Sensors and Materials, Vol. 38, No. 9 (2026) 5243–5258
MYU Tokyo

S & M 4621

Applying Transformers and Knowledge Graphs to Home Care Event Detection

Jui-Feng Yeh,[1*] Chun-Yu Tsai,[1] Bo-Sung Huang,[1] Po-Hsiang Yang,[1] and Chen-Yu Yeh[2]

[1]Department of Computer Science and Information Engineering,
National Chiayi University, No. 300, Xuefu Rd., Chiayi City 600355, Taiwan (R.O.C.)
[2]China Medical University Hospital, Taichung City, Taiwan (R.O.C.)

(Received June 9, 2026; accepted September 9, 2026)

Keywords: home care, transformer, knowledge graph, privacy protection, virtual avatar, digital twin

To address the home-care problems resulting from an increasing number of elderly people living alone and to prevent related accidents, we develop a multimodal event detection system that combines knowledge graphs and transformer architecture for monitoring elderly people by sensor technologies. This system focuses on analyzing behavioral events of older adults in home care settings. In consideration of privacy, a digital twin is adopted in the monitoring video, replacing the original content with virtual avatars. There are mainly two phrases in the proposed system framework. First, the proposed system utilizes the Knowledge-Enabled Language Representation Model enabling language representation with knowledge graph as its core, integrating semantic features from knowledge graphs, and enhancing language understanding through the self-attention mechanism of the transformer. Secondly, the Bootstrapping Language-Image Pre-training model is used for video narration generation, combined with the GroundingDINO model for object detection to enable the narration to correspond to objects in the actual scene. Finally, a text report is output and presented in a visual player. It can also be used as input for AI models to analyze behavioral events in virtual home care spaces. The experimental results indicate that after introducing generative AI, the recall rate for running events increased significantly to 83.3%. These results confirm that the system can effectively improve the accuracy of hazard detection in home care.

1. Introduction

Home care is a broad term that refers to professional support services that allow a person to live safely and independently in their home. Instead of medical/clinical care, home care plays an essential role and focuses on activities of daily living, especially for elderly people. Recently, as numerous nations across Asia and the western world transition into superaged societies, the challenges of providing sustainable home care for the elderly have reached a critical juncture. Isolation significantly escalates the risk of undetected domestic accidents such as falls, disorientation, and hazardous climbing. Therefore, AI approaches combined with sensors are

*Corresponding author: e-mail: ralph@mail.ncyu.edu.tw
https://doi.org/10.18494/SAM6478

ISSN 0914-4935
https://myukk.org/

adapted to monitoring functionalities and achieve good performance. Unlike traditional systems that might simply send raw surveillance clips to family members when an anomaly occurs, our proposed system strictly avoids raw RGB video leakage. The decision to output text reports and simulated alerts rather than directly transmitting raw surveillance clips is driven by the strict need for privacy protection. Transmitting raw video from domestic spaces raises severe privacy concerns; therefore, the system acts as an indispensable privacy-preserving filter, transforming sensitive visual data into secure semantic context, which fully justifies the additional computational processing costs. Consequently, this integrated approach represents a promising potential solution for future applications.

One of the essential features of home care systems for the elderly mainly relies on action recognition technology by sensing video content. Simonyan and Zisserman explored determining an individual's movement type by analyzing changes in skeletal structure in images.[1] However, single-modal visual recognition methods still face three major limitations. The first one is a lack of understanding of scene context and intent. Then, surveillance footage from traditional real-world settings often raises serious privacy concerns, and the cost of collecting real-world data on daily activities in physical spaces is extremely high. Finally, the system struggles to perform multistep logical reasoning on complex events.

To address these challenges, we develop a multimodal event detection system that is based on a transformer combined with a knowledge graph. Considering privacy concerns, the proposed system uses the VirtualHome knowledge graph to process data on daily activities in virtual spaces, thereby enabling inference and prediction.[2] In terms of visual understanding, the system uses the Bootstrapping Language-Image Pre-training (BLIP) model to generate video descriptions, converting visual content into natural language and perform open-set object detection using the GroundingDINO model.[3,4] The GroundingDINO model anchors textual descriptions to specific pixel regions, effectively filtering out potential hallucinations generated by the model and providing a reliable foundation for visual verification. Finally, the system uses the Knowledge-Enabled Language Representation Model (K-BERT) as its core, incorporating related concepts from the knowledge graph into the language model. By leveraging the transformer's self-attention mechanism to fuse external knowledge with textual semantics, it achieves deep behavioral inference.[5]

The main contributions of our proposed approach are summarized as follows.

(1) To address safety concerns regarding elderly individuals living alone in home care settings, we propose a multimodal event detection system that integrates video and semantic analyses. By using the GroundingDINO model to detect relationships between people and objects, and the BLIP model to generate video descriptions, the system can understand the semantic context of a scene and identify potentially dangerous behaviors.
(2) The K-BERT model is combined with knowledge graphs for semantic inference, and the transformer self-attention mechanism is used to enhance event understanding. The association between image and text information is established to achieve more accurate behavioral semantic judgment and risk reasoning.
(3) Generative AI is introduced as the decision optimization layer of the system to perform the

secondary verification and re-ranking of the judgment results of the basic model. This method effectively corrects the misclassification of complex actions (such as climbing and running) and significantly improves the semantic recognition accuracy of the system in real-world scenarios while ensuring a high risk detection rate.

