productivity increasing by only 10% from 2000 to 2022 – just one-fifth of the global average (Construction Dive, n.d.). Large-scale projects routinely face overruns: 98% exceed their budgets by 30%, and 77% experience schedule delays of approximately 40% (Linesight, n.d.). Safety performance remains a concern, with construction accounting for nearly 20% of all U.S. workplace fatalities in 2021 (Bureau of Labor Statistics, n.d.). Moreover, the sector is hindered by fragmented workflows and pervasive information silos that impair collaboration and data sharing (McKinsey, n.d.). Despite growing recognition of the need for digitalization, adoption remains low; between 2005 and 2014, the technology adoption rate in construction was just 1.4% – far below sectors like Information and Communication Technology (ICT) (Madubuike et al., 2022). These persistent challenges highlight the urgent need for transformative technologies to enhance efficiency, safety, and decision-making.

Among emerging solutions, digital twins are increasingly used to resolve data fragmentation and support real-time decisions. A digital twin synchronizes physical assets and processes with a virtual model through sensor networks and IoT, consistent with cyber-physical systems principles (Madubuike et al., 2022). Although adoption in construction is still early, results across the lifecycle are

encouraging. At Frasers Tower in Singapore, more than 900 sensors and 179 Bluetooth beacons support real-time building operations. During construction, reported uses include real-time workflow monitoring (Wang & Chien, 2014), structural health tracking (Wang et al., 2024), and risk warning (Wang et al., 2021), enabling automated progress verification, continuous safety oversight, and more efficient allocation of labor and equipment. Yet most deployments remain monitor-centric. They consolidate data but rarely produce decisions that formally respect safety boundaries, resource coupling, and spatial precedence.

In recent years, foundation models in artificial intelligence – particularly those based on the Transformer architecture – have achieved remarkable progress, offering promising opportunities to enhance perception and intelligent decision-making in the construction industry. Centered on the attention mechanism, Transformer networks excel in multimodal representation and complex temporal learning. Since their initial introduction (Xu et al., 2023), they have consistently set new benchmarks across fields such as natural language processing and computer vision. Representative large-scale models, including BERT, GPT, and the Vision Transformer, can automatically learn hierarchical features and patterns from massive multimodal datasets, demonstrating strong potential for cross-modal fusion, semantic reasoning, and decision support. In the construction domain, Transformer-based methods have been applied to video-based recognition of workers and equipment, sensor time-series prediction, and optimization of scheduling and logistics (Amer et al., 2021), showing advantages in improving predictive foresight, semantic understanding, and decision accuracy under complex on-site conditions. However, when these models are not deeply integrated with engineering constraints, such as safety boundaries, resource coupling, and spatial precedence, their recognition or prediction results often fail to translate into executable on-site actions. This gap represents a key bottleneck in current intelligent construction practices: strong perception but weak decision closure and limited execution feedback.

Building on the above challenges, prior research reveals three gaps that remain insufficiently addressed. First, multimodal semantic alignment is still limited. Data streams from video, physical sensors, BIM, and text are often processed in isolation, and few frameworks achieve reliable synchronization across time, space, and semantics at the component or operation level. This lack of alignment introduces cumulative errors and latency amplification throughout downstream analysis and scheduling. Second, the generation of constraint-aware and feasible strategies remains underdeveloped. Without explicitly encoding safety envelopes, resource limits, and spatial precedence into optimization or policy learning, systems tend to produce unconstrained and idealized recommendations that are impractical or costly to execute on site. Third, the execution and feedback loop is fragile. In most existing studies, decisions end at simulation or visualization; the absence of a measurable recognition–decision–execution–relearning cycle makes it difficult to verify benefits, quantify uncertainties, or define applicability boundaries.

To address these gaps, this study selects the self-climbing platform (SCP) as a representative scenario of high concurrency and strong coupling and develops a DT× Transformer closed-loop system that has been validated on a live project. At the data and semantics layer, video, sensors data, 4D BIM, and task text are unified through a Component-ID plus timestamp alignment to ensure spatiotemporal coherence. At the cognitive layer, a multimodal Transformer is trained for operation-state recognition and short-horizon risk forecasting, with explicit modeling of missing data, occlusion, and noise effects on accuracy and latency. At the decision layer, safety, resource, and spatial precedence constraints are embedded into policy generation through optimization and reinforcement-learning techniques, yielding feasible and safety-compliant task sequences. At the execution and feedback layer, these strategies are delivered to crews via an augmented reality (AR) interface, allowing real-time confirmation and automatic synchronization of on-site actions back to the digital twin. This forms a closed information loop that continuously refines perception and decision performance. Although the framework is tailored for SCP operations, its logic is conditionally transferable to other semi-enclosed, standardized construction settings, provided adequate BIM fidelity, sensor density and power supply, workforce training, and data governance are ensured.

In this study, the focus extends beyond constructing a digital framework to validating its engineering effectiveness and practical feasibility under real construction conditions. To this end, the proposed DT × Transformer system is evaluated through a comprehensive field deployment that links perception, reasoning, and on-site execution into a verifiable feedback loop. The assessment emphasizes measurable improvement rather than theoretical novelty – examining how multi-source sensing, digital-twin integration, and large-model decision support collectively enhance project performance. The evaluation strategy integrates controlled comparisons and stepwise field implementation to quantify gains in safety, schedule stability, communication efficiency, and traceability. Beyond performance metrics, attention is given to reproducibility and transparency: the study documents configuration details, data-alignment protocols, and validation criteria to enable subsequent replication and multi-project generalization in future applications of intelligent construction systems.

2. Literature review

Digitalization in construction remains far less mature than in manufacturing and ICT industries. Data fragmentation and low interoperability cause up to 48 % of project rework and multi-trillion-dollar losses each year (Ammar et al., 2022). Despite the introduction of BIM, IoT, and project-management tools, most digital platforms remain static, manually updated, and disconnected from real-time

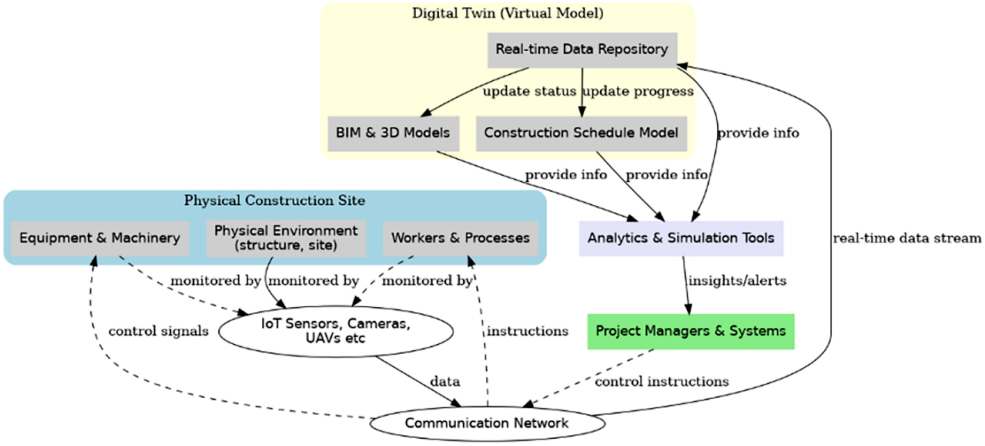

Figure 1. Conceptual architecture of digital twin in the construction phase

operations (Abioye et al., 2021; Oesterreich & Teuteberg, 2016; Zulu et al., 2023). Grieves (2024) emphasized that legacy systems and institutional conservatism still limit transparency and hinder the use of analytics for proactive decision-making. These barriers highlight the need for feedback-driven and context-aware digital ecosystems capable of linking on-site perception with dynamic control.

Originating from aerospace and manufacturing, the Digital Twin (DT) concept synchronizes virtual and physical entities through sensors, IoT, and cyber-physical feedback (Grieves, 2024; Wu et al., 2023; Z. Yang et al., 2024). As shown in Figure 1, DT maturity progresses from static visualization to full bidirectional synchronization, integrating BIM, schedules, and sensor data for real-time reasoning (Deng et al., 2021; Jiang et al., 2024). Early implementations proved effective in workflow monitoring, safety warning, and logistics coordination – Lu et al. (2020) built a campus-scale DT for operational management; Vassena et al. (2023) realized automated progress tracking via IoT–BIM fusion; Xie et al. (2020) detected environmental anomalies from sensor data; and Greif et al. (2020) optimized prefabrication logistics. Despite these advances, most DT systems remain monitor-centric, offering visualization without generating constraint-aware, executable strategies or measurable feedback. Their inability to close the loop between "perception → decision → execution" limits tangible performance gains (Alsakka et al., 2024; Boje et al., 2020; Fan et al., 2021; Sacks et al., 2020).

In parallel, Transformer-based foundation models have revolutionized multimodal representation learning. Introduced by Vaswani et al. (2017), Transformers capture long-range dependencies through self-attention mechanisms, enabling unified processing of text, imagery, and temporal data (Wei et al., 2022). Models such as BERT, GPT, and Vision Transformer demonstrate strong potential for semantic reasoning and cross-modal fusion (Tian et al., 2024; Xu et al., 2023). In construction, they have been applied to worker-behavior recognition (Tian et al., 2024), equipment detection (Eum et al., 2025), and safety prediction (Yang et al., 2023; Yoo et al., 2024); as well as cost forecasting (Shi & Shide, 2026) and schedule optimization (Amer et al., 2021; Li et al., 2025). Figure 2 outlines their expanding role in perception, prediction, and scheduling. However, most studies rely on small, task-specific datasets, lack domain-constraint embedding, and rarely integrate real-time feedback, leaving a gap between algorithmic intelligence and practical executability.

Emerging research now explores the integration of DTs and Transformer-based models to establish intelligent, closed-loop systems. As illustrated in Figure 3, this paradigm continuously collects site data, performs reasoning through large models, and returns executable feedback for adaptive control (Sacks et al., 2020). For instance, Jung et al. (2024) aligned image captions with project schedules within a DT to detect deviations; and Cai et al. (2024) embedded reinforcement learning into a DT for tower-crane path optimization. Nonetheless, persistent issues remain: data heterogeneity and weak multimodal alignment across video, sensor, BIM, and text streams (Huang et al., 2023; Yitmen et al., 2023); limited engineering-knowledge embedding, which leads to idealized yet infeasible outputs (Reja et al., 2022); latency and computational cost restricting on-site deployment (Gharaibeh et al., 2024); and a shortage of field-scale validations demonstrating robustness and scalability.

