Spatio-temporal Relational Reasoning for Video Question Answering

Gursimran Singh
http://gursimar.github.io/
Leonid Sigal
https://www.cs.ubc.ca/~lsigal/
James J. Little
https://www.cs.ubc.ca/~little/

Department of Computer Science
University of British Columbia
Vancouver, Canada

Abstract

Video question answering is the task of automatically answering questions about videos. Among query types which include identification, localization, and counting, the most challenging questions enquire about relationships among different entities. Answering such questions, and many others, require modeling relationships between entities in the spatial domain and evolution of those relationships in the temporal domain. We argue that current approaches have limited capacity to model such long-range spatial and temporal dependencies. To address these challenges, we present a novel spatio-temporal reasoning neural module which enables modeling complex multi-entity relationships in space and long-term ordered dependencies in time. We evaluate our module on two benchmark datasets which require spatio-temporal reasoning: TGIF-QA and SVQA. We achieve state-of-the-art performance on both datasets. More significantly, we achieve substantial improvements on some of the most challenging question types, like counting, which demonstrate the effectiveness of our proposed spatio-temporal relational module.

1 Introduction

Understanding video content is an important topic in computer vision. Relevant tasks include Activity Recognition [3, 22], Temporal Action Localization [20, 24], and, more recently, Video Question Answering [2, 5, 7, 26]. Video Question Answering (aka VideoQA) is the task of answering natural language questions about videos. It is, arguably, the most challenging among video tasks since it may contain a myriad of queries, including those encompassing other video understanding tasks. For instance, simpler questions, similar to ImageQA, involve attribute identification in a single frame of a video [15, 17]. More complex questions, similar to activity recognition and localization, require looking at multiple frames in a local temporal region [5, 17]. The most complex questions require recognizing activities across time, counting them or reasoning about their temporal order [5, 21].

A generic VideoQA algorithm must learn to ground objects of interest in video and reason about their interactions in both the spatial and temporal domains. Conceptually, a VideoQA algorithm can be broken into three different sub-tasks. First, the algorithm should

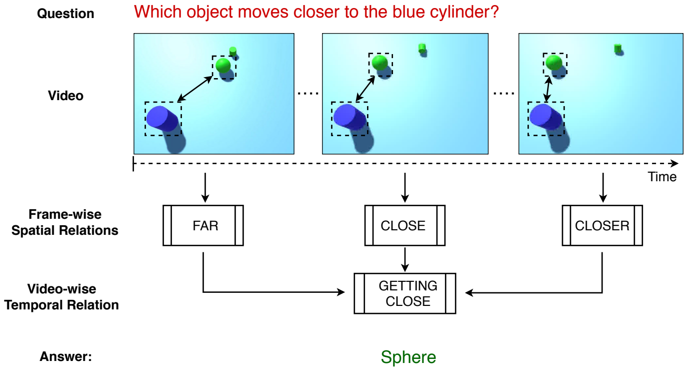

Question
Which object moves closer to the blue cylinder?
Video
Time
Frame-wise Spatial Relations
FAR
CLOSE
CLOSER
Video-wise Temporal Relation
GETTING CLOSE
Answer:
Sphere

Figure 1: The figure shows a question and selected frames of a video in SVQA dataset [21]. We show individual spatial-relations among relevant objects, which are `far`, `close` and `closer`. The change of spatial-relations over time corresponds to the temporal-relation which is `getting close`. Having observed these temporal-relations among object-pairs, the algorithm can infer the answer – `Sphere`.

understand the intent of the query from the natural language description of the question. Second, it should compute relationships which are relevant to the query in the spatial domain. Finally, it should reason how these spatial relations evolve in the temporal domain. For instance, consider the question and the sequence of frames in Figure 1. In order to answer, the algorithm considers spatial relations among all possible object-pairs in each frame individually. In case of the blue cylinder and sphere, these are calculated as `far` in the first frame, `close` and `closer` in the subsequent frames. Having done that, it needs to reason how these spatial relationships change in the temporal domain. This leads to identifying the correct interpretation that cylinder and sphere are `getting close`. Having observed these spatio-temporal relationships among all possible object-pairs, the algorithm can figure out the correct answer which is sphere in this case.

