A Video-grounded Dialogue Dataset and Metric for Event-driven Activities

Wiradee Imrattanatrai[1*], Masaki Asada[1*], Kimihiro Hasegawa[2], Zhi-Qi Cheng[2], Ken Fukuda[1], Teruko Mitamura[2]

[1]National Institute of Advanced Industrial Science and Technology (AIST)
[2]Language Technologies Institute, Carnegie Mellon University
wiradee.imrattanatrai@aist.go.jp, masaki.asada@aist.go.jp

Abstract

This paper presents VDAct, a dataset for a Video-grounded Dialogue on Event-driven Activities, alongside VDEval, a session-based context evaluation metric specially designed for the task. Unlike existing datasets, VDAct includes longer and more complex video sequences that depict a variety of event-driven activities that require advanced contextual understanding for accurate response generation. The dataset comprises 3,000 dialogues with over 30,000 question-and-answer pairs, derived from 1,000 videos with diverse activity scenarios. VDAct displays a notably challenging characteristic due to its broad spectrum of activity scenarios and wide range of question types. Empirical studies on state-of-the-art vision foundation models highlight their limitations in addressing certain question types on our dataset. Furthermore, VDEval, which integrates dialogue session history and video content summaries extracted from our supplementary Knowledge Graphs to evaluate individual responses, demonstrates a significantly higher correlation with human assessments on the VDAct dataset than existing evaluation metrics that rely solely on the context of single dialogue turns.

Resources — https://github.com/aistairc/VDAct

Introduction

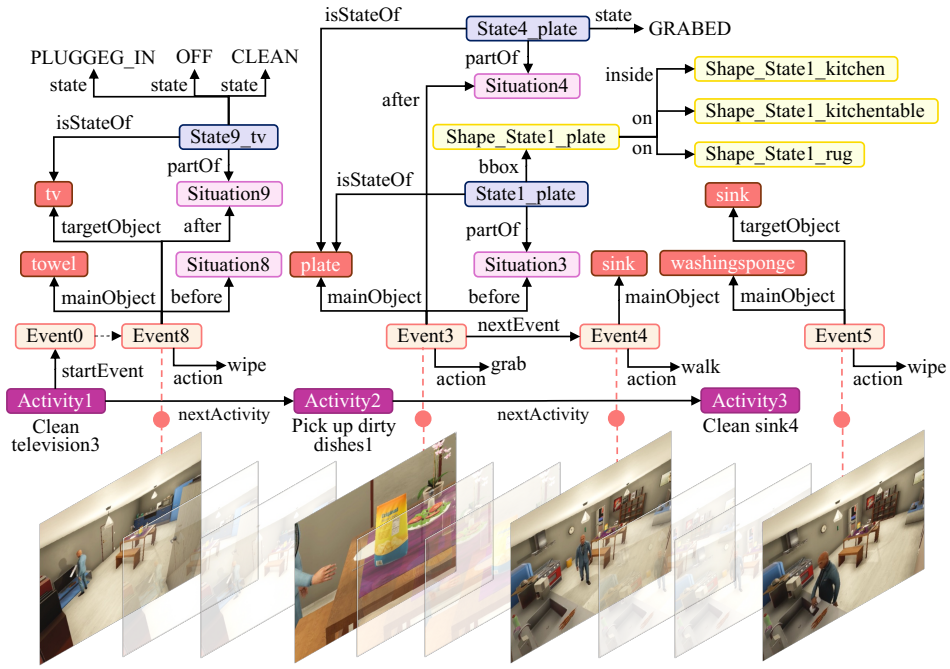
The video-grounded dialogue task involves generating responses to user utterances based on the video content. This task poses significant challenges particularly when dealing with videos presenting compound activities where multiple related events occur in sequence. With these event-driven activities, advanced system capabilities in multimodal understanding, temporal reasoning, and contextual interpretation are required to align visual cues with conversational context and handle dynamic changes in the video. Although several datasets exist for video-based reasoning through question answering, only a few benchmark datasets are available for the video-grounded dialogue task (Alamri et al. 2019; Pasunuru and Bansal 2018). These datasets mostly feature short videos depicting simple activities and a limited range of question types, while in real-world scenarios, dialogue discussions often center around multifaceted activities about how various events with different associated actions are temporally and contextually related. Thus, exposure to dialogues on event-driven activities would expand the system's ability to handle complex interactions and improve its capacity to generate accurate and relevant responses.

Thus, to advance research and development of video-grounded dialogue systems on event-driven activities, we introduce a new dataset named "**V**ideo-grounded **D**ialogue on Event-driven **Act**ivities" (VDAct). This dataset includes dialogues based on daily scenarios where each involves multiple activities with long sequences of events. We opted to utilize virtual simulation videos that allow a variety of activity combinations. Unlike existing datasets which primarily focus on descriptive questions, VDAct includes several other categories to capture a broader range of interactions as shown in Table 1. In addition to the **descriptive** questions which aim to obtain factual information about the activities, we incorporate three other main categories including **temporal** questions, which focus on temporal aspects, such as timing, duration, and sequence of events or activities; **explanatory** questions which explore the reasons or causes behind events or activities; and **quantitative** questions, which seeks numerical or quantitative data. Additionally, VDAct incorporates questions related to the video and dialogue attributes, as well as open-ended and subjective questions.

In addition to videos and dialogues, we enrich our dataset with Knowledge Graphs (KGs) as supplementary information. Given that our target videos represent compound activities involving multiple events, KGs could be useful for both system development and evaluation as they offer detailed information that links visual cues to the structured information. This information includes event sequences, transitions between events, action-object interactions, and changes in agent and object states, as illustrated in Figure 1.

