𝕆𝕃𝕍𝕚𝕋: Multi-Modal State Tracking via Attention-Based Embeddings for Video-Grounded Dialog

Adnen Abdessaied, Manuel von Hochmeister*, Andreas Bulling
University of Stuttgart, Bosch, University of Stuttgart
Germany, Germany, Germany
{adnen.abdessaied, andreas.bulling}@vis.uni-stuttgart.de
manuel.vonhochmeister@de.bosch.com

Abstract

We present the 𝕆bject 𝕃anguage 𝕍ideo 𝕋ransformer (𝕆𝕃𝕍𝕚𝕋) – a novel model for video dialog operating over a multi-modal attention-based dialog state tracker. Existing video dialog models struggle with questions requiring both spatial and temporal localization within videos, long-term temporal reasoning, and accurate object tracking across multiple dialog turns. 𝕆𝕃𝕍𝕚𝕋 addresses these challenges by maintaining a global dialog state based on the output of an Object State Tracker (OST) and a Language State Tracker (LST): while the OST attends to the most important objects within the video, the LST keeps track of the most important linguistic co-references to previous dialog turns. In stark contrast to previous works, our approach is generic by nature and is therefore capable of learning continuous multi-modal dialog state representations of the most relevant objects and rounds. As a result, they can be seamlessly integrated into Large Language Models (LLMs) and offer high flexibility in dealing with different datasets and tasks. Evaluations on the challenging DVD (response classification) and SIMMC 2.1 (response generation) datasets show that 𝕆𝕃𝕍𝕚𝕋 achieves new state-of-the-art performance across both datasets.

Keywords: Multi-Modal Learning, Video Dialog, Dialog State Tracking

1. Introduction

The potential of deep learning for tackling unique challenges at the intersection of computer vision (CV) and natural language processing (NLP) has been demonstrated for a wide range of tasks (Karpathy and Fei-Fei, 2015; Antol et al., 2015; Das et al., 2017; Vo et al., 2019; Rombach et al., 2022). Among these tasks, video dialog is considered to be one of the most challenging. In contrast to visual (Antol et al., 2015) and video (Xu et al., 2016) question answering, which only require reasoning about a single question, video dialog models have to reason over the whole dialog history. Furthermore, in contrast to visual dialog (Das et al., 2017; Abdessaied et al., 2024), video dialog involves reasoning over a dynamic visual input (video) instead of a static image.

While recent video dialog models have improved performance (Hori et al., 2019; Le et al., 2019, 2021a), these gains have largely been marginal, most likely due to the significant challenges posed by this novel task. Current models suffer from several specific limitations: They struggle with questions that require spatial and temporal localization within the video, they suffer from a general inability of long-term reasoning, and they fail to accurately track objects across multiple dialog turns. Moreover, they have only been evaluated on benchmarks that were not explicitly designed to minimize biases and test for higher-order reasoning capabilities (Hori et al., 2019; Zhao et al., 2018). To address these limitations, we propose 𝕆𝕃𝕍𝕚𝕋– the Object Language Video Transformer for video-grounded dialog. At the core of 𝕆𝕃𝕍𝕚𝕋 are two novel components: an object state tracker (OST) and a language state tracker (LST). While the former actively attends to the most relevant objects across different video frames in order to answer the question at hand, the latter encodes the most important linguistic features for a more efficient co-reference resolution (object mentions, attributes, spatial and temporal cues, etc.) across dialog turns. After each turn, both trackers compute con-

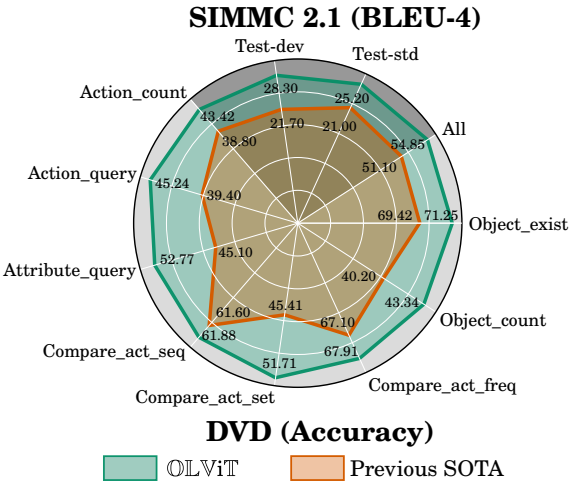

Figure 1: 𝕆𝕃𝕍𝕚𝕋 outperforms strong baselines and achieves new state-of-the-art results on DVD (classification) and SIMMC 2.1 (generation).

* Work conducted while at the University of Stuttgart
Our project web-page is accessible here

tinuous object and language state vectors that are used to update a global dialog state maintained over the course of the entire dialog. We study different ways of integrating the state vectors in an end-to-end manner within LLMs and demonstrate the applicability of our approach in both a discriminative setting, where the model predicts a full answer, and a generative one where response tokens are predicted one after another. In summary, the contributions of our work are three-fold:

• We introduce the Object Language Video Transformer ($\mathbb{OLViT}$) – a novel video dialog model that alleviates key limitations of current methods, specifically joint spatial and temporal localization in videos, long-term reasoning, and accurate object tracking across dialog turns.

• As a key novelty of our approach, we propose two attention-based video dialog state trackers that track the most relevant objects in the visual scene and the most important linguistic coreferences to previous dialog turns as well as multiple integration schemes of these states into an end-to-end trainable model.

• As can be seen in Figure 1, empirical results show the applicability of our model not only in a discriminative setting but also in a generative one as well. Specifically, $\mathbb{OLViT}$ achieves a *test* accuracy of **54**.**85**% on DVD – outperforming the current state of the art by a **3**.**75**% margin. In addition, $\mathbb{OLViT}$ outperforms strong baselines on the SIMMC 2.1 generative task by reaching BLEU-4 scores of **28**.**30** and **25**.**20** on the *test-dev* and *test-std* splits, respectively.

