DVD: A <u>D</u>iagnostic Dataset for Multi-step Reasoning in <u>V</u>ideo Grounded <u>D</u>ialogue

Hung Le[‡§*]**, Chinnadhurai Sankar**[†]**, Seungwhan Moon**[†]**, Ahmad Beirami**[†]**,**
Alborz Geramifard[†]**, Satwik Kottur**[†]

[†]Facebook
{chinnadhurai, shanemoon, beirami, alborzg, skottur}@fb.com
[‡]Singapore Management University
[§]Institute for Infocomm Research, A*STAR
hungle.2018@smu.edu.sg

Abstract

A video-grounded dialogue system is required to understand both dialogue, which contains semantic dependencies from turn to turn, and video, which contains visual cues of spatial and temporal scene variations. Building such dialogue systems is a challenging problem, involving various reasoning types on both visual and language inputs. Existing benchmarks do not have enough annotations to thoroughly analyze dialogue systems and understand their capabilities and limitations in isolation. These benchmarks are also not explicitly designed to minimise biases that models can exploit without actual reasoning. To address these limitations, in this paper, we present ***DVD, a Diagnostic Dataset for Video-grounded Dialogues***. The dataset is designed to contain minimal biases and has detailed annotations for the different types of reasoning over the spatio-temporal space of video. Dialogues are synthesized over multiple question turns, each of which is injected with a set of cross-turn semantic relationships. We use DVD to analyze existing approaches, providing interesting insights into their abilities and limitations. In total, DVD is built from $11k$ CATER synthetic videos and contains 10 instances of 10-round dialogues for each video, resulting in more than $100k$ dialogues and $1M$ question-answer pairs. Our code and dataset are publicly available[1].

1 Introduction

Research in visual question answering (VQA) aims to develop intelligent systems that can reason and answer questions about visual information. Earlier datasets have been introduced to study this problem, focusing on images as the visual input (Antol et al., 2015; Gao et al., 2015; Malinowski

---

[*]Work done when HL was a research intern at Facebook.
[1]github.com/facebookresearch/DVDialogues

Q1: until the end of the cube 's rotation , what types of actions does the big thing undertake the most ? **A1**: flying
Q2: during the same time period , how many sliding objects are there ? **A2**: 2
Q3: among them , there is a ball . during the whole video , what type of action does it undertake second ? **A3**: no action
Q4: how about up until now ? **A4**: sliding
Q5: during the red thing 's last slide , how many things are behind the earlier mentioned large object ? **A5**: 2
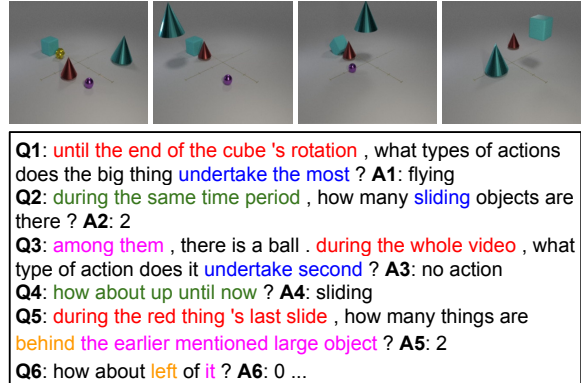**Q6**: how about left of it ? **A6**: 0 ...

Figure 1: **Example DVD dialogue:** We demonstrate an example dialogue in DVD that tests various aspects, including action recognition, temporal reasoning, spatial reasoning, video interval tracking, and dialogue object tracking. Q_i/A_i: question/answer of turn i.

and Fritz, 2014; Zhu et al., 2016) Recently, many QA benchmarks have been proposed to extend the visual information from the image to video domain (Jang et al., 2017; Lei et al., 2018; Zadeh et al., 2019). While image QA problems require a system to learn cross-modality interaction, video QA problems go beyond and capture visual information with temporal variance.

As an orthogonal extension from VQA problems, another line of research investigates image/video QA in a dialogue setting (Das et al., 2017; Seo et al., 2017; De Vries et al., 2017; Chattopadhyay et al., 2017; Alamri et al., 2019). In this problem, questions about a given video or image are positioned in a multi-turn dialogue. In each dialogue turn, a question usually exhibits different types of cross-turn relations to other questions in prior dialogue turns, such as object co-reference and topic alignment. In this work, we investigate the problem of multi-turn video question answering (QA), also known as video-grounded dialogue.

Numerous approaches to video-grounded dialogue have shown remarkable performance in build-

ing intelligent multimodal systems (Hori et al., 2019; Schwartz et al., 2019; Le et al., 2019; Li et al., 2020; Le et al., 2020). However, most of these methods exhibit marginal performance gain, and our ability to understand their limitations is impeded by the complexity of the task. Existing benchmarks are not designed with enough information to determine whether current approaches are capable of sophisticated reasoning and not just exploiting biases, which has been a common concern in vision-language systems (Agrawal et al., 2016; Goyal et al., 2017; Qi et al., 2020).

To address the limitations of existing benchmarks and analyze dialogue systems more efficiently, we propose *DVD, a Diagnostic Dataset for Video-grounded Dialogues*. We demonstrate an example dialogue in DVD in Figure 1. From scene graphs and object action annotation of a CATER video (Girdhar and Ramanan, 2020), we simulate questions based on reasoning structures, also known as functional programs in CLEVR (Johnson et al., 2017). Compared to CLEVR, we introduced 17 novel functional modules, designed for video and dialogue input components. As illustrated in Figure 1, at each dialogue turn, a DVD question tests dialogue systems to perform different types of reasoning on videos, such as action recognition and spatio-temporal reasoning. Across turns, we generate questions to be related to each other by incorporating different types of semantic relationships, including: (1) temporal relation, which requires a system to learn to localize different temporal segments of the video from turn to turn; (2) object reference, which requires a system to resolve visual objects mentioned throughout the dialogue history in either short-term references (pronouns) or long-term references (e.g. "the earlier mentioned large object"); and (3) topic transfer, which requires a system to maintain a memory of the last question turn to solve the question in the current turn.

On DVD, we trained a set of baseline methods and analyzed the results by several aspects of visual and linguistic complexity (Section 4). We found that these methods struggle on questions requiring both video temporal and spatial localization. They are also vulnerable to long-term reasoning in both videos and dialogues as they are not designed to track active visual objects or relevant video segments throughout dialogue context. We hope the DVD dataset will lead to new research avenues to develop intelligent systems capable of complex

reasoning on video and dialogue medium (further discussion in the Supplementary Material). The DVD dataset and code will be made public.

2 Related Work

We compared DVD to existing datasets from the following four angles:

1) Vision-linguistic. Vision-linguistic understanding benchmarks have been proposed, including captioning (Farhadi et al., 2010; Lin et al., 2014; Rohrbach et al., 2015), phrase grounding or object reference (Kazemzadeh et al., 2014; Plummer et al., 2015), scene graph learning (Krishna et al., 2017), and text-to-clip (Anne Hendricks et al., 2017). Our benchmark, DVD, is more related to VQA in which a visual input is given and a system is required to answer a question about this input (Antol et al., 2015; Zhu et al., 2016; Jang et al., 2017; Lei et al., 2018). Another related line of research is the research of navigation systems in a physical environment (Gordon et al., 2018; Wijmans et al., 2019). Compared to the prior benchmarks, one major difference of DVD is the extension of single-turn interaction to a multi-turn human-machine dialogue.

