

Dense Video Captioning Method for Indoor Human Behavior Enhanced with Spatial Information by RGB Depth Camera

陳, 斌

<https://hdl.handle.net/2324/7363820>

出版情報 : Kyushu University, 2024, 博士 (工学) , 課程博士
バージョン :
権利関係 :





Kyushu University

Graduate School of Information Science and Electrical Engineering
Department of Information Science and Technology

Dense Video Captioning Method for Indoor Human Behavior Enhanced with Spatial Information by RGB Depth Camera

By

Chen Bin

*A thesis submitted in partial fulfillment of the requirements for
the degree of Doctor of Engineering*

Supervised by

Professor Yutaka Arakawa

February 2025

Chen Bin

Dense Video Captioning Method for Indoor Human Behavior Enhanced with Spatial Information by RGB Depth Camera

Supervisor: Professor Yutaka Arakawa

Advisory Committee: Prof. Yoshihiro Okada, Assoc. Prof. Mine Tsunenori, Assoc. Prof. Shogo Fukushima, and Prof. Yutaka Arakawa

Kyushu University

Graduate School of Information Science and Electrical Engineering
Department of Information Science and Technology

Abstract

The video medium contains vast amounts of content, serving as an important raw information resource for a wide range of vision-based human-centered applications. For instance, security surveillance systems need to analyze and monitor human action to detect unnatural actions that may indicate potential threats, enabling timely alarms or interventions to prevent harmful outcomes. Similarly, indoor caregiving systems must interpret human action and record relevant activities. These records provide critical information for doctors to devise appropriate medical plans for individuals under care.

However, the large volume of video data poses significant challenges in terms of storage, management, and the efficient retrieval of relevant information for such systems. This necessitates the development of automated methods to summarize videos and extract pertinent information effectively. Fortunately, with advancements in deep learning, dense video captioning—a subfield of video understanding that generates textual captions for target events in videos—has seen substantial progress. This field has garnered significant attention due to its broad applicability. Captions offer a convenient way to search for and record relevant video information, making them highly valuable for video-based human-centered applications.

However, current dense video captioning methods designed for analyzing human action are not well-suited for understanding indoor human action, where the meaning and categorization of actions are closely intertwined with the environment. For instance, in a security surveillance system, a natural behavior such as "the person is cutting something in the kitchen" differs significantly from an unnatural behavior like "the person is cutting something behind the bed." Similarly, in an indoor caregiving system, a normal action such as "the person is sleeping in the bed" contrasts with an abnormal action like "the person is sleeping in front of the door." In summary, current dense video captioning methods lack an effective approach to incorporate spatial

information into generated captions. This limitation hinders their potential for video-based human-centered applications.

To bridge the gap between dense video captioning and video-based human-centered applications, and to enhance the information provided in the generated captions, we conducted three progressively detailed studies.

Foundational Research - Spatial and Ambiguity-Reduced Captioning. The initial phase focused on developing a plug-and-play module to enhance the quality of captions generated by existing methods. By integrating spatial information extracted from depth and color images, this module ensures that captions incorporate essential elements such as actions, locations, and object interactions while reducing ambiguity. Experimental results on a custom dataset validated the module’s effectiveness in incorporating spatial context.

Extended Research - RAG and Knowledge Base Integration. Building on the foundational module, the second phase introduced a system leveraging Retrieval-Augmented Generation (RAG) and Large Language Models (LLMs). By constructing a 3D knowledge base from scene reconstruction and detection data, the system enhanced the contextual relevance of generated captions, particularly for smart home and indoor security applications. Comparative evaluations with methods such as PDVC and GVL demonstrated superior performance in metrics including BLEU-3, METEOR, and CIDEr, confirming the system’s ability to produce detailed and unambiguous captions.

Innovative Framework - Comprehensive Spatial Awareness In the final phase, the research expanded spatial awareness to encompass not only static elements but also moving objects within the scene. This RAG-based framework integrates detailed spatial information of the entire environment, ensuring that captions are both contextually comprehensive and well-suited for human-centric applications. Quantitative and qualitative comparisons with PDVC, GVL, SG-PDVC, and SI-PDVC further highlighted the framework’s superiority in generating spatially enriched captions.

Acknowledgements

The successful completion of this thesis would not have been possible without the support, guidance, and encouragement I have received from numerous individuals and organizations. I am deeply grateful to all who contributed to my academic journey.

First, I would like to express my sincere gratitude to the SPRING Scholarship Program in Japan for providing me with financial support and the opportunity to pursue my studies. This support not only eased the financial burden but also allowed me to dedicate myself fully to my research, fostering both my academic and personal growth.

I am profoundly thankful to my advisors, Professor Arakawa for your invaluable guidance, insightful feedback, and unwavering support throughout my research journey. Your collective mentorship has been instrumental in shaping my academic development. The trust you placed in me and your encouragement during challenging times gave me the confidence to persevere and improve. I am especially grateful for the consistent support you provided during the most difficult moments of my research. I would also like to extend my sincere gratitude to Professor Fukushima and Professor Mine for their thoughtful advice and constructive suggestions, which have greatly contributed to my academic growth. Additionally, I am deeply thankful to Assistant Professor Nakamura for his assistance, which have been invaluable throughout my studies.

I would also like to extend my thanks to my parents, whose unwavering encouragement and understanding have been the bedrock of my achievements. Their sacrifices and steadfast belief in my potential have given me the strength to overcome countless challenges.

Finally, to my fellow students and colleagues, thank you for your collaboration, support, and camaraderie. The discussions, shared ideas, and mutual

encouragement have enriched both my research and my academic experience, creating a strong sense of community.

Contents

1	Introduction	1
1.1	Background	1
1.2	Research Motivations	2
1.3	Contribution	2
1.4	Thesis Outline	2
2	Related Literature	5
2.1	Dense Video Captioning	6
2.1.1	Dense Video Captioning Methods	6
2.1.2	Datasets for dense video caption	7
2.1.3	Metrics for dense video caption	7
2.2	Scene Graph Generation and Image Segmentation	9
2.3	Large Language Model	10
2.4	RAG for LLMs	12
3	Fundamental Concepts	15
3.1	Metrics for Evaluating Dense Video Captioning	15
3.2	Transformer	17
3.3	Common Operations on 3D Point Clouds	19
3.3.1	3D Scene Reconstruction	19
3.3.2	3D Object Detection	19
3.3.3	Point Cloud Registration	19
3.4	Coordinate Transformation	21
4	A Space Information-Enhanced Dense Video Caption for Indoor Human Action Recognition	23
4.1	Background	23
4.2	Method	24
4.2.1	Framework	25

4.3	Experiment and Discussion	28
4.3.1	Experiment Setting	28
4.3.2	Implementation Details	28
4.3.3	Comparison Results	29
4.4	Summary	30
4.5	Main Limitations of This Method and Future Solutions in Next Chapter	31
5	RAG-enhanced Dense Video Caption for Human Indoor Actions: Disambiguating Caption Using Spatial Information Beyond FOV Constraints	33
5.1	Background	33
5.2	Methods	36
5.2.1	3D Reconstruction and 3D Object Detection to build a Knowledge Base (initial part)	37
5.2.2	Design of Spatial-enhanced Query and RAG-enhanced LLM module (caption part)	40
5.3	Experiment and Discussion	42
5.3.1	Knowledge Base of Surrounding Environment	43
5.3.2	Spatial Information Enhanced Query	45
5.3.3	RAG-enhanced LLM Module	46
5.4	Summary	48
5.5	Main Limitations of This Method and Future Solutions in Next Chapter	49
6	Generating Spatially-Aware Dense Video Captions for Indoor Human Action with Position-Based Scene Knowledge	51
6.1	Background	51
6.2	Method	54
6.2.1	Framework	54
6.2.2	Spatial Enhanced Query Extraction	55
6.2.3	Knowledge Base Building	57
6.2.4	RAG Procedure Designing	58
6.3	Experiment and Discussion	60
6.3.1	Experiment Setting	60
6.3.2	Comparison Results	61
6.4	Summary	63

7 Conclusion and Future Work	65
7.1 Conclusion	65
7.2 Limitations and Future Work	66
Bibliography	67
Appendix	71

List of Figures

1.1	The overall structure and contents of the dissertation.	3
2.1	example of human action dataset for dense video caption. This figure was shown in paper [6].	8
2.2	example of cooking dataset for dense video caption. This figure was shown in paper [5].	8
3.1	Architecture of the Transformer. This figure was shown in paper [20].	18
3.2	example of 3D reconstruction	20
3.3	example of 3D detection	21
4.1	Main scheme of the system	25
4.2	Architecture of the System	26
4.3	Spatial and Ambiguity-reduced Information Extraction	27
4.4	Qualitative Comparison Results	29
5.1	Target problem and expected result.	34
5.2	Architecture of the whole system.	36
5.3	Pipeline to align the source point cloud to the target point cloud	39
5.4	(a) Test video and (b) 2D layout of the house with the corresponding camera position.	43
5.5	(a–d) 3D scene reconstruction and detection results of the two rooms.	44
5.6	(a) 3D fragment captured with Azure Kinect in the camera coordinate. (b) 3D scene transforms and registers to the corresponding camera coordinate	45
5.7	Human detection result on RGB-D images pairs	46
5.8	Qualitative comparison result	47

6.1	Illustration of the ability of current methods to extract spatial information and our expected results	53
6.2	Architecture of the System.	54
6.3	Layout of the rooms and the Furniture arrangement.	61

List of Tables

4.1	Captioning and Localization performance	30
5.1	The setting up of Azure Kinect device camera	42
5.2	Quality of registration result	43
5.3	Captioning and Localization performance	48
6.1	Qualitative Comparison Results.	61
6.2	Captioning and Localization performance	63

Acronyms

RGB-D	Red-Green-Blue-Depth
DVC	Dense Video Caption
VC	Video Caption
LLM	Large Language Model
ICP	Iterative Closest Point
FOV	Field of View
Seq2Seq	Sequence to Sequence
RAG	Retrieval-Augmented Generation
LSTM	Long Short-Term Memory
IoU	Intersection over Union
BERT	Bidirectional Encoder Representations from Transformers
AI	Artificial Intelligence
3D	Three Dimensional
DL	Deep Learning
ML	Machine Learning
TTA	Textual Temporal Attention
ICL	In-Context-Learning
CNN	Convolutional Neural Network
AP	Average Precision
SVD	Singular Value Decomposition

RMSE Root Mean Square Error

LCS Longest Common Subsequence

Introduction

1.1 Background

With advancements in sensor devices and deep learning algorithms, vision-based human-centric applications have been widely adopted, significantly enhancing daily life and improving its quality. In all such systems, video content understanding plays a central role. For example, in security surveillance systems, the technology is designed to detect potentially dangerous behaviors and take responsive actions, such as stopping the behavior. Similarly, in remote caregiving systems, it is tasked with identifying abnormal behaviors and performing appropriate responses, such as recording incidents and reporting them to a doctor.

However, video, the primary information source for such systems, is challenging to store and search due to its large volume and complex data format. Nowadays, dense video captioning, a subfield of video understanding, has made significant progress thanks to advancements in deep learning. This technology can generate detailed textual descriptions of interesting events, providing an effective and high-quality textual representation of video content. Such descriptions hold great potential for improving the performance of downstream human-centric pipelines. However, current dense video captioning methods are not specifically designed for analyzing human action. As a result, these methods often lack spatial information, which is crucial for human-centered applications where the meaning of human action is tightly intertwined with the surrounding environment. For instance, in an indoor security surveillance system, a person performing the action category "cutting something" may carry vastly different meanings depending on the environment. Cutting something on a cutting board in the kitchen is a normal action, while cutting something behind a bed in a bedroom is an abnormal and potentially dangerous action that requires reporting. Similarly, in a remote

caregiving system, the action "sleeping" could also vary in significance based on context. Sleeping in a bed is a normal event, but sleeping in front of a door could indicate an abnormal situation that needs to be recorded.

1.2 Research Motivations

As a result, to leverage the advantages of dense video captioning for efficient video content storage and retrieval, and to create spatial-information-enhanced captions for more accurate recording of indoor human actions in downstream human-centric applications, we propose a progressive researches. These researches enhances dense video captioning of indoor human action by addressing challenges such as ambiguity and the lack of spatial details through a series of targeted advancements.

1.3 Contribution

Our contributions are as follows :

1. We designed a system to enhance dense video captioning of indoor human action with spatial awareness, providing more detailed and accurate captions.
2. We provided a plug-and-play module for the existing methods to output the space information enhanced caption for the human indoor activity.
3. We provided a convenient framework based on RAG to incorporate the detailed spatial information about the entire scene and supports the reporting of objects moving within the camera's FOV.

1.4 Thesis Outline

The rest of this dissertation is organized as Figure 1.1 shows. In Chapter 2 we present an overview of related works. In Chapter 3, we provide the

detailed definition of the related concept. In Chapter 4, we present a plug-and-play module designed to enhance existing captioning methods with spatial awareness. In Chapter 5, we propose an RAG-based framework for analyzing indoor human action that incorporates spatial information beyond the field of view (FoV) of individual cameras. In Chapter 6, building upon previous research, we introduce a dense video captioning system for indoor human action, leveraging spatial information from the entire room.

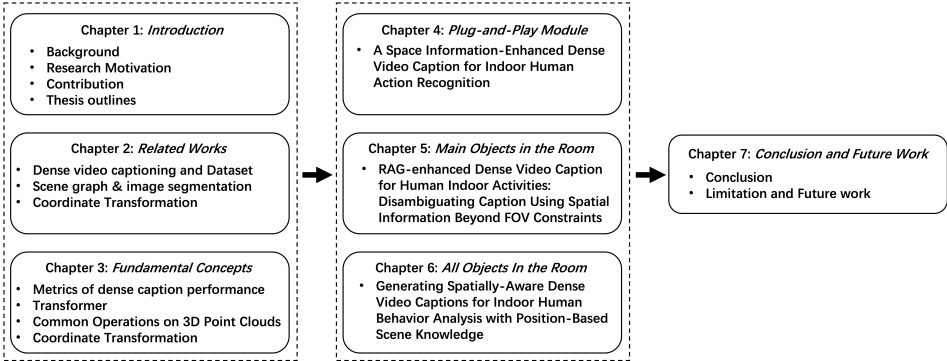


Fig. 1.1: The overall structure and contents of the dissertation.

Related Literature

Dense video captioning has advanced significantly with the development of deep learning, enabling the detection of interesting events and the generation of descriptive content about these events, which shows great potential for human-centric application such as video security applications. And in video security, the primary focus is on detecting abnormal behavior. The complexity of this task can be described from several perspectives. First, human action is inherently complex due to movements of the person. Second, the interpretation of the same behavior may vary across different environments. For example, a person cutting something on a chopping board in the kitchen is considered normal behavior, whereas a person cutting something behind a bed is abnormal behavior that needs to be recorded. Similarly, in a remote caregiving system, a person sleeping in bed is normal, but a person sleeping behind a door is considered abnormal behavior. This thesis focuses on incorporating spatial information into dense video captioning to generate more accurate descriptions of events in videos.

To incorporate spatial information into the generated captions, we utilize popular computer vision techniques, particularly scene graph generation and image segmentation. Additionally, we employ retrieval-augmented generation (RAG) and large language model (LLM)-based technologies to enhance the final captions, ensuring they include spatial information and read more natural.

2.1 Dense Video Captioning

2.1.1 Dense Video Captioning Methods

Dense video caption (DVC) tasks are used to detect interesting events and provide descriptive text for these events from untrimmed videos. [1] Compared with common video captioning methods, DVC has the following two characteristics. (1) It captures all detected events. In the dense video captioning pipeline, it aims to output all detected actions, meaning that it can include multiple action sentences for the same period. (2) It is applicable to untrimmed video. Unlike common video captioning methods, dense video captioning techniques, such as PDVC and GVL, do not require human-annotated boundaries of the video. They can directly use untrimmed video as input and can output the time period along with the captions. As a result, it has great potential to be used in security surveillance and human care systems.

Multimodality-based methods. These methods tend to use multiple source modalities to work together, making the output more accurate. For example, MDVC, a popular dense video captioning method proposed by Iashin and Rahtu in 2020, utilizes audio, speech, and visual information as modalities to design a multimodal generator that outputs a distribution over the vocabulary. [2] Vid2Seq, proposed by Yang et al. in 2023, uses video and transcribed speech as inputs to achieve state-of-the-art results in the DVC task. [3] However, various modalities may contain overlapping or contradictory information, and negative interactions between these modalities can reduce the effectiveness of caption generation.

