Video Dialog as Conversation about Objects Living in Space-Time

Hoang-Anh Pham[1], Thao Minh Le[1], Vuong Le[1], Tu Minh Phuong[2], Truyen Tran[1]

[1]Applied Artificial Intelligence Institute, Deakin University, Australia
[2]Posts and Telecommunications Institute of Technology, Vietnam
[1]{phamhoan, thao.le, vuong.le, truyen.tran}@deakin.edu.au
[2]phuongtm@ptit.edu.vn

Abstract. It would be a technological feat to be able to create a system that can hold a meaningful conversation with humans about what they watch. A setup toward that goal is presented as a video dialog task, where the system is asked to generate natural utterances in response to a question in an ongoing dialog. The task poses great visual, linguistic, and reasoning challenges that cannot be easily overcome without an appropriate representation scheme over video and dialog that supports high-level reasoning. To tackle these challenges we present a new object-centric framework for video dialog that supports neural reasoning dubbed COST–which stands for *Conversation about Objects in Space-Time*. Here dynamic space-time visual content in videos is first parsed into object trajectories. Given this video abstraction, COST maintains and tracks object-associated *dialog states*, which are updated upon receiving new questions. Object interactions are dynamically and conditionally inferred for each question, and these serve as the basis for relational reasoning among them. COST also maintains a history of previous answers, and this allows retrieval of relevant object-centric information to enrich the answer forming process. Language production then proceeds in a step-wise manner, taking the context of the current utterance, the existing dialog, and the current question. We evaluate COST on the AVSD test splits (DSTC7 and DSTC8), demonstrating its competitiveness against state-of-the-arts.

1 Introduction

It is a hallmark of visual intelligence to build a system that can hold a meaningful conversation with humans about a video. A system of this capability would be a strong contender for passing the visual Turing test [11]. Posed as video dialog [1, 12], this task challenges the current arts due to its sheer complexity on many fronts. A dialog has its natural flow through multiple turns, each of which builds upon the previous questions and answers. This demands deep linguistic understanding about, and keeping track of, what has been said and grounded on the visual concepts found in the video, and then analyzing the new question in this newly established context. Given the question semantics

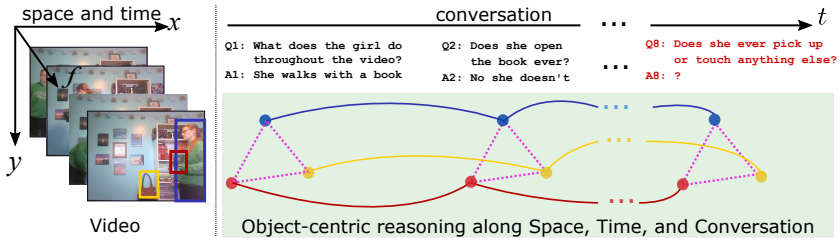

Fig. 1. We introduce COST, an object-centric reasoning framework that collects clues along with the Space and Time dimensions of the Video and the Conversation dimension of the dialog toward reliable QA.

and its constituent words, generating a linguistic answer necessitates symbolic grounding and visual reasoning through the complex space-time structure of the video, where in multiple objects interact in a dynamic manner.

Video dialog is inherently harder than the task of visual dialog over static images [8] due to the temporal dynamics of the scene [1]. It is also harder than the standard setting of video QA [28] since the next question may not be comprehensible without maintaining a history of, and referring to the previous answers. There have been several attempts to tackle these challenges. Early attempts [14, 24, 33, 35] encode the dialog flow using recurrent neural networks. Later methods [22, 29] resort to Transformers for better distant dependencies as well as cross-modality relations. More recent methods make use of graphs as a representation of dialog structures [12, 23] and co-reference [18] which achieved the new state-of-the-art result for this task. However, we have only scratched the surface of what is possible and the principal challenges remain.

A plausible pathway toward solving the remaining video dialog challenges is through high-level, object-centric representation and reasoning as seen in human visual cognition: Humans see objects and agency as the core "living" constructs with natural compositionally, permanency, temporal dynamics, and n-body interactions [39]. Importantly, the object-centric approach has recently been found to be essential for reasoning in Visual QA [27] and Video QA [7], thanks to the ease of binding linguistic concepts to visual regions.

To this end, we propose a new object-centric framework dubbed COST (**C**onversation about **O**bjects in **S**pace-**T**ime) for video dialog. Video is first parsed into a set of object trajectories throughout the video across the spatial dimensions of the frame and the temporal spans of object lives. Central to COST is a model of the object states dynamics as the conversation progresses[1]. In particular, COST maintains a *recurrent system of dialog-induced object states* throughout the course of conversation. For each new question, the object lives will be consulted through semantic word-object grounding and selective frame attention, producing question-guided object representations. These serve as input for the *dialog state recurrent networks* to generate updated dialog states. These

[1] This is related to, but distinct from, the dialog state tracking in typical task-oriented dialogs in NLP [10]

states are used to construct a *question-specific interaction matrix* among objects, enabling relational reasoning among them. The results are coupled with the answers from the previous conversational turns to produce new representations, which are then decoded into the response utterance. See Fig. 1 for an illustration of `COST` in action.This paper makes three contributions:

(i) To the best of our knowledge, we are the first to successfully use object-centric reasoning for video dialog.

(ii) The inductive priors brought by our object-centric framework as *a recurrent system of dialog-induced object states* are *unique* and particularly beneficial for video dialog as empirically evident.

(iii) Experimental results on the AVSD dataset with two different test splits DSTC7 and DSTC8 shows that `COST` is highly competitive against rivals.

2 Related Work

Visual and video dialog The visual dialog task and an accompanied dataset (VisDial) were first introduced in [8]. The task requires both conversational and visual reasoning ability over multiple turns . Like the precursor Visual QA task, visual dialog needs deep understanding of the visual concepts and relations in the images, and reasoning about them in response to the current question. A unique challenge in visual dialog is the co-referrence problem of linguistic information between dialog turns. The early works tried to resolve this by hierarchical encoding [37] or memory network based on attention [36]. However, these works mostly focus on history dialog reasoning. The work of [21] solved the co-reference on both visual and linguistic space using neural module networks. Recently, like the other vision-language tasks, visual dialog is also beneficial from cross-modal pre-training which was employed [13, 32, 43, 53].