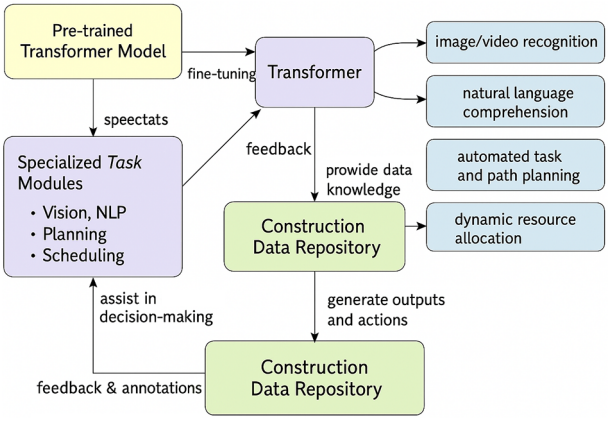

Figure 2. Conceptual architecture of transformer applications in the construction phase

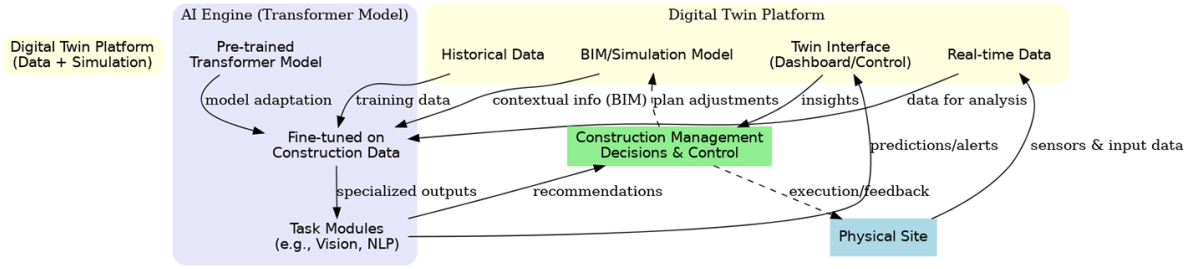

Figure 3. Integration framework of digital twin and foundation model for construction industry

In summary, existing research shows clear progress but also enduring fragmentation. DTs enhance visualization and sensing, while Transformers strengthen perception and prediction; however, their integration into executable and verifiable construction workflows remains nascent. Three critical research gaps persist:

1. Reliable multimodal alignment of heterogeneous data at component and temporal scales;
2. Constraint-aware decision generation respecting safety, resource, and spatial limits; and
3. Executable feedback loops to verify on-site actions and enable continual learning. Addressing these gaps forms the foundation of this study, which develops and validates a DT × Transformer closed-loop intelligent construction system for high-concurrency self-climbing platform (SCP) operations.

3. Methodology

This study takes the high-rise self-climbing platform (SCP) as a representative scenario to construct and validate a closed-loop intelligent construction system based on Digital Twin (DT) and Transformer foundation models. The system aims to achieve semantic fusion and real-time evolution of multi-source data within a unified DT platform, transforming multimodal inputs – sensors, video, BIM, and text – into computable knowledge graphs. These structured representations drive a multimodal Transformer for cognitive reasoning and risk prediction. Guided by engineering constraints, the system generates executable scheduling and safety strategies, which are delivered through AR interfaces for real-time execution and feedback, closing the "perception–cognition–decision–execution–relearning" loop. This chapter outlines the overall architecture, data flow, core technologies, and implementation logic.

3.1. Overall framework

To enable intelligent perception, decision-making, and closed-loop feedback in dynamic construction environments, this study develops a four-layer DT–Transformer architecture (Figure 4), consisting of the physical, data, technology, and functional layers (Nguyen & Adhikari, 2023). Information flows upward through acquisition, integration, reasoning, and execution, while control signals flow downward for feedback and regulation, forming

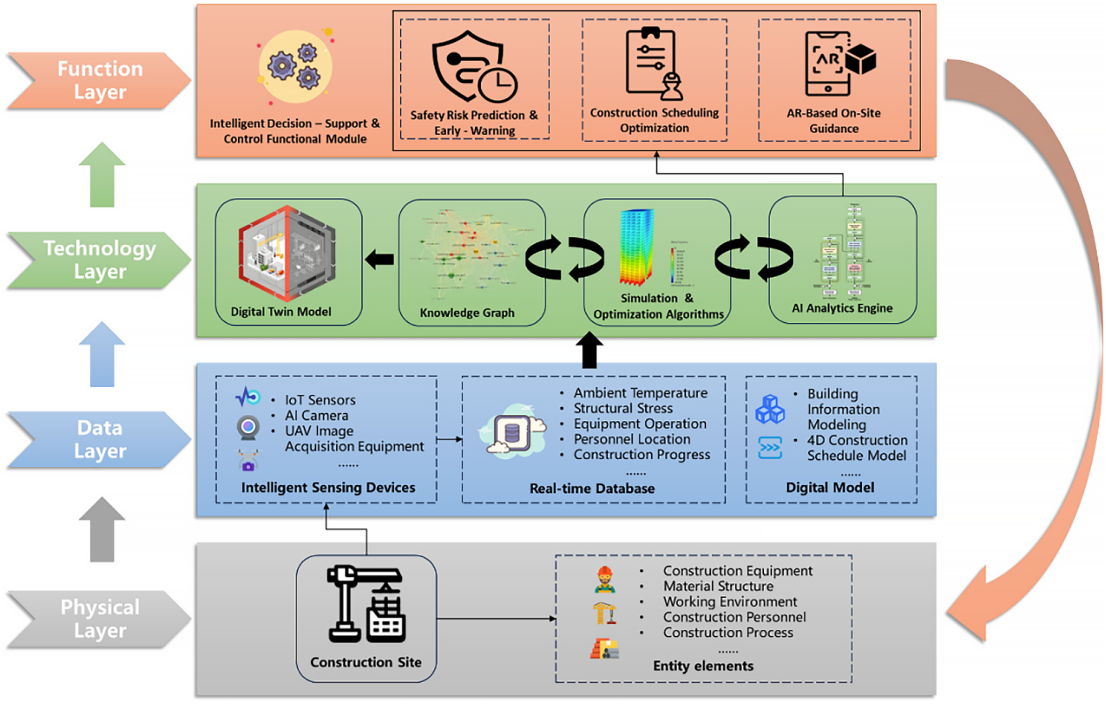

Figure 4. Overall framework of the intelligent construction system

a dynamic loop connecting the physical site and its digital counterpart. This modular and hierarchical framework emphasizes cross-layer data consistency and decision traceability, providing an extensible foundation for intelligent construction.

(1) Physical Layer

The physical layer serves as both the data source and the execution terminal, comprising various sensors, cameras, environmental monitors, and SCP mechanisms. Displacement, acceleration, load, and climate sensors continuously monitor structure, operation, and ambient conditions, while multi-view cameras capture video streams for visual recognition and safety analysis. All devices connect to the data layer in real time via IoT protocols (MQTT/OPC-UA), ensuring high-frequency, low-latency communication. When upper models output scheduling or safety instructions, commands are transmitted through AR devices directly to operators, converting digital decisions into physical actions. Thus, this layer defines both the system's input origin and feedback endpoint, forming the perception–execution boundary of the loop.

(2) Data Layer

Located above the physical layer, the data layer is the core of the DT, responsible for collecting, aligning, and structuring multi-source heterogeneous data (Pan & Zhang, 2021). Using a unified *Component-ID + Timestamp* mechanism, it synchronizes sensor, video, BIM, textual, and environmental data in both space and time. The workflow includes: (1) preprocessing and noise removal via Kalman filtering and IQR detection; (2) sliding-window synchronization for temporal alignment; and (3) semantic labeling embedding multimodal features into the BIM topology to form component-level semantic graphs. This layer transforms fragmented, asynchronous raw data into integrable, computable information flows, supporting DT semantic evolution and Transformer reasoning. It bridges the transition from perception to cognition, providing standardized input interfaces for the upper layers.

(3) Technology Layer

The technology layer functions as the intelligent core, integrating three key modules: DT simulation, knowledge representation, and Transformer reasoning. The DT module builds a real-time BIM-driven virtual scene mapping operational states and risks; it provides a contextual and control layer that integrates data streams with geometric, task, and constraint information to generate operational decisions during construction (Aroquipa et al., 2025). The knowledge module organizes construction logic and spatial constraints in a graph database (Neo4j) (Yoo et al., 2024); and the Transformer processes multimodal inputs – visual, sensory, textual, and structural – to perform state recognition and risk prediction under DT constraints. This layer innovation lies in a data–twin–model synergy, where the DT acts not only as a visualization medium but as a fusion and constraint platform, enabling Transformer reasoning to shift from pure statistical learning toward physically grounded, cognitively informed decision-making.

(4) Functional Layer

The functional layer is the human–machine interaction interface, where AR devices display model outputs – tasks, warnings, and scheduling – directly in the operator's field of view for synchronized virtual–real guidance. Operators provide feedback through gestures, voice, or scanning, while sensors and cameras verify execution and feed results back into the DT for validation. This bidirectional flow supports self-learning: the DT refines state mappings, and the Transformer updates weights through incremental retraining, forming a continuously adaptive cycle. The functional layer thus represents the system's practical intelligence – ensuring that data-driven decisions are both executed and verifiable on site.

3.2. Multi-source sensing and data alignment

In high-rise self-climbing platform (SCP) construction, operations are highly concurrent, spatially congested, and risk conditions change rapidly. Conventional single-source monitoring cannot support such dynamic decision-making. To overcome this, a multi-source sensing network covering five modalities – video, sensor, BIM, textual, and environmental data – is established. A unified *Component-ID + Timestamp* indexing mechanism ensures consistent cross-modal representation and synchronized updates (Figure 5).

The entire process spans four stages – collection, calibration, fusion, and transmission – forming a high-fidelity information flow capable of driving the DT in real time (Figure 6).

At the base level, a dual-channel acquisition architecture integrates video and sensor inputs. On the video side, multi-view HD cameras capture continuous imagery at 25 fps, processed using YOLOv8 and DeepSORT to extract semantic frame sequences of worker behaviors and component motions. On the sensor side, acceleration, displacement, load, and wind-speed nodes connect to an edge gateway via MQTT, sampling at 100 Hz to record transient dynamics. After clock synchronization, both data streams enter a temporal alignment module, where a Cross-Correlation Sliding Window Matching algorithm corrects inter-frequency time offsets. Spatially, PnP reprojection combined with BIM geometry maps pixels to real coordinates, establishing one-to-one correspondence between images and components. After one manual calibration, alignment runs automatically with an average error of ±3 cm.

To maintain robustness and continuity, the pipeline integrates adaptive Kalman filtering and dynamic interpolation. When signals are lost or occluded, similar-window statistics are used for interpolation, while FFT-based spectral constraints suppress high-frequency noise. Construction logs, crew records, and inspection reports are encoded with BERT and keyword-matched with BIM com-

ponent labels and task nodes, binding textual events to their corresponding spatial-temporal entities. This produces a traceable semantic chain linking "video frames–sensor sequences–text descriptions" on a unified temporal-spatial topology.

