

ViFusionTST: Deep Fusion of Time-Series Image Representations from Load Signals for Early Bed-Exit Prediction *

Hao Liu¹, Yu Hu¹, Rakiba Rayhana¹, Ling Bai¹ and Zheng Liu¹

¹The University of British Columbia, Canada

{hao.liu, yu.hu, rakiba.rayhana, ling.bai, zheng.liu}@ubc.ca

Abstract

Bed-related falls remain a major source of injury in hospitals and long-term care facilities, yet many commercial alarms trigger only after a patient has already left the bed. We show that early bed-exit intent can be predicted using only one low-cost load cell mounted under a bed leg. The resulting load signals are first converted into a compact set of complementary images: an RGB line plot that preserves raw waveforms and three texture maps—recurrence plot, Markov transition field, and Gramian angular field—that expose higher-order dynamics. We introduce ViFusionTST, a dual-stream Swin Transformer that processes the line plot and texture maps in parallel and fuses them through cross-attention to learn data-driven modality weights.

To provide a realistic benchmark, we collected six months of continuous data from 95 beds in a long-term-care facility. On this real-world dataset ViFusionTST reaches an accuracy of 0.885 and an F1 score of 0.794, surpassing recent 1D and 2D time-series baselines across F1, recall, accuracy, and AUPRC. The results demonstrate that image-based fusion of load-sensor signals for time series classification is a practical and effective solution for real-time, privacy-preserving fall prevention.

1 Introduction

Falls are a frequent and costly issue among older adults in hospitals and long-term care facilities, with U.S. hospitals reporting 700,000 to 1 million falls annually, and up to 10% resulting in serious injury or death [LeLaurin and Shorr, 2019]. Many falls occur when patients try to leave their beds unassisted [Hitcho *et al.*, 2004; Rose *et al.*, 2020], making reliable bed-exit alarm systems crucial for fall prevention [Seow *et al.*, 2022].

*Accepted for presentation at the AI4TS Workshop of IJCAI 2025. This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC) under ALLRP 571106-21. (Corresponding author: Zheng Liu)

Conventional alarms typically use pressure mats or infrared beams, which often alert caregivers too late, after the patient has already exited the bed [Capezuti *et al.*, 2009]. Recent research has shifted toward predicting bed-exit intentions by detecting pre-exit behaviors, allowing timely warnings to caregivers [Meng *et al.*, 2024].

Various sensing methods, including wearables [Awais *et al.*, 2019], cameras [Chen *et al.*, 2018; Bu *et al.*, 2021; Hao *et al.*, 2014], pressure mats [Meng *et al.*, 2024; Kuwahara and Wada, 2017; Bai *et al.*, 2023], and vibration sensors [Valero *et al.*, 2021], have been explored for bed-exit prediction, but each has practical deployment challenges. Wearables can irritate skin and need frequent charging, making them unsuitable for continuous monitoring. Cameras pose privacy issues and require significant storage and processing power, while also struggling with lighting changes and occlusion [Bisogni *et al.*, 2022]. Pressure mats offer detailed pressure information but are costly and can reduce comfort, especially on adjustable beds. Vibration sensors can be easily disrupted by ambient noises. Consequently, most studies focus on lab experiments rather than real-world applications.

We argue that effective bed-exit intention prediction can be achieved without complex sensor systems. Bed-leg-mounted load sensors offer a low-cost, non-intrusive, and privacy-friendly solution by continuously monitoring vertical force to capture movements leading up to a bed exit, potentially providing sufficient data for accurate real-time prediction. In this study, we convert the one-dimensional (1D) load sensor signals into a set of complementary two-dimensional (2D) images and utilize powerful vision models for time-series modeling, aiming to predict bed exits accurately in real time.

Traditionally, 1D sequence models, ranging from multi-layer perceptrons (MLPs) to deep convolutional neural networks (CNNs)/recurrent neural networks (RNNs), were developed expressly for time-series signals such as bed loads [Zeng *et al.*, 2023; Bai *et al.*, 2018]. In recent years, the field has progressed to Time Series Transformers [Zerveas *et al.*, 2021] and even large language models (LLMs) [Jin and others, 2024]. Meanwhile, Vision Transformers [Dosovitskiy and others, 2021; Liu *et al.*, 2021] have achieved remarkable success in computer vision, motivating researchers to explore whether the same architectures can model temporal data [Chen *et al.*, 2024]. This exploration is enabled by a rich set of techniques that turn 1D time series into 2D

images: line plots for multivariate signals and, for univariate signals, recurrence plots (RP), Markov transition fields (MTF), and Gramian angular fields (GAF) [Li *et al.*, 2024; Costa *et al.*, 2024]. Compared with raw numeric sequences, these images give both humans and neural networks a more intuitive view of temporal patterns. Vision methods for time series bring two practical benefits: 1) by pretraining on a large number of images, vision models have already learned common patterns such as trends, cycles, and peaks; 2) representing a long sequence as a compact image reduces input length, easing the burden of modeling long-range dependencies [Ni *et al.*, 2025].

Yet most prior work [Barra *et al.*, 2020; Li *et al.*, 2024; Jin *et al.*, 2023; Karami *et al.*, 2024] relies on only one image modality—typically a line plot, RP, or GAF—to represent the entire sequence, which limits the information available to the model. FIRTS [Costa *et al.*, 2024] combines RP, MTF, and GAF channels, but it omits the raw-waveform line plot, and its simple channel summation cannot fully integrate the complementary cues each modality provides. These gaps motivate the multi-modal fusion strategy we adopt in this study.