2. Related Work

Herein, we present an in-depth literature review focusing on core areas including home care, privacy protection, digital twins, knowledge graphs, and multimodal vision-language models.

As the demand for aging-in-place solutions accelerates, home care systems have evolved from passive alarm buttons to active intelligent monitoring. However, current systems face a severe privacy-security paradox. To accurately detect falls or abnormal behaviors, systems typically require high-resolution real-time footage, which directly infringes on the privacy of seniors in sensitive areas such as bedrooms and bathrooms. In 2024, a study by Tian *et al.*[6] highlighted that privacy concerns are the primary reason older adults refuse to adopt smart home health technologies, preventing many innovations from practical deployment.

To mitigate this conflict, recent research has shifted from improving image quality to privacy-preserving computing. Recently, Song *et al.*[7] have proposed a real-time behavior recognition system based on edge-cloud collaboration, emphasizing that feature extraction should occur locally to prevent raw video leakage. Similarly, Bettini *et al.*[8] demonstrated through case studies that activity monitoring is only acceptable to users when built upon a rigorous personal data protection framework. Building on this concept of data minimization, Wang *et al.*[9] proposed a privacy-preserving multimodal fall detection system (P2MFDS) that utilizes nonvisual sensing technologies, such as millimeter-wave radar and 3D vibration sensing, to monitor high-risk areas without relying on raw RGB videos. Memon *et al.*[10] proposed a generalized and improved human activity recognition (HAR) model. Their work demonstrated that inertial sensor data from smartphones or wearable devices can accurately identify activities of daily living without the need for cameras, thereby effectively circumventing the privacy infringements associated with visual surveillance. Chang and Lin[11] developed a real-time fall detection system using the AlphaPose model. By extracting human pose keypoints from video sequences, their system intelligently identifies abnormal fall movements to secure high-risk event detection without compromising processing efficiency.

Digital twins are digital simulations of the real world created in the digital world. These simulations can map real-world conditions onto the digital environment through sensors and networks. The application of digital twins in healthcare has become an important research trend. Xames and Topcu[12] analyzed in detail the research trends, technological gaps, and implementation challenges of digital twins in healthcare systems. Momand *et al.*[13] proposed a more specific end-to-end framework for establishing digital twins in elderly care, demonstrating how to integrate multisource physiological data collection to build predictive models. Shankhdhar and Garg[14] proposed a blockchain-enabled secure data transmission framework for

personalized e-healthcare, highlighting the importance of data integrity and privacy in human digital twin well-being systems. Recent advancements have focused on integrating smart home technologies with digital twins to support independent living. Wang *et al.* [15] proposed an intelligent and privacy-preserving digital twin model for aging-in-place, which utilizes unobtrusive sensors to continuously monitor residents' activities and environmental conditions without compromising their privacy. Expanding on the role of smart environments, Adibi *et al.*[16] conducted a comprehensive analysis of sensor-enabled digital twins, emphasizing their transformative potential to enhance real-time health monitoring, optimize resource allocation, and deliver personalized interventions in smart homes and hospitals. In addition to technical implementations, the broader societal implications of this technology are gaining attention. Lehmann *et al.*[17] explored the use of digital twins and cyber-physical systems for virtual coaching to support healthy aging while critically addressing the ethical issues and technological developments required to safely deploy these systems for older adults.

A knowledge graph is a knowledge base that structurally presents real-world concepts and their relationships through nodes (entities) and edges (relationships), supporting accurate semantic search and deep knowledge acquisition. Recently, its effectiveness across various domains has been demonstrated in several studies. Li[18] demonstrated that specialized knowledge representation strategies are essential for reducing hallucinations in retrieval-augmented domain-specific systems. Zhao *et al.*[19] showed that integrating event knowledge graphs (EKGs) into visual language navigation models enhances the understanding of coarse-grained instructions and improves multistep task planning capabilities. Chen *et al.*[20] proposed a knowledge-graph-based framework for smart homes to model complex device interactions and improve automated recommendations on the basis of user habits. Wang *et al.*[21] integrated large language models with medical knowledge graphs to transform fragmented clinical information into interconnected networks, significantly improving semantic reasoning and diagnostic accuracy. Zhu *et al.*[22] constructed global event knowledge graphs to identify causal links and sequence patterns across continuous video streams for advanced spatiotemporal pattern reasoning.

Multimodal vision–language models (VLMs) have significantly advanced the ability to interpret complex visual scenes by bridging the gap between raw visual perception and high-level semantic reasoning. Recent studies have demonstrated their profound impact on video understanding and anomalous event detection. For instance, Huang *et al.*[23] introduced the Vad-R1 framework, a multimodal large language model that explicitly employs a perception-to-cognition chain of thought to perform step-by-step reasoning on complex video anomalies. Lin *et al.*[24] developed Video-LLaVA, a unified visual representation framework that aligns spatial and temporal features to enhance the sequence reasoning capabilities of large language models. Wang *et al.*[25] proposed a multigrained video-language pretraining method that effectively maps dynamic visual changes to textual descriptions, improving the comprehension of continuous, multistep events. To structurally organize this multimodal understanding, Egami *et al.*[26] proposed the virtual home action knowledge graph (VHAKG) method, which deeply integrates simulated multiview video data with event knowledge graphs to continuously record action sequences and environmental changes, effectively linking visual streams with structured semantic knowledge.

