

A Large World Model for Person Overcrowding in Indoor Spaces to Prevent Airborne Diseases

Evangelos D. Spyrou^{1,2} and Vassilios Kappatos¹ and Chrysostomos Stylios^{2,3}

Abstract—Airborne diseases can spread more easily in enclosed and overcrowded indoor environments where prolonged close-proximity interactions occur. Consequently, continuous monitoring and predictive management of person flows can support operational interventions aimed at reducing overcrowding and the associated conditions that facilitate airborne-disease transmission. This work introduces a Large World Model (LWM) framework for monitoring, forecasting, and semantically interpreting person overcrowding in indoor spaces, with an airport used as the experimental case study. The proposed architecture integrates camera-based sensing, a Transformer-based world model, future-state forecasting, semantic alignment, and a lightweight language model for human-readable operational recommendations. The monitored state consists of queue length, incoming person flow, and occupancy measurements. From their temporal evolution, the framework identifies current and predicted overcrowding conditions that are relevant to airborne-disease exposure-risk management. The numerical world state is subsequently mapped to a semantic representation describing queue intensity, incoming-flow conditions, occupancy level, temporal evolution, and overcrowding-related health risk. The framework is evaluated using time-series measurements collected from the monitoring infrastructure. The LWM achieves a forecasting Mean Squared Error (MSE) of 0.3254, improving by a significant percentage compared with a persistence-based predictor (MSE = 0.5512). Linear Regression achieves a slightly lower MSE of 0.3142, showing that the main advantage of the proposed framework is not solely point-forecasting accuracy, but the ability to combine temporal world modeling, health-oriented overcrowding interpretation, and natural-language decision support. Although evaluated in an airport environment, the proposed architecture is applicable to other high-density indoor environments such as railway stations, metro stations, museums, and sporting venues.

Index Terms—Airborne diseases, overcrowding, person-flow management, Large World Model, Large Language Model, indoor monitoring, airport

I. INTRODUCTION

COVID-19 and other airborne diseases can present increased transmission risks in enclosed and overcrowded environments, particularly when large numbers of people remain in close proximity for extended periods. Passenger transport facilities are therefore important environments for continuous person-flow monitoring because temporary queues and local occupancy peaks can generate conditions associated with

increased opportunities for airborne exposure. Reducing such overcrowding does not itself constitute an epidemiological model of disease transmission; rather, it represents an operational mitigation mechanism for reducing environmental and behavioral conditions that can facilitate airborne-disease spread.

Accordingly, the objective of the proposed framework is to continuously monitor person flow, queue and occupancy (change in occupancy), forecast upcoming overcrowding states, and provide operational recommendations that can be used to redistribute or manage people before severe crowd accumulation occurs. The framework therefore addresses airborne-disease prevention through predictive overcrowding management rather than through direct estimation of individual infection probability.

Studies have modeled the spread of COVID-19 and human behavior in indoor spaces to understand transmission better [1], [2]. The virus spreads through respiratory droplets or direct contact. Maintaining a specific number of individuals per square meter and organizing passenger flow from disembarkation or check-in throughout the airport can help minimize transmission risk.

Investigating passenger flow in airport terminals employs a range of methods. Traditionally, on-site surveys, including passenger counting and questionnaires, have been used to validate model results or sensor data. However, this approach is labor-intensive, especially in large airports. To enhance efficiency, various technologies have been implemented, such as Wi-Fi data [3], mobile phone information [4], RFID [5], [6], PIR sensors [7] and surveillance videos [8]. These advancements are not only used in airports but also in other buildings. Additionally, simulation models have been developed to analyze and predict passenger flow within airport terminals.

Given the intricate and expansive nature of airport terminals, agent-based simulation (ABS) models are utilized to understand the ongoing fluctuations in passenger flow and distribution. ABS models consider each passenger as an individual entity, applying a social force model to regulate their movements [9]. Several models exist in the literature that provide flow of people in infrastructure [10]–[15]. However, routing with a real application is a quite challenging task. Effective routing of passengers is crucial in indoor spaces to prevent overcrowding. Therefore, an efficient strategy is essential to guide passengers in airports toward common areas in a way that promotes sparsity. Given the high volume of passenger traffic, maintaining safety and ensuring appropriate

¹Hellenic Institute of Transport, Centre for Research and Technology Hellas, 6th Km Charilaou-Thermi, 57001, Thessaloniki, Greece {espyrou, vkappatos}@certh.gr

²Department of Informatics and Telecommunications, University of Ioannina, Kostaki Artas, 47150, Greece {e.spyrou, stylios}@uoi.gr

³Industrial Systems Institute, Athena Research Center, Patras, Greece stylios@athenarc.gr

distancing in shared spaces is of utmost importance. Also related research is given in this comprising social distancing and person detection [16]–[19].

The rapid growth of sensing technologies, Internet-of-Things (IoT) infrastructures, and cyber-physical systems has resulted in unprecedented volumes of continuously generated environmental data. Modern intelligent systems are increasingly required not only to collect and analyze these observations but also to understand the evolution of the underlying physical processes and support timely decision-making. Traditional machine learning approaches have demonstrated strong performance in specific prediction tasks; however, they are typically designed for narrowly defined objectives and lack a persistent representation of the surrounding world that can be queried for higher-level reasoning.