Furthermore, existing evaluation metrics, such as those designed for text generation (Papineni et al. 2002; Banerjee and Lavie 2005; Lin 2004), and QA tasks (Mañas, Krojer, and Agrawal 2024; Chan et al. 2023; Wada et al. 2024), are insufficient for assessing the quality of generated responses in the context of dialogues. These metrics usually assess system-generated text by comparing it to a reference text, but they often overlook cases where multiple responses can be equally valid. For example, in response to Q7 in Figure 1, both "He grabbed the plates from the kitchen ta-

*These authors contributed equally.



[QNT]
Q1: How many tasks does the man do in the kitchen?
A1: The man has three tasks in the kitchen.

[QNT]
Q2: How many cleaning chores does he do in the kitchen?
A2: Two cleaning chores, one tidying one.

[D-OBJ] [T-SEQ]
Q3: What is the first thing he cleans?
A3: He cleans the television set.

[D-INT]
Q4: Does he clean under the TV?
A4: No he seems pretty focused on the screen part.

[D-INT] [T-SEQ]
Q5: Does he clean another thing after the TV or tidy first?
A5: He tidies first, then does more cleaning.

[D-OBJ]
Q6: What does he tidy up?
A6: He clears two dirty dishes.

[D-LOC] [REF]
Q7: Where did he get them from?
A7: From the kitchen table.

Figure 1: VDAct with example dialogue (turns 1-7) for an activity scenario video with corresponding KG elements for the events. Each input utterance (*i.e.,* question) is labeled with the relevant question types. Note that the object IDs are omitted from the KG illustration.

ble." and "They were on the rug, which was on top of the kitchen table." are valid when considering additional context from dialogue history and KGs. Although neither Q7 nor A7 mentions what objects they are referring to, the first response is deemed correct because it aligns with prior dialogue turns. Additionally, the KG information for the related event (*i.e.,* Event3 of Activity2) shows that the plate was on the rug, making the second response also accurate. However, because the existing metrics only compare responses to the reference and do not consider this additional context, they cannot fully capture the correctness of both responses and as a result, evaluate them as partially or completely inaccurate. Thus, we propose a new LLM-based evaluation metric that integrates *session-based* context including a summary of KG information and dialogue history, rather than relying on *turn-based* context. This inclusion allows the evaluation model to verify the responses more comprehensively.

To this end, our main contributions are three-fold. First, we introduce a new video-grounded dialogue dataset focused on event-driven activities and provide accompanying KGs for the activity scenarios. Second, we propose a new evaluation metric, specifically for the video-grounded dialogue task, to address the unique aspects of dialogues where multiple responses are possible and inference from summary and dialogue history are required. Finally, we selected Vision LMs (VLMs) to evaluate their performance on our newly constructed dataset and evaluate the correlation between the results and our new proposed metric.

Related Works

This section reviews the existing datasets and the existing evaluation metrics for the video-grounded dialogue task.

Video-grounded Dialogue Datasets In terms of video understanding, several video-grounded QA datasets have been introduced. Among them, a few provide KGs such as STAR (Wu et al. 2021) which presents multi-choice QA samples along with situation hypergraphs based on real-world videos, and EgoTaskQA (Jia et al. 2022) which introduced a QA dataset with annotations of object status, human-object and multi-agent relationships, and causal dependency structures between actions, all derived from egocentric videos. However, while QA tasks treat each question independently, dialogue tasks build on prior interactions of questions and answers. This offers a key advantage for system development as the sequential nature of dialogue enables the system to maintain continuity and deliver context-aware responses by referencing previous turns. For the video-grounded dialogue datasets, VisDial (Das et al. 2017) treated the problem of visual dialogue as a multi-turn QA where the system is expected to answer questions given dialogue history and corresponding images. Audio Visual Scene-aware Dialog (AVSD) (Alamri et al. 2019) extended the work from VisDial to include additional modalities including videos with audio signals. Twitch-FIFA (Pasunuru and Bansal 2018) introduced a video-context dialogue dataset based on live-broadcast soccer games and chat from Twitch.tv. Video-grounded Scene and Topic AwaRe dialogue (VSTAR) (Wang et al. 2023) introduced a large-scale

Category		Description
(1) Descriptive		
- Agents	D-AGT	Questions about characteristics or states of the agents
- Actions	D-ACT	General questions related to actions without specifying objects
- Objects	D-OBJ	Questions about objects involved in actions, those upon which actions are performed, or the states of objects
- Interactions	D-INT	Questions about whether actions were performed on specific objects
- Locations	D-LOC	Questions concerning the whereabouts of agents, objects, or where the actions were performed
(2) Temporal		
- Sequence	T-SEQ	Questions related to the temporal sequence of actions w/wo objects or questions related to specific points in time
- Frequency	T-FRQ	Questions about the frequency or duration of actions performed w/wo objects
(3) Explanatory	EXP	Questions seeking explanations for how and why actions were carried out
(4) Quantitative	QNT	Questions related to quantity or number of agents or objects
(5) Other		
- Reference	REF	Questions that reference previous dialogue turns or require answers based on the dialogue history
- Supplementary	SUP	Questions seeking further details or information about the activities
- Video Attributes	VID	Questions about attributes such as audio, quality, length of the video, or the language spoken within the video
- Opinions	OPI	Questions asking opinions or facts about the agents, actions, and objects that are subjective or cannot be verified

Table 1: Categories and types of questions involve in VDAct.

benchmark dataset for understanding the dialogue between characters in a TV series. These datasets usually contain short videos with simple activities and focus mainly on descriptive questions. This constraint limits the system's ability to learn from more complex scenarios that involve multiple related events occurring in sequence. Thus, we propose a new dataset with longer videos depicting complex activities, and dialogues presenting a variety of question types.

