2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

DailySTR: A Daily Human Activity Pattern Recognition Dataset for Spatio-temporal Reasoning

Yue Qiu, Shusaku Egami, Ken Fukuda, Natsuki Miyata, Takuma Yagi, Kensho Hara, Kenji Iwata, Ryusuke Sagawa*

***Abstract*— Recognizing daily human activities is essential for domestic robots to assist humans effectively in indoor environments. These activities typically involve sequences of interactions between humans and objects across different locations and times within a household. Identifying these events and understanding their temporal and spatial relationships is crucial for accurately modeling human behavior patterns. However, most current methods and datasets for human activity recognition focus on identifying singular events at specific moments and locations, neglecting the complexity of activities that span multiple times and places. To address this gap, we collected data on human activity patterns over a single day through crowdsourcing. Based on this, we introduce a novel synthetic video question-answering dataset. Our proposed dataset includes videos of daily activities accompanied by question-answer pairs that require models to reason about sequences of activities in both time and space. We evaluated state-of-the-art methods against our dataset, highlighting their limitations in handling the intricate spatio-temporal dynamics of human activity sequences. To improve upon these methods, we propose a two-stage model. The proposed model initially decodes the detailed content of individual videos using a transformer-based approach, then employs LLMs for advanced spatio-temporal reasoning across multiple videos. We hope our research provides valuable benchmarks and insights, paving the way for advancements in the recognition of daily human activity patterns.**

I. INTRODUCTION

Recognizing daily human activity patterns within indoor environments is essential for various applications aimed at understanding human behavior and providing appropriate assistance. Daily human activity patterns encompass a multitude of individual events, such as "walking to the toilet" or engaging in object interactions like "washing dishes". These events occur sequentially in various locations within a home. While recognizing these individual events is crucial, gaining a holistic understanding of the sequence and spatio-temporal relationships of these activities is equally important. This comprehensive understanding is beneficial for identifying and analyzing individual behavioral patterns, detecting anomalies, inferring unseen events, and predicting future activities. In the realm of robotics, this knowledge is instrumental for assessing users' physical and mental well-being, and providing personalized services. For example, robots could assess if an elderly person is unusually slow in morning routines and accordingly schedule reminders or provide timely assistance, such as prompting for medications during extended bathroom use or aiding in meal preparations when delays occur.

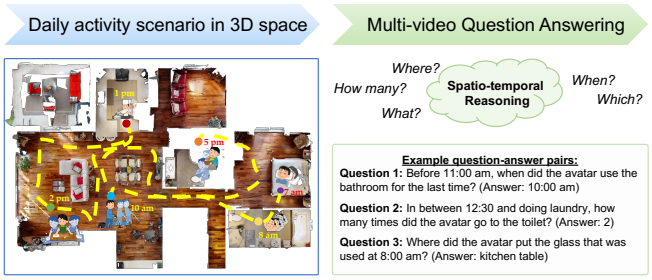

Fig. 1. Illustration of the DailySTR dataset: Each instance in the DailySTR dataset consists of video sequences depicting a full day's activities, accompanied by question-answer pairs designed for assessing spatio-temporal reasoning.

In the recognition of daily human activity patterns, two fundamental aspects are critical: the granular identification of individual events—including their timing, location, the objects involved, and the action type—and the spatio-temporal relationships that interlink these events, such as the sequence in which they occur, causal links, and the frequency of actions or object usage. Traditional video recognition datasets [1], [2], [3] have predominantly been limited to labeling actions within brief video clips, capturing only a momentary snapshot in one place. More contemporary datasets have begun to offer a more nuanced view, annotating dense video content with descriptive text [4], [5] or scene graphs [6], [7], [8] to classify objects and human-object interactions. Nonetheless, these datasets typically address the recognition of isolated events and do not account for the spatio-temporal correlations between multiple events. Consequently, current video recognition techniques [9], [10] have focused primarily on refining feature extraction for action classification. Some newer approaches [11], [12] have attempted spatio-temporal reasoning by synchronizing video with textual data, but they lack a comprehensive reasoning framework, which hampers their performance in interpreting complex activities.

To address the challenges of understanding daily human activity patterns, we have developed a comprehensive dataset, named Daily Human Activity Pattern Recognition Dataset for Spatio-temporal Reasoning (DailySTR), illustrated in Figure 1. This dataset is specifically designed to require spatio-temporal reasoning across multiple events throughout a day, thereby enriching the analysis of

*The authors are with Faculty of Artificial Intelligence Research Center, National Institute of Advanced Industrial Science and Technology (AIST), Japan {qiu.yue, s-egami, ken.fukuda, n.miyata, takuma.yagi, kensho.hara, kenji.iwata, ryusuke.sagawa}@aist.go.jp

daily human behavior. DailySTR was constructed using the VirtualHome-AIST indoor activity simulator [13] and comprises daily routines from 3,090 individuals, collected via crowdsourcing. Each dataset instance represents a chronological sequence of activities depicting a full day's scenario, animated into video sequences with VirtualHome-AIST.