LLMs provide strong capabilities in reasoning, multi-modal processing, and natural-language interaction and can process physical-world information when observations are explicitly provided. However, a stand-alone LLM does not inherently maintain a continuously updated, application-specific temporal representation of a monitored environment. To this end, this work introduces the concept of a LWM as an intermediate intelligence layer between smart camera sensors and an Android application [20] that provide occupancy, queue-length, and incoming person-flow measurements, and language-based reasoning models. Unlike conventional latent-state approaches that compress observations into a task-specific representation, the proposed LWM aims to maintain a rich contextual representation of the observed person traffic by continuously learning from historical camera sensor streams. Through attention-based temporal modeling, the LWM captures complex dependencies among multiple system variables, predicts future states, and provides a structured representation that can be interpreted by downstream reasoning models. The proposed framework combines three fundamental components. First, a camera sensor observation layer continuously collects heterogeneous measurements describing the state of the environment. Second, a Transformer-based Large World Model learns temporal dynamics and generates predictive representations of future system conditions. Third, a semantic alignment mechanism connects the learned world representation with an LLM, enabling natural language reasoning over the physical traffic environment. This integration creates a perception-to-language pipeline where sensor observations can be transformed into forecasts, explanations, and operational recommendations.

The proposed framework does not claim that the Transformer is universally more accurate than conventional forecasting methods. While existing person-flow approaches typically address individual tasks such as counting, occupancy estimation, queue forecasting, congestion detection, or routing, the proposed framework integrates these functions within a persistent representation of the monitored environment. It jointly models queue length, incoming person flow, and occupancy, learns their temporal evolution, predicts

overcrowding, converts numerical states into health-relevant semantic concepts, and generates human-readable recommendations. Its main contribution therefore lies in combining continuous sensing, temporal world modeling, overcrowding-risk interpretation, semantic grounding, and language-based decision support rather than in forecasting alone.

The airport is used as a representative high-throughput indoor environment rather than as a restriction of the proposed methodology. The framework addresses person-flow monitoring, forecasting, and management in enclosed spaces where overcrowding may increase conditions associated with airborne-disease exposure. Its core processing chain of numerical sensing, temporal world-state modeling, overcrowding forecasting, health-oriented semantic interpretation, and recommendation generation can be applied to other high-density environments, including railway and metro stations, terminals, museums, stadiums, and conference venues, by adapting the sensing configuration, occupancy thresholds, and operational responses. Airports are selected as the experimental case study because their variable arrivals, queues, restricted processing areas, and temporary high occupancy provide an appropriate setting for evaluating preventive overcrowding management. Therefore, the proposed LWM represents a general framework for health-aware indoor person-flow management rather than an airport-specific architecture.

II. HARDWARE DESCRIPTION

We placed 4 cameras at the Amygdaleona airport in Kavala in Greece, at specific places close to doors. These cameras were purchases from Footfall (3 cameras) and V-Count (1 camera) respectively. The installations of two indicative cameras can be seen in Fig. 1.

The V-Count cameras occupancy acquisition is simple, since their API requires a request to their cloud with the user name and password, and it returns a single value of the occupancy. The Footfall cameras' API essentially allows for the internal IP of the camera to be hit and the respective data fields are being displayed where the inbound and outbound persons measures can be extracted. This takes place by obtaining an access token and placing dynamically the start and end dates in conjunction with the start and end time. The response is a JSON file with the inbound and outbound values of the camera sensor. In this manner the ability to obtain the occupancy for each spot is provided and the data is sent to the cloud infrastructure and stored in the database there using python scripts. Moreover, in Fig. 2, the camera prototype based on a Raspberry Pi model 4 and a web camera can be seen which is installed close to the aforementioned commercial cameras to detect queues. The software that was built was based on OpenCV and was a simple `ssd_mobilenet_v2_coco_quant_postprocess.tflite` model that counts the persons in the view of the camera and sends the data to the database in the cloud server of the Institute. The data from the queuing cameras are sent to the cloud infrastructure of the Institute using python sockets, whereby the data is stored in the respective fields of the

postgresql database that resides there. The Raspberry Pi device works as a server, in order to accomplish some kind of synchronisation when the client from the server invokes the data transmission. The incoming persons are being measured by an Android application whereby each passenger (person) sets her preferences of which place to visit.



Fig. 1: 1st and 2nd Occupancy Camera Installation.



Fig. 2: Queue Camera Installation

It is important to distinguish the physical sensing modality from the input provided to the proposed LWM. Although cameras constitute the sensing layer, the LWM does not process raw image or video frames. Person detection and counting are performed at the sensing layer, and the resulting numerical measurements are transmitted to the data infrastructure. The LWM therefore operates on time-series values describing queue length, incoming person flow, and occupancy. In this sense, the term camera sensor refers to the source of the numerical observations and not to an image-based input to the Transformer model.

