Multimodal Dialogue State Tracking

Hung Le[†‡§], **Nancy F. Chen**[§], **Steven C.H. Hoi**[†‡]
[†]Salesforce Research Asia
[‡]Singapore Management University
[§]Agency for Science, Technology and Research (A*STAR)
{hungle,shoi}@salesforce.com, nfychen@i2r.a-star.edu.sg

Abstract

Designed for tracking user goals in dialogues, a dialogue state tracker is an essential component in a dialogue system. However, the research of dialogue state tracking has largely been limited to unimodality, in which slots and slot values are limited by knowledge domains (e.g. restaurant domain with slots of restaurant name and price range) and are defined by specific database schema. In this paper, we propose to extend the definition of dialogue state tracking to multimodality. Specifically, we introduce a novel dialogue state tracking task to track the information of visual objects that are mentioned in video-grounded dialogues. Each new dialogue utterance may introduce a new video segment, new visual objects, or new object attributes and a state tracker is required to update these information slots accordingly. We created a new synthetic benchmark and designed a novel baseline, Video-Dialogue Transformer Network (VDTN), for this task. VDTN combines both object-level features and segment-level features and learns contextual dependencies between videos and dialogues to generate multimodal dialogue states. We optimized VDTN for a state generation task as well as a self-supervised video understanding task which recovers video segment or object representations. Finally, we trained VDTN to use the decoded states in a response prediction task. Together with comprehensive ablation and qualitative analysis, we discovered interesting insights towards building more capable multimodal dialogue systems.

1 Introduction

The main goal of dialogue research is to develop intelligent agents that can assist humans through conversations. For example, a dialogue agent can be tasked to help users to find a restaurant based on their preferences of price ranges and food choices. A crucial part of a dialogue system is Dialogue State Tracking (DST), which is responsible for tracking and updating user goals in the form of dialogue states, including a set of *(slot, value)* pairs such as *(price, "moderate")* and *(food, "japanese")*. Numerous machine learning approaches have been proposed to tackle DST, including fixed-vocabulary models (Ramadan et al., 2018; Lee et al., 2019) and open-vocabulary models (Lei et al., 2018b; Wu et al., 2019; Le et al., 2020c), for either single-domain (Wen et al., 2017) or multi-domain dialogues (Eric et al., 2017; Budzianowski et al., 2018).

However, the research of DST has largely limited the scope of dialogue agents to unimodality. In this setting, the slots and slot values are defined by the knowledge domains (e.g. restaurant domain) and database schema (e.g. data tables for restaurant entities). The ultimate goal of dialogue research towards building artificial intelligent assistants necessitates DST going beyond unimodal systems. In this paper, we propose Multimodal Dialogue State Tracking (MM-DST) that extends the DST task in a multimodal world. Specifically, MM-DST extends the scope of dialogue states by defining slots and slot values for *visual objects* that are mentioned in visually-grounded dialogues. For research purposes, following (Alamri et al., 2019), we limited visually-grounded dialogues as ones with a grounding video input and the dialogues contain multiple turns of (question, answer) pairs about this video. Each new utterance in such dialogues may focus on a new video segment, new visual objects, or new object attributes, and the tracker is required to update the dialogue state accordingly at each turn. An example of MM-DST can be seen in Figure 1.

Toward MM-DST, we developed a synthetic benchmark based on the CATER universe (Girdhar and Ramanan, 2020). We also introduced Video-Dialogue Transformer Network (VDTN), a neural network architecture that combines both object-level features and segment-level features in video and learns contextual dependencies between videos

Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 3394 - 3415
July 10-15, 2022 ©2022 Association for Computational Linguistics

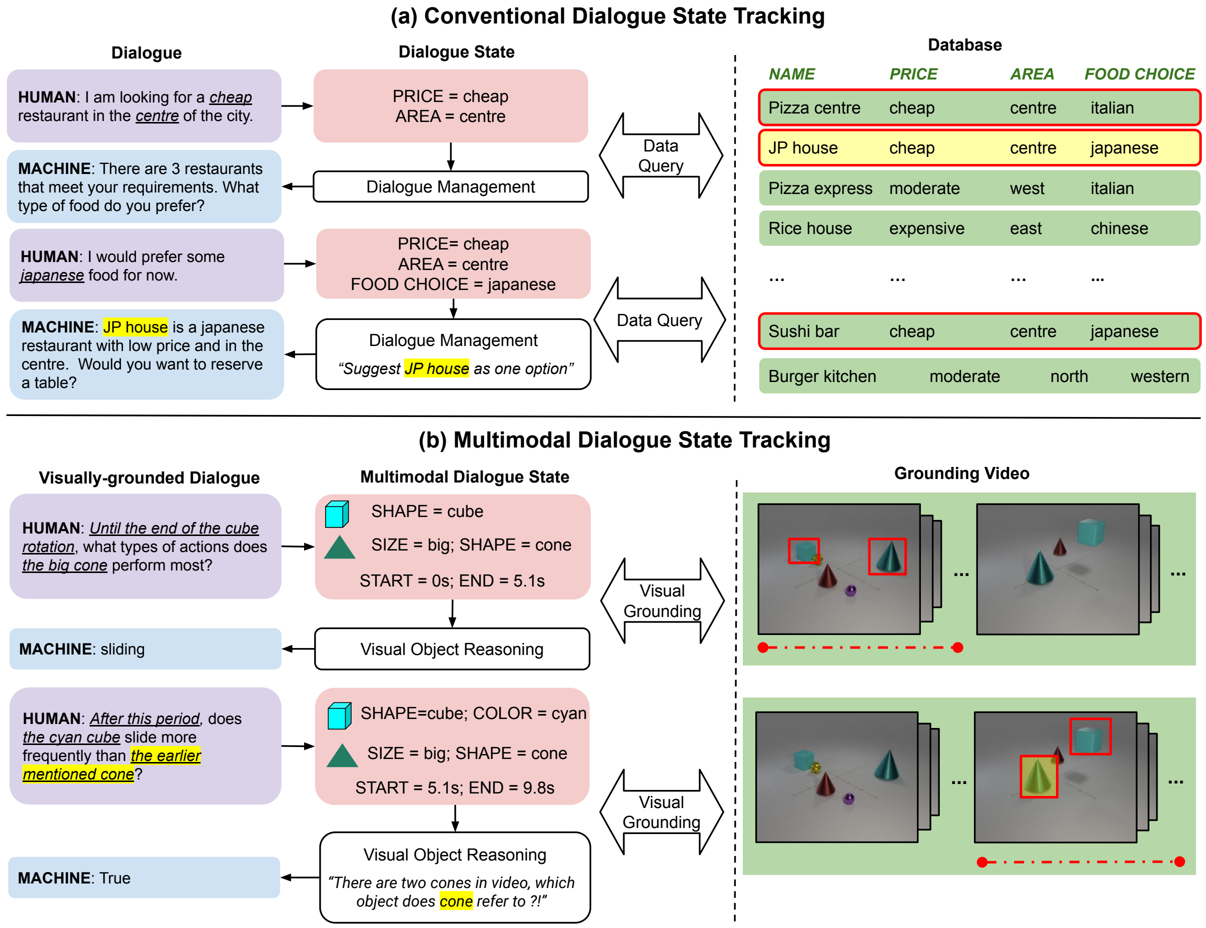

(a) Conventional Dialogue State Tracking
Dialogue
Dialogue State
Database
HUMAN: I am looking for a cheap restaurant in the centre of the city.
PRICE = cheap
AREA = centre
MACHINE: There are 3 restaurants that meet your requirements. What type of food do you prefer?
Dialogue Management
HUMAN: I would prefer some japanese food for now.
PRICE= cheap
AREA = centre
FOOD CHOICE = japanese
MACHINE: JP house is a japanese restaurant with low price and in the centre. Would you want to reserve a table?
Dialogue Management
"Suggest JP house as one option"
Data Query
Data Query
NAME PRICE AREA FOOD CHOICE
Pizza centre cheap centre italian
JP house cheap centre japanese
Pizza express moderate west italian
Rice house expensive east chinese
... ...
Sushi bar cheap centre japanese
Burger kitchen moderate north western
(b) Multimodal Dialogue State Tracking
Visually-grounded Dialogue
Multimodal Dialogue State
Grounding Video
HUMAN: Until the end of the cube rotation, what types of actions does the big cone perform most?
SHAPE = cube
SIZE = big; SHAPE = cone
START = 0s; END = 5.1s
MACHINE: sliding
Visual Object Reasoning
HUMAN: After this period, does the cyan cube slide more frequently than the earlier mentioned cone?
SHAPE=cube; COLOR = cyan
SIZE = big; SHAPE = cone
START = 5.1s; END = 9.8s
MACHINE: True
Visual Object Reasoning
"There are two cones in video, which object does cone refer to ?!"
Visual Grounding
Visual Grounding
...

Figure 1: **Multimodal Dialogue State Tracking (MM-DST):** We proposed to extend the traditional DST from unimodality to multimodality. Compared to traditional DST (a), MM-DST (b) define dialogue states, consisting of slots and slots values for visual objects that are mentioned in dialogues.

and dialogues. Specifically, we maintained the information granularity of visual objects, embedded by object classes and their bounding boxes and injected with segment-level visual context. VDTN enables interactions between each visual object representation and word-level representation in dialogues to decode dialogue states. To decode multimodal dialogue states, we adopted a decoding strategy inspired by the Markov decision process in traditional DST (Young et al., 2010). In this strategy, a model learns to decode the state at a dialogue turn based on the predicted/ observed dialogue state available from the last dialogue turn.

Compared to the conventional DST, MM-DST involves the new modality from visual inputs. Our experiments show that simply combining visual and language representations in traditional DST models results in poor performance. Towards this challenge, we enhanced VDTN with self-supervised video understanding tasks which recovers object-based or segment-based representations. Benchmarked against strong unimodal DST models, we observed significant performance gains from VDTN. We provided comprehensive ablation analysis to study the efficacy of VDTN models. Interestingly, we also showed that using decoded states brought performance gains in a dialogue response prediction task, supporting our motivation for introducing multimodality into DST research.

2 Multimodal Dialogue State Tracking

Traditional DST. As defined by (Mrkšić et al., 2017), the traditional DST includes an input of dialogue $\mathcal{D}$ and a set of slots $\mathcal{S}$ to be tracked from turn to turn. At each dialogue turn t, we denote the dialogue context as $\mathcal{D}_t$, containing all utterances up to the current turn. The objective of DST is for each turn t, predict a value v_i^t of each slot s_i from a predefined set $\mathcal{S}$, conditioned by the dialogue context $\mathcal{D}_t$. We denote the dialogue state at turn t as $\mathcal{B}_t = \{(s_i, v_i^t)\}|_{i=1}^{i=|\mathcal{S}|}$. Note that a majority of traditional DST models assume slots are conditionally independent, given the dialogue context (Zhong et al., 2018; Budzianowski et al., 2018; Wu et al., 2019; Lee et al., 2019; Gao et al., 2019). The learning objective is defined as:

$$\begin{aligned}\hat{\mathcal{B}}_t &= \arg\max_{\mathcal{B}_t} P(\mathcal{B}_t|\mathcal{D}_t,\theta) \\ &= \arg\max_{\mathcal{B}_t} \prod_{i}^{|\mathcal{S}|} P(v_i^t|s_i,\mathcal{D}_t,\theta) \end{aligned} \tag{1}$$

Motivation to Multimodality. Yet, the above definition of DST are still limited to unimodality and our ultimate goal of building intelligent dialogue agents, ideally with similar level of intelligence as humans, inspires us to explore mulitmodality. In neuroscience literature, several studies have analyzed how humans can perceive the world in visual context. (Bar, 2004; Xu and Chun, 2009) found that humans can recognize multiple visual objects and how their contexts, often embedded with other related objects, facilitate this capacity.

Our work is more related to the recent study (Fischer et al., 2020) which focuses on human capacity to create temporal stability across multiple objects. The multimodal DST task is designed to develop multimodal dialogue systems that are capable of maintaining discriminative representations of visual objects over a period of time, segmented by dialogue turns. While computer science literature has focused on related human capacities in intelligent systems, they are mostly limited to vision-only tasks e.g. (He et al., 2016; Ren et al., 2015) or QA tasks e.g. (Antol et al., 2015; Jang et al., 2017) but not in a dialogue task.