During fusion, the system generates component-level multimodal feature vectors, including visual, spectral, topological, and semantic attributes (Figure 7). After alignment, these features are packed into multimodal tensors $X_{i,t} = \left[v_{i,t}, s_{i,t}, b_{i,t}, l_{i,t}\right]$ and streamed into the DT database (Kafka + TimescaleDB). Edge nodes handle preprocess-

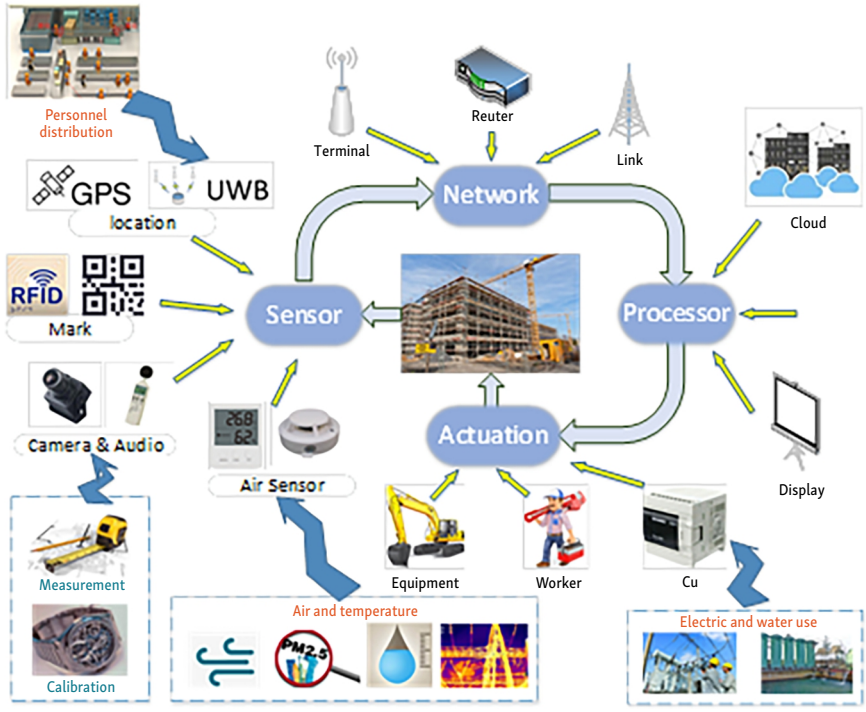

Figure 5. Workflow of multi-source data acquisition and integration

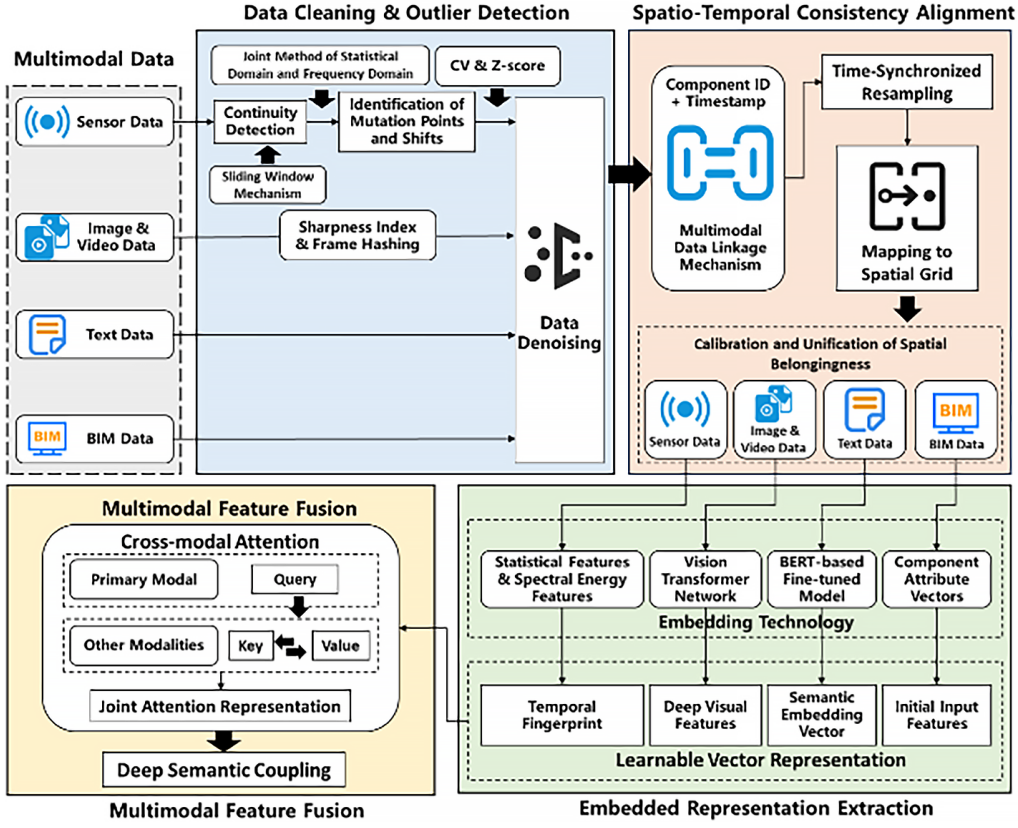

Figure 6. Framework for multi-source data preprocessing and feature representation

ing and encrypted transmission, while the cloud-based DT platform performs asynchronous sequencing and caching. The end-to-end latency averages 1.3 s, satisfying real-time monitoring and inference demands. The resulting semantic dataset continuously feeds the Transformer at the technology layer for task recognition, risk prediction, and strategy optimization, forming a reliable multimodal input foundation for decision-making and feedback.

Through this mechanism, the DT evolves from a static visualization into a central hub for data integration and reasoning. In synergy with the Transformer, it achieves precise semantic alignment, robust perception, and low response latency in complex SCP environments, establishing a solid foundation for closed-loop intelligent construction.

3.3. Semantic fusion and information integration of the digital twin

Following the recent definition by Aroquipa et al. (2025), the Digital Twin (DT) in the construction phase is a *bidirectionally synchronized and semantically integrated* digital representation of on-site assets, processes, and constraints. It enables real-time detection, prediction, and optimization of operations, and issues executable instructions under engineering rules (Aroquipa et al., 2025). Unlike BIM, which provides a mainly static information model, or IoT-based monitoring, which only delivers one-way sensor data, the construction DT forms a *closed information loop* connecting perception, reasoning, and action. In practice, it should ensure (i) low-latency state synchronization, (ii) decision closure from perception to execution with feedback, and (iii) traceable replay of the recognition–decision–execution process.

After completing multi-source sensing and data alignment, a key challenge remains: how to transform heterogeneous data – across modalities, frequencies, and semantics – into interpretable engineering information. Simple data aggregation cannot support real-time cognition and decision-making in complex construction scenarios. To overcome this, a semantic fusion and information integration framework centered on the Digital Twin is established. The DT serves as a unified modeling and evolution platform that integrates video, sensor, BIM, textual, and environmental data in a virtual space, enabling visual cognition, semantic understanding, and logical reasoning of on-site conditions.

Within the overall system, the DT acts as a bridge: downward, it receives multi-source sensing data and structures it into dynamic state streams; upward, it supplies the Transformer with semantically enriched, constraint-aware inputs. This realizes a closed chain of *data perception–semantic modeling–intelligent reasoning–execution feedback*.

```
# Temporal window extraction
frames = frames[-win_v:] if len(frames) >= win_v else np.pad(frames, ((win_v-len(frames),0),(0,0),(0,0),(0,0),(0,0)), mode="edge")
sensor = sensor[-win_s:] if len(sensor) >= win_s else np.pad(sensor, (win_s-len(sensor),0), mode="edge")

# Temporal alignment
vid_signal = frames.mean(axis=(1,2,3))
delay = cross_corr_align(vid_signal, sensor, CFG["video_fps"], CFG["sensor_hz"])
if delay > 0: sensor = np.pad(sensor, (delay,0))[:len(sensor)]
elif delay < 0: sensor = np.pad(sensor, (0,-delay))[(-delay):]

# Spatial alignment to BIM
b_pose = pnp_pixel_to_bim(pixels, depth_m).mean(axis=0)

# Missing data repair
mask_nan = np.random.rand(len(sensor)) < 0.1
sensor[mask_nan] = np.nan
sensor = dynamic_window_interpolate(sensor, max_gap=12)
kf = AdaptiveKalman1D()
sensor = np.array([kf.update(0.0 if np.isnan(v) else float(v)) for v in sensor])

# Multimodal features
v_in = torch.from_numpy(frames[len(frames)//2].transpose(2,0,1)[None].astype(np.float32)/255.).to(device)
v_feat = VisionBackbone().to(device).eval()(v_in).cpu().numpy()[0]
s_feat = sensor_fft_features(sensor, fs=CFG["sensor_hz"], topk=16)
l_feat = TextEncoder().encode(texts).mean(axis=0)[:256]

# TSU assembly
return TSU(
    cid=cid, t0=t_start, t1=t_start + CFG["win_sec"],
    v_feat=v_feat, s_feat=s_feat, b_pose=b_pose, l_feat=l_feat,
    meta={"delay_samples": int(delay),
          "missing_rate": float(mask_nan.mean()),
          "align_win_sec": CFG["win_sec"]}
)
```

Figure 7. Multi-source data fusion implementation code (part)

Constructed with BIM as the core, the DT achieves bidirectional correspondence between the physical site and its virtual counterpart via an IoT sensing network (Figure 8). A multidimensional *Component-ID* + *Timestamp* index continuously writes feature streams from Section 3.2 into the DT repository, unifying video frames, sensor signals, task texts, and component attributes. This achieves geometric, temporal, and semantic alignment, ensuring real-time consistency between the virtual and physical sites (Figure 9). The BIM-based mapping mechanism automatically recognizes component geometry and construction topology, providing the semantic foundation for state evolution and decision reasoning (Peng & Liu, 2023).

Within the DT, a physical–virtual mapping mechanism supports continuous state updating (Figure 10). Time-series sensor data and visual recognition results are written into component nodes. Using Perspective-n-Point (PnP) reprojection, pixel coordinates are mapped to BIM coordinates. Combined with acceleration, displacement, and load inputs, this produces state vectors $S_i(t)$ representing structural behavior and task progress. Through event-driven updates, when signals exceed thresholds, visual anomalies are detected, or delay logs appear, corresponding nodes refresh instantly, and the virtual model synchronously reflects physical changes.

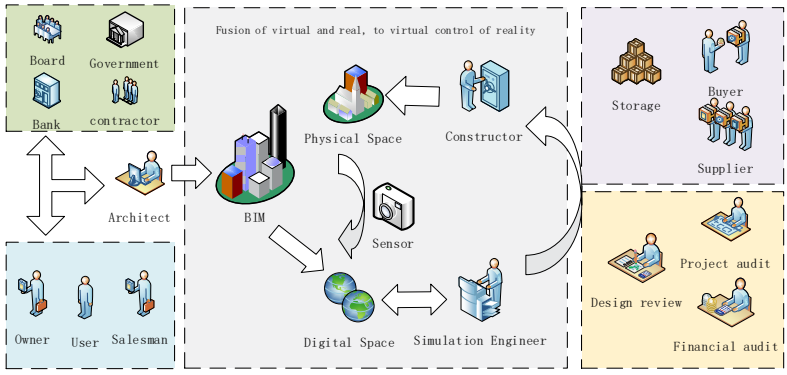

Figure 8. BIM-driven digital twin framework

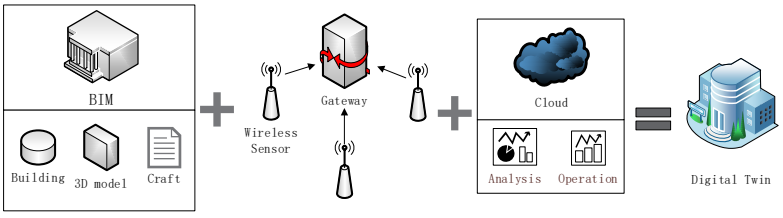

Figure 9. BIM-based mapping mechanism for digital twin implementation

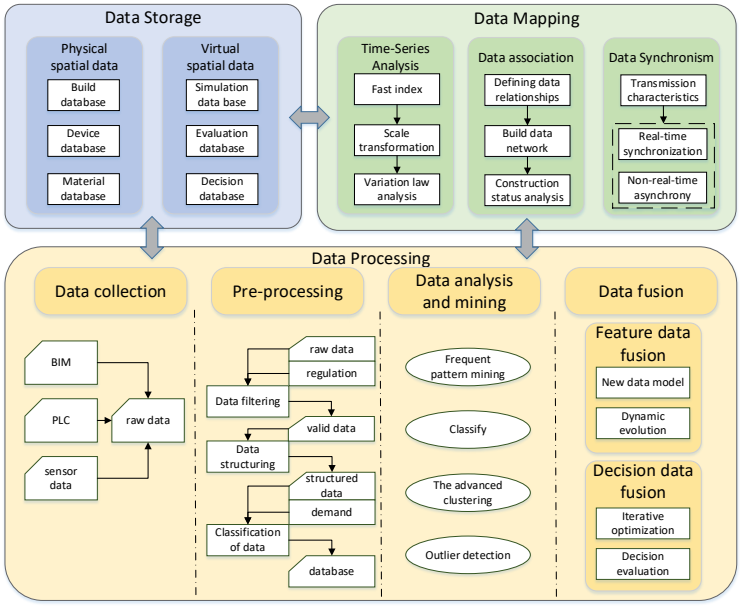

Figure 10. Data mapping between physical and virtual spaces

Building on this foundation, the DT abstracts integrated data into computable knowledge structures. Using BIM topology and construction workflows as the backbone, a multimodal knowledge graph is constructed (Figure 11), where nodes represent components, equipment, and tasks, and edges capture temporal, spatial, and resource dependencies. Each node stores dynamic features, geometric parameters, and semantic descriptions, transforming "data objects" into "knowledge entities". Embedded constraints, such as safety rules and procedural logic, provide a queryable semantic interface for the Transformer.