Vision-modality-based methods. For example, PDVC, an end-to-end framework with parallel decoding proposed by Wang et al., uses video data as the input modality. [1] In their method, the dense video captioning problem is modeled as a set prediction task. They first used a CNN and transformer encoder to extract and encode image features. Then, they applied a parallel decoding phase with a transformer decoder, which features three heads (event counter, localization header, and caption header) to generate the localization and captions independently. [1] Wang et al. proposed the GVL, [4]

which is a grounded vision-language learning framework for untrimmed video aiming to automatically detect informative events and align multi-sentence descriptions with event segments. In their work, they introduced the Event-to-Text generation module and Text-to-Event grounding module to learn event-level representation to increase the accuracy of dense video caption tasks. However, similar to the PDVC, GVL also does not pay enough attention to the relationship between humans and their surrounding environment. This limitation hinders their use in security surveillance and human care applications.

2.1.2 Datasets for dense video caption

The current popular datasets used in dense video captioning include YouCook2 and ActivityNet. YouCook2 is one of the largest task-oriented instructional video datasets, containing 2000 long untrimmed videos from 89 cooking recipes. On average, each distinct recipe has 22 videos [5]. The ActivityNet Captions dataset is based on ActivityNet v1.3 and includes 20,000 YouTube untrimmed videos with 100,000 caption annotations. These videos have an average length of 120 seconds [6]. These datasets are typically used to train end-to-end models and require substantial human effort for collection. A example for human action is shown in the Figure 2.1 taken from ActivityNet. A example for cooking is shown in the Figure 2.2 taken from YouCook2. Hence, there is a need for a more accurate and efficient method to incorporate spatial information and expand the range of application scenarios.

2.1.3 Metrics for dense video caption

Dense video captioning typically involves two sub-tasks: detecting the time period of the interesting event and describing the content of the event. To evaluate the quality of the detected time period, common methods include using Precision and Recall as metrics. To evaluate the quality of the event content, common metrics from the NLP field are used, including METEOR[7], BLEU[8], CIDEr[9], and ROUGE-L[10]. Different metrics have different

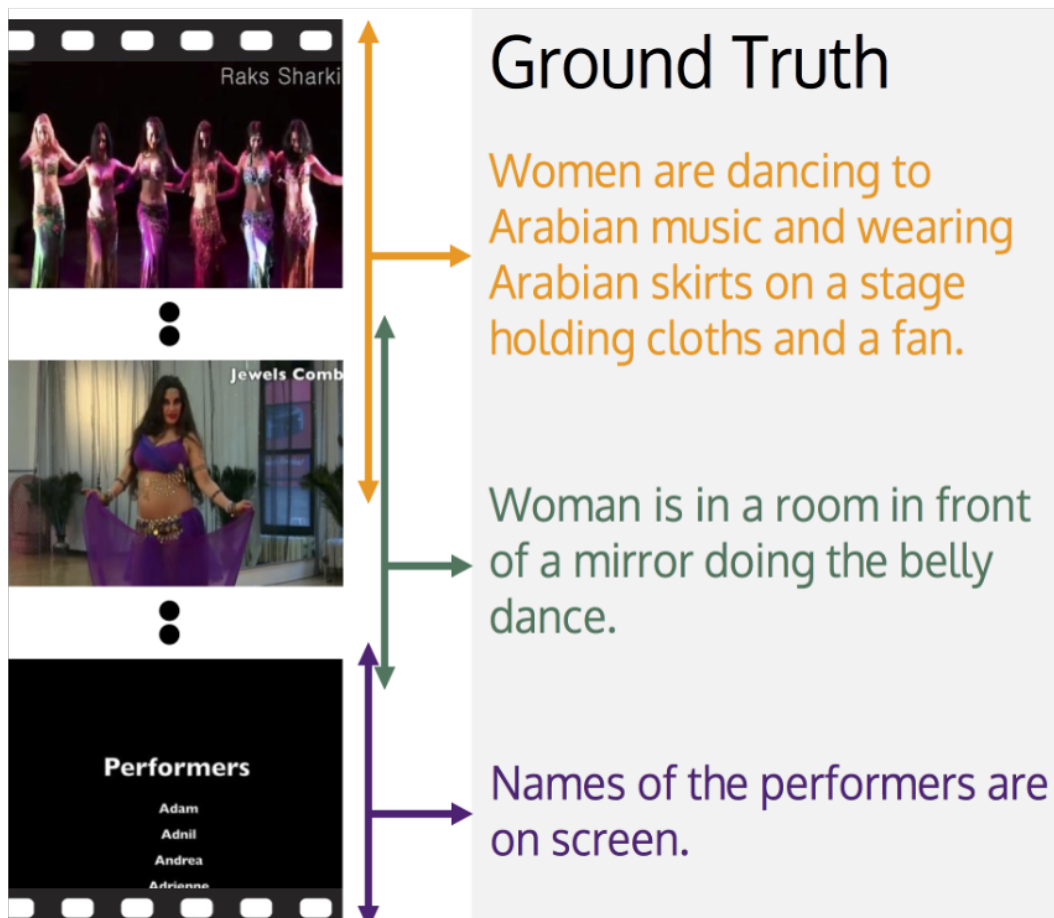


Fig. 2.1: example of human action dataset for dense video caption. This figure was shown in paper [6].

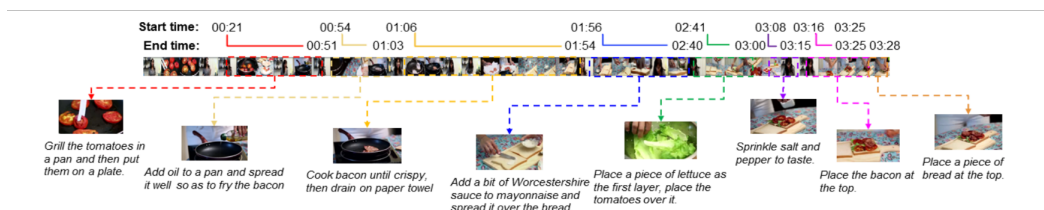


Fig. 2.2: example of cooking dataset for dense video caption. This figure was shown in paper [5].

focuses. Among the four metrics, BLEU was the first to be proposed, designed to address the challenge of non-reusable human evaluation in machine translation tasks. BLEU emphasizes the precision of n-grams. ROUGE-L focuses on recall and utilizes the Longest Common Subsequence (LCS), which goes beyond simple n-gram overlap to capture structural similarity, making it particularly suitable for summarization tasks. METEOR balances precision and recall while penalizing disordered matches, enabling the metric to better capture the coherence and readability of translations. CIDEr introduces TF-IDF weights, assigning importance to n-grams based on their informativeness in the dataset. This metric emphasizes human consensus by prioritizing the relevance of information.

To summary, in human-centric application scenarios where spatial information is necessary, existing methods encounter 3 issues: 1. Less emphasis on spatial information. 2. The spatial information is only extracted within FOV. 3. Predictions are not always accurate.

2.2 Scene Graph Generation and Image Segmentation

Scene Graph Generation (SGG) is an important technology for vision-based scene understanding. Its primary goal is to analyze and recognize visual content and represent their relationship as a graph structure. In previous methods, objects are detected using bounding boxes, followed by the prediction of their pairwise relationships [11]. In 2022, J. Yang et al. introduced Panoptic Scene Graph Generation (PSG), a novel task that challenges models to create a more comprehensive scene graph representation based on panoptic segmentation, as opposed to rigid bounding boxes [11]. In contrast to SGG, the PSG task consistently involves three sub-tasks: semantic segmentation, instance segmentation, and relationship prediction. In their paper, they present a transformer-based method and define 56 predicate relationships. Following their design, Zijian Zhou et al. introduced a new method, Hilo [12], for the PSG task. In their paper, they utilize a DETR-based panoptic segmentation method, Mask2Former [13], as the preprocessing

step to obtain accurate panoptic segmentation results. They subsequently generate more precise outcomes.

2.3 Large Language Model

Large Language Models (LLMs) are advanced artificial intelligence (AI) models built on deep learning (DL) technology to process and generate natural language text. At present, LLMs are widely used in fields such as finance and education, achieving strong performance due to their ability to extract information from vast amounts of data and generate high-quality text.

LLMs are a subset of the language model (LM) field. The development of language models can be categorized into four stages: Statistical Language Models (SLMs), Neural Language Models (NLMs), Pre-trained Language Models (PLMs), and Large Language Models (LLMs).

SLMs are characterized by the idea that word order can be predicted by analyzing the co-occurrence frequency of words in the training data. A key work in this area is "Class-based n-gram models of natural language," published in 1992 by Brown et al. [14]. In their work, they address the problem of predicting a word from the preceding words in a text sample based on the underlying statistical properties. However, this method is relatively simple and does not achieve high accuracy, as it cannot capture the complex semantic and contextual information. With the development of neural networks and advances in computational power in recent years, this field has seen significant progress.

NLM is characterized by its use of neural network architectures. The goal of the language model is to estimate a sentence or sequence of words, as modeled in Equation 2.1. Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks are particularly famous.

$$P(w_1, w_2, \dots, w_T) = \prod_{t=1}^T P(w_t | w_1, w_2, \dots, w_{t-1}) \quad (2.1)$$

The RNN architecture differs from traditional neural networks due to its cyclic structure, which allows information to propagate between time steps. This feature enables the architecture to retain contextual information. The state update formula is shown in Equation 2.2 where the h_t is the current hidden state, x_t is the input at time step t , W_x, W_h are weight matrices, b is the bias and σ is the activation function such as *tanh* and *ReLU*.

$$h_t = \sigma(W_x x_t + W_h h_{t-1} + b) \quad (2.2)$$

The LSTM architecture is an updated model that addresses the limitations of RNNs, such as vanishing gradients and limited long-term memory. Compared to the RNN, the LSTM introduces two key components: the cell state and the forget gate. The gradients of the RNN are shown in Equation 2.3, where the $\frac{\partial L}{\partial h_t}$ represent the gradients from loss L to hidden state h_t .

$$\frac{\partial L}{\partial W} = \sum_{t=1}^T \frac{\partial L}{\partial h_t} \cdot \prod_{k=t}^1 \frac{\partial h_k}{\partial h_{k-1}} \cdot \frac{\partial h_1}{\partial W} \quad (2.3)$$

The LSTM introduces the cell state, which provides a nearly stable gradient stream, as shown in Equation 2.4 where the f_t is the forget gate and the i_t is input gate. The design of the forget and input gates controls the information flow for updating and forgetting, helping to avoid the exponential decay of gradients and maintain longer-term memory.

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t \quad (2.4)$$

However, these methods still lack the ability to effectively understand long texts and are difficult to train, requiring significant time and computational resources.

The PLMs are characterized by the transformer architecture and introduce contextualized embeddings. These methods improve training effectiveness and capture long-distance dependencies due to the design of the transformer architecture. The transformer uses a self-attention mechanism, which can be processed efficiently in parallel, whereas RNNs and LSTMs are calculated step by step. A important example of such methods is Bidirectional Encoder Representations from Transformers (BERT) [15]. In their paper, they introduce a new language representation model designed to pre-train deep bidirectional

representations from unlabeled text. It demonstrates state-of-the-art results on eleven NLP tasks.

LLMs are a subset of PLMs, but due to their significant improvements over traditional PLM models, they are introduced separately here. Most LLMs are based on transformer architecture, though they differ greatly in scale. For example, BERT has around 110M parameters, while GPT-3, a well-known and popular LLM, has 175B parameters. With their larger scale, LLMs can often perform multiple tasks in few-shot scenarios without additional fine-tuning, whereas traditional PLMs are typically fine-tuned for specific tasks, such as sentiment analysis. Current LLMs are trained on vast amounts of data and demonstrate high performance across multiple tasks, such as question answering, translation, and summarization, making them applicable to a wide range of use cases. However, LLMs are also difficult to train, requiring large amounts of data, time, and computational resources. Additionally, LLMs may generate incorrect answers and lack explanations for their responses.

2.4 RAG for LLMs

Large language models have obtained significant attention due to their adaptability across various fields, achieving human-like performance as seen in models like ChatGPT[16] and Gemini[17]. Current research[18] suggests that the success of LLMs is largely due to their ability to act as knowledge bases. This means they store knowledge implicitly in their parameters, learned from training datasets, and generate responses by retrieving this stored information from internal memory. As a result, to enhance the performance of LLMs, the most common approach has been to increase the number of parameters and train them on larger datasets. However, LLMs face two major challenges. The first is the hallucination problem[19], where the model might output plausible but incorrect information. The second challenge is the difficulty of updating the knowledge stored in LLMs, as training or fine-tuning them is a resource-intensive process. Retrieval-Augmented Generation (RAG) is an increasingly popular AI framework aimed at enhancing LLMs by providing a more efficient way to incorporate up-to-date knowledge. The RAG pipeline typically consists of three components: the retriever, the

generator, and the retrieval fusion. Unlike traditional methods that require training the large model, the RAG pipeline extracts relevant information from an external knowledge base and directly fuses it with the LLM to generate the final answer.

Fundamental Concepts

3.1 Metrics for Evaluating Dense Video Captioning

Caption Performance. Dense video captioning methods aim to understand video content and generate captions that describe it accurately. To evaluate performance, the similarity between generated captions and ground truth captions is measured using various metrics. These metrics differ based on the evaluation criteria, with commonly used ones including METEOR, BLEU, CIDEr, and ROUGE-L.

Localization Performance. Dense video captioning also requires recording the time periods of detected events. Localization performance assesses whether the events are correctly identified and whether their temporal boundaries are accurate. Metrics such as precision and recall are commonly used for this evaluation.

- **BLUE** is a commonly used metric for evaluating generated text, particularly in tasks such as text translation and video captioning. It measures the quality of generated captions by calculating the similarity to ground truth captions, based on the n-grams in the generated caption that also appear in the ground truth caption.
- **METEOR** complements BLEU by addressing its limitations. Compared to BLEU, this metric considers semantic matching and assigns equal importance to recall and precision. In other words, METEOR evaluates not only word matching but also semantic alignment and word order. These characteristics make it an important metric for evaluating video captioning results.

- **ROUGE-L** measures the similarity between the generated caption and the ground truth caption based on the Longest Common Subsequence (LCS). Compared to strict n-gram-based metrics like BLEU, this method provides a more flexible approach by also considering sentence structure.
- **CIDer** is an important metric commonly used in image description tasks. It emphasizes semantic similarity and the significance of specific words. Specifically, this method combines n-gram overlap with TF-IDF weighting, where the weight is calculated based on the frequency of specific words in the ground truth caption set.
- **Recall** is a metric used to measure how effectively the system detects event periods from the ground truth. The formula it used is showed as the following equation 3.1

$$Recall = \frac{\text{Number of Correctly detected event}}{\text{Total number of ground truth events}} \quad (3.1)$$

A higher recall value indicates that more events are correctly detected. However, recall does not account for whether extra detections are correct. To evaluate the temporal accuracy of detected events, precision is also required.

- **Precision** is a metric used to measure the proportion of detected events that are correct. The equation to calculate precision is shown below equation. 3.2

$$Precision = \frac{\text{Number of Correctly detected event}}{\text{Total number of detected events}} \quad (3.2)$$

As a result, a higher value indicates that more relevant events have been detected among all detected events. Recall and Precision are always used together to describe the system's ability to achieve temporal accuracy.

3.2 Transformer

The transformer used in this thesis is a novel architecture characterized by its self-attention mechanism. It is often used to replace traditional sequence-to-sequence models, such as recurrent neural networks (RNNs) and long short-term memory (LSTM) networks. The popular architecture of the transformer is shown in the Figure 3.1. The figure is adapted from the paper-Attention Is ALL you Need [20]. From the image, we can summarize that the key components of the architecture are shown below.

- **The Input Embedding and the Output Embedding.** In the input embedding phase, words are first converted into a vector space, with positional encoding applied to preserve the order of the words. The output embedding phase is similar to the input embedding, where the output is embedded into the vector space, allowing the system to understand the output from the previous phase.
- **The Encoder** is the left part of the architecture, and it processes the input sequence to create contextualized representations.
- **The Decoder** has a similar structure to the encoder, but it includes a Masked Multi-Head Self-Attention module. The decoder generates output tokens based on previous tokens as input.

The most innovative aspect is the introduction of the Multi-Head Attention module. The attention score is calculated as shown in Equation 3.3.

$$Attention(Q,K,V) = SoftMax(\frac{QK^T}{\sqrt{d_k}})V \quad (3.3)$$

Where the Q is the query matrix, representing the word we attend to. The K is the key matrix, which represents all tokens we want to compare with. The V is the value matrix, holding the information we extract based on the attention weights. And the d_k is the dimension of the key vectors. Multi-Head Attention module computes multiple attention heads, with each head focusing on a different aspect of the sequence.