The forecasting and semantic-reasoning components operate only on aggregated numerical person-flow measurements and do not require person identity as an input. In particular, attributes such as names, facial identities, or other biometric identifiers are not part of the LWM state vector. This separation between visual sensing and numerical

world modeling reduces the amount of personal information propagated through the processing pipeline. The commercial sensing infrastructure is operated according to its applicable privacy requirements, while the experimental LWM receives only the numerical measurements required for person-flow and overcrowding analysis.

III. SYSTEM MODEL

We consider an indoor person-flow monitoring and reasoning framework operating over discrete time steps $t \in \mathbb{N}$. The physical sensing layer provides three primary numerical measurements: queue length, incoming person flow, and occupancy. The observed state at time t is therefore defined as

$$\mathbf{x}_t = [q_t \quad f_t^{\text{in}} \quad o_t]^\top, \quad (1)$$

where q_t denotes the number of people detected in the monitored queue, f_t^{in} denotes the number of incoming people during the corresponding sampling interval, and o_t denotes the measured change in occupancy of the monitored indoor area during that interval. Accordingly, $o_t > 0$ indicates a net increase in the number of persons present, $o_t < 0$ indicates a net decrease, and $o_t = 0$ indicates no net change in occupancy. Thus, negative values of o_t are physically meaningful and represent departures exceeding arrivals during the corresponding measurement interval, rather than a negative number of persons physically present in the monitored area.

The purpose of forecasting $\mathbf{x}_t$ is not limited to reproducing the next sensor value. Instead, the predicted state is used to determine whether the monitored area is evolving toward an overcrowded condition that may increase close-proximity interactions and therefore create conditions relevant to airborne-disease transmission.

Given a temporal observation window of length W , the input to the LWM at time t is

$$\mathbf{X}_t = [\mathbf{x}_{t-W+1}, \dots, \mathbf{x}_t], \quad (2)$$

and the temporal world representation is

$$\mathbf{h}_t = \mathcal{F}_\theta(\mathbf{X}_t), \quad (3)$$

where $\mathcal{F}_\theta(\cdot)$ denotes the Transformer-based LWM and $\mathbf{h}_t$ summarizes the recent temporal evolution of queue length, incoming flow, and occupancy.

In the present implementation, the LWM uses attention to model dependencies within a finite temporal window. The objective is therefore not to claim unlimited temporal memory, but to learn a contextual representation that preserves interactions among queue evolution, incoming person flow, and occupancy within the selected observation horizon. Compared with a direct point-prediction model, this representation is subsequently reused for both forecasting and semantic interpretation. The world model is parameterized as a causal Transformer operating on the finite observation window $\mathbf{h}_t = \text{Transformer}_\theta(\mathbf{X}_t)$, with self-attention $\text{Attn}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V$, where queries, keys, and values are computed from sensor embeddings.

The temporal representation is mapped to the predicted person-flow state at the next time step $\hat{\mathbf{x}}_{t+1} = \mathcal{G}_\psi(\mathbf{h}_t)$, where $\hat{\mathbf{x}}_{t+1} = [\hat{q}_{t+1} \ \hat{f}_{t+1}^{\text{in}} \ \hat{o}_{t+1}]^\top$. Therefore, the forecasting module jointly predicts the future queue length, incoming person flow, and occupancy of the monitored environment.

During online inference, the observation window is updated whenever a new measurement becomes available by discarding the oldest observation and appending the newest one. The updated window $\mathbf{X}_t$ is then processed using the fixed trained parameters of the LWM; no online retraining is required.

The LWM and the regression head are jointly trained using a one-step multivariate forecasting objective

$$\mathcal{L}_{\text{pred}}(\theta, \psi) = \frac{1}{N} \sum_{t=1}^N \|\mathbf{x}_{t+1} - \hat{\mathbf{x}}_{t+1}\|_2^2, \quad (4)$$

where the prediction error is evaluated jointly over queue length, incoming person flow, and occupancy.

The proposed framework integrates the LWM with an LLM to enable semantic reasoning and human-readable decision support over the monitored indoor person-flow environment. As defined in (3), the LWM produces the contextual representation $\mathbf{h}_t = \mathcal{F}_\theta(\mathbf{X}_t)$ from the most recent W observations. Thus, $\mathbf{h}_t$ summarizes the finite temporal context of queue length, incoming person flow, and occupancy used by both the forecasting and semantic-reasoning components.

To avoid ambiguity, three representations are distinguished throughout this work. The LWM representation $\mathbf{h}_t$ is the learned numerical temporal representation produced by the Transformer. The structured semantic state $\mathbf{z}_t^{\text{sem}}$ consists of interpretable person-flow and overcrowding-risk concepts derived from the monitored state. Finally, the semantic embedding $\mathbf{e}_t^{\text{text}}$ is the continuous Sentence-BERT representation of the textual description of that structured semantic state.

To interface the numerical world model with language-based reasoning, we first construct an explicit semantic state rather than directly exposing the latent Transformer representation to the language model. The semantic state is defined as

$$\mathbf{z}_t^{\text{sem}} = [Q_t, F_t, O_t, T_t, R_t, \hat{R}_{t+1}], \quad (5)$$

where Q_t denotes the semantic queue condition, F_t denotes incoming-flow intensity, O_t denotes the occupancy condition, T_t denotes the temporal trend, R_t denotes the current overcrowding-related health-risk level, and $\hat{R}_{t+1}$ denotes the predicted risk level.