To evaluate spatio-temporal reasoning, we created 64 unique question-answer (QA) templates. These templates automatically generate questions and answers that necessitate a comprehensive understanding of the activity patterns depicted in the videos. The dataset includes a total of 80,573 QA pairs. When we assessed two prior state-of-the-art methods using DailySTR, we observed that these methods struggled with reasoning-intensive questions, often defaulting to biases learned from the dataset. To overcome these shortcomings, we employed a two-stage approach: first, we used a transformer-based decoder to identify individual events in detail, and then we applied Large Language Models (LLMs) to perform explicit spatio-temporal reasoning. This method demonstrated a marked improvement over the existing methods, highlighting the benefits of direct reasoning with LLMs. Although our approach marks a significant step forward, there is still room for model enhancement. We believe that our dataset and the results of our experiments will be invaluable for future research aimed at improving the recognition of human daily activity patterns.

II. RELATED WORK

A. Daily Activity Recognition Dataset

Human activity recognition is vital for various robotic applications. While datasets like MMAct [1] capture single action labels from brief videos, other datasets such as Something-Something-v2 [2] and LEMMA [3] recognize human-object interactions by identifying both the action and the interacted object. Datasets such as Action Genome [6], Home Action Genome [7], and VirtualHome Action Genome [8] delve into parsing activities hierarchically and distinguishing complex activities, such as "cooking", from their component actions, like chopping vegetables. Contrarily, AGQA [4] and AGQA-DECOMP [14] datasets emphasize fine-grained recognition in indoor activity videos through detailed question-answer pairs that assess subjects, objects, actions, and their temporal dynamics. ANetQA [15] extends this to longer, untrimmed videos. Ego-centric video datasets like Ego4D [5] and EMQA [16] offer large-scale collections of daily activities with narratives, supporting various recognition tasks, including episodic memory, and identify activities and 3D locations from ego-centric videos.

Differing from most datasets that focus on single-moment video recognition, our proposed dataset leverages multiple videos from different times of the day to recognize daily activity patterns and facilitate spatio-temporal reasoning across multiple timeframes.

B. Activity Recognition Method

Transformers have become the leading model in activity recognition, with methods like TimeSformer [9] and Video SwinTransformer [10] adapting image-based transformers to video, showcasing the strengths of transformers in recognizing activities. Recent approaches like All-in-one [11] and Lavender [12] have introduced unified frameworks to merge video and language using transformers, facilitating extensive pre-training across various video-language datasets, leading to improved performance in tasks such as video question answering. For more detailed activity recognition, research has been exploring the recognition of actions and relationships [17], objects and locations [18], as well as the use of scene graphs [19] and the examination of hierarchical structures [20] to disentangle video content and boost model efficacy.

Despite the improvements in activity recognition, activity relationship recognition across time and space remains less discussed. Additionally, existing methods deal with reasoning in an implicit way, leaving the reasoning process less transparent. In contrast, we propose a method that addresses spatial and temporal reasoning across time and space, and utilizes LLMs to facilitate reasoning in a more explicit manner.

C. Indoor Scene and Human Activity Simulator

Indoor scene and human activity simulators facilitate the training and assessment of machine learning algorithms and robotics, offering a cost-effective alternative for data collection and preparation for real-world applications. These simulators range in capability, with some, like Replica [21] and 3D-Front [22], supporting the simulation of 3D environments and agent navigation. Others, such as AI2-THOR [23], iGibson 1.0 [24], and 2.0 [25], allow for interaction with objects within the scenes. A few, including VirtualHome [26] and Habitat 3.0 [27], incorporate human avatars, enabling the simulation of human movements and human-object interactions in 3D scenes. VirtualHome-AIST [13], [28] extended the original VirtualHome simulator by implementing 37 new actions, such as "wipe", "vacuum", and "squat" to enhance the range of supportable daily scenarios.

We utilize the VirtualHome-AIST simulator to construct our dataset due to its script-based interface, which allows for the simulation of a range of human activities via text-based activity descriptions.

III. DAILYSTR DATASET

Recognizing daily activity patterns involves observing human activities throughout the day, noting details such as locations, actions, and interactions with objects. However, existing datasets often focus on single-moment, single-location activities. To address this limitation, we introduce DailySTR, a dataset comprising observations of human activities across a day, along with spatio-temporal reasoning QA pairs for model evaluation. Constructed using the VirtualHome-AIST simulator [13], creating DailySTR involves three main steps (Figure 2). The following sections detail the construction process and dataset specifics.