2. Related Work

Video Dialog. Video dialog has recently emerged as a particularly challenging task at the intersection of vision and language. In early work, Zhao et al. proposed to follow a hierarchical attention context learning approach and used a multi-stream network for joint video representation. Subsequently, Hori et al. showed that incorporating audio into an end-to-end video dialog attention-based system could improve performance. More recently, Jin et al. suggested progressively updating query information based on dialog history and video content while Le et al. proposed the multi-modal transformer network (MTN) – an approach that attended to inputs from different embedding spaces and fused the multi-modal information into newly combined representations. Inspired by the success of graph neural networks, Le et al. proposed a framework for discovering information flows among dialog turns through a semantic graph based on lexical components in each question and answer. However, most of these models have only been trained and assessed on datasets that were not specifically created to test for higher-order reasoning capabilities (Hori et al., 2019; Zhao et al., 2018). It is therefore unclear to what extent these models exploit the biases present in these datasets.

Our work differs in two distinctive ways: (1) We actively track the most relevant objects and textual facts via two separate multi-modal dialog state trackers that can be seamlessly integrated into pre-trained LLMs. (2) We specifically focus on datasets with minimal bias (i.e. DVD and SIMMC 2.1) that contain sufficient information to confidently assess a model's reasoning capabilities. In stark contrast, previous models mainly focused on the AVSD dataset (Hori et al., 2019) that does not provide detailed annotations for the different types of reasoning over the spatio-temporal space of video, and thus is not suitable to test for models' higher order reasoning capabilities.

Dialog State Tracking. Previous works (Noroozi et al., 2020) have formulated the problem of dialog state tracking as a slot filling task (Xu and Hu, 2018). Mrkšić et al. proposed the Neural Belief Tracker (NBF) to detect slot-value pairs representing the user's goal by iterating over all possible pairs. Guo et al. and Wu et al. suggested creating a state based on the textual and visual data for visual dialog. They proposed a dialog-based image retrieval model that iteratively interacted with a user in natural language. Pang and Wang proposed an attention-based tracking of visual features to better generate questions for the "GuessWhat?!" dataset (Strub et al., 2017). Most similar to our work is (Le et al., 2022) where the slot filling dialog state paradigm (Xu and Hu, 2018) was simply extended to autoregressively predict the visual attributes of objects in plain text. However, this work still suffers from further limitations. First, it was specifically designed for the DVD dataset. Therefore, it cannot generalize to other datasets and tasks, thus, significantly limiting its applicability. Second, its training paradigm heavily relies on extra supervision labels such as bounding box coordinates extracted from a fine-tuned Mask-RCNN (He et al., 2017) model. Finally, its state tracking approach did not lead to significant performance improvements on the down-stream video dialog task. We instead propose a generic tracking approach that separately tracks the most relevant visual objects and linguistic facts in order to learn continuous state representations that can be seamlessly deployed in conjunction with current pre-trained LLMs.

Multi-modal Reasoning. Previous works have investigated whether models are capable of reasoning by introducing fully-controllable and bias-free datasets in visual question answering (Johnson et al., 2017), video question answering (Yi et al., 2020), visual dialog (Kottur et al., 2019), and video dialog (Le et al., 2021b). Neuro-symbolic models (Yi et al., 2018; Mao et al., 2019; Shi et al., 2019; Han et al., 2019; Yi et al., 2020; Abdessaied et al., 2022) have achieved strong performance on these datasets and have often outperformed fully-connectionist approaches. However, several recent studies have suggested that transformers (Vaswani et al., 2017) can deal equally well with the challenges posed by such tasks that have traditionally been tackled using symbolic approaches (Lample and Charton, 2020; Brown et al., 2020; Hahn et al., 2021). Most similar to our work is the *Aloe* model (Ding et al., 2021) that has been proposed for video question answering and, thus, is not capable of performing multi-turn reasoning within a conversational framework. Our model differs from *Aloe* in that it uses a novel multi-modal two-stream state tracker specifically geared towards video dialog. The state trackers allow our model to jointly attend to the most relevant objects and previous dialog turns for a more efficient co-reference resolution.

3. Multi-modal Attention-based Dialog State Tracking

Our $\mathbb{OLVIT}$ model consists of six main components as shown in Figure 2: An *Object Encoder* uses the unsupervised multi-object network (MONet) (Burgess et al., 2019) to decompose the video frames into multiple masks that are then used to compute the corresponding scene object embeddings. Complementing the object encoder, a *Text Encoder* uses a DistilRoBERTa model (Sanh et al., 2019) to generate a textual embedding of the current question. The *Object and Language State Trackers* take the object and language state vectors $s_o^{(i-1)}$ and $s_l^{(i-1)}$ from the previous dialog turn $(i-1)$ as input and generate updated state vectors $s_o^{(i)}$ and $s_l^{(i)}$. A *Combiner* merges both state vectors with the special `[CLS]` token and the previous object and text embeddings. Finally, a transformer *Encoder* block applies a sequence of self-attention and normalization operations on top of the combiner's outputs and uses the `[CLS]` token to predict the correct answer in the discriminative setting. For the generative task, a *Decoder* block is added on top of the encoder and is used to predict the answer tokens auto-regressively.

3.1. Object Encoder

The encoder first samples T frames equidistantly from the video and uses a frozen MONet to obtain N_o object segmentation masks per frame. Some sample video frames and segmentation masks are illustrated in Figure 2. The masks can be thought of as probabilities of each pixel belonging to a particular object and are encoded in latent variables with mean $\mu_t^n \in \mathbb{R}^{d_{obj}}$, where n indexes the object and t the video frame in which that object occurred. To obtain object embeddings $H_0^{obj} \in \mathbb{R}^{(TN_o)\times d}$, $\{\mu_t^n\}$ are mapped to the same dimension d as the transformer block using a linear layer. Position embeddings are added to learn object-frame relationships:

$$H_0^{obj} = \left[W^{obj}\mu_1^1, .., W^{obj}\mu_1^{N_o}, .., W^{obj}\mu_T^{N_o}\right] + O_{pos}, \tag{1}$$

where $W^{obj} \in \mathbb{R}^{d\times d_{obj}}$ and $O_{pos} \in \mathbb{R}^{(TN_o)\times d}$ are a learnable parameter and the position embedding.