Evaluation Metrics for Video-grounded Dialgoue Developing an effective evaluation metric for the video-grounded dialogue task presents a significant challenge. Previous studies reported that classic evaluation metrics such as BLEU, METEOR, and ROUGE showed low correlation with human evaluations for the video-grounded dialogue task (Liu et al. 2016; Alamri et al. 2019). Considering this point, AVSD employs ranking-based evaluation metrics and a discriminative ranking task setting where the model prediction is selected from candidate answers.

Learning-based metrics such as PAC-S (Sarto et al. 2023) and Polos (Wada et al. 2024) have been proposed and showed a high correlation with human evaluation for the image captioning task. However, these metrics require vast amounts of human scores to train the metrics. This makes it difficult to apply the metrics on the video modality. LLM-based evaluation metrics, such as CLAIR (Chan et al. 2023) for image captioning, and LAVE (Mañas, Krojer, and Agrawal 2024) and LLM-Acc/Rel (Maaz et al. 2024) for VQA have been gaining attention in recent years. However, these metrics only compare the generated text with the reference of a single QA turn without considering additional contextual information that is beneficial for the evaluation.

In this study, we exploit the advantage of the video context being directly linked to event-centric structured KGs and dialogue history to propose a new LLM-based evaluation metric for the video-grounded dialogue task.

VDAct Dataset

Data Collection

We prepare target scenario videos for dialogue generation, followed by employing crowdsourced workers to create dialogues discussing the event-driven activities depicted in the videos. To support the video-grounded dialogue task, we additionally include scenario KGs and their summaries. The following subsections detail the collection process for each data component.

Preparation of Scenario Videos To gather video data representing daily living activities, we leverage the VirtualHome2KG dataset (Egami et al. 2023)[1], which integrates KGs with video data for tasks such as activity recognition. VirtualHome2KG contains simulation videos depicting various daily activities performed by a single agent in a 3D virtual space using the VirtualHome platform (Puig et al. 2018). Each video captures an activity in a unique home environment, varying in room layouts and camera angles. VirtualHome2KG relies on activities of 11 classes defined by HomeOntology (Vassiliades et al. 2020) such as BedTimeSleep, EatingDrinking, FoodPreparation, HouseArrangement, and others, as well as one additional class, Abnormal. Each activity is associated with a program representing a sequence of events involving actions and objects. For example, an event of "[WALK] ⟨television⟩ (297)" where the number denotes the object ID. The dataset consists of 3,530 videos covering 706 activities across 12 categories, 7 environment setups, and 6 viewpoints, including an indoor camera switching view, character rear views, and fixed camera angles.

As the target for our dialogue is to analyze the daily scenario consisting of multiple activities, we combined 2-5 available activities within VirtualHome2KG as an activity scenario. Following are the constraints we imposed to obtain high-quality videos. First, we need to ensure the seamless connectivity of the activity videos as scenario videos. Thus,

[1]https://github.com/KnowledgeGraphJapan/KGRC-RDF/tree/kgrc4si

we chose to combine the activities where each occurs in the same environment setup and that the agent starts and ends the activity in the same room as its previous and succeeding activity, respectively. As there can be too many possible activity combinations, we do not allow the same activity to be in the same scenario, as well as limit the scenario video to 1 to 5 minutes of the combined activity videos. Out of six viewpoints, we selected activity videos from either an indoor camera switching view or a fixed view at a room corner as they show the clearest depiction of activities for a scenario.

With a set of candidate scenarios, we further filtered out scenarios with some criteria to diversify the activity combination. Firstly, we set the limit for the number of individual activities appearing as the first activity as 8. Secondly, we excluded scenarios that have more than half of their activities duplicated across other scenarios. Lastly, we limit each adjacent activity pair to appear in no more than two scenarios. Thus, we obtained 3,021 unique activity combinations as scenarios. From this list, we randomly sample 1,000 scenarios for preparing the dialogues.

Dialogue Data Creation For the creation of dialogue data, we hired six crowdsourced workers through a reputable third-party company specialized in creating and collecting language data for NLP research and development. The goal of this data collection step is to obtain dialogue sessions demonstrating the information exchange of the person's activity scenario between a pair of annotators. For each given scenario, the two annotators were assigned different roles to have a formal discussion about the scenario. One annotator was assigned to act as the investigator with a responsibility to investigate the person's behavior and figure out how the person performs the activities by asking questions to the corresponding annotator. The investigator was not allowed to watch the videos but was given an unordered list of activities to provide some ideas about the scenario. The other annotator, acting as the correspondent, was given a list of activities, and was assigned to watch the videos to provide accurate answers to the investigator.

To cover multiple types of questions, we chose to provide examples from four main categories (*i.e.,* descriptive, temporal, explanatory, and quantitative) that can be incorporated into the dialogues. Moreover, we also provided a few example dialogues to the annotators to have a clear picture of the task. We do not set strict minimum or maximum limits on the number of each question type per dialogue, as such a limitation could negatively impact the natural flow of the conversation. Instead, we instructed annotators to include as many question types within each dialogue to ensure type coverage.

For 1,000 scenarios from the previous data collection step, we formed three different annotator pairs. In half of the scenarios, one annotator was the investigator, while the other was the correspondent. The role of the pair switches for the latter half of the scenarios. This approach ensured that each annotator had the opportunity to play both roles, potentially leading to more diverse and comprehensive dialogues. Moreover, to ensure that the annotators fully understand the task instruction, we manually reviewed the initial dialogues created by each annotator pair before allowing them to proceed with the rest of the process. In total, we obtained 3,000 distinct dialogues with over 30,000 turns (*i.e.,* question-answer pairs) through this process.