The video dialog task pushes the challenges further, thanks to the complexity of analyzing video. Video-grounded dialog system (VGDS) received more interested from community recently with the DSTC7 [51] and DSTC8 [19] challenge. Early attempts [14, 24, 33, 35] employ the recurrent neural network to encode dialog history. Later methods [6, 30, 49] use the Attention mechanism, or [45] design a memory networks to extract the relationship between different modalities, [22, 29, 26] employ Transformer-based network to resolve the cross modality learning. More explicit relationships in dialogs are recently studied, showing promising results [12, 23]. This extends to co-reference graph across visual and textual domains [18]. However, these approach all represent the video by frame features, just lacking the key concepts of of object permanence.

Object-centric visual reasoning Visual QA and visual dialog benefit greatly from object-centric representation of visual content as this fills the gap between low-level visual features and high-level linguistic semantics [7, 27]. The early work proposed the object-centric representation [9] by leveraging the object detection for images and [4, 17, 44, 47, 48] combining with the tracking algorithm for video inputs. In problems that require further reasoning on objects, a relational network

was introduced by [2], and [16, 18, 34, 42, 50, 52] make use of graph-based method for transparent reasoning. However, these methods only use object's features as an additional input on a generic network without any prior reasoning structure in space-time, which is the key of success in further reasoning. The work in [7] introduces a generic reasoning unit through dynamically building interaction graphs between objects in space-time derived by the question, but this is limited to single question.

3 Method

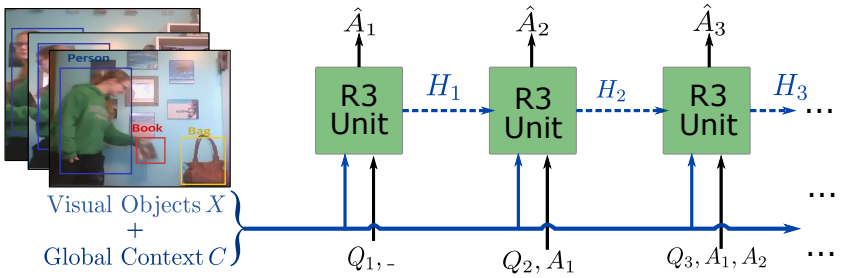

Fig. 2. The architecture of `COST` model features a chain of *Recurrent Relational Reasoning* (`R3`) Units which maintain the *dialog states* $\{H_i\}$ across the turns of the conversation.

Given a video V, the task of video dialog is to hold a smooth conversation of T turns. Each turn t is a textual question-answer pair (Q_t, A_t). We want to estimate a model parameterized by θ that returns the best answer for the corresponding question:

$$\hat{A}_t = \arg\max_A P\left(A \mid V, Q_{1:t-1}, A_{1:t-1}, Q_t; \theta\right), \tag{1}$$

for $t = 1, 2, ..., T$.

The main challenge lies in the coherent answering with respect to the existing conversation of length $t-1$, while reasoning over the space-time of video, which itself demands efficient and effective schemes of representation that supports a high-level dialog. With these constraints in mind, here we treat the video V as a collection of object lives, each of which is a trajectory of spatial positions equipped with object visual features. This object-oriented view enables ease of constructing reasoning paths in response to linguistic queries.

In the following we present `COST`, an *object-centric reasoning model* for video dialog.. Fig. 2 shows the overall architecture of `COST`.

3.1 Preliminaries

Following [7], we parse the video of F frames into N sequences of objects tracked over time. Objects at each frame are position-coded together with their appearance

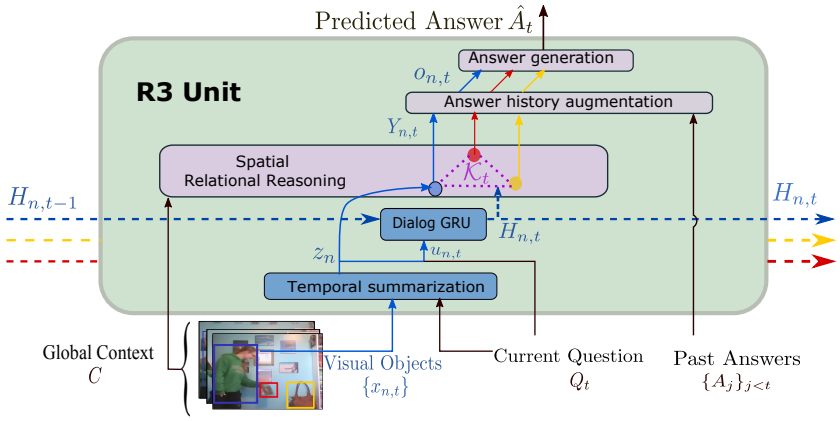

Fig. 3. The architecture of a Recurrent Relational Reasoning (R3) Unit operating at dialog turn t. The space-time-dialog reasoning happens at the three corresponding member blocks. The colors blue/red/yellow indicate terms and operations specific to each object (only drawn for the blue object). Pink indicates cross-object operations.

features. Each *visual object live* throughout the video is therefore represented as a matrix $X_n \in \mathbb{R}^{F\times d}$ for $n = 1, 2, ..., N$. Moreover, each frame is represented by a holistic context vector $c \in \mathbb{R}^{1\times d}$, therefore the holistic context matrix representation of video is denoted as $C \in \mathbb{R}^{F\times d}$. Dialog sentences (questions and answers) are split into words and then embedded into a matrix $S = [w_{1:L}] \in \mathbb{R}^{L\times d}$ where L is sentence length. With a slight abuse of notation, we use d to denote the size of both visual feature vectors and linguistic vectors for ease of reading.