3. Proposed Method

3.1 System architecture

The core of the proposed system is to deeply concatenate visual perception, entity grounding, structured reasoning, and temporal context re-ranking, as shown in Fig. 1. The system framework mainly is composed of three modules: the video analysis module for frame sampling on input continuous home monitoring videos, the BLIP decoder for natural language description generation, and GroundingDINO for the synchronous extraction of local object features and bounding boxes. That is, the video analysis module identifies what is seen. The knowledge reasoning module based on knowledge graph schema acts as the central nervous system. Finally, the texts and labels output by the video analysis module are input into the knowledge graph constructor to convert unstructured data into JavaScript object notation (JSON)- and resource description framework (RDF)-formatted event entities. Subsequently, these entities are combined with the predefined VirtualHome safety knowledge graph and fed into the K-BERT core through a knowledge injection algorithm for self-attention computation to determine potential risks. The BLIP decoder answers the risk determination question "What does this mean?" Finally, generative AI re-ranking and user interface are used to smooth out K-BERT's rigid classification of complex continuous actions, and the system configures a Gemini 2.5 Flash generative model at the backend to perform secondary inference re-ranking on ambiguous predictions. The final risk assessment reports and real-time alerts are integrated and presented on a developed web-based monitoring interface for user reference.

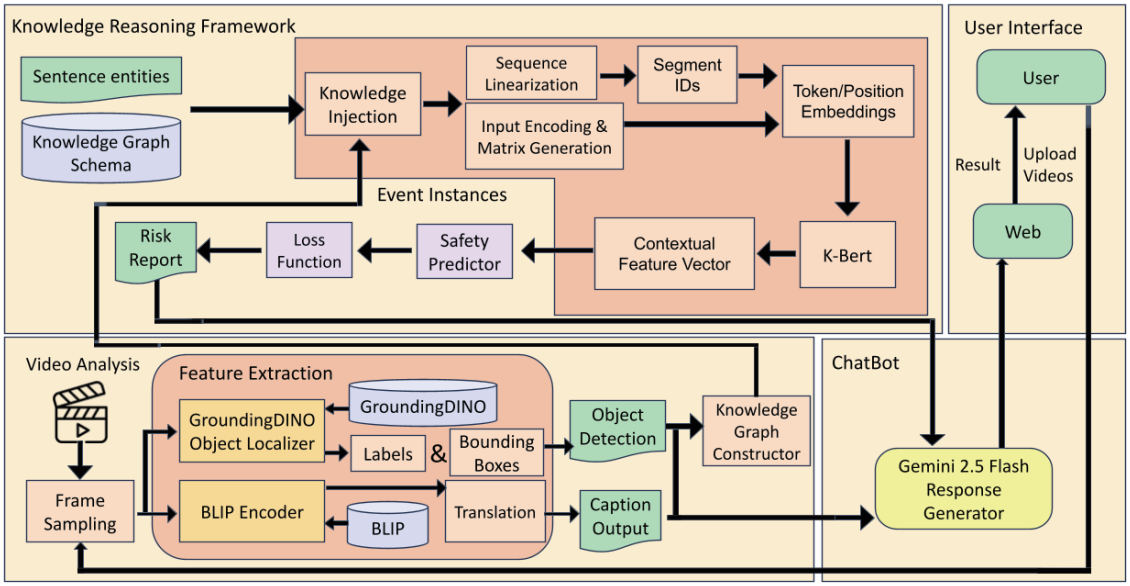

Fig. 1. (Color online) Proposed system framework based on transformer combined with knowledge graph.

3.2 BLIP-based vision language transformation

This system employs BLIP as the primary semantic translator. Compared to traditional models that separate the training of encoders and decoders, BLIP provided the unified architecture shown in Fig. 2. When processing input frames, BLIP's image encoder adopts the vision transformer (ViT) architecture, segmenting the image into multiple patches and extracting global visual features. During the natural language generation phase, BLIP activates its image-grounded text decoder mode. Internally, the decoder replaces standard bi-directional self-attention layers with causal self-attention layers, ensuring that text generation strictly follows an autoregressive sequence from left to right. In the cross-attention layer, the decoder continuously queries the visual features output by the image encoder, thereby ensuring that every generated vocabulary word tightly matches the visual content. During pretraining, BLIP maximizes the likelihood probability of generating the correct text given an image through language modeling loss (LM loss). Its precise caption output serves as the core material for this system to understand the behavioral outlines of elderly individuals.