For example, the semantic state may contain

$$\mathbf{z}_t^{\text{sem}} = [\text{high queue, increasing inflow, high occupancy,} \quad (6)$$

$$\text{increasing, moderate risk, high predicted risk}]. \quad (7)$$

The semantic variables are obtained from the numerical world state through predefined operational discretization rules. Specifically, queue length, incoming person flow, and

occupancy are assigned to categorical operating regions (e.g., low, moderate, or high) according to thresholds defined for the monitored environment. The temporal trend is determined from the recent evolution of the corresponding measurements and predictions. The overcrowding-related health-risk state is then derived from the joint configuration of these variables: simultaneously elevated queue length, incoming flow, and occupancy, particularly under an increasing trend, are mapped to a higher overcrowding-risk category. These categories should not be interpreted as epidemiological estimates of infection probability. Rather, they represent operational indicators of crowding conditions associated with greater opportunities for prolonged close-proximity exposure to airborne pathogens. The thresholds are application-dependent and can therefore be adapted to the monitored indoor environment.

The semantic embedding space therefore does not contain arbitrary linguistic concepts. It specifically represents concepts derived from the monitored person-flow environment: queue intensity, incoming-flow intensity, occupancy, temporal evolution, current overcrowding-related risk, and predicted overcrowding-related risk.

The structured semantic state is converted to a textual description using a deterministic template. For example, “Queue level is high; incoming person flow is increasing; occupancy is high; the current overcrowding-risk level is moderate and is predicted to become high at the next time step.” A pretrained Sentence-BERT encoder maps this textual description to a reference semantic embedding

$$\mathbf{e}_t^{\text{text}} = \text{SBERT}(\mathbf{z}_t^{\text{sem}}). \quad (8)$$

In parallel, a learnable projection network maps the LWM representation to the same embedding dimension

$$\mathbf{e}_t^{\text{LWM}} = g_\phi(\mathbf{h}_t). \quad (9)$$

The semantic alignment objective is then

$$\mathcal{L}_{\text{sem}} = \frac{1}{N} \sum_{t=1}^N \|\mathbf{e}_t^{\text{LWM}} - \mathbf{e}_t^{\text{text}}\|_2^2. \quad (10)$$

The LLM is conditioned on both the user query and the world state representation $\mathbf{y}_t = \text{LLM}_\psi(\mathbf{q}_t, \mathbf{s}_t)$, where $\mathbf{q}_t$ is the natural language query, $\mathbf{s}_t$ is the structured world state, and $\mathbf{y}_t$ is the generated response.

The semantic state and corresponding predicted overcrowding condition are then provided to the lightweight language model as structured textual context. Let $\mathbf{c}_t$ denote this context and $\mathbf{q}_t$ a user or system query. The generated recommendation is $\mathbf{y}_t = \text{LLM}_\psi(\mathbf{q}_t, \mathbf{c}_t)$.

For example, given the structured context “high queue, increasing inflow, high occupancy, and predicted high overcrowding risk”, the language interface can express the corresponding preventive guidance in natural language, e.g., “Consider redirecting incoming persons toward a less occupied area to reduce further crowd accumulation.” The recommendation is intended as human-readable decision

support rather than as an autonomously executed safety-critical control action.

Consequently, the recommendation-generation stage does not independently determine the health-risk level; it verbalizes and contextualizes the overcrowding state already derived from the monitored and predicted person-flow variables. The LLM therefore does not infer airborne-disease transmission from raw sensor measurements. Its role is to express the numerically and semantically derived overcrowding condition in a human-readable form.

The LWM uses a fixed window of $W = 20$ observations rather than an unlimited history, providing bounded inference complexity. Since self-attention scales quadratically with sequence length, longer monitoring horizons would increase memory and computational requirements and could instead employ windowed attention, temporal aggregation, or memory-compressed architectures. Thus, the LWM contribution lies in its reusable contextual world-state representation rather than unlimited temporal memory. The present experiments do not establish superiority over recurrent, state-space, or compressed latent-state models, which may provide similar forecasting performance at lower computational cost. The LWM is adopted because its attention-based representation can support both temporal forecasting and semantic alignment. Systematic comparisons with alternative latent-state models are left for future work.

IV. RESULTS

The framework is evaluated using numerical time-series data from the sensing infrastructure. Although cameras provide physical sensing, the LWM receives only queue length, incoming person flow, and occupancy, with timestamps used solely for chronological ordering. The dataset contains 19,106 observations recorded every second across multiple monitoring sessions, covering stable flow and temporary congestion events. Data are chronologically split into 80% training and 20% testing sets, with missing values forward-filled and all variables standardized using training-set parameters.

Negative values in the forecasting figures may arise either from standardized measurements falling below the training-set mean or, for the occupancy-related variable, from a measured decrease in occupancy during the corresponding sampling interval. Specifically, the occupancy variable represents the change in the number of persons present between measurement intervals; therefore, positive and negative values indicate increases and decreases in occupancy, respectively.