Traditional approaches use 3D CNNs [13, 24], LSTMs [10, 26], or attention [2, 5, 25] to model such relationships. Although successful, they are limited in capacity. For instance, 3D CNNs are useful for identifying local spatio-temporal action events, demonstrated by success in Activity Recognition datasets, however, they struggle in modeling long-range temporal relationships [23, 27]. Similarly, LSTM-based approaches, although known to do well in long-range text sequences, struggle to model videos [2, 9]. This is because, unlike text, videos contain longer and information-richer sequences of spatial data, which LSTMs, or their spatial-attention variants, cannot model in a natural and effective way. More importantly, these network architectures do not provide an effective prior and need to learn relational reasoning from scratch which is inefficient and data-hungry [18].

In this work, we leverage and extend relational networks [14] to model spatio-temporal relationships in videos. Previously, relational reasoning has been used effectively in image question answering [18] and activity recognition [27]. However, it was limited to either the spatial or the temporal domain individually. Inspired by these two works, we present spatio-temporal relational networks which can perform joint relational reasoning in both spatial and temporal domains. Our **contributions** are two-fold: **1)** We present a novel general-purpose neural network module which acts as an effective prior for spatio-temporal relational reasoning in videos. Our network models both spatio-temporal relations (capturing object-interactions) and action-dynamics (capturing how individual objects change over time). To our knowledge, this is the first attempt to perform joint spatio-temporal reasoning using relational networks. **2)** We show the effectiveness of our proposed Spatio-Temporal Relational Network on a variety of VideoQA tasks, which include both real-world (TGIF-QA) and synthetic (SVQA) datasets. Also, we show substantial improvement in the challenging counting task that requires capturing spatio-temporal dynamics in different parts of a video. To best of our knowledge, this is the first attempt to approach VideoQA using relational networks.

2 Related work

2.1 Visual Question Answering

Image question answering [1] is the task of answering queries about images. Typical questions focus on identifying attributes, counting objects or reasoning about their spatial-relations. Video question answering, apart from spatial-queries, also focuses on queries which require spatio-temporal reasoning. These may include identifying a single activity spanning a few contiguous frames, or more generally multiple such activities and inferring relationships between them. The earliest approaches [10, 25, 26] used LSTMs to encode video and text representation, and leveraged temporal attention to selectively attend to important frames in a video. These approaches model temporal reasoning through attention but lack spatial reasoning. Jang *et al*. [5] used both spatial and temporal attention, and utilized motion and appearance features. Although this allowed spatio-temporal reasoning but struggled to model long-range temporal dependencies. To address this problem, Song *et al*. [21] proposed a more granular spatial-attention and a modified-GRU incorporating a temporally-attended hidden state transfer. Similarly, other approaches used memory networks [2, 7, 11] to handle long-term dependencies. Gao *et al*. [2] used a co-memory attention that utilized cues from both appearance and motion, and used conv-deconv features to build multi-level contextual facts. Li *et al*. [8] used a self-attention based technique to exploit global dependencies among words of a question and frames of a video. Although, modified-GRU [21], co-attention [2] and self-attention [8] performed better than LSTM-based approaches, however, they still had to learn relational reasoning from scratch using supervised data. Our approach takes a different route, and instead uses relational networks which provide a prior for relational reasoning, and hence, outperforms the above techniques.

2.2 Relational reasoning

Relational reasoning is the ability to reason about relationships among entities. It is central to general intelligent behavior and is essential to answer complex questions in VQA tasks. Roposo *et al*. [14] and Santoro *et al*. [18] introduced relational networks and showed their effectiveness on scene description data and image question answering, respectively. They demonstrated that even the powerful CNNs or MLPs struggle to solve questions which require relational reasoning. However, when augmented with relational networks (RNs), they achieve superhuman performance even in complex datasets like CLEVR [6]. Zhou *et al*. [27] extended relational networks to temporal domain and introduced Temporal Relational Networks (TRNs). TRN computes temporal relationships among video frames and achieved the state-of-the-art result in Activity Recognition datasets. Our work is different from the above networks in two main ways. First, we show the effectiveness of relational networks in a more challenging setting: Video Question Answering, where interesting events may occur at different parts of the video. Second, and more importantly, we model joint spatio-