Most related work in the dialogue domain is (Pang and Wang, 2020) and almost concurrent to our work is (Kottur et al., 2021). However, (Kottur et al., 2021) is limited to a single object per dialogue, and (Pang and Wang, 2020) extends to multiple objects but does not require to maintain an information state with component slots for each object. Our work aims to complement these directions and address their limitations with a novel definition of multimodal dialogue state.

Multimodal DST (MM-DST). To this end, we proposed to extend conventional dialogue states. First, we use visual object identities themselves as a component of the dialogue state to enable the perception of multiple objects. A dialogue state might have one or more objects and a dialogue system needs to update the object set as the dialogue carries on. Secondly, for each object, we define slots that represent the information state of objects in dialogues (as denoted by (Fischer et al., 2020) as "content" features of objects memorized by humans). The value of each slot is subject-specific and updated based on the dialogue context of the corresponding object. This definition of DST is closely based on the above well-studied human capacities while complementing the conventional dialogue research (Young et al., 2010; Mrkšić et al., 2017), and more lately multimodal dialogue research (Pang and Wang, 2020; Kottur et al., 2021).

We denote a grounding visual input in the form of a video $\mathcal{V}$ with one or more visual objects o_j. We assume these objects are semantically different enough (by appearance, by characters, etc.) such that each object can be uniquely identified (e.g. by an object detection module ω). The objective of MM-DST is for each dialogue turn t, predict a value v_i^t of each slot $s_i \in \mathcal{S}$ for each object $o_j \in \mathcal{O}$. We denote the dialogue state at turn t as $\mathcal{B}_t = |\{(o_j, s_i, v_{i,j}^t)\}|_{i=1,j=1}^{i=|\mathcal{S}|,j=|\mathcal{O}|}$. Assuming all slots are conditionally independent given dialogue and video context, the learning objective is extended from Eq. (1):

$$\begin{aligned}\hat{\mathcal{B}}_t &= \arg\max_{\mathcal{B}_t} P(\mathcal{B}_t|\mathcal{D}_t,\mathcal{V},\theta) \\ &= \arg\max_{\mathcal{B}_t} \prod_{j}^{|\mathcal{O}|}\prod_{i}^{|\mathcal{S}|} P(v_{i,j}^t|s_i,o_j,\mathcal{D}_t,\mathcal{V},\theta)P(o_j|\mathcal{V},\omega)\end{aligned}$$

One limitation of the current representation is the absence of temporal placement of objects in time. Naturally humans are able to associate objects and their temporal occurrence over a certain period. Therefore, we defined two temporal-based slots: s_{start} and s_{end}, denoting the start time and end time of the video segment that an object can be located by each dialogue turn. In this work, we assume that a dialogue turn is limited to a single continuous time span, and hence, s_{start} and s_{end} can be defined turn-wise, identically for all objects. While this is a strong assumption, we believe it covers a large portion of natural conversational interactions. An example of multimodal dialogue state can be seen in Figure 1.

3 Video-Dialogue Transformer Network

A naive adaptation of conventional DST to MM-DST is to directly combine visual features extracted by a pretrained 3D-CNN model. However, as

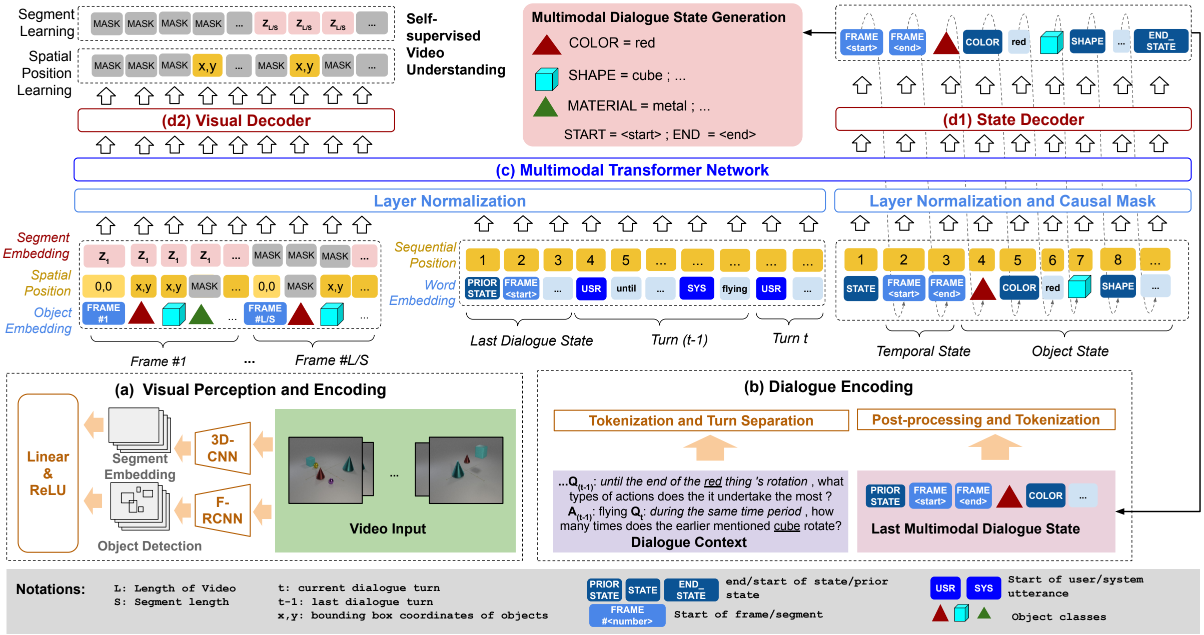

Segment Learning
MASK MASK MASK MASK ... $z_{L/S}$ $z_{L/S}$ $z_{L/S}$ ...
Spatial Position Learning
MASK MASK MASK x,y ... MASK x,y MASK ...
Self-supervised Video Understanding
(d2) Visual Decoder
Multimodal Dialogue State Generation
COLOR = red
SHAPE = cube ; ...
MATERIAL = metal ; ...
START = <start> ; END = <end>
FRAME <start> FRAME <end> COLOR red SHAPE ... END_ STATE
(d1) State Decoder
(c) Multimodal Transformer Network
Layer Normalization
Layer Normalization and Causal Mask
Segment Embedding
Spatial Position
Object Embedding
z_1 z_1 z_1 z_1 ... MASK MASK MASK ...
0,0 x,y x,y MASK ... 0,0 MASK x,y ...
FRAME #1 ... FRAME #L/S ...
Frame #1 ... Frame #L/S
Sequential Position
Word Embedding
1 2 3 4 5 ... ...
PRIOR STATE FRAME <start> ... USR until ... SYS flying USR ...
Last Dialogue State
Turn (t-1)
Turn t
1 2 3 4 5 6 7 8 ...
STATE FRAME <start> FRAME <end> COLOR red SHAPE ...
Temporal State
Object State
(a) Visual Perception and Encoding
Linear & ReLU
Segment Embedding
3D-CNN
Object Detection
F-RCNN
Video Input
(b) Dialogue Encoding
Tokenization and Turn Separation
Post-processing and Tokenization
...$Q_{(t-1)}$: until the end of the red thing 's rotation , what types of actions does the it undertake the most ? $A_{(t-1)}$: flying Q_t: during the same time period , how many times does the earlier mentioned cube rotate?
Dialogue Context
PRIOR STATE FRAME <start> FRAME <end> COLOR ...
Last Multimodal Dialogue State
Notations:
L: Length of Video
S: Segment length
t: current dialogue turn
t-1: last dialogue turn
x,y: bounding box coordinates of objects
PRIOR STATE STATE END_ STATE end/start of state/prior state
FRAME #<number> Start of frame/segment
USR SYS Start of user/system utterance
Object classes

Figure 2: **Video-Dialogue Transformer Network(VDTN)** has 4 key components: (a) *Visual Perception and Encoder* (Section 3.1) (b) *Dialogue Encoder* (Section 3.2) (c) *Transformer Network* (Section 3.3) (d1) *State Decoder* (Section 3.4) and (d2) *Visual Decoder* (Section 3.4)

shown in our experiments, this extension of conventional DST results in poor performance and does not address the challenge of MM-DST. In this paper, we established a new baseline, denoted as Video-Dialogue Transformer Network (VDTN) (Refer to Fig. 2 for an overview):

3.1 Visual Perception and Encoder

Visual Perception. This module encodes videos at both frame-level and segment-level representations. Specifically, we used a pretrained Faster R-CNN model (Ren et al., 2015) to extract object representations. We used this model to output the bounding boxes and object identifiers (object classes) in each video frame of the video. For an object o_j, we denoted the four values of its bounding boxes as $bb_j = (x_j^1, y_j^1, x_j^2, y_j^2)$ and o_j as the object class itself. We standardized the video features by extracting features of up to $N_{obj} = 10$ objects per frame and normalizing all bounding box coordinates by the frame size. Secondly, we used a pretrained ResNeXt model (Xie et al., 2017) to extract the segment-level representations of videos, denoted as $z_m \in \mathbb{R}^{2048}$ for a segment m. Practically, we followed the best practice in computer vision by using a temporal sliding window with strides to sample video segments and passed segments to ResNeXt model to extract features. To standardize visual features, we use the same striding configuration N_{stride} to sub-sample segments for ResNeXt and frames for Faster R-CNN models.

Visual Representation. Note that we do not finetune the visual feature extractors in VDTN and keep the extracted features fixed. To transform these features into VDTN embedding space, we first concatenated all object identity tokens *OBJ<class>* of all frames. An object identity token *OBJ<class>* is the code name of the object class (e.g. a class of small blue metal cones) that a visual object can be unique identified (See Figure 2). Frames are separated by a special token *FRAME<number>*, where *<number>* is the temporal order of the frame. This results in a sequence of tokens X_{obj} of length $L_{obj} = (N_{obj} + 1) \times (|\mathcal{V}|/N_{stride})$ where $|\mathcal{V}|$ is the number of video frames. Correspondingly, we concatenated bounding boxes of all objects, and used a zero vector in positions of *FRAME<number>* tokens. We denoted this sequence as $X_{bb} \in \mathbb{R}^{L_{obj} \times 4}$ where the dimension of 4 is for the bounding box coordinates (x^1, y^1, x^2, y^2). Similarly, we stacked each ResNeXt feature vector by $(N_{obj} + 1)$ for each segment, and obtained a sequence $X_{cnn} \in \mathbb{R}^{L_{obj} \times 2048}$.

Visual Encoding. We passed each of X_{bb} and X_{cnn} to a linear layer with ReLU activation to map their feature dimension to a uniform dimension d.

We used a learnable embedding matrix to embed each object identity in X_{obj}, resulting in embedding features of dimensions d. The final video input representation is the summation of above vectors, denoted as $Z_V = Z_{obj} + Z_{bb} + Z_{cnn} \in \mathbb{R}^{L_{obj} \times d}$.

3.2 Dialogue and State Encoder

Dialogue Encoding. Another encoder encodes dialogue into continuous representations. Given a dialogue context $\mathcal{D}_t$, we tokenized all dialogue utterances into sequences of words, separated by special tokens *USR* for human utterance and *SYS* for system utterance. We used a trainable embedding matrix and sinusoidal positional embeddings to embed this sequence into d-dimensional vectors.

Flattening State into Sequence. Similar to the recent work in traditional DST (Lei et al., 2018b; Le et al., 2020b; Zhang et al., 2020), we are motivated by the DST decoding strategy following a Markov principle and used the dialogue state of the last dialogue turn $\mathcal{B}_{t-1}$ as an input to generate the current state $\mathcal{B}_t$. Using the same notations from Section 2, we can represent B_t into a sequence of o_j, s_i, and $v^t_{i,j}$ tokens, such as "*OBJ4 shape cube OBJ24 size small color red*". This sequence is then concatenated with utterances from $\mathcal{D}_t$, separated by a special token *PRIOR_STATE*. We denoted the resulting sequence as X_{ctx} which is passed to the embedding matrix and positional encoding as described above. As we showed in our experiments, to encode dialogue context, this strategy needs only a few dialogue utterances (that are closer to the current turn t) and $\mathcal{B}_{t-1}$, rather than the full dialogue history from turn 1. Therefore, dialogue representations Z_{ctx} have more compressed dimensions of $|X_{ctx}| \times d$ where $|X_{ctx}| < |\mathcal{D}_t|$.