Since predictions and errors are evaluated at each time step without temporal aggregation, short-term fluctuations and congestion peaks relevant to overcrowding prevention are preserved. Each forecasting sample uses a sliding window of $W = 20$ observations.

The LWM uses a two-layer Transformer encoder with four attention heads and projects each three-dimensional observation into a 64-dimensional representation. A fully connected regression head predicts queue length, incoming

person flow, and occupancy. The model is trained for 50 epochs using Adam with a learning rate of 10^{-3} and an MSE objective. Table I reports losses every five epochs, with epoch 45 being the final reported checkpoint before completion of training.

The semantic alignment module is trained separately for 10 epochs using Adam with a learning rate of 10^{-4} . This shorter training period is sufficient because only a lightweight projection network between the learned LWM representation and fixed Sentence-BERT reference embeddings is optimized, without fine-tuning Sentence-BERT. DistilGPT2 is subsequently applied after forecasting and semantic-state construction to generate human-readable recommendations.

The proposed LWM is compared with two forecasting baselines: a Naive persistence model, which assumes that the next observation is equal to the most recently observed state, and a Linear Regression model trained on the same flattened temporal windows. The reported hyperparameters were selected as a fixed proof-of-concept configuration rather than through an exhaustive hyperparameter optimization procedure. The objective of the present evaluation is to assess the feasibility of the complete sensing-world-model-semantic-reasoning pipeline rather than to claim an optimally tuned Transformer architecture. Systematic optimization of the observation-window length, embedding dimension, number of attention heads, encoder depth, and learning rate is left for future work.

Table I summarizes the convergence behaviour of the proposed LWM, the semantic alignment module, and the overall forecasting performance. The LWM forecasting loss decreases rapidly from approximately 1.16 at initialization to 0.53 within the first five epochs, corresponding to a reduction of nearly 54%. This indicates that the self-attention mechanism quickly learns the dominant temporal correlations among queue length, incoming person flow, and occupancy. Beyond epoch 10, the loss decreases more gradually, reaching approximately 0.474 at epoch 45. The marginal improvements observed after approximately 30 epochs indicate that the forecasting model approaches a stable solution and that further training would provide limited gains at additional computational cost. The semantic alignment module also exhibits stable convergence, with its loss decreasing from approximately 0.058 at initialization to 0.049 after ten epochs, corresponding to an improvement of approximately 15%. This result indicates that the latent representation learned by the LWM can be consistently projected into the semantic embedding space produced by Sentence-BERT, thereby associating numerical sensor observations with meaningful natural-language concepts without affecting the forecasting process. The resulting aligned representation enables the language model to generate context-aware and human-interpretable operational guidance, as illustrated by the recommendation “High incoming load approaching saturation. Suggest action.”

The generated recommendation is included as an illustrative output of the language interface and should not be interpreted as a quantitative evaluation of the LLM component. The present experiments primarily evaluate the forecasting and semantic-alignment stages. A systematic evaluation of recommendation correctness, consistency, usefulness, and sensitivity to upstream forecasting errors requires a dedicated set of reference scenarios and human- or rule-validated target recommendations and is left for future work.

The forecasting results show that the LWM achieves an MSE of 0.3254, compared with 0.5512 for the persistence-based Naive predictor and 0.3142 for Linear Regression. Therefore, the proposed LWM improves substantially over persistence but does not outperform Linear Regression in terms of one-step point-forecasting MSE. Linear Regression is approximately 3.5% better according to this metric, indicating that the present dataset contains strong approximately linear temporal relationships.

The competitive performance of Linear Regression is also consistent with the evaluation setting adopted here. Forecasting is performed one time step ahead from measurements sampled every second, and therefore consecutive observations can exhibit strong short-term temporal correlation. Under such conditions, a simple linear predictor can provide a strong numerical baseline. The current experiment should consequently be interpreted as a proof of concept for the additional world-state and semantic-reasoning functionality of the LWM rather than as evidence that a Transformer is required for one-step forecasting of the present dataset.

The motivation for retaining the LWM is therefore not superior one-step forecasting accuracy on the present dataset. Instead, this experiment investigates whether a learned temporal world-state representation can serve as a common intermediate representation for forecasting, semantic alignment, overcrowding-risk interpretation, and language-based decision support. The results show that this additional functionality can be obtained while maintaining forecasting accuracy close to the Linear Regression baseline, although the present evaluation does not yet establish that this added functionality justifies the additional computational complexity.

Nevertheless, the strong Linear Regression result indicates that a particularly important extension is to investigate a hybrid configuration in which Linear Regression provides the numerical forecast and the same downstream semantic/language interface generates the recommendation. Such a comparison is outside the scope of the present proof-of-concept evaluation and is identified as future work.

Fig 3 illustrates the ability of the proposed LWM to predict future queue evolution based on the historical sequence of queue length, incoming traffic, and available occupancy measurements. During the initial period, up to approximately time step 1700, the system remains in a low-load operating regime corresponding approximately to the training-set mean. This indicates that the Transformer encoder successfully learns the regular temporal patterns of the monitored

TABLE I: Training convergence and forecasting performance of the proposed LWM framework.

