MULTIMODAL DIALOGUE STATE TRACKING

Anonymous authors
Paper under double-blind review

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. Secondly, to facilitate research of this task, we developed DVD-DST, a synthetic video-grounded dialogue benchmark with annotations of multimodal dialogue states. Thirdly, we 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, in the dialogue in Figure 1, a dialogue agent is helping 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. A comparison of traditional DST and our proposed MM-DST can be seen in Figure 1.

Toward MM-DST, we developed a synthetic benchmark based on the CATER universe (Girdhar & Ramanan, 2020) with detailed annotations of dialogue states. In total, our benchmark contains more

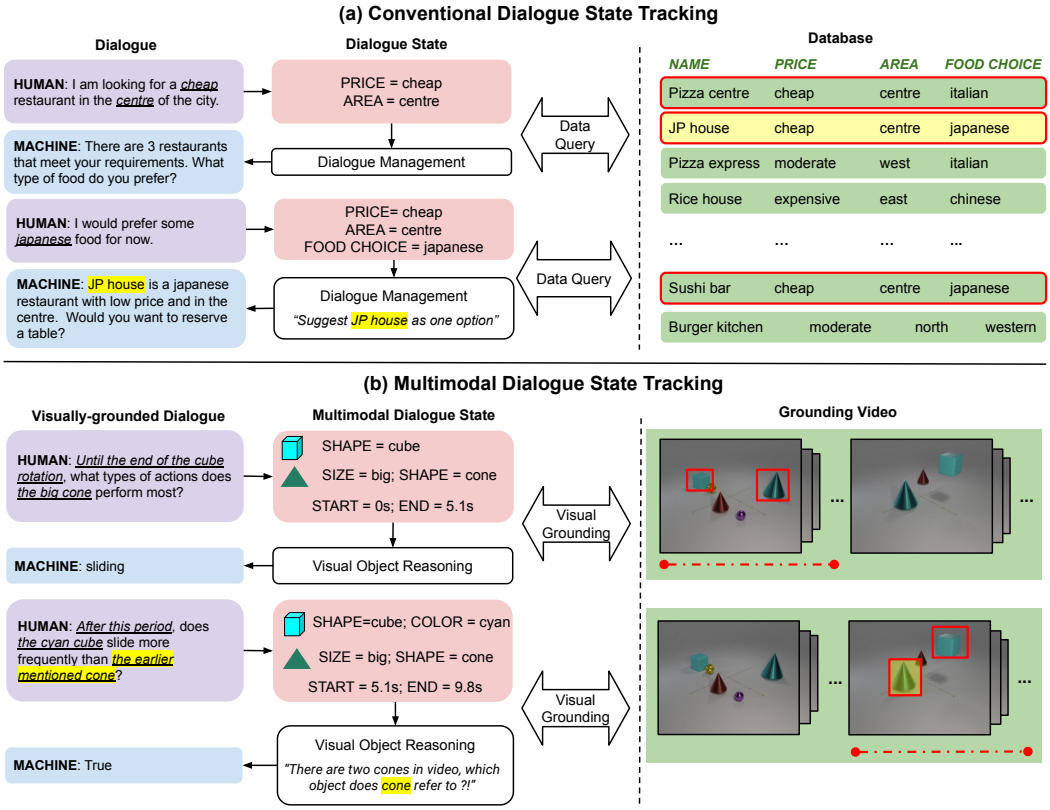


Figure 1: **Multimodal Dialogue State Tracking (MM-DST):** We proposed to extend the traditional DST from unimodality to multimodality. Compared to traditional DST (a), MM-DST (b) define dialogue states, consisting of slots and slots values for visual objects that are mentioned in dialogues.

than $13k$ dialogues, each of which contains 10 dialogue turns, resulting in $130k$ (human, system) utterance pairs and corresponding dialogue states.