We propose ViFusionTST, a dual-stream model that processes (i) the line-plot RGB image and (ii) the three texture maps (RP, MTF, GAF) in two parallel Swin encoders, then fuses them through cross-attention to learn optimal modality weights automatically. Our design preserves the low-cost, privacy-friendly sensing hardware while leveraging state-of-the-art vision models for bed monitoring. To build a realistic benchmark, we collected 6 months of data from 95 beds in a long-term care facility. On this real-world dataset, ViFusionTST outperforms recent state-of-the-art 1D and 2D time series baselines in F1 score, recall, accuracy, and AUPRC. These experimental results demonstrate that ViFusionTST advances image-based time-series modeling and significantly improves prediction performance on the bed-exit task.

2 Related Work

2.1 Bed Exit Prediction

Predicting a bed exit before it happens remains difficult in healthcare practice. Most existing systems only raise an alarm after the patient leaves the bed [Lin *et al.*, 2022; Chang *et al.*, 2021; Inoue and Taguchi, 2020], leaving caregivers little time to act.

Early studies have sought to bridge this gap by employing camera-based human-action recognition methods. Chen *et al.* [Chen *et al.*, 2018] used a depth camera and a 3D CNN to detect two pre-exit motions: torso rotation while sitting and rolling out of bed. However, their study provided only qualitative examples of predictive capability without a quantitative analysis of its accuracy. BED-net [Bu *et al.*, 2021] extended this idea with an R(2+1)D backbone and reached 87.7% accuracy detecting transitional movements, yet performance dropped when the camera angle changed or blankets covered the subject.

Other researchers have explored pressure mats or vibration sensors. Meng *et al.* [Meng *et al.*, 2024] trained a 1-D CNN on air-spring mattress pressure data and achieved 93.5% overall accuracy in a laboratory setting, but did not report

recall or precision. Valero *et al.* [Valero *et al.*, 2021] monitored bed vibrations to recognize one pre-exit posture - sitting. The system was not trained or evaluated to detect other exit patterns, limiting its practical scope.

Across different modalities, existing methods either expose sensitive imagery of patients, require expensive hardware, or cover only a limited set of bed-exit behaviors. We present the first bed-exit prediction system that works with a single low-cost load cell mounted under one bed leg. The sensor fits any bed frame, adds minimal hardware cost, and scales easily. Moreover, our model learns the common cues across diverse bed-exit motions directly from data, without hard-coding specific transitions, making it flexible for real-world use.

2.2 Deep Learning for Time-Series Prediction

1D Time Series

Time series data are most naturally and commonly represented as 1D sequences, where each timestamp corresponds to a vector measurement in the multivariate case. This representation directly models raw signals, thus it is widely applied in diverse time series tasks.

Transformers, since their introduction in 2017 by Vaswani *et al.* [Vaswani and others, 2017], have revolutionized a wide range of domains, including time series analysis [Wen and others, 2023]. Notable adaptations include TST, which applies the vanilla Transformer encoder to multivariate sequences [Zerveas *et al.*, 2021]; iTransformer, which treats each variable as a token to emphasize cross-variable interactions [Liu and others, 2023]. Meanwhile, lightweight non-attention architectures—MLP-based (e.g. DLinear, TSMixer [Zeng *et al.*, 2023; Chen *et al.*, 2023]), CNN-based (e.g. TCN, TimesNet [Bai *et al.*, 2018; Wu *et al.*, 2023]), and state-space-based (e.g. Mamba [Gu and Dao, 2023])—discard global attention to reduce latency and memory usage, yet still achieve competitive benchmark results.

Despite these advances, 1D methods struggle to capture complex hidden patterns in noisy signal data. These methods also do not perform well on long context windows and irregularly sampled sequences. This limitation motivates mapping time series to 2D images to exploit pretrained vision models’ spatial representation ability.

Time Series as 2D Images

2D transformations such as line plot, RP, MTF, and GAF map temporal dynamics into images, turning temporal patterns into spatial textures. These transformations let us exploit the powerful spatial feature extraction of pretrained vision backbones including ResNet [He *et al.*, 2016], ViT [Dosovitskiy and others, 2021] and Swin Transformer [Liu *et al.*, 2021]. These methods have proven useful by previous studies, especially in classification tasks. FIRTS [Costa *et al.*, 2024] converts univariate time series into RP, MTF, and GAF then sums these three representations into a single grayscale image and uses GoogLeNet for time series classification. [Jin *et al.*, 2023] use CNNs on line plot images for univariate classification, and ViTST [Li *et al.*, 2024] extends this approach by applying line plot images and a Swin Transformer to multivariate and irregularly sampled time series. ViTST outperforms

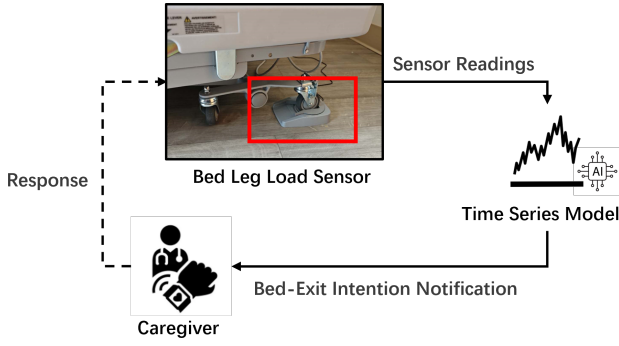


Figure 1: Overview of the strain gauge-based bed monitoring platform.

1D methods on irregular data while remaining competitive on regular sequences.

Although single-modality line plots are simple to generate, they fail to capture the full information of temporal dynamics. The RP, MTF, and GAF modalities provide complementary insights, but naive summation by FIRTs blends their information indiscriminately, which is insufficient to fully fuse and balance different modality contributions. In contrast, ViFusionTST employs dual Swin encoders to extract spatial features from line plot, RP, MTF, and GAF modalities and applies cross-attention fusion to learn optimal modality weights, integrating complementary information for enhanced performance on the time series task.