3.3 Object localization using GroundingDINO

To eliminate the hallucination defects caused by the language model's over-association in BLIP and to resolve its occasional subject blindness, the system introduces GroundingDINO in

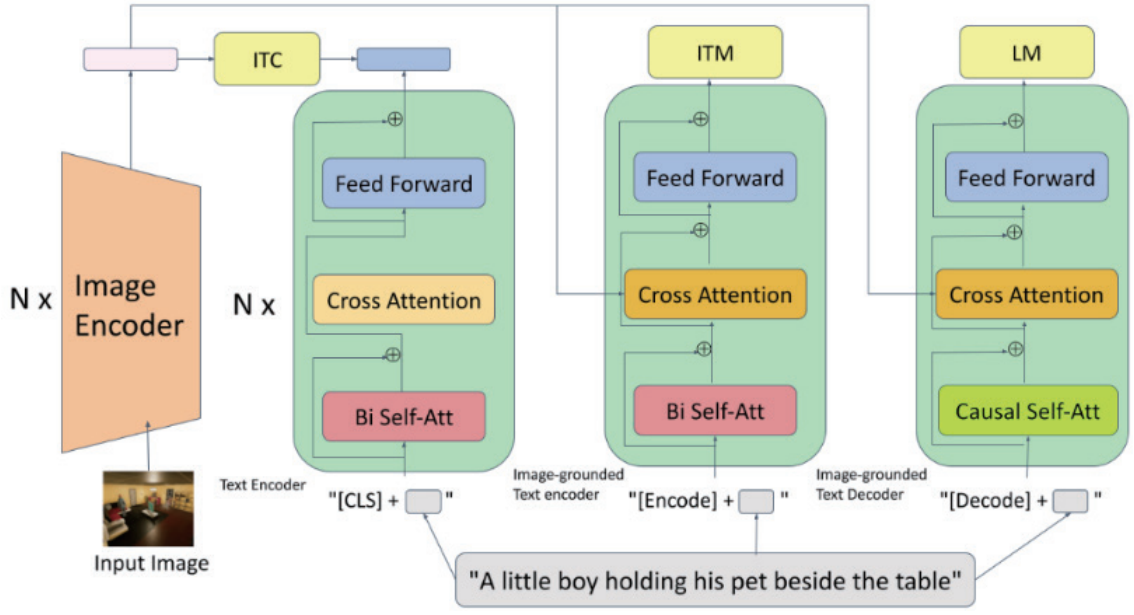

Fig. 2. (Color online) BLIP diagram for vision language transformation.

parallel for visual grounding verification. GroundingDINO's operational mechanism deeply fuses both language and vision modalities. The system performs tokenization on the descriptive text generated by BLIP and feed it into GroundingDINO's text backbone, while the image is input into its image backbone. The crucial feature enhancer and language-guided query selection mechanisms calculate the cross-modal similarity between textual and visual features. Through a cross-modality decoder, the model directly searches the image for targets described by the text. For example, when the text prompt includes person and table, GroundingDINO can accurately output the bounding boxes and confidence scores of these entities without needing to be retrained for this specific environment. This means zero-shot. If BLIP describes an object but GroundingDINO's confidence score falls below a preset threshold and fails to output a bounding box, the object will be marked as a hallucination and filtered out, ensuring that information entering the knowledge graph possesses absolute physical reality.

3.4 Knowledge-enhanced reasoning using K-BERT

One of the core innovations lies in utilizing K-BERT for deep semantic and commonsense reasoning. First, knowledge injection and sentence tree generation are developed to scan the entities in the input sequence and then automatically use the entities to query the VirtualHome knowledge graph to extract relevant triplet knowledge. These knowledge nodes are directly attached to the entities in the original input, forming a sentence tree with a multibranch structure. Secondly, sequence linearization and soft-position embedding are formed and fed into transformers. Herein, K-BERT is used to flatten the sentence tree as a linear sequence. To prevent the loss of the original semantic hierarchy after flattening, we introduce soft-position embedding in K-BERT. Unlike hard-position indexes that simply mark the absolute position of a sequence, soft-position indexes allow knowledge nodes on branches to share the same relative position weights as the entities in the main sentence. This ensures that regardless of how much background knowledge is injected, the chronological logic of the original action will not be distorted. Finally, a mask transformer and a visible matrix are used to resolve the knowledge noise issue brought by massive domain knowledge. Moreover, K-BERT constructs a visible matrix (M) that specifies whether any two tokens in the sequence possess a logical connection as visible or invisible. In the neural network's calculations, the mask-transformer computation at layer i is as follows. The system first computes the Query, Key, and Value vectors via trainable weight matrices. Equation (1) describes the generation of Query (Q), Key (K), and Value (V) matrices from the hidden state h^i:

$$Q^{i+1},\ K^{i+1},\ V^{i+1} = h^i W_q,\ h^i W_k,\ h^i W_v, \tag{1}$$

where W_q, W_k, and W_v are trainable model parameters, h^i is the output vector of the i-th mask transformer, and d_k is used to scale the attention weights. M is the visible matrix, which is used to zero out the self-attention weights between invisible tokens. For the self-attention score (S^{i+1}), the visible matrix M is directly superimposed onto the scaled dot-product result as shown below.

$$S^{i+1} = \text{softmax}\left(\frac{Q^{i+1}K^{i+1^\top}}{\sqrt{d_k}} + M\right) \tag{2}$$

Finally, the contextual feature vector is calculated as

$$h^{i+1} = S^{i+1}V^{i+1}. \tag{3}$$

For the unconnected tokens in the sentence tree, the corresponding value in matrix M is an extremely small negative number. After softmax(·), the attention weight will approach zero. K-BERT can further leverage the implicit safety connections in the knowledge graph while avoiding mutual interference from irrelevant knowledge, ultimately outputting the precise risk label for the event.