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 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 Task

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 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 $\mathcal{B}_t = \{(s_i, v_i^t)\}|_{i=1}^{i=|\mathcal{S}|}$. An example of dialogue state ground-truth can be seen in Figure 1a. 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 objective is then defined as:

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

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 & 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 study shows that some object features are carried over across memory episodes as a mechanism to maintain stable representations of objects over time. The multimodal DST task is designed to develop a similar capacity in multimodal dialogue systems. Specifically, we require systems to maintain a recurring information state of multiple objects, including their own features, over a time period 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 closely related work in the dialogue domain is (Pang & 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 & 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 account for the perception of multiple objects (Bar, 2004; Xu & Chun, 2009). 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 & 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 objective of MM-DST is defined as:

$$\hat{\mathcal{B}}_t = \arg\max_{\mathcal{B}_t} P(\mathcal{B}_t|\mathcal{D}_t, \mathcal{V}, \theta) = \arg\max_{\mathcal{B}_t} \prod_{j}^{|\mathcal{O}|} \prod_{i}^{|\mathcal{S}|} P(v_{i,j}^t|s_i, o_j, \mathcal{D}_t, \mathcal{V}, \theta) P(o_j|\mathcal{V}, \omega) \tag{2}$$

One limitation of this representation is the assumption of a universal slot ontology, with a predefined set of slots. However, this limitation is not just limited to multimodal dialogues, but has been noted and addressed to some extent in unimodal DST (Rastogi et al., 2020). We assume a universal slot set in this work and will reserve future work to tackle this limitation.

Another 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

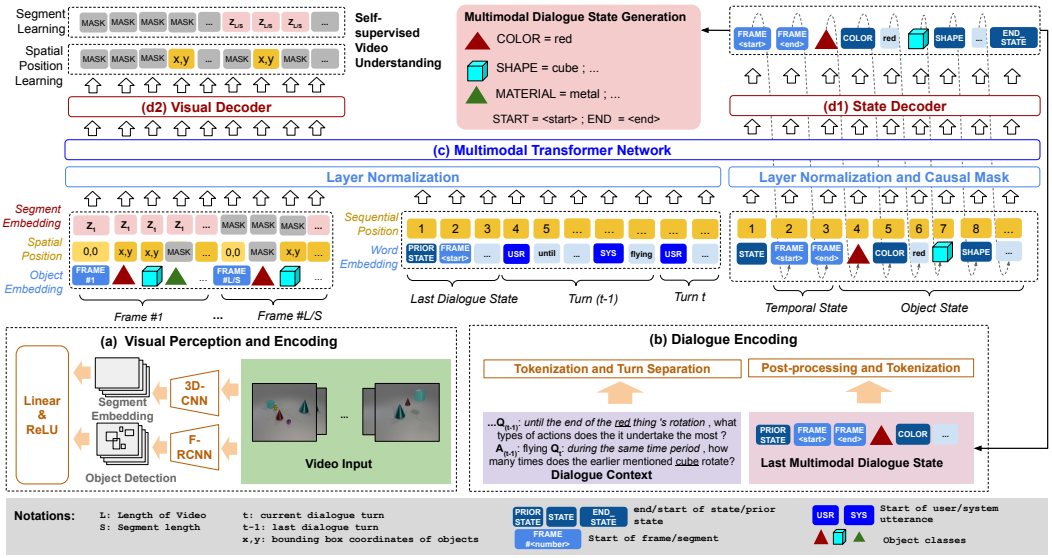


Figure 2: **Video-Dialogue Transformer Network(VDTN)** has 4 key components: (a) *Visual Perception and Encoder* (Section 3.1) (b) *Dialogue Encoder* (Section 3.2) (c) *Transformer Network* (Section 3.3) (d1) *State Decoder* (Section 3.4) and (d2) *Visual Decoder* (Section 3.4)

period. Similarly, in dialogue, we want the dialogue agent to achieve this capacity over the length of a conversation. 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 1b.

3 Video-Dialogue Transformer Network

Compared to traditional DST, MM-DST involves additional information from visual inputs. A naive adaptation of conventional DST to MM-DST is to directly combine visual features extracted by a pretrained visual model. Most often a 3D-CNN model can be used to extract sub-clips from videos and extracted feature vectors are concatenated to dialogue context representations. However, as shown in our experiments, this extension of conventional DST results in poor performance and does not address the challenge of visual object reasoning in visually-grounded dialogues. In this paper, we established a strong baseline for MM-DST and called this model Video-Dialogue Transformer Network (VDTN). VDTN is composed of 4 major components (refer to Figure 2 for an overview):

3.1 Visual Perception and Encoder

This module encodes videos at both frame-level and segment-level representations. Frame-level representations consist of visual-object embeddings and their spatial locations. In each frame, each visual object representation is combined with the representations of the video segment corresponding to that frame to facilitate both spatial and temporal perception.