Training Stage	Epoch	Loss / Metric
LWM Training	0	1.1597
	5	0.5303
	10	0.4932
	15	0.4916
	20	0.4832
	25	0.4797
	30	0.4772
	35	0.4762
	40	0.4747
	45	0.4735
Semantic Alignment	0	0.0580
	5	0.0534
	10	0.0492
LWM Forecasting MSE		0.3254
Naive Baseline MSE		0.5512
Linear Regression MSE		0.3142
LLM Recommendation	High incoming load approaching saturation. Suggest action.	

system. Three significant congestion events are observed around time steps 1800–2200 and 3000–3200, where the measured queue reaches values close to 3.8. The LWM successfully identifies the occurrence and temporal location of these congestion periods; however, the predicted peaks are smoother, reaching approximately 2.0. This behavior results from the temporal representation learning mechanism of the LWM, which prioritizes consistent evolution of the system state and avoids excessive sensitivity to isolated measurements. The LWM therefore captures the approximate temporal occurrence of the major queue bursts but underestimates their peak amplitude. In addition, a smaller predicted increase that is not matched by a corresponding ground-truth event constitutes a false-positive congestion indication. These observations show that the current model is more reliable for detecting broad changes in flow regime than for reproducing the exact magnitude of abrupt queue peaks. This limitation is important for disease-prevention applications because underestimating extreme overcrowding may delay an intervention, and it is therefore explicitly considered in the evaluation.

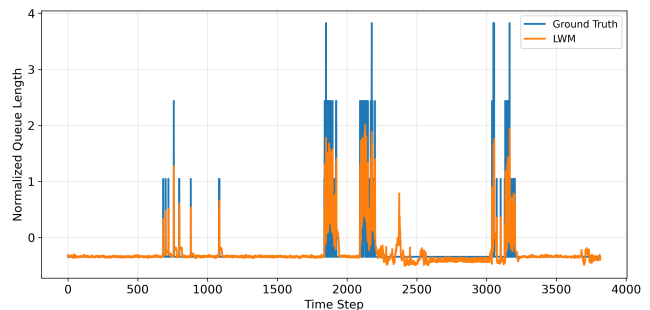


Fig. 3: Queue Forecast

Fig. 4 evaluates the proposed LWM against a persistence-based Naive predictor and a Linear Regression model. The Naive baseline remains close to the previous system state and is therefore unable to anticipate sudden congestion increases.

This limitation becomes evident during the major traffic events, where the difference between the actual queue value and the prediction approaches approximately 4 normalized units. The Linear Regression baseline follows the general trend of the data more effectively but produces additional fluctuations, particularly after time step 2200, where several oscillations appear despite the absence of significant congestion changes. In comparison, the LWM provides smoother and more consistent predictions while preserving the timing of congestion events. The obtained MSE values confirm this behavior, with the LWM achieving an error of 0.3254 compared with 0.5512 for the Naive model. Although Linear Regression achieves a slightly lower MSE of 0.3142, the difference is only approximately 3.5%, while the LWM additionally provides a learned latent representation suitable for semantic interpretation and interaction with a large language model.

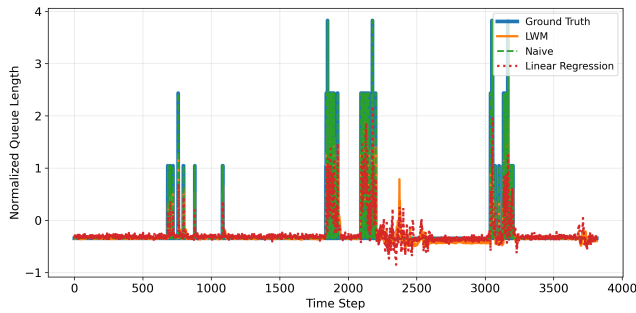


Fig. 4: Baseline Comparison

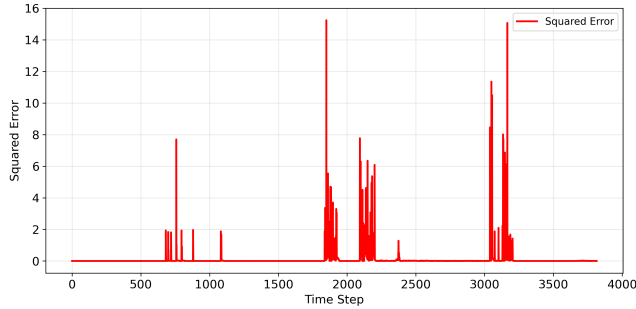
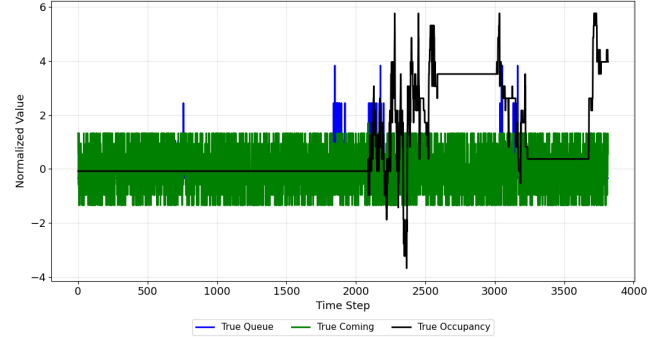