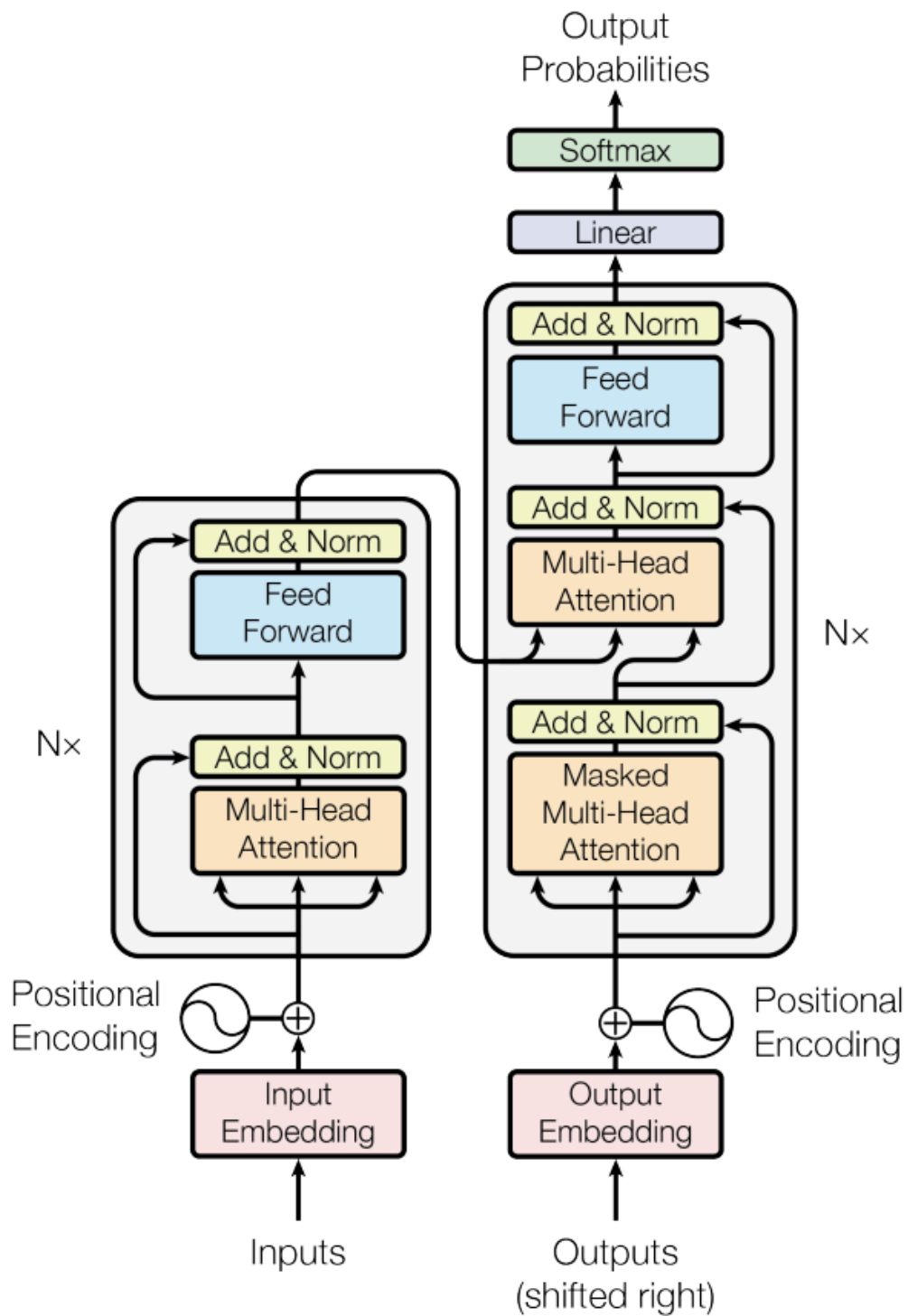


Fig. 3.1: Architecture of the Transformer. This figure was shown in paper [20].

3.3 Common Operations on 3D Point Clouds

3.3.1 3D Scene Reconstruction

3D scene reconstruction is an important and commonly used method for providing a 3D spatial representation of real-world scenes. This is crucial for enabling machines to understand and interact with the real world. The typical inputs for 3D scene reconstruction include RGB data, depth data, and video data. During the reconstruction process, the operator is usually required to move the sensor around the scene to capture it from different viewpoints. The resulting output is commonly saved as point clouds or meshes to represent the surfaces of objects in the scene or the scene itself.

In this thesis, we use the Scaniverse application to get the cloud point of the test rooms. The generated example results which are saved as point clouds are showed in the Figure 3.2. The picture show the reconstructed point clouds in 4 different directions.

3.3.2 3D Object Detection

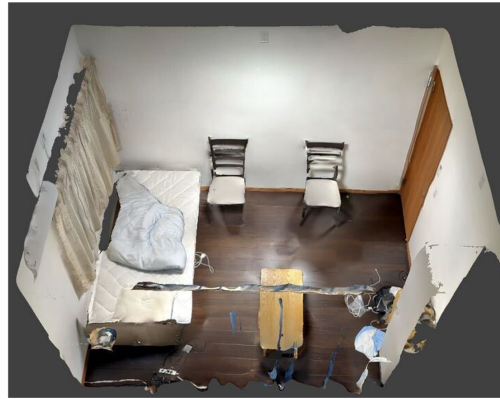
3D object detection is a task that requires the system to provide both the category and position of objects in 3D space. In contrast to 2D object detection, which is typically performed on RGB images and outputs 2D bounding boxes, 3D object detection generally outputs a 3D bounding box that represents the object's position and orientation in 3D space. An example of 3D object detection results is shown in Figure 3.3. The result is generated by the VoteNet model, using the point cloud of the scene as input. In the figure, different colors represent different categories, and the 3D bounding boxes indicate the positions and orientations of the objects.

3.3.3 Point Cloud Registration

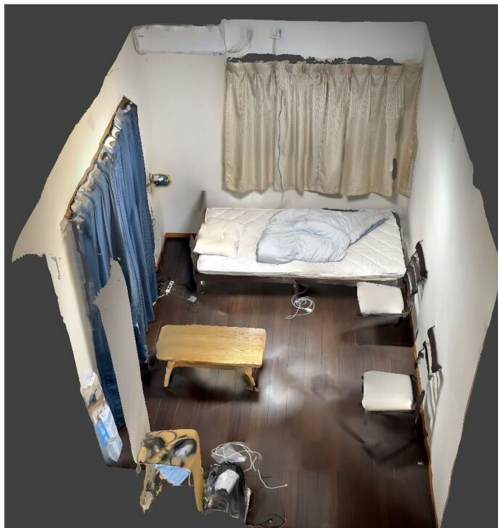
In real-world applications, the scene can sometimes be very large, making it difficult to capture a complete point cloud of the scene. As a result, to create



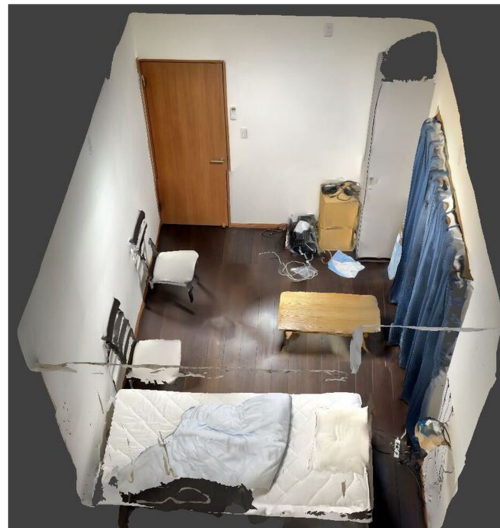
(a)



(b)



(c)



(d)

Fig. 3.2: example of 3D reconstruction

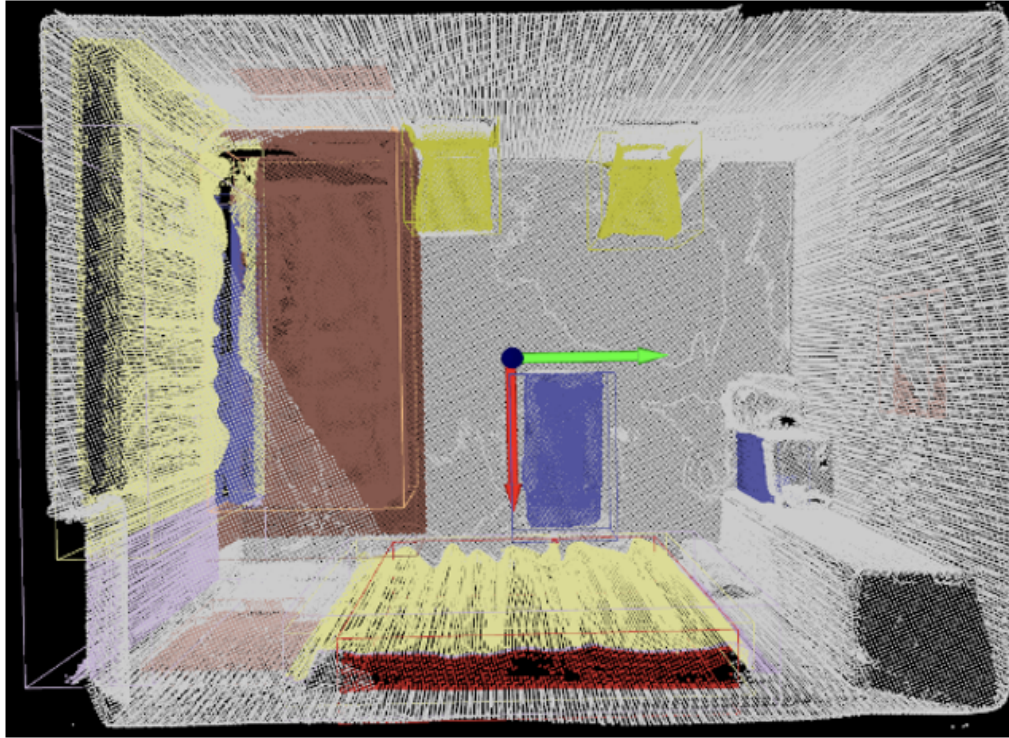


Fig. 3.3: example of 3D detection

a complete point cloud, it is often necessary to combine point clouds from different viewpoints or time periods. Point cloud registration is the technique used for this task. It processes the point clouds to merge them into a unified coordinate system. The key idea is to find a transformation matrix and a rotation matrix that minimize the error between corresponding points in the two point clouds. One of the most common methods to achieve this is by using the Open3D library. The Open3D library provides both global and fine-tune registration pipelines, which can meet most application scenarios.

3.4 Coordinate Transformation

When using cameras to detect objects in the world, it is often necessary to transform the results obtained by different methods, each in different coordinates, into a unified coordinate system for meaningful comparison. In this thesis, three coordinates are involved: the pixel coordinate, the camera

coordinate, and the world coordinate. All detection results in the thesis are transformed into the camera coordinate system for comparison.

2D Pixels coordinate to 3D camera coordinates. You Only Look Once(YOLO) performs 2D object detection on the RGB image and provides the results in pixel coordinates. To transform the information from 2D pixel coordinates to 3D camera coordinates, we typically need the intrinsic matrix and the depth values of the points in the RGB image. Specifically, we use the following Equation 3.4 to perform the coordinate transformation.

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = d \cdot \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} p_x \\ p_y \\ 1 \end{bmatrix} \quad (3.4)$$

Where the f_x, f_y, c_x and c_y are the intrinsic parameter of the camera, the p_x and p_y is the result generated by the YOLO methods in the pixel coordinate and the d is the corresponding depth value to the point in the RGB image. And the x, y and z are the corresponding position in 3D camera coordinate.

3D World Coordinate to 3D Camera coordinate. To transform from one 3D coordinate system to another, we need to obtain the transformation matrix and the orientation matrix. Specifically, the transformation method can be modeled using the following Equation 3.5.

$$\begin{bmatrix} x_c \\ y_c \\ z_c \end{bmatrix} = \mathbf{R} \begin{bmatrix} x_w \\ y_w \\ z_w \end{bmatrix} + \mathbf{t} \quad (3.5)$$

Where the $\mathbf{R}$ is a 3x3 rotation matrix and the $\mathbf{t}$ is a 3x1 matrix to represent the translate vector.

A Space Information-Enhanced Dense Video Caption for Indoor Human Action Recognition

4.1 Background

Video is an important information medium in security monitoring and smart home applications. The large volume of video content demands the need for automated methods to summarize and compactly represent the essential content [2]. One promising approach to creating content summaries is to use dense video captions, a technique that generates descriptive text for each event of the video [1]. Unlike classification methods that utilize skeleton joints or frames as input to output a specific defined category, as seen in the method proposed by Naresh et al [21], and the method proposed by Mostafa et al [22], the captions generated by dense video caption methods can always include additional information.

However, existing dense video captioning methods and related datasets have often overlooked spatial information and object relationships within the visual contents of videos. This limitation hinders the adaptability of these methods to other application scenarios, such as generating activity log files for indoor individuals or indoor security monitoring, where the main focus is on tracking abnormalities [23] and the same activity can have varying interpretations with different indoor positions. Intuitively, to clearly

and unambiguously describe indoor human action, captions should include information about the action class, location, and associated objects.

In this paper, we introduce a plug-and-play module to enhance spatial information in captions, building upon the current popular approach. As illustrated in Figure 4.1, we use Red-Green-Blue-Depth (RGB-D) data as input and directly extract accurate spatial information from the RGB-D images. Subsequently, we combine this spatial information with object relationships using a fine-tuned sequence-to-sequence (Seq2Seq) text generation model to generate the unambiguous and detailed captions. In our scheme, the Region of Interest (RoI) to which we pay particular attention include: 1. The human's position in relation to the main object. 2. The position of the object interacting with the human. 3. The ambiguous object in relation to other objects of the same class. We collected a dense video caption dataset enhanced with spatial information and object relationships of interesting objects and then evaluated our system on the dataset to show its effectiveness. Our results demonstrate that the system is capable of effectively generating dense video captions with space information.

The contribution of this paper is here:

1. We collected a dataset for dense video captioning, enriched with spatial information and the relationships among interesting objects.
2. We designed a plug-and-play module for existing methods to generate space-enhanced unambiguous captions.

4.2 Method

Dense video captioning aims to identify a set of temporally-localized captions of suitable length. In our paper, we model this task with the following Equation 4.1, where t_j^s , t_j^e , S_j represent the starting time of the event, ending time of an event, and the caption of an event, respectively. f_e , f_o , f_l represent the video caption module, spatial and ambiguity-reduced information extraction module, and sequence-to-sequence text generation module, respectively. In our evaluation section, we will use t_j^s and t_j^e to measure localization performance, while S_j will be used to measure dense video caption performance.

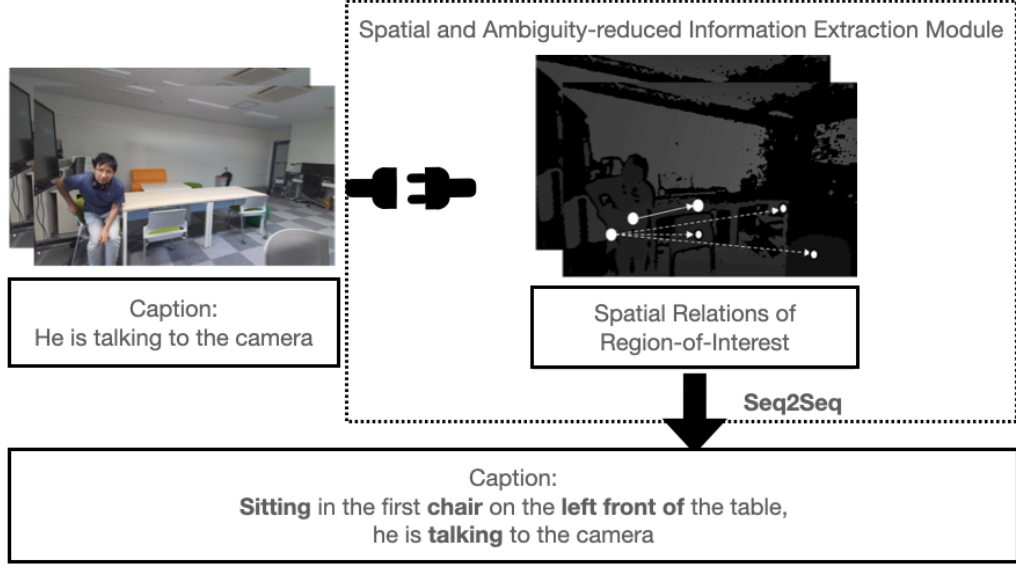


Fig. 4.1: Main scheme of the system

We model the constitute of S_j as Equation 4.2, where S_j^s , S_j^a , S_j^o represent space information, action information, and object information.

$$\{(t_j^s, t_j^e, S_j)\}_{j=1}^{N_{set}} = f_l(f_e(RGB), f_o(RGB - D)) \quad (4.1)$$

$$S_j = f(S_j^s, S_j^a, S_j^o) \quad (4.2)$$

4.2.1 Framework

The core insight of our system is to design a system that can incorporate spatial information to generate detailed and unambiguous captions. The entire pipeline is illustrated in Figure 4.2. The system comprises 3 main components: video caption generation, spatial and ambiguity-reduced information extraction, and sequence-to-sequence(seq2seq) text generation.