Specifically, we used a Faster R-CNN model (Ren et al., 2015) finetuned on the CATER universe 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 ResNeXt model (Xie et al., 2017) finetuned on Kinetics dataset (Kay et al., 2017). We used this model 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.

Note that we do not finetuned 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 identities tokens of *OBJ<class>*) of all frames, 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 $X_{bb} \in \mathbb{R}^{L_{obj} \times 4}$. Similarly, we stacked each ResNeXt feature vector by $(N_{obj} + 1)$ for each segment, and obtained a sequence $X_{cnn} \in \mathbb{R}^{L_{obj} \times 2048}$.

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 token in X_{obj}, resulting in embedding features of dimensions d. A video input representation is the element-wise summation of all above vectors, resulting in a vector $Z_V \in \mathbb{R}^{L_{obj} \times d}$.

3.2 Dialogue Encoder

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 representation dimension d.

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 (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 is 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 Generation and Self-supervised Video Understanding

State Decoder. 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. The ground-truth dialogue states are flattened into sequences as described in Section 3.2. 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 that transforms representations to state vocabulary embedding space. The decoder applies the same procedure for the subsequent positions to decode dialogue states auto-regressively.

During training, we directly used the ground-truth dialogue states and applied a causal mask to simulate the auto-regressive process. We applied a softmax layer and optimize the model by minimizing the negative log-likelihood. Note that this decoder design partially avoids the assumption of conditionally independent slots from (1) and (2). 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:

$$P(\mathcal{B}_t|\mathcal{D}_t, \mathcal{V}) = \prod_{k}^{|\mathcal{B}_t|} P(b_{k,t}|b_{<k,t}, X_{ctx}, X_{obj}) \quad \mathcal{L}_{dst} = -\sum \log P(b_{k,t}|b_{<k,t}, X_{ctx}, X_{obj}) \quad (3)$$

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 Decoder. Finally, moving away from conventional unimodal DST, we proposed to enhance our DST model with a *Visual Decoder* that learns to recovers 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 used this masking strategy to train the models to learn the dynamics of an object based on its visual context. 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:

$$\mathcal{L}_{obj} = \sum_{j} \mathbf{1}_{\text{masked}} \times l(f_{bb}(Z'_{V,j}), X_{bb,j}), j \in [1, L_{obj}] \quad (4)$$

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:

$$\mathcal{L}_{seg} = \sum_{j} \mathbf{1}_{\text{masked}} \times l(f_{cnn}(Z'_{V,j}), X_{cnn,j}), j \in [1, L_{obj}] \quad (5)$$

4 EXPERIMENTS

DVD-DST Benchmark. To study the multimodal DST task as defined in this paper, there are not many available suitable benchmarks. As mentioned in Section 2, the closest possible studies to our task are (Pang & Wang, 2020; Kottur et al., 2021) but each contains its own shortfalls. In existing popular benchmarks of multimodal dialogues such as VisDial (Das et al., 2017a), 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), or annotation bias, in which the causal link connecting dialogue history and current turn question is actually harmful (Qi et al., 2020). These biases would obviate the need for a DST task.

We found that a recent benchmark called DVD (Le et al., 2021b) can address both biases. The dialogues are grounded on videos from CATER (Shamsian et al., 2020), which contain visually simple yet highly varied objects. The dialogues in DVD 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. Moreover, as shown by our experiments in Table 1, models with access to videos only do not perform as well as models with access to dialogue only. This indicates less distribution bias in dialogue context in DVD than other benchmarks and models can generalize better by relying on dialogue features.

However, we noted that DVD is designed for response prediction and only contains the dialogue states up to the second last turn in each dialogue. Therefore, we generated new dialogues following (Le et al., 2021b) but included state annotation at all dialogue turns. Moreover, we chose to generate dialogue data based on an extended CATER video split (Shamsian et al., 2020) rather than the original CATER video data (Girdhar & Ramanan, 2020). The extended CATER split (Shamsian et al., 2020) includes additional annotations of bounding box boundaries of each visual object in video frames. This annotation facilitates 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 elaborate further in Appendix B.