3.2. Text Encoder

The current question at turn i is first tokenized and the resulting N_w sub-word tokens are processed by a frozen pre-trained DistilRoBERTa model (Sanh et al., 2019). We opted for this approach to improve training efficiency: We found that a full RoBERTa model (Liu et al., 2019) only marginally improved performance (see Section 5.3). The d_w-dimensional embeddings from the last DistilRoBERTa layer are again mapped to match the input dimension d of the subsequent transformer layers and used as the final word embeddings $H_0^w \in \mathbb{R}^{N_w\times d}$, that is:

$$H_0^w = [W^w h_1, ..., W^w h_{N_w}] + W_{pos}, \tag{2}$$

where $W^w \in \mathbb{R}^{d\times d_w}$ is a learnable parameter, h_j is the embedding of the j-th token and $W_{pos} \in \mathbb{R}^{N_w\times d}$ is a learnable positional embedding.

3.3. Dialog State Tracking

The dialog state tracker is a key novelty of our method. Its purpose is to track relevant objects within the video over the course of the dialog and to remember co-references to previous dialog turns. These goals are tackled by two separate sub-components: an object (OST) and a language (LST) state tracker.

Object State Tracker. As shown in the bottom left corner of Figure 2, the OST takes the object state vector $s_o^{(i-1)}$ and the k most important objects from the previous turn as input. It consists of

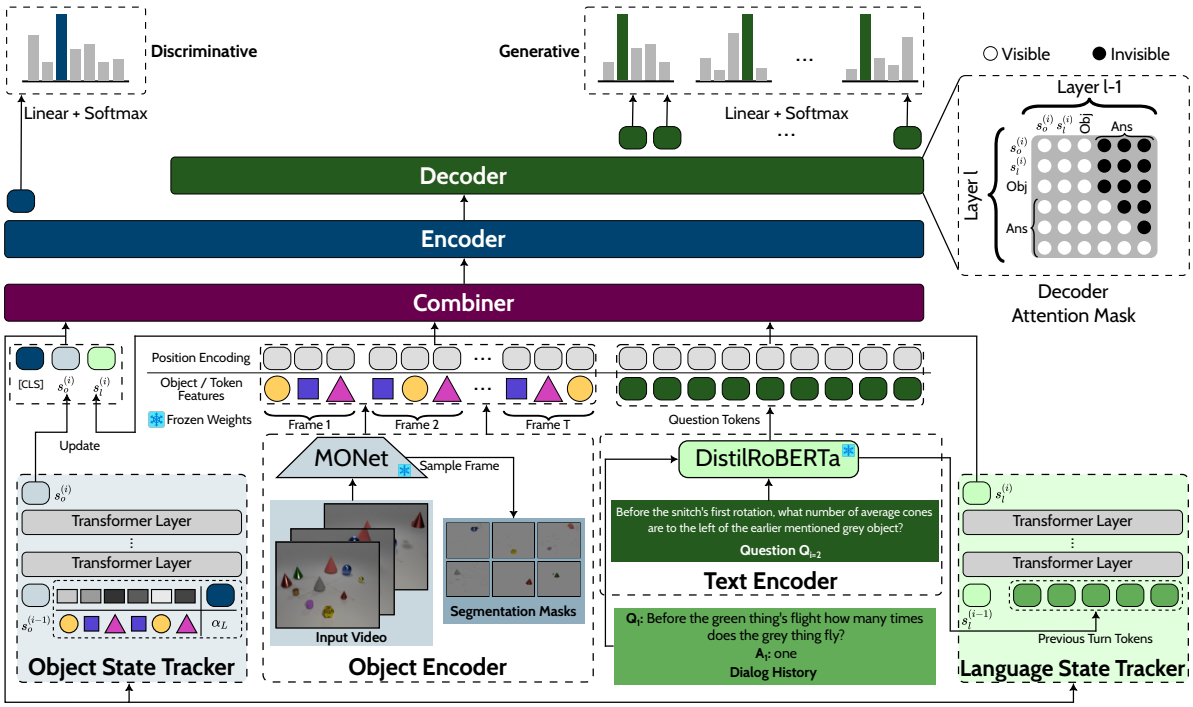

Figure 2: Architecture overview of our $\mathbb{OLVIT}$ model. It uses MONet and DistilRoBERTa-Base to generate the object embeddings for each frame and the text embeddings, respectively. Then, we add position encoding and append the special `[CLS]` token. Finally, we combine the object and language state vectors of the current $i-$th turn ($s_o^{(i)}$ and $s_l^{(i)}$) with the rest of the input, which will be processed by the subsequent transformer layers. In the generative setting, a decoder block is added to predict the answer token auto-regressively.

L_{ost} transformer layers and uses multi-head self-attention (MSA) with layer normalization (LN) to output an updated $s_o^{(i)}$:

$$H_0^{ost} = \left[s_o^{(i-1)}, h_{L,1}^{obj}, ..., h_{L,k}^{obj}\right], \quad (3)$$

$$h_{L,j}^{obj} = H_L^{obj}[j,:] \; \forall j \in top_k(\alpha_L(h_L^{[\text{CLS}]}, H_L^{obj})), \quad (4)$$

$$H_l^{ost} = \text{MSA}(\text{LN}(H_{l-1}^{ost})) + H_{l-1}^{ost}, \quad (5)$$

$$s_o^{(i)} = H_{L_{ost}}^{ost}[0,:], \quad (6)$$

where L is the number of subsequent transformer layers and $\alpha_L(h_L^{[\text{CLS}]}, H_L^{obj}) \in \mathbb{R}^{TN_o}$ is the attention values between the final embedding of the `[CLS]` token and the object embeddings. As such, the updated object state vector $s_o^{(i)}$ holds the most relevant objects for the current turn i. For the first turn, the object state vector is initialized with a zero vector, i.e. $s_o^{(0)} = 0$.