3.5 Generative AI re-ranking

Although K-BERT possesses strong graph constraints, when handling actions with highly overlapping physical features, such as climbing and running, the rigidity of graph rules often leads to conservative misjudgments. Therefore, after K-BERT outputs its initial judgment, the system incorporates a re-ranking layer based on the Gemini 2.5 Flash large language model. The system combines a continuous JSON state sequence including BLIP descriptions, GroundingDINO coordinates, and K-BERT's initial labels within a certain time window into a high-density prompt. Gemini uses its massive commonsense memory bank to analyze the rationality of continuous actions. For instance, if an elderly person's vertical height drops, but the previous second's state showed them approaching a chair, Gemini can infer that this is a continuous intention to sit down or attempt to climb, rather than an instantaneous out-of-control fall. Through context-level inference re-ranking by the LLM, the system manages to maintain high sensitivity while significantly correcting semantic recognition accuracy.

4. Experiment

4.1 Dataset preparation and evaluation metrics

To evaluate the proposed event detection system, we curated a comprehensive dataset encompassing 1150 video sequences. These sequences are categorized into five core behavioral events critical for home care monitoring: normal, fall, disoriented, climb, and running. To ensure both the structural reasoning reliability and real-world generalizability of the system, the dataset was further partitioned into three functional subsets, as shown in Table 1.

Table 1
Data distribution and utilization.

	Data type	Description	Amount of data
Videos	MP4	Multiperspective virtual scenario video	3530
Episodes	JSON	List of all scheduled activities for the whole day	70
Knowledge graph	RDF (TTL)	RDF definitions of environment, objects, and event states	710

The test set serves as a standardized baseline, featuring clear semantic boundaries and minimal visual noise within the digital twin environment. In contrast, the upload set introduces complex environmental variables typical of actual home settings, such as partial occlusions and varying lighting conditions, to critically assess the system's re-ranking performance.

To evaluate the performance of the hybrid architecture in home care scenarios comprehensively and objectively, two failure scenario metrics are used for the visual modules and four standard machine learning metrics for end-to-end event classification. The visual subset was utilized to determine the subject blindness rate (*SBR*) and hallucination rate (HR) between different captioning models to ensure stable input for the knowledge graph. Since the core of home-care event is the person, *SBR* measures how often the model completely omits the description of the entity when a person is clearly in the frame as the follows.

$$SBR = \frac{Number\ of\ \ frames\ missing\ \ person}{Total\ number\ of\ \ frames\ containing\ \ person} \tag{4}$$

HR is used to measure the frequency at which generative models fabricate nonexistent objects. A high hallucination rate severely pollutes the knowledge graph, leading to erroneous inferences. The calculation formula is shown as Eq. (5).

$$HR = \frac{Number\ of\ \ frames\ containing\ hallucinated\ \ object}{Total\ number\ of\ \ all\ \ frames} \tag{5}$$

Additionally, the metrics including accuracy, precision, recall, and F1-score are used to evaluate the performance of the proposed approach.

4.2 Evaluations about captioning model

To ensure that the most stable natural language descriptions are fed into the knowledge graph in the first phase of the experiment, an A/B test is conducted on the lightweight BLIP and BLIP-2 (2.7B) models. The evaluation is based on *SBR*.

Table 2 reveals a crucial insight regarding deploying large models in hardware-constrained environments. Although BLIP-2 is architecturally advanced, under the lightweight inference environment of this study, its *SBR* reached a staggering 69.28%. This means that in nearly 70% of the frames, BLIP-2 completely ignored the most important entity in the scene—the human being cared for. The in-depth analysis of its mechanisms shows that large parametric models, constrained by hardware computing power or quantization compression, tend to capture background elements with the largest area and most obvious statistical features, thereby overlooking elderly individuals who might be partially obscured by a sofa or table. In contrast, the BLIP model has a higher segment accuracy (67.63%) with the visual hallucinations that generate non-existent objects. Evaluated from a systems engineering perspective, the loss of subject information is an irreversible fatal error, whereas superfluous hallucinated objects can be filtered out through backend cross-validation mechanisms. Therefore, to maximize the preservation of subject information, BLIP is selected ultimately as the foundational visual-text translator.

4.3 Visual grounding architecture validation

To address the BLIP hallucination issue confirmed in the previous stage and compensate for occasional subject omission, in this experiment, GroundingDINO's filtering and remedial capabilities are validated within the hybrid architecture. The system inputs entity labels erroneously generated by BLIP into GroundingDINO for physical localization testing.

The results in Table 3 clearly confirm the absolute necessity of integrating zero-shot visual grounding. When BLIP is affected by language statistical bias, it fabricates descriptions, such as that of a "dog" in an empty room out of thin air, and GroundingDINO attempts to anchor this "dog" text prompt into specific pixel spaces. Because of the lack of corresponding visual features, GroundingDINO's recall rate is zero percent, and the system immediately blocks this

Table 2
Captioning model selection results.

Model	Test video count	Total number of samples	Segment accuracy	Subject blindness	Key failure mode
BLIP	50	1201	67.63%	36.14%	Hallucination
BLIP-2	50	1201	28.22%	**69.28%**	**Subject blindness**

Table 3
GroundingDINO object localization validation results.