Triplet	Sentence
(event0, *from*, bedroom75) (event0, *action*, walk) (event0, *mainObject*, bathroom11)	The person is in the bedroom. He walks to the bathroom.
(door53, *inside*, bathroom) (door53, *state*, OPEN) (event1, *action*, walk) (event1, *mainObject*, door53)	The door is inside the bathroom. The door is OPEN. He walks to the door.
(event2, *action*, close) (event2, *mainObject*, door53)	He closes the door.
(door53, *state*, CLOSED) (toilet, *inside*, bathroom11) (event3, *action*, walk) (event3, *mainObject*, toilet46)	The door is CLOSED. The toilet is inside the bathroom. He walks to the toilet.
(toilet46, *close*, character1) (character1, *inside*, bathroom11) (event4, *action*, sit) (event4, *mainObject*, toilet46)	The toilet is next to the person. The person is inside the bathroom. He sits on the toilet.

Table 2: Example triplets and the corresponding sentences for template-based video summaries from KGs.

Knowledge Graph Collection for Scenarios As supplementary information for the dataset, we collected KGs for scenarios by relying on KGs for activities provided by VirtualHome2KG. Each activity KG consists of 9 main node types including Activity, Event, Action, Situation, Object, State, Attribute, StateVal, and Shape. The Activity such as "Drink wine while watching television3" links to Event nodes such as "event0" and "event1" which indicate events happening in sequence. Each Event node connects to the main Object node (*e.g.,* wine465) and the Action node (*e.g.,* grab) associated with the event. Additionally, Event nodes are connected to Situation nodes that describe the state of each object (*i.e.,* a State node) in the environment before and after the event using the StateVal nodes (*e.g.,* ON, OFF, CLEAN and DIRTY). Meanwhile, Shape nodes represent the 3D coordinates of agents and object states and are linked from State nodes by bbox relations and to each other through spatial relations (*e.g., inside*, *on*, and *close*).

Since the average number of triplets for a single activity in VirtualHome2KG is over 29,237, we selectively curated and combined only the most relevant triplets within activity KGs as a scenario KG to suit the purpose of our dataset for the task. Instead of including all objects present in the environment, we focused on the main and target objects directly involved in the events of each activity. This approach aligns with our goal of constructing dialogues centered around activities, as most dialogue turns typically pertain to the main and target objects on which actions are performed. Additionally, to enhance the temporal coherence of the scenario, we introduced additional triplets with the **nextActivity** relationship to link adjacent activities in chronological order.

Dataset	#Videos	#Dialogues	#QA Pairs	Video Source	Avg. Video Length	Avg. Question Length	Avg. Answer Length	KG
VisDial	120k (images)	120k	1.2M	-	-	5.1	8.2	✗
Twitch-FIFA	49	15,083	15,083	Soccer match	30 secs	68	6.3	✗
AVSD	11,816	11,816	118,160	Crowdsourced	30 secs	7.9	9.4	✗
VDAct	1,000	3,000	30,095	VirtualHome	248 secs	7.8	10.2	✓

Table 3: Comparison of our dataset with the existing datasets.

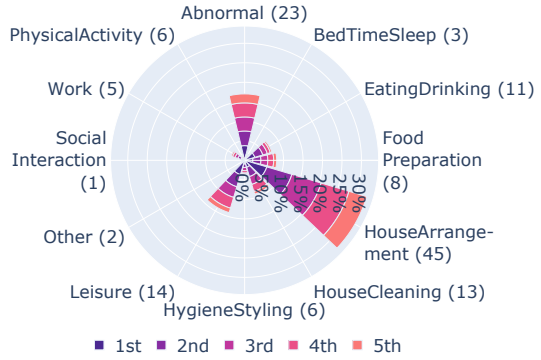
Preparation of Scenario Video Summaries We generated video summaries for each scenario by translating triplets from the scenario KGs using a **template**-based approach. The examples are shown in Table 2. After generating the linearized KG-to-text summaries, we further **refined** the text using a commercial LLM, specifically GPT-4o-mini, with the prompt: "Please summarize the following text without adding any extra information: {text}," where {text} is the placeholder for the linearized summary. This additional step is necessary to remove redundant sentences that might arise from triplets where object states remain unchanged throughout the event sequence. For example, the sentence "the stove is inside the kitchen." might appear repeatedly for multiple events, so this step helps eliminate such redundancies. After the refinement, we checked the quality of a few summaries whether they contained necessary information and were free from fabricated details.

Dataset Analysis

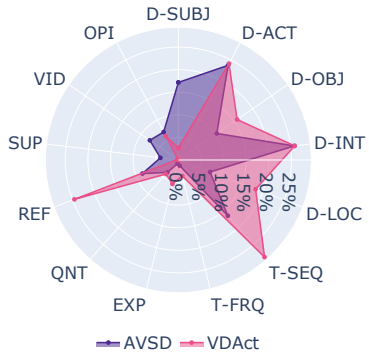
Table 3 compares the existing datasets for the visual dialogue task with ours. While our dataset includes a smaller number of videos and dialogues, it presents target videos with a longer average duration and represents a diverse combination of activities. Additionally, unlike existing datasets, our dataset includes KGs that provide structured, event-centric information linking activities, events, and other relevant details. These KGs are a valuable resource for developing and evaluating video-grounded dialogue systems.