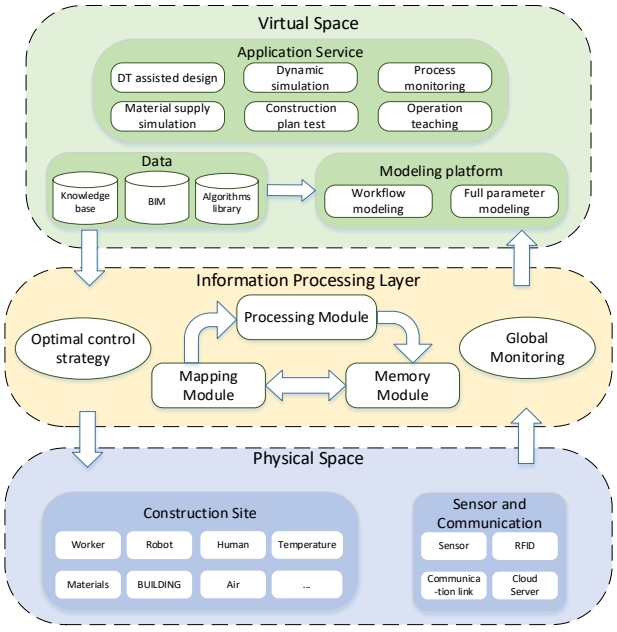

Figure 11. BIM-driven digital twin mapping and coordination process

Through this mechanism, the DT achieves multi-level interoperability: At the bottom, it fuses sensing data with BIM to realize spatiotemporal mapping; In the middle, event-driven updates maintain real-time synchronization between virtual and physical domains; At the top, structured semantics in knowledge-graph form supply contextualized inputs and constraints to the large model (Han & Golparvar-Fard, 2017). Consequently, the Transformer reasons and schedules within the DT's semantic and engineering framework, ensuring that decisions are data-grounded, physically feasible, and operationally meaningful in real construction contexts.

3.4. Cognitive reasoning and decision generation of the foundation model

After semantic fusion and knowledge mapping through the DT, the system advances to intelligent cognition and decision generation. A multimodal Transformer serves as the core reasoning engine, leveraging dynamic state streams and knowledge-graph information from the DT to recognize process states, predict risks, and perform constraint-aware scheduling inference. This establishes a cognitive–decision chain integrating semantic cognition, risk assessment, and strategy generation.

The reasoning process operates within the contextual environment of the DT (Figure 12). Inputs include the dynamically updated component state vectors $S_i(t)$, task semantics, and constraint conditions, unified into multimodal temporal sequences. The Transformer comprises four modules: input encoding, cross-modal attention fusion, constraint-aware reasoning, and strategy generation.

In the input encoding phase, features from each modality are embedded and serialized. Visual semantics are extracted through a dual-channel convolution–temporal

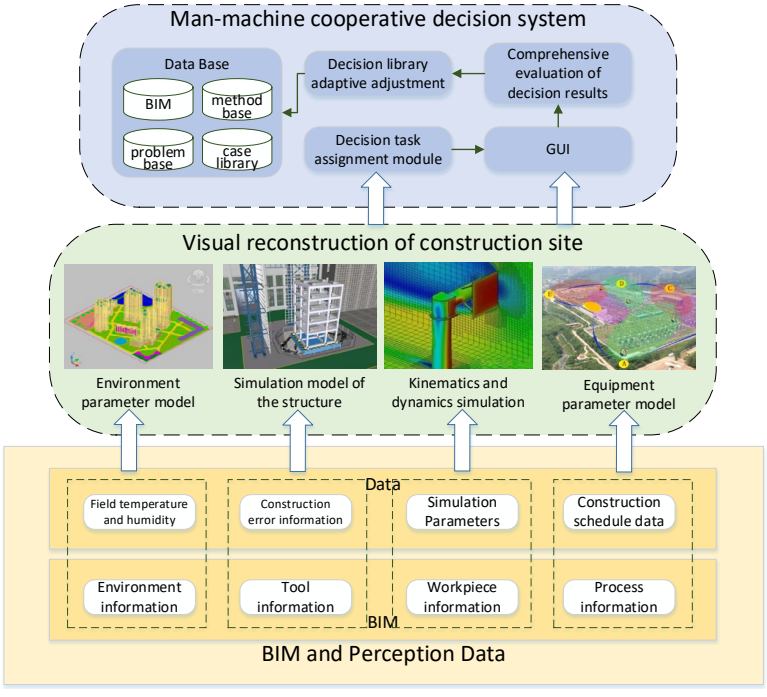

Figure 12. Digital twin decision-making mechanism

encoder; sensor signals are represented by frequency-domain transformation with positional encoding; textual semantics from logs and safety notes are encoded by a pretrained BERT model; and BIM topology is converted into graph-structured positional vectors. All encoded inputs are unified in the Transformer's latent feature space, forming the foundation for attention-based fusion.

During attention fusion and contextual modeling, an enhanced Cross-modal Multi-head Attention mechanism is employed, guided by spatial and procedural priors from the DT's knowledge graph. Dynamic adjustment of attention weights allows the model to focus on the most risk-relevant features across time, achieving collaborative perception among visual, sensor, text, and structural modalities. The resulting latent representation captures both temporal evolution and embedded physical and contextual constraints.

In the risk prediction phase, the model estimates short-term risks using time-windowed state sequences. Exploiting the Transformer's self-attention and long-range dependency modeling, the system identifies hazards arising from compound factors, such as local instability induced by combined vibration and wind load, and outputs process-level risk scores $R_i(t)$. These predictions are written back to the DT for state evolution updates and automatic alert triggering.

During constraint-aware strategy generation, constraint data from the DT's knowledge graph serve as conditional inputs for feasibility filtering and optimization of Transformer outputs. A Constraint Satisfaction–based post-processing module encodes safety, resource, and spatial logic constraints as differentiable objectives. Thus, the generated task sequences are both efficient and executable under real conditions. For instance, when a high-risk node is detected, the scheduling module automatically delays related tasks and reallocates resources (Jung et al., 2024), producing optimized instruction sets for field operations (Figure 13).

The system output comprises three components: (1) process-state recognition labels for real-time progress verification; (2) risk-prediction results for proactive intervention and safety alerts; and (3) constraint-compliant task sequences providing operational instructions for subsequent AR-based execution. All outputs include timestamps, confidence intervals, and semantic tags to ensure traceability and verification. Through this mechanism, the Transformer operates as a cognitive decision engine embedded in the DT environment, using contextual data for semantic understanding, performing reasoning under explicit constraints, and generating structured outputs that drive execution. Ultimately, the DT–Transformer synergy enables a closed logical loop from data cognition to engineering decision-making, endowing the construction process with self-perception, adaptive reasoning, and continuous optimization, and providing a robust foundation for subsequent AR-guided execution and feedback.

3.5. Closed-loop decision generation and AR-based execution

Building on described in Section 3.3 and 3.4, the system integrates cognitive reasoning, decision optimization, and physical feedback to establish a closed-loop intelligent construction framework. The multimodal Transformer functions as the cognitive core, while the Digital Twin (DT) provides constraint-aware semantics ensuring that generated strategies are executable under real-world conditions. Dynamic component-state vectors $S_i(t)$, task semantics, and constraint conditions from the DT are unified into multimodal temporal sequences. Through Cross-modal Multi-head Attention architecture, the Transformer encodes visual, sensor, textual, and BIM features into a shared latent space, guided by DT-derived spatial and procedural priors. This enables risk-relevant feature extraction and contextual reasoning under explicit safety and resource constraints.

Given tasks $(\mathcal{I})$, resources $(\mathcal{K})$, precedence pairs $(\mathcal{P})$, and Transformer-inferred short-horizon risk scores $(\hat{\rho}_t \in [0,1])$, the DT-constrained scheduler optimizes start/completion times and assignments under safety, resource, and spatial envelopes. (x_i) (start of task i), (C_i) (completion), $(y_{ik} \in \{0,1\})$ (task–resource assignment), $(z_{ij} \in \{0,1\})$ (ordering: i precedes j), $(r_t \in [0,1])$ (risk throttle at time t).

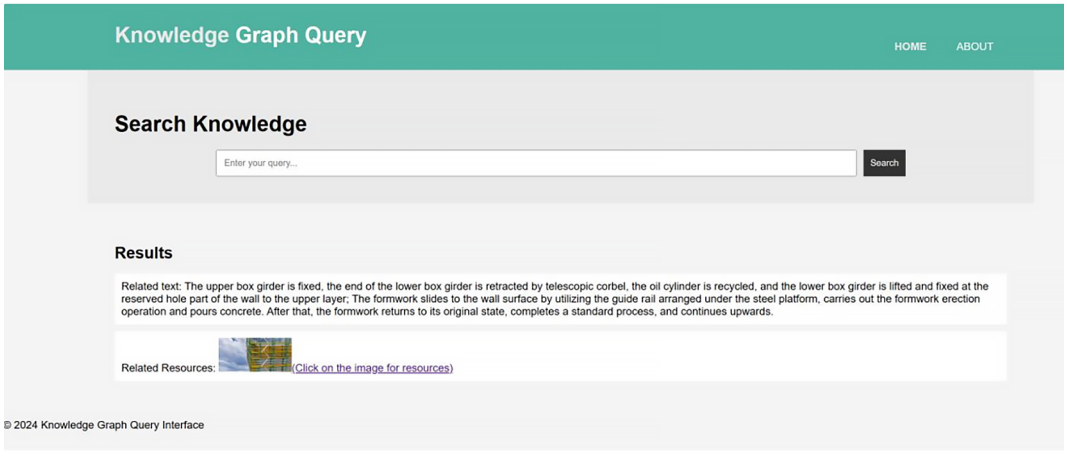

Figure 13. Intelligent construction knowledge query and decision support

Objective:

$$\min \alpha \max_i C_i + \beta \sum_i (x_i - ready_i)^+ + \gamma \sum_t r_t \hat{\rho}_t, \quad (1)$$

where $(\alpha, \beta, \gamma > 0)$ are tuned by cross-validated policy preference.

Constraints:

(1) Precedence & spatial sequencing:

$$C_i \le x_j + M\left(1 - z_{ij}\right),\ z_{ij} + z_{ji} = 1,\ \forall (i,j) \in \mathcal{P}. \quad (2)$$

(2) Resource capacity:

$$\sum_i y_{ik} \mathbf{1}_{t \in [x_i, C_i)} \le \mathrm{cap}_k, \forall k, \forall t. \quad (3)$$

(3) Transformer-informed safety:

$$r_t \ge \mathbf{1}_{\hat{\rho}_t \ge \tau},\ \mathbf{1}_{t \in [x_i, C_i)} \Rightarrow \hat{\rho}_t < \tau_i. \quad (4)$$

(4) Lift/collision avoidance (platform/crane):

$$\mathrm{dist}(i,j,t) \ge d_{\min},\ \forall t \in [x_i, C_i) \cap [x_j, C_j). \quad (5)$$

(5) Work windows & shifts:

$$x_i \in \bigcup_{s \in \mathcal{S}_i} [\underline{w}_{is}, \bar{w}_{is}],\ \sum_k y_{ik} = 1. \quad (6)$$

We use a two-stage hybrid: the Transformer emits $(\hat{\rho}_t)$ and a valid-action sampler; then a MILP/CP-SAT layer ensures feasibility & optimality under DT constraints (see Algorithm 1). This preserves learned priors while guaranteeing constraint satisfaction before AR dispatch.