3.3 Multimodal Transformer Network

We concatenated both video and dialogue representations, denoted as $Z_{VD} = [Z_V; Z_D]$. Z_{VD} has a length of $L_{obj} + L_{ctx}$ and embedding dimension d. We pased Z_{VD} to a vanilla Transformer network (Vaswani et al., 2017) through multiple multi-head attention layers with normalization (Ba et al., 2016) and residual connections (He et al., 2016). Each layer allows multimodal interactions between object-level representations from videos and word-level representations from dialogues.

3.4 Dialogue State Decoder and Self-supervised Video Denoising Decoder

State Decoding. This module decodes dialogue state sequence auto-regressively, i.e. each token is conditioned on all dialogue and video representations as well as all tokens previously decoded. At the first decoding position, a special token *STATE* is embedded into dimension d (by a learned embedding layer and sinusoidal positional encoding) and concatenated to Z_{VD}. The resulting sequence is passed to the Transformer network and the output representations of *STATE* are passed to a linear network layer that transforms representations to state vocabulary embedding space. The decoder applies the same procedure for the subsequent positions to decode dialogue states auto-regressively. Denoting $b_{k,t}$ as the k^{th} token in $\mathcal{B}_t$, i.e. token of slot, object identity, or slot value, we defined the DST loss function as the negative log-likelihood:

$$\mathcal{L}_{dst} = -\sum \log P(b_{k,t}|b_{<k,t}, X_{ctx}, X_{obj})$$

Note that this decoder design partially avoids the assumption of conditionally independent slots. During test time, we applied beam search to decode states with the maximum length of 25 tokens in all models and a beam size 5. An *END_STATE* token is used to mark the end of each sequence.

Visual Denoising Decoding. Finally, moving away from conventional unimodal DST, we proposed to enhance our DST model with a *Visual Decoder* that learns to recover visual representations in a self-supervised learning task to improve video representation learning. Specifically, during training time, we randomly sampled visual representations and masked each of them with a zero vector. At the object level, in the m^{th} video frame, we randomly masked a row from $X_{bb}(m) \in \mathbb{R}^{N_{obj} \times 4}$. Since each row represents an object, we selected a row to mask by a random object index $j \in [1, N_{obj}]$ such that the same object has not been masked in the preceding frame or following frame. We denote the Transformer output representations from video inputs as $Z'_V \in \mathbb{R}^{L_{obj} \times d}$. This vector is passed to a linear mapping f_{bb} to bounding box features $\mathbb{R}^4$. We defined the learning objective as:

$$\mathcal{L}_{obj} = \sum_o \mathbf{1}_{\text{masked}} \times l(f_{bb}(Z'_{V,o}), X_{bb,o})$$

where l is a loss function and $\mathbf{1}_{\text{masked}} = \{0, 1\}$ is a masking indicator. We experimented with both

L1 and L2 loss and reported the results. Similarly, at the segment level, we randomly selected a segment to mask such that the preceding or following segments have not been chosen for masking:

$$\mathcal{L}_{seg} = \sum_{s} \mathbf{1}_{\text{masked}} \times l(f_{cnn}(Z'_{V,s}), X_{cnn,s})$$

4 Experiments

4.1 Experimental Setup

Dataset. In existing benchmarks of multimodal dialogues such as VisDial (Das et al., 2017a) and AVSD (Alamri et al., 2019), we observed that a large number of data samples contain strong distribution bias in dialogue context, in which dialogue agents can simply ignore the whole dialogue and rely on image-only features (Kim et al., 2020). Another observed bias is the annotator bias that makes a causal link between dialogue context and output response actually harmful (Qi et al., 2020) (as annotator's preferences are treated as a confounding factor). The above biases would obviate the need for a DST task.

To address the above biases, Le et al. (2021b) developed a Diagnostic Benchmark for Video-grounded Dialogues ("DVD"), by synthetically creating dialogues that are grounded on videos from CATER videos (Shamsian et al., 2020). The videos contain visually simple yet highly varied objects. The dialogues are synthetically designed with both short-term and long-term object references. These specifications remove the annotation bias in terms of object appearances in visual context and cross-turn dependencies in dialogue context.

Extension from DVD (Le et al., 2021b) . We generated new dialogues following Le et al. (2021b)'s procedures but based on an extended CATER video split (Shamsian et al., 2020) rather than the original CATER video data (Girdhar and Ramanan, 2020). We chose the extended CATER split (Shamsian et al., 2020) as it includes additional annotations of ground-truth bounding box boundaries of visual objects in video frames. This annotation facilitates experiments with Faster-RCNN finetuned on CATER objects and experiments with models of perfect visual perception, i.e. $P(o_j|\mathcal{V}, \omega) \approx 1$. As shown in (Le et al., 2021b), objects can be uniquely referred in utterances based on their appearance by one or more following aspects: "size", "color", "material", and "shape". We directly reuse these and define them as slots in our dialogue states, in addition to 2 temporal slots for s_{start} and s_{end}. We denote the new benchmark as DVD-DST and summarize the dataset in Table 1 (for more detail, please refer to Appendix B).

Split	# Videos	# Dialogues	# Turns
DVD-DST-Train	9300	9295	92950
DVD-DST-Val	3327	3326	33260
DVD-DST-Test	1371	1371	13710
DVD-DST-All	13998	13992	139920

Table 1: Summary of the DVD-DST dataset

Baselines. To benchmark VDTN, we compared the model with following baseline models, including both rule-based models and trainable models:

- *Q-retrieval (tf-idf)*, for each test sample, directly retrieves the training sample with the most similar question utterance and use its state as the predicted state;
- *State prior* selects the most common tuple of (object, slot, value) in training split and uses it as predicted states;
- *Object (random)*, for each test sample, randomly selects one object predicted by the visual perception model and a random (slot, value) tuple (with slots and values inferred from object classes) as the predicted state;
- *Object (all)* is similar to the prior baseline but selects all possible objects and all possible (slot, value) tuples as the predicted state;
- *RNN(+Att)* uses RNN as encoder and an MLP network as decoder. Another variant of the model is enhanced with a vanilla dot-product attention at each decoding step;
- We adapted and experimented with strong unimodal DST baselines, including: *TRADE* (Wu et al., 2019), *UniConv* (Le et al., 2020b) and *NADST* (Le et al., 2020c).

We implemented these baselines and tested them on dialogues with or without videos. When video inputs are applied, we embedded both object and segment-level features (See Section 3.1). The video context features are integrated into baselines in the same techniques in which the original models treat dialogue context features.

Training. We trained VDTN by jointly minimizing $\mathcal{L}_{dst}$ and $\mathcal{L}_{bb/seg}$. We trained all models using the Adam optimizer (Kingma and Ba, 2015) with a warm-up learning rate period of 1 epoch and the learning rate decays up to 160 epochs. Models are selected based on the average $\mathcal{L}_{dst}$ on the validation set. To standardize model sizes, we selected embedding dimension $d = 128$ for all models, and experimented with both shallow ($N = 1$) and deep networks ($N = 3$) (by stacking attention or RNN blocks), and 8 attention heads in Transformer backbones. We implemented models in Pytorch and released the code and model checkpoints [1]. Refer to Appendix C for more training details.

Evaluation. We followed the unimodal DST task (Budzianowski et al., 2018; Henderson et al., 2014a) and used the state accuracy metric. The prediction is counted as correct only when all the component values exactly match the oracle values. In multimodal states, there are both discrete slots (object attributes) as well as continuous slots (temporal start and end time). For continuous slots, we followed (Hu et al., 2016; Gao et al., 2017) by using Intersection-over-Union (IoU) between predicted temporal segment and ground-truth segment. The predicted segment is counted as correct if its IoU with the oracle is more than p, where we chose $p = \{0.5, 0.7\}$. We reported the joint state accuracy of discrete slots only ("Joint Acc") as well as all slot values ("Joint Acc IoU@p"). We also reported the performance of component state predictions, including predictions of object identities o_j, object slot tuples $(o_j, s_i, v_{i,j})$, and object state tuples $(o_j, s_i, v_{i,j}) \forall s_i \in \mathcal{S}$. Since a model may simply output all possible object identities and slot values and achieve 100% component accuracies, we reported the F1 metric for each of these component predictions.

4.2 Results

Overall results. From Table 2, we have the following observations:

- we noted that simply using naive retrieval models such as *Q-retrieval* achieved zero joint state accuracy only. *State prior* achieved only about 15% and 8% F1 on object identities and object slots, showing that a model cannot simply rely on distribution bias of dialogue states.

Model	Dial.	Vid.	Obj Ident. F1	Obj Slot F1	Obj State F1	Joint Acc	Acc IoU @.5	Acc IoU @.7
Q-retrieval	✓	-	6.7	3.3	2.7	1.0	0.8	0.7
State prior	-	-	14.9	7.7	0.1	0.0	0.0	0.0
Object (rand.)	-	✓	19.8	14.1	0.4	0.0	0.0	0.0
Object (all)	-	✓	60.5	27.2	1.5	0.0	0.0	0.0
RNN(V)	-	✓	21.2	10.8	8.3	1.0	0.1	0.1
RNN(D)	✓	-	57.8	43.3	38.0	4.8	1.1	0.6
RNN(V+D)	✓	✓	63.2	48.5	42.8	6.8	2.6	2.3
RNN(V+D)+Att	✓	✓	73.4	59.0	46.8	8.5	3.3	2.0
TRADE (N=1)	✓	-	75.3	63.2	47.8	8.7	2.2	1.1
TRADE (N=1)	✓	✓	75.8	63.8	48.0	9.2	3.3	2.5
TRADE (N=3)	✓	-	74.2	62.6	47.2	8.3	2.1	1.1
TRADE (N=3)	✓	✓	76.1	64.5	48.2	8.9	3.2	2.4
UniConv (N=1)	✓	-	70.6	58.0	44.7	11.1	4.5	3.2
UniConv (N=1)	✓	✓	73.6	60.5	46.2	11.6	6.1	5.4
UniConv (N=3)	✓	-	76.4	62.7	52.5	15.0	6.4	4.6
UniConv (N=3)	✓	✓	76.4	62.7	50.5	14.5	7.8	7.0
NADST (N=1)	✓	-	78.0	63.8	44.9	11.6	4.6	3.2
NADST (N=1)	✓	✓	78.4	64.0	47.7	12.7	6.1	5.5
NADST (N=3)	✓	-	80.6	67.3	50.2	15.3	6.3	4.3
NADST (N=3)	✓	✓	79.0	65.1	49.2	13.3	6.3	5.5
VDTN	✓	✓	**84.5**	**72.8**	**60.4**	**28.0**	**15.3**	**13.1**
VDTN+GPT2	✓	✓	**85.2**	**76.4**	**63.7**	**30.4**	**16.8**	**14.3**

Table 2: Performance (in %) of VDTN vs. baselines on the test split of DVD-DST. ✓on "Dial" or "Vid" column indicates whether we use dialogue context or video context respectively.

- The results of *Object (random/all)* show that in DVD-DST, dialogues often focus on a subset of visual objects and an object perception model alone cannot perform well.

- The performance gains of *RNN* models show the benefits of neural network models compared to retrieval models. The higher results of *RNN(D)* against *RNN(V)* showed the dialogue context is essential and reinforced the above observation.

- Comparing TRADE and UniConv, we noted that TRADE performed slightly better in component predictions, but was outperformed in joint state prediction metrics. This showed the benefits of UniConv which avoids the assumptions of conditionally independent slots and learns to extract the dependencies between slot values.

- Results of TRADE, UniConv, and NADST all displayed minor improvement when adding video inputs to dialogue inputs, displaying their weakness when exposed to cross-modality learning.

- VDTN achieves significant performance gains and achieves the SOTA results in all component or joint prediction metrics.