Activities for Scenarios We analyzed the diversity of activity combinations within scenarios by examining the distribution of activities across categories, as shown in Figure 2a. The analysis indicates a balanced distribution of activities among 1^{st} to 4^{th} placement within scenarios, categorized by their respective types. Although the HouseArrangement category shows a noticeably higher number of activities in scenarios compared to other categories, this is justified by the fact that it contains nearly twice as many distinct activities (45) as the second-highest category (23).

Question Types for Dialogues In Figure 2b, we conducted a comprehensive analysis of the question types used in the dataset by categorizing them into 13 pre-defined types based on their distinct characteristics, as detailed in Table 1. To compare our dataset with the closely related AVSD dataset, we randomly selected 60 dialogues from each and labeled each question with one or more of our question types. It is important to note that each question can be assigned multiple types. Our analysis revealed that the AVSD dataset primarily features the question types D-ACT, D-INT,



(a) The occurrence of activities by categories based on their placements in scenarios. The number in parentheses indicates the total count of activities within that category.



(b) Comparison of the number of question types.

Figure 2: Statistics on activities as scenarios in the VDAct dataset and percentages of different questions types for sample dialogues in comparison with the AVSD dataset.

and D-AGT, whereas our dataset highlights T-SEQ, D-INT, REF, and D-ACT as the top types. This difference emphasizes the realistic nature of our dialogues, which frequently reference dialogue history in a natural conversational style. Additionally, our dataset emphasizes discussions about activities involving actions and interactions between actions and objects, including temporal inferences. Particularly, it contains a significantly higher number of questions classified as D-OBJ, D-LOC, T-SEQ, EXP, and REF compared to the AVSD dataset. Meanwhile, as opposed to our dataset, the AVSD dataset includes a significantly higher number of questions related to D-AGT and VID types. This is because the AVSD dataset employs crowdsourced videos with vary-

	Activity	Scenario
Events	10.2	40.26
Situations	11.2	44.56
States	14.92	60.36
StateVals	10.84	44.04
Shapes	86.56	360.17
Main Objects	4.08	13.56
Target Objects	0.78	2.57
Triplets	320.78	1,317.51

Table 4: Comparison of Scenario and Activity KGs by the average number of different components.

ing agents, leading to a greater focus on questions concerning agent states such as ages, appearances, and emotions and the video attributes such as audio and language spoken.

Scenario KGs Table 4 shows the statistics of our scenario KGs after merging the activity KGs. From this table, it suggests that each scenario KG contains more number of Events, Situations, States, and Objects nodes than the activity KGs. This increase highlights the greater complexity and detail of our dataset in providing compound activities with multiple related events for the dialogue task.

VDEval Metric

We introduce a new evaluation metric for the video-grounded dialogue task to overcome the shortcomings of the existing metrics. Particularly due to the lack of sufficient context, the existing metrics failed to verify the content in generated responses as they are only presented with the user utterances (*i.e.,* questions) and references. Typically, in the context of dialogue, different responses can be considered accurate, even if they do not precisely match the references, as long as they align with the video context and dialogue history. In addition, in some cases, responses may include extra information that needs to be assessed for relevance.

Thus, for our proposed metric, we extended the existing LLM-based metric which demonstrates a high correlation with humans for the VQA task, LAVE (Mañas, Krojer, and Agrawal 2024). LAVE uses rationales and a scaled score rating of 1-3 to assess the generated answer (*i.e.,* response) given the question and the reference answer. Our new metric introduces two key improvements over LAVE. First, instead of evaluating the generated response using the individual turn-based context, we include the entire dialogue history, which provides context from the previous question-answer pairs with evaluated scores and rationale. Second, we add a video summary from scenario KGs that offers a high-level overview of the video content for evaluation. Our approach to incorporating session-based dialogue context is based on the idea that this additional context provides details and information that are difficult to capture with just the question and reference alone. By including the full dialogue history and video content summary, LLMs can use this richer context to accurately assess the correctness of the current response. Figure 3 illustrates how our enhanced metric compares to the existing one.

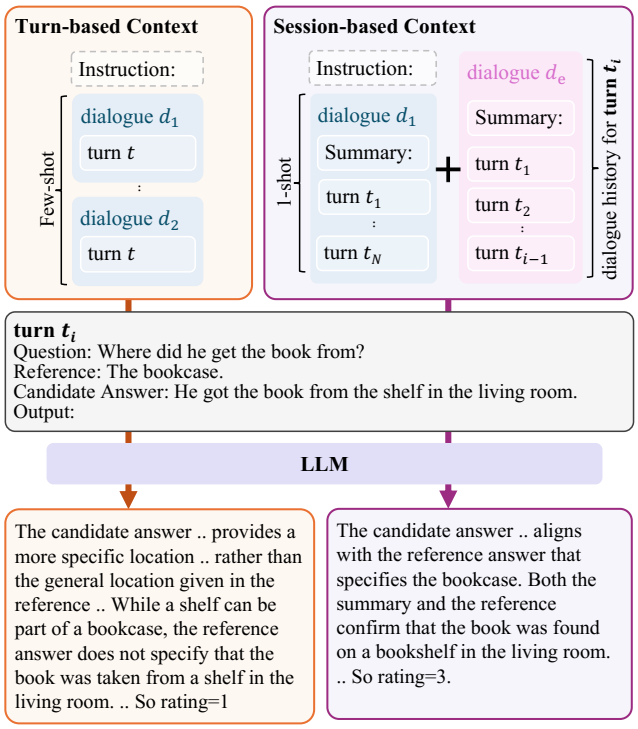


Figure 3: Comparison of turn-based and session-based contexts for evaluation metrics. The input prompt includes turn- or session-based context with the information of turn t_i.

