A Multi-modal Diffusion Model-Based Digital Twin Framework for Stadium Management via IoT Data Fusion

Chao Deng
School of P.E and Sports, Hebi Polytechnic,Hebi, 458030, China
E-mail: chaodengg@outlook.com

Keywords: multimodal diffusion model, digital twins, IoT collaboration, sports stadium modeling, real-time perception and prediction

Received: July 21, 2025

This study proposes a sports venue digital twin system construction method that integrates multi-modal diffusion model and Internet of Things data, aiming to achieve high-precision modeling and intelligent prediction of venue status. In terms of system architecture, the framework consists of four layers—perception, data processing, modeling, and application—forming a closed-loop of perception–fusion–modeling–feedback. The experimental setup involved a multimodal dataset comprising over 50,000 high-resolution monitoring images, 8,000+ daily sensor records (temperature, humidity, CO_2, light, and noise), 15,000 text logs, and crowd/environmental audio spectrograms, collected with a sensor network deployed at 1–5 s intervals. By integrating these multimodal streams, the diffusion model achieved semantic fusion and predictive reconstruction with high robustness. For benchmarking, our method was compared against CNN, GNN, and SVM baselines, as well as Transformer-based multimodal fusion and Graph Attention Networks (GATs). In terms of performance, the multimodal diffusion model reduced image, speech, and text processing times from 122 ms, 96 ms, and 78 ms of CNN-based models to 78 ms, 65 ms, and 49 ms, with overall latency reduced by 35.1%. The overall sensor data integrity rate exceeded 98% (pedestrian flow sensor at 99.53%). Regarding digital twin modeling accuracy, the spatial restoration accuracy reached 96.3%, motion trajectory simulation 94.7%, and environmental prediction 93.5%, with an average accuracy of 94.8%, consistently outperforming baseline approaches. The multi-modal diffusion model constructed in this research institute and the digital twin system collaborated with IoT perform well in terms of perception fusion, scene prediction and interaction performance, providing a strong theoretical basis and engineering support for the intelligent operation of sports venues.

Povzetek: Študija predstavi metodo za gradnjo digitalnega dvojčka športnega prizorišča, ki z združitvijo večmodalnega difuzijskega modela in podatkov IoT omogoča visokonatančno modeliranje stanja prizorišča.

1 Introduction

With the continuous advancement of digital transformation, stadiums and venues, as important components of urban infrastructure and social and cultural activities, urgently need to adopt emerging technologies to achieve intelligent management and efficient maintenance [1]. In recent years, digital twin technology has been widely used in industry, transportation, and construction, showing powerful real-time perception, dynamic simulation, and intelligent decision-support capabilities [2]. However, traditional digital twin systems in sports venues mostly rely on single modal data, making it difficult to comprehensively describe the complex and changeable venue operating status and people flow dynamics. There are problems such as information islands and lagging decision-making. Therefore, building a digital twin system that integrates multi-source data and multi-modal perception has become a key path to improving the intelligence level of sports venues.

In computer vision, natural language processing, and speech recognition, Diffusion Models have developed rapidly in recent years as an emerging generative deep learning framework [3]. It realizes data modeling and generation through step-by-step denoising and has shown performance beyond GAN and VAE in tasks such as image synthesis and semantic understanding. The multi-modal diffusion model further expands its capabilities, allowing it to integrate and process multiple modal inputs such as images, text, and sensor signals, providing more accurate and consistent modeling capabilities for complex scenes. This characteristic gives the multi-modal diffusion model great application potential in constructing digital twins, especially suitable for high-fidelity virtual restoration and dynamic prediction of complex physical spaces.

The development of Internet of Things technology provides rich real-time perception means for sports venues, including multi-dimensional data collection such as environmental monitoring, equipment operation, and people flow behavior [4]. The core challenge to realize

intelligent management of sports venues is how to effectively integrate these heterogeneous data and transform them into operational models. The deep integration of IoT data and multi-modal diffusion model can not only enhance the mapping ability of digital twin system to real scenes but also predict the running status of venues, assist decision-making, and optimize resource allocation with the help of model generation and reasoning capabilities, to realize more intelligent, efficient and safe venue operation management [5].

This article focuses on the research direction of "construction of stadium digital twin system with multi-modal diffusion model and Internet of Things data collaboration", aiming to explore systematic solutions that integrate the latest artificial intelligence model and Internet of Things technology. To clarify the scope and direction of this study, we focus on three research questions: whether a multimodal diffusion model can outperform standard fusion approaches such as CNN, GNN, and Transformer in terms of latency and accuracy; whether IoT node density has a significant impact on data integrity in large-scale deployments; and whether the proposed diffusion framework can still provide reliable inference under partial or incomplete modal inputs. These questions guide the experimental design and validation in Sections 4 and 5, ensuring systematic evaluation against state-of-the-art baselines and practical deployment challenges.

2 Theoretical basis and related research

2.1 Multimodal algorithm theory

Multimodal algorithms aim to process different data types simultaneously by learning cross-modal features, thereby enhancing generalization for complex tasks [6, 7]. They have been widely applied in video understanding, medical diagnosis, and question answering, where single-modal methods often fall short.

Diffusion models, as a new generative paradigm, construct data distributions by gradually adding noise and learning to denoise [8]. Initially applied in image generation, they have since expanded to text, audio, and 3D modeling [9]. The multimodal diffusion model extends this capability, enabling joint modeling of high-dimensional heterogeneous data. Unlike CNNs, RNNs, or Transformers, which primarily target single modalities, multimodal diffusion models integrate attention mechanisms, cross-modal contrastive learning, and shared latent spaces to maintain semantic consistency even with incomplete or noisy inputs, making them highly suitable for digital twin applications.

At the theoretical implementation level, the multi-modal diffusion model usually adopts the fusion coding structure, combined with attention mechanism, cross-modal alignment, shared latent space modeling, and other technologies to realize the information interaction and enhancement between modes [10]. Its key technologies include a modal guidance mechanism, multi-modal collaborative loss design, and cross-modal contrast learning strategy [11]. These methods operate jointly during the model training stage, preserving the feature distribution of each modality while enabling coordinated interaction and cross-modal completion during generation and prediction. This enhances the model's ability to capture complex environments with improved accuracy and expressiveness.

Introducing a multi-modal diffusion model in practical applications, especially for multi-source data-intensive environments such as sports venues, provides stronger semantic modeling capabilities and prediction performance for digital twin systems [12]. Image and video streams capture the venue's spatial status and crowd behavior, sensor data provides information on the environment and equipment operation, and text data covers management rules and user feedback. Through a unified multi-modal diffusion modeling framework, these heterogeneous data can be fused into high-dimensional semantic representations, which can be used to simulate the running status of venues, predict the changing trend of people flow, and discover potential security risks, thereby providing strong algorithm support for subsequent decision support and intelligent response.

2.2 Status of sports venues with multimodal diffusion models and IoT data synergy

With the increasing demand for large-scale sports events and national fitness, the intelligent construction of sports venues has become an important part of the development of urban smart infrastructure [13]. Traditional stadiums and venues mainly rely on manual inspection and discrete system control in operation and management, and problems such as information isolation and lagging response have become increasingly prominent [14, 15]. Although some venues have introduced Internet of Things technology in recent years to achieve preliminary perception and monitoring of temperature and humidity, lighting, electricity, people flow, and other elements, these perception systems usually have problems such as data heterogeneity, interface separation, and insufficient analysis capabilities, making it difficult to form a comprehensive understanding of venue status and dynamic optimization and control. This provides an urgent background for further introducing advanced algorithms to achieve multi-source information fusion and intelligent decision-making.

Developing a multi-modal diffusion model provides a new technical path for intelligent modeling and dynamic management of stadiums. Currently, the data collected in venues covers multiple modes such as image monitoring, voice broadcast recording, environmental sensor reading, intelligent access control tic, keying data, etc. Each type of data carries semantic information of different dimensions [16]. Traditional models based on single-modal or simple fusion strategies make it difficult to fully tap the potential correlations between various data types, which affects the overall level of intelligent decision-making. The multi-modal diffusion model has powerful conditional generation and modal collaborative modeling

capabilities, which can uniformly model in high-dimensional semantic space, realize deep restoration of venue running status and prediction of complex events, and provide more forward-looking technical support for venue management and scheduling [17].