3 Methodology

3.1 Bed Monitoring Platform Overview

In this study, we developed a bed monitoring platform that enables unobtrusive, non-contact monitoring of hospital beds, as illustrated in Fig. 1. The bed sensor employs a strain gauge-based load cell housed in an enclosure with a built-in ramp and platform, allowing a bed wheel to roll up and remain securely on top. The sensor is powered by a standard power cord and transmits data wirelessly to our AI platform through an IoT (Internet of Things) network.

The bed sensor records three primary signals: load, vibration, and bed occupancy status. The load signal is directly measured by the load cell. The vibration signal is derived from the load signal using a band-pass filter between 0.5 Hz and 10 Hz. Bed occupancy status is determined by an adaptive thresholding method, which dynamically calculates and distinguishes between empty bed and occupied bed weights, thus accurately indicating bed occupancy.

In healthcare practice, patients frequently exhibit a distinct transition phase before leaving their beds, as demonstrated in Fig. 2. During this transition, patients typically move towards the bed edge, causing noticeable fluctuations in load sensor measurements and heightened vibration signals. The transition period generally spans from 1 to 10 minutes prior to an actual bed exit, varying according to individual patient physical abilities. Conversely, non-active periods represent phases of minimal movement or minor repositioning, occupying the

majority of the in-bed duration, and preceding the transition phase and subsequent bed exit.

The primary objective of our approach, implemented through the proposed ViFusionTST model, is to accurately distinguish between genuine pre-exit behaviors and routine movements such as repositioning or resting. ViFusionTST analyzes patterns within the load sensor signals to identify transitions indicating a patient’s intention to exit the bed. Upon detecting these pre-exit patterns, the system promptly alerts caregivers through connected devices, facilitating timely interventions.

3.2 Image Representations of Time Series

To use powerful modern pretrained vision backbones as spatial feature extractors for modeling complex temporal correlations, we transform the 1D time-series signals from the bed monitoring platform into various 2D image representations. Specifically, we employ four distinct transformations: Line Plot, RP, MTF, and GAF. Each representation captures unique aspects of the temporal data, enhancing the model’s ability to identify critical patterns associated with bed-exit behavior.

Line Plot

The line plot representation offers a direct, human-intuitive visualization of multivariate time series, similar to how one inspects temporal patterns by eye. Following the ViTST methodology [Li *et al.*, 2024], we render each time points as a marker and connect markers with straight lines. We visualize each variable separately as distinct subplots and differentiate them by color. In our study, besides the three primary signals—load, vibration, and bed occupancy status—we additionally calculate a derived feature, in-bed duration, as a fourth variable to capture longer-term occupancy behavior. These four variables are arranged into a 2×2 subplot configuration, as illustrated in Fig. 3.

Recurrence Plot (RP)

RP [Eckmann *et al.*, 1995] encodes the times at which a dynamical system revisits a similar state. Given a univariate time series $\{x_t\}_{t=1}^N$, we first use the 1D values directly to reconstruct a phase-space trajectory. The recurrence matrix $\mathbf{R} \in \{0, 1\}^{N \times N}$ is defined as

$$R_{i,j} = \Theta(\varepsilon - \|x_i - x_j\|), \quad (1)$$

where $\Theta(\cdot)$ is the Heaviside step function and ε is a distance threshold. A black pixel at (i, j) indicates that the states at times i and j are within ε of each other.

Markov Transition Field (MTF)

MTF [Wang *et al.*, 2015; Wang and Oates, 2015] captures the transition probabilities of quantized states over time, projected back onto the temporal axis. We first quantize each x_t into one of Q discrete bins by

$$b_t = \text{bin}(x_t) \in \{1, 2, \dots, Q\}. \quad (2)$$

We estimate the $Q \times Q$ state-transition matrix \mathbf{P} , where

$$P_{k,\ell} = \Pr(b_{t+1} = \ell \mid b_t = k) \quad (k, \ell = 1, \dots, Q). \quad (3)$$

The MTF is then the $N \times N$ matrix \mathbf{M} with entries

$$M_{i,j} = P_{b_i, b_j}, \quad (4)$$

so that each element encodes the transition likelihood between the quantized values at times i and j .