We also make use of the attention function [40] defined over the triplet of *query* $q \in \mathbb{R}^{1\times d}$, *keys* $K \in \mathbb{R}^{M\times d}$ and *values* $V \in \mathbb{R}^{M\times d}$:

$$\text{Attn}\left(q, K, V\right) := \sum_{m=1}^{M} \text{softmax}_m \left(\frac{K_m W_k \left(qW_q\right)^\top}{\sqrt{d}} \right) V_m W_v \in \mathbb{R}^{1\times d}. \tag{2}$$

where W_k, W_q, and W_v are learnable parameters. In essence, this function reads out the most relevant values whose keys match the query. Likewise, a linear projection $W \in \mathbb{R}^{d_1\times d_2}$ from $x \in \mathbb{R}^{1\times d_1}$ to $\mathbb{R}^{1\times d_2}$ is denoted as

$$\text{Linear}(x) := xW. \tag{3}$$

3.2 Recurrent Relational Reasoning over Space-Time and Turns

The COST is a recurrent system that keeps track of the evolution of the dialog states over turns. Each step is a reasoning unit R3 (short for Recurrent Relational Reasoning) which takes as input a question $Q_t \in \mathbb{R}^{S\times d}$ of the present turn t, the previous dialog states $H_{t-1} \in \mathbb{R}^{N\times d}$, a holistic context representation $C \in \mathbb{R}^{F\times d}$, a set of N object sequences $X = \left\{X_n \mid X_n \in \mathbb{R}^{F\times d}\right\}_{n=1}^{N}$, and past answers $\{A_j\}_{j<t}$. The outputs of a R3 unit are the representations of object

lives conditioned on the current query and information from past turns. Fig. 3 illustrates the structure of `R3` unit.

Query-induced Temporal Summarization At each *conversational turn* t, we produce query-specific object representation. Firstly, we generate a *query-specific* summary of each object sequence $X_n \in \mathbb{R}^{F\times d}$ into a vector $z_n \in \mathbb{R}^{1\times d}$ using temporal attention over frames. The frame attention weights are driven by the words in the query $Q_t = \left\{Q_{s,t} \in \mathbb{R}^{1\times d}\right\}_{s=1}^{S_t}$. These produce a query-specific *object resume* as follows:

$$z_n = \frac{1}{S_t}\sum_{s=1}^{S_t} \text{Attn}\left(Q_{s,t}, X_n, X_n\right) \in \mathbb{R}^{1\times d}. \tag{4}$$

To handle the case where an object cannot be detected at particular frames, we place a binary mask in the appropriate place.

Next we generate an *object-specific* embedding of the question:

$$q_n = \text{Attn}\left(z_n, Q_t, Q_t\right) \in \mathbb{R}^{1\times d},$$

Finally the object embedding is modulated by the question as:

$$u_{n,t} = \tanh\left([z_n, q_n, q_n \odot z_n]\right) \in \mathbb{R}^{1\times 3d}. \tag{5}$$

This embedding serves as input for the recurrent network, which is presented in the next subsection.

Recurrent Dialog States In dialog, the question at a turn typically advances from, and co-refers to, the previous questions and answers. In video dialog, questions are semantically related to objects appearing in video. We hence maintain a **dialog state** at turn t in the form of a matrix $H_t \in \mathbb{R}^{N\times d}$, i.e., each row corresponds to an object. The state dynamics is modeled in a set of N parallel recurrent networks:

$$H_{n,t} = \text{GRU}\left(H_{n,t-1}, u_{n,t}\right) \in \mathbb{R}^{1\times d}, \tag{6}$$

for $n = 1, 2, ..., N$, where GRU is a standard Gated Recurrent Unit [5], $u_{n,t}$ is turn-specific embedding of the object n at turn t calculated in Eq. (5).

As the dialog states are object-centric, co-references between questions are *indirectly* and *distributionally* captured into the current multi-object states through the integration of previous states $H_{n,t-1}$ (which contains information of the previous questions) and the current question as part of $u_{n,t}$. In what follows, we will show how H_t is used for relational reasoning.

Relational Reasoning between Objects Equipped with dialog states, we now model the inter-object interaction which describes the behavior of objects with their neighbors driven by the current question Q_t. We employ a spatial graph whose vertices are the object resumes $Z = \{z_n\}_{n=1}^{N}$ computed in Eq. (4), and the edges are represented by an adjacency matrix $\mathcal{K}_t \in \mathbb{R}^{N \times N}$ which is calculated dynamically as the *question-specific interaction matrix* between objects:

$$\mathcal{K}_t = \text{softmax}\left(\frac{H_t H_t^\top}{\sqrt{d}}\right), \tag{7}$$

where H_t is computed in Eq. (6). This matrix serves as a backbone for a Deep Graph Convolutional Network (DGCN) [27] to refine object representations by taking into account the relations with their neighboring nodes:

$$\bar{Z} = \text{DGCN}\left(Z; \mathcal{K}_t\right) \tag{8}$$

Details of $\text{DCGN}\left(\cdot;\cdot\right)$ is given in the Supplement.

Utilizing visual context To utilize the underlying background scene information and compensate for possible undetected objects, we augment the object representations with the holistic context information $C \in \mathbb{R}^{F \times d}$. We summarize the context sequence into a vector as follows:

$$\bar{c} = \frac{1}{S_t}\sum_{s=1}^{S_t} \text{Attn}\left(Q_{s,t}, C, C\right) \in \mathbb{R}^{1 \times d}. \tag{9}$$

Finally, the final output of this component is

$$Y_{n,t} = \text{Linear}\left(\left[\bar{Z}[n]; \bar{c}\right]\right) \in \mathbb{R}^{1 \times d}, \tag{10}$$

where $\bar{Z}[n]$ is the n-th row of $\bar{Z}$ in Eq. (8).

Leveraging Answer History So far, the `R3` unit has used the recurrent dialog states to generate the representation of objects for the current question, however, part of historical information has been forgotten after a long conversation. It is therefore necessary to maintain a dynamic history of previous turns, which will be queried when seeking an answer to a new question. This would help to mitigate the co-reference effect by borrowing parts of previous answers into the current answer when deemed relevant.