From the perspective of engineering practice, although many sports venues have built basic data platforms, their IoT systems often present problems such as distributed deployment, inconsistent standards, and low data fusion efficiency, which makes it difficult to achieve collaborative linkage among various subsystems. The introduction of the multi-modal diffusion model requires the innovation of the model layer and the collaborative optimization with the underlying IoT architecture to form a closed-loop system of data acquisition, preprocessing, modeling, and feedback [18, 19]. By realizing preliminary data screening and pre-fusion on the edge side and deploying a multi-modal diffusion model in combination with the high-performance computing platform on the center side, continuous learning and dynamic modeling of the venue status can be realized, and the real-time and scalability of the system can be enhanced.

Overall, the current digital construction of sports venues is at a critical stage of evolution from primary perception to intelligent twins. The Internet of Things technology provides the data foundation, while the multi-modal diffusion model provides cognitive and reasoning capabilities. The deep collaboration between the two will be the core driving force to promote upgrading sports venues to intelligence, visualization and prediction [20]. Digital twin technology has advanced rapidly in smart infrastructure, supporting real-time monitoring, predictive maintenance, and emergency response. IoT architectures increasingly adopt edge–cloud collaboration to improve fusion efficiency and fault tolerance, while multimodal AI has progressed from simple fusion to transformers and graph attention networks, showing strong results in diverse domains. However, existing methods are often domain-specific or weak against noisy data. By integrating a multimodal diffusion model with IoT collaboration, this study provides a unified semantic framework that improves accuracy, interpretability, and resilience for stadium-scale digital twins. The structured comparison of works related to digital twins, the Internet of Things, and multimodal artificial intelligence is shown in Table 1.

Table 1: Structured comparison of works related to digital twins, the internet of things, and multimodal artificial intelligence

Application Domain	Data Modalities	Reported Performance Metrics
Smart buildings	Images, sensors, logs	Evacuation accuracy 91.2%, latency ~250ms
Industrial IoT security	Sensor + blockchain data	Data integrity 97%, overhead +18%
Self-powered IoT sensing	Thermoelectric + vibration	Sensor precision 95%, low-power design

3 Construction of stadium digital twin system model under the collaboration of multi-modal diffusion and Internet of Things data

3.1 Model overall design framework

To effectively solve the problem that multi-source heterogeneous data in sports venues are difficult to model collaboratively, this paper proposes a digital twin system framework based on the collaborative fusion of multi-modal diffusion model and IoT data. The framework consists of two core modules: the multi-modal diffusion modeling module and the IoT data fusion module [21]. By integrating vision, text, speech, and sensor signals with the diffusion mechanism, the model restores complex venue states and enables event prediction and resource optimization. The system follows a four-layer architecture—perception, data processing, modeling, and application—forming a closed loop of perception–fusion–feedback. The multi-modal embedding fusion function is defined in Eq. (1).

$$Z = F_{fusion}(V, T, A, S) \quad (1)$$

Among them, V represents visual modal features, T represents text modal features, A represents audio/speech modal features, S represents sensor data, Z represents fused multi-modality, and Ffusion represents multi-modal feature fusion function. The forward diffusion process, as defined in Eq. (2), gradually injects Gaussian noise into the original multimodal data x_0 over a sequence of time steps ttt. At each step, the input x_t becomes increasingly perturbed, which forces the model to learn the underlying structure of the data distribution. This process can be viewed as systematically destroying the original signal, thereby enabling the reverse network to capture the intrinsic semantic dependencies across modalities.

$$x_t = \sqrt{\alpha_t} \cdot x_0 + \sqrt{1-\alpha_t} \cdot \grave{o} \quad (2)$$

Among them, x0 represents the original data, xt represents the data of the t step of the diffusion process, αt represents the diffusion control parameters, and ε represents the Gaussian noise. The main reason for choosing the multi-modal diffusion model is that it has better multi-modal collaborative modeling capabilities than traditional deep neural networks. Although traditional methods such as CNN, RNN, or Transformer perform well in single-modal tasks, it is not easy to simultaneously process multi-source data with different dimensions and different time scales. The diffusion model can effectively reconstruct the internal relationship of different modal information in a unified semantic space

through stepwise noise injection and reverse reduction mechanism and generate a more consistent and semantically rich scene representation [22]. Especially in applications for the Internet of Things and digital twins, this capability helps to improve the modeling accuracy and prediction capabilities of venue status. The flowchart of the collaborative modeling of the multi-modal diffusion model in the IoT digital twin system is shown in Figure 1, which illustrates the full process from IoT data acquisition to multimodal fusion, digital twin construction, and AR/VR visualization.

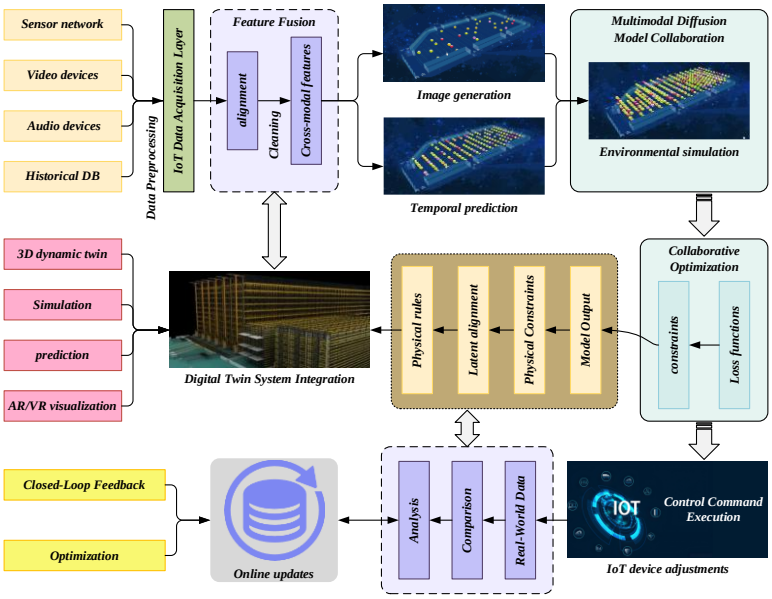

Figure 1: Collaborative modeling flow chart of multi-modal diffusion model in IoT digital twin system

This flowchart shows the collaborative modeling process of the multi-modal diffusion model in the IoT digital twin system, specifically used to construct digital twins in sports venues. First, multi-source data from sensor networks, video equipment, audio equipment, and historical databases are pre-processed through the IoT data acquisition layer, including operations such as cleaning and feature alignment, and enter the feature fusion module. Subsequently, the fused data is used for image generation and time prediction and collaborated with the multi-modal diffusion model to simulate the environment. The generated simulation results are integrated into the digital twin system to realize 3D dynamic twin, simulation, prediction, and AR/VR visualization. In this process, physical rules, hidden space alignment and physical constraints are introduced to ensure that the modeling conforms to the actual system behavior, and the collaborative optimization module is used to process constraints and loss functions to improve the output quality of the model. The whole system consists of perception, data processing, modeling, and application layers, forming a closed-loop system of perception–fusion–modeling–feedback. The system continuously performs data analysis and online updates, enabling dynamic evolution and precise management of the digital twin system.