Daily activity scenario collection: To capture natural human daily activity patterns, we employed crowd-sourcing services to gather typical one-day activity lists from 3,500

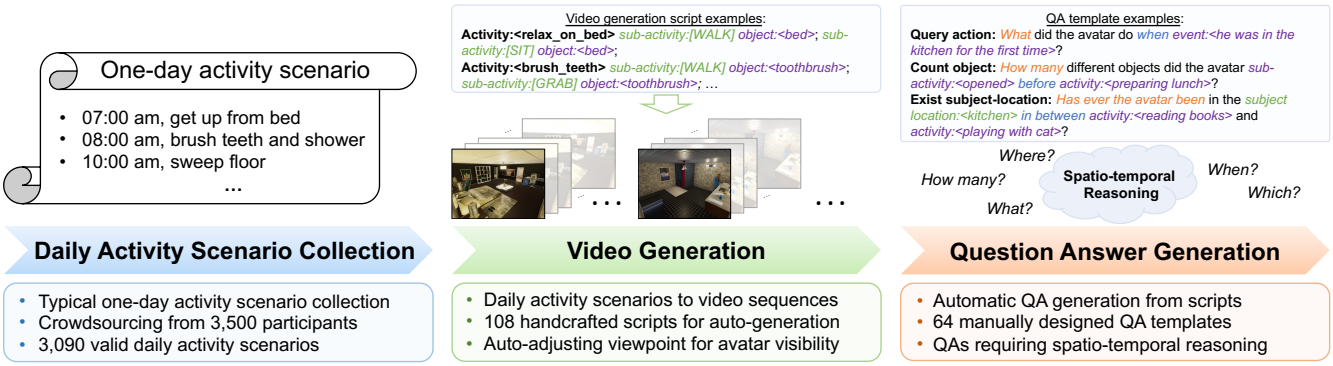

Fig. 2. DailySTR dataset generation process.

individuals. Additionally, we utilized the VirtualHome-AIST simulator [13], which allows the execution of 24 activity types, categorized into five groups: leisure, personal hygiene, work, eating and cooking, and house chores. Each participant was asked to select 12 time points from a total of 48 time points, spaced at 30-minute intervals throughout the day. Subsequently, they were tasked with choosing one typical activity from the 24 available options for each selected time point. Following the data collection phase, a separate group of experimenters assessed the validity of the 3,500 one-day activity lists. Ultimately, 3,090 of these lists were confirmed to represent eligible daily activity patterns.

Video generation: This step generates video sequences (12 videos per scenario) for each daily activity scenario collected by the previous step. Each daily activity scenario comprises 12 sequentially occurring activities chosen from 24 distinct activity types. To enhance dataset diversity, we designed 108 detailed activities for these activity types. For instance, the "prepare breakfast" activity includes variations such as "cook cold cereal", "cook hot cereal", and "cook fried bread", with one variation randomly selected for each scenario. Scripts listing sub-activities, including types and objects interacted with (Figure 2, middle), guide the VirtualHome-AIST simulator in producing videos from specific viewpoints. We handcrafted scripts for all 108 detailed activities. To ensure clear visibility of the avatar's actions, we have installed two viewpoints on the ceiling of each room, which automatically adjust when the avatar transitions to another room. Throughout the video collection process, we recorded critical data such as the locations of the avatars and objects to facilitate the automatic generation of QA pairs.

QA generation: The DailySTR dataset focuses on spatio-temporal reasoning in daily activities, detailing the recognition of objects, locations, activities, and their temporal orders, as well as spatial or visual similarities among them. We created QA templates that automatically generate QA pairs from scripts and recorded scenario data. These templates fall into three categories: "query", for specifics like action type, and timing; "count", for tallying actions or occurrences of objects/subjects/locations; and "query the existence", for checking the presence of actions/subjects/objects/locations. Temporal relationships are defined using terms like "before", "after", and "in-between". Our dataset includes 64 such QA templates (Figure 2, right). Each template randomly incorporates details from activities, objects, locations, and times during the automatic QA generation.

TABLE I
DAILYSTR DATASET STATISTICS.

# Scenario	# Vocabulary	# QAs		Per-scenario	
		Total	Unique	# Videos	Avg. length
3,090	360	80,573	43,570	12	342.2

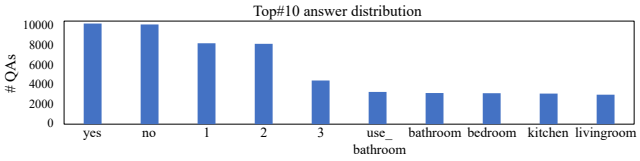

Fig. 3. Top#10 answer distribution of the DailySTR dataset.