Recall that the reasoning by the `R3` unit results in each object having a turn-specific representation $Y_{n,j} \in \mathbb{R}^{1 \times d}$ for turn $j = 1, 2, ..., t-1$. Let $A_{j-1} \in \mathbb{R}^{L_{j-1} \times d}$ be the embedding the answer at turn $j-1$. The *object-specific answer embedding* is computed as:

$$a_{n,j-1} = \text{Attn}\left(Y_{n,j}, A_{j-1}, A_{j-1}\right) \in \mathbb{R}^{1 \times d}. \tag{11}$$

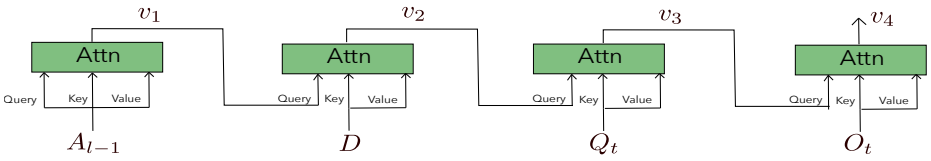

Fig. 4. Four-step Transformer decoder used in the Answer generator for the question at turn t at generation step l. A_{l-1}: embedding matrix of the unfinished utterance of length $l-1$; D: embedding matrix of the dialog history; Q_t : embedding of the question at turn t; O_t: output of `COST`.

This is combined with the object representation and turn position to generate a new object representation:

$$G_{n,j} = \text{Linear}\left([Y_{n,j}, a_{j-1}, p_j]\right) \in \mathbb{R}^{1\times d}, \tag{12}$$

where p_j is a positional encoding feature for each turn.

Thus the collection over turns $\mathfrak{M}_{n,t} = [G_{n,j}]_{j=1}^{t-1} \in \mathbb{R}^{(j-1)\times d}$ is a history of answer-guided representation for object n over previous turns. The answer history enables the retrieval of relevant pieces w.r.t. the current question at turn t:

$$\mathcal{H}_{n,t} = \text{Attn}\left(Y_{n,t}, \mathfrak{M}_{n,t}, \mathfrak{M}_{n,t}\right). \tag{13}$$

This is then augmented with the current object representation $Y_{n,t}$ in Eq. (10) to produce the final form:

$$O_{n,t} = \text{Linear}\left([\mathcal{H}_{n,t}, Y_{n,t}]\right); \qquad O_t = [O_{n,t}]_{n=1}^{N} \in R^{N\times d}. \tag{14}$$

This serves as input for the answer generation module, which we present next.

3.3 Answer Generation

To generate the response utterance, we employ the standard autoregressive framework to produce one word at a time by iteratively estimating the conditional word distribution $P\left(w \mid w_{t,1:l-1}; V, Q_{1:t-1}, A_{1:t-1}, Q_t\right)$ at generation step l. Inspired by the decoders in [25, 26], we use a four-step Transformer decoder, as illustrated in Fig. 4.

Let $A_{t,l} \in \mathbb{R}^{l\times d}$ be the embedding matrix of the unfinished utterance of length l; $D = [Q_{1:t-1}, A_{1:t-1}] \in \mathbb{R}^{L_D\times d}$ the embedding of the dialog history of length L_D (words); $Q_t \in \mathbb{R}^{L_Q\times d}$ the embedding of the question, and $O_t \in \mathbb{R}^{N\times d}$ the output of generated by the `COST` in Eq. (14). The decoder generates a representation v_4 through a step-wise manner:

$$\begin{aligned} v_1 &= \text{Attn}\left(a_l, A_{t,l-1}, A_{t,l-1}\right); & v_2 &= \text{Attn}\left(v_1, D, D\right); \\ v_3 &= \text{Attn}\left(v_2, Q_t, Q_t\right); & v_4 &= \text{Attn}\left(v_3, O_t, O_t\right). \end{aligned} \tag{15}$$

where $a_l = A_{t,l}[l]$, e.g., the last row of $A_{t,l}$. In essence, the sequence of current utterance, the existing dialog, and the current question forms the context to query the object representations O_t.

The retrieved information v_4 is used to generate the next word through the word distributions:

$$P_{vocab} = \text{softmax}\left(\text{Linear}(v_4)\right) \in \mathbb{R}^{1\times N_{vocab}}, \tag{16}$$

$$P_q = \text{Ptr}\left(Q_t, v_4\right) \in \mathbb{R}^{1\times N_{vocab}}, \tag{17}$$

$$P_l = \alpha P_q + (1-\alpha) P_{vocab}, \tag{18}$$

where $\alpha \in (0,1)$, and Ptr is a trainable Pointer Network [41] that "points" to all the tokens in question Q_t that are related to v_4, and P_l is a short-form for $P\left(w \mid w_{t,1:l-1}, V, Q_{1:t-1}, A_{1:t-1}, Q_t; \theta\right)$. Ptr seeks to reuse the relevant question words for the answer; and this is useful when facing rare words or word repetition is required. The gating function α is learnable (detailed in the Supplement).

3.4 Training

Given ground-truth answers $A^*_{1:T}$ of a full conversation of T turns, where $A^*_t = \left(w_1, w_2, ..., w_{L^a_t}\right)$. To make our generator more robust, we also add the log-likelihood of re-generated current turn's question $Q^*_t = \left(w_1, w_2, ..., w_{L^q_t}\right)$ to our loss. The network is trained with the cross-entropy loss w.r.t parameter θ, θ_q :

$$\mathcal{L} = \mathcal{L}(\theta) + \mathcal{L}(\theta_q), \text{where} \tag{19}$$

$$\mathcal{L}(\theta) = \sum_{t=1}^{T} \log P\left(A^*_t \mid V, Q_{1:t-1}, A^*_{1:t-1}, Q_t; \theta\right) \tag{20}$$

$$= \sum_{t=1}^{T}\sum_{l=1}^{L^a_t} \log P_l\left(w_l \mid w_{t,1:l-1}, V, Q_{1:t-1}, A^*_{1:t-1}, Q_t; \theta\right); \text{and} \tag{21}$$

$$\mathcal{L}(\theta_q) = \log P_{vocab}\left(Q^*_t \mid V, Q_t; \theta_q\right) \tag{22}$$

$$= \sum_{l=1}^{L^q_t} \log P_{vocab}\left(w^q_l \mid w^q_{t,1:l-1}, V, Q_t; \theta_q\right); \tag{23}$$

where $P_l(\cdot), P_{vocab}(\cdot)$ are computed in Eqs. (16–18).