Validated instructions are remapped via the Component-ID + Timestamp mechanism into structured AR packets containing spatial coordinates, task windows, and safety cues. Through low-latency transmission, these are visualized on AR terminals (e.g., HoloLens 2) as overlaid component highlights and hazard zones (Figure 14).

During execution, AR devices with RGB-D and IMU sensors capture operator actions and site dynamics. A lightweight Vision Transformer matches actual behaviors with DT task nodes, detecting deviations or unsafe operations. Feedback events – tagged with time, space, and confidence – are written into the DT for state updates and incremental model retraining, completing the perception–decision–execution–relearning loop (Alkan & Basaga, 2023).

Figure 14. AR-based on-site construction guidance and human–machine interaction system

The DT provides structured semantics and constraints; the Transformer performs reasoning and optimization; AR enables human–machine collaboration and feedback (Wang et al., 2014). Together they realize a verifiable detect–predict–optimize–execute–learn cycle, transforming cognitive insights into auditable on-site actions and advancing construction toward autonomous, self-correcting operation.

4. Case study: Self-climbing platform for high-rise construction

4.1. Background and system overview

This study selects a landmark high-rise office building located in the Central Business District of Wuhan, China, as the case project. Situated along the Yangtze River and surrounded by dense pedestrian and vehicular traffic, the site imposes stringent constraints on logistics organization during construction. The building reaches approximately 200 meters in height, comprising 60 stories above ground and four basement levels. Its complex structure, long construction period, and high-quality requirements make it a representative testbed for intelligent construction applications. The building's central core adopts a cast-in-place reinforced concrete structure equipped with an advanced Self-Climbing Platform (SCP) system. The SCP integrates a hydraulic synchronization module, intelligent lifting mechanism, high-strength formwork system, and centralized control interface, as illustrated in Figure 15.

In such large-scale and complex high-rise projects, achieving real-time perception, coordinated control, and closed-loop optimization across multiple concurrent operations remain a longstanding challenge. Conventional scaffolding and manual scheduling methods struggle to meet the stringent safety, efficiency, and coordination requirements of high-rise core construction. As a hydraulic-driven system anchored to the core wall, the SCP integrates formwork support, material transport, personnel

Figure 15. Schematic of self-climbing platform operation in high-rise construction

passage, and equipment operation, autonomously climbing as the structure progresses. This mechanized system allows concrete pouring, formwork assembly, and hoisting operations to proceed efficiently without reliance on tower cranes, significantly improving safety and spatial utilization. However, the SCP operation environment is characterized by high density, high risk, and high integration. Multiple trades – including rebar tying, formwork assembly, concrete pouring, and electrical installation – must operate within confined elevated spaces, often in overlapping workflows. These conditions amplify the likelihood of hazards such as hydraulic asynchrony, lifting collisions, or overloading. Consequently, developing an intelligent system capable of dynamic monitoring, predictive reasoning, and adaptive scheduling is essential for improving both safety and productivity.

The DT × Transformer intelligent system developed in this study (see Figure 16) provides an ideal framework for such verification. To achieve bidirectional connectivity between the physical site and the virtual decision platform, the system establishes a closed-loop process encompassing *perception–modeling–reasoning–scheduling–feedback*. Through this workflow, it enables real-time monitoring, state recognition, risk prediction, task optimization, and on-site execution in a unified manner. It should be noted that the system validation in this study focuses primarily on the localized SCP operation unit rather than the entire construction life cycle or multi-trade coordination. Due to limitations in sensor coverage density, computational resources, and the available construction window, the experimental validation centers on multi-source data fusion and intelligent decision-making within a confined structural area. The objective is to demonstrate the feasibility and performance potential of the proposed technical framework rather than to fully replace existing construction management systems.

4.2. System functional modules

Within the integrated architecture, the proposed system is organized into four functional modules (see Figure 17) that collectively support an intelligent closed loop from real-time monitoring to decision-making.

(1) Structural and Safety Status Monitoring and Prediction

This module continuously acquires multi-source sensor and video data to identify process states and issue early warnings through image recognition and deep learning–based predictive models. By leveraging an attention mechanism for spatiotemporal semantic fusion, the system performs joint analysis of hydraulic cylinder displacement, platform load, and operation video streams. When detecting anomalies such as displacement deviation, asynchronous vibration, or hazardous postures, the model automatically triggers alerts and generates event reports that are synchronized in real time with the digital twin platform.

(2) Construction Scheduling Optimization and Resource Reallocation

Building on the perception outputs, the system constructs a dynamic operation graph within the DT and performs constraint-aware task sequencing and resource allocation optimization for key processes such as hoisting, pouring, and formwork assembly (Yao et al., 2024). Through the Transformer-based policy generation module, the system explicitly incorporates safety envelopes, resource capacity limits, and spatial precedence constraints (Boje et al., 2020) to ensure that all generated schedules are both efficient and executable. The scheduling logic and optimization scenario for the SCP operation zone are shown in Figure 18.

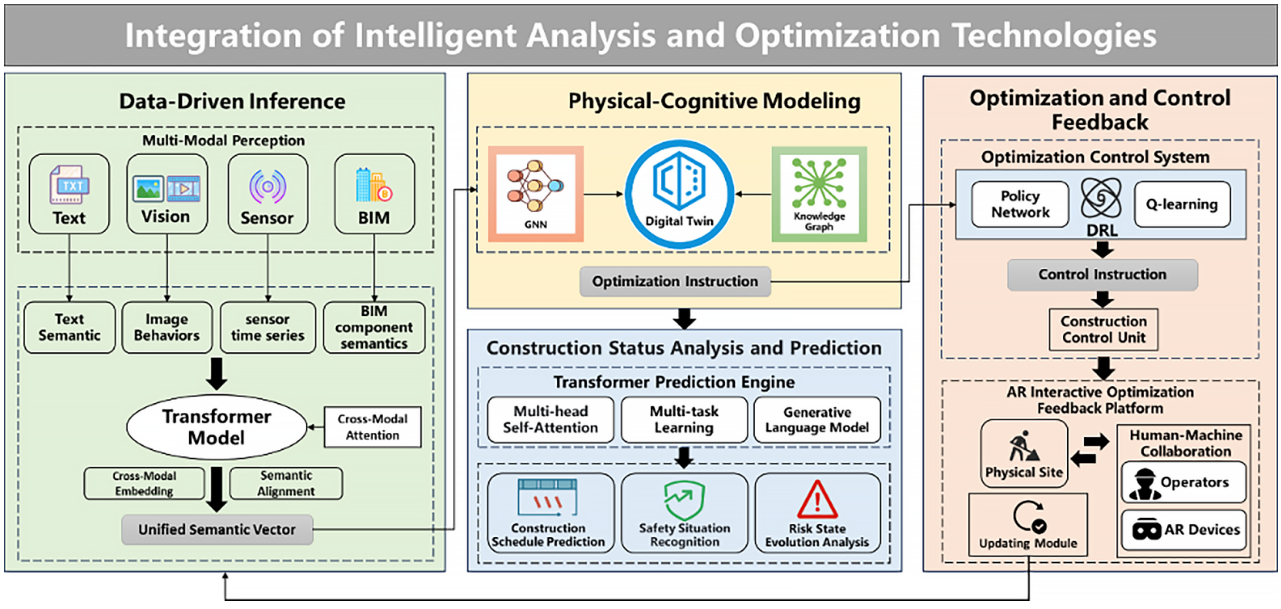

Figure 16. Integrated system for intelligent analysis and optimization in construction

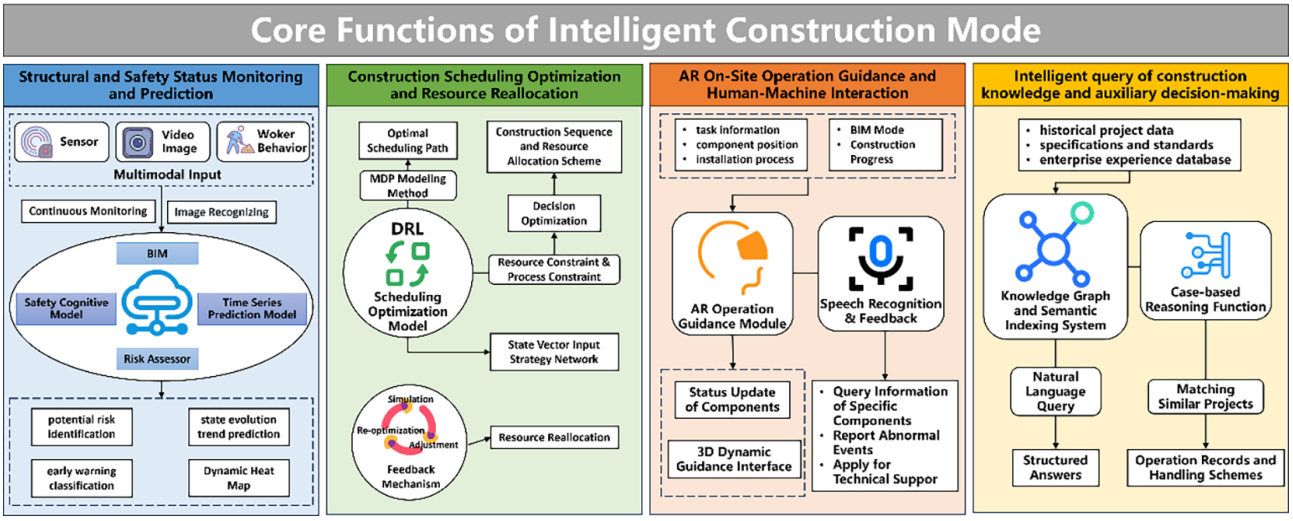

Figure 17. Schematic of core functional modules in the intelligent construction model

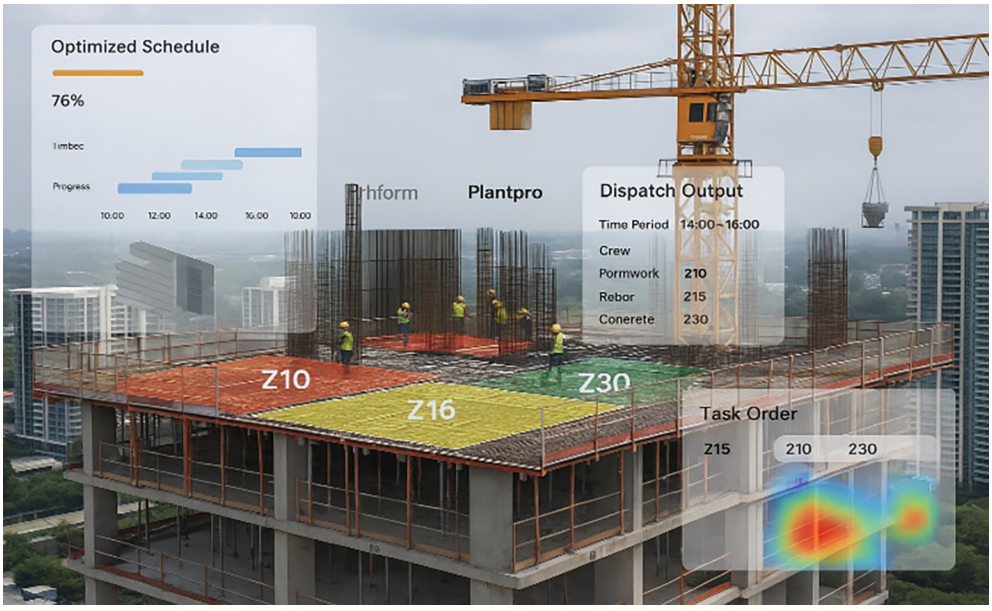

Figure 18. Schematic diagram of the application scenario of the scheduling optimization system in the platform operation area

(3) AR-based On-site Operation Guidance and Human–Machine Interaction

On site, operators wear AR glasses (e.g., HoloLens 2) or use tablet terminals to receive real-time task instructions and safety notifications issued by the system. The AR interface overlays semi-transparent visual elements, such as key components, lifting points, load paths, and safety zone boundaries – directly onto the operator's field of view (see Figure 19). It also supports voice confirmation and gesture feedback to ensure operational consistency between the virtual plan and on-site execution (Jiang et al., 2021). All operation data and feedback results are uploaded synchronously to the DT platform, forming a "command–execution–feedback" closed loop.