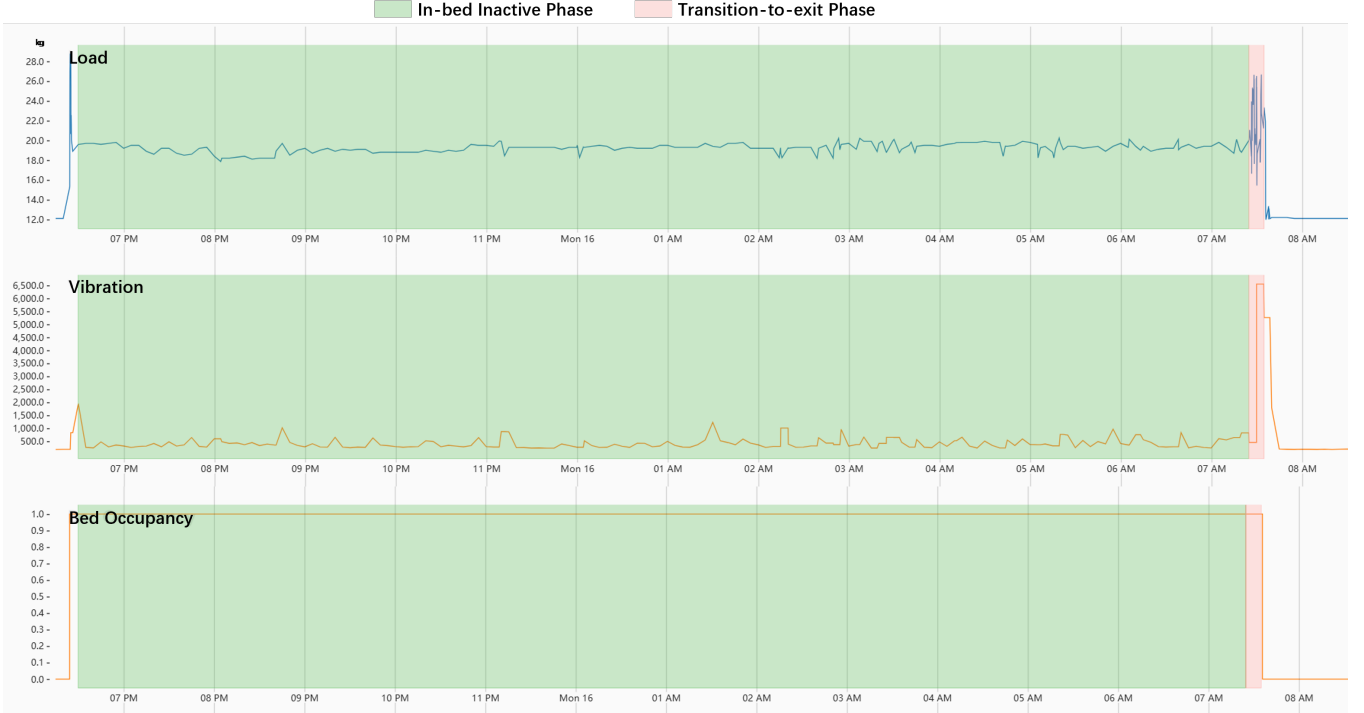


Figure 2: Sensor signals capturing the in-bed period and bed-exit transition.

Gramian Angular Field (GAF)

GAF [Wang *et al.*, 2015; Wang and Oates, 2015] represents time-series values in polar coordinates and captures pairwise angular relationships. First, we rescale the series x_t to $[-1, 1]$, obtaining \tilde{x}_t . Then encode each \tilde{x}_t as an angle

$$\phi_t = \arccos(\tilde{x}_t) \in [0, \pi], \quad t = 1, \dots, N. \quad (5)$$

The Gramian Angular Summation Field (GASF) is defined by

$$G_{i,j}^{(\text{sum})} = \cos(\phi_i + \phi_j). \quad (6)$$

The resulting $N \times N$ matrix encodes global temporal dependencies via angular correlations between every pair of time points.

Since the RP, MTF, and GAF methods inherently apply to univariate time series, we select the most informative feature, load, for these transformations. The resulting three representations—RP, MTF, and GAF—are combined as distinct image channels to form a unified 3-channel input to our deep learning model, as shown in Fig. 3.

3.3 ViFusionTST: Deep Fusion for Bed Exit Prediction

Given the image representations derived from the time-series signals, we propose ViFusionTST, a deep fusion model that consists of two parallel Swin Transformer vision encoders, each processing distinct image representation modalities, and a deep cross-attention fusion module to effectively integrate their complementary information. Specifically, one vision encoder processes the line plot composed of RGB channels, while the other processes a three-channel image created from

the RP, MTF, and GAF. The overall model architecture of ViFusionTST is illustrated in Fig. 3.

The motivation for this design comes from the observation that line plots and field-based representations emphasize different temporal characteristics. Line plots inherently preserve the raw waveform details—such as peaks, slopes, and local fluctuations—allowing vision models to interpret these patterns akin to spatial textures and edges in conventional images. In contrast, field-based encodings (RP, MTF, and GAF) reveal higher-order temporal dynamics, including state recurrences, transition probabilities between quantized states, and global angular correlations, respectively. By assigning each modality group to its own vision encoder, we let each network specialize on the unique statistical patterns in its input. The encoder handling RGB line plots specializes in capturing fine-grained, local temporal patterns, whereas the encoder processing RP, MTF, and GAF representations is optimized for detecting global and structural dependencies within the time series. Subsequently, a cross-attention fusion module dynamically re-weights and integrates the extracted features from both streams, producing joint representations.

Moreover, our approach utilizes the Swin-Tiny Transformer encoder rather than the Swin-Base backbone used in ViTST. We empirically observed that Swin-Tiny is better suited for simpler image representations derived from time series data, significantly mitigating overfitting while reducing computational overhead. This choice achieves a beneficial balance between model complexity and performance, making ViFusionTST efficient, accurate, and practical for real-world bed-exit prediction scenarios.

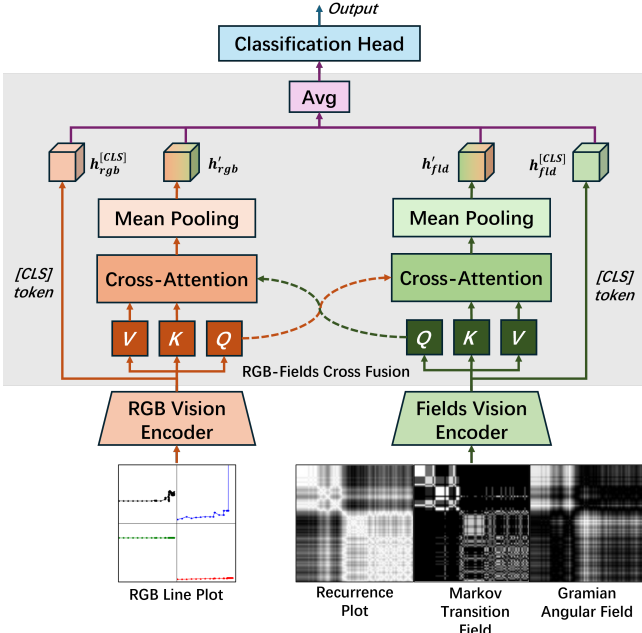


Figure 3: ViFusionTST model architecture.