We also experimented with a version of VDTN in which the transformer network (Section 3.3) was

[1] https://github.com/henryhungle/mm_dst

Video self-supervision	Loss	Acc	IoU @0.5	IoU @0.7
None	N/A	24.8	13.8	11.8
$\mathcal{L}_{obj}$	L1	26.0	14.4	12.4
$\mathcal{L}_{obj}$	L2	24.1	13.3	11.4
$\mathcal{L}_{obj}$ (tracking)	L1	27.2	14.7	12.6
$\mathcal{L}_{obj}$ (tracking)	L2	22.9	12.7	10.9
$\mathcal{L}_{seg}$	L1	**28.0**	**15.3**	**13.1**
$\mathcal{L}_{seg}$	L2	27.4	14.7	12.7
$\mathcal{L}_{obj} + \mathcal{L}_{seg}$	L1	23.7	13.0	11.2
$\mathcal{L}_{obj} + \mathcal{L}_{seg}$	L2	24.3	13.4	11.6

Table 3: Accuracy (in %) by self-supervised objectives. $\mathcal{L}_{obj}$ (tracking) assumes access to oracle bounding box labels and treats the self-supervised learning task as an object tracking task.

Video Features	Dialogue State	Video loss	Acc	IoU @0.5	IoU @0.7
X_{bb}	$\mathcal{B}\backslash$time	-	17.9	N/A	N/A
$X_{bb} + X_{cnn}$	$\mathcal{B}\backslash$time	-	22.4	N/A	N/A
X_{bb}	$\mathcal{B}$	-	19.3	11.0	9.5
$X_{bb} + X_{cnn}$	$\mathcal{B}$	-	24.8	13.8	11.8
X_{bb}	$\mathcal{B}$	$\mathcal{L}_{obj}$	24.0	12.9	11.0
$X_{bb} + X_{cnn}$	$\mathcal{B}$	$\mathcal{L}_{obj}$	26.0	14.4	12.4
$X_{bb} + X_{cnn}$	$\mathcal{B}$	$\mathcal{L}_{seg}$	**28.0**	**15.3**	**13.1**

Table 4: Accuracy (in %) by video features and state formulations

initialized from a GPT2-base model (Radford et al., 2019) with a pretrained checkpoint released by HuggingFace[2]. Aside from using BPE to encode text sequences to match GPT2 embedding indices, we keep other components of the model the same. VDTN+GPT2 is about 36× bigger than our default VDTN model. As shown in Table 2, the performance gains of VDTN+GPT2 indicates the benefits of large-scale language models (LMs). Another benefit of using pretrained GPT2 is faster training time as we observed the VDTN+GPT2 converged much earlier than training it from scratch. From these observations, we are excited to see more future extension of SOTA unimodal DST models (Lin et al., 2021; Dai et al., 2021) and large pretrained LMs (Brown et al., 2020; Raffel et al., 2020), especially ones with multimodal learning such as (Lu et al., 2019; Zhou et al., 2020), to MM-DST task.

Impacts of self-supervised video representation learning. From Table 3, we noted that compared to a model trained only with the DST objective $\mathcal{L}_{dst}$, models enhanced with self-supervised video understanding objectives can improve the results. However, we observe that L1 loss works more consistently than L2 loss in most cases. Since L2 loss minimizes the squared differences between predicted and ground-truth values, it may be susceptible to outliers (of segment features or bounding boxes) in the dataset. Since we could not control these outliers, an L1 loss is more suitable.

We also tested with $\mathcal{L}_{obj}$ (tracking), in which we used oracle bounding box labels during training, and simply passed the features of all objects to VDTN. This modification treats the self-supervised learning task as an object tracking task in which all output representations are used to predict the ground-truth bounding box coordinates of all objects. Interestingly, we found $\mathcal{L}_{obj}$ (tracking) only improves the results insignificantly, as compared to the self-supervised learning objective $\mathcal{L}_{obj}$. This indicates that our self-supervised learning tasks do not strongly depend on the availability of object boundary labels.

Finally, we found combining both segment-level and object-level self-supervision is not useful. This is possibly due to our current masking strategy that masks object and segment features independently. Therefore, the resulting context features might not be sufficient for recovering masked representations. Future work can be extended by studying a codependent masking technique to combine segment-based and object-based representation learning.

Impacts of video features and time-based slots. Table 4 shows the results of different variants of VDTN models. We observed that:

- Segment-based learning is marginally more powerful than object-based learning.
- By considering the temporal placement of objects and defining time-based slots, we noted the performance gains by "Joint Obj State Acc" ($\mathcal{B}$ vs. $\mathcal{B}\backslash time$). The performance gains show the interesting relationships between temporal slots and discrete-only slots and the benefits of modelling both in dialogue states.
- Finally, even with only object-level features X_{bb}, we still observed performance gains from using self-supervised loss $\mathcal{L}_{obj}$, confirming the benefits of better visual representation learning.

[2] https://huggingface.co/gpt2

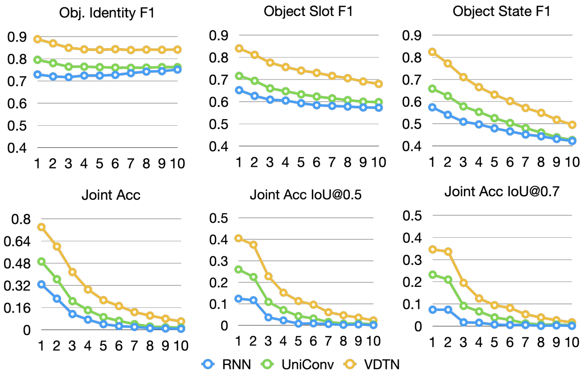


Figure 3: Ablation results of VDTN and baselines by dialogue turn positions (x axis)

Ablation analysis by turn positions. Figure 3 reported the results of VDTN predictions of states that are separated by the corresponding dialogue positions. The results are from the VDTN model trained with both $\mathcal{L}_{dst}$ and $\mathcal{L}_{seg}$. As expected, we observed a downward trend of results as the turn position increases. We noted that state accuracy reduces more dramatically (as shown by "Joint Acc") than the F1 metrics of component predictions. For instance, "Object Identity F1" shows almost stable performance lines through dialogue turns. Interestingly, we noted that the prediction performance of dialogue states with temporal slots only deteriorates dramatically after turn 2 onward. We expected that VDTN is able to learn short-term dependencies (1-turn distance) between temporal slots, but failed to deal with long-term dependencies ($>$ 1-turn distance) between temporal slots. In all metrics, we observed VDTN outperforms both RNN baseline and UniConv (Le et al., 2020b), across all turn positions. However, future work is needed to close the performance gaps between lower and higher turn positions.

Impacts on downstream response prediction task. Finally, we tested the benefits of studying multimodal DST for a response prediction task. Specifically, we used the best VDTN model to predict dialogue states across all samples in DVD-DST. We then used the predicted slots, including object identities and temporal slots, to select the video features. The features are the visual objects and segments that are parts of the predicted dialogue states. We then used these selected features as input to train new Transformer decoder models which are added with an MLP as the response prediction layer. Note that these models are trained only with a cross-entropy loss to predict answer candidates.

Dialogue State	Response Accuracy
No state	43.0
$\mathcal{B}\setminus$time	46.8/47.1
$\mathcal{B}$	48.7/48.9

Table 5: Accuracy (in %) of response predictions (by greedy/beam search states)

From Table 5, we observed the benefits of filtering visual inputs by predicted states, with up to 5.9% accuracy score improvement [3]. Note that there are more sophisticated approaches such as neural module networks (Hu et al., 2018) and symbolic reasoning (Chen et al., 2020) to fully exploit the decoded dialogue states. We leave these extensions for future research.

For more experiment results, analysis, and qualitative examples, please refer to Appendix D.

5 Discussion and Conclusion

Compared to conventional DST (Mrkšić et al., 2017; Lei et al., 2018b; Gao et al., 2019; Le et al., 2020c), we show that the scope of DST can be further extended to a multimodal world. Compared to prior work in multimodal dialogues (Das et al., 2017a; Hori et al., 2019; Thomason et al., 2019) which focuses more on vision-language interactions, our work was inspired from a dialogue-based strategy with a formulation of a dialogue state tracking task. For more comparison to related work, please refer to Appendix A.

We noted the current work are limited to a synthetic benchmark with a limited video domain (3D objects). However, we expect that MM-DST task is still applicable and can be extended to other video domains (e.g. videos of humans). We expect that MM-DST is useful in dialogues centered around a "focus group" of objects. For further discussion of limitations, please refer to Appendix E.

In summary, in this work, we introduced a novel MM-DST task that tracks visual objects and their attributes mentioned in dialogues. For this task, we experimented on a synthetic benchmark with videos simulated in a 3D environment and dialogues grounded on these objects. Finally we proposed VDTN, a Transformer-based model with self-supervised learning objectives on object and segment-level visual representations.

[3] The response prediction performance is lower than the results reported by DVD (Le et al., 2021b) as the training splits are not the same; DVD-DST has about 10x smaller training data than (Le et al., 2021b).

6 Broader Impacts

During the research of this work, there is no human subject involved and hence, no ethical concerns regarding the experimental procedures and results. The data is used from a synthetically developed dataset, in which all videos are simulated in a 3D environment with synthetic non-human visual objects. We intentionally chose this dataset to minimize any distribution bias and make fair comparisons between all baseline models.

However, we wanted to emphasize on ethical usage of any potential adaptation of our methods in real applications. Considering the development of AI in various industries, the technology introduced in this paper may be used in practical applications, such as dialogue agents with human users. In these cases, the adoption of the MM-DST task or VDTN should be strictly used to improve the model performance and only for legitimate and authorized purposes. It is crucial that any plan to apply or extend MM-DST in real systems should consider carefully all potential stakeholders as well as the risk profiles of application domains. For instance, in case a dialogue state is extended to human subjects, any information used as slots should be clearly informed and approved by the human subjects before the slots are tracked.

References

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

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.

Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. 2016. Layer normalization. *arXiv preprint arXiv:1607.06450*.

Moshe Bar. 2004. Visual objects in context. *Nature Reviews Neuroscience*, 5(8):617–629.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium. Association for Computational Linguistics.

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)*.

Xinyun Chen, Chen Liang, Adams Wei Yu, Denny Zhou, Dawn Song, and Quoc V. Le. 2020. Neural symbolic reader: Scalable integration of distributed and symbolic representations for reading comprehension. In *International Conference on Learning Representations*.

Yinpei Dai, Hangyu Li, Yongbin Li, Jian Sun, Fei Huang, Luo Si, and Xiaodan Zhu. 2021. Preview, attend and review: Schema-aware curriculum learning for multi-domain dialogue state tracking. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 879–885, Online. Association for Computational Linguistics.

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

Abhishek Das, Satwik Kottur, José M. F. Moura, Stefan Lee, and Dhruv Batra. 2017b. Learning cooperative visual dialog agents with deep reinforcement learning. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2970–2979.

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.

Mihail Eric, Lakshmi Krishnan, Francois Charette, and Christopher D. Manning. 2017. Key-value retrieval networks for task-oriented dialogue. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 37–49, Saarbrücken, Germany. Association for Computational Linguistics.

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.

Tharindu Fernando, Simon Denman, Sridha Sridharan, and Clinton Fookes. 2018. Tracking by prediction: A deep generative model for mutli-person localisation and tracking. In *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1122–1132. IEEE.

Cora Fischer, Stefan Czoschke, Benjamin Peters, Benjamin Rahm, Jochen Kaiser, and Christoph Bledowski. 2020. Context information supports serial dependence of multiple visual objects across memory episodes. *Nature communications*, 11(1):1–11.

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. 2017. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*, pages 5267–5275.

Shuyang Gao, Abhishek Sethi, Sanchit Aggarwal, Tagyoung Chung, and Dilek Hakkani-Tur. 2019. Dialog state tracking: A neural reading comprehension approach. *arXiv preprint arXiv:1908.01946*.

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

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Matthew Henderson, Blaise Thomson, and Jason D Williams. 2014a. The second dialog state tracking challenge. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 263–272.

Matthew Henderson, Blaise Thomson, and Steve J. Young. 2014b. Robust dialog state tracking using delexicalised recurrent neural networks and unsupervised adaptation. *2014 IEEE Spoken Language Technology Workshop (SLT)*, pages 360–365.

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.

Ronghang Hu, Jacob Andreas, Trevor Darrell, and Kate Saenko. 2018. Explainable neural computation via stack neural module networks. In *Proceedings of the European conference on computer vision (ECCV)*, pages 53–69.

Ronghang Hu, Huazhe Xu, Marcus Rohrbach, Jiashi Feng, Kate Saenko, and Trevor Darrell. 2016. Natural language object retrieval. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4555–4564.