4 Experiments

4.1 Experimental Settings

Dataset: We train our proposed method `COST` on the Audio-Visual Scene-Aware Dialogue (AVSD) [1], a large-scale video grounded dialog dataset. The dataset provides text-based dialogs visually grounded on untrimmed action videos from

Table 1. Statistics of the AVSD dataset and the test splits used at DSTC7 and DSTC8.

	Training	Validation	DSTC7 Test	DSTC8 Test
No. Videos	7,659	1,787	1,710	1,710
No. Dialog turns	153,180	35,740	13,490	18,810

the popular Charades dataset [38]. Each annotated dialog consists of 10 rounds of question-answer about both static and dynamic scenes in a video, including objects, actions, and audio content. We benchmark against existing methods on video dialog using two different test splits used at the Seventh Dialog System Technology Challenges (DSTC7) [51] and the Eighth Dialog System Technology Challenges (DSTC8) [19]. See Table 1 for detailed statistics of the AVSD dataset and the two test splits at DSTC7 and DSTC8.

Often state-of-the-art methods in video dialog rely on different sources of information, including visual dynamic scenes, text description in the form of caption/video summary, and audio content. While the textual data containing high-level information of visual content can significantly attribute to the model's performance, they provide shortcuts due to linguistic biases that the models can exploit. As the ultimate goal of the video dialog task is to benchmark if a model can gather the visual clues to produce an appropriate response for a smooth conversation with humans, our experiments deliberately aim at challenging the visual reasoning capability of the models. In particular, we assume that the models only have access to the visual content and dialog history to answer a question at a specific turn while ignoring other additional textual data and audio data used by many other methods [25, 26, 33].

Implementation details: All models are trained by optimizing the multi-label cross-entropy loss over generated tokens using Adam optimizer [20] with cosine learning rate scheduler [31]. At the training stage, we also use the auto-encoder loss function for the current question from [25]. We use a batch size of 128 samples distributed on 4 GPUs and train all the models for 50 epochs. Unless stated otherwise, each attention component in Eq. (15) is composed of a stack of 3 identical attention layers in Eq. (2). For each attention layer, we also use 4 parallel heads as suggested by [40]. Model parameters are selected based on the convergence of validation loss. At inference time, we adopt a beam search algorithm with a beam size of 3 for our answer generator.

Regarding object lives extraction, we strictly follow [7] and extract 30 object sequences per video. We further apply frame sub-sampling at an overall ratio of 4:1 to reduce the computational expensiveness. On average, each object live is composed of 176 time steps. For the context features C used in Eq. (9), we use I3D features [3] similar to other existing methods [26, 22]. Pytorch implementation of our model is available online.[2]

[2] https://github.com/hoanganhpham1006/COST

Evaluation metrics: We adopt the same word-overlap-based metrics, including BLEU, METEOR, ROUGE-L, and CIDEr, as used by [51] to evaluate the effectiveness of the models.

4.2 Comparison against SOTAs

We compare against the state-of-the-art methods, including MTN [25], FA+HRED [33], Student-Teacher [15], SCGA [18] and BiST [26] on both AVSD@DSTC7 and AVSD@DSTC8 test splits. For fair comparisons, all models only make use of the video content and dialog history. Results are shown in Table 2 and Table 3 for DSTC7 and DSTC8 test splits, respectively. In particular, `COST` consistently sets new SOTA performance on all evaluation metrics against existing methods on both test splits. The results strongly demonstrate the efficiency of our object-centric reasoning model with recurrent relational reasoning compared to methods that rely only on holistic visual features such as I3D [3] and ResNeXt [46].

Table 2. Experimental results on the AVSD@DSTC7 test split. All models only have access to video content and dialog history. †Models use visual features other than holistic video features such as I3D or ResNeXt. `COST` use object sequences and I3D features.

Methods	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE-L	CIDEr
FA+HRED [33]	0.648	0.505	0.399	0.323	0.231	0.510	0.843
MTN (I3D) [25]	0.654	0.521	0.420	0.343	0.247	0.520	0.936
MTN (ResNeXt) [25]	0.688	0.55	0.444	0.363	0.260	0.541	0.985
Student-Teacher [15]	0.675	0.543	0.446	0.371	0.248	0.527	0.966
BiST [26]	0.711	0.578	0.475	0.394	0.261	0.550	1.050
SCGA† [18]	0.702	0.588	0.481	0.398	0.256	0.546	1.059
`COST` (Ours)	**0.723**	**0.589**	**0.483**	**0.400**	**0.266**	**0.561**	**1.085**

Table 3. Experimental results on the AVSD@DSTC8 test split.

Methods	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE-L	CIDEr
MTN (I3D) [25]	0.611	0.496	0.404	0.336	0.233	0.505	0.867
MTN (ResNeXt) [25]	0.643	0.523	0.427	0.356	0.245	0.525	0.912
BiST [26]	0.684	0.548	0.457	0.376	0.273	0.563	1.017
SCGA† [18]	0.675	0.559	0.459	0.377	0.269	0.555	1.024
`COST` (Ours)	**0.695**	**0.559**	**0.465**	**0.382**	**0.278**	**0.574**	**1.051**

4.3 Model Analysis

Object-centric Representation Facilitates Space-Time-Dialog Reasoning To better highlight the effectiveness of our object-centric representation for reasoning over space-time to explore the semantic structure in videos, we