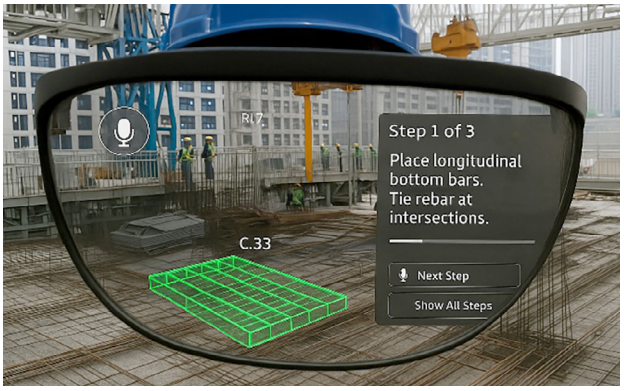

Figure 19. Schematic of the AR interface from the operator's point of view

(4) Intelligent Construction Knowledge Query and Decision Support

This module enables intuitive management and decision-making through natural-language interaction with the project knowledge graph and large language model. Managers can query the system to retrieve risk patterns, equipment maintenance logs, task execution records, and historical comparison reports using plain language. The system integrates multimodal information from the DT and LLMs to deliver comprehensive decision support (Chen et al., 2023). In the experimental setup, both hardware and software modules are centrally managed through the DT middleware, establishing a complete data–model–execution feedback chain that integrates data acquisition, model inference, task generation, and AR-based confirmation. To ensure reproducibility and transparency, all model parameters, scheduling constraints, and sensor calibration coefficients are automatically recorded in the system log, forming a traceable and verifiable experimental archive.

4.3. System verification

To validate the feasibility and overall benefits of the proposed DT × Transformer intelligent construction system in a real construction environment, a comprehensive verification was conducted focusing on the system's four functional modules and their collective enhancement of construction performance–covering safety management, operational efficiency, communication and execution, and information transparency. The experiment adopted a staged implementation under consistent working conditions, enabling controlled comparisons and cross-period interventions.

The validation was carried out within the self-climbing platform (SCP) operation area of the high-rise core tube, where a closed-loop experimental environment featuring observability, interactivity, and verifiability was established. The selected test zone, located between the 18th and 20th floors, covered approximately 210 m^2 and included the hydraulic synchronization system, formwork hoisting area, material staging zone, and operation corridor. A digital twin system integrating multi-source data interfaces and foundation-model reasoning modules was deployed, as illustrated in Figure 20.

The deployment comprised both hardware and software components. On the hardware side, eight 1080p industrial cameras and twelve sets of acceleration, displacement, and load sensors were installed, operating respectively at 25 fps and 100 Hz sampling rates. On the software side, a 4D BIM model, task schedules, and log texts were integrated into the DT using a unified *Component-ID + Timestamp* indexing mechanism. All data streams were aggregated via an on-site edge gateway and processed on a local cloud platform for fusion and inference. Bidirectional communication between the physical and virtual layers was achieved through MQTT and WebSocket protocols. To evaluate system performance comprehensively, four categories of key performance indicators (KPIs) were defined: (1) Safety leading indicators, including warning lead time (the time gap between system warning and event occurrence), violation frequency per thousand work-hours, and near-miss reporting rate; (2) Schedule and resource coordination indicators, such as average cycle time per floor, process waiting ratio, and variance of labor and equipment utilization; (3) Communication and execution efficiency indicators, including end-to-end decision latency (from risk identification to confirmation), instruction misunderstanding rate, and redundant command frequency; and (4) Transparency indicators, including traceability rate with complete evidence chains and task-level audit coverage.

4.3.1. Safety leadership

Data were obtained from three sensor types (displacement, load, acceleration) and video anomaly triggers. A "risk event" was defined as any signal exceeding its process threshold for ≥ 1 s, as specified in technical manuals. "Warning time" corresponded to the system's first issuance

Figure 20. Architecture of the digital twin and multi-source integration for self-climbing platform operations

of a risk alert on the DT platform. The warning lead time was thus calculated as:

$$Lead = Event\ Time - Warning\ Time, \quad (7)$$

stratified by event type.

To prevent selection bias, an event database was frozen before experimentation, accepting only automatically logged and numbered events. A 20% random sample underwent double-blind review, jointly verifying event timestamps using both video frames and raw sensor curves. Among 68 auditable events (27 formwork hoisting, 14 jacking, 12 material transfer, 15 edge operations), the median warning lead time increased from 1.7 s (IQR 0.8–2.6) in the baseline phase to 4.5 s (IQR 3.2–5.7) during intervention, yielding a difference-in-differences (DiD) effect of 2.6 s (95% CI 1.9–3.3, $p < 0.01$). The violation rate per thousand work-hours declined from 6.6 to 3.1. As shown in Figure 21, on June 12 at 15:42, a sudden 0.6 s step was detected in the P12 lifting-point load curve at 2.3 t. The system issued an alert at 15:42:18.4, while the manually verified event time was 15:42:22.9, confirming a 4.5 s warning lead.

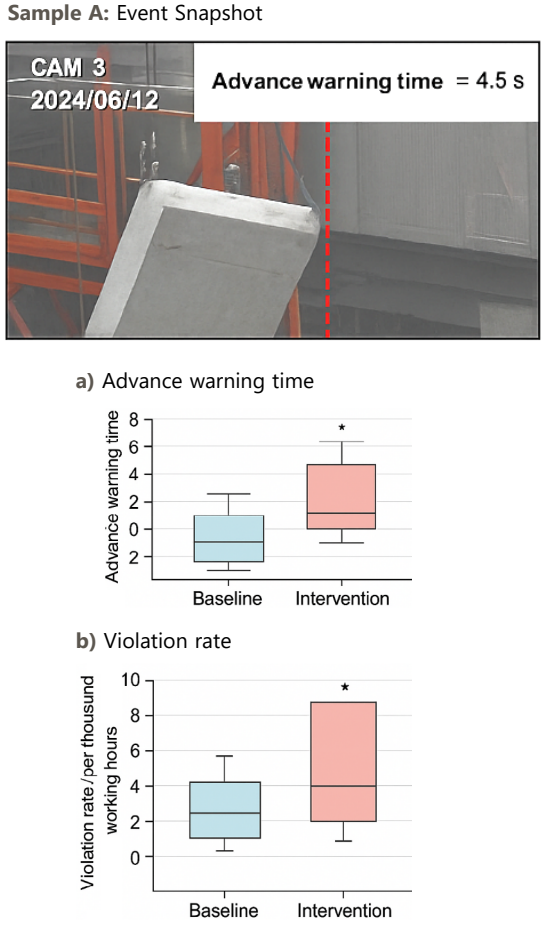

Figure 21. Safety leading indicators: early warning advance and violation rate

4.3.2. Schedule and resource synergy

Before system deployment, the site relied on manual scheduling and CPM-based plans, which frequently led to task overlaps, tower-crane conflicts, and idle waiting times. After implementation, the DT platform continuously synchronized video, sensor, and BIM progress data, automatically identifying current operational states. Predictions were then fed into the constraint-aware scheduling module to generate next-day task plans, incorporating safety, spatial, and resource boundaries. Plans were reviewed by supervisors and distributed to crews through the AR interface for immediate execution.

A stepwise wedge design was adopted, progressively deploying the system from the 17th to 20th floors to maintain consistency in crew composition, weather conditions, and concrete batch supply. Each floor's data collection period covered a complete three-day work cycle. Task logs, crane loads, and labor attendance curves were automatically sampled and verified on site. Results showed that the average single-floor cycle time was shortened to 36.4 hours, and workflow variance decreased significantly. The mean process waiting time dropped from 10.8 hours to 6.2 hours. A representative case (see Figure 22) occurred in the third week of June: on the 19th floor, rebar tying and formwork assembly overlapped, causing 45 minutes of waiting due to tower-crane contention. In contrast, the system-optimized plan for the 20th floor automatically detected this risk, rescheduled the hoisting window by 40 minutes, and adjusted the concrete-pouring slot, reducing the crane load peak from 0.92 to 0.74 and eliminating task interference.

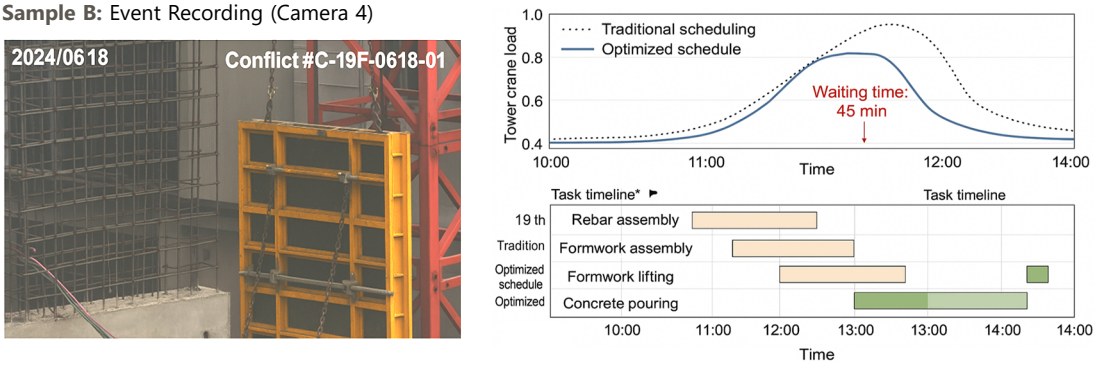

Figure 22. Schedule and resource synergy under the new system

4.3.3. Communication and execution efficiency

Information delays and misinterpretations are common in high-rise operations, often resulting in redundant or ineffective commands. To assess improvements in decision timeliness, end-to-end decision time – the duration from risk identification (system or manual) to on-site acknowledgment – was used as a key metric. The experiment was conducted across the 18th–20th floors in sequential phases. The baseline phase used traditional voice calls and instant-messaging coordination; during the intervention phase, the system automatically generated task cards, verified by safety officers, and issued them through AR terminals. The system automatically logged timestamps for recognition, decision generation, and acknowledgment, alongside voice/video archives for validation. Random stratified sampling (120 events per floor) was applied to minimize bias. Taking the 20th floor as an example, the median end-to-end time decreased from 11.6 s (baseline) to 4.8 s (intervention), as shown in Figure 23. Over 80% of the improvement stemmed from the elimination of delays in decision transmission and acknowledgment. Instant visual prompts and confirmation via AR greatly reduced redundant or misinterpreted instructions: the rate of repeated or withdrawn commands dropped from 0.33 to 0.18 per task.