Feature Extractor. Humam Alwassel et al. [24], highlight that traditional feature extractors are often trained for trimmed action classification tasks. Consequently, to obtain features more suitable for untrimmed videos, we

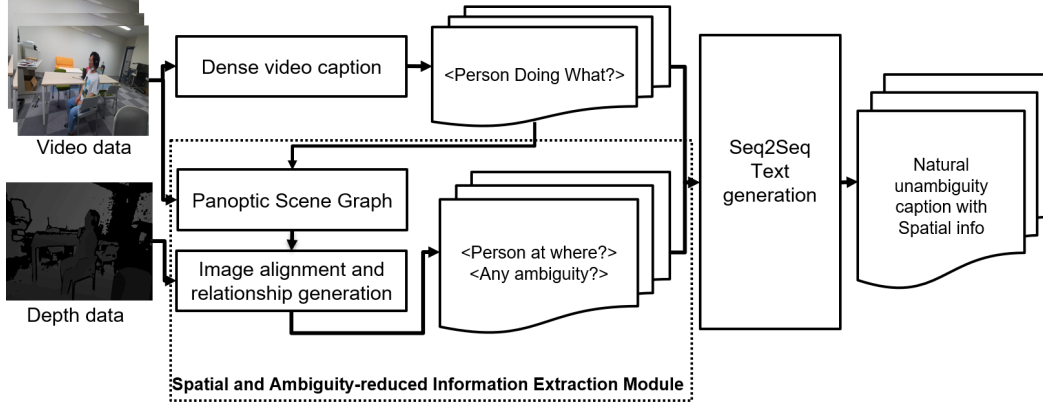


Fig. 4.2: Architecture of the System

have implemented the TSP feature extractor to extract the corresponding features, replacing the use of the C3D[25] features.

Video Captioning Generation. The Parallel Decoding architecture for dense video captioning is a novel design introduced by Teng Wang et al. [1]. In contrast to the traditional two-stage pipeline, they employ a transformer encoder and decoder architecture to simultaneously predict the location and caption, followed by merging the location and caption with the predicted scores.

The localization head is designed to perform box prediction and binary classification tasks using a multi-layer perceptron. The results will be saved as sets of tuples $\{t_j^s, t_j^e, c_i^{loc}\}_{j=1}^N$, where c_i^{loc} stand for the confidence of the event. The caption head is responsible for generating sentences that describe the event. As this head cannot determine the boundaries of the event, the deformable soft attention (DSA) and LSTM are used to generate a sentence $S_j^a = \{w_{j1}, \dots, w_{jM_j}\}$, where the w_{jt} means the word in the j proposal and M_j is the sentence length.

Spatial and Ambiguity-reduced Information Extraction. The module is shown in Figure 4.3. In order to extract spatial and unambiguous information from RGB-D video sequences and generate concise sentences S_j^s and S_j^o for descriptions, we utilize the previous output (t_j^s, t_j^e) as the input and consider the RGB and depth images with a time stamp computed as $(t_j^s + t_j^e)/2$ as the key frames for the j^{th} sentence.

We implement a Mask2Former to obtain a more accurate image segmentation result and based on that, we implement a fine-tuned OpenPSG [11] module to generate all the relationships of the visual contents.

To extract spatial and unambiguous information from keyframes, we adhere to the following design principles: 1. The main object is chosen from the three largest objects and centered. 2. Additional clarification is introduced only when ambiguity arises with other objects. We reuse the ‘human action’ class from OpenPSG, employ panoptic segmentation from Mask2Former, and align RGB-D images to directly generate information about: [human position related to the main object], [human interact with what] and [which object he interact with] based on the depth of the interest objects.

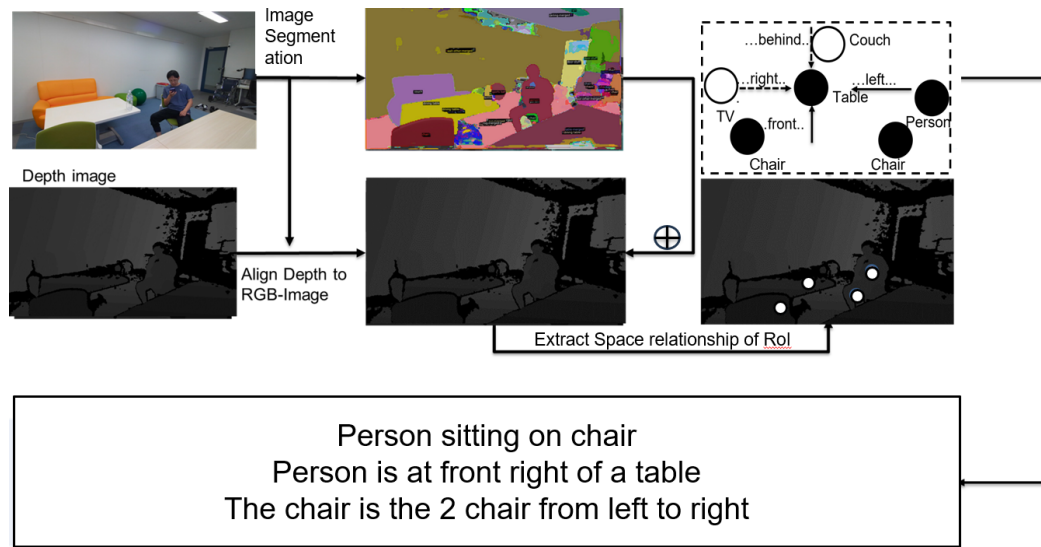


Fig. 4.3: Spatial and Ambiguity-reduced Information Extraction

Seq2Seq Text Generation. To generate natural sentences, we utilize a Large Language Model (LLM) to combine and summarize the short sentences. ChatGPT is a well-known model in the field of natural language processing and has produced excellent results for a variety of tasks. However, ChatGPT is not freely available. To make the system applicable to a wide range of applications, we implemented a fine-tuned ‘falcon-7b¹’ model that integrates all information into longer sentences with a more natural structure and human comprehension.

¹Falcon LLM <https://falconllm.tii.ae>

4.3 Experiment and Discussion

4.3.1 Experiment Setting

Dataset. There is currently no benchmark dataset that comprises RGB-D images of indoor scenes and annotations in the form of natural sentences describing human action. To address this gap, we have gathered a dataset consisting of 30 untrimmed, long videos for training and testing. Within the dataset, we have designed 3 different scenes, 4 position classes (front, back, left, right), and 2 subjects. Additionally, each video includes both unambiguous and ambiguous action classes.

Metrics. The method is evaluated in two aspects: 1. For location performance, the average precision and average recall across IOU (Intersection over Union) at [0.3, 0.5, 0.7, 0.9] are used. 2. For dense captioning performance, we calculate the BLEU[8], METEOR[7], and CIDEr[9] of the matched pairs between generated captions and ground with IOU thresholds of [0.3, 0.5, 0.7, 0.9].

4.3.2 Implementation Details

We collect the dataset with one Microsoft Azure Kinect. The video is 30 FPS with 720P. We implemented a TSP feature-updated PDVC and pre-trained it on ActivityNet dataset as the backbone. In the spatial and ambiguity-reduced information extraction module, we implement a Mask2Former model which is trained on the coco panoptic dataset for detection and segmentation of daily objects. We design 4 position relations (left, right, front, back) and 4 human-object interaction classes (walking on, standing on, lying on, sitting on) to generate based on OpenPSG.

4.3.3 Comparison Results

We provide qualitative comparison results and quantitative comparison results to the existing methods on localization performance and dense video caption performance.

Localization performance. To create a plug-and-play module aimed at improving the quality of captions generated by existing methods, we reuse the time intervals. As a result, we achieve the same precision and recall scores as the PDVC method, as demonstrated in the Table 4.1.

		
Ground Truth	He is sitting on the chair in the front-left of the table. The chair is the first one from the left.	He is standing on the floor behind the right side of the table, playing with a ball.
PDVC	He is seen speaking to the camera lead into he talk	He then put the ball down and down the ball
GVL	A man is talking to the camera	The man then begins to play the ball and then begins to play the ball
Ours	He is sitting in the front left-hand position of the table, with his chair being the first chair . He is facing the camera and engaging in a conversation.	He is standing on the floor to the right of the table, and he then puts a ball down and kicks it.

Fig. 4.4: Qualitative Comparison Results

Dense video caption performance. In Figure 4.4, we present the ground truth of our dataset in a video file, along with the corresponding caption results from existing methods(PDVC [1] and GVL[4]) and ours. In this qualitative comparison, our methods successfully add spatial information and eliminate ambiguity.

The quantitative comparison results are presented in Table 4.1. All metrics are evaluated across IOU at [0.3, 0.5, 0.7, 0.9] and averaged. Our system achieves significant improvements in METEOR, which provides a comprehensive evaluation considering word choice and order, as well as in BLEU4, which

measures text similarity based on the precision of 4-grams. It also reaches the same level as GVL in CIDEr, which assesses diversity and consensus in captions.

Tab. 4.1: Captioning and Localization performance

	<i>PDVC</i>	<i>GVL</i>	<i>Ours</i>
BLEU1	10.3	12.5	13.6
BLEU2	6.3	8.0	8.2
BLEU3	4.1	5.2	5.3
BLEU4	1.3	0.9	3.3
METEOR	4.7	5.8	8.4
CIDEr	4.5	12.2	12.2
Recall	55.1	40.9	55.1
Precision	26.4	34.3	26.4

In other words, our system excels in generating space-enhanced and unambiguous captions. Compared to existing methods, it is better suited for application scenarios that require space information and clarity in human action recognition.

4.4 Summary

The indoor human-centered application scenario, where the meaning of human action is related to the surrounding environment, requires spatial information enhanced captions to better record indoor human action. However, current dense video captioning methods lack a convenient way to incorporate spatial information. To address this limitation, we propose a plug-and-play module based on OpenPSG with an RGB-D camera. In our approach, the system extracts spatial information from RGB-D pairs using OpenPSG. This spatial information is then merged with the original captions through large language models (LLMs) to generate the final spatial information enhanced captions, making them more natural and fluent.

We collected a dataset to evaluate our method and compared it with other popular methods, PDVC and GVL, in terms of caption performance and localization performance. The results show that our method achieves a BLEU-4 score of **3.3**, a METEOR score of **8.4**, a CIDEr score of **12.2**, a Recall of **55.1**,

and a Precision of 26.4. These results demonstrate that our system excels in generating spatial information enhanced and unambiguous captions. Compared to existing methods, it is better suited for application scenarios that require spatial information and clarity in human action recognition.

4.5 Main Limitations of This Method and Future Solutions in Next Chapter

The method proposed in this chapter provides a convenient plug-and-play solution for existing systems, enhancing the quality of generated captions by incorporating the relationship between humans and objects within the field of view (FOV). However, the description of object order is highly dependent on the viewpoint, and it cannot capture information about objects outside the FOV.

In the next chapter, we propose a system based on Retrieval-Augmented Generation (RAG) that utilizes key objects—whether in or beyond the FOV—to describe human positions, thereby making the captions more readable and natural.

RAG-enhanced Dense Video Caption for Human Indoor Actions: Disambiguating Caption Using Spatial Information Beyond FOV Constraints

5.1 Background

Video medium serves as an important input in many vision-based security monitoring and smart home applications. However, the video stream is very difficult to search and store so the large volume of video content demands the need for automated methods to summarize and compactly represent the essential contents. [2] Fortunately, with the development of deep learning, the dense video caption, which is a technique that generates descriptive text for each event of the video, [1] provides a convenient way to create content summaries and record the content in the video.

However, current dense video captioning methods are not designed for recording indoor human action. The generated captions often include ambiguity and have not been paid enough attention to detect the relationship between the person and their surrounding environment. These limitations can be critical in specific scenarios, such as generating activity log files for indoor security monitoring, where the focus is on tracking abnormalities,

and different indoor positions can lead to diverse interpretations of the same activity. For instance, in indoor security systems, ambiguous captions might fail to distinguish between harmless activities such as a person walking and potentially concerning behaviors such as unauthorized access attempts. In Figure 5.1, the problem is illustrated where a person in a video performs a single action across three different positions, each with various environmental information. Current existing methods such as PDVC and GVL, popular in the dense video field and commonly trained on the ActivityNet dataset, focus solely on human action categories. Consequently, they often fail to distinguish between the three events, resulting in ambiguous captions.

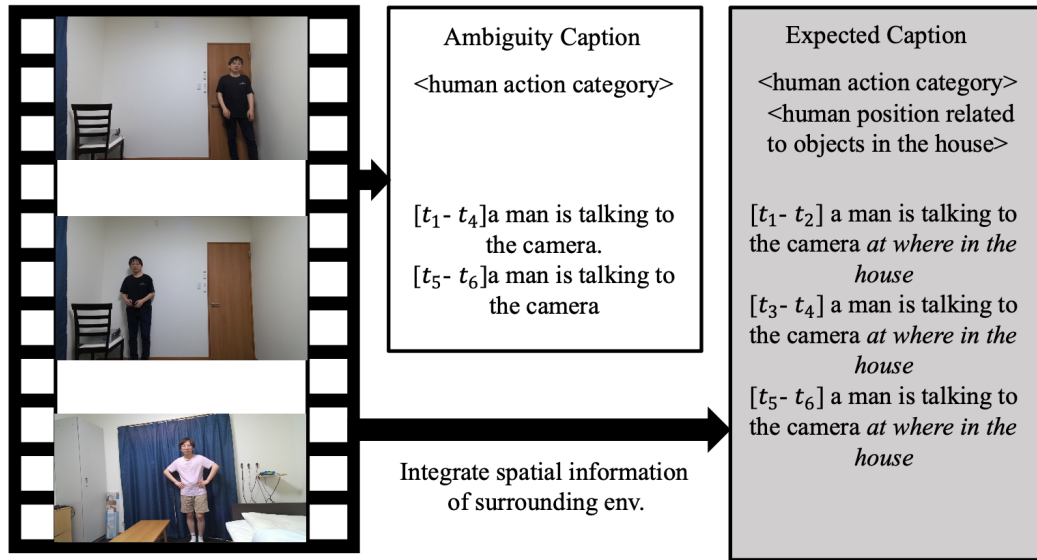


Fig. 5.1: Target problem and expected result.

To update the quality of caption to reduce ambiguity, in this research, we utilize large datasets to train and construct large models, such as Vid2Seq, [3] which is pretrained on YT-Temporal-1B, [26] a collection of 20 million YouTube videos with English subtitles covering diverse topics. Training and adapting such end-to-end models to specific scenarios are challenging, as they require retraining the model to incorporate contextual knowledge. Tu et al. devised a textual-temporal attention (TTA) model that extends the conventional temporal attention model by adding another textual attention model to reduce the ambiguity in the video caption task. [26] Specifically, they extracted object information from video frames and used the labels of these objects to enhance the quality of the captions. Following their work, the SG-PDVC [27] that we proposed also utilizes object information detected

in the frames to reduce ambiguity, and more details were added for improved caption results. However, this type of model lacks knowledge of the surrounding environment owing to its limited field of view (FOV), making it difficult to extract the relationship between the person and their surroundings. As a result, it is not suitable for certain human-serving applications where understanding the spatial relationship between the person and their surroundings is crucial.

To create a more convenient system that can be easily adapted to various application scenarios and introduce the relationship between human actions and the surrounding environment, we designed a system based on the retrieval-augmented generation (RAG), which is able to combine retrieval and in-context learning (ICL) techniques to enhance the results of large language models (LLMs). This system incorporates information about surrounding objects in the house into the current captions to provide high-quality, natural descriptions of indoor human action. The architecture of our system can be simply divided into three parts: (1) 3D reconstruction and 3D object detection are used to build a knowledge base, which only needs to be run once when setting up the system (initial setup). (2) A spatial-information-enhanced query that integrates the spatial information of humans and the original captions is designed, allowing RAG to extract related information from the knowledge base. (3) A RAG-enhanced LLM module that fuses and summarizes all related information in query and context to output the spatial-information-enhanced caption for indoor human action is designed. It is convenient to adapt our system to different rooms, as the initial setup only needs to be run once. This effectively reduces the computational burden.