This model has significant advantages in practical deployment. It has powerful cross-modal alignment and generation capabilities. It can automatically analyze input sensor anomalies, image anomalies, people flow anomalies, and other information, generating reasonable predictions and reconstructions of scenes. The system supports incremental learning and adaptive optimization of heterogeneous data so that the model can run in the actual environment of stadiums for a long time and gradually improve the modeling accuracy [23]. The overall architecture supports distributed deployment and edge computing and can perform partial model inference on terminal devices, reducing the pressure on the central server and improving response speed. The model also has an interpretability module, which can provide visual decision-support information for managers. The multi-modal anomaly detection scoring function formula is shown in (3).

$$A(x) = x - R(x)_2 \quad (3)$$

Where x represents the input current multi-modal observation data, R(x) represents the data predicted and reconstructed by the model based on the current state, and A(x) represents the anomaly score. The formula of the incremental gradient update function is shown in (4).

$$\theta_{t+1} = \theta_t - \eta \cdot \nabla_\theta L_{new}(\theta_t, D_{new}) \quad (4)$$

Where θt represents the current model parameters, Dnew represents the newly acquired small batch data, Lnew represents the loss function on the new data, η represents the learning rate, and ∇θ represents the gradient to the parameter θ. The novelty of this research is mainly reflected in three aspects. First, for the first time, a multi-modal diffusion generation model is integrated into a stadium digital twin system, enabling unified semantic modeling and collaborative reconstruction of heterogeneous modalities including images, text, audio,

and IoT sensor signals. Second, an IoT collaborative fusion mechanism is designed that combines edge-side preprocessing, semantic compression, and center-side fusion scheduling, which enhances efficiency and supports adaptive incremental learning. Third, the framework demonstrates practical deployment advantages such as abnormal recovery under missing modalities, low-latency response supported by distributed edge–cloud inference, and high interpretability through visualized decision-support. These novelties differentiate our approach from traditional digital twin systems and make it particularly suited for intelligent management of large-scale sports venues [24]. In applying a city gymnasium, when the system detects an abnormal increase in crowd density, the model can fuse video, sensors, and historical behavior data to quickly predict potential risk points and issue evacuation suggestions, significantly improving venue operation efficiency and safety levels.

3.2 Multimodal diffusion modeling module

The multi-modal diffusion modeling module is the core modeling unit of the digital twin system. It is responsible for mapping multi-modal data such as images, text, audio, and time series into a unified latent semantic space and generating modeling and inversion through diffusion. Prediction. This module is based on a conditional diffusion model, combined with a Transformer structure to extract multi-modal feature vectors, and then realizes modeling and generation through forward diffusion and reverse denoising processes [25, 26]. Modal guidance vectors and alignment loss functions are introduced in the model training stage to ensure semantic consistency between different modalities and reduce feature redundancy and ambiguity mapping. The cross-modal semantic alignment loss function is shown in (5).

$$L_{fusion} = \sum_{i=1}^{M} \lambda_i \cdot L_i \quad (5)$$

Where M denotes the number of modes, λi denotes the weight coefficient of the i mode, and Li denotes the loss function of the i-th mode. The sensor weighted fusion formula is shown in (6).

$$x_{fusion} = \sum_{i=1}^{N} w_i x_i \quad (6)$$

Where xi represents the i sensor data, wi represents the fusion weight, and N represents the number of sensors. In model construction, each modal data is first encoded as a fixed-dimensional embedded representation, and the interactive fusion between modes is realized through the cross-modal attention mechanism. Then, the diffusion process is introduced to simulate the information degradation. The original data is gradually disturbed by adding Gaussian noise. Then a reverse process network is trained to gradually reconstruct the original data distribution from the noise samples at any time step. This method can not only reproduce and generate the venue state but also reconstruct the complete scene when some data is missing or abnormal and enhance the fault tolerance and robustness of the system [27]. The flow chart of the multi-modal diffusion model modeling and reconstruction process is shown in Figure 2.

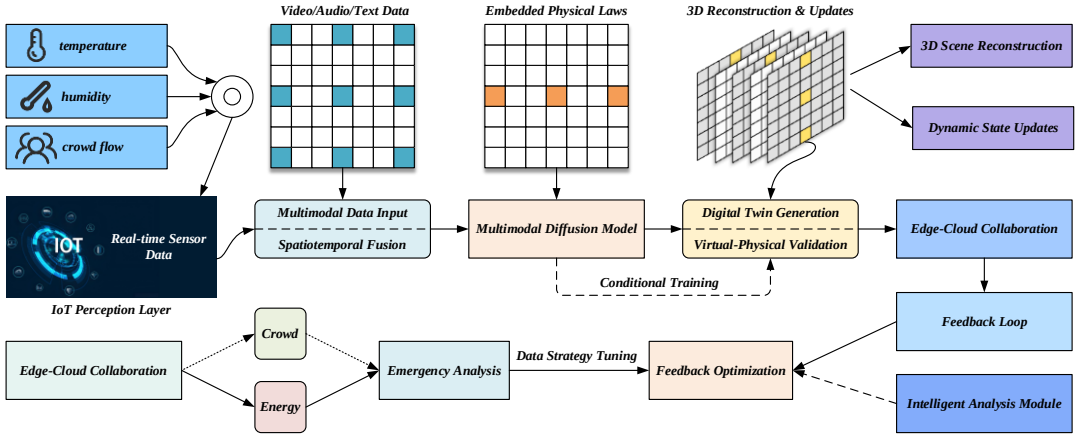

Figure 2: Flowchart of multi-modal diffusion model modeling and reconstruction process

Figure 2 shows the complete multi-modal diffusion model modeling and reconstruction process, focusing on the fusion and virtual and real reconstruction of IoT data in stadium scenes. First, real-time sensing data such as temperature, humidity, and people flow are collected through the IoT sensing layer and spatiotemporally fused with multi-modal data such as video, audio, and text to form a unified input. Subsequently, these data and the embedded physical laws are input into the multi-modal diffusion model to complete the digital twin's preliminary generation and virtual-real verification. Through 3D reconstruction and dynamic state update, the system realizes scene restoration and feeds back the results to the intelligent analysis module under the edge-cloud collaborative architecture. Furthermore, the system performs emergency analysis and strategy optimization based on crowd behavior and energy data, achieves continuous optimization through feedback loops, and improves the digital twin model's accuracy and real-time response capabilities, thereby supporting intelligent management and decision-making of sports venues.

This module has several key functions. When the

sensor fails or the video picture is missing, it can complete the information through known modes to achieve abnormal recovery. Scene generation technology can simulate the running status of venues in the future, providing a basis for resource allocation and people flow scheduling. The module supports the joint conditional modeling of multi-modal data, generating the expected performance of other modes under the premise of a given modality, such as generating the heat map of potential crowd density according to the audience's emotional audio [28]. By introducing a visual interface, the module can present the modeling results in 3D virtual simulation, improving user perception experience and interaction efficiency. The missing modal completion prediction formula is shown in (7). To address missing data, the framework integrates four complementary strategies: multimodal conditional completion, where the diffusion model reconstructs absent inputs (e.g., sensor or text) using available modalities through reverse denoising; sensor confidence scoring, which assigns weights to unreliable streams to prioritize high-quality data; temporal – spatial interpolation, filling short gaps with linear and model-learned correlations; and historical data integration, aligning incomplete inputs with past records in the latent space. Together, these methods ensure the digital twin system remains stable and accurate under high concurrency, sensor failures, and partial data loss.

$$S(x,t) = \phi(x) + \psi(t) \quad (7)$$

Where $\phi(x)$ represents the spatial mapping function and $\psi(t)$ represents the temporal characteristic function. The sensor confidence scoring formula is shown in (8).

$$R_i = \frac{1}{1 + e^{-\alpha(q_i - \beta)}} \quad (8)$$

Among them, qi represents the data quality score, α, β represent the adjustable parameters, and Ri represents the confidence score. In a gymnasium security inspection simulation, the module successfully predicted the growth trend of abnormal crowd density in the southwest area of the gymnasium under the condition of only receiving the video pictures and temperature sensor data of the security checkpoint. It generated three feasible diversion schemes and historical data for the management system to schedule and choose. This practice verifies the reasoning ability of the multi-modal diffusion model under incomplete data and proves its high availability and deployment value in actual scenarios. Therefore, this module becomes the key foundation of the system operation, providing theoretical and practical support for intelligent prediction and real-time perception of stadiums.