Failure case	Total number of samples	Prompt example	GroundingDINO test result	Result analysis
BLIP hallucination	1201	"Dog", "wallet" Not in video	Object Not Detected (0% Recall)	**Filtered successfully**
BLIP-2 subject blindness	1201	Person Exists in video	Object Detected (98% Recall)	Corrected successfully

false information from entering the knowledge graph. Conversely, when directly inputting "person" as a prompt, GroundingDINO demonstrates a 98% precise localization capability, vastly compensating for the deficiencies of relying purely on generative text. This not only validates GroundingDINO as a powerful object detector but also establishes its core position as the system's "physical spatial safety net," ensuring that all inferences are established upon visual reality and are visually grounded.

4.4 System performance using K-BERT

After establishing the high reliability of visual inputs, in this stage, the complete system integrated with K-BERT is tested. The system aims to classify elderly behaviors in standardized scenarios (Test Set) and complex real-world scenarios (Upload Set) as Normal or Anomaly/Risk that includes fall, disoriented action, climbing, and running.

In Table 4, K-BERT achieves near-perfect performance across all metrics when processing the built-in standardized dataset (Test Set). More crucially, when faced with unpredictable real-world scenarios (Upload Set), K-BERT demonstrates an extreme 100% recall rate for the most fatal risk, "fall". This proves that K-BERT, through the safety knowledge loaded via the sentence tree and visible matrix, possesses an extremely robust risk interception capability, ensuring that the system never misses a single falling incident. However, the data also reveals the limitations of deterministic rule-based reasoning. In real-world scenarios, the recall rate for "climb" drops sharply to 16.7% and "running" also suffers classification confusion. The fundamental reason is that when a person attempts to climb a chair, their instantaneous vertical displacement and physical contortion are physically highly similar to a "fall". Bound by strict safety knowledge constraints, K-BERT adopts highly conservative decision logic, directly classifying ambiguous climbing actions as high-risk "falls". While this mechanism maximizes safety, it markedly reduces the precision of falls (dropping to 50.0%). In clinical applications, frequently misreporting climbing as falling will severely exacerbate the invalid alarm problem of care systems.

Table 4
System performance with K-BERT.

Event case	Total number of samples (test/upload)	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Normal behavior	200/30	99.5/91.3	99.0/75.8	100/83.3	99.5/79.4
Fall	200/30	99.0/80.0	98.5/50.0	99.5/100	99.0/66.7
Disoriented action	200/30	98.5/90.7	98.0/78.6	99.0/73.3	98.5/75.9
Climbing	200/30	98.0/83.3	0.0/100	**0.0/16.7**	0.0/28.6
Running	200/30	98.5/83.3	98.5/100	98.0/80.0	98.2/88.9

4.5 Enhancing inference via Generative AI re-ranking

To resolve K-BERT's semantic rigidity when facing complex actions, we further introduce Gemini 2.5 Flash for secondary inference re-ranking in the experiment. Generative AI will acquire event sequences within a continuous time window, thereby evaluating the temporal logic of the actions.

As shown in Table 5, the intervention of the large language model brings about a crucial semantic breakthrough. In real-world scenarios, the recall rate for "running" steadily improves to 83.3% and the recall rate for the previously near-failing "climbing" markedly soars from 16.7 to 83.3%. This is attributed to the LLM's excellent contextual understanding; it can recognize that if a marked change in vertical height occurs following a continuous action approaching furniture with elevation differences, it should belong to climbing intent rather than an instantaneous fall. However, this re-ranking mechanism also demonstrates the price of "high sensitivity" in LLMs under healthcare settings. To prevent missed reports, the LLM tends to strictly judge certain semantically ambiguous normal actions (such as seniors stretching widely or turning irregularly) as "disoriented" or potential hazards. This results in a drop in precision for normal behaviors and falls. Overall, this data distribution indicates that the algorithm design of this system fully conforms to the supreme safety guiding principle in home healthcare: "Better to falsely alarm than to falsely ignore." In real-world home care settings, missed detections of critical events, such as falls, carry catastrophic consequences. However, to actively mitigate the potential issue of alarm fatigue among family members or caregivers caused by frequent false positive alerts, the system generates specific semantic text reports via the LLM. This allows caregivers to rapidly evaluate situations and dismiss false alarms without the cognitive burden of reviewing continuous, ambiguous warning videos.

4.6 Ablation experiment: Importance of knowledge graphs

While the integration of generative AI has significantly optimized semantic re-ranking, a core architectural question remains: can the system rely entirely on the visual reasoning

Table 5
System performance with generative AI re-ranking.

Event case	Total number of samples (test/upload)	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Normal behavior	200/30	99.5/90.0	99.0/71.4	100/83.3	99.5/76.9
Fall	200/30	99.0/56.7	98.5/26.7	99.5/66.7	99.0/38.1
Disoriented action	200/30	98.5/86.7	98.0/62.5	99.0/83.3	98.5/71.4
Climbing	200/30	98.0/96.7	99.0/100	**97.5/83.3**	98.2/90.9
Running	200/30	98.5/96.7	98.5/83.3	**98.0/83.3**	98.2/83.3

capabilities of large language models for risk assessment, thereby omitting the complex K-BERT and knowledge graph frameworks? To address this issue, we conducted an ablation study, completely removing the knowledge reasoning module and retaining only the generative AI layer to perform event classification.