We evaluated our method on a custom video dataset collected from two houses with eight different camera positions. We compared our method with PDVC and GVL, two popular methods in the dense video captioning field, as well as a similar system, SG-PDVC using the same class and temporally sensitive pretraining (TSP) features. The evaluation metrics include Recall, Precision, BLEU, METEOR, ROUGE-L, and CIDEr. Our method outperforms the three compared methods in BLEU-3, METEOR, and CIDEr metrics, but shows a small decline in ROUGE-L metrics compared with GVL. For localization performance, our method achieves the highest Precision score among the compared methods but shows only a slight improvement in Recall, which

is still lower than GVL. These results demonstrate that our method effectively incorporates spatial information and reduces ambiguity for indoor human action caption applications. Our contributions are as follows:

1. We collected a dataset for dense video captioning focused on spatial information of indoor human action.
2. We designed an updated dense video captioning system to record indoor human action without FOV constraints, leveraging a knowledge base of objects in the surrounding environment.

5.2 Methods

We designed a system on the basis of the RAG framework to enhance the captions generated by the existing dense video caption system, aiming to create a specific system for better human indoor action recognition and recording. It can be divided into two parts: (1) the initial part for building a knowledge database for the LLM and (2) the caption generation part that will use the spatial-information-enhanced query and enriched context extracted by an RAG process to generate the final caption result. The whole system is shown in Figure 5.2.

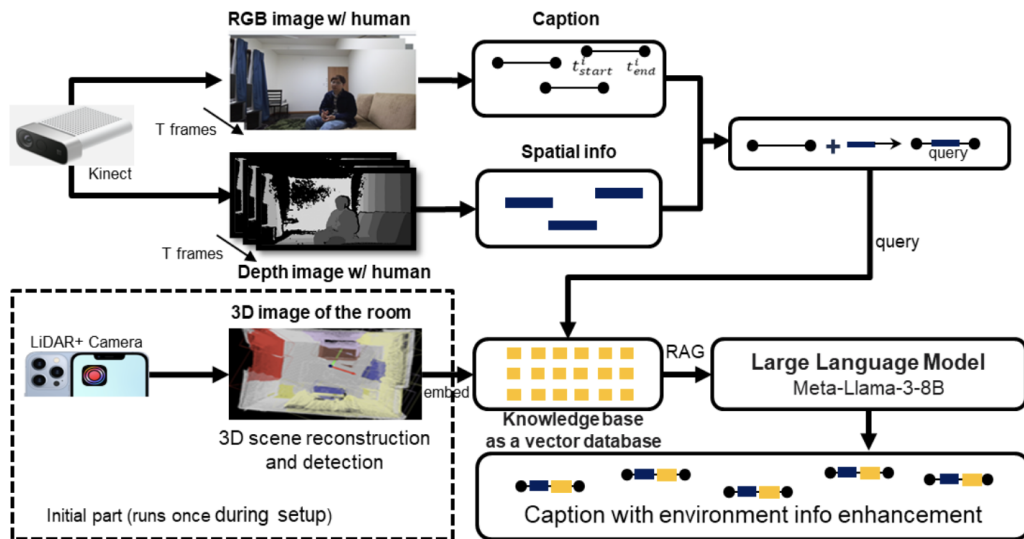


Fig. 5.2: Architecture of the whole system.

5.2.1 3D Reconstruction and 3D Object Detection to build a Knowledge Base (initial part)

In this step, our goal is to extract spatial information about the objects in the house, and this section can be divided into four sub-tasks: (1). 3D scene reconstruction for the house, (2). 3D object detection to extract the object information in the house, (3). calculation of the position of the camera and alignment of the scene and 3D object detection results to the camera coordinate, and (4). Building of the knowledge base for the RAG pipeline.

3D scene reconstruction. We use camera sensors and LiDAR sensors from the iPhone 14 pro as the device to generate the cloud points of the room by using the Scaniverse application. Theoretically, the cloud points representation of the room can be generated by any method.

3D object detection. We implemented the methods VoteNet, [28] which is trained on the ScanNet 3D dataset [29] to generate the 3D object detection results. The VoteNet is an end-to-end 3D object detection network that operates directly on point cloud data and has achieved state-of-the-art 3D detection on the two large real indoor 3D datasets, namely, ScanNet and SUN RGB-D. We take the 3D points cloud of the house as the representation of the 3D space and use it as the input for the VoteNet. The effectiveness of the VoteNet is evaluated on the Scannet with the Average Precision (AP) metric, which is calculated as the area under the Precision—Recall curve and gives a single score to summarize the precision–recall trade-off of a detector. The $AP@_\alpha$ stands for that the AP that is calculated for an Intersection over Union (IoU) threshold of α .

Coordinate transformation. In this step, we aim to determine the position of the camera and transform all contextual information to the camera's coordinates. To increase the accuracy of the registration results, we use global registration and a fine-tune registration process, the Iterative Closest Point algorithm (ICP). The whole pipeline is showed in the Figure 5.3.

In the global part, we utilize six corresponding points of the scene points cloud and the 3D fragment (points cloud) generated by the RGB-D image pairs taken by the depth camera to calculate the initial 3D transformation matrix. We use the reconstruction system pipeline in the open3D

library [30] [31] to build the points cloud to represent the 3D fragment obtained by the camera. Here, to provide a good initial 3D transformation matrix for the fine-tune registration process, we calculate the rotation and translation matrices through the following mathematical steps: (1). Calculate the covariance matrix which captures the relationship between the centered source and target points,. (2). perform Singular Value Decomposition (SVD) on the covariance matrix to find rotation-related matrices,. and (3). calculate the rotation matrix and translation matrices to construct the initial 3D transformation matrix. In the first step, we calculate the covariance matrix C based on the Equation 5.1, where p and q represent the selected source points and target points, respectively. P and Q are the matrices whose columns are the source points and target points, respectively. Q' is the matrix whose columns are the centered target points, and P' is a matrix whose columns are the centered source points. In the second step, we do perform an Singular Value Decomposition (SVD) on the covariance matrix C to calculate rotation-related matrices based on Equation 5.2, where U and V are the orthogonal matrices to represent rotations, and S is a diagonal matrix containing the singular values. In the third step, the initial 3D transformation matrix is constructed and calculated based on Equation 5.3, where R and t represent the rotation matrix and the translation vector, respectively.

$$C = Q'^T P' = \left(Q - \frac{1}{6} \sum_{i=1}^6 q_i \right)^T \left(P - \frac{1}{6} \sum_{i=1}^6 p_i \right) \quad (5.1)$$

$$C = USV^T \quad (5.2)$$

$$M = \begin{bmatrix} R & t \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} VU^T & \frac{1}{6} \sum_{i=1}^6 p_i - VU^T \frac{1}{6} \sum_{i=1}^6 q_i \\ 0 & 1 \end{bmatrix} \quad (5.3)$$

In the fine-tune registration process, we use the ICP algorithm to calculate the alignment between two point clouds. The algorithm is modeled as Equation 5.4.

$$\min_{\mathbf{T}} \sum_{i=1}^N \|\mathbf{T}p_i - q_i\|^2 \quad (5.4)$$

Here, T is the final fine-tuned 3D transformation matrix, and N is the number of points. All points in the two points clouds are utilized to determine the final 3D transformation matrix.

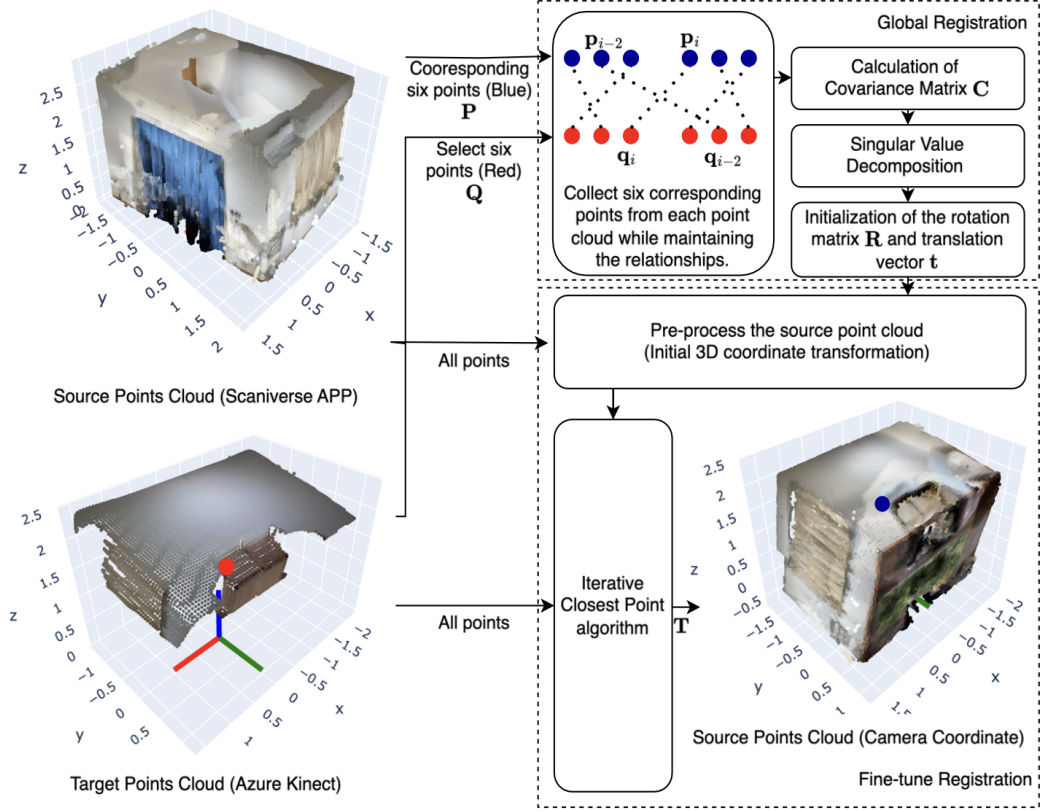


Fig. 5.3: Pipeline to align the source point cloud to the target point cloud

Building the knowledge base. In this step, we aim to build context information of the environment, including the camera position and the positions of objects in the house, within the camera's coordinates. This information will serve as a knowledge base that can be searched with a query. To build a knowledge base for the RAG process, there are always two sub-tasks: (1). Embedding the knowledge. and (2). Providing suitable distance criteria for the knowledge vectors. Here, to balance the accuracy and the model size, we utilize the gte-small model, [32] which is mainly based on the BERT framework and is trained on a large-scale corpus of relevant text pairs, covering a wide range of domains and scenarios as our embedding model. Then we utilize the FAISS and Euclidean distance, which is the distance between the ends of vectors to embed the all knowledge base as a vector database

5.2.2 Design of Spatial-enhanced Query and RAG-enhanced LLM module (caption part)

In this step, our goal is to design a module to use the spatial-enhanced query and extract related information from the knowledge base to output the spatial-enhanced caption for indoor human action.

Spatial- enhanced query. In this sub-task, we use the caption generated by PDVC as the original caption and extract the position of the human in the camera coordinate. The PDVC is a popular and efficient vision-based dense video caption method. In our work, the PDVC we re-implemented is trained with ActivityNet [6] and the Temporally-Sensitive Pretraining [24] feature, which improves the temporal sensitivity and is more suitable for the Dense Video Captioning task than the traditional feature, such as C3D. The original caption is representation as $\{s_j, t_j^s, t_j^e, c_i^{loc}\}_{j=1}^N$, where t_j^s and t_j^e represent the start time and end time of the j event, respectively. And the s_j, c_i^{loc} represents the caption and confidence of the j the event, where $s_j = \{w_j1, w_j2, \dots, w_jM_j\}$. w_jt means the j proposal and M_j is the sentence length.

The position of the person is obtained in two parts: (1) train You Only Look Once (YOLO) v8 for human detection and (2) align the RGB result to the depth image and transform the coordinates to the camera coordinate.

YOLO v8 is the latest version of the YOLO [33] series and is popularly used for real-time object detection and classification in the computer version. In this work, we use the YOLO v8 as the backbone to detect the position of the person in pixel coordinates. The YOLO v8 is trained with the COCO8 dataset, which includes the “person” class. In our work, we aim to disambiguate the original captions with spatial information. Specifically, we compare the position of the human at the start and end times to decide whether the event should be divided. The schema can be represented by Equation 5.5

$$Event = \begin{cases} \{(s_j, t_j^s, t_j^c), (s_j, t_j^c, t_j^e)\} & \text{if } d_h > thr, \text{ where } t_j^c = \frac{1}{2}(t_j^s + t_j^e) \\ \{(s_j, t_j^s, t_j^e)\} & \text{if } d_h \leq thr \end{cases} \quad (5.5)$$

where t_j^c is the center time stamp, d_h is the distance of the two positions of the person, and thr represents the threshold value.

We utilize the trained YOLO v8 to output the detection results of the person in pixel coordinates, $[p_x, p_y]$. Then, we align the result of the RGB image to the depth image $[p_x, p_y, d]$, where d is the value of the pixel in the depth image. We use the following Equation 5.6 to map $[p_x, p_y, d]$ to the camera coordination.

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = d \cdot \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} p_x \\ p_y \\ 1 \end{bmatrix} \quad (5.6)$$

Here, $[x, y, z]^T$ represents the detection result of human in the camera coordinate. f_x and f_y represent the focal lengths in the x and y directions, respectively. And $[c_x, c_y]$ is the principal points (the pixel coordinates of the camera center).

LLM to output the final spatial-information-enhanced caption. In this subtask, we use the designed query and the extracted knowledge as the context for the LLM to generate the spatial- information-enhanced caption. Llama 3 is the latest version of the open-access LLM released by Meta, consisting of two architectures: the 8 billion (8B) and 70 billion (70B) parameter models. Considering the summary and reasoning capabilities of the model, as well as its size, we implement the Llama 3 8B model with a Pydantic parser to standardize the input and output. To improve the accuracy of the output, we design a two-subtask pipeline. In the first step, we utilize the query to extract information about the related objects in the house. In the second step, we input the human spatial enhanced query and context of the objects to the LLM to merge all the information and output the final spatial-information-enhanced natural caption that records what the person is doing at a given location and time period.

5.3 Experiment and Discussion

In our work, the 3D scene represented as a point cloud of the house can theoretically be captured by any device. Thus, to show the convenience, we utilize the LiDAR and camera sensors of iPhone 14 Pro and the Scaniverse application from the APP store to generate the 3D scene point cloud of the house.

We utilize Azure Kinect to capture the input video. Azure Kinect is a device that integrates an RGB camera, a spatial microphone array, an orientation sensor (IMU), and a depth sensor operating on time-of-flight technology. The pictures of the device, the sensors used, and the modes of the sensors are shown in Table 5.1. In the table, MJPG_720P means that we use Motion-JPEG as the video compression format with a 720p resolution for the color image. WFOV_2x2BINNED means that the device is used in Wide Field of View mode, where the camera combines adjacent sensor pixels into a bin, effectively reducing the resolution but increasing the field of view. We use the device to capture two types of data. The first type of data is RGB-D image pairs without any people. This type of data is used to generate the partial 3D scene fragment that is used to calculate the transformation matrix between the 3D scene point cloud of the house and the 3D scene fragment in order to determine the relationship between the camera coordinates and the world coordinates. The second type of data consists of RGB-D image pairs that include a person, which are used to design the spatial query.

Tab. 5.1: The setting up of Azure Kinect device camera

<i>Related Sensor</i>	<i>Mode and Parameter</i>
<i>RGB camera</i>	<i>MJPG_720P</i>
<i>Depth sensor</i>	<i>WFOV_2x2BINNED</i>

In the evaluation dataset, we use two indoor rooms as our test rooms. In each room, we set up four camera positions to capture the video data. In each video, the participant is instructed to stand in different positions in front of the camera. We present the layout of the two houses, generated by the 3D object detection results of the houses along with their corresponding 3D transformation matrices to the camera coordinate, and the videos captured

by Azure Kinect to two human participants to obtain the ground truth caption. We ask them to describe the positions and actions of the participant in the video. An example of the video and the generated 2D layout are shown in Figure 5.4 where the dotted line represents the position and FOV of the camera.

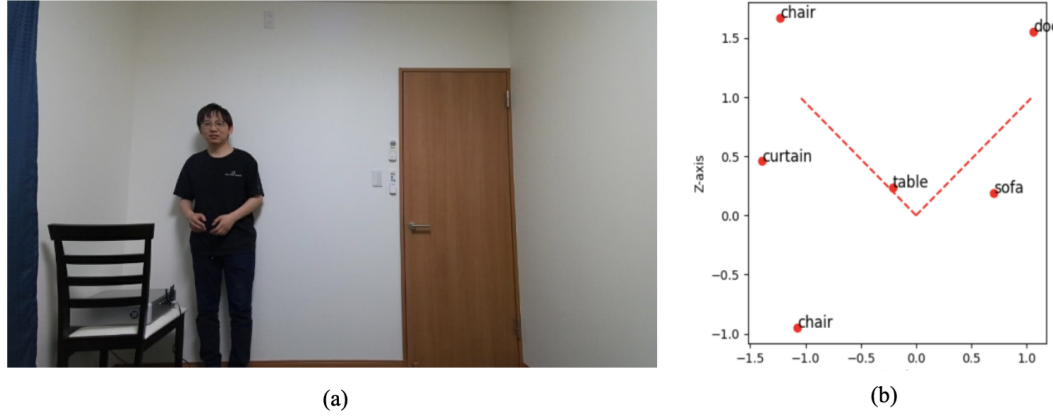


Fig. 5.4: (a) Test video and (b) 2D layout of the house with the corresponding camera position.