We also want to highlight that like other synthetic benchmarks such as CLEVR (Johnson 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.

Baselines. We benchmarked VDTN on DVD-DST with the following baseline models: (1) *Q-retrieval (tf-idf)*, for each test sample, directly retrieves the training sample the with most similar question utterance and use its state as the predicted state; (2) *State prior* selects the most common tuple of (object, slot, value) in training split and uses it as predicted states; (3) *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; (4) *Object (all)* is similar to (3) but selects all possible objects and all possible (slot, value) tuples as the predicted state; (5) *RNN(+Attn)* uses RNN as encoder and an MLP as decoder with a vanilla dot-product attention; We experimented with strong unimodal DST baselines, including: (6) *TRADE* (Wu et al., 2019); (7) *UniConv* (Le et al., 2020b); and (8) *NADST* (Le et al., 2020c). We implemented baselines (5) to (8) and tested them on dialogues with or without videos. When video inputs are applied, we embedded both object and segment-level features using the same method as described in Section 3.1. The embedded features are integrated into baselines in the same techniques in which the original models treat dialogue representations. Refer to Appendix C for our training details.

Evaluation. We followed the evaluation metrics from unimodal DST benchmarks (Budzianowski et al., 2018; Henderson et al., 2014a). In these benchmarks, a joint state accuracy compared the predicted state and ground-truth state per dialogue turn. 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 these 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 Obj State Acc") as well as all slot values ("Joint State 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.

Overall results. From Table 1, we have the following observations: (1) 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. (2) 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 predict dialogue states well. (3) 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 (2). (4) 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. (5) 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. (6) VDTN achieves significant performance gains and achieves the SOTA results in all component or joint prediction metrics.

(7) We also experimented with a version of VDTN in which the transformer network (Section 3.3) was initialized from a GPT2-base model (Radford et al., 2019) with a pretrained checkpoint released by HuggingFace[1]. Asides 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 $36\times$ bigger than

[1] `https://huggingface.co/gpt2`

Table 1: Overall performance of all models on the test split of DVD-DST

Model	Dial	Video	Obj Identity F1	Obj Slot F1	Obj State F1	Joint Obj State Acc	Joint State IoU@0.5	Joint State IoU@0.7
Q-retrieval (tf-idf)	Q only	-	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 (random)	-	Objs	19.8%	14.1%	0.4%	0.0%	0.0%	0.0%
Object (all)	-	Objs	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)+Attn	✓	✓	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 (ours)	✓	✓	**84.5%**	**72.8%**	**60.4%**	**28.0%**	**15.3%**	**13.1%**
VDTN+GPT2(ours)	✓	✓	**85.2%**	**76.4%**	**63.7%**	**30.4%**	**16.8%**	**14.3%**

Table 2: Ablation results by joint state predictions, using greedy or beam search decoding styles

Video Features	Dialogue State	Video loss	Greedy			Beam Search		
			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
X_{bb}	$\mathcal{B}\backslash\text{time}$	-	17.3%	N/A	N/A	17.9%	N/A	N/A
$X_{bb} + X_{cnn}$	$\mathcal{B}\backslash\text{time}$	-	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%**

our default VDTN model. As shown in Table 1, 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 adaptation of large pretrained LMs, such as (Brown et al., 2020; Raffel et al., 2020), or of pretrained multimodal transformer models, such as (Lu et al., 2019; Zhou et al., 2020), in the MM-DST task. (8) Finally, while we noted that using synthetic benchmarks such as DVD-DST might result in over-estimated performance of models, we argue that the current reported results of state accuracy are reasonable due to the strict measurement of this metric in the MM-DST task. We noted that this metric basically treat the MM-DST task as a classification task. In DVD, there are roughly 7200 classes, each of which is a distinct set of objects, each with many possible slot combinations. Combined with the upstream error from object perception, we expect the current results are justifiable (See refer to Appendix E for more discussion).

Ablation analysis. Table 2 shows the results of different variants of VDTN models. We observed that: (1) 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). (2) 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. (3) The results of using self-supervised losses displayed the benefits of enhancing models with better video representations. We observed that segment-based learning is marginally more powerful than object-based learning.