Dataset statistics: The DailySTR dataset includes over 80K questions, more than half of which are unique (Table I). The average video length for each scenario is 342.2 seconds, exceeding the duration of most prior video QA datasets. The 3,090 scenarios, set across seven different virtual houses in the VirtualHome-AIST simulator, follow a training-to-testing ratio of 4:1. Figure 3 illustrates the frequency distribution of the top 10 answers from a pool of 180. Notably, in each question type, at least the top two answers occur with comparable frequency, thereby reducing the risk of models relying on dataset biases through memorization.

IV. METHOD

To accurately recognize detailed daily human activities across multiple videos, it is crucial to understand both the specifics of each video and the spatio-temporal relationships between them. Current methods like All-in-one [11] process videos holistically, not separating content or focusing solely on object extraction as TranSTR [18] does. Additionally, most methods employ a black box transformer structure for reasoning, making the reasoning process less transparent. To overcome these limitations, we propose a two-stage method (Figure 4). The first stage involves detailed recognition of activities, objects, and locations within videos using a decoder

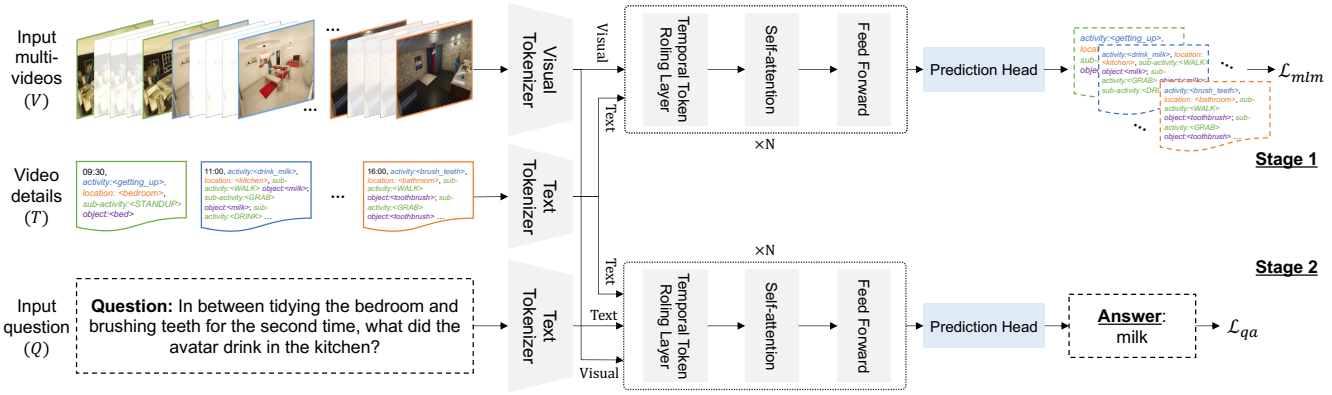

Fig. 4. Illustration of the proposed two-stage method. Stage 1 decodes the masked texts from the input of videos V and masked video details T, while Stage 2 predicts answers from an answer list based on the input of V, T, and questions Q.

structure. The second stage then performs spatio-temporal reasoning based on the detailed information identified. The following sections elaborate on these two stages.

Video detail decoder (Stage 1): The purpose of this stage is to extract detailed information from each video input, including the activities, sub-activities, objects interacted with by humans, and the humans' locations. We utilize a masked language model (MLM)-based sentence decoder to effectively capture all aforementioned details. The inputs for this stage are the videos, denoted as V, and texts, denoted as T, detailing the video content. In line with standard MLM procedures, we mask 15% of the words in each sentence for prediction. As depicted in Figure 4 (top), V and T are first encoded using visual and text tokenizers, respectively. Subsequently, an N-layer transformer structure with a prediction head is employed to predict the masked text segments. Vision Transformer (ViT) [29] and BERT [30] serve as visual and text tokenizers, respectively, with the All-in-one [11] model forming the transformer backbone. We used cross-entropy loss for masked word prediction. The loss is denoted as $\mathcal{L}_{mlm}$, and with W_{gt} representing the ground truth and W_p the predicted words, the loss is calculated as the following equation:

$$\mathcal{L}_{mlm} = CE(W_{gt}, W_p) \tag{1}$$

Spatio-temporal reasoning module (Stage 2): This stage involves predicting answers from a list of answer options using video V, video details T, and input questions Q. Time tag (e.g., 09:00) for each video is omitted in T during Stage 1 and is then utilized in Stage 2. The DailySTR dataset requires models to discern inter-video relationships, including temporal sequences, similarity in occurrences between objects, actions, locations, and the connection between questions and videos. To tackle these challenges, we implemented two methods. The first utilizes a transformer structure to integrate different inputs, as shown in Figure 4 (bottom). Specifically, V, T, and Q are tokenized using respective tokenizers, and then we employ the All-in-one architecture with N layers as the backbone to interlink these inputs. An answer A is deduced via a prediction head, and the network is refined using a cross-entropy loss, $\mathcal{L}_{qa}$, as defined in the subsequent equation.

$$\mathcal{L}_{qa} = CE(A_{gt}, A_p) \tag{2}$$

For training, we use the ground truth video details T, while for evaluation, we apply the predicted video details obtained from Stage 1.