2) Visually-grounded Dialogue. Extended from the vision-linguistic understanding research, this line of research focuses on answering questions sequentially positioned over multiple turns (De Vries et al., 2017; Das et al., 2017; Chattopadhyay et al., 2017; Hori et al., 2019; Thomason et al., 2019). A system has to understand the dialogue context and resolve cross-turn semantic dependencies. However, due to the complexity of the tasks, involving cross-modality and cross-turn information, prior benchmarks are often subject to bias that models often exploit without actual reasoning (Qi et al., 2020). In this work, we design a diagnostic benchmark with minimal bias and incorporate a set of specific reasoning requirements.

3) Diagnostic. Our work is related to MNIST Dialogue (Seo et al., 2017) and CLEVR Dialog (Kottur et al., 2019). They involve synthetic images to develop image-grounded dialogues. Compared to them, DVD questions are extended from the image to the video domain and injected with more diverse cross-turn semantics. As shown in Table 1 DVD contains a higher proportion of unique questions than related benchmarks. DVD is also inspired by the dialogue state tracking task (DST) (Mrkšić et al., 2017; Bordes et al., 2017; Kottur et al., 2021; Moon et al., 2020). DST requires a system to detect

Split	#Videos/ Images	#Dialogs	#Questions	# Unique Questions
DVD-Train	6,157	61,551	615,510	360,334
DVD-Val	1,540	15,396	153,960	99,211
DVD-Test	3,299	32,978	329,780	200,346
DVD-Total	10,996	109,925	1,099,250	620,739
CLEVR	100K	N/A	1M	854K
CLEVRER	20K	N/A	305K	26.4K
VisDial	123K	123K	1.2M	380K
AVSD	11.1K	11.1K	101.2K	59K
MNIST Dialog	50K	150K	1.5M	355
CLEVR Dialog	85K	425K	4.25M	73K

Table 1: **Statistics for DVD**: Compared to synthetic dialogue benchmarks, MNIST Dialog and CLEVR Dialog, majority of questions in DVD are unique. Questions are generated from question templates and incorporated with various cross-turn semantics.

all information slots mentioned in dialogue, such as restaurant name and booking date. Instead, in DVD, for each turn, we introduce an object tracking state, defined as visual objects and their attributes mentioned in dialogue context.

4) Multi-step reasoning. A multi-step reasoning question is typically represented by a reasoning structure, also known as functional programs. Earlier efforts (Andreas et al., 2016; Johnson et al., 2017) designed questions that are expressed as elementary operation programs. More related to our work, Song et al. (2018); Yi* et al. (2020) extended the prior work to the video domain with questions focusing on the temporal variance of video frames. A major difference between our work and these approaches is the extension of functional programs to a dialogue task with context-based operations, such as object tracking and interval tracking. This extension brings a step toward more transparent dialogue systems capable of performing reasoning operations across question turns.

3 The DVD Dataset

Our benchmark provides a dataset that can be used to conduct rich diagnostics to better understand the reasoning capabilities of dialogue systems. Table 1 and Figure 3 to 6 give an overview of DVD.

3.1 Objects, Spatial Relations, and Intervals

Objects. Objects are identified by their attributes, including object shapes, sizes, materials, and colors. One unique characteristic of CATER objects is that each object can move multiple times in a single video. From the CATER universe, we define 4 types of object actions: "flying", "rotating", "sliding", and "no action" (object being stationary). Another characteristic of CATER objects is that one
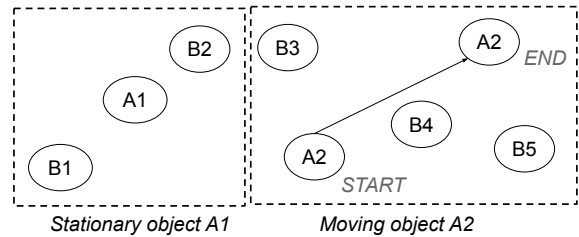

Figure 2: **Example spatial relationship**: We demonstrate the projection of objects and their movements on the ground plane. Considering the "left" relationship, "A1 is left of B2" and "A2 is left of B5".

object can be contained by another object, resulting in a visual problem called *object containment*. In our experiments, current dialogue systems are still vulnerable to this problem, making it hard to apply to the open world (See Section 4.3).

Video intervals. We define video intervals as continuous video frames, limited by a start and end point, each of which can be the start or end of an object's action or the start or end of the whole video. We formulate two types of video intervals:

1) Atomic intervals. In these intervals, all objects have at most one action and they can be in only one of the two states: in motion or stationary. To find atomic intervals, we simply collate the start and end timestamps of all object actions in a CATER video and sort them chronologically. By definition, any non-overlapping interval between two timestamps is considered atomic. This constraint allows us to identify the *relative spatial relationships* ("left", "right", "behind", and "front") between any two objects by using their coordinates at the start and end of the interval. Note that in the CATER universe, all actions can be projected either as a straight line ("flying" and "sliding") or a single point ("rotating" and "no action"). Practically, we focus on spatial reasoning only when one of the two objects is stationary. Figure 2 demonstrates the "left" spatial relation, and Figure 3 (Top) shows an example question of atomic interval with spatial relation.

2) Compositional intervals. Compositional intervals are all other intervals that are not atomic. In these intervals, an object can have more than one actions, i.e. be in more than one states such as "flying" then "no action". Therefore, its movement projections are not linear and we do not identify spatial relations in these cases. Instead, we focus on information such as *action set* and *action sequence* to generate questions. Figure 3 (Bottom) presents an example question of compositional interval.

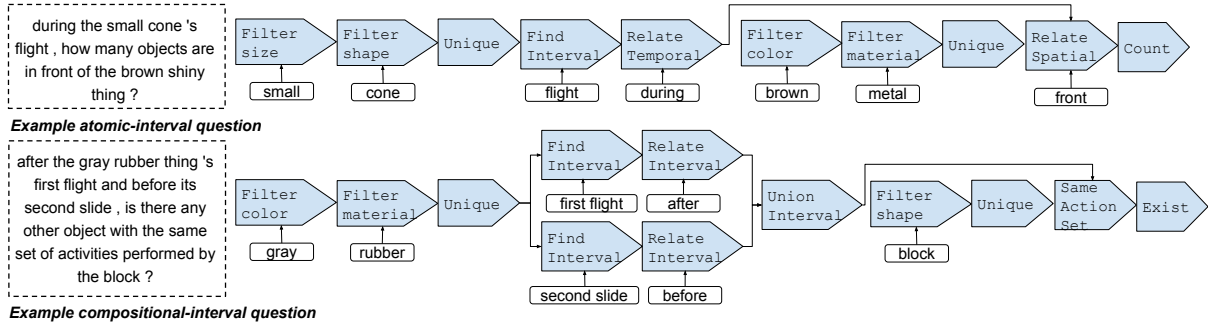To create DVD questions, we first identify all in-

Figure 3: **Example questions and their functional programs:** Top: A question of atomic interval with relative spatial relationship. Bottom: A question of compositional interval with action set comparison semantic.
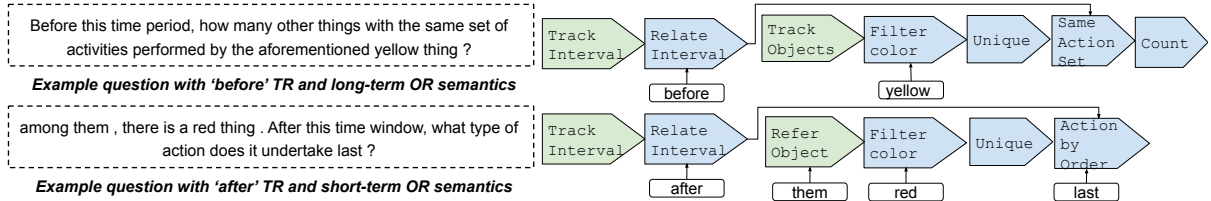

Figure 4: **Examples questions positioned in dialogue and their functional programs**: Each question contained references to past dialogue turns, through video temporal relation (TR) or dialogue object reference (OR).

tervals in a video (with a minimum duration of about 0.5s), then randomly sample one interval, and proceed to create questions based on object movements and locations in this interval. Figure 5-(a) shows the percentages of DVD questions by video interval types. Overall, more than 60% of questions are of compositional intervals and among the atomic-interval questions, the majority of them contain a spatial relation. We still maintain a small percentage of temporal-agnostic instances ("none" type) to keep the dialogue flow natural.