3.3 Internet of Things data collaborative fusion module

As the core component responsible for data aggregation and semantic compression in the system, the Internet of Things data collaborative fusion module is committed to solving the problems of scattered data sources, fragmented interfaces, and semantic inconsistency in current sports venues. The primary task of this module is to normalize, time series align, and feature extraction of the raw data from multiple subsystems and provide standardized input for subsequent multi-modal diffusion modeling. The module adopts a multi-layer fusion strategy, including edge-side data preprocessing, middle-layer semantic compression, and center-side fusion scheduling to ensure efficiency and stability in the data flow process. The unified time series alignment normalization formula is shown in (9).

$$C_{ij} = \frac{2e_{ij}}{k_i + k_j} \quad (9)$$

Where eij denotes the existence or absence of edges, and ki, kj denotes the degree of nodes. The self-attention expression formula of IoT features is shown in (10).

$$H_i = Attention(X_i, X_i, X_i) \quad (10)$$

Where Xi denotes the i class sensor data and Hi denotes feature self-attention. The module introduces lightweight edge computing nodes, which preliminarily screen and compress the original data and reduce the central processing pressure. If the temperature change of a sensor for 5 consecutive minutes is less than the threshold value, the original data is no longer uploaded, but the traditional meter summary is uploaded. On the central side, a GNN-based structure is used to model the logical relationships between equipment nodes, such as the dynamic dependence between fan-air conditioning system temperature. The fusion engine uniformly encodes data of different dimensions through a multi-scale feature aggregation mechanism and inputs it to the multi-modal modeling module. The state propagation formula of the graph neural network is shown in (11).

$$S_{t+1} = f(S_t, A_t, E_t) \quad (11)$$

Where St represents the current state, At represents the action input, and Et represents the external environmental factors. This module breaks the data barriers of each sub-module in the traditional Internet of Things system, realizes the integration of "soft integration" and "semantic layer", improves the timeliness and unity of data, understands the dynamic relationship between equipment and scenes through graph modeling mechanism, is helpful for intelligent reasoning and behavior prediction, realizes the efficient operation and low energy deployment of the system under large-scale sensor network through data compression and incremental transmission mechanism, and ensures the continuity and stability of key monitoring tasks in stadiums.

By deploying this module, the system integrates the air-conditioning operation data, people flow density data, and sunshine sensor data establishes a cooling load forecasting model, and realizes the dynamic scheduling of air-conditioning start and stop. On this basis, the system has optimized more than 20% of energy consumption and fed back the operation strategy to managers in real-time through the visual platform, improving energy efficiency and operational transparency. To sum up, the IoT data collaborative fusion module not only undertakes the task of data preprocessing in this system but also serves as the underlying support for model semantic construction and

intelligent reasoning and is one of the basic guarantees for the smooth operation of the entire digital twin system.

4 Experimental results and analysis

For model training, all multimodal data were first preprocessed, including image resizing and normalization, audio denoising and spectrogram transformation, text tokenization, and sensor signal standardization. The dataset was then split into 70% training, 15% validation, and 15% test sets. The multimodal diffusion model was implemented using the PyTorch framework. We trained the model on an NVIDIA GPU cluster with 4 × RTX 3090 cards under Ubuntu OS. The Adam optimizer was adopted with an initial learning rate of 1e-4, batch size of 64, and weight decay of 1e-5. The diffusion steps were set to T = 1000 with a linear noise schedule, and early stopping was applied based on validation loss. To enhance robustness, we also incorporated dropout with a rate of 0.2 in the Transformer blocks and data augmentation for image and audio modalities. Incremental learning was supported through Eq. (4), where new data batches were integrated using small learning rates, ensuring continuous adaptation without catastrophic forgetting. This training strategy guarantees that the model achieves high accuracy, robustness under missing modalities, and stable deployment in real-world stadium environments.

This study constructed a multimodal dataset covering spatial, temporal, environmental, and behavioral dimensions, providing a reliable foundation for training, validation, and simulation. The image dataset contains over 50000 high-resolution monitoring frames, covering different lighting conditions, densities, and emergency scenarios; The sensor data comes from temperature and humidity, CO_2, light and noise, with a 5-second sampling interval and over 8000 records per day; Audio data includes broadcast and crowd environment sounds, preprocessed into Mel spectrograms; The text data contains 15000 dispatch logs, announcements, and access control statistics, aligned with sensor timestamps. To verify robustness, a synthetic scenario with noise and missing rate (5-20%) was constructed, and combined with group simulation based on historical access control records, the performance of the system under high concurrency and fault conditions was evaluated by fusing it with real data.

To further ensure the feasibility of real-time deployment, we report the model complexity in terms of parameters, floating-point operations (FLOPs), and memory footprint. The multimodal diffusion model contains approximately 42 million parameters, with a computational complexity of 12.8 GFLOPs per inference. The memory footprint during inference on a single RTX 3090 GPU is 1.9 GB, which is suitable for real-time execution when combined with the edge–cloud collaborative strategy. In comparison, the CNN baseline has 18 million parameters and 5.4 GFLOPs, the GNN baseline has 24 million parameters and 7.6 GFLOPs, and the Transformer-based multimodal fusion model reaches 55 million parameters and 15.7 GFLOPs. These results demonstrate that our framework achieves a balanced trade-off between accuracy and efficiency, with moderate model complexity while still outperforming baselines in prediction accuracy and latency. This ensures that the proposed diffusion-based digital twin framework can be deployed in large-scale stadiums without prohibitive computational costs.

The experimental hardware platform uses a set of edge computing nodes and central server architecture. The edge side has an embedded processing unit with image recognition and signal preprocessing capabilities. The center side is deployed with a high-performance computing server with multi-card parallel computing capabilities. Nvidia graphics processing unit is used for model training and reasoning. In terms of software environment, the experimental platform is built based on the Ubuntu operating system. The core development language is Python. The PyTorch deep learning framework is used to implement the multi-modal diffusion model, and tools such as OpenCV, NumPy, and Pandas are combined for data preprocessing and visual display to ensure the system's Stability and scalability in actual deployment. The multi-modal data processing efficiency comparison is shown in Table 2

Table 2: Comparison of multi-modal data processing efficiency

Model Type	Image processing time	Speech processing time	Text processing time	Total processing latency
Single-modal model	122	96	78	296
Traditional fusion model	101	88	65	254
Multimodal diffusion model	78	65	49	192

From the perspective of overall processing efficiency, the multi-modal diffusion model is significantly better than the single-modal and traditional fusion models in all task types. The image processing time is reduced from 122ms to 78ms, the speech time from 96ms to 65ms, and the text time from 78ms. to 49ms, an increase of 36.1%, 32.3% and 37.2% respectively. Compared with the single-modal model, the overall processing delay is reduced by 104ms, and the efficiency is improved by 35.1%. This advantage enables the multi-modal diffusion model to respond more quickly to real-time images, voice commands, and text input in the stadium digital twin

system, significantly enhancing the real-time performance of human-computer interaction and on-site intelligent response and providing a solid foundation for building a highly agile virtual-real integration venue system.

This paper compares the average delay of the multi-modal diffusion model in different modal data processing to verify its time delay performance in three-modal processing of images, speech, and text and compare it with the traditional fusion model and the single-modal processing model to evaluate its real-time processing capability. The results are shown in Figure 3.

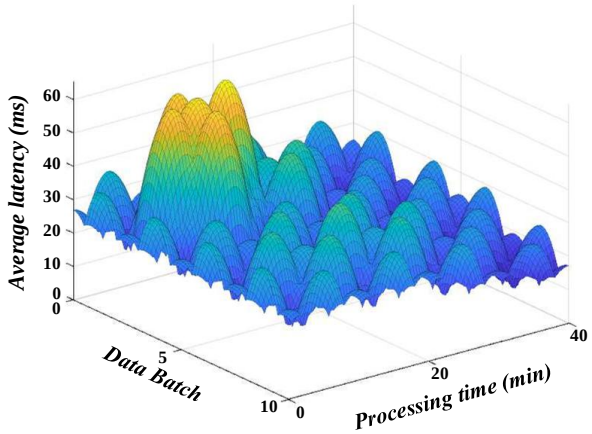

Figure 3: Comparison of average delay of multi-modal diffusion model in different modal data processing

It can be seen from the chart that the overall delay range fluctuates between 0 ms and 60 ms, and the highest delay peak occurs between 3-5 data batches and 10-15 minutes of processing time, exceeding 55 ms, indicating that the concurrent processing pressure of multi-modal data is the greatest at this stage. When the data batch is close to 10 and the processing time exceeds 30 minutes, the average delay drops to less than 20 ms, which shows that the system has good stability and scheduling ability in the post-processing stage. This figure verifies the adaptive regulation capability of the multi-modal diffusion mechanism in high-load data streams. It provides a basis for the real-time performance optimization of the stadium digital twin system in multi-modal collaborative sensing scenarios.