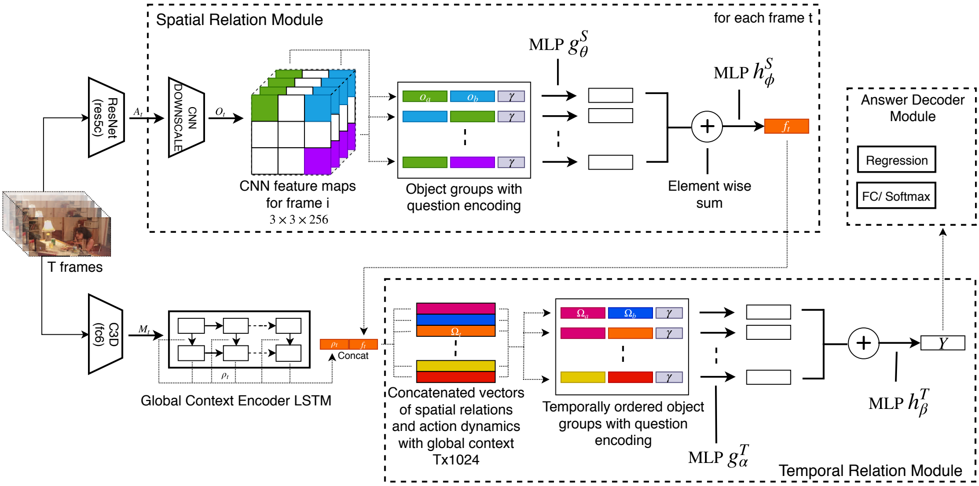


Figure 2: **Spatio-temporal Relational Network architecture**. The Spatial Relations Module (top) models arbitrary spatial-relations among all possible groups of objects for each frame individually. The Global Context Encoder LSTM (bottom left) models the action-dynamics with global context at time t. The concatenated output of these modules is then fed to the Temporal Relations Module (bottom right) which computes temporal relations among a temporally-ordered group of frames. Notice that, for simplicity, we have shown object-groups as pairs, however, in general they can be more than two.

temporal relationships, unlike [18] and [27] which work either in spatial or temporal domains individually. Similar to ours, a recent work by Wang *et al.* [23] models spatio-temporal relationships in Activity Recognition. However, they use Graph Neural Networks [19] which depend on structured data like bounding boxes extracted using Region Proposal Networks (RPNs) [16]. In comparison, we use relation networks which are simpler and flexible to work with unstructured data like raw RGB values or CNN outputs.

3 Approach

The input to our model consists of a video, a question and optionally answer-options (only in the case of multiple-choice questions). We extract appearance ($\{A^t\}_{t=1}^T \in \mathbb{R}^{7\times7\times2048}$) and motion ($\{M^t\}_{t=1}^T \in \mathbb{R}^{4096}$) features using ResNet-152 [4] ($res5c$) and C3D [22] ($fc6$), respectively; T is the sequence length. The overall architecture of our Spatio Temporal Relational Network (STRN) is shown in Figure 2. There are three main components: (a) Spatial Relation Module (SRM), (b) Global Context Encoder LSTM (GCE), and (c) Temporal Relations Module (TRM). The Spatial Relation Module takes appearance features ($\{A^t\}_{t=1}^T$) as input and computes spatial relations among various objects. This can be seen as modeling object interactions in each frame individually. The Global Context Encoder LSTM takes motion features ($\{M^t\}_{t=1}^T$) as input and captures action-dynamics with global context at time t. Finally, the Temporal Relation Module takes the concatenated SRM-encoding (f_t) and the GCE-encoding (ρ_t) as input and computes how the spatial-relations and action-dynamics change over time. This corresponds to modeling temporal changes in both the interactions among different objects and the motion-dynamics of individual objects. The output (Y) of the TRM is passed to the Answer Decoder Module whose exact form depends on the specific question answering task (Section 3.2).

3.1 Spatio-temporal Relational Network

Inspired by relational networks [18, 27], we encode the ability to model spatio-temporal relationships right in the formulation of STRN. Hence, it acts as an effective prior for situations which require joint reasoning over both spatial and temporal domains. The input consists of an ordered temporal sequence of spatial frame-descriptors $\{O^t\}_{t=1}^T$, where each O^t contains L spatial object-representations $\{o_x\}_{x=1}^L$. In general, the spatial frame-descriptors can be any representation of interest. It can be structured in the case of bounding boxes or unstructured in the case of CNN feature maps. In this work, we use CNN feature maps $(\{O^t\}_{t=1}^T \in \mathbb{R}^{3\times3\times256})$ which we obtain from ResNet-152 features $(\{A^t\}_{t=1}^T \in \mathbb{R}^{7\times7\times2048})$ using a Downscale-CNN layer (see Figure 2). Given $\{O^t\}_{t=1}^T$, we define the basic-STRN as a composite function below:

$$\text{STRN\_B}(O) = h_\beta^T\left(\sum_{a<b} g_\alpha^T(f_a, f_b)\right) \tag{1}$$