For the generative setting where we did not use the `[CLS]` token in our architecture, we summed the attention values of the object embeddings over all textual tokens in order to determine the most important objects for the given question, i.e.

$$h_{L,j}^{obj} = H_L^{obj}[j,:] \; \forall j \in top_k\left(\alpha_L(h_L^{txt}, H_L^{obj})\right), \quad (7)$$

where $h_L^{txt} = \sum_{k=1}^{N_w} h_k^w$ and $\{h_k^w\}$ are the question token embeddings of the last decoder layer.

Language State Tracker. As shown in the bottom right corner of Figure 2, the LST is a transformer-based module with L_{lst} layers that stores important information previously mentioned in the dialog history which is necessary to resolve potential co-references to previous dialog turns. The input of its first layer is the previous language state vector $s_l^{(i-1)}$ and turn embeddings. To calculate the updated language state vector $s_l^{(i)}$ for the current turn i, the LST has to be executed on the dialog history comprising of the previous $(i-1)$ turn embeddings. Each turn is composed of a question-answer pair and its embeddings are generated using the same frozen DistilRoBERTa model as before, that is:

$$H_0^{lst} = \left[s_l^{(i-1)}, W_{lst}^w h_1^w, .., W_{lst}^w h_{n_{i-1}}^w\right], \quad (8)$$

$$H_l^{lst} = \text{MSA}(\text{LN}(H_{l-1}^{lst})) + H_{l-1}^{lst} \quad (9)$$

$$s_l^{(i)} = H_{L_{lst}}^{lst}[0,:], \quad (10)$$

where W_{lst}^w is a learnable parameter, h_j^w is the j-th token embedding and n_i is the length of i-th dialog turn. For the first turn, the language state vector is initialized with a zero vector, i.e. $s_l^{(0)} = 0$.

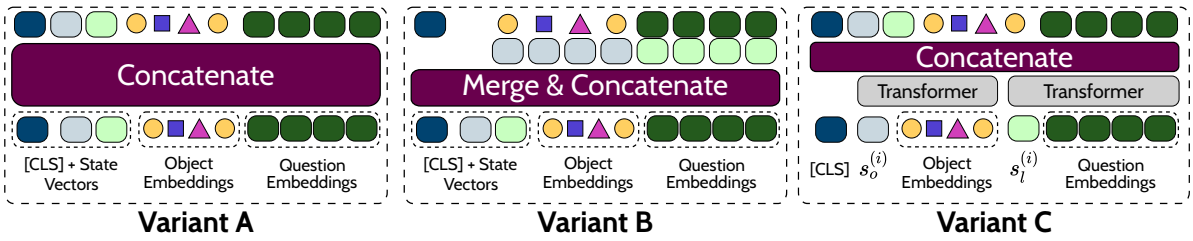

Figure 3: Overview of the different variants of our combiner.

3.4. Combiner

The combiner is responsible for merging the [CLS] token embedding $h_{\texttt{[CLS]}}$ and the two dialog state vectors with the rest of the multi-modal input, i.e. the object and text embeddings. We propose three variants of how this merging is performed, as illustrated in Figure 3.

Variant A. This variant concatenates all inputs to form the input of the current dialog turn, i.e.

$$I_A = \left[h^{\texttt{[CLS]}}, s_o^{(i)}, s_l^{(i)}, H_0^{obj}, H_0^{w}\right]. \quad (11)$$

Variant B. This variant first appends the current object state vector $s_o^{(i)}$ to the object embeddings and the current language state vector $s_l^{(i)}$ to the text embeddings. It then maps them to the d-dimensional space using a linear layer and concatenates all inputs like in Variant A, i.e.

$$I_B = \left[h^{\texttt{[CLS]}}, \tilde{H}_0^{obj}, \tilde{H}_0^{w}\right], \quad (12)$$

$$\tilde{H}_0^{obj} = \left[W^b[s_o^{(i)}; h_1^{obj}], ..., W^b[s_o^{(i)}; h_{TN_o}^{obj}]\right], \quad (13)$$

$$\tilde{H}_0^{w} = \left[W^b[s_l^{(i)}; h_1], ..., W^b[s_l^{(i)}; h_{N_w}]\right], \quad (14)$$

where $W^b \in \mathbb{R}^{d\times(2d)}$ is a learnable parameter.

Variant C. This variant uses two small transformers operating on the concatenation of the object state vector with the object embeddings and the language state vector with the text embeddings, respectively. Then, it concatenates the outputs of their final layers (H^{obj} and H^{w}) with the [CLS] token embeddings and the state vectors, i.e.

$$I_C = \left[h^{\texttt{[CLS]}}, s_o^{(i)}, s_l^{(i)}, H^{obj}, H^{w}\right]. \quad (15)$$

3.5. Encoder/Decoder

The last component of our model, the encoder, consists of L transformer layers. It takes the output of the combiner as input and applies multi-head self-attention operations with layer normalization. The output of its final layer is used to either predict or generate the answer to the current question.

Prediction. We use a linear layer with softmax to map the final d-dimensional [CLS] token embeddings to the N-dimensional answer space. We train our model end-to-end using cross-entropy loss. During testing, we choose the answer $\hat{a}$ with the highest score, that is:

$$\hat{a} = \underset{a\in A}{\arg\max}\left[logP\left(a|s_o^{(i)}, s_l^{(i)}, H^{obj}, H^{w}\right)\right], \quad (16)$$

where A is the set of all candidate answers.