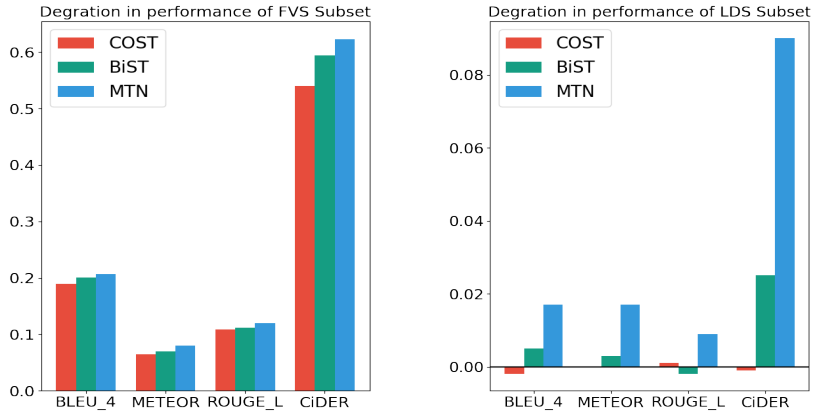

Fig. 5. Performance degradation points from DSTC7 full test splits to (a) FVS subset and (b) LDS subset. The FVS requires fine-grained visual understanding to answer questions while the LDS challenges the capability to handle long-distance dependencies questions of the models. The lower the better. Negative degradation points indicate improvements in performance. `COST` demonstrates its robustness against degradation in performance compared to BiST and MTN on these challenging subsets.

design a subset of the AVSD@DSTC7 test split that poses challenges for models heavily relying on linguistic biases but undervaluing fine-grained visual information. First, we train a variant of the MTN model [25] with all the visual and audio components removed on the AVSD dataset. Next, we only pick dialog turns and their associated videos having BLEU4 score lower than 0.05 on the DSTC7 test split. This eliminates any questions whose answers can be guessed by the linguistic biases. Eventually, we obtain a subset of 1,062 videos with 8,782 associated dialog turns referred to as *Fine-grained Visual Subset (FVS)*. We evaluate the degradation of `COST` and the current state-of-the-art BiST and MTN on the FVS subset and report the results in Fig. 5(a). As shown, `COST` is more robust to questions in FVS while BiST, MTN struggle as it is degraded by larger margins on all the evaluation metrics. The results are clearly evident the benefits of the fine-grained video understanding brought by the object-centric representation compared to the holistic video representation used by MTN, BiST.

Recurrent Modeling Supports Long-Distance Dependencies One of the main advantages of `COST` against SOTA methods is that it maintains a *recurrent system of dialog states,* which offers the better capability of handling long-distance dependencies questions. These questions require models to maintain and retrieve information appearing far in earlier turns. Methods without an explicit mechanism to propagate long-distance dependencies would struggle to generalize. In order to verify this, we design another subset of the AVSD@DSTC7 test split where we only collect questions at turns greater than 3. This results in a subset of 950 videos and 11,210 associated dialog turns. We refer to this

Table 4. Ablation studies on AVSD@DSTC7 test split. AHR: Answer history retrieval.

Effects of	BLEU1	BLEU2	BLEU3	BLEU4	METEOR	ROUGE-L	CIDEr
recurrent design							
w/o recurrence	0.710	0.577	0.471	0.388	0.261	0.554	1.041
object modeling							
w/o object-centric	0.709	0.574	0.467	0.385	0.260	0.553	1.042
attention in AHR							
w/o self-attention	0.719	0.584	0.477	0.395	0.263	0.558	1.062
pointer networks							
w/o pointer	0.715	0.583	0.477	0.394	0.260	0.557	1.045
Full model	**0.723**	**0.589**	**0.483**	**0.400**	**0.266**	**0.561**	**1.085**

as *Long-distance Dependencies Subset (LDS)*. Fig. 5(b) details the degradation in performance of COST, BiST, and MTN on the LDS. As shown, while COST achieves slight improvements in performance (negative degradation) thanks to it recurrent design, BiST, MTN experience consistent losses across the evaluation metrics. The results verify our hypothesis that maintaining the recurrent dialog object states is beneficial for handling long-distance dependencies between turns.

4.4 Ablation Studies

We conduct an extensive set of ablation studies to examine the contributions of each components in COST (See Table 4). These include ablating the recurrent design of COST, the use of object-centric video representation, and the use of self-attention layers for answer history retrieval for answer generation. We find that ablating any of these components would take the edge off the model's performance. The results are consistent with our analysis in Sec. 4.3 on the crucial effects of our recurrent design and the object-centric modeling by COST on the overall performance. We detail the effects as follows.

Effect of recurrent design: We remove the GRU in Eq. (6) and use query-specific object representation outputs of Eq. (5) as direct input to compute the adjacency matrix $\mathcal{K}_t$ in Eq. (7). By doing this, we ignore the effect of past dialog states on the current turn. As clearly seen, the model's performance considerably degrades on all evaluation metrics.

Effect of object-centric modeling: In this experiment, we use only the context feature C (I3D features) and remove all the effects of the object representation. This leads to performance degradation by nearly 2% on all BLEU scores. The fine-grained object-centric representation clearly has an effect on improving the understanding of hidden semantic structure in video.

Effect of self-attention-based answer history retrieval: This experiment ablates the role of prior generated answer tokens on information retrieval as in Eq. (13). We instead use the turn-specific visual representation as a direct input for the answer generator. The results show that this slightly affects the overall performance of the model.

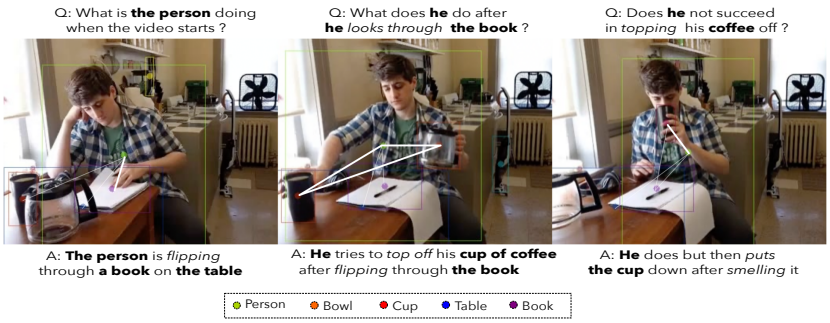

Fig. 6. Visualization of the question-specific interaction matrices between objects $\mathcal{K}_t$ of Eq. (7). Each frame/question pair (left to right) represents a dialog turn, where the frame is selected to reflect the moment being queried. The visibility of edges denotes the relevance of visual objects' relations to the questions and desired answers. Detected objects are named by Faster RCNN. `COST` succeeds in constructing turn-specific graphs of relevant visual objects that facilitate answering questions. Sample is taken from DSTC7 test split - Best viewed in colors.