4 Experiments and Results

In this section, we describe our experimental setup and evaluate our model against state-of-the-art methods. We also conduct ablation and performance analysis to validate our approach.

4.1 Experimental Setup

Data Collection & Preparation

For this study, we collected sensor readings from our bed sensor platform deployed across 95 beds at a long-term care facility in British Columbia, Canada. Data was recorded over a period spanning December 21, 2023, to May 30, 2024. To ensure a diverse representation of patient behaviors, mobility levels, and sensor configurations, we manually reviewed and selected data from 57 beds, totaling 579 days of sensor readings for labeling.

The manual labeling process focused on accurately defining transition periods and non-active periods, as illustrated in Fig. 2. A total of 1211 transition periods were annotated, along with corresponding non-active periods. These samples were then split 60%/20%/20% for training, validation, and testing, respectively.

Implementation Details

The proposed ViFusionTST is implemented in Pytorch and trained using the AdamW optimizer with a batch size of 256 and a learning rate of 2×10^{-5} . Input image representations of time series, including line plots, RP, MTF, and GAF, are all resized to 224×224 to match the Swin-tiny backbone pretrained on ImageNet. A 3-hour look-back window is employed for all time series inputs. We apply early stopping with a patience of 10 evaluation steps, and the model checkpoint with the highest validation accuracy is then used for the final

test evaluation. All experiments are conducted on a machine equipped with dual NVIDIA RTX 2080 Ti GPUs.

4.2 Comparative Study

To evaluate the effectiveness of our ViFusionTST model for bed-exit prediction, we benchmarked it against 10 state-of-the-art methods: eight 1D time-series models—iTransformer [Liu and others, 2023], TST [Zerveas *et al.*, 2021], DLinear [Zeng *et al.*, 2023], LightTS [Zhang *et al.*, 2022], TSMixer [Chen *et al.*, 2023], Temporal Convolutional Networks (TCN) [Bai *et al.*, 2018], TimesNet [Wu *et al.*, 2023], and Mamba [Gu and Dao, 2023]; two 2D time-series models—ViTST [Li *et al.*, 2024] and FIRTS [Costa *et al.*, 2024].

The comparative results are presented in Table. 1. The best results are in **red**, and the second-best results are underlined. Our ViFusionTST model outperforms previous baseline models on 4 out of 5 metrics. Specifically, our method achieves an F1 score of 79.4%, which is 6.0 pp higher than the second-best ViTST (73.4%), corresponding to an 8.2% relative improvement. Similarly, recall increases by 14.8 pp (21.8%), accuracy by 2.1 pp (2.4%), and AUPRC by 3.6 pp (4.3%). Meanwhile, our fusion-based method maintains high precision, preventing excessive false alerts and thus avoiding alarm fatigue among caregivers.

This consistent gain results from our cross-attention fusion module, which dynamically re-weights features from both line plots and RP/MTF/GAF representations, leading to more discriminative embeddings during short transition periods. Compared to FIRTS which relies solely on RP/MTF/GAF, and ViTST which processes only line plots, our multi-modal fusion approach substantially increases recall, thereby improving overall predictive performance. Moreover, both ViTST and our ViFusionTST—which incorporate line-plot representations as input—outperform the 1D baselines. We attribute this to the vision-based architectures’ better ability to capture long-term dependencies from extended irregular context windows in line plots—patterns that 1D models struggle to learn from raw time-series data.

4.3 Ablation Study

To quantify the individual contributions of our fusion strategy and the choice of input modalities, we carried out two ablation experiments. First, we compared our cross-attention fusion against a simple feature-concatenation baseline, and then evaluated models trained on different input representations.

Effect of Fusion Method

To isolate the impact of our cross-attention fusion, we compare against a feature-concatenation baseline that uses the identical backbone and training protocol as ViFusionTST. As shown in Table 2, replacing feature concatenation with our cross-attention fusion module yields large gains across all metrics. In particular, F1 score increases from 0.658 to 0.794 (+0.136, +20.7%), recall from 0.582 to 0.827 (+0.245, +42.1%), and AUPRC from 0.761 to 0.868 (+0.107, +14.1%). This confirms that our cross-attention module more effectively integrates complementary features than naive concatenation.

Dim.	Model	F1 Score ↑	Recall ↑	Precision ↑	Accuracy ↑	AUPRC ↑
1D	DLinear	0.584	0.519	0.667	0.802	0.686
	LightTS	0.718	0.678	0.764	0.858	0.801
	TSMixer	0.733	0.701	0.769	0.863	0.825
	TCN	0.688	0.738	0.644	0.821	0.721
	TimesNet	0.690	0.631	0.759	0.848	0.750
	Mamba	0.701	0.631	0.787	0.855	0.787
	iTransformer	0.730	0.707	0.754	0.860	0.807
	TST	0.716	0.672	0.766	0.857	0.809
2D	FIRTS	0.659	0.640	0.680	0.823	0.783
	ViTST	0.734	0.679	0.797	0.868	0.832
	ViFusionTST (Ours)	0.794	0.827	0.763	0.885	0.868

Table 1: Performance comparison of different models for bed-exit prediction.

Fusion Method	F1 Score ↑	Recall ↑	Precision ↑	Accuracy ↑	AUPRC ↑
Feature Concatenation	0.658	0.582	0.756	0.838	0.761
Cross Fusion (Ours)	0.794	0.827	0.763	0.885	0.868
% Improvement	+20.7%	+42.1%	+0.9%	+5.6%	+14.1%

Table 2: Evaluation of fusion methods.