3.2 Question and Dialogue Generation

Question representation. We use question templates to materialize questions in natural language. Each template associates with an applicable type of video interval and a functional program. Compared to CLEVR functional programs (Johnson et al., 2017), we introduce 17 new functional modules, of which 13 are extended for video-based inputs and 4 are extended for dialogue-based inputs. Overall, we utilize 26 question templates for 8 question types. Figure 3 illustrates two sample questions with corresponding reasoning structures and Figure 5-(b) shows the statistics of question type distribution. Please refer to the supplementary material for the full details of functional modules and question types and examples.

Dialogue Generation. We generated dialogues with a fixed length of 10 turns. In each turn, we adopted a Depth First Search (DFS) approach, as

similarly used in CLEVR (Johnson et al., 2017), to instantiate questions by sequentially traversing and executing functional programs. To generate linguistic dependencies between dialogue turns, at each turn, we randomly sample and incorporate one or more of the 3 semantic relations below. Figure 4 and 6 present examples of 2 questions and a dialogue with these semantic relations.

Type I: Video Temporal Relation (TR): This type of semantic relation tests a system to localize video intervals in relation to past dialogue turns. We randomly select one of three types of relation: (1) *"during"* relation reuses the same time interval as the last dialogue turn, e.g. the Q4 in Figure 6; (2) *"before"* and (3) *"after"* relations simulate a dialogue flow with references to the earlier and subsequent video segments. TR synthesizes scenarios when humans either maintain or shift their attention temporally from one video segment to a related part.

Type II: Dialogue Object Reference (OR): We incorporate object references into a question by replacing original object phrase, such as "the large rubber cone", with pronouns, such as "it", to refer to object(s) mentioned in the earlier part of the dialogue. The distance of reference is one turn and we call this a *short-term memory* OR. Additionally, we simulate *long-term memory* OR by injecting unique objects mentioned further in the past dialogue turns. We simulate this behavior by maintaining a *dialogue object state* at each turn. To choose an object for references, we randomly

(a) Question distribution by video interval

(b) Question distribution by question type

(c) Number of active objects per dialogue turn

(d) Question distribution by turn distance of object references

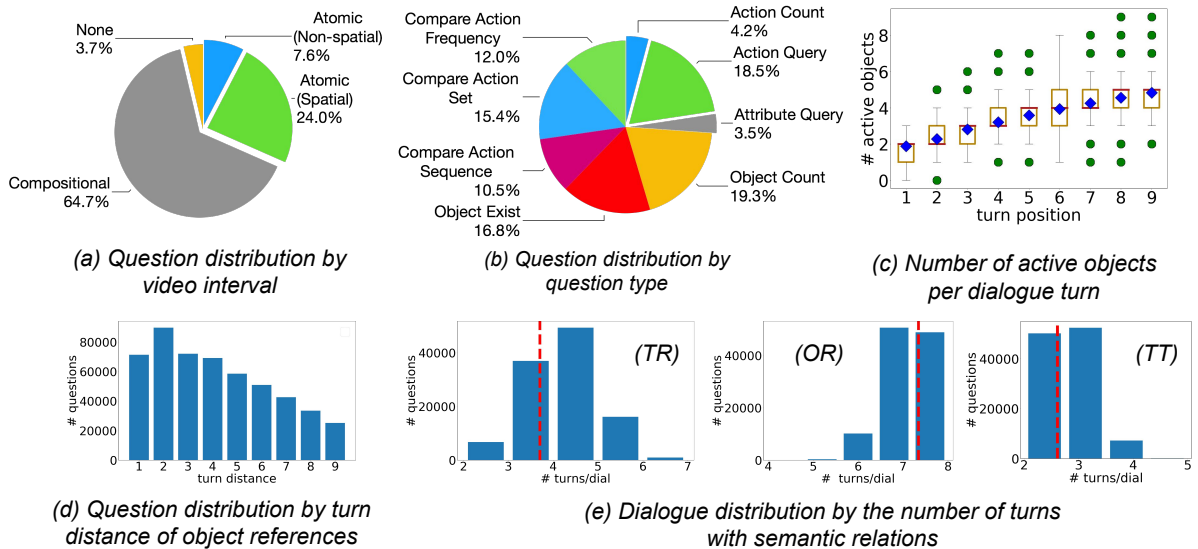*(e) Dialogue distribution by the number of turns with semantic relations*

Figure 5: **Data analysis of DVD:** (a) and (b): Questions are distributed by 8 question types and 3 video interval types. (c): The boxplot displays the distribution of active objects mentioned in each dialogue turn position. (d): At turn position i, an old object originally mentioned in a prior turn position j might be reused, resulting in reference of turn distance $i - j$. (e): Each dialogue turn is incorporated with semantic relations, including TR (temporal relation), OR (object references), and TT (topic transfer). The dotted line indicates the overall average.

sample a past dialogue turn position and sample an object introduced in this turn. This object then replaces the original object phrases in the question of the current turn. For example, in question Q3 in Figure 6, "the earlier mentioned small thing" is identified from the object originally introduced in Q1. Following this method, our dialogue simulates scenarios in which humans only focus on a subset of objects rather than all objects in the video scene and they can refer to those objects again over multiple dialogue turns. Figure 5-(c) displays the boxplot of the number of active objects involved in each turn position. Out of 10 objects (the maximum number of objects in a CATER video), 2 to 5 objects are involved on average per dialogue. Figure 5-(d) shows the question distribution by the turn distance of long-term memory OR, with the majority of questions containing 2-turn distance references.

Type III: Topic Transfer (TT): This relation tests the model ability to memorize and reuse the context of the last dialogue turn to the current turn through 3 types of topic transfers: (1) Attribute transfer and (2) spatial transfer reuse the same question from the prior dialogue turn with a modification of object attribute or spatial relation (e.g. Q2 and Q5 in Figure 6). Compared to TR, these two types of topic transfers focus on human attention shifts in spatial space rather than temporal space; (3) Temporal transfer introduces a unique setting of *situated dialogue*

in DVD. Instead of using a fixed video input for each dialogue instance, at the first dialogue turn, we shorten a CATER video by a cutoff point, e.g. T_0. At each later turn, for 30% of time, we update the current video input to a new cutoff point later than the previous one e.g. $T_{i+1} > T_i$. We do not update when the cutoff reaches the end of the original CATER video $\mathcal{T}$ i.e. $T_{i+1} = \mathcal{T}$. For instance, in Figure 6, at Q7, we reuse the same context from Q6 but with new extended visual content. We introduce temporal transfer as a preliminary step to challenge dialogue systems in a dynamic environment with a continuous visual stream.