We also incorporated GPT-4 [31] as a reasoning module. Leveraging the information contained in the video details, we employ GPT-4's zero-shot capabilities to predict answers based on T and Q. This approach not only yields predictions but also could be used to elucidate the reasoning process, thereby enhancing the transparency of how answers are derived.

V. EXPERIMENTS

We conducted experiments using the DailySTR dataset to evaluate the spatio-temporal reasoning capabilities of the previous method and our proposed approach. Initially, we introduce the two prior methods utilized for benchmarking purposes. Subsequently, we detail the training and evaluation protocols employed in our experiments. The following sections provide an ablation study of our model design, along with both quantitative and qualitative comparisons with existing methods.

A. Baseline Models

We selected two methods, All-in-one [11] and TranSTR [18], for comparative analysis. These methods have achieved state-of-the-art results on existing video question answering datasets, including MSR-VTT [32] and NExT-QA [33]. The All-in-one method employs a unified transformer structure to correlate video and text information, facilitating large-scale pretraining across a variety of video and language datasets. In contrast, TranSTR is designed to extract objects from videos and correlate them with video frames and questions within a transformer framework. This approach is particularly suited for answering questions from videos featuring multiple objects and events, a characteristic central to the proposed DailySTR dataset.

B. Implementation Details

For both two stages of our proposed method, we utilized the ViT (ViT-B-32) model [29], pre-trained on the ImageNet dataset [34], as the visual tokenizer, along with a pre-trained BERT model [30] for text tokenization. The transformer backbone was adapted from the official implementation of the All-in-one method [11], setting both layers and heads to 12. To standardize the setup for both the proposed and previous methods, we adjusted the image size to 224×224 and limited the number of video frames to four per video (randomly sampled). The initial learning rates were set at 10^{-4}, with models undergoing training for 20 epochs in each experiment.

C. Evaluation Metrics

For the evaluation of model performance in question-answering tasks, we employed the standard accuracy metric, encompassing both overall accuracy and detailed accuracies for each question type. The performance of all existing methods, as well as our proposed Transformer-based method (Our (Trans)) that utilized a transformer as the reasoning module in Stage 2, was assessed across the entire test split of the dataset for both the ablation study (Table II) and quantitative comparisons (Table III). For the GPT-4-based approach (Our (GPT-4)), we selected a random sample of 50 examples from the test split for the ablation study, and 350 examples for the quantitative comparison experiment.

TABLE II
ABLATION STUDY OF MODEL DESIGN.

Activity	Sub-activity	Object	Location	Accuracy	
				Trans	GPT-4
	✓	✓	✓	60.2	68.0
✓		✓	✓	61.5	64.0
✓	✓		✓	59.0	68.0
✓	✓	✓		61.1	66.0
✓	✓	✓	✓	62.9	**70.0**

D. Ablation Study

Table II presents our experimental evaluation of model designs, differentiating between input video details (the first four columns) and the choice of spatio-temporal reasoning modules (the last two columns). Our optimized model processes video details, encompassing the activity category, sub-activities, object categories involved, and the avatar's start and end locations during the activity. We applied a leave-one-out approach in Table II for evaluation.

The findings indicate that for both the transformer-based (Trans) and GPT-4-based (GPT-4) models, incorporating all four information types yielded the highest accuracy. For the transformer-based model, omitting one type of input resulted in a minor impact on performance. This method, which learns correlations among the four information types during Stage 2, may compensate for a missing input based on the remaining three. In GPT-4-based model, omitting sub-activities led to a relatively noticeable decline in performance. GPT-4's capacity for commonsense reasoning could assist the model in inferring missing information, such as activities, objects, and locations. However, deducing sub-activities from other information proved to be more challenging.

GPT-4 outperforms the transformer-based method, suggesting that its spatio-temporal compositional and commonsense reasoning capabilities are more adept for this task.

E. Quantitative Results

Table III presents the overall accuracies and detailed accuracies across various question types within the DailySTR dataset for the existing methods, All-in-one and TranSTR, alongside the proposed methods incorporating all four types of input information. All-in-one and TranSTR, originally designed for single video recognition, lack mechanisms to process time tags associated with each video. To address this issue, we integrated a 2-layer multilayer perceptron (MLP) to embed time information into All-in-one (serving as input for the Temporal Token Rolling Layer) and into TranSTR (as input for the Answer Decoder).