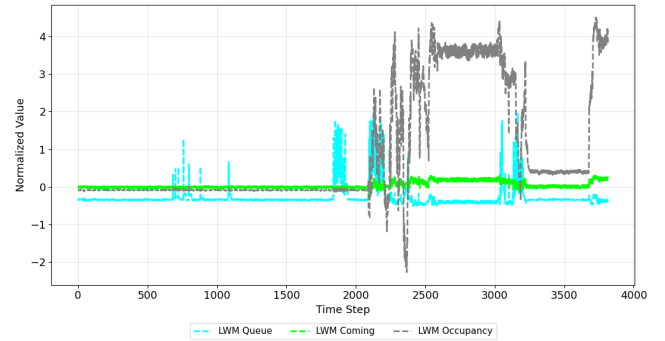
Fig. 5: Prediction Error

Fig. 5 depicts the temporal distribution of forecasting errors produced by the LWM. The error remains almost zero during normal operating conditions, demonstrating that the model accurately represents the regular dynamics of the system. The dominant error contributions occur only during abrupt transitions, particularly around time steps 1800 and 3100, where the squared error increases to approximately 15. These peaks correspond to the high-congestion periods, where the LWM smooths the maximum queue values. Importantly, the largest errors are concentrated around a relatively small number of abrupt congestion events, whereas

substantially smaller errors are observed throughout the more stable portions of the evaluation sequence. This indicates that the LWM is highly effective during normal operation and that future improvements should focus primarily on rare-event and peak-demand prediction.



(a) True evolution of queue length, incoming person flow, and occupancy.



(b) LWM-predicted evolution of queue length, incoming person flow, and occupancy.

Fig. 6: Multivariate forecasting results.

Fig. 6a presents the ground-truth evolution of queue length, incoming person flow, and occupancy over approximately 3,800 time steps. Incoming flow fluctuates mainly between approximately -1.4 and 1.3 , while the queue exhibits distinct peaks reaching about 3.8, particularly around time steps 1,800–2,500 and 3,050. Occupancy shows the largest variations, ranging from approximately -3.7 to 5.8 , with pronounced changes after time step 2,100. Periods where increased queue activity coincides with high occupancy are particularly relevant to the proposed framework, as they indicate overcrowding conditions that may increase close-proximity interactions and are subsequently used for health-oriented semantic interpretation. Fig. 6b presents the corresponding LWM predictions for queue length, incoming person flow, and occupancy. The model captures the main temporal changes observed in the ground truth, particularly the increased activity between approximately time steps 2,100–2,600 and around 3,050. Predicted occupancy shows the strongest variation, reaching approximately 4.5 and maintaining an elevated level of about 3.5 between time steps

2,600 and 3,000. The predicted queue also responds to the main congestion periods, although its peaks are smoother and lower than those observed in the ground truth, while incoming person flow remains comparatively stable around zero.

V. CONCLUSIONS

This paper introduced an LWM framework for monitoring and predicting indoor person overcrowding to support preventive actions relevant to airborne-disease transmission. Using camera-derived numerical measurements of queue length, incoming person flow, and occupancy, the Transformer-based LWM maintains an updated representation of the environment, predicts future person-flow states, and maps overcrowding conditions associated with increased close-proximity interactions to a health-risk state. Rather than estimating individual infection probability, this state supports preventive interventions before severe crowd accumulation occurs. The LWM achieves an MSE of 0.3254, compared with 0.5512 for the persistence baseline and 0.3142 for Linear Regression. Thus, although it does not outperform Linear Regression in one-step forecasting accuracy, the LWM additionally provides a reusable temporal world representation, semantic interpretation of numerical person-flow states, identification of health-relevant overcrowding conditions, and human-readable recommendations. The airport serves as an experimental case study rather than a methodological restriction. By adapting the sensing configuration and overcrowding thresholds, the same sensing–forecasting–risk interpretation–semantic reasoning architecture can be applied to other high-density indoor environments, including railway and metro stations, museums, stadiums, and event venues.

Future work will evaluate the LWM on larger and more diverse datasets, including abrupt queue bursts, concept drift, and different indoor environments, and compare it with stream-mining, adaptive online, recurrent, state-space, compressed latent-state, and lightweight temporal models. A hybrid Linear-Regression–semantic–language pipeline and rule-based and fuzzy-logic recommendation approaches will also be examined to isolate the LWM’s benefits and assess the added complexity of language-based interaction. End-to-end evaluation will consider semantic and recommendation accuracy, false-positive and false-negative overcrowding detection, error propagation, latency, memory, and energy consumption. Finally, different temporal aggregation intervals and environmental variables, including ventilation, indoor-air-quality, and CO₂, will be investigated to strengthen the connection between overcrowding monitoring and airborne-disease exposure assessment.