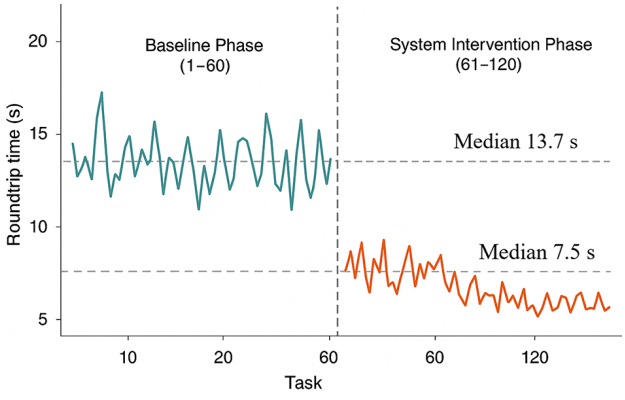

Figure 23. Statistics of decision time for single-layer tasks (taking 20 layers as an example)

4.3.4. Transparency and reproducibility

To verify the system's advantages in data governance and auditability, the platform automatically generated a digital evidence-chain manifest for each medium- and high-risk task. Each manifest included data-slice indices (sensor and video frame numbers), model version, strategy-generation parameters, constraint summaries, AR instruction screenshots, and execution timestamps. All records were uploaded to the local cloud storage with unique cryptographic hashes to prevent tampering (see Figure 24).

For transparency assessment, 50 decision records were randomly selected for playback verification. Using the replay function, the DT could reconstruct the exact perception context, reasoning path, and feedback at the time of decision. Manual cross-validation confirmed 48 fully consistent results (96%). Compared with traditional manual logs, the new method located complete decision evidence within two minutes, transforming verification from result review to process traceability.

4.4. Quantitative evaluation

To further verify the performance of the proposed DT × Transformer intelligent construction system, this section conducts a quantitative comparison with the conventional planning approach. The traditional method combines manual experience with the Critical Path Method (CPM) for scheduling. Both approaches were tested under identical construction conditions to evaluate improvements in efficiency, accuracy, and traceability.

Experimental data were collected from six consecutive construction cycles, with 120 representative task events sampled per floor. For each indicator, the mean ±95% confidence interval (CI) was estimated using the bootstrap method (1,000 resamples), and statistical significance was tested via permutation tests. The evaluation covered seven key performance indicators (KPIs): *warning lead time*, *violation frequency per 1,000 work-hours*, *floor-cycle duration*, *average waiting time*, *end-to-end decision latency*, *repeated-command rate*, and *traceability coverage*. The results are summarized in Table 1.

Figure 24. System continuity files (Template installation examples)

Table 1. Comparison of key performance indicators (Conventional vs. proposed system, 95% CI)

KPI	Conventional (mean ± CI)	Proposed System (mean ± CI)	Δ (Abs)	Δ%	*p*-value
Warning lead time (s)	1.8 ± 0.5	4.6 ± 0.7	+2.8	+156%	<0.01
Violation / 1,000 work-h	6.5 ± 0.8	3.1 ± 0.6	−3.4	−52%	0.01
Floor cycle (h)	43.1 ± 2.3	36.2 ± 1.9	−6.9	−16%	0.02
Waiting time (h)	8.1 ± 1.1	5.3 ± 0.8	−2.8	−34%	0.01
Decision latency (s)	12.3 ± 2.8	6.2 ± 1.7	−6.1	−50%	<0.01
Reissued commands (%)	7.5 ± 1.3	3.1 ± 0.8	−4.4	−59%	0.01
Traceability coverage (%)	68.4 ± 4.1	93.5 ± 2.3	+25.1	+37%	<0.01

The statistical results demonstrate that the proposed system significantly outperforms the conventional method across all metrics ($p < 0.05$). Specifically, the average warning lead time increased from 1.8 ± 0.5 s to 4.6 ± 0.7 s (+156%), the average floor-cycle duration was reduced by 16%, and the waiting time decreased by 34%. The violation frequency per 1,000 work-hours dropped by 52%, while decision latency was reduced from 12.3 ± 2.8 s to 6.2 ± 1.7 s, indicating a substantially faster perception–decision–execution process. Meanwhile, the repeated-command rate fell to 3.1%, and traceability coverage improved to 93.5%, ensuring full-process transparency and data consistency.

At the model level, representative tasks such as worker posture recognition and short-term risk prediction also exhibited high accuracy and stability. Posture recognition achieved an F1-score of 0.88, while short-horizon risk prediction yielded an AUROC of 0.91 and a Brier score of 0.094, indicating well-calibrated confidence outputs consistent with actual risk levels. By embedding semantic fusion and constraint optimization within the scheduling process, the system maintained safety constraints while enhancing both task rationality and response speed.

Overall, the quantitative verification confirms that the proposed Digital Twin–foundation model fusion approach achieves substantial improvements in efficiency, safety, and decision accuracy compared with traditional planning. These findings provide strong empirical evidence for the method's practical value in intelligent construction and its potential for scalable engineering deployment.

5. Discussion

This study developed and validated a DT × Transformer–based intelligent construction framework within a high-rise self-climbing platform (SCP) project. Rather than proposing a universal or finalized solution, the framework offers a technically feasible and verifiable pathway that connects perception, reasoning, and execution in a closed feedback loop. The field validation demonstrates that integrating multi-source perception, digital-twin modeling, Transformer-based reasoning, and AR-assisted execution can produce measurable improvements in safety and operational efficiency. Gains such as earlier risk warnings, reduced violation rates, and smoother task coordination stem from transforming fragmented sensory data into coherent, actionable knowledge. The digital twin provides a semantic and physical backbone that aligns multimodal data under consistent spatiotemporal indices, while the Transformer model enhances contextual understanding and adaptive prediction under explicit safety and resource constraints. Together, they realize a form of operational intelligence that is traceable, explainable, and auditable throughout the construction process.

Despite these promising results, several limitations and trade-offs were observed during practical deployment. High-resolution perception and large-model inference improve precision but increase computational demand and maintenance costs, while strict constraint embedding enhances safety compliance at the expense of scheduling flexibility. Edge computing mitigates latency yet depends on stable power and network conditions. The use of AR interfaces enhances communication accuracy but also relies on workers' adaptation and site discipline. These findings suggest that intelligent construction should aim for augmentation rather than automation – leveraging digital intelligence to enhance human expertise and organizational coordination rather than replace them.

From a methodological perspective, the study emphasizes transparency and reproducibility. All data alignment rules, configuration parameters, and evaluation procedures were recorded within a manifest-driven archive that allows experiment re-execution under identical settings. Although raw data cannot be fully disclosed due to industrial confidentiality, the study provides replicable configuration templates and audit-ready manifests containing dataset hashes, model versions, and parameter seeds. This ensures traceability and methodological openness while respecting data security constraints common in field-scale construction research. Such practices support cumulative progress by enabling other researchers to verify, adapt, or extend the presented framework.

Ethical and privacy considerations are equally central. The system processes visual and positional information that may include personal data; hence, data handling strictly follows principles of minimization and purpose limitation. Video streams are anonymized through on-device blurring and pose-keypoint abstraction, and all access to the digital-twin environment is controlled through role-based authorization and activity logging. Data retention is restricted to the legally required period for safety auditing and then anonymized or deleted. The overall framework

complies with relevant data-protection regimes, including the EU General Data Protection Regulation (GDPR) and China's Personal Information Protection Law (PIPL), ensuring that innovation in AI-driven construction remains ethically responsible and socially acceptable.

Finally, scalability remains a key consideration for extending the framework beyond the tested SCP context. High-fidelity digital-twin synchronization and large-model inference demand considerable computational resources, which may limit adoption in smaller or resource-constrained projects. To address this, a modular edge–cloud hybrid architecture was introduced, where low-level perception is processed locally and global reasoning is executed in the cloud. Future research will explore model compression, lightweight deployment, and adaptive zoning to improve scalability across diverse project types.

Overall, the DT × Transformer system represents an early but meaningful step toward data-driven and constraint-aware construction management. Its significance lies not in full automation but in creating a verifiable feedback structure where planning and execution become mutually transparent. By coupling digital twins with foundation models to achieve measurable, explainable, and repeatable improvements, the study contributes both a conceptual foundation and an empirical reference for the next generation of intelligent construction systems.

6. Conclusions

This study has introduced a next-generation intelligent construction paradigm that integrates digital twin technology with a Transformer-based policy network, establishing an end-to-end system capable of state perception, risk prediction, adaptive scheduling, and AR-enabled human–machine collaboration. The system is distinguished by its ability to unify multi-source sensing, semantic modeling, and AI-driven decision-making into a closed-loop workflow, tailored to the operational demands of complex high-rise self-climbing platform scenarios.

The major contributions of this research are threefold: (1) proposing a scalable integration framework that bridges the current gap between digital twin applications and Transformer-based analytics in construction; (2) demonstrating the feasibility of real-time, data-informed decision-making through on-site validation in a high-risk, high-density work environment; and (3) establishing a foundation for human–machine collaborative execution via AR-enabled interfaces.

Beyond its immediate implementation, this work offers a transferable methodology for diverse construction typologies, contributing to the digital transformation of the architecture, engineering, and construction (AEC) sector. By providing both a conceptual framework and a validated system application, the study lays the groundwork for future intelligent construction systems that are adaptive, interoperable, and capable of continuous self-optimization.

Data availability statement

Some or all data, models, or code generated or used during the study are proprietary or confidential in nature and may only be provided with restrictions.

Funding

This work was supported by National Key R&D Program of China (grant No. 2022YFC3802201).

References

Abioye, S. O., Oyedele, L. O., Akanbi, L., Ajayi, A., Davila Delgado, J. M., Bilal, M., Akinade, O. O., & Ahmed, A. (2021). Artificial intelligence in the construction industry: A review of present status, opportunities and future challenges. *Journal of Building Engineering*, *44*, Article 103299. https://doi.org/10.1016/j.jobe.2021.103299

Alkan, I. B., & Basaga, H. B. (2023). Augmented reality technologies in construction project assembly phases. *Automation in Construction*, *156*, Article 105107. https://doi.org/10.1016/j.autcon.2023.105107

Alsakka, F., Yu, H., El-Chami, I., Hamzeh, F., & Al-Hussein, M. (2024). Digital twin for production estimation, scheduling and real-time monitoring in offsite construction. *Computers & Industrial Engineering*, *191*, Article 110173. https://doi.org/10.1016/j.cie.2024.110173

Amer, F., Jung, Y., & Golparvar-Fard, M. (2021). Transformer machine learning language model for auto-alignment of long-term and short-term plans in construction. *Automation in Construction*, *132*, Article 103929. https://doi.org/10.1016/j.autcon.2021.103929

Ammar, A., Nassereddine, H., AbdulBaky, N., AbouKansour, A., Tannoury, J., Urban, H., & Schranz, C. (2022). Digital twins in the construction industry: A perspective of practitioners and building authority. *Frontiers in Built Environment*, *8*, Article 834671. https://doi.org/10.3389/fbuil.2022.834671

Aroquipa, H., Hurtado, A., Murga, C., De La Cruz, R., & Tarque, N. (2025). Towards smart cities: Foundational methodology for implementing intelligent circular resilience in heritage buildings through structural health monitoring and digital-twins – part a. *International Journal of Disaster Risk Reduction*, *128*, Article 105749. https://doi.org/10.1016/j.ijdrr.2025.105749