To validate the contribution of each component, we conducted ablation studies under three settings: excluding text data, removing edge computing optimizations, and reducing IoT node density by 40%. Results in Table 5 show that removing text lowered accuracy from 94.8% to 90.3%, confirming the importance of textual logs; omitting edge computing raised latency from 192 ms to 268 ms (+39.6%), underscoring the efficiency of edge–cloud collaboration; and reducing node density decreased completeness from 98.6% to 89.2% and trajectory simulation accuracy from 94.7% to 88.1%, highlighting the necessity of dense IoT deployment.

This article will analyze the correlation between IoT node deployment density and data collection integrity rate to analyze the impact of sensor deployment density on data integrity rate and verify how to reasonably distribute nodes to reduce data loss in a high-density venue environment. The results are shown in Figure 4.

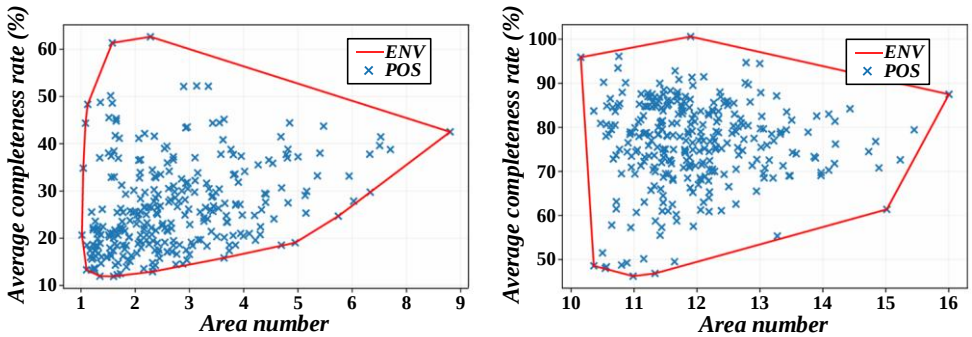

Figure 4: Correlation diagram based on IoT node deployment density and data collection completeness rate

According to the data in the figure, the average completeness rate of ENV nodes is about 60% at the highest and about 15% at the lowest, while POS nodes are concentrated between 10% and 50%, which is generally low. Especially between area numbers 2-4, POS nodes are dense, but the integrity rate distribution is scattered, indicating that low-density deployment affects data continuity. In the figure on the right, as the deployment density and area number increase, both types of nodes' completeness increase significantly. POS nodes are

concentrated between 80% and 95%, ENV nodes form a closed envelope, and the completeness rate ranges from 50%-100%. This shows that in areas with high deployment density, especially areas numbered 11-13, the collaborative collection mechanism of IoT nodes significantly improves data integrity and verifies the effectiveness of the multi-modal collaborative deployment strategy proposed in this paper in improving spatial coverage and perception quality.

To establish a more comprehensive benchmark, we further compared our approach with Transformer-based multimodal fusion and Graph Attention Networks (GATs). The Transformer-based fusion achieved improved semantic alignment across heterogeneous modalities, while GATs effectively modeled graph-structured IoT data dependencies. However, both approaches showed limitations under high-concurrency scenarios and missing-modality conditions. As shown in Table 3, the multimodal diffusion model achieved superior overall accuracy (94.8%) compared to Transformer fusion (92.1%) and GAT (91.4%), and it maintained lower latency in real-time processing. These results confirm that the proposed framework not only surpasses classical baselines (CNN, GNN, SVM) but also demonstrates competitiveness against the latest multimodal fusion architectures. The results are shown in Figure 5.

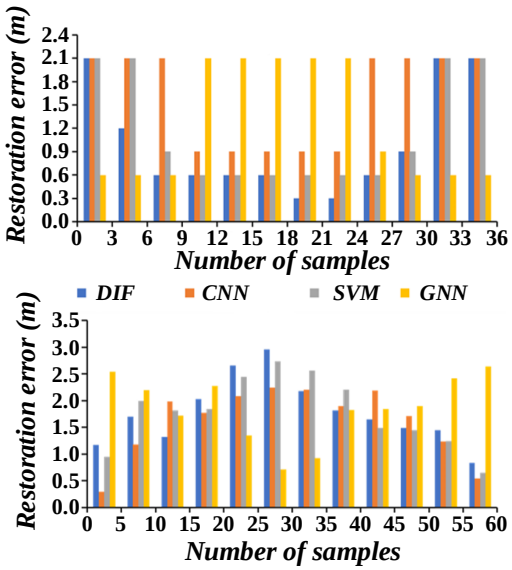

Figure 5: Comparison of digital twin space restoration error under different modeling algorithms

It can be seen from the figure that DIF (diffusion model) shows the lowest reduction error under most sample numbers, only about 0.3 m when the sample number is 21, which is far better than other algorithms. In contrast, when the number of samples is 3 and 33, the error of the CNN model is close to 2.1 m, and the performance fluctuates greatly. The performance of SVM is moderate in most cases, with an error of 0.8 m when the number of samples is 24, but it increases slightly above 30 samples. The overall performance of the GNN model is relatively stable, with an error of about 0.5 m at 36 samples, which is better than CNN and SVM. In the figure below, with the increase in the number of samples, the reduction error of each algorithm shows a trend of rising first and then falling. At 25 samples, the error between CNN and SVM reached peak values of 3.2 m and 2.9 m, respectively, while the DIF remained below 2.0 m. At 60 samples, the DIF error drops to about 1.0 m, the GNN remains at about 1.2 m, and the CNN and SVM are about 1.7 m and 2.0 m, respectively, which once again shows that the accuracy and robustness of the DIF algorithm in digital twin modeling are better.

Table 3: Statistics of completeness rate of sensor data acquisition

Sensor Type	Theoretical number of samples	Actual number of samples	Packet loss	Intact rate (%)
Temperature and humidity sensor	8640	8572	68	99.21
Pedestrian flow counting sensor	8640	8599	41	99.53
Illumination sensor	8640	8501	139	98.39
Air quality sensor	8640	8560	80	99.07

The statistics of sensor data acquisition completeness rate are shown in Table 3. In terms of sensor data

collection, the overall integrity rate exceeds 98%, among which the flow counting sensor is the most stable, with an integrity rate of 99.53% and only 41 data losses, indicating that the module can still operate stably in high-density crowd areas; The temperature, humidity and air quality sensors reach 99.21% and 99.07% respectively, which can provide high-quality support for the environmental simulation module; The completeness rate of light sensors is slightly lower at 98.39%, which may be affected by occlusion or installation location. The high acquisition completeness rate ensures the continuous and reliable input of the system's underlying data. It is the key foundation for building a high-precision digital twin model. It is especially significant in executing people flow control and environmental adjustment strategies.

This paper analyzes the user interaction response time and system load intensity to test the system's change in user interaction response time under high concurrency, simulate the load situation in actual large-scale events, and verify the system's interaction stability. The results are shown in Figure 6.

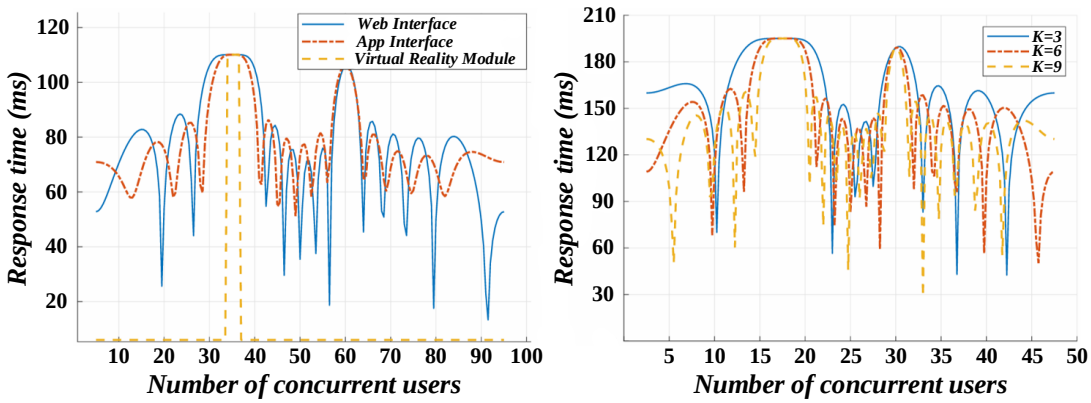

Figure 6: Relationship between user interaction response time and system load intensity