After sampling question templates and semantic dependencies, the ground-truth answers are obtained by executing corresponding functional programs. For each question template, we discard dominating instances to maintain an approximate uniform distribution of answer values, minimizing bias resulting from question-conditioned data distributions. Additionally, at each turn, we remove any question that is ill-posed or becomes redundant when positioned in dialogue. For instance, the question *"how many red rubber objects are there?"* is removed if in a prior dialogue turn, the question is *"how many red objects are there?"* and the answer is already *"1"*. To do this, we perform a check at every dialogue turn to determine whether involving objects and their attributes are already mentioned in the dialogue object state. Finally, we only keep dia-

Dialogue	Dialogue Object State	TR	OR	TT A	TT S	TT T
Q1: before the large thing 's first flight , what color is the average thing that is in front of the small thing? **A1**: yellow						
Q2: what about its material ? **A2**: rubber	{obj1: size=large}, {obj2: size=average, color=yellow}, {obj3: size=small}		✓	✓		
Q3: during the earlier mentioned small thing 's first slide , what shape is the stationary thing to the right of the aforementioned average object? **A3**: cube	{obj1: size=large}, {obj2: size=average, color=yellow, material=rubber}, {obj3: size=small}			✓		
Q4: during the same time period , how many average cyan shiny things are behind the gray object? **A4**: 1	{obj1: size=large}, {obj2: size=average, color=yellow, material=rubber}, {obj3: size=small}, {obj4: shape=cube}	✓				
Q5: how about to the left of it ? **A5**: 0	{obj1: size=large}, {obj2: size=average, color=yellow, material=rubber}, {obj3: size=small}, {obj4: shape=cube}, {obj5: color=gray}, {obj6: color=cyan, size=average, material=metal}	✓	✓		✓	
Q6: throughout the whole video , does the earlier cube object fly more frequently than the earlier mentioned average object slides ? **A6**: True	{obj1: size=large}, {obj2: size=average, color=yellow, material=rubber}, {obj3: size=small}, {obj4: shape=cube}, {obj5: color=gray}, {obj6: color=cyan, size=average, material=metal}			✓		
Q7: what about up until now ? **A7**: False	{obj1: size=large}, {obj2: size=average, color=yellow, material=rubber}, {obj3: size=small}, {obj4: shape=cube}, {obj5: color=gray}, {obj6: color=cyan, size=average, material=metal}					✓

Figure 6: **Dialogue generation:** In each dialogue turn, we generate questions with randomly sampled cross-turn dependencies: temporal relation (TR), object reference (OR), and topic transfers (TT), including attribute (A), spatial (S), and temporal (T) transfer. We maintain a dialogue object state of active objects which are color-coded.

logues that have cross-turn dependencies in 9 out of 10 turns, considering the first turn semantically independent. Figure 5-(e) provides the distribution of dialogues by the number of TR, OR, and TT relations. For more analysis of DVD, please refer to the supplementary material.

4 Dialogue Systems on DVD

The video-grounded dialogue task in DVD is defined as a turn-based retrieval task from multiple-choice candidate answers. At each dialogue turn i ($i = 1, 2, ..., 10$), video input V_i, the ground-truth dialogue context, including question and answer pairs up to the last dialogue turn, $\mathcal{C}_i = (\mathcal{Q}_k, \mathcal{A}_k)|_{k=1}^{k=i-1}$, the question of the current turn $\mathcal{Q}_i$, are provided. The system is given a set of candidate answers $\mathcal{A}$, predefined as all possible answer values for all question types, with $|\mathcal{A}| = 40$ in DVD, and is required to select one answer from $\mathcal{A}$. We evaluate models by the accuracy of predicted answers against the ground-truth answers. For a system denoted as θ, the objective function is: $\hat{\mathcal{A}}_i = \arg\max_{\mathcal{A}} P(\mathcal{A}_i | \mathcal{V}_i, \mathcal{Q}_i, \mathcal{C}_i; \theta)$.

4.1 Experimental Setup

Baselines. We experimented with a representative set of baseline approaches on DVD, including: (1) *Answer Prior*, which selects the most popular answer option as predicted answers; (2) *Q-type (Random/Frequency)*, which assume known question types and select a random or most popular answer from the corresponding answer space; (3) *Q-retrieval (TF-IDF)*, which retrieves the most similar question from the training set and use its

answer as the predicted answer; (4) *RNN(Q)* and *HRNN(C+Q)*, which encode dialogue-only components without seeing visual information to predict answers; (5) *HRNN(C+Q)+CNN(V)/TA(V)*, same as (4) but with access to visual information which is encoded by pretrained CNN models and temporal attention (TA) (Jang et al., 2017; Lei et al., 2018; Hori et al., 2019); (6) *TF(C+Q+V)*, which uses a Transformer-based architecture to encode visual and language information (Schwartz et al., 2019; Le et al., 2019; Li et al., 2020). Finally, we conducted internal human evaluation on a subset of the DVD test split. For each test sample, a human received an input video, dialogue history, and the question for the current turn. The human was required to select an answer from the list of 40 candidates $\mathcal{A}$ to answer the question.

Experiments. Video-grounded dialogues entail a lot of visio-linguistic and reasoning challenges that are not easy to be studied in isolation using existing datasets. To address this issue with DVD, we exploit the rich annotations of DVD in our experiments during evaluation. We designed our experiments to systematically analyze model capabilities and shortcomings through unique challenges in video-grounded dialogue systems. Specifically, in Section 4.2, we analyzed the results of all models overall as well as by each question type. In Section 4.3, we leverage the spatio-temporal annotation of visual objects to analyze model performance by related video interval types, spatial reasoning (results by object containment), and temporal reasoning (results by relative interval length). In terms of dialogue contextual complexity, in Section 4.4,

Accuracy	Answer Prior	Q-type (Random)	Q-type (Freq)	Q-retrieval (TF-IDF)	RNN (Q)	HRNN (C+Q)	HRNN (C+Q)+ CNN(V)	HRNN (C+Q)+ TA(V)	TF (C+Q +V)	Human
All	21.3	27.8	35.3	32.1	39.7	45.8	49.3	50.2	51.1	89.3
Action count	0.0	9.3	23.4	19.8	16.3	28.2	37.8	36.0	38.8	87.5
Action query	0.0	12.7	23.7	20.6	25.8	33.1	36.7	38.6	39.4	88.1
Attribute query	0.0	32.9	38.7	39.4	38.1	39.2	43.3	45.1	43.1	98.0
Compare action seq	33.4	34.1	37.3	35.1	45.5	52.5	58.2	57.5	61.6	91.5
Compare action set	25.1	28.2	36.3	28.2	32.8	40.0	43.0	44.3	45.4	82.9
Compare action freq	48.5	50.0	50.5	44.4	58.4	56.9	62.3	65.2	67.1	88.5
Object count	0.0	9.1	23.3	18.8	26.2	38.6	40.0	40.2	39.9	90.6
Object exist	48.9	49.8	51.1	54.4	66.4	67.0	69.2	69.4	69.0	92.3
None	0.0	32.1	38.3	39.0	38.3	39.5	43.1	45.1	43.4	99.1
Atomic (non-spatial)	18.8	26.3	31.9	42.4	47.2	47.8	49.9	50.7	48.9	83.3
Atomic (spatial)	21.2	27.3	35.5	27.6	36.8	46.0	47.5	47.6	47.1	93.9
Compositional	22.8	28.0	35.4	32.1	40.0	45.8	50.2	51.4	53.2	87.1
Transfer (attribute)	0.0	30.7	45.5	37.1	40.8	45.7	54.5	57.3	57.7	100.0
Transfer (spatial)	49.8	42.4	44.9	26.4	29.6	48.1	47.7	47.4	48.0	90.5
Transfer (temporal)	28.9	38.4	22.6	3.0	30.2	53.5	62.2	64.6	69.0	79.8

Table 2: **Experiment results on the DVD test split**: Models are evaluated by overall accuracy and by question types (Top), accuracy by video intervals in question (Center), and transferability accuracy (Bottom).

we use cross-turn relation annotations to analyze model performance by temporal-based attention shift (TR), dialogue turn distance (OR), and short-term transferability (TT).