The results in Table 6 bring forth the most disruptive and critical insight of the entire research. Upon losing the strict structured definitions of the K-BERT knowledge graph, the system's detection capability for the core risk of "falling" completely collapses. In real-world scenarios, the recall rate for falls plummets from 100% in the complete system to a mere 50.0%. This means that if relying solely on generative AI, half of the actual fall events would be ignored by the system or interpreted as other normal behaviors. This disastrous data decline reveals a fatal flaw of large language models in safety monitoring applications. Generative AI possesses overly abundant associative capabilities and divergent logic. When it sees a person lying on the ground, without the mandatory constraints of medical rules, it might over interpret and independently deduce that the person is "doing yoga" or "looking for dropped items," thereby overlooking the most direct fall risk. The only category unaffected is "running" (maintaining a 93.3% in recall rate), proving that generative AI excels at recognizing dynamic and highly continuous actions, but is extremely unreliable for momentary danger boundaries with high degrees of ambiguity. The results of this ablation study perfectly confirm that in pursuing high safety in the medical and care fields, deterministic knowledge graph reasoning based on K-BERT is an absolutely irreplaceable foundational defense line, while generative AI can only serve as an auxiliary tool for boundary fine-tuning. A hybrid neuro-symbolic architecture consisting of both is the optimal solution. Furthermore, the integration of the LLM serves as a fundamental semantic reasoning layer, not merely a localized patch for specific misclassifications. Distinguishing ambiguous complex actions remains a significant bottleneck in deterministic rule-based systems. The LLM provides the necessary contextual and temporal understanding, proving itself as an irreplaceable component of the hybrid neuro-symbolic architecture.

Table 6
Ablation study results (generative AI only).

Event case	Total number of samples (test/upload)	Accuracy (%)	Precision (%)	Recall (%)	F1-score (%)
Normal behavior	200/30	79.5/86.7	49.3/66.7	92.5/66.7	64.3/66.7
Fall	200/30	66.0/76.0	37.0/41.7	**100/50.0**	54.1/45.5
Disoriented action	200/30	78.5/92.0	48.1/90.9	95.0/66.7	63.9/76.7
Climbing	200/30	75.5/76.0	49.4/41.2	97.5/46.7	65.5/43.7
Running	200/30	89.2/98.7	65.0/100	97.5/93.3	78.0/96.6

5. Conclusions

We proposed a home care event detection system combining knowledge graphs and the transformer to address the limitations of traditional single-modal vision sensors. By integrating BLIP and GroundingDINO, the system effectively anchors semantic context to physical reality, providing a framework that is both highly accurate and privacy-preserving. Then, the K-BERT model ensures a high recall rate for critical events such as falls, while the introduction of generative AI rearrangement layers resolves the semantic ambiguity of complex actions such as climbing and running, significantly reducing invalid alerts. The results demonstrate that our hybrid neural symbolic architecture provides a robust and highly reliable solution for security monitoring. However, current methods primarily rely on visual and structured data evaluated in digital twin environments. Future research will focus on expanding the scale of knowledge graphs and integrating additional multimodal features such as speech and acoustic sensors. In addition, more datasets to validate this architecture improve its generalization capabilities and support the development of the next generation of smart home security ecosystems.

References

1 K. Simonyan and A. Zisserman: Proc. 2014 28th Neural Information Processing Systems (2014) 568–576.
2 S. Egami, T. Ugai, M. Oono, K. Kitamura, and K. Fukuda: IEEE Access **11** (2023) 23857. https://doi.org/10.1109/ACCESS.2023.3253807
3 J. Li, D. Li, C. Xiong, and S. Hoi: Proc. 2022 39th Int. Conf. Machine Learning (2022) 12888–12900.
4 S. Liu, Z. Zeng, T. Ren, F. Li, H. Zhang, J. Yang, C. Li, J. Yang, H. Su, J. Zhu, and L. Zhang: Lect. Notes Comput. Sci. **15105** (2024) 38. https://doi.org/10.1007/978-3-031-72970-6_3
5 Q. Huang, Z. Gan, A. Celikyilmaz, D. Wu, J. Wang, and X. He: Proc. AAAI Conf. Artificial Intelligence. (2019) 8465–8472. https://doi.org/10.1609/aaai.v33i01.33018465
6 Y. J. Tian, V. Duong, E. Buhr, N. A. Felber, D. R. Schwab, and T. Wangmo: AJOB Empirical Bioethics **16** (2025) 61. https://doi.org/10.1080/23294515.2024.2416121
7 H. Song, S. Tian, J. Hao, C. Yuan, Z. Jia, J. Shao, and X. Li: IEEE Access **14** (2026) 3658575. https://doi.org/10.48550/arXiv.2601.22938
8 C. Bettini, A. Moradbeikie, and G. Civitarese: Proc. IEEE Int. Conf. Pervasive Computing and Communications Workshops Workshops (IEEE, 2025) 458–463. https//doi.org/10.1109/PerComWorkshops65533.2025.00108
9 H. Wang, Y. Wang, X. Miao, Y. Zhang, Y. Zhang, and A. Mansoor: Proc. Int. Conf. Artificial Intelligence of Things and Systems (2025) 129.
10 Q. Memon, M. Al Ameri, and N. Musthafa: Proc. Eng. Technol. Innov. **28** (2024) 30. https://doi.org/10.46604/peti.2024.13900
11 Y.-S. Chang and G.-Y. Lin: Sens. Mater. **37** (2025) 1639. https://doi.org/10.18494/SAM5419
12 M. D. Xames and T. G. Topcu: IEEE Access **12** (2024) 4099. https://doi.org/10.1109/ACCESS.2023.3349379
13 Z. Momand, P. Mongkolnam, J. H. Chan, N. Charoenkitkarn, and D. Pal: IEEE Access **13** (2025) 169415. https://doi.org/10.1109/ACCESS.2025.3607603
14 A. Shankhdhar and H. Garg: Cluster Comput. **28** (2025) 956. https://doi.org/10.1007/s10586-025-05602-8
15 Y. Wang, J. C. Leung, M. Chen, Z. Zeng, B. T. H. Tan, Y. Qiu, and Z. Shen: Proc. 2025 IEEE Region 10 Symposium (IEEE, 2025) 1–8.
16 S. Adibi, A. Rajabifard, D. Shojaei, and N. Wickramasinghe: Sensors **24** (2024) 2793. https://doi.org/10.3390/s24092793
17 J. Lehmann, L. Granrath, R. Browne, T. Ogawa, K. Kokubun, Y. Taki, K. Jokinen, S. Janboecke, C. Lohr, R. Wieching, and G. M. Revel: Sustainability **16** (2024) 3064. https://doi.org/10.3390/su16073064
18 C.-H. Li: Int. J. Eng. Technol. Innovation **15** (2025) 476. https://doi.org/10.46604/ijeti.2025.15708