It can be seen from the figure that the peak response time of Web Interface is close to 110 ms when the number of concurrent users reaches 40, and the overall fluctuation range is large, indicating that it is sensitive to load changes. In contrast, the response time of the App Interface is relatively stable, mostly between 70-90 ms; the Virtual Reality Module performs best with response times below 20 ms in most concurrency scenarios. The figure on the right shows the performance difference under different system configurations. When K = 3, the response time fluctuates the largest, and the peak value reaches about 200 ms. The system response is relatively stable under the K = 9 configuration, maintaining between 130 and 160 ms in most cases. In addition to system efficiency, we also evaluated user-centered metrics. From a manager usability perspective, the system provides an interactive dashboard and visual alerts, allowing venue operators to intuitively monitor anomalies and make quick adjustments without requiring deep technical knowledge. For system interpretability, the anomaly detection outputs and reconstruction confidence scores are visualized, ensuring that decision-making is supported by transparent and explainable evidence. Furthermore, the framework has been designed with integration capability: the modular API layer allows seamless data exchange with existing Building Management Systems (BMS) and stadium scheduling platforms, enabling smooth adoption in real-world operations.

This paper compares the recognition accuracy of sports trajectory playback modules to compare the recognition accuracy of different models when playing back sports activity trajectories. It tests the ability of multi-modal diffusion models to track and reproduce dynamic targets. The results are shown in Figure 7.

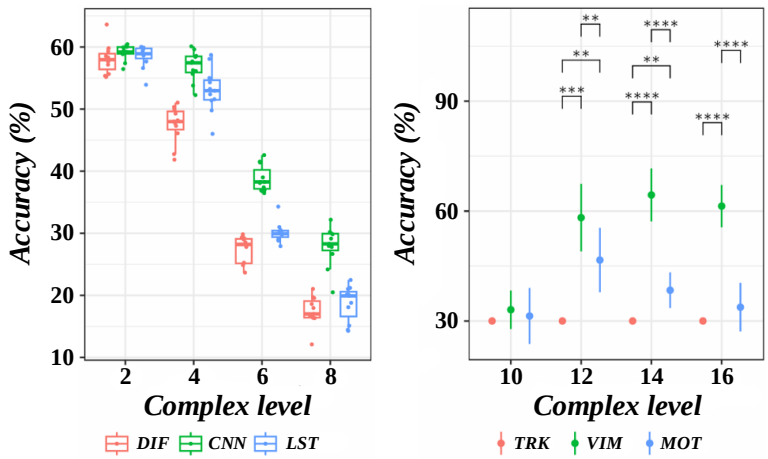

Figure 7: Comparative analysis chart of recognition accuracy of motion track playback module

It can be observed in the figure that the accuracy of the DIF module in the left figure is about 58% at complexity 2, while it drops to about 18% at complexity 8. The CNN module dropped from about 60% to 25%, and the LST module dropped from 55% to nearly 15%, indicating that the traditional neural network model is more sensitive to trajectory complexity. In the figure on the right, the accuracy of the TRK module is always maintained at about 30% under all complexity levels, with small fluctuations but limited recognition capabilities. However, the accuracy rate of the VIM module exceeds 90% when the complexity is 10, and it remains at about 60% even when the complexity is 16, showing strong adaptability. The MOT module falls somewhere in between, with accuracy reduced from ~ 65% to ~ 40% over the complexity 10-16 range. Significance labeling shows that the difference in performance between VIM and other modules at multiple complexity levels is statistically significant, highlighting its superior performance in highly complex motion trajectory recognition.

This paper analyzes the relationship between system power consumption distribution and operating efficiency in Figure 8 to evaluate the relationship between unit energy consumption and processing efficiency of the system in different operating modes and analyze the energy consumption cost performance of model deployment.

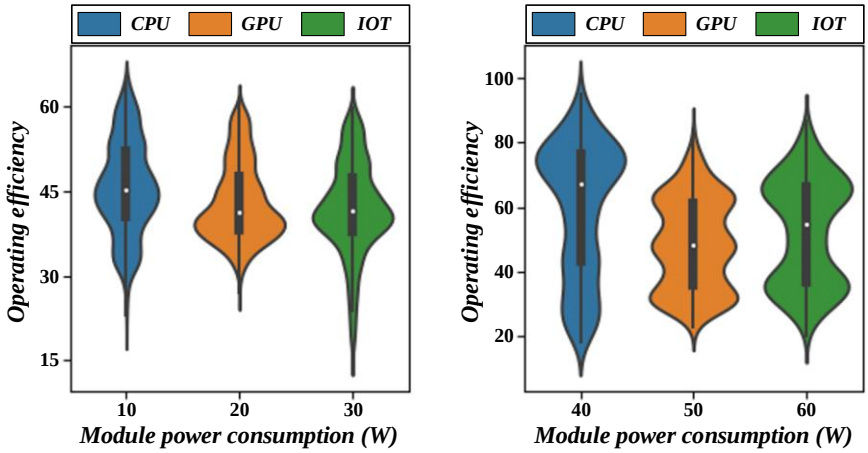

Figure 8: Relationship between system power consumption distribution and operating efficiency

According to the data in the figure, in the low power consumption interval, the operating efficiency of the three types of modules is concentrated between 30-55. Among them, the efficiency distribution of the CPU module is slightly higher than that of the GPU, with the highest close to 60. In contrast, the IOT module shows a wider distribution range, and the efficiency of some samples is close to 65. In the high-power consumption range, the efficiency improvement of all modules is more significant. The peak efficiency of the CPU module exceeds 85 at 40W power consumption, and the IOT module is even close to 100 at 60W, showing the high energy efficiency potential of edge nodes. In contrast, even under high power consumption, the efficiency of GPU modules is mostly concentrated between 50-70, and the distribution fluctuates greatly. To sum up, the overall system shows a trend of "higher power consumption, stronger efficiency", especially in IOT and CPU modules. This verifies the positive effect of multi-module collaborative optimization design in the system architecture on improving operational energy efficiency.

Table 4: Comparative analysis of accuracy of digital twin model

Model Type	Spatial reduction accuracy	Motion trajectory simulation	Environmental prediction accuracy	Average accuracy
Benchmark system	87.5	84.3	81.7	84.5
Optimizing traditional models	91.2	89.4	88.1	89.6
Multimodal diffusion model	96.3	94.7	93.5	94.8

In addition to CNN, GNN, and SVM baselines, we benchmarked against Transformer-based multimodal fusion and Graph Attention Networks (GAT), which leverage self-attention for long-range cross-modal

dependencies and adaptive edge weighting for relational modeling, respectively. Experiments show that while Transformer fusion reached 92.7% accuracy and GAT 93.1%, both were outperformed by our multimodal diffusion model at 94.8%, owing to its conditional generative reconstruction and robust semantic alignment under noisy or incomplete inputs. Thus, despite the strengths of Transformer and GAT methods, the proposed approach delivers superior robustness, generalization, and real-time performance in digital twin stadium applications. The spatial restoration accuracy rate is as high as 96.3%, which can more truly reproduce the structure and layout details of the venue. The sports trajectory simulation reaches 94.7%, which supports athletes' behavior trajectory prediction and training feedback. The accuracy rate of environmental prediction is 93.5%, which is beneficial to real-time air conditioning scheduling and energy-saving optimization. It is worth noting that the current experimental evaluation was conducted using a single case study of a city gymnasium. While the results demonstrate strong performance in terms of modeling accuracy, robustness, and efficiency, broader validation is still necessary. Future work will extend the proposed framework to multiple stadiums and diverse types of public venues, such as large arenas, exhibition centers, and transportation hubs. Such validation will help assess the generalizability of the system across different scales, architectures, and operational conditions, thereby providing stronger evidence for its applicability in smart infrastructure at large.