The results demonstrate that the proposed methods significantly outperform the existing ones by considerable margins, highlighting the critical role of extracting and analyzing detailed information from videos in the DailySTR dataset. Notably, the method utilizing GPT-4 for reasoning achieved the highest overall scores, showcasing its exceptional capability in spatio-temporal reasoning within daily activity contexts. Incorporating a time encoder led to improved overall accuracy for both existing methods, with notable enhancements in accuracy for "query time" questions, where direct time information is essential. Between the two existing methods, TranSTR, which explicitly handles disentangled information like objects in videos, delivered better outcomes than All-in-one, which does not explicitly separate information within videos.

Among the three types of questions, all methods showed higher accuracy for "Exist" questions, which have simpler "yes" or "no" answers, unlike "Query" questions that require specific labels and "Count" questions that demand counting skills. The two proposed methods excelled in most question types, particularly in "Query, Count, and Exist action" due to their ability to extract action labels from videos. However, extracting locations or objects from videos proved difficult, limiting these methods' accuracy in some areas. Notably, the method using GPT-4 outperformed others in "Query time" questions, showcasing superior temporal reasoning. All methods struggled with location-related questions, especially "Count subject location" highlighting the ongoing challenge of accurately recovering detailed location information from videos. Improving spatial reasoning could enhance performance on location-related questions.

F. Qualitative Results

Two scenarios from the DailySTR dataset are illustrated in Figure 5, each accompanied by two questions and answers evaluated across different methods. For conciseness, only

TABLE III

QUANTITATIVE RESULTS ON THE DAILYSTR DATASET.

Methods	Overall	Query					Count				Exist			
		action	time	subject location	object	object location	action	subject location	object	object location	action	subject location	object	object location
All-in-one [11]	49.7	37.6	13.2	45.1	29.4	41.3	37.1	20.0	43.7	50.6	67.9	82.6	70.5	71.4
All-in-one [11] + time	50.0	38.9	24.1	46.2	25.0	65.5	39.2	18.0	44.3	49.8	66.7	73.6	67.8	67.5
TranSTR [18]	51.4	43.7	8.3	46.3	**32.3**	51.7	39.4	15.6	50.9	53.1	73.3	75.9	78.1	79.7
TranSTR [18] + time	51.7	44.2	10.3	45.4	27.9	44.8	38.8	15.1	52.0	54.8	75.0	73.5	79.2	79.6
Our (Trans)	62.9	57.2	24.0	48.0	29.4	55.1	56.8	24.1	61.4	**63.4**	91.9	**83.0**	**90.3**	**91.0**
Our (GPT-4)	**70.8**	**72.5**	**82.1**	**51.4**	32.0	**68.0**	**65.2**	**25.0**	**75.0**	54.5	**92.3**	82.9	76.1	90.4