4.2 Results

From Table 2 (Top), we observe that "blind" systems that use answers only or questions only, achieve quite poor results up to 39% accuracy. By selecting the most popular answer option, *Answer Prior* only achieves 21% accuracy. When a "blind" model has access to dialogue history, the performance increases up to 45%. This increment shows that dialogue context contains useful information for a dialogue system to infer answers. We note that on average there are nearly 3 out of 10 question turns with a topic transfer per dialogue (see Figure 5-(e)). In such cases, a model can randomly make a good guess by just reusing the answer of the last question turn. When a system is presented with the visual input, we observe model performance increases up to 51%. However, in the best system, the performance is still far below the human level with a performance gap of 38 absolute points.

In Table 2 (Top), from the results of *Q-type(Random)* per question type, we observed that answers are balanced in each question type. The table also shows performance drops between pairs of object-oriented vs. action-oriented question types. For instance, *TF(C+Q+V)* achieves 38% accuracy in Action count vs. 39% in Object count, and 39% accuracy in Action query vs. 43% in Attribute query. In comparison-based questions, comparing action sets tend to be more challenging than comparing action sequences. To compare action sets of two objects in a video interval, a system needs to process the interval completely. However, to compare action sequences, in most cases, the system can determine the answer after the first few action steps the objects perform. For more analysis of question types and sub-types, please refer to the supplementary material.

4.3 Analysis by Visual Complexity

To understand the drive of the performance by visual inputs, we investigated the results by the visual complexity in questions. In Table 2 (Center), compared to *HRNN(C+Q)+CNN(V)*, models using attention, either through *TA(V)* or Transformer, show more improvement in compositional interval questions with increments up to 3 absolute points. In other types of intervals, the performance gains are not very significant. Particularly, in atomic-interval questions that require spatial localization, the performance does not change when applying attention. This observation necessitates systems that focus on both spatial and temporal space of visual inputs.

In Figure 7 (Left), we analyzed model performance by the number of objects mentioned in questions that are contained in video scenes. We noted that current models are vulnerable to visual object containment, as the accuracy decreases by the number of contained objects. This observation is consistent with the results of CATER action recognition tasks (Girdhar and Ramanan, 2020). In Figure 7 (Right), we investigated model performance
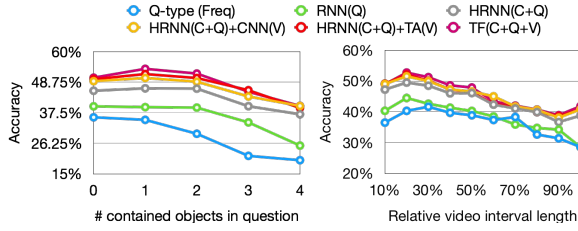
Figure 7: **Experiment results by visual properties**: Left: results by the number of objects mentioned in question that are contained in video scenes. Right: results by the relative length of video interval in question.

by the relative length of ground-truth video interval in question, measured as the percentage of the whole video length. To make a fair analysis, we removed cases in which a question can be solved correctly without localizing the specific video interval but simply using the whole video. We observed that model performance decreases as the interval length increases, demonstrating the challenge of long-term video understanding in video scenes. We noted that there is a drop in performance in the lowest range of interval lengths, $0-10\%$. As this range often represents atomic intervals, the majority of which include questions with spatial relations, systems are negatively affected and the curve drops initially in this low range.

4.4 Analysis by Cross-turn Relations

We examined model performance in a multi-turn setting by cross-turn semantic relations. First, we investigated the effect of TR. In a TR-injected question, a system is required to learn to retrieve a video segment related to the last used segment. However, some questions may be correctly answered without localizing the correct segments. For instance, at the current dialogue turn, a question is of interval (t_m, t_n) and at the next turn, a question with an "after" TR is of interval (t_n, t_q) (s.t. $t_m < t_n < t_q$) might be solved if the visual context is the same in both intervals. We separate such question turns and measured the results of the remaining questions with TR relations "after" and "before". From Figure 8, we observed that current systems are not optimal to learn to shift attention to related intervals, depending on the type of questions. In action-based questions (AC, AQ, CASeq, CASet, and CAF), the results of "before" and "after" TR are lower than those without a TR relation, but in object-based questions (OC, OE), we observed differently. This difference can be explained by the dynamics of actions vs. objects. Between video intervals, information about object actions (e.g. fre-
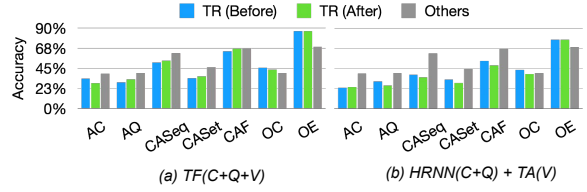

Figure 8: **Experiment results by temporal relations**: Action count (AC), Action query (AQ), Attribute query (AttQ), Compare action sequence (CASeq), Compare action set (CASet), Compare action frequence (CAF), Object count (OC), and Object exist (OE).
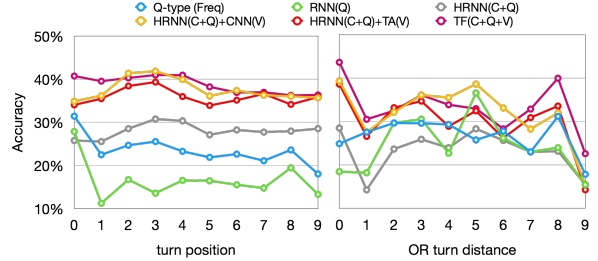

Figure 9: **Experiment results for cross-turn reasoning**: Results of Action count questions by turn position (Left) and by turn distance of object references (Right).

quency, types) tends to change more easily than objects themselves. Action-based questions challenge systems through cross-turn temporal reasoning more than object-based questions.

Secondly, we analyzed the impacts of *long-term memory* OR. From Figure 9 (Left), we noticed that model performance becomes more stable in systems where dialogue history is introduced as an input. For instance, compared to *RNN(Q)*, the performance curve of *TF(C+Q+V)* follows a more gentle downward trend from low to high dialogue turn positions. To fairly analyze performance by OR turn distance, we discard any instances that do not require systems to use dialogue context to resolve the references, but simply rely on the input video. For example, a question with a reference "the earlier mentioned red object" is removed if there is indeed only one "red object" in the video scene. From results by OR turn distance in Figure 9 (Right), we observed all systems are relatively unstable, even as dialogue history is introduced as an input. This difference against the results by turn position exhibits a limitation of current systems as they struggle to resolve object references by existing dialogue encoding techniques.

Finally, to analyze the effect of TT relations, we investigate a new metric, called *transferability*, in Table 2 (Bottom). When a system is presented with a question turn with a topic transfer, it should learn

to derive the answer in relation to the context of the last dialogue turn. If the last answer is right, an intelligent system should be able to consistently answer in the current turn correctly. For instance, given a question-answer pair "what is the color of the sliding cube? red", a human can often infer the answer to a TT(A)-injected question "what about its material?" based on the same visual object. We gather questions that precede questions containing topic transfers and call this set $\mathcal{Q}_{\text{prior}}^{\text{tt}}$. For each question $q_{\text{prior}}^{\text{tt}}$ that the model answered correctly, we measure the accuracy over the corresponding transferred question q_{tt} and average the scores. We observed a clear performance gain from *RNN(Q)* to *HRNN(C+Q)* in terms of transferability metric, demonstrating the impacts of dialogue context on TT questions. A chance-based system can achieve approximately 50% transferability by just recycling answers from prior turns. The best system results, however, are still far from human-level performance. This observation necessitates systems designed with a better contextual memory to adapt past context in new dialogue turns.