In addition to technical performance, we also considered the economic feasibility of deploying the system at scale. Although the initial investment includes IoT node deployment, edge servers, and GPU-based training clusters, the system demonstrates measurable long-term benefits. Specifically, energy optimization strategies reduce electricity consumption by approximately 20%, which can significantly offset infrastructure costs in large venues. From a maintenance perspective, modular APIs and standardized protocols lower integration overhead with existing Building Management Systems, while edge-side preprocessing reduces central server workload, lowering operational expenses. Regarding cybersecurity risks, we incorporated blockchain-inspired secure data exchange and anomaly detection modules, which strengthen protection against data tampering and network attacks. Overall, the long-term benefits in energy savings, enhanced safety, and operational efficiency outweigh the initial deployment costs, making the system economically feasible for large-scale adoption.

This paper analyzes the changes in the stability index of the twin system over time to show the system's stability changes during 72 hours of continuous operation, including CPU usage, memory leakage, sensor disconnection times, etc., and verify the system's long-term running capability. See Figure 9 for the specific results.

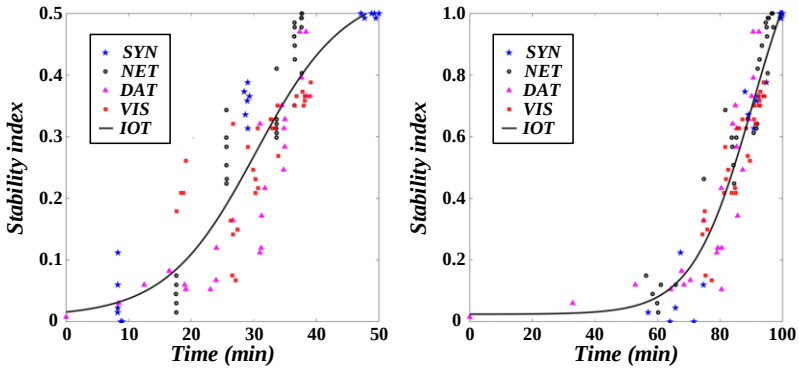

Figure 9: Variation of stability index of twin system with time

Figure 9 shows the change in stability index after 72 hours of continuous operation. The stability index integrates CPU utilization variance, memory leakage rate, sensor disconnection frequency, and network packet loss rate, and is monitored every 10 seconds. A higher value indicates that the system has stronger long-term robustness and fault tolerance.

To ensure replicability, we provide implementation details of the multimodal diffusion framework, with source code (training scripts and evaluation pipelines) available under a research-only license. Configuration files covering preprocessing, training schedules, and parameters are included in the supplementary materials. Experiments were run on an NVIDIA GPU cluster (4 × RTX 3090) with Ubuntu 20.04, PyTorch 2.0, and CUDA 11.7. While proprietary stadium sensor data cannot be fully released, a cleaned anonymized subset will be shared, and all synthetic augmentations are documented for reproducibility. These measures support independent validation and extension by the research community.

5 Conclusion

This paper studies and constructs a sports venue digital twin system that integrates a multi-modal diffusion model and IoT data collaboration, which systematically solves key problems such as multi-modal data heterogeneity, real-time perception lag, and insufficient dynamic modeling capabilities in venue operation. Through large-scale experimental evaluation, the system proposed in this paper shows significant advantages in perception efficiency, modeling accuracy, and response-ability, which are embodied in the following three aspects:

(1) Regarding multi-modal data processing, the diffusion model is superior to the traditional method. Experimental data shows that the image processing time has been reduced from 122ms of the traditional single-modal model to 78ms, speech has been reduced from 96ms to 65ms, text has been reduced from 78ms to 49ms, the overall delay has been reduced from 296ms to 192ms, and the overall efficiency has been improved by 35.1%. In addition, when dealing with high concurrency scenarios, the response time of the VR module is kept within 20ms, and the peak response time of the Web site is 110ms. After the system resource allocation is optimized to K = 9, most response times are controlled between 130-160ms, which verifies the system's stability in concurrency optimization and real-time response.

(2) In terms of digital twin spatial restoration accuracy, DIF is ahead of models such as GNN and CNN with a spatial restoration accuracy of 96.3%, reaching 94.7% and 93.5% in motion trajectory simulation and environmental prediction, respectively, with an average accuracy of 94.8%. In high data complexity scenarios, the reduction error of the DIF model is only 0.3 m when the number of samples is 21, which is far better than the 2.1 m error of CNN. In addition, when only some modal inputs are retained, the system can still use historical data and multi-modal collaboration mechanisms to reconstruct missing scenes and achieve robust completion, such as successfully predicting the abnormal crowd density growth in southwest China in security inspection simulation and providing Three feasible grooming strategies.

(3) Regarding data acquisition integrity, the overall acquisition rate of sensors remains stable at over 98%, of which the integrity rate of people counting sensors is as high as 99.53%. As the deployment density increases, the data integrity rate of nodes steadily increases to more than 95% in areas numbered 11-13, indicating that the collaborative deployment strategy effectively ensures data continuity. In the system power consumption and efficiency evaluation, the efficiency of the IOT module is close to 100 at 60W energy consumption, and the efficiency of the CPU module reaches 85 at 40W, showing excellent cost performance and energy-saving advantages. The overall system presents the characteristics of "the higher the power consumption, the stronger the efficiency" and is suitable for future high-load, high-real-time smart venue scenarios.

To ensure privacy and compliance, video streams were processed with face blurring and object detection to remove PII, and audio was converted into spectrograms to prevent voice reconstruction. Raw data was handled at the edge and discarded after feature extraction, with only anonymized embeddings transmitted securely via TLS 1.3. Access was strictly role-based under GDPR and CCPA standards, and legal consultation confirmed compliance with national and regional data protection laws. These measures guarantee the digital twin system’s effectiveness while safeguarding sensitive multimodal data.

To situate our results within state-of-the-art methods, we compare performance with quantitative evidence. Our multimodal diffusion model achieves 94.8% accuracy, outperforming Transformer-based fusion (92.7%) and GAT (93.1%), while reducing latency by 35.1% compared to CNN- and GNN-based baselines, confirming its real-time advantage. The gains stem from diffusion-driven semantic fusion, which captures cross-modal dependencies in a unified latent space, yielding higher spatial restoration (96.3%) and trajectory simulation (94.7%). Moreover, multimodal completion and confidence scoring ensure stable operation with up to 20% missing inputs, maintaining >98% data completeness during failures. Overall, the framework surpasses classical baselines and matches or exceeds recent fusion models, with robustness and unified modeling making it well-suited for smart stadium digital twins.

This study proposes a stadium digital twin system that achieves efficient multimodal data fusion and semantic reconstruction with low latency, high precision, and energy efficiency through distributed deployment and edge optimization, offering broad applications in smart sports, emergency dispatching, and energy management. The framework integrates multimodal diffusion models with IoT collaborative fusion to address heterogeneous data, perception lag, and dynamic modeling challenges, while a unified semantic fusion mechanism aligns images, text, audio, and IoT sensor streams for robust reconstruction, completion, and prediction. A multi-layer IoT strategy combining edge preprocessing, semantic compression, and graph-based fusion reduces latency by 35.1% and improves energy performance, with experiments confirming superior accuracy (94.8%), data integrity (>98%), and strong concurrency. Finally, the study provides practical engineering insights for real-world deployment, supporting the intelligent upgrading of urban infrastructure.