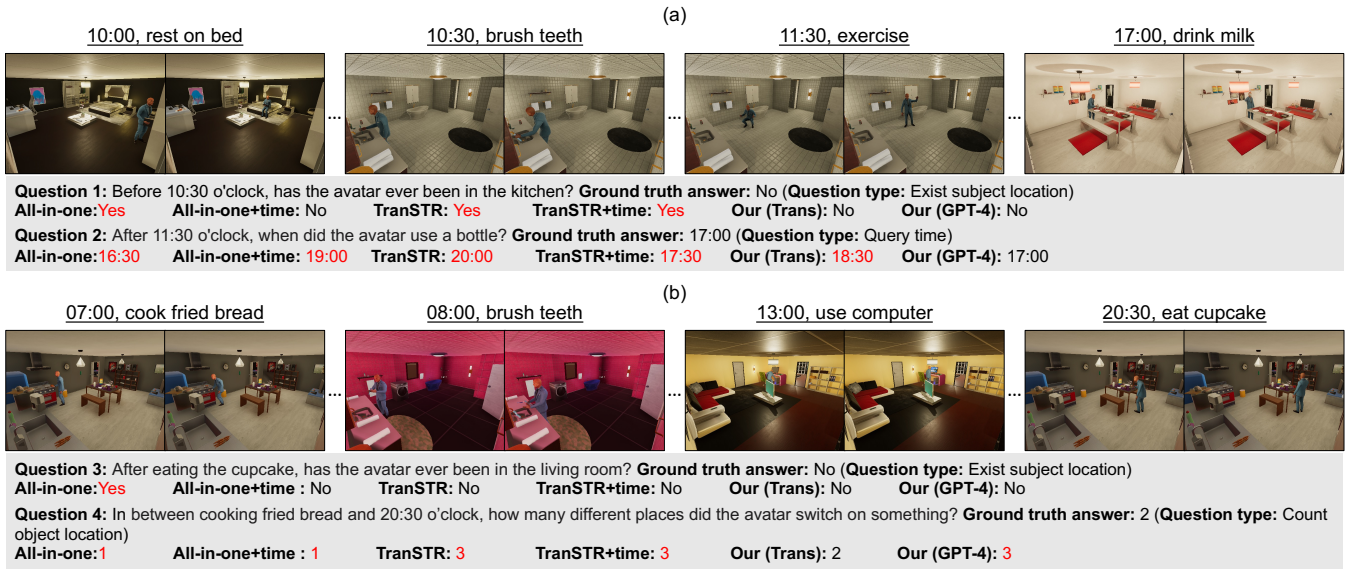

Fig. 5. Examples from the DailySTR dataset. Both scenarios (a) and (b) contain a video sequence of 12 videos each, with eight videos from each sequence omitted in the figure. For each single video represented, two frames, the time, and the activity are shown. Incorrect answers are marked in red.

four out of twelve activities are displayed for each scenario. Unlike existing methods, which have a minimum of two incorrect answers out of four, our two proposed methods each recorded only one incorrect prediction, demonstrating superior performance in these scenarios.

Specifically, for the "Exist subject location" questions (Questions 1 and 3), at least half of the methods achieved correct answers, suggesting these questions were comparatively less challenging. These questions merely required a "Yes" or "No" response, without necessitating the prediction of specific labels or other abilities, such as counting. For Question 2, focused on "Query time", only the proposed method incorporating GPT-4 reasoning achieved the correct answer, aligning with findings in Table III that highlight GPT-4's capability in temporal reasoning. Meanwhile, for Question 4, which involves "Count object location", only Our (Trans) method succeeded, underscoring the challenge in counting tasks for most methods, including the method based on GPT-4. Introducing specialized counting modules could potentially enhance model performance.

VI. CONCLUSIONS

Understanding daily human activity scenarios is vital for discerning patterns in human behavior, identifying abnormal activities, and providing personalized services. Recognizing these scenarios is critical across various applications, such as domestic robotics, medical diagnosis, and nursing care. However, current datasets and methods predominantly focus on single-moment, single-location activities and lack depth in spatio-temporal reasoning across multiple times and locations throughout the day. Our novel dataset, DailySTR, along with a baseline model, addresses this gap by recognizing video details and facilitating spatio-temporal reasoning over these details. Our experiments demonstrate that while previous state-of-the-art methods underperform on DailySTR, the application of LLMs yields promising reasoning results. We aim for our dataset and findings to serve as a benchmark and provide insights for future advancements in the comprehension of daily scenarios.

Looking ahead, we plan to enrich the dataset by incorporating a variety of 3D environments and a broader spectrum of human activities, and by leveraging Generative AI to enhance the dataset's realism, complexity, and scope. For model enhancement, integrating advanced multimodal LLMs to improve commonsense reasoning and action prediction is an exciting avenue for future research. Additionally, we are interested in adapting our datasets and models for practical real-world applications.

ACKNOWLEDGMENT

Computational resource of AI Bridging Cloud Infrastructure (ABCI) provided by National Institute of Advanced Industrial Science and Technology(AIST) was used.

REFERENCES

[1] Kong, Q., Wu, Z., Deng, Z., Klinkigt, M., Tong, B., Murakami, T. Mmact: A large-scale dataset for cross modal human action understanding. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019.

[2] Goyal, R., Ebrahimi Kahou, S., Michalski, V., Materzynska, J., Westphal, S., Kim, H., Memisevic, R. The "something something" video database for learning and evaluating visual common sense. In Proceedings of the IEEE international conference on computer vision, 2017.

[3] Jia, B., Chen, Y., Huang, S., Zhu, Y., Zhu, S. C. LEMMA: A Multi-view Dataset for LEarning Multi-agent Multi-task Activities. In European Conference on Computer Vision, 2020.

[4] Grunde-McLaughlin, M., Krishna, R., Agrawala, M. Agqa: A benchmark for compositional spatio-temporal reasoning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021.

[5] Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., ..., Malik, J. Ego4d: Around the world in 3,000 hours of egocentric video. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022.

[6] Ji, J., Krishna, R., Fei-Fei, L., Niebles, J. C. Action genome: Actions as compositions of spatio-temporal scene graphs. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020.

[7] Rai, N., Chen, H., Ji, J., Desai, R., Kozuka, K., Ishizaka, S., Adeli, E., Niebles, J. C. Home action genome: Cooperative compositional action understanding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021.

[8] Qiu, Y., Nagasaki, Y., Hara, K., Kataoka, H., Suzuki, R., Iwata, K., Satoh, Y. VirtualHome Action Genome: A Simulated Spatio-Temporal Scene Graph Dataset with Consistent Relationship Labels. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023.

[9] Bertasius, G., Wang, H., Torresani, L. Is space-time attention all you need for video understanding?. In International Conference on Machine Learning, 2021.