5 Discussion and Conclusion

We have introduced DVD, a diagnostic dataset designed to analyze video-grounded dialogue systems. DVD dataset is generated with tight control of data bias through balancing the question and answer distribution and questions are built based on a principled approach to reflect the complexity in videos and dialogues. Our results have shown that DVD can provide interesting insights into system abilities and limitations. Specifically, our analysis has revealed some key shortcomings of current models, including: (1) limited ability to efficiently integrate visual information from both spatial and temporal space; (2) limited ability to recognize and compile multiple actions in long-ranged video intervals; (3) inconsistent performance across dialogue turns, especially in cases when systems are required to switch attention temporally; and (4) unstable performance to resolve object co-reference in the dialogue context, especially when the turn distance of the object references increases.

These insights provide potential avenues where we hope DVD will be a useful benchmark to explore new ideas. Specifically, we discuss two research directions:

Dialogue object tracking. To further diagnose a dialogue system, we aim to study their *long-term* *memory* reasoning ability to track objects and their attributes mentioned in the dialogue context. We are inspired by research work of dialogue state tracking in task-oriented dialogues (Bordes et al., 2017) and propose to use tracking accuracy metric in video-grounded dialogue systems. At each turn t, a video-grounded dialogue system should be able to track and update a dialogue state $\mathcal{S}_t$, defined as a set of all mentioned objects o_i^t and their attributes, including sizes z_i^t, colors c_i^t, materials m_i^t, and shapes s_i^t: $\mathcal{S}_t = (o_1^t, o_2^t, ...) = ((z_1^t, c_1^t, m_1^t, s_1^t), (z_2^t, c_2^t, m_2^t, s_2^t), ...)$. We define two tracking metrics, including *joint accuracy*, measuring the accuracy of prediction of all objects and attributes as a set, and *slot accuracy*, measuring the accuracy of predicted attributes individually. The introduction of these evaluation metrics necessitates a new learning task, dialogue object tracking (DOT) in video-grounded dialogue systems, to better understand current systems' long-term reasoning ability.

Video interval tracking. Another aspect of dialogue systems that we want to diagnose is their ability to localize video segments in a multi-turn setting. Each question turn often focuses on different parts of the video as the dialogue extends over time. It is important to learn how a system can localize the right segments of the video from turn to turn. Similar to DOT, we define a new learning task for video interval tracking (VIT) in a similar nature as text-to-clip tasks (Anne Hendricks et al., 2017). The task can be defined as a ranking task of segment candidates to choose the relevant segments in each question turn. This task is evaluated by ranking metrics such as Rank@1 or Rank@2, and *mean intersection over union* (mIoU). Alternatively, we can adapt *grounding*, a simple metric used by Hudson et al. (2019) to assess spatial attention of image regions. in DVD, *grounding* can be used in temporal attention-based approaches to determine model ability to localize the right position of video intervals in question.

Finally, we want to emphasize that DVD is designed as a synthetic dataset for diagnosis purposes to systematically evaluate model capabilities. The benchmark should not be used to replace data of human dialogues but be used to supplement real-world dialogue datasets.

References

Aishwarya Agrawal, Dhruv Batra, and Devi Parikh. 2016. Analyzing the behavior of visual question answering models. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1955–1960, Austin, Texas. Association for Computational Linguistics.

Huda Alamri, Vincent Cartillier, Abhishek Das, Jue Wang, Anoop Cherian, Irfan Essa, Dhruv Batra, Tim K Marks, Chiori Hori, Peter Anderson, et al. 2019. Audio visual scene-aware dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7558–7567.

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. 2016. Neural module networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 39–48.

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. 2017. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pages 5803–5812.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Antoine Bordes, Y-Lan Boureau, and Jason Weston. 2017. Learning end-to-end goal-oriented dialog. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net.

Prithvijit Chattopadhyay, Deshraj Yadav, Viraj Prabhu, Arjun Chandrasekaran, Abhishek Das, Stefan Lee, Dhruv Batra, and Devi Parikh. 2017. Evaluating visual conversational agents via cooperative human-ai games. In *Proceedings of the Fifth AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 326–335.

Harm De Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. 2017. Guesswhat?! visual object discovery through multi-modal dialogue. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5503–5512.

Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. 2010. Every picture tells a story: Generating sentences from images. In *European conference on computer vision*, pages 15–29. Springer.

Haoyuan Gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei Xu. 2015. Are you talking to a machine? dataset and methods for multilingual image question. *Advances in neural information processing systems*, 28:2296–2304.

Rohit Girdhar and Deva Ramanan. 2020. Cater: A diagnostic dataset for compositional actions and temporal reasoning. In *International Conference on Learning Representations*.

Daniel Gordon, Aniruddha Kembhavi, Mohammad Rastegari, Joseph Redmon, Dieter Fox, and Ali Farhadi. 2018. Iqa: Visual question answering in interactive environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4089–4098.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6904–6913.

C. Hori, H. Alamri, J. Wang, G. Wichern, T. Hori, A. Cherian, T. K. Marks, V. Cartillier, R. G. Lopes, A. Das, I. Essa, D. Batra, and D. Parikh. 2019. End-to-end audio visual scene-aware dialog using multimodal attention-based video features. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2352–2356.

Drew Hudson, A., and Christopher D. Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. 2017. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2758–2766.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2901–2910.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798.

Satwik Kottur, Seungwhan Moon, Alborz Geramifard, and Babak Damavandi. 2021. SIMMC 2.0: A task-oriented dialog dataset for immersive multimodal conversations. *CoRR*, abs/2104.08667.

Satwik Kottur, José M. F. Moura, Devi Parikh, Dhruv Batra, and Marcus Rohrbach. 2019. CLEVR-dialog: A diagnostic dataset for multi-round reasoning in visual dialog. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 582–595, Minneapolis, Minnesota. Association for Computational Linguistics.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73.

Hung Le, Doyen Sahoo, Nancy Chen, and Steven Hoi. 2019. Multimodal transformer networks for end-to-end video-grounded dialogue systems. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5612–5623, Florence, Italy. Association for Computational Linguistics.

Hung Le, Doyen Sahoo, Nancy Chen, and Steven C.H. Hoi. 2020. BiST: Bi-directional spatio-temporal reasoning for video-grounded dialogues. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1846–1859, Online. Association for Computational Linguistics.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. 2018. TVQA: Localized, compositional video question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1369–1379, Brussels, Belgium. Association for Computational Linguistics.

Zekang Li, Zongjia Li, Jinchao Zhang, Yang Feng, Cheng Niu, and Jie Zhou. 2020. Bridging text and video: A universal multimodal transformer for video-audio scene-aware dialog. *DSTC Workshop @ AAAI*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Mateusz Malinowski and Mario Fritz. 2014. A multi-world approach to question answering about real-world scenes based on uncertain input. *Advances in neural information processing systems*, 27:1682–1690.

Seungwhan Moon, Satwik Kottur, Paul Crook, Ankita De, Shivani Poddar, Theodore Levin, David Whitney, Daniel Difranco, Ahmad Beirami, Eunjoon Cho, Rajen Subba, and Alborz Geramifard. 2020. Situated and interactive multimodal conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 1103–1121, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Nikola Mrkšić, Diarmuid Ó Séaghdha, Tsung-Hsien Wen, Blaise Thomson, and Steve Young. 2017. Neural belief tracker: Data-driven dialogue state tracking. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1777–1788, Vancouver, Canada. Association for Computational Linguistics.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649.