Input Modality	F1 Score ↑	Recall ↑	Precision ↑	Accuracy ↑	AUPRC ↑
Line Plot	0.771	0.824	0.725	0.869	0.835
RP/MTF/GAF	0.659	0.640	0.680	0.823	0.732
Line Plot + RP/MTF/GAF (Ours)	0.794	0.827	0.763	0.885	0.868

Table 3: Impact of different input Modalities on prediction performance.

Effect of Input Modalities

Table 3 reports results for three input configurations: line plots only, RP/MTF/GAF only, and their fusion. The line-plot-only variant achieves an F1 of 0.771 and AUPRC of 0.835, outperforming the RP/MTF/GAF-only variant (F1 = 0.659, AUPRC = 0.732), which indicates that raw curves carry stronger temporal cues. When both modalities are fused, the model reaches F1 = 0.794 (+2.3 pp over line plot, +13.5 pp over RP/MTF/GAF). These results demonstrate that line plots and RP/MTF/GAF representations provide complementary information and that their joint modeling yields the best predictive performance.

Effect of Vision Encoders

To quantify the impact of different pretrained vision backbones on bed-exit prediction, we replace our Swin-Tiny encoders with a variety of alternatives, while keeping all other components fixed. Table 4 reports the performance for each encoder. The Swin-Tiny encoder achieves the best overall performance, surpassing all other encoders, including its larger variants. Among non-Swin architectures, the MambaVision delivers the next best results, with MambaVision-Tiny achieving F1 = 0.759 and AUPRC = 0.830. This suggests that replacing self-attention layers with Mamba blocks can also effectively capture temporal image features while reducing computational cost.

Notably, within each model family, the smaller variant consistently outperforms its larger counterpart (e.g., ViT-Base vs. ViT-Large). We attribute this to the relatively low complexity of temporal image representations compared to natural images. Larger models tend to overfit on these simpler patterns, leading to degraded generalization. Overall, these findings emphasize the importance of selecting appropriately sized encoders for time-series image inputs, since larger models are not necessarily better.

Prediction Visualization

Figure 4 presents a representative bed-exit event, where the load signal (blue) and the predicted bed-exit probability (green) are plotted over time. During the initial stable-in-bed phase, the patient remains motionless. Consequently, the load is flat and the model outputs a probability of 0, suppressing unnecessary alarms. As the patient transitions toward bed exit, subtle posture shifts introduce high-frequency load oscillations. These precursor movements cause the probability to rise steadily. The ward’s nurse call system employs a fixed alarm threshold of 0.5 (dashed line). The model first exceeds this threshold six minutes before the true exit time. This advance notice is clinically valuable: caring staff receive a single precise alert rather than a series of premature or fluctuating warnings. These observations demonstrate that the ViFusionTST model can identify preparatory movements

Vision Encoder	F1 Score \uparrow	Recall \uparrow	Precision \uparrow	Accuracy \uparrow	AUPRC \uparrow
Resnet-50 [He <i>et al.</i> , 2016]	0.727	0.663	0.804	0.866	0.817
MambaVision-Base [Hatamizadeh and Kautz, 2025]	0.731	0.665	0.812	0.869	0.813
MambaVision-Small [Hatamizadeh and Kautz, 2025]	0.729	0.655	0.824	0.870	0.821
MambaVision-Tiny [Hatamizadeh and Kautz, 2025]	0.759	0.702	0.825	0.880	0.830
SwinV2-Base [Liu <i>et al.</i> , 2022]	0.701	0.656	0.753	0.850	0.785
SwinV2-Small [Liu <i>et al.</i> , 2022]	0.734	0.815	0.668	0.842	0.782
ViT-Large [Dosovitskiy and others, 2021]	0.721	0.649	0.811	0.865	0.783
ViT-Base [Dosovitskiy and others, 2021]	0.731	0.690	0.778	0.864	0.822
Swin-Base [Liu <i>et al.</i> , 2021]	0.766	0.797	0.738	0.870	0.817
Swin-Small [Liu <i>et al.</i> , 2021]	0.775	0.783	0.767	0.878	0.822
Swin-Tiny [Liu <i>et al.</i>, 2021]	0.794	0.827	0.763	0.885	0.868

Table 4: Performance Comparison of Vision Encoders

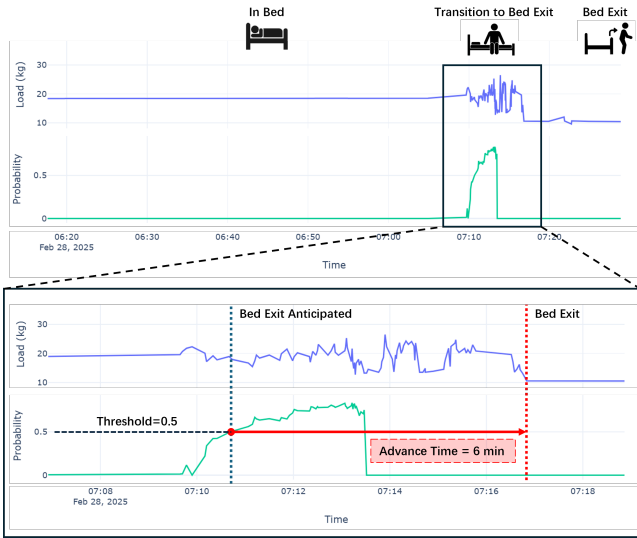


Figure 4: ViFusionTST prediction probability over time for a bed-exit event.

that simple load-drop heuristics cannot detect, while providing interpretable and actionable predictive results suitable for deployment in real world.