5.3.1 Knowledge Base of Surrounding Environment

We use the Scaniverse app to obtain the point cloud and use the VoteNet from Mdetect3D to obtain the 3D object detection result. We input the point cloud to the VoteNet, which is pretrained and evaluated on the ScanNet dataset that is a dataset including daily indoor objects. Here, we utilize the checkpoint weights where the AP@0.25 metric is 62.34 and the AP@0.5 metric is 40.08, which are accurate enough for our system. The detection results of the two rooms are shown in Figure 5.5.

Tab. 5.2: Quality of registration result

	<i>Room 1</i>	<i>Room 2</i>
Fitness(Avg.)	0.122726	0.115741
Inlier RMSE (Avg.)	0.0175551	0.0177646
Corresponding Set Size (Avg.)	8694.25	7818.75

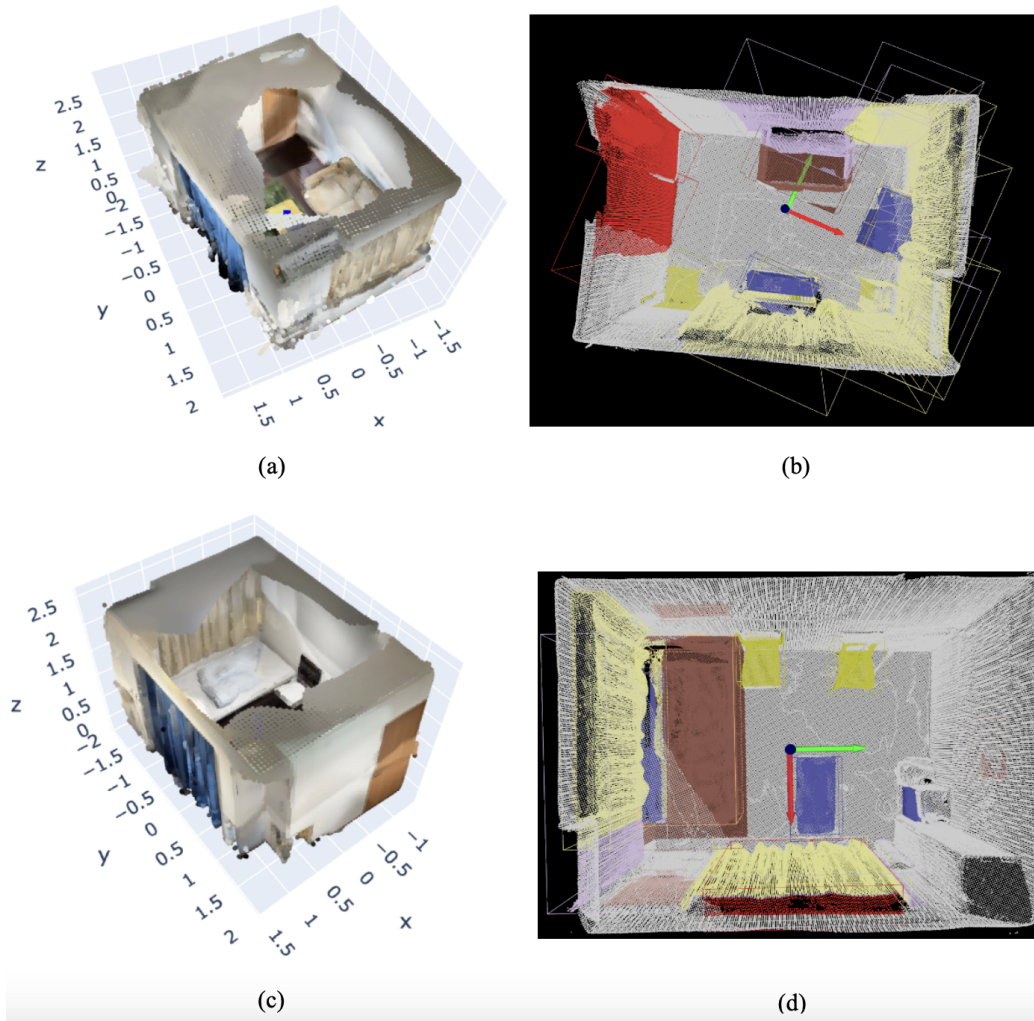


Fig. 5.5: (a–d) 3D scene reconstruction and detection results of the two rooms.

Next, we aim to find the camera's position and transform all the information into the camera's coordinate. The results of the ICP process are shown in Table 5.2 and Figure 5.6. The fitness and correspondence set size measure the alignment and matched points between the source point cloud and the target point cloud, respectively. The inlier root mean square error (inlier RMSE) measures the average distance error between corresponding points. We utilize the volume value and position value of the objects and the corresponding transformation matrix to build the knowledge base of the surrounding environment.

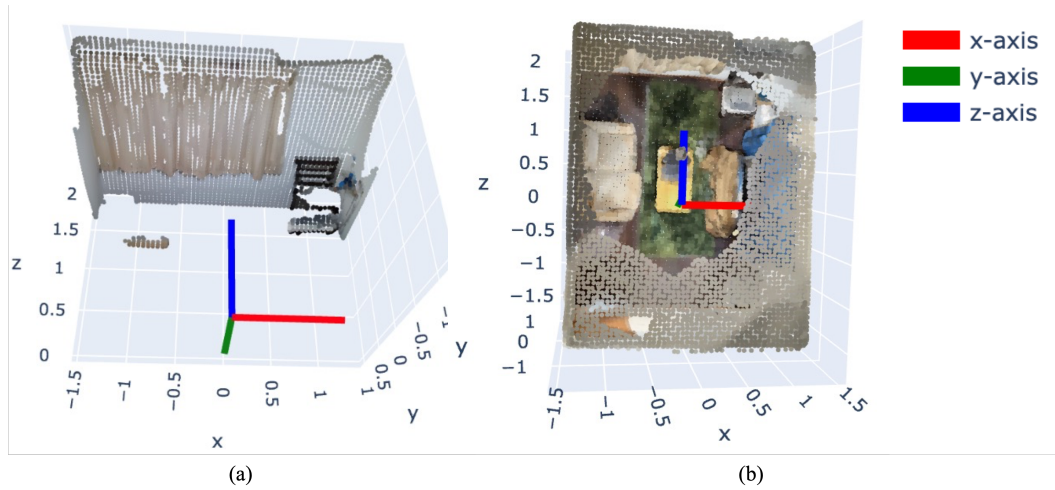


Fig. 5.6: (a) 3D fragment captured with Azure Kinect in the camera coordinate. (b) 3D scene transforms and registers to the corresponding camera coordinate

5.3.2 Spatial Information Enhanced Query

We use YOLO v8 to obtain the position of the person. We utilize the pretrained checkpoints of YOLO v8 on the COCO8 dataset, which achieve a mAP@50 metric of 0.872. Because the FOV of the depth image is larger than that of the RGB image, we align the depth image to the RGB image and map the human detection results from the RGB image to the depth image to find the corresponding depth value of the person. The result is shown in Figure 5.7. The pixel coordinates with depth values will be used to calculate the human position in the camera coordinate.

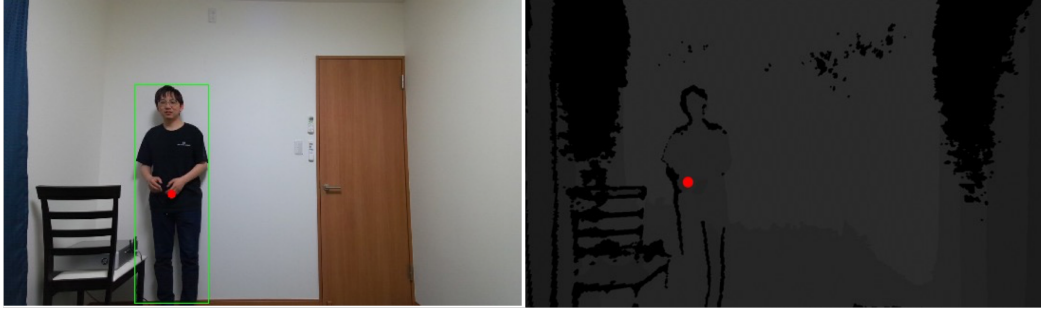


Fig. 5.7: Human detection result on RGB-D images pairs

5.3.3 RAG-enhanced LLM Module

We compare the effect of our system with PDVC, GVL, and the similar work published by Chen et al. [27] with TSP features in the same human action class in the custom dataset and evaluate our system on localization performance and dense video caption performance with qualitative and quantitative analyses.

Qualitative analyses. The results are shown in Figure 5.8. From the results, we can see that both PDVC and GVL do not pay attention to the surrounding environment and cannot include spatial information in the final caption. The SG-PDVC proposed by Chen et al.[27] can enhance the caption with spatial information, but it is limited by the FOV of the camera. It can only output the human position relative to the chair with words like "first" or "second," which are unnatural. The captions generated by our system can use the information of the two largest objects, the curtain and sofa, in the house beyond the FOV to provide spatial information. Our system also maintains consistency in direction by using the same camera coordination. Additionally, our system detects changes in position and redetermines the time boundary of the event, updating the caption to eliminate ambiguity.

Quantitative Analysis. All metrics are calculated across IoU at [0.3,0.5,0.7,0.9] and averaged as shown in Table 5.3. To evaluate the localization performance, Precision and Recall metrics are used. Precision is the fraction of relevant instances among the retrieved instances, and Recall is the fraction of relevant instances that are retrieved. Our method achieves the highest Precision score among the compared methods but shows only a slight improvement in Recall, which is still lower than GVL. For evaluating dense

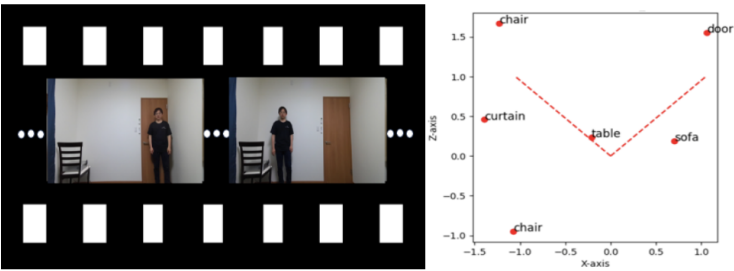
Evaluation Video and Extracted Knowledge	
Ground Truth	[0–20 s] A person stands at the right-front of the curtain while talking to the camera. [30–40 s] A person is positioned to the right-front of the curtain and to the left-front of the sofa while talking to the camera.
PDVC	[3–47 s] A woman is seen to talking to the camera.
GVL	[13–36 s] The man then walks up and down the room and speaks to the camera.
SG-PDVC	[3–47 s] He is talking to the camera on the right of the chair, which is the first chair.
Ours	[3–25 s] The person is seen speaking to the camera, who is situated at the right-front of both the curtain and the sofa. [25–47 s] The person is seen speaking to the camera; the person is at both the left-front of the curtain and the left-front of the sofa.

Fig. 5.8: Qualitative comparison result

video caption performance, the metrics METEOR, BLEU, CIDEr, and ROUGE-L are used. METEOR is based on the harmonic mean of unigram precision and recall, BLEU compares candidate sentences to one or more reference sentences using the geometric mean of n-gram precision, CIDEr measures the similarity between machine-generated captions and human-produced reference captions, typically used in image captioning tasks, and ROUGE-L is based on the longest common subsequence (LCS) between the model output and reference, often used in text summarization tasks. Our method achieves significant improvements in CIDEr metrics, slight improvements in METEOR and BLEU-3, but a small decline in ROUGE-L compared to with GVL. And Moreover, benefiting from the successful integration of the spatial information of humans and their surrounding environment, our method achieves higher scores in all four metrics compared to with PDVC, the backbone of our system, and the similar work SG-PDVC.

Tab. 5.3: Captioning and Localization performance

	<i>GVL</i>	<i>PDVC</i>	<i>SG-PDVC</i>	<i>Ours</i>
BLEU3	7.63	1.81	1.98	7.90
METEOR	6.82	2.32	3.18	7.56
ROUGE_L	11.0	6.18	6.13	8.79
CIDEr	10.8	3.05	3.79	14.9
Recall	82.8	75.0	75.0	77.3
Precision	35.6	33.4	33.4	36.8

In other words, our system successfully generates spatial-information-enhanced and unambiguous captions of indoor human actions. Compared with previous methods, our approach is more suitable for smart home or indoor security systems where both the position and action of a human are important.

5.4 Summary

In this chapter, we presented a system based on RAG and LLM to enhance the quality of captions generated by existing methods for smart home and indoor security systems. In summary, we utilized 3D reconstruction and detection results to build a knowledge base of the environment. We then designed an RAG pipeline to extract relevant information on the basis of

spatial cues, enhancing queries for an LLM. Our objective was to ensure that the final caption includes information on the person’s location and activity while minimizing ambiguity. We compared our system with PDVC and GVL, two popular dense video caption methods, as well as a similar method proposed in our previous work, using a custom dataset. According to the results, our method outperforms the three compared methods in BLEU-3, METEOR, CIDEr, and Precision metrics, but shows a small decline in ROUGE-L and Recall metrics compared with GVL. These results demonstrate that our system successfully generates spatial-information-enhanced and unambiguous captions of indoor human action.

5.5 Main Limitations of This Method and Future Solutions in Next Chapter

The method proposed in this chapter introduces a system based on Retrieval-Augmented Generation that utilizes main objects—whether within or beyond the field of view (FOV)—to describe human positions. However, this approach faces the following challenges:

- Identifying the main objects: In scenes with numerous objects, it becomes difficult, even for a human, to determine which objects should be highlighted for describing human positions.
- Visualizing the scene layout: The caption only provides spatial information for the main objects, making it challenging for the reader to imagine the entire scene.

To address these issues, we introduce a novel framework based on RAG. Unlike the method in this chapter, the new framework incorporates spatial information for all objects in the room, enabling the generation of more detailed and comprehensive captions.

Generating Spatially-Aware Dense Video Captions for Indoor Human Action with Position-Based Scene Knowledge

6.1 Background

The video medium serves as a rich source of content in many vision-based human-centric applications, such as security surveillance and a remote indoor caregiving systems. However, the large volume of video data makes it challenging to store and search, highlighting the need for automated methods to summarize and compactly represent the essential content of the videos. Fortunately, dense video captioning, which aims to generate descriptive captions and time periods for each target activity in the video, offers a promising solution to address this need.

However, most current dense video captioning methods are not designed to capture the spatial information of individuals and the surrounding environment[2, 1, 4], which is crucial in certain human-centric applications, such as indoor security surveillance and remote indoor caregiving systems. In these contexts, the focus is on detecting abnormal behavior, and the same behavior may have different interpretations due to varying spatial information within the scene. For example, in a remote indoor caregiving system, the abnormal behavior-"a man sleeping in front of the door" is completely

different from the normal behavior-"a man sleeping in bed". In an indoor security surveillance system, the normal behavior-"a man chopping something on the chopping board" is also different from the abnormal behavior-"a man chopping something behind the bed". There are also works, such as SI-PDVC [34], that recognize the importance of spatial information for dense video captioning in human-centric applications and focus on providing the enhanced caption with spatial information to reduce ambiguity. However, the captions generated by these methods are not yet detailed enough, and the problem remains far from solved. In Figure 6.1, we provide an example to illustrate the remaining issues. To fairly compare the performance of spatial information extraction, the video shows an individual performing the same action-talking to the camera-but in three different locations. The existing methods fail to detect the three distinct events, defined as target human actions along with its corresponding time period, and to provide natural and detailed captions.

To obtain high quality and detailed captions, we propose a system for enhancing dense video captioning of indoor human action with spatial awareness based on RAG [35], text classification, 3D scene reconstruction, and 3D object detection. The architecture of our system is shown in Figure 6.2. In summary, our system is divided into two main parts: 1. Initial part: building the knowledge base with 3D reconstruction and 3D object detection. This step only needs to be performed once during system setup. 2. Runtime part: we design a RAG framework that uses the information extracted from RGB-D pairs as prompts and the knowledge from the knowledge base as context to generate the final spatial information enhanced captions for indoor human action.