Experiments

Experimental Settings

Data Splits Our dataset comprises 3,000 dialogues created from 1,000 scenarios, with each scenario created by three pairs of annotators. To prevent the occurrence of dialogues based on identical scenarios across training and validation/test sets, the dataset is split at the scenario level. We set the fraction of the train, test, and validation as 80% (2,400 dialogues), 15% (450), and 5% (150), respectively.

Baselines We adopted state-of-the-art large-scale vision-language foundation models, including both open models such as Video-LLaVA (Lin et al. 2023), Video-ChatGPT (Maaz et al. 2024), and VideoLLaMA2 (Cheng et al. 2024), as well as proprietary models GPT-4o and Gemini-1.5-pro, as baseline models for our dataset. We evaluated open models both with frozen pretrained parameters and with LoRA fine-tuned parameters on the train set of VDAct. We followed the same video frame sampling method used during its pretraining, extracting a fixed number of frames at regular intervals. For fine-tuning with LoRA and model inferences, we report the parameter settings on the supplementary material.

Baseline Metrics We adopted widely used metrics for evaluating generated text: BLEU (Papineni et al. 2002), ROUGE (Lin 2004), METEOR (Banerjee and Lavie 2005). We also included the standard image captioning metric SPICE (Anderson et al. 2016), and similarity-based metrics:

		BLEU	ROUGE	METEOR	SPICE	LLM-Acc	LLM-Rel	LAVE	VDEval
Frozen	Video-LLaVA (8)	6.56	26.68	29.24	25.15	19.74	31.94	30.14	29.01
	Video-ChatGPT (100)	4.38	24.09	24.88	22.87	18.99	30.89	29.48	29.56
	VideoLLaMA2 (8)	**6.96**	**28.86**	**32.20**	**27.33**	**20.99**	**33.63**	**32.20**	**32.89**
	VideoLLaMA2 (16)	6.80	27.73	30.55	26.17	19.67	32.02	31.17	31.07
LoRA finetuned	Video-LLaVA (8)	11.05	**37.16**	**42.08**	34.59	30.02	43.13	41.06	42.19
	Video-ChatGPT (100)	8.64	33.65	38.87	31.86	23.92	36.53	34.79	35.62
	VideoLLaMA2 (8)	**11.80**	36.63	40.87	**35.30**	**30.92**	**44.14**	**41.84**	**43.13**
	VideoLLaMA2 (16)	10.68	35.56	40.45	34.41	30.06	42.91	41.03	42.39
Proprietary	GPT-4o (8)	5.95	26.93	27.55	25.02	27.32	40.07	38.46	37.30
	GPT-4o (16)	**6.37**	**27.60**	**28.14**	**25.58**	**30.39**	**43.25**	**41.38**	**41.26**
	Gemini-1.5-pro-002 (8)	1.92	21.60	19.28	21.73	21.97	36.85	37.71	34.41
	Gemini-1.5-pro-002 (16)	1.80	21.86	19.02	21.97	24.03	39.39	40.08	37.77

Table 5: Performances of baselines across various metrics in terms of dialogue generation. The numbers in parentheses represent the number of video frames used by each baseline. The scores of LLM-Rel, LAVE, and VDEval have been scaled to the 0-1 range and are reported in percentages to ensure consistency with the other metrics. **Bold** scores represent the highest scores within each baseline group, while underlined scores denote the highest scores across all baselines.

Metric type		VL2 frozen	VL2 finetuned
Classic	BLEU	26.13	34.27
	ROUGE	32.65	37.39
	METEOR	31.35	42.73
	SPICE	30.02	38.94
Similarity-based	BERTScore	22.43	29.59
	BARTScore	22.29	38.45
LLM-based	LAVE	66.94	69.35
	LLM-Rel	62.46	70.08
VDEval (Ours)			
Context	KG summary		
Turn-based	✗	66.85	70.09
	Template	70.37	70.71
	Refined	70.68	71.75
Session-based	✗	68.25	72.59
	Template	69.12	71.63
	Refined	**72.30***	**73.62***

Table 6: Kendall's rank correlation coefficient between various evaluation metrics and human judgments. VL2 stands for VideoLLaMA2. Bold indicates the highest correlation among all metrics, while underline indicates the highest correlation among existing metrics. * indicates statistically significant improvements in correlation from the underlined scores at the significance level of $p < 0.05$.

BERTScore (Zhang et al. 2020), and BARTScore (Yuan, Neubig, and Liu 2021).

Additionally, we adopted LLM-based evaluation metrics LLM-Acc/Rel (Maaz et al. 2024) and LAVE (Mañas, Krojer, and Agrawal 2024), specifically used for the VQA task. LLM-Acc assesses the binary correctness of generated text by feeding QA pairs and predicted answers to LLMs. LLM-Rel rates text quality on a 0-5 scale, with higher scores indicating better quality. LAVE scores 1-3 and employs a few-shot examples as an instruction prompt.

It is noted that the only difference between VDEval with turn-based context without KG summaries and LAVE is a selection of few-shot examples. VDEval uses a 1-shot dialogue example drawn from the VDAct dataset, while LAVE employs few-shot examples provided by the original paper for the general VQA task. We selected GPT-4o-mini for the LLM-based evaluation metrics LAVE, LLM-Acc/Rel, and VDEval.