Generation. For answer generation, we couple the encoder with a decoder (with the same number of transformer layers L and attention heads) and append the ground truth answer to the question. We then train the model end-to-end using the teacher forcing strategy (Williams and Zipser, 1989) while making sure that only the left part is visible to each answer token when attention is applied as shown in the attention mask of Figure 2. While testing, we select the token $\hat{y}_j$ at the current step j with the highest score until the [EOS] token is predicted or if a maximum length of 40 tokens is reached, i.e.

$$\hat{y}_j = \underset{y\in V}{\arg\max}\left[logP\left(y|\hat{Y}_{j-1}, s_o^{(i)}, s_l^{(i)}, H^{obj}, H^{w}\right)\right], \quad (17)$$

where $\hat{Y}_{j-1} = [\hat{y}_1, ..., \hat{y}_{j-1}]$ is the set of previously predicted tokens and V is the vocabulary.

4. Experiments

Datasets & Metrics. To evaluate the performance of our model for both discriminative and generative task settings we used the DVD (Le et al., 2021b) and SIMMC 2.1 (Kottur et al., 2021) benchmark datasets. DVD was recently proposed with the goal of assessing higher-order spatio-temporal reasoning capabilities of video dialog models. It is based on 11k videos from the challenging CATER dataset (Girdhar and Ramanan, 2020) and contains over 100k dialogs and 1M question-answer pairs with detailed spario-temporal annotations. SIMMC 2.1 is a task-oriented dataset that was proposed for realistic virtual assistance scenarios. It contains 11k dialogs from the shopping domain. We used accuracy and

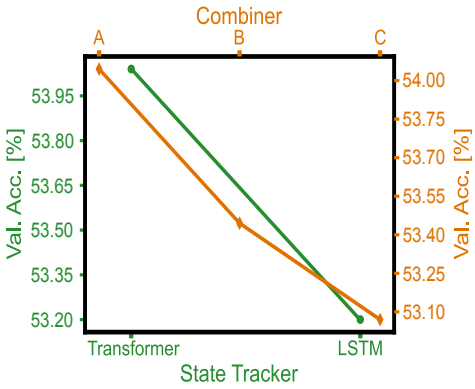

Figure 4: Performance comparison of 𝕆𝕃𝕍𝕚𝕋 with different combiners and state tracker variants.

Figure 5: Performance comparison of 𝕆𝕃𝕍𝕚𝕋 with different numbers of objects and history turns.

BLEU-4 scores to assess the performance of our model on DVD and SIMMC 2.1, respectively.

Combiner. As can be seen from Figure 4, variant A of the combiner – despite having the simplest architecture – outperformed all other variants with a validation accuracy of 54.01% on DVD. We hypothesize that this is because concatenating the states vectors to all respective embeddings (variant B) prevents the model from applying attention over the raw state vectors that contain rich information about the previous relevant objects and dialog turns. On the other hand, using additional transformer layers increases complexity and the risk of over-fitting (variant C). We also experimented with an LSTM-based state tracker which reached a validation accuracy of 53.20% compared to 54.01% achieved by the transformer-based state trackers.

Number of Objects and History Turns. To find the best hyper-parameters (number of the most important previous objects and history turns) of the state tracker variant A, we first used the full history, i.e. all previous turns, and varied the number of objects. Similarly to the previous experiment, we considered the two variables independently to keep the size of the search space tractable. Once the optimal number of objects was found in the OST, we optimized the number of previous turns in the LST. As can be seen from Figure 5, we increased the number of object embeddings in each OST step (i.e. None, 1, 2, 3, and 4) while keeping the entire dialog history. 𝕆𝕃𝕍𝕚𝕋 achieved its peak validation accuracy of 55.10% when using two object embeddings. With more embeddings, performance started to decrease and reached 54.86% and 54.72% with three and four object embeddings, respectively. We then fixed the number of object embeddings to two and varied the history length. The best validation accuracy of 55.39% was reached when we used a history of seven previous turns. Using less or more turns resulted in reduced performance, i.e. 54.53% and 54.71% with 10 and 5 turns, respectively. These optimal values were then fixed for the rest of the experiments unless it was explicitly stated otherwise.

Baselines. We compared 𝕆𝕃𝕍𝕚𝕋 using the previous optimal values against strong baselines that (1) were recently published for video dialog state tracking, i.e. VDTN (Le et al., 2022), and (2) hold the current state-of-the-art performance on both DVD and SIMMC 2.1 datasets, i.e. MTN (Le et al., 2019) and the GPT-2 (Radford et al., 2019) based MM-DST model (Moon et al., 2020).

5. Results

5.1. Results on DVD

Quantitative Analysis. We first evaluated 𝕆𝕃𝕍𝕚𝕋 on the discriminative video dialog task. Given a video, a dialog history, and a question, the model needs to predict the correct answer from a pool of $N = 40$ candidate answers. As can be seen from Table 1, our model reached an overall test accuracy of 54.85%, thereby outperforming the state of the art model by 3.75% absolute points (last column of Table 1). Not only does our model reached a new overall state-of-the-art test accuracy, it did so by improving the performance across *all* question categories. This is in stark contrast to previous methods (e.g. MTN (Le et al., 2019) and more importantly the recent VDTN model (Le et al., 2022)) that typically only improved performance for a subset of categories. Moreover, our method performed particularly well on challenging categories that require accurate object tracking and that existing models tend to struggle with. Specifically, 𝕆𝕃𝕍𝕚𝕋 improved the performance of *action count, action query* and *attribute query* by 4.62%, 5.84%, and 7.67% over the state of art, respectively. Furthermore, our model outperformed the state of the art across