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 assumed bounding box annotations during training, and simply passed unmasked visual features to VDTN for an object tracking task. 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 significantly, as compared to the self-supervised learning objective $\mathcal{L}_{obj}$. This indicates that our self-supervised learning tasks do not heavily rely on object boundary annotations. Finally, we found combining both segment and object-level self-supervision is not useful. This is possible 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.

Table 3: Results by self-supervised objectives

Video self-supervision	Loss	Joint Obj State Acc	Joint State IoU@0.5	Joint State 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%

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 decode dialogue states across all samples in DVD-DST. We then used the predicted slots, including object identities and temporal slots, to filter the video inputs by objects and segments. We then used these filtered videos as input to train new VDTN models with an MLP as the response prediction layer. Note that these models are not trained with $\mathcal{L}_{dst}$ or self-supervised objectives, but only with a cross-entropy loss to predict answer candidates. From Table 4, we observed the benefits of visual inputs filtered by states, resulting in accuracy improvement of up to 5.9% accuracy score. Note that there are more sophisticated approaches such as neural module networks (Andreas et al., 2016; Hu et al., 2018) and symbolic reasoning (Yi et al., 2018; Chen et al., 2020) to fully exploit the decoded dialogue states. We leave this extension for future research.

Table 4: Results of response predictions (by greedy/beam search states):

Dialogue State	Accuracy
No state	43.0%
$\mathcal{B}\backslash\text{time}$	46.8%/47.1%
$\mathcal{B}$	48.7%/48.9%

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

5 Discussion and Conclusion

Related Work. Our work is related to the research of unimodal DST and visually-grounded dialogues. We show that the scope of DST (Young et al., 2010; Mrkšić et al., 2017; Lei et al., 2018b; Gao et al., 2019; Le et al., 2020c), can be further extended to a multimodal world. Within the research of visually-grounded dialogues, our work is related to (De Vries et al., 2017; Das et al., 2017a; Chattopadhyay et al., 2017; Hori et al., 2019; Thomason et al., 2019). However, these approaches are not designed to track objects across dialogue turns, and they do not maintain a recurring memory/state of these objects and their features throughout dialogues. Most of the prior approaches introduced techniques inspired by vision-language tasks such as VQA (Rohrbach et al., 2015; Antol et al., 2015; Jang et al., 2017; Lei et al., 2018a). Our work, instead, was inspired from a dialogue-based angle, with a new learning task for multimodal DST. For more detailed related work, please refer to Appendix A.

Limitations and Conclusion. 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, including a potential extension of MM-DST to videos of humans, please refer to Appendix E. In summary, in this work, we formally define a novel multimodal DST task to test models ability to track visual objects and their attributes in dialogues. For this task, we introduced a new benchmark, and proposed VDTN as a strong baseline with a video self-supervised learning strategy. Our experiments indicate the multimodal reasoning capacities of VDTN and the potentials of MM-DST in a dialogue system.

6 ETHICS STATEMENT

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 from turn to turn.

7 REPRODUCIBILITY STATEMENT

In this paper, we reported the full technical details of VDTN in Section 3, the dataset and evaluation details in Section 4. Due to the page limit of the conference, we included more data preprocessing and training details in Appendix B and C. To improve the reproducibility of this work, we will fully release the dataset MM-DST as well as the complete implementation of VDTN. We will also release the pretrained models of VDTN to replicate the experiment results reported in this paper. 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.

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. Audio-visual scene-aware dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural module networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 39–48, 2016.

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

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

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

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

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 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*, pp. 5016–5026, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1547. URL https://aclanthology.org/D18-1547.

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

Xinyun Chen, Chen Liang, Adams Wei Yu, Denny Zhou, Dawn Song, and Quoc V. Le. Neural symbolic reader: Scalable integration of distributed and symbolic representations for reading comprehension. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=ryxjnREFwH.

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

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

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

Mihail Eric, Lakshmi Krishnan, Francois Charette, and Christopher D. Manning. Key-value retrieval networks for task-oriented dialogue. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pp. 37–49, Saarbrücken, Germany, August 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-5506. URL https://aclanthology.org/W17-5506.

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

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

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

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

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

Rohit Girdhar and Deva Ramanan. Cater: A diagnostic dataset for compositional actions and temporal reasoning. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=HJgzt2VKPB.