[10] Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., Hu, H. Video swin transformer. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022.

[11] Wang, J., Ge, Y., Yan, R., Ge, Y., Lin, K. Q., Tsutsui, S., Lin, X., Cai, G., Wu, J., Shan, Y., Qie, X., Shou, M. Z. All in one: Exploring unified video-language pre-training. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023.

[12] Li, L., Gan, Z., Lin, K., Lin, C. C., Liu, Z., Liu, C., Wang, L. Lavender: Unifying video-language understanding as masked language modeling. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023.

[13] Ugai, T., Egami, S., Htun, S. N. N., Kozaki, K., Kawamura, T., Fukuda, K. Synthetic Multimodal Dataset for Empowering Safety and Well-being in Home Environments. arXiv preprint arXiv:2401.14743, 2024.

[14] Gandhi, M., Gul, M. O., Prakash, E., Grunde-McLaughlin, M., Krishna, R., Agrawala, M. Measuring compositional consistency for video question answering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022.

[15] Yu, Z., Zheng, L., Zhao, Z., Wu, F., Fan, J., Ren, K., Yu, J. ANetQA: A Large-scale Benchmark for Fine-grained Compositional Reasoning over Untrimmed Videos. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023.

[16] Datta, S., Dharur, S., Cartillier, V., Desai, R., Khanna, M., Batra, D., Parikh, D. Episodic memory question answering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022.

[17] Urooj, A., Kuehne, H., Wu, B., Chheu, K., Bousselham, W., Gan, C., ..., Shah, M. Learning Situation Hyper-Graphs for Video Question Answering. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023.

[18] Li, Y., Xiao, J., Feng, C., Wang, X., Chua, T. S. Discovering spatio-temporal rationales for video question answering. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023.

[19] Zhou, H., Martín-Martín, R., Kapadia, M., Savarese, S., Niebles, J. C. Procedure-aware pretraining for instructional video understanding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023.

[20] Tan, C., Lin, Z., Hu, J. F., Zheng, W. S., Lai, J. Hierarchical Semantic Correspondence Networks for Video Paragraph Grounding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023.

[21] Straub, J., Whelan, T., Ma, L., Chen, Y., Wijmans, E., Green, S., ..., Newcombe, R. The Replica dataset: A digital replica of indoor spaces. arXiv preprint arXiv:1906.05797, 2019.

[22] Fu, H., Cai, B., Gao, L., Zhang, L. X., Wang, J., Li, C., ..., Zhang, H. 3d-front: 3d furnished rooms with layouts and semantics. In Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021.

[23] Kolve, E., Mottaghi, R., Han, W., VanderBilt, E., Weihs, L., Herrasti, A., ..., Farhadi, A. Ai2-thor: An interactive 3d environment for visual ai. arXiv preprint arXiv:1712.05474, 2017.

[24] Shen, B., Xia, F., Li, C., Martín-Martín, R., Fan, L., Wang, G., ..., Savarese, S. iGibson 1.0: A simulation environment for interactive tasks in large realistic scenes. In IEEE/RSJ International Conference on Intelligent Robots and Systems, 2021.

[25] Li, C., Xia, F., Martín-Martín, R., Lingelbach, M., Srivastava, S., Shen, B., ..., Savarese, S. igibson 2.0: Object-centric simulation for robot learning of everyday household tasks. arXiv preprint arXiv:2108.03272, 2021.

[26] Puig, X., Ra, K., Boben, M., Li, J., Wang, T., Fidler, S., Torralba, A. Virtualhome: Simulating household activities via programs. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018.

[27] Puig, X., Undersander, E., Szot, A., Cote, M. D., Yang, T. Y., Partsey, R., ..., Mottaghi, R. Habitat 3.0: A co-habitat for humans, avatars and robots. arXiv preprint arXiv:2310.13724, 2023.

[28] Htun, S. N. N., Egami, S., Fukuda, K. Activity scenarios simulation by discovering knowledge through activities of daily living datasets. SICE Journal of Control, Measurement, and System Integration, 17(1), 87-105, 2024.

[29] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ..., Houlsby, N. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In International Conference on Learning Representations, 2020.

[30] Kenton, J. D. M. W. C., Toutanova, L. K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of NAACL-HLT, 2019.

[31] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., ..., McGrew, B. Gpt-4 technical report. arXiv preprint arXiv:2303.08774, 2023.

[32] Xu, J., Mei, T., Yao, T., Rui, Y. Msr-vtt: A large video description dataset for bridging video and language. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2016.

[33] Xiao, J., Shang, X., Yao, A., Chua, T. S. Next-qa: Next phase of question-answering to explaining temporal actions. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021.

[34] Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In IEEE conference on computer vision and pattern recognition, 2009.