$$f_t = h_\phi^S\left(\sum_{a,b} g_\theta^S(o_a, o_b)\right) \tag{2}$$

Equation (2) corresponds to SRM and is responsible for computing spatial relations (f_t) for each $O^t = \{o_x\}_{x=1}^L$. In particular, spatial-relation function g_θ^S infers whether and how the two inputs are related to each other. The relations are computed for all possible input combinations $o_a, o_b \in \{O^t\}$. The individual object-object relations are then agglomerated and reasoned over by the function h_ϕ^S. In a similar way, Equation (1) corresponds to TRM and computes the temporal relations among a sequence of ordered inputs $\{f_t\}_{t=1}^T$ obtained from $\{O^t\}_{t=1}^T$ using Equation (2). The temporal-relation function g_α^T computes the individual frame-frame relations, which are agglomerated, and reasoned over by the function h_β^T. Hence, the combination of Equations (1) and (2) models the temporal relations among the spatial relations, achieving spatio-temporal relational reasoning. In other words, STRN_B models the interactions among objects and how they evolve over time. Following previous work [18, 27], we use fully-connected layers to represent the functions $g_\alpha^T, h_\beta^T, g_\theta^S$, and h_ϕ^S, which are parameterized by α, β, θ, and ϕ.

Capturing Action Dynamics: In STRN_B, we used the f_i's in Equation (1) to be spatial relations, which helped us model evolving object interactions. However, apart from interactions, some queries may also inquire about changes in motion (or appearance) of individual objects. In order to capture action-dynamics, we leverage motion features $(\{M^t\}_{t=1}^T \in \mathbb{R}^{4096})$, which represent course motion information corresponding to each object in a video [22]. However, both C3D and Flow features are known to encode only short-term temporal information [23]. Hence, we additionally make use of a Global Context Encoder LSTM to capture long-term global context. Instead of using only the SRM-encoding, we consider the concatenation of the SRM-encoding (f_t) and the GCE_LSTM-encoding (ρ_t) while computing temporal relations. The resultant model, STRN_GC, captures both the interactions among object-groups and the action-dynamics of individual objects with global-context.

$$\text{STRN\_GC}(O) = h_\beta^T\left(\sum_{a<b} g_\alpha^T(\Omega_a, \Omega_b)\right) \tag{3}$$

$$\rho_t = \text{LSTM}(M^t, \rho_{t-1}) \tag{4}$$

where Ω_t is obtained by concatenating f_t (Eq. 2) and ρ_t (Eq. 4), ρ_t is the hidden state of the Global Context Encoder LSTM at time t and M^t is the t^{th} motion embedding.

Conditioning and Multi-scale Relations: For tasks like video question answering, different questions may require different types of relations. Hence, we model dependence on questions by conditioning the relation-functions g^T_α, and g^S_θ to obtain query-specific variants. For instance, functions $g^T_\alpha(f_a, f_b)$, and $g^S_\theta(o_a, o_b)$ transform to $g^T_\alpha(f_a, f_b, \gamma)$, and $g^S_\theta(o_a, o_b, \gamma)$, where γ is the question encoding obtained through a text-encoder LSTM, similar to one used in Jang *et al.* [5].

Additionally, inspired by multi-scale temporal relational networks [27], instead of computing relations among only two possible frames/objects at a time, we generalize the relation-functions $g^T_\alpha(f_a, f_b, \gamma)$, and $g^S_\theta(o_a, o_b, \gamma)$ to consider multiple frames/objects: $g^T_\alpha(f_a, f_b, ..f_m, \gamma)$, and $g^S_\theta(o_a, o_b, ..o_n, \gamma)$, for m frames and n objects, respectively. Then, we consider multiple relation-functions each specializing to capture relationships for a given value of (m,n) frames/ objects at a time. This allows modelling relationships at multiple scales. We define the **M** multi-scale, **N** multi-object Spatio-Temporal Relational Network (STRN) as:

$$\text{STRN\_S}(O,m,n) = h^T_\beta\left(\sum_{a<b..<m} g^T_\alpha(\Omega_a, \Omega_b, ..\Omega_m, \gamma)\right) \tag{5}$$

$$f_t = h^S_\phi\left(\sum_{a,b..n} g^S_\theta(o_a, o_b, ..o_n, \gamma)\right) \tag{6}$$

$$\text{STRN}(O,M,N) = \sum_{m=2,n=2}^{M,N}\Big(\text{STRN\_S}(O,m,n)\Big) \tag{7}$$

Each STRN_S(O,m,n) in Equation (5) computes relationships among a given value of m temporal-objects and n spatial-objects and has its own h and g functions. Additionally, we consider the temporal-relation function, g^T_α (from Eq. 3 and 4), which captures both object-interactions and action-dynamics. STRN(O,M,N) in Equation (7) accumulates relationships from multiple STRN_S(O,m,n) for all values of (m,n), ranging from $(2,2)$ to (M,N). Hence, we obtain the M-multi-scale and N multi-object Spatio-Temporal Relation Network (STRN) which we use as our final model.