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

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

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

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

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. 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)*, pp. 2352–2356, May 2019. doi: 10.1109/ICASSP.2019.8682583.

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

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

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

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2901–2910, 2017.

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

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

Hyounghun Kim, Hao Tan, and Mohit Bansal. Modality-balanced models for visual dialogue. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8091–8098, Apr. 2020. doi: 10.1609/aaai.v34i05.6320. URL https://ojs.aaai.org/index.php/AAAI/article/view/6320.

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

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

Satwik Kottur, Seungwhan Moon, Alborz Geramifard, and Babak Damavandi. 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*, pp. 4903–4912, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.emnlp-main.401.

Gakuto Kurata, Bing Xiang, Bowen Zhou, and Mo Yu. Leveraging sentence-level information with encoder LSTM for semantic slot filling. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2077–2083, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1223. URL https://www.aclweb.org/anthology/D16-1223.

Hung Le, Doyen Sahoo, Nancy Chen, and Steven Hoi. Multimodal transformer networks for end-to-end video-grounded dialogue systems. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 5612–5623, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1564. URL https://www.aclweb.org/anthology/P19-1564.

Hung Le, Doyen Sahoo, Nancy Chen, and Steven C.H. Hoi. BiST: Bi-directional spatio-temporal reasoning for video-grounded dialogues. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1846–1859, Online, November 2020a. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.145. URL https://www.aclweb.org/anthology/2020.emnlp-main.145.

Hung Le, Doyen Sahoo, Chenghao Liu, Nancy Chen, and Steven C.H. Hoi. 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)*, pp. 1860–1877, Online, November 2020b. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.146. URL https://aclanthology.org/2020.emnlp-main.146.

Hung Le, Richard Socher, and Steven C.H. Hoi. Non-autoregressive dialog state tracking. In *International Conference on Learning Representations*, 2020c. URL https://openreview.net/forum?id=H1e_cC4twS.

Hung Le, Nancy F. Chen, and Steven Hoi. Learning reasoning paths over semantic graphs for video-grounded dialogues. In *International Conference on Learning Representations*, 2021a. URL https://openreview.net/forum?id=hPWj1qduVw8.

Hung Le, Chinnadhurai Sankar, Seungwhan Moon, Ahmad Beirami, Alborz Geramifard, and Satwik Kottur. 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)*, pp. 5651–5665, Online, August 2021b. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.439. URL https://aclanthology.org/2021.acl-long.439.

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

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. TVQA: Localized, compositional video question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 1369–1379, Brussels, Belgium, October-November 2018a. Association for Computational Linguistics. doi: 10.18653/v1/D18-1167. URL https://aclanthology.org/D18-1167.

Wenqiang Lei, Xisen Jin, Min-Yen Kan, Zhaochun Ren, Xiangnan He, and Dawei Yin. 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)*, pp. 1437–1447, 2018b.

Zekang Li, Zongjia Li, Jinchao Zhang, Yang Feng, and Jie Zhou. Bridging text and video: A universal multimodal transformer for video-audio scene-aware dialog. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 1–1, 2021. doi: 10.1109/TASLP.2021.3065823.

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

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/c74d97b01eae257e44aa9d5bade97baf-Paper.pdf.

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

Nikola Mrkšić, Diarmuid Ó Séaghdha, Tsung-Hsien Wen, Blaise Thomson, and Steve Young. 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)*, pp. 1777–1788. Association for Computational Linguistics, 2017. doi: 10.18653/v1/P17-1163. URL http://www.aclweb.org/anthology/P17-1163.

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

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

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

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

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

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

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

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8689–8696, Apr. 2020. doi: 10.1609/aaai.v34i05.6394. URL https://ojs.aaai.org/index.php/AAAI/article/view/6394.

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

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

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

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

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

Yangyang Shi, Kaisheng Yao, Hu Chen, Dong Yu, Yi-Cheng Pan, and Mei-Yuh Hwang. 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*, pp. 393–399, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1044. URL https://www.aclweb.org/anthology/N16-1044.

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

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 30*, pp. 5998–6008. Curran Associates, Inc., 2017. URL http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf.

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

Tsung-Hsien Wen, David Vandyke, Nikola Mrkšić, Milica Gašić, Lina M. Rojas-Barahona, Pei-Hao Su, Stefan Ultes, and Steve Young. 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*, pp. 438–449, Valencia, Spain, April 2017. Association for Computational Linguistics. URL https://aclanthology.org/E17-1042.