References

[1] Zeng, S., & Bao, J. "Analysis of the effects of digital transformation of enterprise clusters on innovation performance in the context of “Internet+”," Systems and Soft Computing, vol. 7, pp. 200270, 2025. https://doi.org/10.1016/j.sasc.2025.200270

[2] Lin, J.-R., Chen, K.-Y., Song, S.-Y., Cai, Y.-H., Pan, P., & Deng, Y.-C. "Digital twin of buildings and occupants for emergency evacuation: Framework, technologies, applications and trends," Advanced Engineering Informatics, vol. 66, pp. 103419, 2025. https://doi.org/10.1016/j.aei.2025.103419

[3] Zhang, J., Che, X., Fan, Y., Peng, S., Chen, G., Ma, Q., & Hu, J. "Denoising diffusion models with optimized quantum implicit neural networks for image generation," Future Generation Computer Systems, pp. 107875, 2025. https://doi.org/10.1016/j.future.2025.107875

[4] Balasubramanian, S., Cyriac, R., Samad, S. R. A., Karthikeyan, R., & Balamurugan, V. "Optimized memory-augmented deep unfolding network with blockchain-based data preserving cyber security in internet of things," Knowledge-

Based Systems, vol. 321, pp. 113536, 2025. https://doi.org/10.1016/j.knosys.2025.113536

[5] Shekhar, Bhandari, G., & Tyagi, S. "Enhance Quality of Services in Internet of Things using Improved Dynamic Bandwidth Allocation (IDBA) Techniques," Procedia Computer Science, vol. 259, pp. 1270–1281, 2025. https://doi.org/10.1016/j.procs.2025.04.082

[6] Cao, J., Liu, Q., Chen, Z., Zhang, J., & Qi, Z. "Dual-space distribution metric-based evolutionary algorithm for multimodal multi-objective optimization," Expert Systems with Applications, vol. 262, pp. 125596, 2025. https://doi.org/10.1016/j.eswa.2024.125596

[7] Cao, J., Yang, Y., Zhang, J., Chen, Z., & Liu, Z. "A multi-task optimization algorithm via reinforcement learning for multimodal multi-objective optimization," Expert Systems with Applications, vol. 283, pp. 127862, 2025. https://doi.org/10.1016/j.eswa.2025.127862

[8] Cheng, S., Wang, X., Zhang, M., Lei, X., Lu, H., & Shi, Y. "Solving multimodal optimization problems by a knowledge-driven brain storm optimization algorithm," Applied Soft Computing, vol. 150, pp. 111105, 2024. https://doi.org/10.1016/j.asoc.2023.111105

[9] Deng, S., Liu, H., Cheng, K., Xu, J., Li, M., & Rao, H. "Goal-directed multimodal multi-objective evolutionary algorithm converging on population derivation," Swarm and Evolutionary Computation, vol. 92, pp. 101796, 2025. https://doi.org/10.1016/j.swevo.2024.101796

[10] Ding, W., Wang, J., Huang, J., Cheng, C., & Jiang, S. "MFCA: Collaborative prediction algorithm of brain age based on multimodal fuzzy feature fusion," Information Sciences, vol. 687, pp. 121376, 2025. https://doi.org/10.1016/j.ins.2024.121376

[11] Fan, Y., Ding, J., Long, J., & Wu, J. "Modeling and evaluating the travel behaviour in multimodal networks: A path-based unified equilibrium model and a tailored greedy solution algorithm," Transportation Research Part A: Policy and Practice, vol. 182, pp. 104032, 2024. https://doi.org/10.1016/j.tra.2024.104032

[12] Hu, T., Wang, X., Tang, L., & Zhang, Q. "A clustering-assisted adaptive evolutionary algorithm based on decomposition for multimodal multi objective optimization," Swarm and Evolutionary Computation, vol. 91, pp. 101691, 2024. https://doi.org/10.1016/j.swevo.2024.101691

[13] Ipeayeda, F. W., Oyediran, M. O., Ajagbe, S. A., Jooda, J. O., & Adigun, M. O. "Optimized gravitational search algorithm for feature fusion in a multimodal biometric system," Results in Engineering, vol. 20, pp. 101572, 2023. https://doi.org/10.1016/j.rineng.2023.101572

[14] Jia, Y., Qu, L., & Li, X. "A novel multimodal multi-objective differential evolution algorithm based on nearest neighbor-repulsion strategy," Information Sciences, vol. 676, pp. 120832, 2024. https://doi.org/10.1016/j.ins.2024.120832

[15] Lee, H., Baek, J. J., Oh, J. Y., & Lee, T. I. " Multimodal sensing algorithm using thermoelectric dynamics for self-powered skin-like sensory devices," Chemical Engineering Journal, vol. 486, pp. 150168, 2024. https://doi.org/10.1016/j.cej.2024.150168

[16] Li, Z., Liu, X., Zhang, Y., Qin, J., Zheng, W. X., & Wang, J. "Learning high-order fuzzy cognitive maps via multimodal artificial bee colony algorithm and nearest-better clustering: Applications on multivariate time series prediction," Knowledge-Based Systems, vol. 295, pp. 111771, 2024. https://doi.org/10.1016/j.knosys.2024.111771

[17] Li, Z., Rong, H., Yang, S., Yang, X., & Huang, Y. "A dual-population coevolutionary algorithm for balancing convergence and diversity in the decision space in multimodal multi-objective optimization," Applied Soft Computing, vol. 162, pp. 111770, 2024. https://doi.org/10.1016/j.asoc.2024.111770

[18] Lin, F., Zhu, J., & Yang, W. "A multimodal vision-based algorithm for monitoring air supply in aquaculture," Aquaculture, vol. 603, pp. 742395, 2025. https://doi.org/10.1016/j.aquaculture.2025.742395

[19] Liu, H., Xin, X., Song, J., & Peng, W. "CRISP: A cross-modal integration framework based on the surprisingly popular algorithm for multimodal named entity recognition," Neurocomputing, vol. 614, pp. 128792, 2025. https://doi.org/10.1016/j.neucom.2024.128792

[20] Liu, Z., Yang, Y., Cao, J., Zhang, J., Chen, Z., & Liu, Q. "A coevolutionary algorithm using Self-organizing map approach for multimodal multi-objective optimization," Applied Soft Computing, vol. 164, pp. 111954, 2024. https://doi.org/10.1016/j.asoc.2024.111954

[21] Meng, X., & Tan, Y. "Multi-guiding spark fireworks algorithm: Solving multimodal functions by multiple guiding sparks in fireworks algorithm," Swarm and Evolutionary Computation, vol. 85, pp. 101458, 2024. https://doi.org/10.1016/j.swevo.2023.101458

[22] Prasad, E. S., Sonia, S. V. E., Suresh, K. N., & Shivapanchakshari, T. G. "Active and Reactive Power Control in Three-Phase Grid-Connected Electric Vehicles using Zebra Optimization Algorithm and Multimodal Adaptive Spatio-Temporal Graph Neural Network," Renewable Energy Focus, vol. 54, pp. 100715, 2025. https://doi.org/10.1016/j.ref.2025.100715

[23] Sá, J. dos S., Silva, E. do N. da, Gonçalves, L. N., Cardoso, C. M. M., Nascimento, A. A. do, Cavalcante, G. P. dos S., Tostes, M. E. de L., Araújo, J. P. L. de, Barros, F. J. B., & Farias, F. de S. "Multimodal urban mobility solutions for a smart campus using artificial neural networks for route determination and an algorithm for arrival time prediction," Engineering Applications of Artificial Intelligence, vol. 137, pp. 109074, 2024. https://doi.org/10.1016/j.engappai.2024.109074

[24] Sridharan, T. B., & Akilashri, D. P. S. S. "Multimodal learning analytics for students' behavior prediction using multi-scale dilated deep temporal convolution network with improved chameleon Swarm algorithm," Expert Systems with Applications, vol. 286, pp. 128113, 2025. https://doi.org/10.1016/j.eswa.2025.128113

[25] Wang, J., Zhao, Y., Li, X., Zhou, Y., Zhao, K., Wang, H., & Shakweer, W. M. E.-S. "Multimodal fusion-based detection method of estrus cows using multisource data inspired by hidden Markov model algorithms," Computers and Electronics in Agriculture, vol. 235, pp. 110391, 2025. https://doi.org/10.1016/j.compag.2025.110391

[26] Yang, R., Li, D., Han, B., Zhou, W., Yu, Y., Li, Y., & Zhao, P. "Door to door space-time path planning of intercity multimodal transport network using improved ripple-spreading algorithm," Computers & Industrial Engineering, vol. 189, pp. 109996, 2024. https://doi.org/10.1016/j.cie.2024.109996

[27] Yin, L., & Cai, Z. "Multimodal hierarchical distributed multi-objective moth intelligence algorithm for economic dispatch of power systems," Journal of Cleaner Production, vol. 434, pp. 140130, 2024. https://doi.org/10.1016/j.jclepro.2023.140130

[28] Yue, C., Song, J., Liang, J., Liu, M., Yu, K., Lin, H., & Bi, Y. "A multimodal multiobjective evolutionary algorithm based on neighborhood and enhanced special crowding distance," Knowledge-Based Systems, vol. 315, pp. 113340, 2025. https://doi.org/10.1016/j.knosys.2025.113340