Correlation with Human Judgment To examine the validity of existing evaluation metrics and our VDEval, we prepared human evaluation judgments and calculated the correlation coefficient between human scores and scores by automatic evaluation metrics. For human evaluation, we randomly selected 20% of the test set, comprising 90 dialogues with approximately 900 turns. We defined scores from 1 to 3 as 1: incorrect, 2: partially correct, 3: correct. Three independent annotators were employed, with each annotator scoring all 90 dialogues individually. We instructed annotators to comprehensively score the quality of answers by comparing them with gold references and checking the content of the video and dialogue history. To determine the final score for each sample, we took a majority vote among the scores from three annotators. In cases where all three annotators gave different scores, we excluded those samples from the correlation calculation. Following previous evaluation metrics (Anderson et al. 2016; Wada et al. 2024), we adopted the Kendall-B rank correlation coefficient as a correlation metric with human evaluation. To assess the statistical significance of the improvement in correlation achieved by VDEval compared to existing metrics, we employed a permutation test (Good 2000) with 10,000 iterations.

Results and Discussion

Baseline Performances on VDAct Table 5 shows a performance comparison of various baseline models on VDAct test set. Notably, all baseline models exhibited performance below 50% on VDEval metric, indicating that the quality of the generated dialogue answers does not even exceed a score of 2, which means "partially correct". These results highlight the difficulty of the VDAct task and underline the need for further advancements in video-grounded dialogue systems for event-driven activities. When the model parameters are frozen, VideoLLaMA2 8-frame model consistently outperformed other baselines across all evalua-

tion metrics. When model parameters are fine-tuned, VideoLLaVa 8-frame model achieves the highest performance on ROUGE and METEOR metrics, while the VideoLLaMA2 8-frame model showed the best performance on the remaining metrics. In addition, increasing the number of input video frames for VideoLLaMA2 from 8 to 16 resulted in a performance decrease across all evaluation metrics. This result aligns with a consistent trend reported in the VideoLLaMA2 paper on open-ended VQA benchmarks. These results highlight the current limitations of open VLMs and indicate the need for methodological improvements for better video understanding.

Regarding proprietary models, the GPT-4o 16-frame model achieved comparable performance to the fine-tuned VideoLLaMA2 with both 8 and 16 frames, across all LLM metrics except for VDEval (i.e., LLM-Acc, LLM-Rel, and LAVE). When compared to VideoLLaMA2 with frozen parameters, both GPT-4o variants showed large improvements in LLM metrics. In contrast, for Gemini-1.5-pro, both 8-frame and 16-frame configurations underperformed across every evaluation metric when compared to fine-tuned VideoLLaMA2, although they surpassed the frozen-parameter models in LLM metrics. However, similar to open-source VLMs, the proprietary VLMs still struggled to reach even 50% on any evaluation metric. This further underscores the difficulty of the VDAct dataset.

Correlation between Human Judgments Table 6 shows the rank correlation coefficients between human assessment and various evaluation metrics, including our proposed VDEval. We can see that LLM-based evaluation metrics showed higher correlations compared to classic metrics and similarity-based metrics, indicating the effectiveness of using LLMs for evaluating dialogue systems. Regarding VDEval, we first found that incorporating session-based context improved correlation coefficients regardless of the use of KG summaries. Regarding KG summaries, incorporating template-based summaries led to a decrease in correlation coefficients for the fine-tuned VideoLLaMA2 with the session-based context, while refined summaries consistently enhanced correlation coefficients across all experimental settings. In particular, the variation that uses both the session-based context and refined KG summaries showed the highest correlation with human evaluation, achieving a significant improvement compared to the existing state-of-the-art metrics, substantiating the effectiveness of VDEval in more accurately reflecting human judgments in video-grounded dialogue evaluation.

Performances on Different Question Types We evaluated the performance of VideoLLaMA2, the best fine-tuned baseline on VDAct, across various question types. Our analysis revealed that the baseline performed less effectively on T-SEQ, T-FRQ, and EXP question types, with average VDEval scores of 1.49 (24.5%), 1.56 (28%), and 1.62 (31%), respectively. In contrast, the average scores for questions not falling into these types were higher at 1.7 (35%), 1.66 (33%), and 1.66 (33%). For QNT questions, the baseline showed better performance, with an average score of 1.78 (39%) compared to 1.65 (32.5%) for questions that do not belong to QNT. This better performance is likely because many QNT questions involve determining a small number of items (1 or 2) as the correct answer. Thus, this simplifies the system's task for the QNT question type. Additionally, the baseline performed better on binary questions as it achieved an average score of 1.78 (39%), compared to 1.59 (29.5%) for other question types.

The Effect of Question and Answer Length We examined how the length of questions and answers affects the performance of the best fine-tuned baseline. Our findings show that the average word counts of reference answers where their generated responses achieved VDEval scores of 1 (0%), 2 (50%), and 3 (100%) are 11.51, 12.83, and 7.52 words, respectively. This indicates that the model struggles to provide accurate responses for more complex questions as they require more elaborate answers. To further support the above claim, the third quartile regarding the word counts of answers for questions with a score of 3 (100%) is 10 words, while for others with scores of 1 (0%) and 2 (50%), the third quartiles are 15 and 16 words, respectively. This suggests that the model's accuracy declines as the length of the reference answers increases.

Conclusion

This paper presents VDAct, a novel dataset for video-grounded dialogue on event-driven activities, and VDEval, a specialized evaluation metric for this task. VDAct features longer and more complex video sequences depicting diverse activity scenarios that demand advanced contextual understanding. Experimental results showed that VDAct task is challenging and includes several question types that are difficult to answer. Additionally, VDEval, which incorporates dialogue session history and video content summaries from KGs for evaluating dialogue responses, demonstrates a significantly higher correlation with human assessments compared to existing metrics. Future work can explore a model architecture that utilizes KGs for video-grounded dialogue.