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.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*.

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.

Hyounghun Kim, Hao Tan, and Mohit Bansal. 2020. Modality-balanced models for visual dialogue. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8091–8098.

Diederick P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*.

Satwik Kottur, Seungwhan Moon, Alborz Geramifard, and Babak Damavandi. 2021. SIMMC 2.0: A task-oriented dialog dataset for immersive multimodal conversations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4903–4912, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Satwik Kottur, José MF Moura, Devi Parikh, Dhruv Batra, and Marcus Rohrbach. 2018. Visual coreference resolution in visual dialog using neural module networks. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 153–169.

Gakuto Kurata, Bing Xiang, Bowen Zhou, and Mo Yu. 2016. Leveraging sentence-level information with encoder LSTM for semantic slot filling. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2077–2083, Austin, Texas. Association for Computational Linguistics.

Hung Le, Nancy F. Chen, and Steven Hoi. 2021a. Learning reasoning paths over semantic graphs for video-grounded dialogues. In *International Conference on Learning Representations*.

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. 2020a. 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.

Hung Le, Doyen Sahoo, Chenghao Liu, Nancy Chen, and Steven C.H. Hoi. 2020b. UniConv: A unified conversational neural architecture for multi-domain task-oriented dialogues. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1860–1877, Online. Association for Computational Linguistics.

Hung Le, Chinnadhurai Sankar, Seungwhan Moon, Ahmad Beirami, Alborz Geramifard, and Satwik Kottur. 2021b. DVD: A diagnostic dataset for multi-step reasoning in video grounded dialogue. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5651–5665, Online. Association for Computational Linguistics.

Hung Le, Richard Socher, and Steven C.H. Hoi. 2020c. Non-autoregressive dialog state tracking. In *International Conference on Learning Representations*.

Hwaran Lee, Jinsik Lee, and Tae yoon Kim. 2019. Sumbt: Slot-utterance matching for universal and scalable belief tracking. In *ACL*.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. 2018a. 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.

Wenqiang Lei, Xisen Jin, Min-Yen Kan, Zhaochun Ren, Xiangnan He, and Dawei Yin. 2018b. Sequicity: Simplifying task-oriented dialogue systems with single sequence-to-sequence architectures. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1437–1447.

Zekang Li, Zongjia Li, Jinchao Zhang, Yang Feng, and Jie Zhou. 2021. Bridging text and video: A universal multimodal transformer for video-audio scene-aware dialog. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pages 1–1.

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.

Weizhe Lin, Bo-Hsiang Tseng, and Bill Byrne. 2021. Knowledge-aware graph-enhanced GPT-2 for dialogue state tracking. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7871–7881, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Xiangyang Mou, Brandyn Sigouin, Ian Steenstra, and Hui Su. 2020. Multimodal dialogue state tracking by qa approach with data augmentation. *arXiv preprint arXiv:2007.09903*.

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. Association for Computational Linguistics.

Wei Pang and Xiaojie Wang. 2020. Visual dialogue state tracking for question generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11831–11838.

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.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Osman Ramadan, Paweł Budzianowski, and Milica Gasic. 2018. Large-scale multi-domain belief tracking with knowledge sharing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 432–437.

Abhinav Rastogi, Dilek Z. Hakkani-Tür, and Larry P. Heck. 2017. Scalable multi-domain dialogue state tracking. *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 561–568.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28:91–99.

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.

Aviv Shamsian, Ofri Kleinfeld, Amir Globerson, and Gal Chechik. 2020. Learning object permanence from video. In *European Conference on Computer Vision*, pages 35–50. Springer.

Yangyang Shi, Kaisheng Yao, Hu Chen, Dong Yu, Yi-Cheng Pan, and Mei-Yuh Hwang. 2016. Recurrent support vector machines for slot tagging in spoken language understanding. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 393–399, San Diego, California. Association for Computational Linguistics.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. 2016. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Tana Wang, Yaqing Hou, Dongsheng Zhou, and Qiang Zhang. 2021. A contextual attention network for multimodal emotion recognition in conversation. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7. IEEE.

Tsung-Hsien Wen, David Vandyke, Nikola Mrkšić, Milica Gašić, Lina M. Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 2017. A network-based end-to-end trainable task-oriented dialogue system. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 438–449, Valencia, Spain. Association for Computational Linguistics.

Chien-Sheng Wu, Andrea Madotto, Ehsan Hosseini-Asl, Caiming Xiong, Richard Socher, and Pascale Fung. 2019. Transferable multi-domain state generator for task-oriented dialogue systems. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 808–819, Florence, Italy. Association for Computational Linguistics.

Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. 2017. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500.

Puyang Xu and Qi Hu. 2018. An end-to-end approach for handling unknown slot values in dialogue state tracking. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1448–1457. Association for Computational Linguistics.

Yaoda Xu and Marvin M Chun. 2009. Selecting and perceiving multiple visual objects. *Trends in cognitive sciences*, 13(4):167–174.

Steve Young, Milica Gašić, Simon Keizer, François Mairesse, Jost Schatzmann, Blaise Thomson, and Kai Yu. 2010. The hidden information state model: A practical framework for pomdp-based spoken dialogue management. *Computer Speech and Language*, 24(2):150–174.

Jianguo Zhang, Kazuma Hashimoto, Chien-Sheng Wu, Yao Wang, Philip Yu, Richard Socher, and Caiming Xiong. 2020. Find or classify? dual strategy for slot-value predictions on multi-domain dialog state tracking. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, pages 154–167, Barcelona, Spain (Online). Association for Computational Linguistics.

Victor Zhong, Caiming Xiong, and Richard Socher. 2018. Global-locally self-attentive encoder for dialogue state tracking. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1458–1467, Melbourne, Australia. Association for Computational Linguistics.

Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. 2020. Unified vision-language pre-training for image captioning and vqa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 13041–13049.

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 Details of Related Work

Our work is related to the following domains:

A.1 Dialogue State Tracking

Dialogue State Tracking (DST) research aims to develop models that can track essential information conveyed in dialogues between a dialogue agent and human (defined as hidden information state by (Young et al., 2010) or belief state by (Mrkšić et al., 2017)). DST research has evolved largely within the domain of task-oriented dialogue systems. DST is conventionally designed in a modular dialogue system (Wen et al., 2017; Budzianowski et al., 2018; Le et al., 2020b) and preceded by a Natural Language Understanding (NLU) component. NLU learns to label sequences of dialogue utterances and provides a tag for each word token (often in the form of In-Out-Begin representations) (Kurata et al., 2016; Shi et al., 2016; Rastogi et al., 2017). To avoid credit assignment problems and streamline the modular designs, NLU and DST have been integrated into a single module (Mrkšić et al., 2017; Xu and Hu, 2018; Zhong et al., 2018). These DST approaches can be roughly categorized into two types: fixed-vocabulary or open-vocabulary. Fixed-vocabulary approaches are designed for classification tasks which assume a fixed set of *(slot, value)* candidates and directly retrieve items from this set to form dialogue states during test time (Henderson et al., 2014b; Ramadan et al., 2018; Lee et al., 2019). More recently, we saw more approaches toward open-vocabulary strategies which learn to generate candidates based on input dialogue context (Lei et al., 2018b; Gao et al., 2019; Wu et al., 2019; Le et al., 2020c). Our work is more related to open-vocabulary DST, but we essentially redefined the DST task with multimodality. Based on our literature review, we are the first to formally extend DST and bridge the gap between traditional task-oriented dialogues and multimodal dialogues.

A.2 Visually-grounded Dialogues

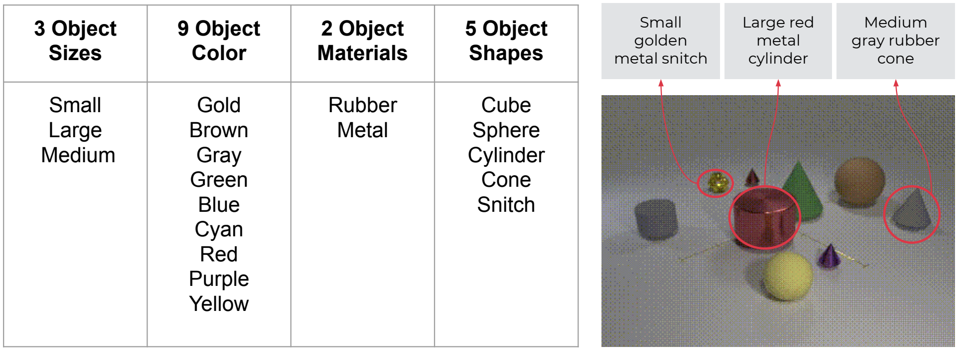
A novel challenge to machine intelligence, the intersection of vision and language research has expanded considerably in the past few years. Earlier benchmarks test machines to perceive visual inputs, and learn to generate captions (Farhadi et al., 2010; Lin et al., 2014; Rohrbach et al., 2015), ground text phrases and objects (Kazemzadeh et al., 2014; Plummer et al., 2015), and answer questions about the visual contents (Antol et al., 2015; Zhu et al., 2016; Jang et al., 2017; Lei et al., 2018a). As an orthogonal development from Visual Question Answering problems, we noted recent work that targets vision-language in dialogue context, in which an image or video is given and the dialogue utterances are centered around its visual contents (De Vries et al., 2017; Das et al., 2017a; Chattopadhyay et al., 2017; Hori et al., 2019; Thomason et al., 2019; Le et al., 2021b). Recent work has addressed different challenges in visually-grounded dialogues, including multimodal integration (Hori et al., 2019; Le et al., 2019; Li et al., 2021), cross-turn dependencies (Das et al., 2017b; Schwartz et al., 2019; Le et al., 2021a), visual understanding (Le et al., 2020a), and data distribution bias (Qi et al., 2020). Our work is more related to the challenge of visual object reasoning (Seo et al., 2017; Kottur et al., 2018), but focused on a multi-turn tracking task over multiple turns of dialogue context. The prior approaches are not well designed to track objects and maintain a recurring memory or state of these objects from turn to turn. This challenge becomes more obvious when a dialogue involves multiple objects of similar characters or appearance. We directly tackles this challenge as we formulated a novel multimodal state tracking task and leveraged the research development from DST in task-oriented dialogue systems. As shown in our experiments, baseline models that use attention strategies similar to (Seo et al., 2017; Kottur et al., 2018) did not perform well in MM-DST.

A.3 Multimodal DST

We noted a few studies have attempted to integrate some forms of state tracking in multimodal dialogues. In (Mou et al., 2020), however, we are not convinced that a dialogue state tracking task is a major focus, or correctly defined. In (Pang and Wang, 2020), we noted that some form of object tracking is introduced throughout dialogue turns. The tracking module is used to decide which object

Split	# Videos	# Dialogues	# Turns	# Slots
DVD-DST-Train	9300	9295	92950	6
DVD-DST-Val	3327	3326	33260	6
DVD-DST-Test	1371	1371	13710	6
DVD-DST-All	13998	13992	139920	6
MultiWOZ (Budzianowski et al., 2018)	N/A	8438	115424	25
CarAssistant (Eric et al., 2017)	N/A	2425	12732	13
WOZ2 (Wen et al., 2017)	N/A	600	4472	4
DSTC2 (Henderson et al., 2014a)	N/A	1612	23354	8

Table 6: Dataset summary: statistics of related benchmarks are from (Budzianowski et al., 2018)



3 Object Sizes	9 Object Color	2 Object Materials	5 Object Shapes
Small Large Medium	Gold Brown Gray Green Blue Cyan Red Purple Yellow	Rubber Metal	Cube Sphere Cylinder Cone Snitch

Figure 4: Synthetic visual objects in the CATER universe

the dialogue centers around. This method extends to multi-object tracking but the objects are only limited within static images, and there is no recurring information state (object attributes) maintained at each turn. Compared to our work, their tracking module only requires object identity as a single-slot state from turn to turn. Almost concurrent to our work, we noted (Kottur et al., 2021) which formally, though very briefly, focuses on multimodal DST. However, the work is limited to the task-oriented domain, and each dialogue is only limited to a single goal-driven object in a synthetic image. While this definition is useful in the task-oriented dialogue domain, it does not account for the DST of multiple visual objects as defined in our work.