Jiaxin Qi, Yulei Niu, Jianqiang Huang, and Hanwang Zhang. 2020. Two causal principles for improving visual dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10860–10869.

Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. 2015. A dataset for movie description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3202–3212.

Idan Schwartz, Seunghak Yu, Tamir Hazan, and Alexander G Schwing. 2019. Factor graph attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2039–2048.

Paul Hongsuck Seo, Andreas Lehrmann, Bohyung Han, and Leonid Sigal. 2017. Visual reference resolution using attention memory for visual dialog. In *Advances in neural information processing systems*, pages 3719–3729.

Xiaomeng Song, Yucheng Shi, Xin Chen, and Yahong Han. 2018. Explore multi-step reasoning in video question answering. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 239–247.

Jesse Thomason, Michael Murray, Maya Cakmak, and Luke Zettlemoyer. 2019. Vision-and-dialog navigation. In *Conference on Robot Learning (CoRL)*.

Erik Wijmans, Samyak Datta, Oleksandr Maksymets, Abhishek Das, Georgia Gkioxari, Stefan Lee, Irfan Essa, Devi Parikh, and Dhruv Batra. 2019. Embodied Question Answering in Photorealistic Environments with Point Cloud Perception. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kexin Yi*, Chuang Gan*, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. 2020. Clevrer: Collision events for video representation and reasoning. In *International Conference on Learning Representations*.

Amir Zadeh, Michael Chan, Paul Pu Liang, Edmund Tong, and Louis-Philippe Morency. 2019. Social-iq: A question answering benchmark for artificial social intelligence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8807–8817.

Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. 2016. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4995–5004.

A A Comparison of DVD to Related Benchmarks

In Table 3, we compare DVD with related benchmarks by 4 aspects: spatial reasoning (SR), temporal reasoning (TR), dialogue object tracking (DOT), and video interval tracking (VIT). SR and TR are visual-related reasoning types. SR refers to the reasoning requirement to localize information within an image. SR is the most popular reasoning type, being involved in most vision-language benchmarks such as VQA (Antol et al., 2015) and TGIF-QA (Jang et al., 2017). TR is often present when a video is used as input, which requires systems to localize the relevant temporal location in the video. However, TR is not just limited to video understanding tasks but also refers to problems with dynamic visual inputs such as navigation systems or embodied QA. DOT and VIT refer to cross-turn semantic relations in a multi-turn dialogue problem setting. DOT refers to the use of object references, requiring systems to learn to resolve these references in dialogue context. DOT can be seen clearly in most dialogue benchmarks as object references are used frequently in traditional dialogues. VIT is a new reasoning requirement in video-grounded dialogue tasks. It requires systems to localize temporal parts of the video from turn to turn. VIT is less obvious in prior benchmarks as it is challenging to simulate. It is mostly present in specific tasks such as AVSD (Hori et al., 2019) and CVDN (Thomason et al., 2019) where a video input is introduced and at each turn, only a specific temporal part of the video is relevant. Compared to existing benchmarks, DVD is the first diagnostic benchmark that combines all 4 aspects, SR, TR, DOT, and VIT, together.

B DVD Functional Program Modules

In Table 4 and 5, we describe all data types and functional program modules in DVD. In total, there are 20 data types and 32 functional modules. Among the functional modules, compared to CLEVR (Johnson et al., 2017), we introduced 17 novel modules that are designed to be executed on dialogue or video components. Within these modules, there are 13 video-based modules (*Count Action, Filter Action, Same Action Set, Same Action Sequence, Find Interval, Union Interval, Relate Spatial, Relate Temporal, Query Action Set, Query Action Sequence, Action by Frequency, Action by Order, Equal Action*) and 4 dialogue-based modules (*Refer Object, Track Object, Refer Interval, Track Interval*).

C DVD Question Types, Sub-types, and Examples

In Table 6, we detail all 8 question types for DVD. In each question type, we described the types of video intervals applicable, including *Atomic* interval, *Compositional* interval, or *None*. *None* type is used in temporal agnostic questions, such as questions to query object attributes or count objects. In each question type, we further classify questions by question sub-types. Figure 10 presents the distribution of questions by question sub-types. We observed that per each question type, question sub-types are balanced in most cases. For instance, the question type *Compare Action Frequency* include 3 sub-types: *equal*, *less*, and *more*, and each is about 4% of the total questions. Similar observations can be seen in other question types, including *Compare Action Sequence*, *Compare Action Set*, and *Attribute Query*.

Benchmarks	Diagnostic benchmark	Visual reasoning		Language reasoning	
		SR	TR	DOT	VIT
Image/video QA, embodied QA					
VQA (Antol et al., 2015), Visual7W (Zhu et al., 2016)	✗	✓	✗	✗	✗
TGIF-QA (Jang et al., 2017), TV-QA (Lei et al., 2018)	✗	✓	✓	✗	✗
IQA (Gordon et al., 2018), EQA (Wijmans et al., 2019)	✗	✓	✓	✗	✗
Image/video grounded dialogues, navigation dialogues					
VisDial (Das et al., 2017), GuessWhat (De Vries et al., 2017)	✗	✓	✗	✓	✗
AVSD (Hori et al., 2019), CVDN (Thomason et al., 2019)	✗	✓	✓	✓	✓
Synthetic image/video QA					
SHAPE (Andreas et al., 2016), CLEVR (Johnson et al., 2017)	✓	✓	✗	✗	✗
SVQA (Song et al., 2018), CLEVRER (Yi* et al., 2020)	✓	✓	✓	✗	✗
Synthetic dialogues					
bAbI (Bordes et al., 2017)	✓	✗	✗	✓	✗
MNIST Dialog (Seo et al., 2017), CLEVR-Dialog (Kottur et al., 2019)	✓	✓	✗	✓	✗
DVD (Ours)	✓	✓	✓	✓	✓

Table 3: **Comparison to related benchmarks:** Compared to existing datasets for vision-language understanding, DVD is the first diagnostic benchmark designed for both spatial reasoning (SR) and temporal reasoning (TR) and explicit requiring dialogue object tracking (DOT) and video interval tracking (VIT) in a multi-turn setting.

Data type	Description
Object	A dictionary storing the attributes of an object, including its shape, size, color, and material, and details of its actions, including start and end points
Objects	A list of of *Objects*
Spatial Relation	A value from the set: "left", "right", "front", and "behind"
Temporal Relation	A value from the set: "before", "after", and "during"
Reference	Pronoun, such as "it", "its", "them", "the first one", used to refer to an object or action mentioned in the last dialogue turn
Last Turn	The last dialogue turn, including the last question and answer
Object Tracker	A list of *objects*, storing all objects involved and their attributes mentioned so far up to the last dialogue turn
Interval Tracker	A list of video intervals mentioned so far up to the last dialogue turn
Interval	A tuple containing the start and end time of a video segment
Action	Any value from "sliding", "flying", "rotating", and "no action"
Action Set	Any combination of *actions*, except for "no ation", without duplication. A standalone "no action" is acceptable.
Action Sequence	Any combination of *actions*, except for "no action", that can form a sequence. A standalone "no action" is acceptable.
Frequency	A positive integer that indicates the number of times an action is undertaken. Frequency can also be expressed by superlatives such as "least" or "most".
Order	An ordinal number that indicates the order of an action during a video interval.
Color	A string that indicates an object color: "gold", "gray", "green", "purple", "red", "cyan", "cylinder", "blue", "brown", "yellow"
Material	A string that indicates an object material, including "metal" and "rubber"
Shape	A string that indicates an object shape, including "cone", "cube", "sphere", "snitch"
Size	A string that indicates an object size, including "large", "medium", and "small"
Binary	A binary value, either "False" or "True"
Integer	An integer value $>= 0$

Table 4: **Data types in DVD:** In total, there are 20 data types, which can be categorized by the following groups (from Top to Bottom): object-based, relation-based, cross-turn based, action-based, attributes, and binary/integer.