Effect of pointer networks for answer generation: We remove the use of the pointer networks as in Eqs. (17 and 18) during answer generation. The results show that the removal of the pointer slightly hurt the model's performance.

4.5 Qualitative Analysis

We visualize a representative example from the AVSD@DSTC7 test split as a showcase to analyze the internal operation of the proposed method `COST`. We present the question-induced interaction matrix as it is a crucial component of our model design in Eq. (7). Fig. 6 presents the evolution of the relationships between objects in video from turn to turn (left to right). `COST` succeeds in constructing turn-specific graphs of relevant visual objects that reflect the relationships of interest by respective questions and answers. The interpretability and strong quantitative results (Sec. 4.2 and 4.3) by `COST` are evident to the appropriateness of the object-centric representation towards solving video dialog task.

5 Conclusion

Addressing the highly challenging task of video dialog, we have proposed `COST`, a new *recurrent object-centric* system that learns to reason through multiple dialog turns, object dynamics, and interactions over space-time in video. `COST` maintains and tracks dialog states over the course of conversation. It treats objects in video as primitive constructs whose "lives" and relations to others throughout the video are dynamically examined through the guidance of the questions, conditioned on the *dialog states* and *answer history*. Tested on the challenging AVSD dataset, `COST` demonstrates its effectiveness against state-of-the-art models.

References

1. Alamri, H., Cartillier, V., Das, A., Wang, J., Cherian, A., Essa, I., Batra, D., Marks, T.K., Hori, C., Anderson, P., et al.: Audio visual scene-aware dialog. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7558–7567 (2019)
2. Baradel, F., Neverova, N., Wolf, C., Mille, J., Mori, G.: Object level visual reasoning in videos. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 105–121 (2018)
3. Carreira, J., Zisserman, A.: Quo vadis, action recognition? a new model and the kinetics dataset. In: CVPR. pp. 6299–6308 (2017)
4. Chao, Y.W., Vijayanarasimhan, S., Seybold, B., Ross, D.A., Deng, J., Sukthankar, R.: Rethinking the faster r-cnn architecture for temporal action localization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1130–1139 (2018)
5. Cho, K., Van Merriënboer, B., Bahdanau, D., Bengio, Y.: On the properties of neural machine translation: Encoder-decoder approaches. arXiv preprint arXiv:1409.1259 (2014)
6. Chu, Y.W., Lin, K.Y., Hsu, C.C., Ku, L.W.: Multi-step joint-modality attention network for scene-aware dialogue system. arXiv preprint arXiv:2001.06206 (2020)
7. Dang, L.H., Le, T.M., Le, V., Tran, T.: Hierarchical object-oriented spatio-temporal reasoning for video question answering. IJCAI (2021)
8. Das, A., Kottur, S., Gupta, K., Singh, A., Yadav, D., Lee, S., Moura, J.M., Parikh, D., Batra, D.: Visual Dialog. IEEE Transactions on Pattern Analysis and Machine Intelligence **41**(5), 1242–1256 (2019). https://doi.org/10.1109/TPAMI.2018.2828437
9. Desta, M.T., Chen, L., Kornuta, T.: Object-based reasoning in vqa. In: 2018 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 1814–1823. IEEE (2018)
10. Gao, S., Sethi, A., Agarwal, S., Chung, T., Hakkani-Tur, D.: Dialog state tracking: A neural reading comprehension approach. In: Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue. pp. 264–273 (2019)
11. Geman, D., Geman, S., Hallonquist, N., Younes, L.: Visual turing test for computer vision systems. Proceedings of the National Academy of Sciences **112**(12), 3618–3623 (2015)
12. Geng, S., Gao, P., Chatterjee, M., Hori, C., Le Roux, J., Zhang, Y., Li, H., Cherian, A.: Dynamic graph representation learning for video dialog via multi-modal shuffled transformers. In: Proc. AAAI Conference on Artificial Intelligence (2021)
13. Hong, Y., Wu, Q., Qi, Y., Rodriguez-Opazo, C., Gould, S.: Vln bert: A recurrent vision-and-language bert for navigation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 1643–1653 (June 2021)
14. Hori, C., Alamri, H., Wang, J., Wichern, G., Hori, T., Cherian, A., Marks, T.K., Cartillier, V., Lopes, R.G., Das, A., et al.: 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. IEEE (2019)
15. Hori, C., Cherian, A., Marks, T.K., Hori, T.: Joint student-teacher learning for audio-visual scene-aware dialog. In: INTERSPEECH. pp. 1886–1890 (2019)
16. Huang, D., Chen, P., Zeng, R., Du, Q., Tan, M., Gan, C.: Location-aware graph convolutional networks for video question answering. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 34, pp. 11021–11028 (2020)