B DVD-DST Dataset Details

For each of CATER videos from the extended split (Shamsian et al., 2020), we generated up to 10 turns for each CATER video. In total, DVD-DST contains more than $13k$ dialogues, resulting in more $130k$ (human, system) utterance pairs and corresponding dialogue states. A comparison of statistics of DVD-DST and prior DST benchmarks can be seen in Table 6. We observed that DVD-DST contains a larger scale data than the related DST benchmark. Even though the number of slots in DVD-DST is only 6, lower than prior state tracking datasets, our experiments indicate that most current conventional DST models perform poorly on DVD-DST.

CATER universe. Figure 4 displays the configuration of visual objects in the CATER universe. In total, there are 3 object sizes, 9 colors, 2 materials, and 5 shapes. These attributes are combined randomly to synthesize objects in each CATER video. We directly adopted these attributes as slots in dialogue states, and each dialogue utterance frequently refers to these objects by one or more attributes. In total, there are 193 (size, color, material, shape) valid combinations, each of which corresponds to an object class in our models.

Sample dialogues. Please refer to Figure 5, Table 14 and Table 15.

Usage. We want to highlight that the DVD-DST dataset should only be used for its intended purpose, i.e. to diagnose dialogue systems on their tracking abilities. Any derivatives of the data should be limited within the research contexts of MM-DST.

C Additional Training Details

In practice, we applied label smoothing (Szegedy et al., 2016) on state sequence labels to regularize the training. As the segment-level representations are stacked by the number of objects, we randomly selected only one vector per masked segment to apply $\mathcal{L}_{seg}$. We tested both L1 and L2 losses on $\mathcal{L}_{bb/seg}$. All model parameters, except pretrained visual perception models, are initialized by a uniform distribution (Glorot and Bengio, 2010).

Video	Dialogue	Video	Greedy			Beam Search		
Features	State	loss	Joint Obj	Joint State	Joint State	Joint Obj	Joint State	Joint State
			State Acc	IoU@0.5	IoU@0.7	State Acc	IoU@0.5	IoU@0.7
X_{bb}	$\mathcal{B}\backslash$time	-	17.3%	N/A	N/A	17.9%	N/A	N/A
$X_{bb} + X_{cnn}$	$\mathcal{B}\backslash$time	-	20.0%	N/A	N/A	22.4%	N/A	N/A
X_{bb}	$\mathcal{B}$	-	16.6%	9.6%	8.3%	19.3%	11.0%	9.5%
$X_{bb} + X_{cnn}$	$\mathcal{B}$	-	22.4%	12.7%	10.8%	24.8%	13.8%	11.8%
X_{bb}	$\mathcal{B}$	$\mathcal{L}_{obj}$	21.7%	11.7%	10.0%	24.0%	12.9%	11.0%
$X_{bb} + X_{cnn}$	$\mathcal{B}$	$\mathcal{L}_{obj}$	23.1%	13.2%	11.3%	26.0%	14.4%	12.4%
$X_{bb} + X_{cnn}$	$\mathcal{B}$	$\mathcal{L}_{seg}$	**24.3%**	**13.4%**	**11.4%**	**28.0%**	**15.3%**	**13.1%**

Table 7: Ablation results by joint state predictions, using greedy or beam search decoding styles

Video Features	Dialogue State	Video self-supervision	Obj Identity F1	Obj Slot F1	Obj State F1	Size F1	Color F1	Material F1	Shape F1
X_{bb}	$\mathcal{B}\backslash$time	-	79.4%	64.2%	48.5%	55.9%	76.6%	41.4%	63.5%
$X_{bb} + X_{cnn}$	$\mathcal{B}\backslash$time	-	81.4%	66.9%	52.5%	58.0%	79.4%	39.5%	66.6%
X_{bb}	$\mathcal{B}$	-	78.5%	63.6%	49.8%	56.5%	76.4%	38.8%	63.1%
$X_{bb} + X_{cnn}$	$\mathcal{B}$	-	83.3%	69.4%	55.1%	56.7%	81.8%	47.0%	69.8%
X_{bb}	$\mathcal{B}$	$\mathcal{L}_{obj}$	82.2%	69.5%	56.2%	61.4%	81.0%	44.9%	69.9%
$X_{bb} + X_{cnn}$	$\mathcal{B}$	$\mathcal{L}_{obj}$	**84.7%**	72.0%	58.6%	59.7%	83.5%	**52.3%**	71.7%
$X_{bb} + X_{cnn}$	$\mathcal{B}$	$\mathcal{L}_{seg}$	84.5%	**72.8%**	**60.4%**	**64.1%**	**84.2%**	50.9%	**71.9%**

Table 8: Ablation results by component predictions of object identities, slots, and object states

For fair comparison among baselines, all models use both object-level and segment-level feature representations, encoded by the same method as Describe in Section 3.1. In *TRADE*, the video representations are passed to an RNN encoder, and the output hidden states are concatenated to the dialogue hidden states. Both are passed to the original pointer-based decoder. In *UniConv* and *NADAST*, we stacked another Transformer attention layer to attend on video representations before the original state-to-dialogue attention layer. We all baseline models, we replaced the original (domain, slot) embeddings as (object class, slot) embeddings and kept the original model designs.

Note that in our visual perception model, we adopted the finetuned Faster R-CNN model used by (Shamsian et al., 2020). The model was finetuned to predict object bounding boxes and object classes. The object classes are derived based on object appearance, based on the four attributes of size, color, material, and shape. In total, there are 193 object classes. For segment embeddings, we adopted the ResNeXt-101 model (Xie et al., 2017) finetuned on Kinetics dataset (Kay et al., 2017). For all models (except for VDTN ablation analysis), we standardized $N_{obj} = 10$ and $N_{stride} = 12$ to sub-sample object and segment-level embeddings.

Resources. Note that all experiments did not require particularly large computing resources as we limited all model training to a single GPU, specifically on a Tesla V100 GPU of 16G configuration.

D Additional Results

Greedy vs. Beam Search Decoding. Table 7 shows the results of different variants of VDTN models. We observed that compared to greedy decoding, beam search decoding improves the performance in all models. As beam search decoding selects the best decoded state by the joint probabilities of tokens, this observation indicates the benefits of considering slot values to be co-dependent and their relationships should be modelled. This is consistent with similar observations in later work of unimodal DST (Lei et al., 2018b; Le et al., 2020c).

Ablation analysis by component predictions. From Table 8, we have the following observations: (1) In ablation results by component predictions, we noted that models can generally detect object identities well with F1 about 80%. However, when considering object and slot tuples, F1 reduces to $48 - 60\%$, indicating the gaps are caused by slot value predictions. (2) By individual slots, we noted "color" and "shape" slots are easier to track than

Video Features	Dialogue State	Video self-supervision	Obj Identity Recall	Obj Identity Precision	Obj Slot Recall	Obj Slot Precision	Obj State Recall	Obj State Precision
X_{bb}	$\mathcal{B}\backslash$time	-	77.2%	81.8%	65.0%	63.4%	47.1%	50.0%
$X_{bb} + X_{cnn}$	$\mathcal{B}\backslash$time	-	75.1%	88.8%	63.1%	71.3%	48.5%	57.3%
X_{bb}	$\mathcal{B}$	-	73.6%	84.1%	61.7%	65.7%	46.7%	53.4%
$X_{bb} + X_{cnn}$	$\mathcal{B}$	-	78.2%	89.1%	66.2%	73.0%	51.7%	58.9%
X_{bb}	$\mathcal{B}$	$\mathcal{L}_{obj}$	76.4%	88.9%	67.4%	71.7%	52.2%	60.8%
$X_{bb} + X_{cnn}$	$\mathcal{B}$	$\mathcal{L}_{obj}$	80.1%	**90.0%**	69.1%	75.2%	55.4%	62.2%
$X_{bb} + X_{cnn}$	$\mathcal{B}$	$\mathcal{L}_{seg}$	**80.5%**	89.0%	**70.2%**	**75.6%**	**57.6%**	**63.6%**

Table 9: Ablation results by individual object identity/slot/state

Video Features	Dialogue State	Video self-supervision	Size Recall	Size Precision	Color Recall	Color Precision	Material Recall	Material Precision	Shape Recall	Shape Precision
X_{bb}	$\mathcal{B}\backslash$time	-	60.1%	52.2%	76.8%	76.4%	43.2%	39.7%	61.4%	65.6%
$X_{bb} + X_{cnn}$	$\mathcal{B}\backslash$time	-	52.0%	65.6%	76.2%	82.9%	34.8%	45.8%	65.5%	67.8%
X_{bb}	$\mathcal{B}$	-	52.0%	61.9%	72.0%	81.2%	40.8%	37.1%	63.3%	63.0%
$X_{bb} + X_{cnn}$	$\mathcal{B}$	-	49.4%	66.5%	79.2%	84.6%	45.0%	49.2%	68.9%	70.6%
X_{bb}	$\mathcal{B}$	$\mathcal{L}_{obj}$	59.6%	63.4%	79.3%	82.9%	43.8%	46.0%	66.6%	73.5%
$X_{bb} + X_{cnn}$	$\mathcal{B}$	$\mathcal{L}_{obj}$	54.1%	66.6%	82.4%	84.7%	**48.8%**	**56.3%**	**69.3%**	74.3%
$X_{bb} + X_{cnn}$	$\mathcal{B}$	$\mathcal{L}_{seg}$	**60.9%**	**67.7%**	**83.2%**	**85.4%**	48.6%	53.4%	67.9%	**76.5%**

Table 10: Ablation results by individual slot type

"size" and "material" slots. We noted that in the CATER universe, the latter two slots have lower visual variances (less possible values) than the others. As a result, objects are more likely to share the same size or material and hence, discerning objects by those slots and tracking them in dialogues become more challenging.

Table 9 and 10 display the ablation results by component predictions, using precision and recall metrics. We still noted consistent observations as described in Section 4. Notably, we found that current VDTN models are better in tuning the correct predictions (as shown by high precision metrics) but still fail to select all components as a set (as shown by low recall metrics). This might be caused by the upstream errors coming from the visual perception models, which may fail to visually perceive all objects and their attributes.

Results by turn positions. Table 11 reported the results of VDTN predictions of states that are separated by the corresponding dialogue positions. The results are from the VDTN model trained with both $\mathcal{L}_{dst}$ and $\mathcal{L}_{seg}$. As expected, we observed a downward trend of results as the turn position increases.

Impacts of dialogue context encoder. In Table 12a, we observed the benefits of using the Markov process to decode dialogue states based on the dialogue states of the last turn $\mathcal{B}_{t-1}$. This strategy allow us to discard parts of dialogue history that is already represented by the state. We noted that the optimal design is to use at least 1 last dialogue turn as the dialogue history. In a hypothetical scenario, we applied the oracle $\mathcal{B}_{t-1}$ during test time, and noted the performance is improved significantly. This observation indicates the sensitivity of VDTN to a turn-wise auto-regressive decoding process.

Impacts of frame-level and segment-level sampling. As expected, Table 12b displays higher performance with higher object limits N_{obj}, which increases the chance of detecting the right visual objects in videos. We noted performance gains when sampling strides increase up to 24 frames. However, in the extreme case, when sampling stride is 300 frames, the performance on temporal slots reduce (as shown by "Joint State IoU@p"). This raises the issue to sample data more efficiently by balancing between temporal sparsity in videos and state prediction performance. We also observed that in a hypothetical scenario with a perfect object perception model, the performance improves significantly, especially on the predictions of discrete slots, although less effect on temporal slots.