Chien-Sheng Wu, Andrea Madotto, Ehsan Hosseini-Asl, Caiming Xiong, Richard Socher, and Pascale Fung. Transferable multi-domain state generator for task-oriented dialogue systems. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 808–819, Florence, Italy, July 2019. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/P19-1078.

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

Puyang Xu and Qi Hu. 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)*, pp. 1448–1457. Association for Computational Linguistics, 2018. URL http://aclweb.org/anthology/P18-1134.

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

Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Josh Tenenbaum. Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. In *Advances in Neural Information Processing Systems*, pp. 1031–1042, 2018.

Steve Young, Milica Gašić, Simon Keizer, François Mairesse, Jost Schatzmann, Blaise Thomson, and Kai Yu. The hidden information state model: A practical framework for pomdp-based spoken dialogue management. *Computer Speech and Language*, 24(2):150–174, 2010. ISSN 0885-2308. doi: https://doi.org/10.1016/j.csl.2009.04.001. URL https://www.sciencedirect.com/science/article/pii/S0885230809000230.

Jianguo Zhang, Kazuma Hashimoto, Chien-Sheng Wu, Yao Wang, Philip Yu, Richard Socher, and Caiming Xiong. 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*, pp. 154–167, Barcelona, Spain (Online), December 2020. Association for Computational Linguistics. URL https://aclanthology.org/2020.starsem-1.17.

Victor Zhong, Caiming Xiong, and Richard Socher. 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)*, pp. 1458–1467, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1135. URL https://www.aclweb.org/anthology/P18-1135.

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

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

A Details of Related Work

Our work is related to two domains: dialogue state tracking and visually-grounded dialogues.

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 & 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 & Wang, 2020), we noted that some form of object

tracking is introduced throughout dialogue turns. The tracking module is used to decide which object 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. Moreover, the scope of our work is towards visually-grounded dialogues rather than in the task-oriented paradigm.

B DVD-DST Dataset Details

Table 5: Dataset summary: statistics of related benchmarks are from (Budzianowski et al., 2018)

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

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

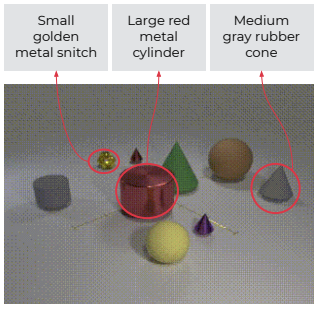


Figure 3: Synthetic visual objects in the CATER universe

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 5. 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 3 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 12 and Table 13.

C Training Details

We trained VDTN by jointly minimizing $\mathcal{L}_{dst}$ and $\mathcal{L}_{bb/cnn}$. 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 $\mathcal{L}_{bb/cnn}$. We trained all models using the Adam optimizer (Kingma & 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. All model parameters, except pretrained visual perception models, are initialized by a uniform distribution (Glorot & Bengio, 2010). 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.

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.

D Additional Results

Ablation analysis by component predictions. From Table 6, 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 "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 6: Ablation results by component predictions of object identities, slots, and object states

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}	$\mathcal{B}\backslash\text{time}$	-	79.4%	64.2%	48.5%	55.9%	76.6%	41.4%	63.5%
$X_{bb} + X_{cnn}$	$\mathcal{B}\backslash\text{time}$	-	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 7 and 8 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.

Ablation analysis by turn positions. Table 9 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. Figure 4 shows that state accuracy reduces more dramatically (as shown by

Table 7: Ablation results by individual object identity/slot/state

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}	$\mathcal{B}\backslash\text{time}$	-	77.2%	81.8%	65.0%	63.4%	47.1%	50.0%
$X_{bb} + X_{cnn}$	$\mathcal{B}\backslash\text{time}$	-	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 8: Ablation results by individual slot type

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}	$\mathcal{B}\backslash\text{time}$	-	60.1%	52.2%	76.8%	76.4%	43.2%	39.7%	61.4%	65.6%
$X_{bb} + X_{cnn}$	$\mathcal{B}\backslash\text{time}$	-	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%**

"Joint Obj State 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.