ACKNOWLEDGEMENT

”This research work was supported by the project : “UNCAN-CONNECT: Decentralized Collaborative Network for Advancing Cancer Research and Innovation” funding from the European Union’s Horizon Europe Research

and Innovation programme under Grant Agreement No 101215206 HORIZON-MISS-2024-CANCER-01-01.

REFERENCES

- [1] E. Cuevas, “An agent-based model to evaluate the covid-19 transmission risks in facilities,” *Computers in biology and medicine*, vol. 121, p. 103827, 2020.
- [2] T. Harweg, D. Bachmann, and F. Weichert, “Agent-based simulation of pedestrian dynamics for exposure time estimation in epidemic risk assessment,” *Journal of Public Health*, pp. 1–8, 2021.
- [3] H. Nikoue, A. Marzuoli, J.-P. Clarke, E. Feron, and J. Peters, “Passenger flow predictions at sydney international airport: a data-driven queuing approach,” *arXiv preprint arXiv:1508.04839*, 2015.
- [4] G. Zhong, X. Wan, J. Zhang, T. Yin, and B. Ran, “Characterizing passenger flow for a transportation hub based on mobile phone data,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 18, no. 6, pp. 1507–1518, 2016.
- [5] N. Li, G. Calis, and B. Becerik-Gerber, “Measuring and monitoring occupancy with an rfid based system for demand-driven hvac operations,” *Automation in construction*, vol. 24, pp. 89–99, 2012.
- [6] E. Soltanaghaei and K. Whitehouse, “Practical occupancy detection for programmable and smart thermostats,” *Applied Energy*, vol. 220, pp. 842–855, 2018.
- [7] Y. Benezeth, H. Laurent, B. Emile, and C. Rosenberger, “Towards a sensor for detecting human presence and characterizing activity,” *Energy and Buildings*, vol. 43, no. 2-3, pp. 305–314, 2011.
- [8] J. Zou, Q. Zhao, W. Yang, and F. Wang, “Occupancy detection in the office by analyzing surveillance videos and its application to building energy conservation,” *Energy and Buildings*, vol. 152, pp. 385–398, 2017.
- [9] P. P.-Y. Wu and K. Mengersen, “A review of models and model usage scenarios for an airport complex system,” *Transportation Research Part A: Policy and Practice*, vol. 47, pp. 124–140, 2013.
- [10] S. Alodhaibi, R. L. Burdett, and P. K. Yarlagadda, “Framework for airport outbound passenger flow modelling,” *Procedia Engineering*, vol. 174, pp. 1100–1109, 2017.
- [11] K. Li, K. Zhang, X. Li, J. Xi, H. Fang, and Z. Jia, “Building occupancy estimation with people flow modeling in anylogic,” in *2016 12th IEEE International Conference on Control and Automation (ICCA)*. IEEE, 2016, pp. 669–672.
- [12] Y. Yang, J. Li, and Q. Zhao, “Study on passenger flow simulation in urban subway station based on anylogic,” *Journal of Software*, vol. 9, no. 1, pp. 140–146, 2014.
- [13] P. F. i Casas, J. Casanovas, and X. Ferran, “Passenger flow simulation in a hub airport: an application to the barcelona international airport,” *Simulation Modelling Practice and Theory*, vol. 44, pp. 78–94, 2014.
- [14] H. S. Ghali, Y. A. Essawy, A. Abdullah, and K. Nassar, “Analyzing passenger flow lines and optimizing the facility layout management for cairo metro elevated stations,” in *Construction Research Congress 2022*, 2022, pp. 1327–1336.
- [15] D. Duma and R. Aringhieri, “Mining the patient flow through an emergency department to deal with overcrowding,” in *Health Care Systems Engineering: HCSE, Florence, Italy, May 2017 3*. Springer, 2017, pp. 49–59.
- [16] S. N. S. Al-Humairi and A. A. A. Kamal, “Design a smart infrastructure monitoring system: a response in the age of covid-19 pandemic,” *Innovative Infrastructure Solutions*, vol. 6, no. 3, p. 144, 2021.
- [17] N. S. Pun, S. K. Sonbhadra, S. Agarwal, and G. Rai, “Monitoring covid-19 social distancing with person detection and tracking via fine-tuned yolo v3 and deepsort techniques,” *arXiv preprint arXiv:2005.01385*, 2020.
- [18] M. Rezaei and M. Azarmi, “Deepsocial: Social distancing monitoring and infection risk assessment in covid-19 pandemic,” *Applied Sciences*, vol. 10, no. 21, p. 7514, 2020.
- [19] S. Saponara, A. Elhanashi, and A. Gagliardi, “Implementing a real-time, ai-based, people detection and social distancing measuring system for covid-19,” *Journal of Real-Time Image Processing*, vol. 18, no. 6, pp. 1937–1947, 2021.
- [20] E. D. Spyrou and V. Kappatos, “A retrieval-based ai chatbot for indoor building passenger information targeting airborne disease spreading prevention,” *IEEE Internet of Things Journal*, 2026.