Accuracy [%]	Answer Prior	Q-type (Random)	Q-type (Freq)	Q-retrieval	RNN (Q)	HRNN (C+Q)	HRNN (C+Q+V)	VDTN† (C+Q+V)	MTN (C+Q+V)	𝕆𝕃𝕍𝕚𝕋	Δ [%]
Action count	0.0	9.30	23.40	19.80	16.30	28.20	36.00	38.78	38.80	**43.42**	+4.62
Action query	0.0	12.70	23.70	20.60	25.80	33.10	38.60	39.37	39.40	**45.24**	+5.84
Attribute query	0.0	32.90	38.70	39.40	38.10	39.20	45.10	42.93	43.10	**52.77**	+7.67
Compare action seq	33.40	34.10	37.30	35.10	45.50	52.50	57.50	61.57	61.60	**61.88**	+0.28
Compare action set	25.10	28.20	36.30	28.20	32.80	40.00	44.30	45.41	45.40	**51.71**	+6.30
Compare action freq	48.50	50.00	50.50	44.40	58.40	56.90	65.20	66.42	67.10	**67.91**	+0.81
Object count	0.0	9.10	23.30	18.80	26.20	38.60	40.20	39.86	39.90	**43.34**	+3.14
Object exist	48.90	49.80	51.10	54.40	66.40	67.00	69.40	69.42	69.00	**71.25**	+1.83
None	0.0	32.10	38.30	39.00	38.30	39.50	45.10	43.51	43.40	**52.74**	+7.64
Atomic (non-spatial)	18.80	26.30	31.90	42.40	47.20	47.80	50.70	48.88	48.90	**56.54**	+5.84
Atomic (spatial)	21.20	27.30	35.50	27.60	36.80	46.00	47.60	47.12	47.10	**49.61**	+2.01
Compositional	22.80	28.00	35.40	32.10	40.00	45.80	51.40	53.18	53.20	**56.70**	+3.50
Transfer (attribute)	0.0	30.70	45.50	37.10	40.80	45.70	57.30	57.70	57.70	**61.28**	+3.58
Transfer (spatial)	49.80	42.40	44.90	26.40	29.60	48.10	47.40	47.86	48.00	**50.50**	+2.40
Transfer (temporal)	28.90	38.40	22.60	3.00	30.20	53.50	64.60	68.72	69.00	**74.83**	+5.83
All	21.30	27.80	35.30	32.10	39.70	45.80	50.20	51.02	51.10	**54.85**	+3.75

Table 1: Performance comparison on DVD *test* split. Best and second best performances are in **bold** and underlined, respectively. Q, C, and V denote question, context, and visual input, respectively, and † denotes training with additional supervision.

	test-dev			test-std		
	MM-DST	MTN	𝕆𝕃𝕍𝕚𝕋	MM-DST	MTN	𝕆𝕃𝕍𝕚𝕋
BLEU-4	19.20	21.70	**28.30**	19.20	21.00	**25.20**
Rel. Δ [%]	+47.4	+30.4	–	+31.3	+20.0	–

Table 2: Performance comparison on SIMMC 2.1 *test-dev* and *test-std* splits.

Ablated 𝕆𝕃𝕍𝕚𝕋				State Trackers	
DVD		SIMMC 2.1			
Acc. [%]	Δ[%]	BLEU-4	Rel. Δ [%]	OST	LST
44.11	11.28	21.50	22.79	✗	✗
46.52‡	8.87	21.70†	21.66	✓	✗
54.94‡	0.45	26.30‡	0.38	✗	✓
55.39†	–	**26.40**†	–	✓	✓

Table 3: Performance comparison of different ablated versions on DVD and SIMMC 2.1 *val* splits. † and ‡ represent significant improvement with $p < 0.05$ and $p < 0.01$ compared to the second best score in each column, respectively.

all *transfer* categories (penultimate section of Table 1) matching our hypothesis that it has superior spatio-temporal reasoning capabilities with accuracies of 61.28%, 50.50%, and 74.83% on the *transfer (attribute)*, *transfer (spatial)*, and *transfer (temporal)* categories, respectively. This corresponds to respective improvements of 3.58%, 2.40%, and 5.83% over the state of the art.

Qualitative Analysis. Figure 6 shows sample predictions of 𝕆𝕃𝕍𝕚𝕋 on the DVD test split together with some video frames and MONet object masks. While our model answered the first three questions correctly, it failed at the last two. For instance, when asked about the actions the red metal cone performs, it predicted "*flying*" instead of "*flying, sliding*". This can most likely be attributed to the difficulty of the CATER videos – deciding between "*flying*" and "*sliding*" requires the model to reason about the object's shadow which are hard to acquire from the object embeddings.

5.2. Results on SIMMC 2.1

For the generative setting, we evaluated 𝕆𝕃𝕍𝕚𝕋 on the SIMMC 2.1 dataset that comes with two test splits (*test-dev* and *test-std*). The ground-truth answers of the former are publicly available but those of the latter are withheld by the creators of the dataset. As can be seen in Table 2, our model outperformed strong baselines on the test-dev split by reaching a BLEU-4 score of 28.30 compared to 19.20 and 21.70 achieved by MM-DST and MTN, respectively. This corresponds to relative improvements of 47.4% and 30.4%, respectively. Table 2 also shows the performance of 𝕆𝕃𝕍𝕚𝕋 as well as the baselines on the *test-std* split of SIMMC 2.1 where it outperformed both baselines by a considerable margin achieving a BLEU-4 score of 25.20. This corresponds to relative improvements of 31.3% and 20.0% over MM-DST and MTN, respectively.

To qualitatively assess the performance of our model, we show in Figure 7 the generated answer of our model on a randomly sampled example from the SIMMC 2.1 *test-dev* split. Although there is not a big overlap in the used words between our model's prediction (green) and the ground-truth (red), the two responses are semantically similar and lead to the same action of the virtual assistant, i.e. ask for more information from the user.