Acknowledgements

This paper is based on results obtained from: (1) a project, Programs for Bridging the gap between R&D and the IDeal society (society 5.0) and Generating Economic and social value (BRIDGE)/Practical Global Research in the AI × Robotics Services, implemented by the Cabinet Office, Government of Japan, and (2) a project, JPNP20006, commissioned by the New Energy and Industrial Technology Development Organization (NEDO).

In addition, we would like to thank Susan Holm, our knowledge engineer, for her insightful feedback during the data collection process.

References

Alamri, H.; Cartillier, V.; Das, A.; Wang, J.; Cherian, A.; Essa, I.; Batra, D.; Marks, T. K.; Hori, C.; Anderson, P.; et al. 2019. Audio visual scene-aware dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7558–7567.

Anderson, P.; Fernando, B.; Johnson, M.; and Gould, S. 2016. Spice: Semantic propositional image caption evaluation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, 382–398. Springer.

Banerjee, S.; and Lavie, A. 2005. 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*, 65–72.

Chan, D.; Petryk, S.; Gonzalez, J.; Darrell, T.; and Canny, J. 2023. CLAIR: Evaluating Image Captions with Large Language Models. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 13638–13646. Singapore: Association for Computational Linguistics.

Cheng, Z.; Leng, S.; Zhang, H.; Xin, Y.; Li, X.; Chen, G.; Zhu, Y.; Zhang, W.; Luo, Z.; Zhao, D.; and Bing, L. 2024. VideoLLaMA 2: Advancing Spatial-Temporal Modeling and Audio Understanding in Video-LLMs. *arXiv preprint arXiv:2406.07476*.

Das, A.; Kottur, S.; Gupta, K.; Singh, A.; Yadav, D.; Moura, J. M.; Parikh, D.; and Batra, D. 2017. Visual dialog. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 326–335.

Egami, S.; Ugai, T.; Oono, M.; Kitamura, K.; and Fukuda, K. 2023. Synthesizing event-centric knowledge graphs of daily activities using virtual space. *IEEE Access*, 11: 23857–23873.

Good, P. 2000. Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses. *Springer Series in Statistics*.

Jia, B.; Lei, T.; Zhu, S.-C.; and Huang, S. 2022. EgoTaskQA: Understanding Human Tasks in Egocentric Videos. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Lin, B.; Zhu, B.; Ye, Y.; Ning, M.; Jin, P.; and Yuan, L. 2023. Video-LLaVA: Learning United Visual Representation by Alignment Before Projection. *arXiv preprint arXiv:2311.10122*.

Lin, C.-Y. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, 74–81.

Liu, C.-W.; Lowe, R.; Serban, I.; Noseworthy, M.; Charlin, L.; and Pineau, J. 2016. How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation. In Su, J.; Duh, K.; and Carreras, X., eds., *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2122–2132. Austin, Texas: Association for Computational Linguistics.

Maaz, M.; Rasheed, H.; Khan, S.; and Khan, F. S. 2024. Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models. arXiv:2306.05424.

Mañas, O.; Krojer, B.; and Agrawal, A. 2024. Improving Automatic VQA Evaluation Using Large Language Models. In *Proceedings of The Thirty-Eighth AAAI Conference on Artificial Intelligence (AAAI-24)*, 4171–4179.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, 311–318.

Pasunuru, R.; and Bansal, M. 2018. Game-Based Video-Context Dialogue. In Riloff, E.; Chiang, D.; Hockenmaier, J.; and Tsujii, J., eds., *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 125–136. Brussels, Belgium: Association for Computational Linguistics.

Puig, X.; Ra, K.; Boben, M.; Li, J.; Wang, T.; Fidler, S.; and Torralba, A. 2018. Virtualhome: Simulating household activities via programs. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 8494–8502.

Sarto, S.; Barraco, M.; Cornia, M.; Baraldi, L.; and Cucchiara, R. 2023. Positive-augmented contrastive learning for image and video captioning evaluation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 6914–6924.

Vassiliades, A.; Bassiliades, N.; Gouidis, F.; and Patkos, T. 2020. A knowledge retrieval framework for household objects and actions with external knowledge. In *Semantic Systems. In the Era of Knowledge Graphs: 16th International Conference on Semantic Systems, SEMANTiCS 2020, Amsterdam, The Netherlands, September 7–10, 2020, Proceedings 16*, 36–52. Springer International Publishing.

Wada, Y.; Kaneda, K.; Saito, D.; and Sugiura, K. 2024. Polos: Multimodal Metric Learning from Human Feedback for Image Captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13559–13568.

Wang, Y.; Zheng, Z.; Zhao, X.; Li, J.; Wang, Y.; and Zhao, D. 2023. VSTAR: A Video-grounded Dialogue Dataset for Situated Semantic Understanding with Scene and Topic Transitions. In Rogers, A.; Boyd-Graber, J.; and Okazaki, N., eds., *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 5036–5048. Toronto, Canada: Association for Computational Linguistics.

Wu, B.; Yu, S.; Chen, Z.; Tenenbaum, J. B.; and Gan, C. 2021. STAR: A Benchmark for Situated Reasoning in Real-World Videos. In *Thirty-fifth Conference on Neural Information Processing Systems (NeurIPS)*.

Yuan, W.; Neubig, G.; and Liu, P. 2021. BARTScore: Evaluating Generated Text as Text Generation. In Ranzato, M.; Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*, volume 34, 27263–27277. Curran Associates, Inc.

Zhang, T.; Kishore, V.; Wu, F.; Weinberger, K. Q.; and Artzi, Y. 2020. BERTScore: Evaluating Text Generation with BERT. In *International Conference on Learning Representations*.