5 Conclusion

This study shows that a simple load-sensor installation, combined with state-of-the-art vision backbones, is enough to achieve high accuracy, real-time bed-exit prediction in a long-term care setting. We introduced ViFusionTST, a dual-stream Swin Transformer that separately encodes RGB line-plot images and three texture maps (RP, MTF, GAF), then unifies them with cross-attention to learn data-driven modality weights. Evaluated on six months of data from 95 beds, ViFusionTST achieved an accuracy of 0.885 and an F1 score of 0.794, outperforming recent 1D and 2D time-series baselines across all major metrics, establishing a new benchmark for load-sensor-based fall prevention.

Our work contributes (i) a scalable, privacy-preserving sensing pipeline that integrates seamlessly with existing nurse-call systems, (ii) a large real-world benchmark for bed load signal modeling, and (iii) a deep fusion architecture that can be adapted to other regular or irregular time series. Collectively, these advances move image-based time-series modeling closer to real-world deployment and offer a practical path toward reducing fall-related injuries in hospitals and long-term care facilities.

References

- [Awais *et al.*, 2019] Muhammad Awais, Mohsin Raza, Kamran Ali, Zulfiqar Ali, Muhammad Irfan, Omer Chughtai, Imran Khan, Sunghwan Kim, and Masood Ur Rehman. An internet of things based bed-egress alerting paradigm using wearable sensors in elderly care environment. *Sensors*, 19(11):2498, 2019.
- [Bai *et al.*, 2018] S. Bai, J. Z. Kolter, and V. Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling, Apr 2018. arXiv:1803.01271.
- [Bai *et al.*, 2023] Dorothy Bai, Mu-Chieh Ho, Bhukumuzi M Mathunjwa, and Yeh-Liang Hsu. Deriving multiple-layer information from a motion-sensing mattress for precision care. *Sensors*, 23(3):1736, 2023.
- [Barra *et al.*, 2020] Silvio Barra, Salvatore Mario Carta, Andrea Corrigan, Alessandro Sebastian Podda, and Diego Reforgiato Recupero. Deep learning and time series-to-image encoding for financial forecasting. *IEEE/CAA Journal of Automatica Sinica*, 7(3):683–692, 2020.
- [Bisogni *et al.*, 2022] Carmen Bisogni, Aniello Castiglione, Sanoar Hossain, Fabio Narducci, and Saiyed Umer. Impact of deep learning approaches on facial expression recognition in healthcare industries. *IEEE Transactions on Industrial Informatics*, 18(8):5619–5627, 2022.
- [Bu *et al.*, 2021] F. Bu, Q. Lin, J. Allebach, Q. Lin, and J. Allebach. Bed exit detection network (BED Net) for patients bed-exit monitoring based on color camera images. *Electron. Imaging*, 33:1–8, Jan 2021.

- [Capezuti *et al.*, 2009] Elizabeth Capezuti, Barbara L Brush, Stephen Lane, Hannah U Rabinowitz, and Michelle Secic. Bed-exit alarm effectiveness. *Archives of gerontology and geriatrics*, 49(1):27–31, 2009.
- [Chang *et al.*, 2021] W.-J. Chang, L.-B. Chen, Y.-K. Ou, A. R. See, and T.-C. Yang. A bed-exit and bedside fall warning system based on deep learning edge computing techniques. In *2021 IEEE 3rd Global Conference on Life Sciences and Technologies (LifeTech)*, pages 142–143, Mar 2021.
- [Chen *et al.*, 2018] T.-X. Chen, R.-S. Hsiao, C.-H. Kao, D.-B. Lin, and B.-R. Yang. Bed-exit prediction based on 3D convolutional neural network. In *2018 IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation (SmartWorld/SCALCOM/UIC/ATC/CBDCom/IOP/SCI)*, pages 1185–1188, Oct 2018.
- [Chen *et al.*, 2023] S.-A. Chen, C.-L. Li, N. Yoder, S. O. Arik, and T. Pfister. TSMixer: An all-MLP architecture for time series forecasting, Sep 2023. arXiv:2303.06053.
- [Chen *et al.*, 2024] Mouxiang Chen, Lefei Shen, Zhuo Li, Xiaoyun Joy Wang, Jianling Sun, and Chenghao Liu. Visions: Visual masked autoencoders are free-lunch zero-shot time series forecasters. *arXiv preprint arXiv:2408.17253*, 2024.
- [Costa *et al.*, 2024] Henrique V Costa, André GR Ribeiro, and Vinicius MA Souza. Fusion of image representations for time series classification with deep learning. In *International Conference on Artificial Neural Networks*, pages 235–250. Springer, 2024.
- [Dosovitskiy and others, 2021] A. Dosovitskiy et al. An image is worth 16x16 words: Transformers for image recognition at scale, Jun 2021. arXiv:2010.11929.
- [Eckmann *et al.*, 1995] J-P Eckmann, S Oliffson Kamphorst, and David Ruelle. Recurrence plots of dynamical systems. In *Turbulence, Strange Attractors and Chaos*, pages 441–445. World Scientific, 1995.
- [Gu and Dao, 2023] A. Gu and T. Dao. Mamba: Linear-time sequence modeling with selective state spaces, Dec 2023. arXiv:2312.00752.
- [Hao *et al.*, 2014] Jun Hao, Xiaoxiao Dai, Amy Stroder, Jun Jason Zhang, Bradley Davidson, Mohammad Mahoor, and Neil McClure. Prediction of a bed-exit motion: Multi-modal sensing approach and incorporation of biomechanical knowledge. In *2014 48th Asilomar Conference on Signals, Systems and Computers*, pages 1747–1751. IEEE, 2014.
- [Hatamizadeh and Kautz, 2025] Ali Hatamizadeh and Jan Kautz. Mambavision: A hybrid mamba-transformer vision backbone. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 25261–25270, 2025.
- [He *et al.*, 2016] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- [Hitcho *et al.*, 2004] Eileen B Hitcho, Melissa J Krauss, Stanley Birge, William Claiborne Dunagan, Irene Fischer, Shirley Johnson, Patricia A Nast, Eileen Costantinou, and Victoria J Fraser. Characteristics and circumstances of falls in a hospital setting: a prospective analysis. *Journal of general internal medicine*, 19(7):732–739, 2004.
- [Inoue and Taguchi, 2020] M. Inoue and R. Taguchi. Bed exit action detection based on patient posture with long short-term memory. In *2020 42nd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 4390–4393, Jul 2020.
- [Jin and others, 2024] M. Jin et al. Time-LLM: Time series forecasting by reprogramming large language models, Aug 2024. arXiv: 2310.01728.
- [Jin *et al.*, 2023] Shuzu Jin, Soumya Mohanty, Qunying Xie, Hanzhi Wang, and Xue-Hao Zhang. Classification of time series as images using deep convolutional neural networks: application to glitches in gravitational wave data. In *5th International Conference on Advances in Signal Processing and Artificial Intelligence (ASPAI’2023)*, 2023.
- [Karami *et al.*, 2024] Hojjat Karami, Mary-Anne Hartley, David Atienza, and Anisoara Ionescu. Timehr: Image-based time series generation for electronic health records. *arXiv preprint arXiv:2402.06318*, 2024.
- [Kuwahara and Wada, 2017] Noriaki Kuwahara and Kengo Wada. Bed-leaving prediction using a sheet-type pressure-sensitive sensor base with deep-learning. *Journal of Fiber Science and Technology*, 73(12):343–347, 2017.
- [LeLaurin and Shorr, 2019] Jennifer H LeLaurin and Ronald I Shorr. Preventing falls in hospitalized patients: state of the science. *Clinics in geriatric medicine*, 35(2):273–283, 2019.
- [Li *et al.*, 2024] Zekun Li, Shiyang Li, and Xifeng Yan. Time series as images: Vision transformer for irregularly sampled time series. *Advances in Neural Information Processing Systems*, 36, 2024.
- [Lin *et al.*, 2022] C.-J. Lin, T.-S. Wei, P.-T. Liu, B.-H. Chen, and C.-H. Shih. Bed-exit behavior recognition for real-time images within limited range. *Sensors*, 22(15), Jan 2022.
- [Liu and others, 2023] Y. Liu et al. iTransformer: Inverted transformers are effective for time series forecasting, Oct 2023. arXiv:2310.06625.
- [Liu *et al.*, 2021] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10012–10022, 2021.
- [Liu *et al.*, 2022] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng

- Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12009–12019, 2022.
- [Meng *et al.*, 2024] Fanchao Meng, Teng Liu, Chuizhou Meng, Jianjun Zhang, Yifan Zhang, and Shijie Guo. Method of bed exit intention based on the internal pressure features in array air spring mattress. *Scientific Reports*, 14(1):27273, 2024.
- [Ni *et al.*, 2025] Jingchao Ni, Ziming Zhao, ChengAo Shen, Hanghang Tong, Dongjin Song, Wei Cheng, Dongsheng Luo, and Haifeng Chen. Harnessing vision models for time series analysis: A survey. *arXiv preprint arXiv:2502.08869*, 2025.
- [Rose *et al.*, 2020] Georgie Rose, Veerle Decalf, Karel Everaert, and Wendy F Bower. Toileting-related falls at night in hospitalised patients: The role of nocturia. *Australasian journal on ageing*, 39(1):e70–e76, 2020.
- [Seow *et al.*, 2022] Jason Phil Seow, Tse Lert Chua, Fazila Aloweni, Shu Hui Lim, and Shin Yuh Ang. Effectiveness of an integrated three-mode bed exit alarm system in reducing inpatient falls within an acute care setting. *Japan Journal of Nursing Science*, 19(1):e12446, 2022.
- [Valero *et al.*, 2021] Maria Valero, Jose Clemente, Fangyu Li, and WenZhan Song. Health and sleep nursing assistant for real-time, contactless, and non-invasive monitoring. *Pervasive and mobile computing*, 75:101422, 2021.
- [Vaswani and others, 2017] A. Vaswani et al. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.
- [Wang and Oates, 2015] Zhiguang Wang and Tim Oates. Imaging time-series to improve classification and imputation. *arXiv preprint arXiv:1506.00327*, 2015.
- [Wang *et al.*, 2015] Zhiguang Wang, Tim Oates, et al. Encoding time series as images for visual inspection and classification using tiled convolutional neural networks. In *Workshops at the twenty-ninth AAAI conference on artificial intelligence*, volume 1, pages 1–7. Austin, TX, 2015.
- [Wen and others, 2023] Q. Wen et al. Transformers in time series: A survey, May 2023. arXiv:2202.07125.
- [Wu *et al.*, 2023] Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. Timesnet: Temporal 2d-variation modeling for general time series analysis. In *International Conference on Learning Representations*, 2023.
- [Zeng *et al.*, 2023] Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 11121–11128, 2023.
- [Zerveas *et al.*, 2021] George Zerveas, Srideepika Jayaraman, Dhaval Patel, Anuradha Bhamidipaty, and Carsten Eickhoff. A transformer-based framework for multivariate time series representation learning. In *Proceedings of the 27th ACM SIGKDD conference on knowledge discovery & data mining*, pages 2114–2124, 2021.
- [Zhang *et al.*, 2022] Tianping Zhang, Yizhuo Zhang, Wei Cao, Jiang Bian, Xiaohan Yi, Shun Zheng, and Jian Li. Less is more: Fast multivariate time series forecasting with light sampling-oriented mlp structures, 2022.