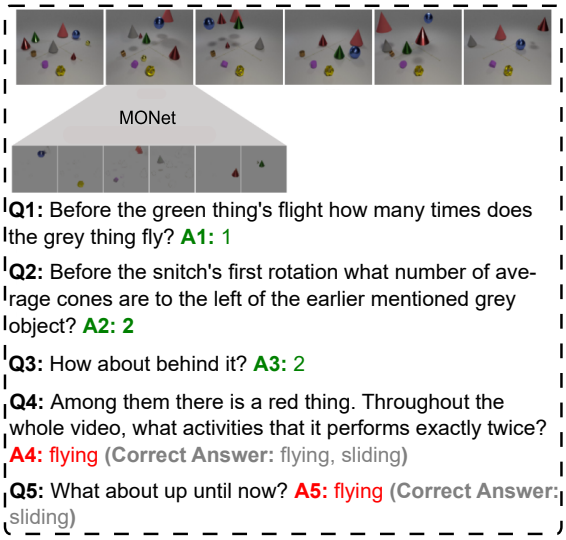

Figure 6: 𝕆𝕃𝕍𝕚𝕋 predictions on a random DVD sample from the test split.

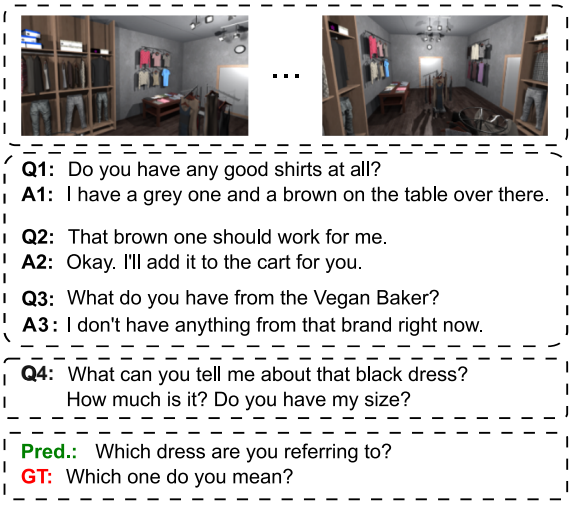

Figure 7: 𝕆𝕃𝕍𝕚𝕋 predictions on a random SIMMC 2.1 sample from the test split.

LM	# of LM param.	# of Layers L	Accuracy [%]
None	-	4	54.46
None	-	16	54.54
DistillRoBERTa	81.5M	4	55.39
RoBERTa	354M	**4**	**55.41**

Table 4: The effect of pre-trained language models on the performance of our model. Performance is measured on the DVD *val* split.

5.3. Ablation Study

Performance without OST & LST. We first evaluated a version of our model without any state trackers (first row in Table 3). It is important to note that in this case, the model still has access to the MONet object embeddings but not to the previous history turns. This ablated version performed poorly, only reaching 44.11% validation accuracy on DVD and 21.50 BLEU-4 score on SIMMC.

Performance with the OST. Adding the OST (second row in Table 3) resulted in a notable performance improvement on both datasets, i.e. 46.52% on DVD and 21.70 on SIMMC. We emphasize that this version still does not have access to the previous history turns and treats the task as a simple single-turn question answering task. This finding indicates the positive influence of the OST in helping the model to focus on the most important objects when answering a question.

Performance with the LST. Adding the LST (third row in Table 3) lifted the performance on both DVD and SIMMC 2.1 by a considerable margin compared to the version with only the OST. We note here that this version, in stark contrast to the previous one, not only had access to the visual input in the form of MONet embeddings but also to the previous dialog turns. This variant reached a validation accuracy of 54.94% and a BLEU-4 score of 26.30. However, it still under-performed our full model with both dialog trackers in action which achieved the best validation performance on both datasets, i.e. 55.39% accuracy on DVD and 26.40 BLEU-4 score on SIMMC 2.1.

Performance with Pre-trained LLM. To assess the effect of pre-trained language models on 𝕆𝕃𝕍𝕚𝕋, we completely removed the DistillRoBERTa model and trained the encoder layers from scratch on the DVD dataset. Thereby, we increased the number of the encoder transformer layers from 4 to 16 to compensate for the removal of the pre-trained language model. As can be seen from Table 4, our model's DVD validation accuracy dropped from 55.41% to 54.46% when DistillRoBERTa was replaced by 4 encoder layers. However, its performance improved when we increased the number of these layers to 16 and reached a validation accuracy of 54.54%. Finally, using a full RoBERTa model only improved the model's accuracy by 0.02%. As a result, we decided to use a DistillRoBERTa model in order to train efficiently.

6. Conclusion

In this work, we proposed 𝕆𝕃𝕍𝕚𝕋 – a novel model for video dialog operating over a novel two-stream multi-modal attention-based dialog state tracker that jointly learns visual object representations and textual embeddings. Through extensive experiments on two challenging datasets, we demon-

strated significant improvements over strong baselines for discriminative and generative tasks. Our results are highly encouraging and underline the importance of performing multi-modal dialog state tracking for a more reliable higher-order reasoning of (video) dialog systems. We strongly believe that real-world multi-modal dialog tasks can equally benefit from our novel multi-modal state tracking approach. We leave this for future work.

Acknowledgment

A. Bulling was funded by the European Research Council (ERC; grant agreement 801708).

7. Bibliographical References

Adnen Abdessaied, Mihai Bâce, and Andreas Bulling. 2022. Neuro-Symbolic Visual Dialog. In *COLING*.

Adnen Abdessaied, Lei Shi, and Andreas Bulling. 2024. VD-GR: Boosting Visual Dialog With Cascaded Spatial-Temporal Multi-Modal Graphs. In *WACV*.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *ICCV*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, et al. 2020. Language models are few-shot learners. In *NeurIPS*.

Christopher P. Burgess, Loic Matthey, Nicholas Watters, Rishabh Kabra, Irina Higgins, Matt Botvinick, and Alexander Lerchner. 2019. MONet: Unsupervised Scene Decomposition and Representation. In *arXiv, 1901.11390*.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M.F. Moura, Devi Parikh, and Dhruv Batra. 2017. Visual Dialog. In *CVPR*.

David Ding, Felix Hill, Adam Santoro, Malcolm Reynolds, and Matthew Botvinick. 2021. Attention over Learned Object Embeddings Enables Complex Visual Reasoning. In *NeurIPS*.