Boje, C., Guerriero, A., Kubicki, S., & Rezgui, Y. (2020). Towards a semantic construction digital twin: Directions for future research. *Automation in Construction*, *114*, Article 103179. https://doi.org/10.1016/j.autcon.2020.103179

Bureau of Labor Statistics. (n.d.), *Construction deaths due to falls, slips, and trips increased 5.9 percent in 2021*. https://www.bls.gov/opub/ted/2023/construction-deaths-due-to-falls-slips-and-trips-increased-5-9-percent-in-2021.htm

Cai, B., Ye, Z., Chen, S., & Liang, X. (2024). Reducing safety risks in construction tower crane operations: A dynamic path planning model. *Applied Sciences*, *14*(22), Article 10599. https://doi.org/10.3390/app142210599

Chen, Q., Long, D., Yang, C., & Xu, H. (2023). Knowledge graph improved dynamic risk analysis method for behavior-based safety management on a construction site. *Journal of Management in Engineering*, *39*(4), Article 4023023. https://doi.org/10.1061/JMENEA.MEENG-5306

Construction Dive. (n.d.). *Why construction productivity growth is lagging—And what to do about it | construction dive*. https://www.constructiondive.com/news/why-construction-productivity-lags-mckinsey/736082/

Deng, M., Menassa, C. C., & Kamat, V. R. (2021). From BIM to digital twins: A systematic review of the evolution of intelligent building representations in the AEC-FM industry. *Journal of Information Technology in Construction*, *26*, 58–83. https://doi.org/10.36680/j.itcon.2021.005

Eum, I., Kim, J., Wang, S., & Kim, J. (2025). Heavy equipment detection on construction sites using you only look once (YOLO-version 10) with transformer architectures. *Applied Sciences*, *15*(5), Article 2320. https://doi.org/10.3390/app15052320

Fan, C., Zhang, C., Yahja, A., & Mostafavi, A. (2021). Disaster city digital twin: A vision for integrating artificial and human intelligence for disaster management. *International Journal of Information Management*, *56*, Article 102049. https://doi.org/10.1016/j.ijinfomgt.2019.102049

Gharaibeh, L., Matarneh, S., Lantz, B., & Eriksson, K. (2024). Quantifying the influence of BIM adoption: An in-depth methodology and practical case studies in construction. *Results in Engineering*, *23*, Article 102555. https://doi.org/10.1016/j.rineng.2024.102555

Greif, T., Stein, N., & Flath, C. M. (2020). Peeking into the void: Digital twins for construction site logistics. *Computers in Industry*, *121*, Article 103264. https://doi.org/10.1016/j.compind.2020.103264

Grieves, M. (2024). Intelligent digital twins and the development and management of complex systems. *Digital Twin*, *1*(1), Article 8. https://doi.org/10.12688/digitaltwin.17574.1

Han, K. K., & Golparvar-Fard, M. (2017). Potential of big visual data and building information modeling for construction performance analytics: An exploratory study. *Automation in Construction*, *73*, 184–198. https://doi.org/10.1016/j.autcon.2016.11.004

Huang, Y., Tao, J., Sun, G., Wu, T., Yu, L., & Zhao, X. (2023). A novel digital twin approach based on deep multimodal information fusion for aero-engine fault diagnosis. *Energy*, *270*, Article 126894. https://doi.org/10.1016/j.energy.2023.126894

Jiang, F., Ma, L., Broyd, T., & Chen, K. (2021). Digital twin and its implementations in the civil engineering sector. *Automation in Construction*, *130*, Article 103838. https://doi.org/10.1016/j.autcon.2021.103838

Jiang, Y., Su, S., Zhao, S., Zhong, R. Y., Qiu, W., Skibniewski, M. J., Brilakis, I., & Huang, G. Q. (2024). Digital twin-enabled synchronized construction management: A roadmap from construction 4.0 towards future prospect. *Developments in the Built Environment*, *19*, Article 100512. https://doi.org/10.1016/j.dibe.2024.100512

Jung, Y., Hockenmaier, J., & Golparvar-Fard, M. (2024). Transformer language model for mapping construction schedule activities to uniformat categories. *Automation in Construction*, *157*, Article 105183. https://doi.org/10.1016/j.autcon.2023.105183

Li, F., Lang, S., Tian, Y., Hong, B., Rolf, B., Noortwyck, R., Schulz, R., & Reggelin, T. (2025). A transformer-based deep reinforcement learning approach for dynamic parallel machine scheduling problem with family setups. *Journal of Intelligent Manufacturing*, *36*(7), 4735–4768. https://doi.org/10.1007/s10845-024-02470-8

Lu, Q., Parlikad, A. K., Woodall, P., Don Ranasinghe, G., Xie, X., Liang, Z., Konstantinou, E., Heaton, J., & Schooling, J. (2020). Developing a digital twin at building and city levels: Case study of West Cambridge campus. *Journal of Management in Engineering*, *36*(3), Article 5020004. https://doi.org/10.1061/(ASCE)ME.1943-5479.0000763

Madubuike, O. C., Anumba, C. J., & Khallaf, R. (2022). A review of digital twin applications in construction. *Journal of Information Technology in Construction*, *27*, 145–172. https://doi.org/10.36680/j.itcon.2022.008

McKinsey. (n.d.). *Challenges in the construction industry*. Retrieved June 29, 2025, from https://www.linkedin.com/pulse/mckinsey-challenges-construction-industry-buromatei-uttrc

Linesight. (n.d.). *Is efficiency eluding the construction industry?* https://www.linesight.com/en-us/insights/is-efficiency-eluding-the-construction-industry

Nguyen, T. D., & Adhikari, S. (2023). The role of BIM in integrating digital twin in building construction: A literature review. *Sustainability*, *15*(13), Article 10462. https://doi.org/10.3390/su151310462

Oesterreich, T. D., & Teuteberg, F. (2016). Understanding the implications of digitisation and automation in the context of industry 4.0: A triangulation approach and elements of a research agenda for the construction industry. *Computers in Industry*, *83*, 121–139. https://doi.org/10.1016/j.compind.2016.09.006

Pan, Y., & Zhang, L. (2021). A BIM-data mining integrated digital twin framework for advanced project management. *Automation in Construction*, *124*, Article 103564. https://doi.org/10.1016/j.autcon.2021.103564

Peng, J., & Liu, X. (2023). Automated code compliance checking research based on BIM and knowledge graph. *Scientific Reports*, *13*(1), Article 7065. https://doi.org/10.1038/s41598-023-34342-1

Reja, V. K., Varghese, K., & Ha, Q. P. (2022). Computer vision-based construction progress monitoring. *Automation in Construction*, *138*, Article 104245. https://doi.org/10.1016/j.autcon.2022.104245

Sacks, R., Brilakis, I., Pikas, E., Xie, H. S., & Girolami, M. (2020). Construction with digital twin information systems. *Data-Centric Engineering*, *1*, Article e14. https://doi.org/10.1017/dce.2020.16

Shi, T., & Shide, K. (2026). A comparative analysis of LSTM, GRU, and transformer models for construction cost prediction with multidimensional feature integration. *Journal of Asian Architecture and Building Engineering*, *25*(1), 634–649. https://doi.org/10.1080/13467581.2025.2455034

Tian, C., Chen, Y., Feng, Y., & Zhang, J. (2024). Fine-tuning vision transformer (ViT) to classify highway construction workers' activities. In *Proceedings of Construction Research Congress 2024* (pp. 1140–1148). ASCE. https://doi.org/10.1061/9780784485262.116

Vassena, G. P. M., Perfetti, L., Comai, S., Mastrolembo Ventura, S., & Ciribini, A. L. C. (2023). Construction progress monitoring through the integration of 4D BIM and SLAM-based mapping devices. *Buildings*, *13*(10), Article 10. https://doi.org/10.3390/buildings13102488

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *31st Conference on Neural Information Processing Systems* (*NIPS 2017*), Long Beach, CA, USA.

Wang, C. C., & Chien, O. (2014). The use of BIM in project planning and scheduling in the australian construction industry. In *Proceedings of ICCREM 2014* (pp. 126–133). ASCE. https://doi.org/10.1061/9780784413777.015

Wang, X., Truijens, M., Hou, L., Wang, Y., & Zhou, Y. (2014). Integrating augmented reality with building information modeling: Onsite construction process controlling for liquefied natural gas industry. *Automation in Construction*, *40*, 96–105. https://doi.org/10.1016/j.autcon.2013.12.003

Wang, C. C., Wang, M., Sun, J., & Mojtahedi, M. (2021). A safety warning algorithm based on axis aligned bounding box meth-

od to prevent onsite accidents of mobile construction machineries. *Sensors*, *21*(21), Article 7075. https://doi.org/10.3390/s21217075

Wang, M., Wang, C. C., Zlatanova, S., Shen, X., & Brilakis, I. (2024). A streamlined laser scanning verticality check method for installation of prefabricated wall panels. *Journal of Construction Engineering and Management*, *150*(11), Article 4024159. https://doi.org/10.1061/JCEMD4.COENG-14989

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., Chi, E. H., Hashimoto, T., Vinyals, O., Liang, P., Dean, J., & Fedus, W. (2022). *Emergent abilities of large language models*. arXiv. https://doi.org/10.48550/arXiv.2206.07682

Wu, J., Peng, L., Sheng, W., Wang, C. C., & Sun, J. (2023). Track gauge measurement based on model matching using UAV image. *Automation in Construction*, *155*, Article 105070. https://doi.org/10.1016/j.autcon.2023.105070

Xie, X., Lu, Q., Rodenas-Herraiz, D., Parlikad, A. K., & Schooling, J. M. (2020). Visualised inspection system for monitoring environmental anomalies during daily operation and maintenance. *Engineering, Construction and Architectural Management*, *27*(8), 1835–1852. https://doi.org/10.1108/ECAM-11-2019-0640

Xu, P., Zhu, X., & Clifton, D. A. (2023). Multimodal learning with transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *45*(10), 12113–12132. https://doi.org/10.1109/TPAMI.2023.3275156

Yang, M., Wu, C., Guo, Y., Jiang, R., Zhou, F., Zhang, J., & Yang, Z. (2023). Transformer-based deep learning model and video dataset for unsafe action identification in construction projects. *Automation in Construction*, *146*, Article 104703. https://doi.org/10.1016/j.autcon.2022.104703

Yang, Z., Tang, C., Zhang, T., Zhang, Z., & Doan, D. T. (2024). Digital twins in construction: Architecture, applications, trends and challenges. *Buildings*, *14*(9), Article 2616. https://doi.org/10.3390/buildings14092616

Yao, Y., Tam, V. W. Y., Wang, J., Le, K. N., & Butera, A. (2024). Automated construction scheduling using deep reinforcement learning with valid action sampling. *Automation in Construction*, *166*, Article 105622. https://doi.org/10.1016/j.autcon.2024.105622

Yitmen, I., Kovacic, I., & Tagliabue, L. C. (2023). Cognitive digital twins for facilitating construction 4.0: Challenges and opportunities for implementation. *Frontiers in Built Environment*, *9*, Article 1130115. https://doi.org/10.3389/fbuil.2023.1130115

Yoo, B., Kim, J., Park, S., Ahn, C. R., & Oh, T. (2024). Harnessing generative pre-trained transformers for construction accident prediction with saliency visualization. *Applied Sciences*, *14*(2), Article 664. https://doi.org/10.3390/app14020664

Zulu, S. L., Saad, A. M., & Omotayo, T. (2023). The mediators of the relationship between digitalisation and construction productivity: A systematic literature review. *Buildings*, *13*(4), Article 4. https://doi.org/10.3390/buildings13040839