Impacts of object-level representation. Table 13 reported the results when only segment-level features are used. We observed that both VDTN and *RNN(V+D)* are affected significantly, specifically by 24% and 3.1% "Joint Obj State Acc" score respectively. Interestingly, we noted that *RNN(V)*, using only video inputs, are not affected by the removal of object-level features. These observations

Turn Position	Obj Identity F1	Obj Slot F1	Obj State F1	Joint Obj State Acc	Joint State IoU@0.5	Joint State IoU@0.7
1	**88.8%**	**84.0%**	**82.4%**	**74.0%**	**40.5%**	**34.6%**
2	86.9%	81.1%	77.2%	60.0%	37.5%	33.6%
3	84.9%	77.6%	71.0%	41.6%	22.8%	19.5%
4	84.2%	75.6%	66.5%	29.0%	15.2%	12.5%
5	84.0%	74.0%	63.1%	21.3%	11.3%	9.4%
6	84.3%	73.0%	60.2%	17.1%	9.6%	8.2%
7	83.9%	71.6%	57.1%	12.7%	6.1%	5.3%
8	84.1%	70.6%	54.9%	10.2%	4.7%	3.9%
9	84.0%	69.1%	51.8%	7.9%	3.6%	2.6%
10	84.1%	68.0%	49.5%	6.0%	2.3%	1.7%
Average	84.9%	74.5%	63.4%	28.0%	15.3%	13.1%

Table 11: Ablation results by dialogue turn positions

$\mathcal{B}_{t-1}$	Max turns	Joint Obj State Acc	Joint State IoU@0.5	Joint State IoU@0.7
✓	10	22.5%	11.5%	10.1%
✓	7	22.0%	11.8%	10.4%
✓	1	24.8%	13.8%	11.8%
✓	0	22.3%	12.3%	10.5%
-	10	18.5%	9.4%	8.6%
-	7	19.0%	9.5%	8.7%
-	1	7.8%	4.5%	4.1%
-	0	1.3%	0.7%	0.7%
✓*	1	29.3%	18.6%	16.4%

(a) dialogue encoding by prior states and dialogue sizes: $*$ denotes using oracle values.

N_{obj}	N_{stride}	Joint Object State Acc	Joint State IoU@0.5	Joint State IoU@0.7
10	12	24.8%	13.8%	11.8%
7	12	18.0%	10.1%	9.0%
3	12	4.9%	2.9%	2.6%
0	12	1.5%	0.7%	0.7%
10	300	28.2%	6.0%	3.7%
10	24	27.8%	14.8%	12.6%
10	15	26.3%	14.4%	12.4%
10	12	24.8%	13.8%	11.8%
10*	12	29.2%	15.6%	13.4%

(b) video encoding by number of objects and sampling strides: $*$ denotes perfect object perception.

Table 12: Ablation results by encoding strategies: All models are trained only with $\mathcal{L}_{dst}$.

indicate that current MM-DST requires object-level information. We expected that existing 3DCNN models such as ResNeXt still fail to capture such level of granularity.

Qualitative analysis. Table 14 and 15 display 2 sample dialogues and state predictions. We displayed the corresponding video screenshots for these dialogues in Figure 5. To cross-reference between videos and dialogues, we displayed the bounding boxes and their object classes in video screenshots. These object classes are indicated in ground-truth and decoded dialogue states in dialogues. Overall, we noted that VDTN generated temporal slots of start and end time such that the resulting periods better match the ground-truth temporal segments. VDTN also showed to maintain the dialogue states better from turn to turn.

E Further Discussion

Synthetic datasets result in overestimation of real performance and don't translate to real-world usability. We agree that the current state accuracy seems to be quite low at about 28%. However, we want to highlight that state accuracy used in this paper is a very strict metric, which only considers a prediction as correct if it completely matches the ground truth. In DVD, assuming the average 10 objects per video with the set of attributes as in Figure 4 (+ 'none' value in each slot), we can roughly equate the multimodal DST as a 7200-class classification task, each class is a distinct set of objects, each with all possible attribute combinations. Combined with the cascading error from object perception models, we think the current reported results are reasonable.

Moreover, we want to highlight that the reported performance of baselines reasonably matches their own capacities in unimodal DST. We can consider Object State F1 as the performance on single-object state and it can closely correlate with the joint state accuracy in unimodal DST (remember that unimodal DST such as MultiWOZ (Budzianowski et al., 2018) is only limited to a single object/entity per dialogue). As seen in Table 2, the Object State F1 results of TRADE (Wu et al., 2019), UniConv (Le et al., 2020b), and NADST (Le et al., 2020c) are between 46-50%. This performance range is indeed not very far off from the performance of these baseline models in unimodal DST in the MultiWOZ benchmark (Budzianowski et al., 2018).

Finally, we also want to highlight that like other synthetic benchmarks such as CLEVR (Johnson

Model	$X_{bb} + X_{cnn}$			X_{cnn} only		
	Joint Obj State Acc	Joint State IoU@0.5	Joint State IoU@0.7	Joint Obj State Acc	Joint State IoU@0.5	Joint State IoU@0.7
VDTN	28.0%	15.3%	13.1%	4.0%	2.2%	2.0%
RNN(V)	1.0%	0.1%	0.1%	1.5%	0.4%	0.4%
RNN(V+D)	6.8%	2.6%	2.3%	3.7%	1.8%	1.6%

Table 13: Results with and without object representations

et al., 2017), we want to use DVD in this work as a test bed to study and design better multimodal dialogue systems. However, we do not intend to use it as a training data for practical systems. The DVD-DST benchmark should be used to supplement real-world video-grounded dialogue datasets.

MM-DST in practical applications e.g. with videos of humans. While we introduced MM-DST task and VDTN as a new baseline, we noted that the existing results are limited to the synthetic benchmark. For instance, in the real world, there would be many identical objects with the same (size, color, material, shape) tuples, which would make the current formulation of dialogue states difficult. In such object-driven conversations, we would recommend a dialogue agent not focus on all possible objects in each video frame, but only on a "focus group" of objects. These objects, required to be semantically different, are topical subjects of the conversations.

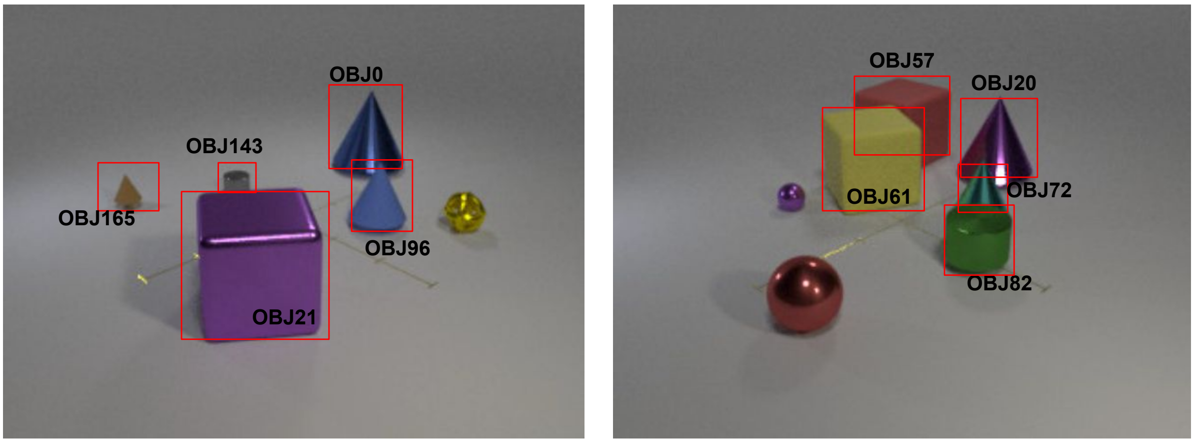
Say we want to scale to a new domain e.g. videos of humans, the first challenge from the current study is the recognition of human objects, which often have higher visual complexity than moving objects as in DVD. We also noted that it is impossible to define all human object classes as in CATER object classes, each of which is unique by its own appearance. To overcome this limitation, we would want to explore multimodal DST with the research of human object tracking, e.g. (Fernando et al., 2018), and consider human object identities uniquely defined per video. Another limitation is the definition of slots to track in each human object. While this requires careful considerations, for both practical and ethical reasons, we noted several potential papers that investigate human attributes in dialogues such as human emotions (Wang et al., 2021). Along these lines, we are excited to see interesting adaptations of multimodal dialogue states grounded on videos of humans.

#1	HUMAN:	after the cube 's second rotation , how many other things perform the same sequence of activities as the brown thing ?
	Gold:	STAR=102, END=138, (OBJ21, SHAPE, cube), (OBJ165, COLOR, brown)
	RNN+Attn:	STAR=73, END=73, (OBJ165, COLOR, brown)
	UniConv:	STAR=73, END=145, (OBJ142, SHAPE, cube), (OBJ165, COLOR, brown)
	VDTN:	STAR=97, END=145, (OBJ21, SHAPE, cube), (OBJ165, COLOR, brown)
#2	SYSTEM:	3
	HUMAN:	what about up until now ?
	Gold:	STAR=102, END=300, (OBJ21, SHAPE, cube), (OBJ165, COLOR, brown)
	RNN+Attn:	STAR=1, END=300, (OBJ165, COLOR, brown)
	UniConv:	STAR=145, END=300, (OBJ142, SHAPE, cube), (OBJ165, COLOR, brown)
	VDTN:	STAR=97, END=300, (OBJ21, SHAPE, cube), (OBJ165, COLOR, brown)
#3	SYSTEM:	0
	HUMAN:	throughout the whole video, do the large cone and the earlier mentioned brown thing perform the same set of activities ?'
	Gold:	STAR=1, END=300, (OBJ0, SIZE, large), (OBJ0, SHAPE, cone), (OBJ21, SHAPE, cube), (OBJ165, COLOR, brown)
	RNN+Attn:	STAR=1, END=300, (OBJ165, COLOR, brown)
	UniConv:	STAR=1, END=300, (OBJ142, SHAPE, cube), (OBJ165, COLOR, brown) (OBJ165, SHAPE, cone)
	VDTN:	STAR=1, END=300, (OBJ21, SHAPE, cube), (OBJ165, COLOR, brown)
#4	SYSTEM:	FALSE
	HUMAN:	during the earlier mentioned block 's last rotation , what is the earlier mentioned brown thing doing ?
	Gold:	STAR=175, END=200, (OBJ0, SIZE, large), (OBJ0, SHAPE, cone), (OBJ21, SHAPE, cube), (OBJ165, COLOR, brown)
	RNN+Attn:	STAR=1, END=300, (OBJ165, COLOR, brown)
	UniConv:	STAR=265, END=300, (OBJ142, SHAPE, cube), (OBJ165, COLOR, brown) (OBJ165, SHAPE, cone)
	VDTN:	STAR=241, END=265, (OBJ21, SHAPE, cube), (OBJ165, COLOR, brown)
#5	SYSTEM:	no action
	HUMAN:	considering that object, after its slide is there a flying small yellow thing ?'
	Gold:	STAR=291, END=300, (OBJ0, SIZE, large), (OBJ0, SHAPE, cone), (OBJ21, SHAPE, cube), (OBJ165, COLOR, brown)
	RNN+Attn:	STAR=1, END=300, (OBJ165, COLOR, brown)
	UniConv:	STAR=277, END=300, (OBJ142, SHAPE, cube), (OBJ165, COLOR, brown) (OBJ165, SHAPE, cone)
	VDTN:	STAR=288, END=300, (OBJ21, SHAPE, cube), (OBJ165, COLOR, brown)
#6	SYSTEM:	FALSE
	HUMAN:	during the purple thing 's second rotation , what number of big metal things are on the right side of the earlier mentioned brown thing ?
	Gold:	STAR=80, END=102, (OBJ0, SIZE, large), (OBJ0, SHAPE, cone), (OBJ21, COLOR, purple), (OBJ21, SHAPE, cube), (OBJ165, COLOR, brown)
	RNN+Attn:	STAR=1, END=300, (OBJ21, COLOR, purple), (OBJ165, COLOR, brown)
	UniConv:	STAR=277, END=300, (OBJ142, SHAPE, cube), (OBJ165, COLOR, brown) (OBJ165, SHAPE, cone)
	VDTN:	STAR=73, END=97, (OBJ21, COLOR, purple), (OBJ21, SHAPE, cube), (OBJ165, COLOR, brown)
#7	SYSTEM:	1
	HUMAN:	what about left of it ?'
	Gold:	STAR=80, END=102, (OBJ0, SIZE, large), (OBJ0, SHAPE, cone), (OBJ21, COLOR, purple), (OBJ21, MATERIAL, metal), (OBJ21, SHAPE, cube), (OBJ165, COLOR, brown)
	RNN+Attn:	STAR=1, END=300, (OBJ21, COLOR, purple), (OBJ165, COLOR, brown)
	UniConv:	STAR=73, END=97, (OBJ21, COLOR, purple), (OBJ142, SHAPE, cube), (OBJ165, COLOR, brown) (OBJ165, SHAPE, cone)
	VDTN:	STAR=73, END=97, (OBJ21, SIZE, large), (OBJ21, COLOR, purple), (OBJ21, SHAPE, cube), (OBJ143, MATERIAL, metal), (OBJ165, COLOR, brown)
#8	SYSTEM:	0
	HUMAN:	before this time window, how many other objects undertake the same types of actions as the earlier mentioned cone ?'
	Gold:	STAR=1, END=80, (OBJ0, SIZE, large), (OBJ0, SHAPE, cone), (OBJ21, COLOR, purple), (OBJ21, MATERIAL, metal), (OBJ21, SHAPE, cube), (OBJ165, COLOR, brown)
	RNN+Attn:	STAR=1, END=300, (OBJ21, COLOR, purple), (OBJ165, COLOR, brown)
	UniConv:	STAR=1, END=73, (OBJ21, COLOR, purple), (OBJ21, SHAPE, cube), (OBJ165, COLOR, brown) (OBJ165, SHAPE, cone)
	VDTN:	STAR=1, END=73, (OBJ21, SIZE, large), (OBJ21, COLOR, purple), (OBJ21, SHAPE, cube), (OBJ143, MATERIAL, metal), (OBJ165, COLOR, brown)
#9	SYSTEM:	2
	HUMAN:	among them, there is a cone . throughout the whole video, is there any other thing that has the same types of actions as it ?'
	Gold:	STAR=1, END=300, (OBJ0, SIZE, large), (OBJ0, SHAPE, cone), (OBJ21, COLOR, purple), (OBJ21, MATERIAL, metal), (OBJ21, SHAPE, cube), (OBJ96, SHAPE, cone), (OBJ165, COLOR, brown)
	RNN+Attn:	STAR=1, END=300, (OBJ21, COLOR, purple), (OBJ165, COLOR, brown)
	UniConv:	STAR=1, END=300, (OBJ21, COLOR, purple), (OBJ21, SHAPE, cube), (OBJ96, COLOR, blue), (OBJ165, COLOR, brown) (OBJ165, SHAPE, cone)
	VDTN:	STAR=1, END=300, (OBJ21, SIZE, large), (OBJ21, COLOR, purple), (OBJ21, SHAPE, cube), (OBJ143, MATERIAL, metal), (OBJ165, COLOR, brown)
#10	SYSTEM:	FALSE
	HUMAN:	until the end of the blue shiny thing 's last flight , does the earlier mentioned brown object fly as frequently as the cylinder rotates ?
	Gold:	STAR=1, END=228, (OBJ0, SIZE, large), (OBJ0, COLOR, blue), (OBJ0, MATERIAL, metal), (OBJ0, SHAPE, cone), (OBJ21, COLOR, purple), (OBJ21, MATERIAL, metal), (OBJ21, SHAPE, cube), (OBJ96, SHAPE, cone), (OBJ143, SHAPE, cylinder), (OBJ165, COLOR, brown)
	RNN+Attn:	STAR=1, END=300, (OBJ21, COLOR, purple), (OBJ96, COLOR, blue), (OBJ143, SHAPE, cylinder), (OBJ165, COLOR, brown)
	UniConv:	STAR=1, END=300, (OBJ21, COLOR, purple), (OBJ96, COLOR, blue), (OBJ143, SHAPE, cylinder), (OBJ165, COLOR, brown)
	VDTN:	STAR=1, END=241, (OBJ21, SIZE, large), (OBJ21, COLOR, purple), (OBJ21, SHAPE, cube), (OBJ143, MATERIAL, metal), (OBJ165, COLOR, brown)