Rohit Girdhar and Deva Ramanan. 2020. CATER: A diagnostic dataset for Compositional Actions and TEmporal Reasoning. In *ICLR*.

Xiaoxiao Guo, Hui Wu, Yu Cheng, Steven Rennie, Gerald Tesauro, and Rogerio Feris. 2018. Dialog-based interactive image retrieval. In *NeurIPS*.

Christopher Hahn, Frederik Schmitt, Jens U. Kreber, Markus Norman Rabe, and Bernd Finkbeiner. 2021. Teaching Temporal Logics to Neural Networks. In *ICLR*.

Chi Han, Jiayuan Mao, Chuang Gan, Joshua B. Tenenbaum, and Jiajun Wu. 2019. Visual Concept Metaconcept Learning. In *NeurIPS*.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask R-CNN. In *ICCV*.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Comput.*

Chiori Hori, Huda Alamri, Jue Wang, Gordon Wichern, Takaaki Hori, Anoop Cherian, Tim K. Marks, Vincent Cartillier, Raphael Gontijo Lopes, Abhishek Das, Irfan Essa, Dhruv Batra, and Devi Parikh. 2019. End-to-end audio visual scene-aware dialog using multimodal attention-based video features. In *ICASSP*.

Weike Jin, Zhou Zhao, Mao Gu, Jun Xiao, Furu Wei, and Yueting Zhuang. 2019. Video dialog via progressive inference and cross-transformer. In *EMNLP-IJCNLP*.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning. In *CVPR*.

Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *CVPR*.

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

Satwik Kottur, José M. F. Moura, Devi Parikh, Dhruv Batra, and Marcus Rohrbach. 2019. CLEVR-Dialog: A Diagnostic Dataset for Multi-Round Reasoning in Visual Dialog. In *NAACL*.

Guillaume Lample and François Charton. 2020. Deep learning for symbolic mathematics. In *ICLR*.

Hung Le, Nancy F. Chen, and Steven C. H. Hoi. 2021a. Learning reasoning paths over semantic graphs for video-grounded dialogues. In *ICLR*.

Hung Le, Nancy F. Chen, and Steven C.H. Hoi. 2022. Multimodal Dialogue State Tracking. In *NAACL*.

Hung Le, Doyen Sahoo, Nancy Chen, and Steven Hoi. 2019. Multimodal transformer networks for end-to-end video-grounded dialogue systems. In *ACL*.

Hung Le, Chinnadhurai Sankar, Seungwhan Moon, Ahmad Beirami, Alborz Geramifard, and Satwik Kottur. 2021b. DVD: A diagnostic dataset for multi-step reasoning in video grounded dialogue. In *ACL*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. In *arXiv, 1907.11692*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *ICLR*.

Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B. Tenenbaum, and Jiajun Wu. 2019. The Neuro-Symbolic Concept Learner: Interpreting Scenes, Words, and Sentences From Natural Supervision. In *ICLR*.

Seungwhan Moon, Satwik Kottur, Paul Crook, Ankita De, Shivani Poddar, Theodore Levin, David Whitney, Daniel Difranco, Ahmad Beirami, Eunjoon Cho, Rajen Subba, and Alborz Geramifard. 2020. Situated and interactive multimodal conversations. In *COLING*.

Nikola Mrkšić, Diarmuid Ó Séaghdha, Tsung-Hsien Wen, Blaise Thomson, and Steve Young. 2017. Neural belief tracker: Data-driven dialogue state tracking. In *ACL*.

Vahid Noroozi, Yang Zhang, Evelina Bakhturina, and Tomasz Kornuta. 2020. A fast and robust bert-based dialogue state tracker for schema-guided dialogue dataset. In *Proceedings of the Workshop on Conversational Systems Towards Mainstream Adoption, KDD*.

Xudong Pan, Mi Zhang, Shouling Ji, and Min Yang. 2020. Privacy risks of general-purpose language models. In *Symposium on Security and Privacy*.

Wei Pang and Xiaojie Wang. 2020. Visual dialogue state tracking for question generation. In *AAAI*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zach DeVito, Martin Raison, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. In *Proceedings od the Workshop on Energy Efficient Machine Learning and Cognitive Computing, NeurIPS*.

Haoyue Shi, Jiayuan Mao, Kevin Gimpel, and Karen Livescu. 2019. Visually grounded neural syntax acquisition. In *ACL*.

Congzheng Song and Ananth Raghunathan. 2020. Information leakage in embedding models. In *ACM CCS*.

Florian Strub, Harm De Vries, Jeremie Mary, Bilal Piot, Aaron Courville, and Olivier Pietquin. 2017. End-to-end optimization of goal-driven and visually grounded dialogue systems. In *IJCAI*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *NeurIPS*.

Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. 2019. Composing text and image for image retrieval-an empirical odyssey. In *CVPR*.

Ronald J Williams and David Zipser. 1989. A learning algorithm for continually running fully recurrent neural networks. *Neural Comput.*

Yaxiong Wu, Craig Macdonald, and Iadh Ounis. 2022. Multi-Modal Dialog State Tracking for Interactive Fashion Recommendation. In *ACM RecSys*.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*.

Puyang Xu and Qi Hu. 2018. An End-to-end Approach for Handling Unknown Slot Values in Dialogue State Tracking. In *ACL-findings*.

Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. 2020. CLEVRER: Collision Events

for Video Representation and Reasoning. In *ICLR*.

Kexin Yi, Jiajun Wu, Chuang Gan, Antonio Torralba, Pushmeet Kohli, and Joshua B. Tenenbaum. 2018. Neural-symbolic vqa: Disentangling reasoning from vision and language understanding. In *NeurIPS*.

Zhou Zhao, Xinghua Jiang, Deng Cai, Jun Xiao, Xiaofei He, and Shiliang Pu. 2018. Multi-turn video question answering via multi-stream hierarchical attention context network. In *IJCAI*.

Ligeng Zhu, Zhijian Liu, and Song Han. 2019. Deep leakage from gradients. In *NeurIPS*.