We collect a custom dataset for spatial information enhanced dense video captioning of indoor human action to evaluate the effectiveness of our system. We compare our system with PDVC [1], GVL [4], SG-PDVC [27], and SI-PDVC [34] in both quantitative and qualitative analyses. In quantitative analyses, we utilize METEOR, BLEU3, ROUGE_L, Recall and Precision as evaluation metrics. The evaluation results show that our system effectively generates human action recordings with spatial information, significantly reducing ambiguity in the captions.

Our contributions are as follows :

1. We collected a dataset for dense video captioning of indoor human action with spatial information, characterized by the spatial relationships of the individual and objects in the room.
2. We designed a system to enhance dense video captioning of indoor human action with spatial awareness, providing more detailed and accurate captions.
3. The proposed system provides a convenient way to incorporate the detailed spatial information about the entire scene and supports the reporting of objects moving within the camera's FOV.

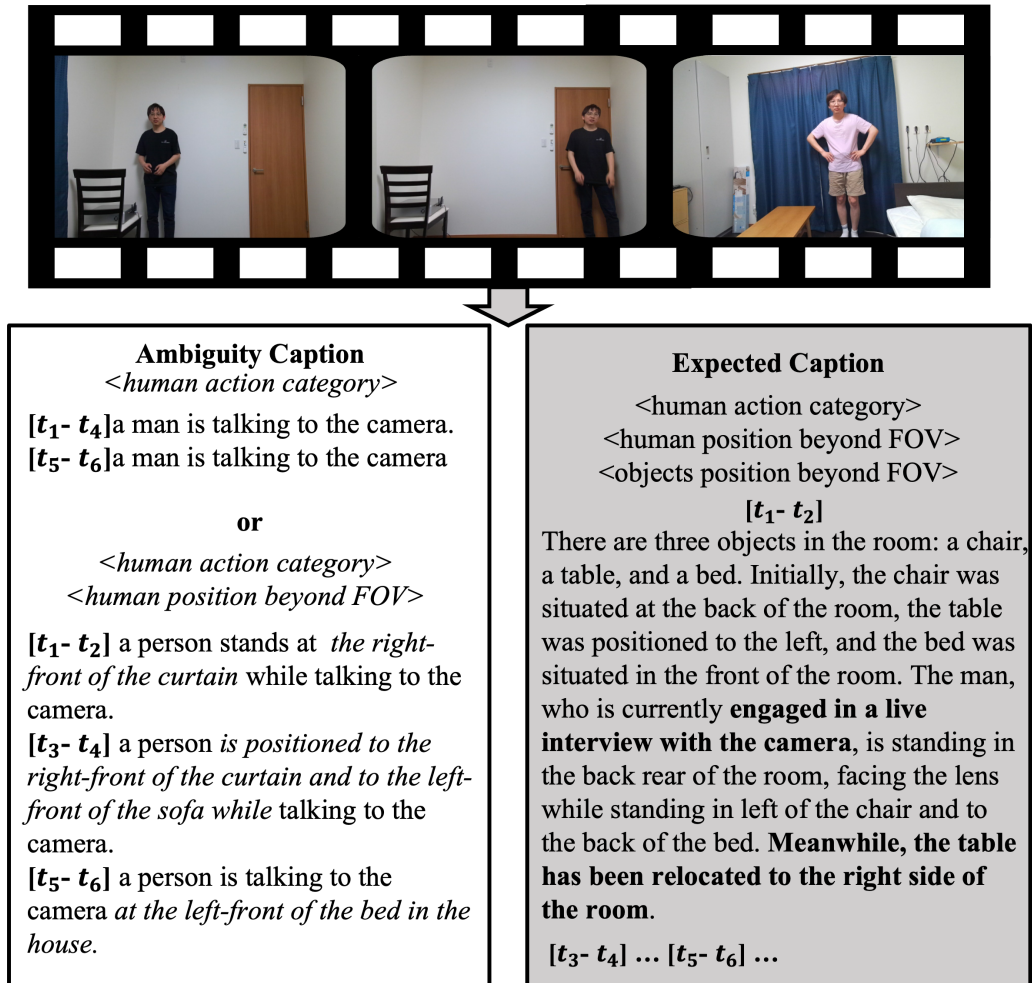


Fig. 6.1: Illustration of the ability of current methods to extract spatial information and our expected results

6.2 Method

The captions generated by existing methods lack detailed spatial information. In this research, we propose a system for enhancing dense video captioning of indoor human action by incorporating spatial information about an individual and objects in the environment. This generates more detailed and appropriate captions, which can be applied in human-centric applications such as vision-based security surveillance and remote indoor caregiving.

6.2.1 Framework

The entire architecture is shown in Figure 6.2 and is divided into three parts: 1. Designing the spatial enhanced query based on the individual’s action description, the position information of the individual and the objects in the FOV extracted from the RGB-D pairs. 2. Build a knowledge base using 3D reconstruction and 3D object detection. 3. Develop a RAG procedure to extract the object information of the room and fuse all information to generate the final caption with enhanced spatial details.

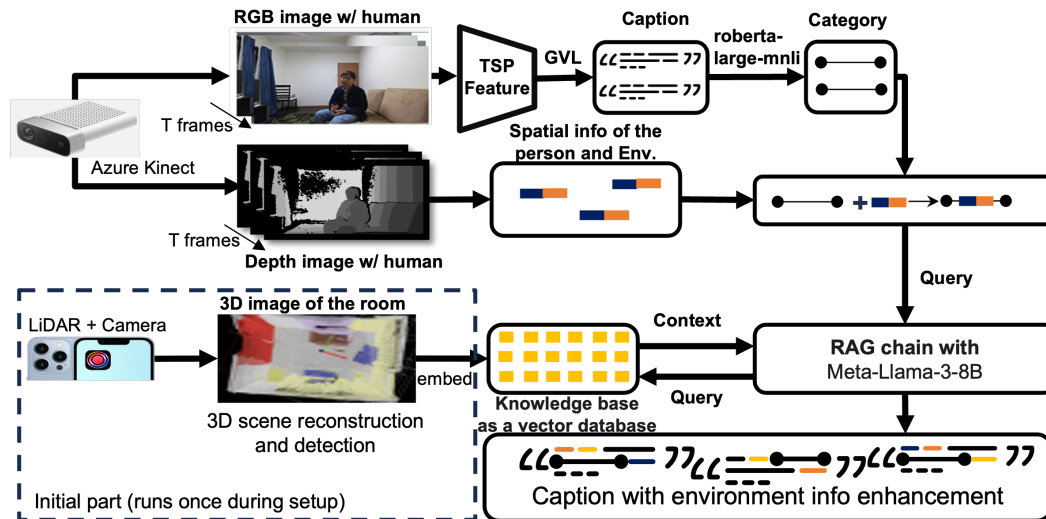


Fig. 6.2: Architecture of the System.

6.2.2 Spatial Enhanced Query Extraction

In this step, our goal is to extract the human action, the event duration, and the spatial information of both the human and the objects within the FOV, in order to obtain a query enhanced with spatial information for the LLMs.

In order to extract the human action and the event duration, we design a processing module based on the text classification and Intersection over Union(IoU) calculation to refine the original captions generated by the GVL[4] method, identifying the target human action and its corresponding time period. Specifically, the process consists of four steps. At first, we implement a Temporally-Sensitive Pretraining(TSP) extractor to obtain TSP features from the RGB video. TSP features are video features that significantly enhance temporal sensitivity and are commonly used to train models that also consider background clips and global video information. These features can be extracted using the pre-trained extractor provided in the work[24]. Compared to the C3D feature[25], the TSP feature is extracted from a pre-trained encoder using a novel supervised pretraining paradigm proposed in[24], which enhances temporal sensitivity and makes it more suitable for dense video captioning tasks. Next, the TSP features will be fed into the GVL model trained on ActivityNet captions database[6]. Third, to correct the sentences and remove the redundant information from the original caption, we implement a roberta-large-mnli model, which is the RoBERTa [36] large model fine-tuned on the Multi-Genre Natural Language Inference and belongs to text classification field. This model determines whether the target action is included in the caption. If the target human action is detected, the caption is updated using a fixed template: "The individual is { target action }". Fourth, we calculate the IoU based on the Equation 6.1 to exclude fully contained events under the conditions specified in Equation 6.2). The IoU is a metric commonly used to measure the overlap between two 2D bounding box, defined as the ratio of the intersection area to the union of the two boxes. In our case, IoU is used to compare 1D time periods to determine how much one is contained within the other, while excluding cases where the compared event is the same as itself. Here, s_1 , e_1 and L_1 represent the start time, end time and duration of the first event, while s_2 , e_2 and L_2 are the start time, end time and duration of the second event. If all conditions are all met, we will retain only the first event and exclude the second.

$$IoU = \frac{Intersection}{Union} \quad (6.1)$$

$$(s_1 - T \leq s_2) \wedge (e_2 \leq e_1 + T) \wedge ((L_1 > L_2)) \wedge (IoU \neq 1) \quad (6.2)$$

In order to extract the relevant spatial information, we implement a YOLO v8, a state-of-art objects detection solution, to detect the 2D pixel coordinates of the individual and objects within the FOV. The movement of objects within the FOV is determined by tracking changes in their 2D pixel coordinates over time. And then, we transform the spatial information to the camera coordinate and use the spatial information of the individual to update the events in the untrimmed video. Specifically, we implement a YOLO v8, which is trained on the COCO8 dataset, to effectively detect the individual and objects in the scene. The detected 2D pixel coordinates are then transformed to the 3D camera coordinates using the Equation 6.3, where the p_x, p_y are the 2D pixel coordinates, the f_x and f_y are the focal length of the camera, the c_x and c_y are principal point of the camera and the $[x, y, z]^T$ represent the 3D coordinate of the position in the camera coordinate. Given that different spatial information can lead to varying interpretations of human action, we update the events as shown in Equation 6.4) based on the consideration that if the position of an individual changes significantly, we should treat the events as a distinct one. Here, $t_j^c = \frac{1}{2}(t_j^s + t_j^e)$ represents the center timestamp of the time period, where the t_s is the start time and t_e is the end time. d_h is the distance between two position of the individual.

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = d \cdot \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix}^{-1} \begin{bmatrix} p_x \\ p_y \\ 1 \end{bmatrix} \quad (6.3)$$

$$Event = \begin{cases} \{(s_j, t_j^s, t_j^c), (s_j, t_j^c, t_j^e)\} & \text{if } d_h > thr \\ \{(s_j, t_j^s, t_j^e)\} & \text{if } d_h \leq thr \end{cases} \quad (6.4)$$

6.2.3 Knowledge Base Building

In this step, our goal is to build a knowledge base of the spatial information of the objects in the scene for the RAG procedure. Following the previous work, SI-PDVC, we use the Scaniverse app with an iPhone 14 Pro to generate the points cloud of the room as input of 3D detection module.

We utilize VoteNet[37], an end-to-end system for 3D object detection, as the 3D detection module to generate detection results that include the 3D positions of the objects in the world coordinates and their categories. The VoteNet implementation we utilized is trained and evaluated on the ScanNet 3D dataset[38], which is an instance-level indoor RGB-D dataset, achieving state-of-the-art detection results.

The results from VoteNet will be processed using the corresponding transformation matrix to convert the coordinates from world to camera coordinates. The 2D positions, excluding the height of the objects, along with their categories, will be stored in a text file as the database. Specifically, we obtain the corresponding transformation matrix through two sub-steps. First, we select six corresponding points from the point cloud generated by the Scaniverse app using an iPhone 14 Pro and the point cloud produced by the RGB-D pairs captured by the depth camera during the camera setup with Reconstruction System in the Open3D library. This allows us to derive the initial rotation and translation matrices based on Equation 6.5, Equation 6.6, Equation 6.7, Equation 6.8). Specifically, we calculate the covariance matrix of the two points clouds using Equation 6.5 and perform singular value decomposition on the covariance matrix to obtain rotation-related matrices as described in Equation 6.6. Here, the p and q represent the selected corresponding points, P , Q are the matrices whose columns consist of the selected points from two point clouds, and U , V are orthogonal matrices, while S represent the diagonal matrix. The initial rotation and transformation matrices are calculated based on Equation 6.7 and Equation 6.8), where R , t are the rotation matrix and transformation matrix, respectively. Second, we designed a fine-tuning process based on the ICP method, modeled in Equation 6.9, to obtain the corresponding transformation matrix. Here the T is the final fine-tuned 3D transformation matrix, and the N is the number of points in the points cloud. After obtaining the transformation matrix, we use the following template

to format the detection result. “This test is done in the {room name}. The camera is set in the {position number }. There are {number of detected objects} objects in the room. The first object is {category}, which is located {position of the object}. The second object is ...”

$$\mathbf{C} = \mathbf{Q}'^T \mathbf{P}' = \left(\mathbf{Q} - \frac{1}{6} \sum_{i=1}^6 \mathbf{q}_i \right)^T \left(\mathbf{P} - \frac{1}{6} \sum_{i=1}^6 \mathbf{p}_i \right) \quad (6.5)$$

$$\mathbf{U} \mathbf{S} \mathbf{V}^T = \mathbf{C} \quad (6.6)$$

$$\mathbf{R} = \mathbf{V} \mathbf{U}^T \quad (6.7)$$

$$\mathbf{t} = \frac{1}{6} \sum_{i=1}^6 \mathbf{p}_i - \mathbf{V} \mathbf{U}^T \frac{1}{6} \sum_{i=1}^6 \mathbf{q}_i \quad (6.8)$$

To build a knowledge base which can be used in a RAG procedure, we embed the knowledge with GTE-small model[32], which is based on the Bidirectional Encoder Representations from Transformers (BERT) framework and has been pre-trained on a large-scale corpus of relevant text pairs, covering a wide range of domains and scenarios and calculate the Euclidean distance metrics for the vectors in the knowledge base for searching in the knowledge base.

$$\min_{\mathbf{T}} \sum_{i=1}^N \|\mathbf{T} \mathbf{p}_i - \mathbf{q}_i\|^2 \quad (6.9)$$

6.2.4 RAG Procedure Designing

In this step, our goal is to generate the spatial enhanced caption for the untrimmed video based on the spatial information enhanced query and the extracted information from the knowledge base using a RAG procedure.

Considering the model’s capability and size, we implement a pre-trained Llama 3 8b model, a state-of-the-art open source LLM, as the backbone of the RAG procedure. The template we used to extract the information from the knowledge base is shown in the following text.

```

1 <s>[INST]
2 Please answer the following question and
3 output the result into a structured data.
4 Give me the result, NOT the code.
5 Only output one json dict, NO any otherwords.
6 NO Prompt word such as "output" in your answer,
7 just give me result.
8
9 here is the format instructions:
10 {format_instructions}
11
12 here is the context:
13 {context_obj}
14
15 here is the question:
16 {question}
17 [/INST]
18 </s>

```

And the question we used is as follows:

```

1 tell me all the information of the objects
2 when the test is done on the {bed room} and
3 the camera is on the {second} position.

```

The extracted information will be output as a dictionary, including the categories of the objects and their positions in the camera coordinates. We then regularize and compare the coordinates of the objects and the individual's coordinates from the query to format the spatial information using the following template: "The person is in {position} of the {nearest object}. The {object} is in the {position} area of the room..." Here, the "position" is expressed in natural language, such as "left" or "back". We define the positive direction of the x coordinate as "right" and the y coordinate as "front." The "nearest object" is determined by the distance between the regularized coordinates. The formatted spatial information and the query are input into the Llama3 8B model along with the following question to generate the captions.

```

1 proofreading the short sentences into long
2 sentence to make it Concise suitable for
3 a conference paper. Please include ALL the
4 information.

```

6.3 Experiment and Discussion

6.3.1 Experiment Setting

Device and Dataset. We utilize the Azure Kinect depth camera to obtain RGB-D pairs for our system. The RGB images are captured in Motion-JPEG format at a resolution of 720p and a frame rate of 30 FPS. Depth images are captured with “WFOV_2x2BINNED” mode of the depth sensor. To align the depth and color images, we employ the Open3D Azure Kinect MKV Reader, which accurately warps the depth image to match the RGB image.

To evaluate our system, we collected a dataset from three indoor rooms, with cameras positioned at ten different locations in the rooms. This dataset includes 34 RGB-D video recordings of human actions. Three annotators, aged 20 to 30, stood in the camera’s positions to generate ground truth captions, describing the human actions and the positions of the individual and objects in the whole scene. To ensure a fair comparison with SG-PDVC and SI-PDVC—both of which focus on a single action category-“talking to the camera”, we required the individual to perform this predefined action. The layouts arrangements of the recording rooms, illustrated in Figure 6.3, were designed to create a realistic indoor setting.