17. Kalogeiton, V., Weinzaepfel, P., Ferrari, V., Schmid, C.: Action tubelet detector for spatio-temporal action localization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 4405–4413 (2017)
18. Kim, J., Yoon, S., Kim, D., Yoo, C.D.: Structured co-reference graph attention for video-grounded dialogue. AAAI (2021)
19. Kim, S., Galley, M., Gunasekara, C., Lee, S., Atkinson, A., Peng, B., Schulz, H., Gao, J., Li, J., Adada, M., et al.: The eighth dialog system technology challenge. arXiv preprint arXiv:1911.06394 (2019)
20. Kingma, D., Ba, J.: Adam: A method for stochastic optimization. International Conference on Learning Representations (ICLR) (2014)
21. Kottur, S., Moura, J.M., Parikh, D., Batra, D., Rohrbach, M.: Visual coreference resolution in visual dialog using neural module networks. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 153–169 (2018)
22. Le, H., Chen, N.F.: Multimodal transformer with pointer network for the dstc8 avsd challenge. arXiv preprint arXiv:2002.10695 (2020)
23. Le, H., Chen, N.F., Hoi, S.C.: Learning reasoning paths over semantic graphs for video-grounded dialogues. arXiv preprint arXiv:2103.00820 (2021)
24. Le, H., Hoi, S., Sahoo, D., Chen, N.: End-to-end multimodal dialog systems with hierarchical multimodal attention on video features. In: DSTC7 at AAAI2019 workshop (2019)
25. Le, H., Sahoo, D., Chen, N., Hoi, S.: 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 (2019)
26. Le, H., Sahoo, D., Chen, N., Hoi, S.C.: 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 (2020)
27. Le, T.M., Le, V., Venkatesh, S., Tran, T.: Dynamic language binding in relational visual reasoning. In: IJCAI. pp. 818–824 (2020)
28. Le, T.M., Le, V., Venkatesh, S., Tran, T.: Hierarchical conditional relation networks for video question answering. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 9972–9981 (2020)
29. Lee, H., Yoon, S., Dernoncourt, F., Kim, D.S., Bui, T., Jung, K.: Dstc8-avsd: Multimodal semantic transformer network with retrieval style word generator. arXiv preprint arXiv:2004.08299 (2020)
30. Lin, K.Y., Hsu, C.C., Chen, Y.N., Ku, L.W.: Entropy-enhanced multimodal attention model for scene-aware dialogue generation. arXiv preprint arXiv:1908.08191 (2019)
31. Loshchilov, I., Hutter, F.: Sgdr: Stochastic gradient descent with warm restarts. International Conference on Learning Representations (2017)
32. Murahari, V., Batra, D., Parikh, D., Das, A.: Large-scale pretraining for visual dialog: A simple state-of-the-art baseline. In: European Conference on Computer Vision. pp. 336–352. Springer (2020)
33. Nguyen, D.T., Sharma, S., Schulz, H., Asri, L.E.: From film to video: Multi-turn question answering with multi-modal context. DSTC7 workshop at AAAI 2019 (2019)
34. Pan, B., Cai, H., Huang, D.A., Lee, K.H., Gaidon, A., Adeli, E., Niebles, J.C.: Spatio-temporal graph for video captioning with knowledge distillation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10870–10879 (2020)

35. Sanabria, R., Palaskar, S., Metze, F.: CMU Sinbad's submission for the DSTC7 AVSD challenge. In: DSTC7 at AAAI2019 workshop. vol. 6 (2019)
36. Seo, P.H., Lehrmann, A., Han, B., Sigal, L.: Visual reference resolution using attention memory for visual dialog. arXiv preprint arXiv:1709.07992 (2017)
37. Serban, I., Sordoni, A., Lowe, R., Charlin, L., Pineau, J., Courville, A., Bengio, Y.: A hierarchical latent variable encoder-decoder model for generating dialogues. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 31 (2017)
38. Sigurdsson, G.A., Varol, G., Wang, X., Farhadi, A., Laptev, I., Gupta, A.: Hollywood in homes: Crowdsourcing data collection for activity understanding. In: ECCV. pp. 510–526. Springer (2016)
39. Spelke, E.S., Kinzler, K.D.: Core knowledge. Developmental science **10**(1), 89–96 (2007)
40. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Advances in neural information processing systems. pp. 5998–6008 (2017)
41. Vinyals, O., Fortunato, M., Jaitly, N.: Pointer networks. Advances in neural information processing systems **28** (2015)
42. Wang, X., Gupta, A.: Videos as space-time region graphs. In: Proceedings of the European conference on computer vision (ECCV). pp. 399–417 (2018)
43. Wang, Y., Joty, S., Lyu, M.R., King, I., Xiong, C., Hoi, S.C.: Vd-bert: A unified vision and dialog transformer with bert. arXiv preprint arXiv:2004.13278 (2020)
44. Wojke, N., Bewley, A., Paulus, D.: Simple online and realtime tracking with a deep association metric. In: ICIP. pp. 3645–3649. IEEE (2017)
45. Xie, H., Iacobacci, I.: Audio visual scene-aware dialog system using dynamic memory networks. In: DSTC8 at AAAI2020 workshop (2020)
46. Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K.: Aggregated residual transformations for deep neural networks. In: CVPR. pp. 1492–1500 (2017)
47. Xie, S., Sun, C., Huang, J., Tu, Z., Murphy, K.: Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In: Proceedings of the European conference on computer vision (ECCV). pp. 305–321 (2018)
48. Yang, Z., Garcia, N., Chu, C., Otani, M., Nakashima, Y., Takemura, H.: Bert representations for video question answering. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1556–1565 (2020)
49. Yeh, Y.T., Lin, T.C., Cheng, H.H., Deng, Y.H., Su, S.Y., Chen, Y.N.: Reactive multi-stage feature fusion for multimodal dialogue modeling. arXiv preprint arXiv:1908.05067 (2019)
50. Yi, K., Gan, C., Li, Y., Kohli, P., Wu, J., Torralba, A., Tenenbaum, J.B.: Clevrer: Collision events for video representation and reasoning. arXiv preprint arXiv:1910.01442 (2019)
51. Yoshino, K., Hori, C., Perez, J., D'Haro, L.F., Polymenakos, L., Gunasekara, C., Lasecki, W.S., Kummerfeld, J.K., Galley, M., Brockett, C., et al.: Dialog system technology challenge 7. arXiv preprint arXiv:1901.03461 (2019)
52. Zeng, R., Huang, W., Tan, M., Rong, Y., Zhao, P., Huang, J., Gan, C.: Graph convolutional networks for temporal action localization. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7094–7103 (2019)
53. Zhuge, M., Gao, D., Fan, D.P., Jin, L., Chen, B., Zhou, H., Qiu, M., Shao, L.: Kaleido-bert: Vision-language pre-training on fashion domain. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12647–12657 (June 2021)