Module Type	Module Name	Input Type	Output Type	Module Description
Count	Count Object	*Objects*	*Integer*	Return number of objects
	Count Action	*(Interval, Object, Action)*	*Integer*	Return number of times an object undertakes a specific type of actions during an interval
Exist	Exist	*Objects*	*Binary*	Return whether there is at least one resulting object from the last module
Object-based	Filter Color	*(Objects, Color)*	*Objects*	Return objects of a specific color
	Filter Material	*(Objects, Material)*	*Objects*	Return objects of a specific material
	Filter Shape	*(Objects, Shape)*	*Objects*	Return objects of a specific shape
	Filter Size	*(Objects, Size)*	*Objects*	Return objects of a specific size
	Filter Action	*(Interval, Objects, Action)*	*Objects*	Return objects performing a specific action during an interval
	Same Action Set	*(Interval, Object)*	*Objects*	Return objects performing the same action set as another object during an interval
	Same Action Sequence	*(Interval, Object)*	*Objects*	Return objects performing the same action sequence as another object during an interval
	Unique	*Objects*	*Object*	Return the unique object from resulting objects
	Scene		*Objects*	Return all objects in the current video
Interval-based	Find Interval	*(Object, Action)*	*Interval*	Return the start and end point of the interval of an action performed by an object
	Union Interval	*(Interval$_1$, Interval$_2$)*	*Interval*	Return the overlapping interval from Interval$_1$ and Interval$_2$
Relate	Relate Spatial	*(Interval, Object, Spatial Relation)*	Objects	Return objects located in relation to another object during a specific interval
	Relate Temporal	*(Interval, Temporal Relation)*	*Interval*	Return interval in relation to another interval
Integer-based	Greater Than	*(Integer$_1$, Integer$_2$)*	*Binary*	Return whether Integer$_1$ > Integer$_2$
	Less Than	*(Integer$_1$, Integer$_2$)*	*Binary*	Return whether Integer$_1$ < Integer$_2$
	Equal	*(Integer$_1$, Integer$_2$)*	*Binary*	Return whether Integer$_1$ = Integer$_2$
Multi-turn	Refer Object	*(Reference, Last Turn)*	*Objects*	Resolve object reference based on the last dialogue turn
	Track Object	*Object Tracker*	*Objects*	Return all objects mentioned so far in the dialogue
	Refer Interval	*(Reference, Last Turn)*	*Interval*	Resolve interval reference to an action mentioned in the last dialogue turn
	Track Interval	*Interval Tracker*	*Interval*	Return the interval used in the last dialogue turn
Action-based	Query Action Set	*(Interval, Object)*	*Action Set*	Return the set of actions performed by an object during an interval
	Query Action Sequence	*(Interval, Object)*	*Action Sequence*	Return the sequence of actions performed by an object during an interval
	Action by Frequency	*(Interval, Object, Frequency)*	*Action Set*	Return the set of action performed by an object for a fixed number of times during an interval
	Action by Order	*(Interval, Object, Order)*	*Action*	Return an specific action performed by an object during an interval in an ordinal position (e.g. 1st, 2nd)
	Equal Action	*(Action Set/Sequence, Action Set/Sequence)*	*Binary*	Return whether two set of actions are the same or two sequences of actions are the same
Other Modules	Query Color	*Object*	*Color*	Obtain the color of a specific object
	Query Material	*Object*	*Material*	Obtain the material of a specific object
	Query Shape	*Object*	*Shape*	Obtain the shape of a specific object
	Query Size	*Object*	*Size*	Obtain the size of a specific object

Table 5: **Details of functional program modules:** In total, there are 32 functional program modules, of which 17 are modules introduced for video-based and dialogue-based components.

Question Type	Question Interval	Question Subtype	Example
Compare action frequency	Compositional	more	until the end of the snitch 's rotation , does the blue thing fly more frequently than the purple object flies ?
		equal	during the whole video , does the sphere rotate as frequently as the cylinder slides ?
		less	after the large thing 's first flight , does the cylinder fly less frequently than the green object slides ?
Compare action sequence	Compositional	count	before the large matte thing 's flight , how many other things perform the same sequence of activities as the cyan object ?
		exist	until the end of the metal sphere 's slide , is there any other thing with the same sequence of activities performed by the average purple thing ?
Compare action set	Compositional	count	throughout the whole video , how many other things undertake the same types of actions as the large block ?
		exist	until the end of the cyan shiny thing 's last slide , is there any other object that has the same types of actions as the large rubber object ?
Action query	Compositional	by frequency	during the gray thing 's flight , what activities that the big thing perform the least ?
		by order	until the end of the average green thing 's second flight , what is the purple thing doing first ?
		all actions	during the whole video , what is the brown thing doing ?
	Atomic (Non-Spatial)	all actions	after the red cube 's second slide , what actions does the green sphere undertake ?
	Atomic (Spatial)	all actions	during the small thing 's second rotation , what actions does the average rubber thing that is in front of the red thing undertake ?
Attribute query	None	size	how big is the cylinder ?
		color	what color is the cylinder ?
		material	what material is the brown cone ?
		shape	what is the shape of the cyan object ?
Action count	Compositional	-	throughout the whole video , how many times does the metal cylinder spin in total ?
Object count	Compositional	-	throughout the whole video , what number of sliding matte cones are there ?
	Atomic (Spatial)	-	during the cylinder 's first rotation , what number of objects are in front of the purple thing ?
	Atomic (Non-spatial)	-	before the cylinder 's slide , how many stationary metallic objects are there ?
	None	-	what number of purple things are there ?
Object exist	Compositional	-	throughout the whole video , is there any sliding large rubber cone ?
	Atomic (Spatial)	-	during the brown thing 's rotation , is there any cone in front of the purple cube ?
	Atomic (Non-spatial)	-	since the start of the big red thing 's flight , is there a contained small red metal cylinder ?

Table 6: **Question types and examples:** In total, there are 8 question types, each of which is designed for one or more types of video intervals (Atomic, Compositional, or None). In each question type, we also classify further into question sub-types.
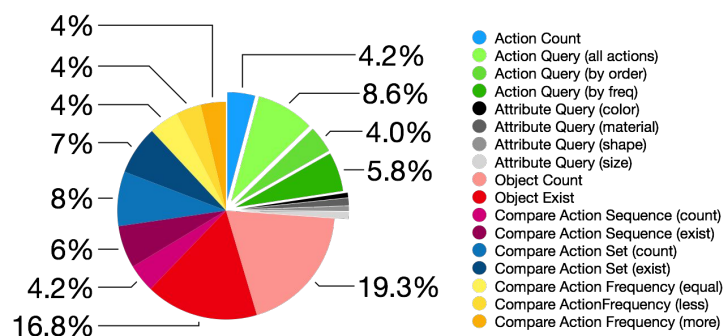

Figure 10: **Distribution of questions by question sub-types:** For each question type, we classify questions further into corresponding sub-types. In total, from 8 question types, there are 17 question sub-types.