Table 9: Ablation results by dialogue turn positions

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%

Impacts of dialogue context encoder. In Table 10a, 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 10b 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.

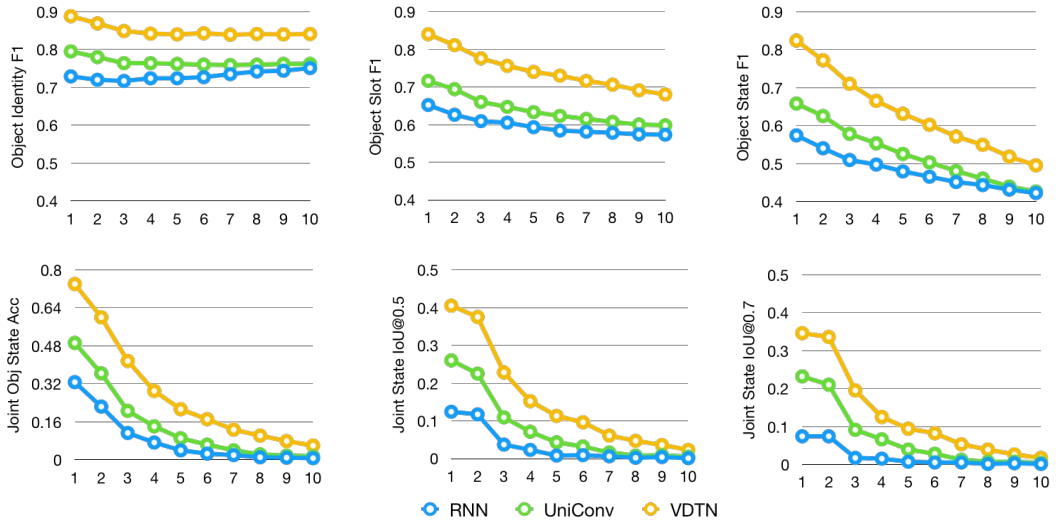


Figure 4: Ablation results of VDTN and baselines by dialogue turn positions (x axis)

Table 10: Ablation results by encoding strategies: All models are trained only with $\mathcal{L}_{dst}$.

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

$\mathcal{B}_{t-1}$	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%

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

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%

Impacts of object-level representation. Table 11 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 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.

Table 11: Results with and without object representations

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%

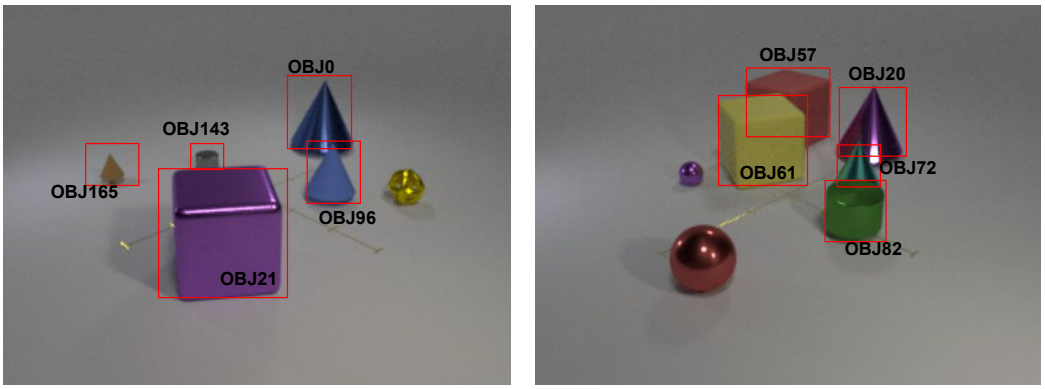
Qualitative analysis. Table 12 and 13 display 2 sample dialogues and state predictions. We displayed the corresponding video screenshots for these dialogues in Figure 5. To cross-reference between videos and dialogues, we displayed the bounding boxes and their object classes in video screenshots. These 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.

Table 12: Example predictions from a sample dialogue (CATER video ID #002660)

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

#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)



(a) Video #002660 (b) Video #001441

Figure 5: Example screenshots of CATER videos for dialogues in Table 12 (Video #002660) and 13 (Video #001441). We showed example bounding boxes and their object classes in each video.

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 3 (+ '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 1, 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).

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.