19 K. Zhao, Y. Song, H. Zhao, H. Liu, T. Li, and Z. Li: Proc. 33rd ACM Int. Conf. Information and Knowledge Management (2024) 3320-3330.
20 W. Chen, H. Sun, M. You, J. Jiang, and M. Rivera: Energies **18** (2025) 833. https://doi.org/10.3390/en18040833
21 Z. Wang, D. Shi, J. Zhao, X. Diao, X. Tang, and Y. Qin: arXiv preprint arXiv:2511.13526 (2025). https://doi.org/10.48550/arXiv.2511.13526
22 N. Zhu, Y. Dong, T. Wang, X. Li, S. Deng, Y. Wang, and S. Jiao: arXiv preprint arXiv:2508.19542 (2025). https://doi.org/10.48550/arXiv.2508.19542
23 C. Huang, B. Wang, J. Wen, C. Liu, W. Wang, L. Shen, and X. Cao: Proc. 2025 38th Neural Information Processing Systems (2025) 118486–118518. https://doi.org/10.52202/085713-3952
24 B. Lin, Y. Ye, B. Zhu, J. Cui, M. Ning, P. Jin, and L. Yuan: Proc. 2024 Conf. Empirical Methods in Natural Language Processing (2024) 5971–5984. https://doi.org/10.18653/v1/2024.emnlp-main.342
25 Y. Wang, K. Li, X. Li, J. Yu, Y. He, G. Chen, B. Pei, R. Zheng, J. Huang, and L. Wang: European Conf. Computer Vision (2024) 396–416. https://doi.org/10.48550/arXiv.2403.15377
26 S. Egami, T. Ugai, S. N. N. Htun, and K. Fukuda: Proc. 33rd ACM Int. Conf. Information and Knowledge Management (2024) 5360. https://doi.org/10.1145/3627673.3679175

About the Authors

Jui-Feng Yeh received his B.S. degree in computer science and information engineering from National Chiao-Tung University in 1993, and his M.S. and Ph.D. degrees in computer science and information engineering from National Cheng Kung University in 1995 and 2006, respectively. He is currently a professor of the Department of Computer Science and Information Engineering, National Chiayi University, Republic of China (ROC). His research interests include artificial intelligence, speech signal processing, natural language processing, and affective computing.
(ralph@mail.ncyu.edu.tw)

Chun-Yu Tsai received his B.S. degree from the Department of Computer Science and Information Engineering, National Chiayi University, Taiwan, in 2025. He is currently pursuing his M.S. degree at National Chiayi University, Taiwan. He also holds a part-time position at the university's Computing Center, where he is primarily responsible for the management and maintenance of the computer network in student dormitories. His research interests are in AI and computer vision. (s1140335@mail.ncyu.edu.tw)

Bo-Sung Huang is currently studying at the Department of Computer Science and Information Engineering, National Chiayi University, Taiwan. He has also worked part-time at the university's Computer Center for several years, primarily responsible for the management and maintenance of computer equipment. His research interest is mainly machine learning.
(s1112952@mail.ncyu.edu.tw)

Po-Hsiang Yang is currently pursuing his B.S. degree in the Department of Computer Science and Information Engineering, National Chiayi University, Taiwan. He is also an intern at Kdan Mobile Software Ltd., where he is responsible for software testing. His research interests include deep learning and event detection. (s1112956@mail.ncyu.edu.tw)

Chen-Yu Yeh received his B.S. degree from the School of Medicine, China Medical University, Taiwan (R.O.C.) in 2025. He is currently serving as an M.D. in a medical physician team at China Medical University Hospital, Taiwan (R.O.C.). His research interests include medicine and AI applications in biomedical information. (100028@tool.caaumed.org.tw)