Table 14: Example predictions from a sample dialogue (CATER video ID #002660)

#1	HUMAN:	during the red block 's rotation, how many matte things are to the left of the green object ?
	Gold:	STAR=143, END=163, (OBJ57, COLOR, red), (OBJ57, SHAPE, cube), (OBJ82, COLOR, green)
	RNN+Attn:	STAR=37, END=109, (OBJ57, SHAPE, cube), (OBJ82, COLOR, green), (OBJ91, COLOR, red)
	UniConv:	STAR=73, END=109, (OBJ57, MATERIAL, rubber), (OBJ82, COLOR, green), (OBJ91, COLOR, red)
	VDTN:	STAR=73, END=109, (OBJ57, COLOR, red), (OBJ57, SHAPE, cube), (OBJ82, COLOR, green)
#2	SYSTEM:	2
	HUMAN:	how about in front of it ?
	Gold:	STAR=143, END=163, (OBJ57, COLOR, red), (OBJ57, SHAPE, cube), (OBJ82, COLOR, green)
	RNN+Attn:	STAR=37, END=25, (OBJ82, COLOR, green), (OBJ91, COLOR, red)
	UniConv:	STAR=73, END=109, (OBJ57, MATERIAL, rubber), (OBJ82, COLOR, green), (OBJ91, COLOR, red)
	VDTN:	STAR=73, END=109, (OBJ57, COLOR, red), (OBJ57, SHAPE, cube), (OBJ82, COLOR, green)
#3	SYSTEM:	0
	HUMAN:	before this time window, how many moving things are there ?
	Gold:	STAR=1, END=143, (OBJ57, COLOR, red), (OBJ57, SHAPE, cube), (OBJ82, COLOR, green)
	RNN+Attn:	STAR=1, END=1, (OBJ82, COLOR, green), (OBJ91, COLOR, red)
	UniConv:	STAR=1, END=73, (OBJ57, COLOR, red), (OBJ57, MATERIAL, rubber), (OBJ82, COLOR, green), (OBJ91, COLOR, red)
	VDTN:	STAR=1, END=73, (OBJ57, COLOR, red), (OBJ57, SHAPE, cube), (OBJ82, COLOR, green)
#4	SYSTEM:	5
	HUMAN:	among them, there is a purple thing . during the whole video, how many other things with the same sequence activities performed by it ?
	Gold:	STAR=1, END=168, (OBJ20, COLOR, purple), (OBJ57, COLOR, red), (OBJ57, SHAPE, cube), (OBJ82, COLOR, green)
	RNN+Attn:	STAR=1, END=169, (OBJ82, COLOR, green), (OBJ91, COLOR, red)
	UniConv:	STAR=1, END=169, (OBJ57, COLOR, red), (OBJ57, MATERIAL, rubber), (OBJ82, COLOR, green), (OBJ91, COLOR, red), (OBJ152, COLOR, purple)
	VDTN:	STAR=1, END=169, (OBJ20, COLOR, purple), (OBJ57, COLOR, red), (OBJ57, SHAPE, cube), (OBJ82, COLOR, green)
#5	SYSTEM:	1
	HUMAN:	how about up until now ?
	Gold:	STAR=1, END=247, (OBJ20, COLOR, purple), (OBJ57, COLOR, red), (OBJ57, SHAPE, cube), (OBJ82, COLOR, green)
	RNN+Attn:	STAR=1, END=253, (OBJ82, COLOR, green), (OBJ91, COLOR, red)
	UniConv:	STAR=1, END=253, (OBJ57, COLOR, red), (OBJ57, MATERIAL, rubber), (OBJ82, COLOR, green), (OBJ91, COLOR, red), (OBJ152, COLOR, purple)
	VDTN:	STAR=1, END=253, (OBJ20, COLOR, purple), (OBJ57, COLOR, red), (OBJ57, SHAPE, cube), (OBJ82, COLOR, green)
#6	SYSTEM:	0
	HUMAN:	during the earlier mentioned purple thing 's flight , is there a flying green object ?
	Gold:	STAR=45, END=69, (OBJ20, COLOR, purple), (OBJ57, COLOR, red), (OBJ57, SHAPE, cube), (OBJ82, COLOR, green)
	RNN+Attn:	STAR=1, END=241, (OBJ82, COLOR, green), (OBJ91, COLOR, red)
	UniConv:	STAR=73, END=97, (OBJ57, COLOR, red), (OBJ57, MATERIAL, rubber), (OBJ82, COLOR, green), (OBJ91, COLOR, red), (OBJ152, COLOR, purple)
	VDTN:	STAR=73, END=97, (OBJ20, COLOR, purple), (OBJ57, COLOR, red), (OBJ57, SHAPE, cube), (OBJ82, COLOR, green)
#7	SYSTEM:	FALSE
	HUMAN:	during the yellow thing 's second rotation , does the cylinder rotate more frequently than the aforementioned purple object flies ?
	Gold:	STAR=74, END=98, (OBJ20, COLOR, purple), (OBJ57, COLOR, red), (OBJ57, SHAPE, cube), (OBJ61, COLOR, yellow), (OBJ82, COLOR, green), (OBJ82, SHAPE, cylinder)
	RNN+Attn:	STAR=1, END=241, (OBJ61, COLOR, yellow), (OBJ82, COLOR, green), (OBJ82, SHAPE, cylinder), (OBJ91, COLOR, red)
	UniConv:	STAR=73, END=97, (OBJ57, COLOR, red), (OBJ57, MATERIAL, rubber), (OBJ82, COLOR, green), (OBJ91, COLOR, red), (OBJ152, COLOR, purple)
	VDTN:	STAR=73, END=97, (OBJ20, COLOR, purple), (OBJ57, COLOR, red), (OBJ57, SHAPE, cube), (OBJ61, COLOR, yellow), (OBJ82, COLOR, green)
#8	SYSTEM:	TRUE
	HUMAN:	during the aforementioned cube 's rotation , what types of actions does the cone left of the aforementioned yellow object undertake ?
	Gold:	STAR=143, END=163, (OBJ20, COLOR, purple), (OBJ57, COLOR, red), (OBJ57, SHAPE, cube), (OBJ61, COLOR, yellow), (OBJ72, SHAPE, cone), (OBJ82, COLOR, green), (OBJ82, SHAPE, cylinder)
	RNN+Attn:	STAR=1, END=193, (OBJ20, COLOR, purple), (OBJ20, SHAPE, cone), (OBJ57, COLOR, red), (OBJ61, COLOR, yellow), (OBJ82, COLOR, green), (OBJ82, SHAPE, cylinder), (OBJ91, COLOR, red)
	UniConv:	STAR=73, END=97, (OBJ57, MATERIAL, rubber), (OBJ72, SHAPE, cone), (OBJ82, COLOR, green), (OBJ82, SHAPE, cylinder), (OBJ91, COLOR, red), (OBJ152, COLOR, purple)
	VDTN:	STAR=73, END=97, (OBJ20, COLOR, purple), (OBJ20, SHAPE, cone), (OBJ57, COLOR, red), (OBJ57, SHAPE, cube), (OBJ61, COLOR, yellow), (OBJ82, COLOR, green)
#9	SYSTEM:	flying
	HUMAN:	throughout the whole video, is there anything else that performs the same set of activities as the earlier mentioned yellow thing ?
	Gold:	STAR=1, END=247, (OBJ20, COLOR, purple), (OBJ57, COLOR, red), (OBJ57, SHAPE, cube), (OBJ61, COLOR, yellow), (OBJ72, SHAPE, cone), (OBJ82, COLOR, green), (OBJ82, SHAPE, cylinder)
	RNN+Attn:	STAR=1, END=241, (OBJ20, COLOR, purple), (OBJ20, SHAPE, cone), (OBJ57, COLOR, red), (OBJ61, COLOR, yellow), (OBJ82, COLOR, green), (OBJ82, SHAPE, cylinder), (OBJ91, COLOR, red)
	UniConv:	STAR=1, END=253, (OBJ57, MATERIAL, rubber), (OBJ57, SHAPE, cube), (OBJ72, SHAPE, cone), (OBJ82, COLOR, green), (OBJ82, SHAPE, cylinder), (OBJ91, COLOR, red), (OBJ152, COLOR, purple)
	VDTN:	STAR=1, END=253, (OBJ20, COLOR, purple), (OBJ20, SHAPE, cone), (OBJ57, COLOR, red), (OBJ57, SHAPE, cube), (OBJ61, COLOR, yellow), (OBJ82, COLOR, green)

Table 15: Example predictions from a sample dialogue (CATER video ID #001441)



(a) Video #002660

(b) Video #001441

Figure 5: Example screenshots of CATER videos for dialogues in Table 14 (Video #002660) and 15 (Video #001441). We showed example bounding boxes and their object classes in each video.