Metrics. We evaluated our system using qualitative and quantitative analyses. In the qualitative analysis, we provide examples of the ground truth captions alongside the captions generated by our method and the compared methods. For the quantitative analysis, we utilize the METEOR[7], BLEU3, ROUGE_L, recall, and precision metrics to compare our results with those from PDVC, GVL, SG-PDVC and SI-PDVC, demonstrating the effectiveness of our system in localizing events and generating captions. All metrics are calculated and averaged across IoU at [0.3,0.5,0.7,0.9].



Fig. 6.3: Layout of the rooms and the Furniture arrangement.

Tab. 6.1: Qualitative Comparison Results.

Videos	
Grand Truth	[0-46 s] the person is talking to the camera standing at the front area of the room which is the right of the bed . there are 7 objects in the room. In the left ,there are bed and curtain . In the right , there is a sofa and door . In the back , there are a chair and a curtain .
PDVC	[22-77 s] he is talking to the camera.
GVL	[0-51 s] a man is seen speaking to the camera and leads into a man speaking to the camera.
SG-PDVC	[22-77 s] he is at the front right of the chair , standing. He is talking to the camera
SI-PDVC	[22-77 s] He is talking to the camera from his position at both the right-front of the table and the left-front of the sofa .
Ours	[0-51 s]A person is talking to camera. In the scene, the person is situated in the right-front of a bed , where a chair has been moved . There are ten objects present in the room, including a door located in the right-front area, a table situated in the left area, a bed positioned in the left area, and a window also in the left area. Additionally, there are curtains (curtain_1 and curtain_2) in the right-back area, a garbagebin in the same region, and a sofa placed in the right area.

6.3.2 Comparison Results

We present both qualitative and quantitative comparison results against existing methods on localization and captioning performance.

Qualitative analysis. We present qualitative results comparing ours with others and the ground truth captions in Table 6.1. The ground truth captions describe a event with a person performing the behavior. The PDVC and GVL methods, focused on detecting human action, fail to incorporate the spatial information of individuals and objects, resulting in ambiguous captions. In addition, SG-PDVC and SI-PDVC lacks detailed spatial information and does

not address moved objects. In contrast, our method generates accurate and detailed captions that enhance spatial information, incorporating details beyond the camera’s FOV and accounting for moved objects, making it more suitable for human-centric applications.

Quantitative analysis. Our system is characterized by its ability to fuse spatial information from the entire room, enabling the generation of detailed and accurate indoor human action recordings. Currently, no existing system can provide such a wide range of spatial information in their captions. We compare our system with the most similar and widely used systems, PDVC and GVL, both state-of-the-art methods in the dense video captioning field that focus on human actions and do not use Llama 3. SG-PDVC and SI-PDVC are spatial information enhanced dense video captioning system and focus on human actions and positions within the FOV.

To evaluate the captioning results, we employ popular natural language processing metrics—BLEU3, METEOR and ROUGE_L—where METEOR calculates the harmonic mean of unigram precision and recall, BLEU evaluates n-gram precision against reference sentences, and ROUGE-L measures the longest common subsequence between generated and reference captions. Additionally, we assess the quality of the localization of the generated events using recall and precision metrics, expressed as percentages. The results, shown in Table 6.2, demonstrate that our method achieves higher scores across all caption quality evaluation metrics, indicating that our system is an effective approach for generating spatial information-enhanced captions. In contrast, other methods struggle to effectively incorporate detailed spatial information into their captions. Regarding localization evaluation metrics, our system exhibits a slight decline in recall scores but achieves the highest precision, indicating that the proposed method effectively detects events in untrimmed videos.

In summary, our system effectively generates accurate and detailed spatial information enhanced captions for indoor human action, supporting downstream human-centric applications such as security surveillance and remote indoor caregiving systems.

Tab. 6.2: Captioning and Localization performance

	<i>PDVC</i>	<i>GVL</i>	<i>SG-PDVC</i>	<i>SI-PDVC</i>	<i>Ours</i>
BLEU3	0.213	1.03	0.77	1.40	9.49
METEOR	1.20	2.09	2.12	3.47	15.7
ROUGE_L	1.98	1.23	2.63	3.87	20.3
Recall	72.7	79.7	72.7	75.8	52.3
Precision	31.7	13.3	31.7	34.7	52.4

6.4 Summary

In this chapter, we present a system for enhancing dense video captioning of indoor human action with enhanced spatial awareness based on a RAG framework, designed to integrate scene knowledge and reduce ambiguity in the generated captions. Compared to existing methods, our approach incorporates comprehensive spatial details of the entire scene and supports the reporting of objects moving within the camera’s field of view, making the generated captions more suitable for human-centric applications. We evaluate our system on a custom dataset, comparing it with PDVC, GVL, SG-PDVC, and SI-PDVC through both quantitative and qualitative analyses. The results demonstrate that our system effectively produces detailed and accurate spatial information-enhanced captions.

Conclusion and Future Work

7.1 Conclusion

This thesis presents progressive frameworks for enhancing dense video captioning of indoor human action, addressing challenges of ambiguity and insufficient spatial details through a series of advancements. Each phase of the research builds upon its predecessor, forming a coherent trajectory of innovation:

Foundational Research - Spatial and Ambiguity-Reduced Captioning: The initial phase focused on developing a plug-and-play module to enhance the quality of captions generated by existing methods. By integrating spatial information extracted from depth and color images, the module ensures that captions incorporate key elements such as actions, locations, and object interactions while minimizing ambiguity. Experimental results on a custom dataset validated the module's effectiveness in integrating spatial context.

Extended Research - RAG and Knowledge Base Integration: Building on the foundational module, the second phase introduced a system leveraging Retrieval-Augmented Generation (RAG) and Large Language Models (LLMs). By constructing a 3D knowledge base from scene reconstruction and detection data, the system improved the contextual relevance of generated captions, particularly for smart home and indoor security applications. Comparative evaluations with PDVC, GVL, and earlier methods demonstrated superior performance in metrics such as BLEU-3, METEOR, and CIDEr, establishing the system's ability to generate detailed and unambiguous captions.

Innovative Framework - Comprehensive Spatial Awareness: In the final phase, the research extended the scope of spatial awareness to encompass

not only static elements but also moving objects within the scene. This RAG-based framework integrates detailed spatial information of the entire environment, ensuring captions are both contextually comprehensive and suitable for human-centric applications. Quantitative and qualitative comparisons with PDVC, GVL, SG-PDVC, and SI-PDVC further highlighted the system's superiority in producing spatially detailed captions.

Through these progressive developments, this thesis establishes a robust framework for generating spatially aware, unambiguous dense video captions for indoor human action.

7.2 Limitations and Future Work

While our framework demonstrates significant potential, it has certain limitations. One key constraint is that it only reports movements of objects within the camera's field of view, leaving the spatial information of objects outside this range outdated. Furthermore, the system is resource-intensive due to the complexity and number of integrated modules.

In the future, we aim to address these limitations in two main ways. First, we plan to develop a multi-camera system that provides comprehensive coverage of the entire room. This enhancement will enable continuous spatial updates for all objects, even when they move beyond the FOV of individual cameras, improving the accuracy and quality of captions in dynamic environments. Second, we will focus on optimizing the framework's efficiency by streamlining its modules and reducing computational overhead, making it more suitable for resource-constrained applications.

These advancements will further enhance the robustness and applicability of our framework in diverse real-world scenarios.

Bibliography

- [1]Teng Wang, Huicheng Zheng, Mingjing Yu, Qian Tian, and Haifeng Hu. “Event-centric hierarchical representation for dense video captioning”. In: *IEEE Transactions on Circuits and Systems for Video Technology* 31.5 (2020), pp. 1890–1900 (cit. on pp. 6, 23, 26, 29, 33, 51, 52).
- [2]Vladimir Iashin and Esa Rahtu. “Multi-modal dense video captioning”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*. 2020, pp. 958–959 (cit. on pp. 6, 23, 33, 51).
- [3]Antoine Yang, Arsha Nagrani, Paul Hongsuck Seo, et al. “Vid2seq: Large-scale pretraining of a visual language model for dense video captioning”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2023, pp. 10714–10726 (cit. on pp. 6, 34).
- [4]Teng Wang, Jinrui Zhang, Feng Zheng, et al. “Learning grounded vision-language representation for versatile understanding in untrimmed videos”. In: *arXiv preprint arXiv:2303.06378* (2023) (cit. on pp. 6, 29, 51, 52, 55).
- [5]Luowei Zhou, Chenliang Xu, and Jason Corso. “Towards automatic learning of procedures from web instructional videos”. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 32. 1. 2018 (cit. on pp. 7, 8).
- [6]Jingwen Wang, Wenhao Jiang, Lin Ma, Wei Liu, and Yong Xu. “Bidirectional attentive fusion with context gating for dense video captioning”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2018, pp. 7190–7198 (cit. on pp. 7, 8, 40, 55).
- [7]Satanjeev Banerjee and Alon Lavie. “METEOR: An automatic metric for MT evaluation with improved correlation with human judgments”. In: *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*. 2005, pp. 65–72 (cit. on pp. 7, 28, 60).
- [8]Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. “Bleu: a method for automatic evaluation of machine translation”. In: *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 2002, pp. 311–318 (cit. on pp. 7, 28).

- [9]Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. “Cider: Consensus-based image description evaluation”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2015, pp. 4566–4575 (cit. on pp. 7, 28).
- [10]Chin-Yew Lin. “Rouge: A package for automatic evaluation of summaries”. In: *Text summarization branches out*. 2004, pp. 74–81 (cit. on p. 7).
- [11]Jingkang Yang, Yi Zhe Ang, Zujin Guo, et al. “Panoptic scene graph generation”. In: *European Conference on Computer Vision*. Springer. 2022, pp. 178–196 (cit. on pp. 9, 27).
- [12]Zijian Zhou, Miaojing Shi, and Holger Caesar. “Hilo: Exploiting high low frequency relations for unbiased panoptic scene graph generation”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 21637–21648 (cit. on p. 9).
- [13]Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. “Masked-attention mask transformer for universal image segmentation”. In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2022, pp. 1290–1299 (cit. on p. 9).
- [14]Peter F Brown, Vincent J Della Pietra, Peter V Desouza, Jennifer C Lai, and Robert L Mercer. “Class-based n-gram models of natural language”. In: *Computational linguistics* 18.4 (1992), pp. 467–480 (cit. on p. 10).
- [15]Jacob Devlin. “Bert: Pre-training of deep bidirectional transformers for language understanding”. In: *arXiv preprint arXiv:1810.04805* (2018) (cit. on p. 11).
- [16]Long Ouyang, Jeffrey Wu, Xu Jiang, et al. “Training language models to follow instructions with human feedback”. In: *Advances in neural information processing systems* 35 (2022), pp. 27730–27744 (cit. on p. 12).
- [17]Gemini Team, Rohan Anil, Sebastian Borgeaud, et al. “Gemini: a family of highly capable multimodal models”. In: *arXiv preprint arXiv:2312.11805* (2023) (cit. on p. 12).
- [18]Badr AlKhamissi, Millicent Li, Asli Celikyilmaz, Mona Diab, and Marjan Ghazvininejad. “A review on language models as knowledge bases”. In: *arXiv preprint arXiv:2204.06031* (2022) (cit. on p. 12).
- [19]David Dale, Elena Voita, Loïc Barrault, and Marta R Costa-jussà. “Detecting and mitigating hallucinations in machine translation: Model internal workings alone do well, sentence similarity even better”. In: *arXiv preprint arXiv:2212.08597* (2022) (cit. on p. 12).

- [20]A Vaswani. “Attention is all you need”. In: *Advances in Neural Information Processing Systems* (2017) (cit. on pp. 17, 18).
- [21]Naresh Kumar and Nagarajan Sukavanam. “Motion trajectory for human action recognition using fourier temporal features of skeleton joints”. In: *Journal of Image and Graphics* 6.2 (2018), pp. 174–180 (cit. on p. 23).
- [22]Mostafa A Abdelrazik, Abdelhaliem Zekry, and Wael A Mohamed. “Efficient hybrid algorithm for human action recognition”. In: *Journal of Image and Graphics* 11.1 (2023), pp. 72–81 (cit. on p. 23).
- [23]Muhammad Hassan, Tasweer Ahmad, Nudrat Liaqat, et al. “A review on human actions recognition using vision based techniques”. In: *Journal of Image and Graphics* 2.1 (2014), pp. 28–32 (cit. on p. 23).
- [24]Humam Alwassel, Silvio Giancola, and Bernard Ghanem. “Tsp: Temporally-sensitive pretraining of video encoders for localization tasks”. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2021, pp. 3173–3183 (cit. on pp. 25, 40, 55).
- [25]Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. “Learning spatiotemporal features with 3d convolutional networks”. In: *Proceedings of the IEEE international conference on computer vision*. 2015, pp. 4489–4497 (cit. on pp. 26, 55).
- [26]Yunbin Tu, Chang Zhou, Junjun Guo, Shengxiang Gao, and Zhengtao Yu. “Enhancing the alignment between target words and corresponding frames for video captioning”. In: *Pattern Recognition* 111 (2021), p. 107702 (cit. on p. 34).
- [27]CHEN Bin, Yugo NAKAMURA, Shogo FUKUSHIMA, and Yutaka ARAKAWA. “A Space Information-Enhanced Dense Video Caption for Indoor Human Action Recognition”. In: *2024 8th International Conference on Robotics, Control and Automation (ICRCA)*. IEEE. 2024, pp. 423–427 (cit. on pp. 34, 46, 52).
- [28]Charles R. Qi, Or Litany, Kaiming He, and Leonidas Guibas. “Deep Hough Voting for 3D Object Detection in Point Clouds”. In: *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. 2019, pp. 9276–9285 (cit. on p. 37).
- [29]Angela Dai, Angel X. Chang, Manolis Savva, et al. “ScanNet: Richly-Annotated 3D Reconstructions of Indoor Scenes”. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017, pp. 2432–2443 (cit. on p. 37).
- [30]Sungjoon Choi, Qian-Yi Zhou, and Vladlen Koltun. “Robust reconstruction of indoor scenes”. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2015, pp. 5556–5565 (cit. on p. 38).

- [31]Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. “Colored Point Cloud Registration Revisited”. In: *2017 IEEE International Conference on Computer Vision (ICCV)*. 2017, pp. 143–152 (cit. on p. 38).
- [32]Zehan Li, Xin Zhang, Yanzhao Zhang, et al. “Towards general text embeddings with multi-stage contrastive learning”. In: *arXiv preprint arXiv:2308.03281* (2023) (cit. on pp. 39, 58).
- [33]Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. “You Only Look Once: Unified, Real-Time Object Detection”. In: *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 2016, pp. 779–788 (cit. on p. 40).
- [34]Bin Chen, Yugo Nakamura, Shogo Fukushima, and Yutaka Arakawa. “Retrieval-augmented-generation-enhanced Dense Video Caption for Human Indoor Activities: Disambiguating Caption Using Spatial Information Beyond Field of View Constraints”. In: *Sensors and Materials* 36.10 (2024), pp. 4549–4566 (cit. on p. 52).
- [35]Patrick Lewis, Ethan Perez, Aleksandra Piktus, et al. “Retrieval-augmented generation for knowledge-intensive nlp tasks”. In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 9459–9474 (cit. on p. 52).
- [36]Yinhan Liu. “Roberta: A robustly optimized bert pretraining approach”. In: *arXiv preprint arXiv:1907.11692* 364 (2019) (cit. on p. 55).
- [37]Zhipeng Ding, Xu Han, and Marc Niethammer. “VoteNet: A deep learning label fusion method for multi-atlas segmentation”. In: *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer. 2019, pp. 202–210 (cit. on p. 57).
- [38]Angela Dai, Angel X Chang, Manolis Savva, et al. “Scannet: Richly-annotated 3d reconstructions of indoor scenes”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2017, pp. 5828–5839 (cit. on p. 57).

Appendix

Publish Works

Journal

1. B. Chen, Y. NAKAMURA, S. Fukushima and A. Yutaka, "Retrieval-augmented-generation-enhanced Dense Video Caption for Human Indoor Activities: Disambiguating Caption Using Spatial Information Beyond Field of View Constraints". In: *Sensors and Materials* 36 (2024), pp. 4549-4566.

Conference

1. B. Chen, Y. NAKAMURA, S. Fukushima and A. Yutaka, "A Space Information-Enhanced Dense Video Caption for Indoor Human Action Recognition". In: *Proc. 8th International Conference on Robotics, Control and Automation (ICRCA)*. Shanghai, China, 2024.
2. B. Chen, Y. NAKAMURA, S. Fukushima and A. Yutaka, "Generating Spatially-Aware Dense Video Captions for Indoor Human Behavior Analysis with Position-Based Scene Knowledge". In: *2024 international conference on intelligent computing and its emerging application (ICEA)*. Tokyo, Japan, 2024.