3.2 Answer Decoder Module

Depending on the question type (Section 4.1) we have three different types of modules:

Multiple-choice: We define a linear regression function which takes the TRM-encoding (Y) as input and outputs a real-valued score for each multiple-choice answer-candidate: $s = W^T_{MC}Y$, W_{MC} are model parameters. To optimize, we use hinge-loss: $max(0, 1 + s_n - s_p)$, where s_p and s_n are scores of the correct and incorrect answer, respectively. We use this decoder for repeating action and state transition tasks in the TGIF-QA dataset.

Open-ended numbers: Similar to above, we define a linear regression function: $s = [W^T_N Y + b]$, where [.] denotes rounding, Y is the TRM-encoding, W_N are model parameters and b is the bias. We optimize the network using l_2 loss between the ground truth and the predicted value. This decoder is used for the repetition count task in the TGIF-QA dataset.

Open-ended word: We define a linear classifier which selects an answer from a vocabulary V: $o = softmax(W^T_w Y + b)$, where W_w are model parameters and b is the bias. We use cross-entropy loss and the final answer is obtained using: $y = argmax_{y \in V}(o)$. We use this decoder for the SVQA dataset and also for the FrameQA task in the TGIF-QA dataset.

3.3 Implementation details

We implement our model and design our experiments using PyTorch. Following previous work [2, 6, 8, 21], we train separate models for each task of the TGIF-QA dataset and one model for the entire SVQA dataset. Similarly, we set the maximum number of uniformly-sampled frames in a video to 35. To encode text, we use the 300D Glove [12] word embeddings and take the output of the final layer of a text-encoder LSTM as the question-encoding (taken as answer-encoding in case of multiple-choice questions). Both the text-encoder and global-context-encoder are two layer LSTMs with 512 hidden units. In all our experiments, we use a batch size of 64. We train our networks in an end-to-end fashion using Adam optimizer with an initial learning rate of 0.001. Wherever applicable, we use a dropout of 0.2. The functions $g^T_\alpha, h^T_\beta, g^S_\theta, h^S_\phi$ in Equation (5) and (6) are fully-connected networks with 2, 1, 2, 2 layers and and 256, 256, 256, 256 hidden units, respectively. In Equation (7), we choose M=10 different scales, which means we consider 2-10 frames at a time while computing temporal relations. Since the number of possible combinations of frames can be large, we follow Zhou *et al.* [27] and randomly sample $S = 3$ possible frame-sequences for each separate scale. Similarly, we choose N=3, which means we consider 2-3 spatial-objects at a time while computing spatial relations. We do not subsample spatial-relations but we downscale the appearance-features from $\{A^t\}_{t=1}^T \in \mathbb{R}^{7\times7\times2048}$ to $\{O^t\}_{t=1}^T \in \mathbb{R}^{3x3x256}$ using a Downscale-CNN having (384, 192, 256) filters and (2,3,2) kernels. The code and pre-trained models are available at https://github.com/gursimar/STRN.

4 Evaluation

4.1 Dataset

TGIF-QA [6] is a large-scale dataset containing 165K QA pairs collected from 71K real-world animated Tumblr GIFs. The questions are categorized into four separate tasks. 1) Repeating Action (Action) aims to name the event that happened a specific number of times in the video. This is a multiple-choice task where the correct answer is one of the five available options (Fig 3a). 2) State Transition (Trans), similarly, is a multiple-choice task with five options. Questions ask about state transitions like facial expressions (from happy to sad), among others (Fig 3b). 3) FrameQA is an open-ended task which, similar to image-QA, can be answered by looking at one of the "appropriate" frames in the video. However, the range of possible answers span the entire vocabulary (Fig 3c). 4) Repetition Count (Count) aims to count the number of times a given event happens in the video. This is an open-ended task and answers lie in a range of integers: 0 to 10 (Fig 3d).

SVQA [21] is synthetically generated dataset designed to control and minimize language biases in existing videoQA datasets. It contains 120K questions asked on 12K videos with moving objects like sphere, cylinder or cube (Fig 1). Similar to FrameQA, answers span the entire vocabulary. Questions are compositional and require a series of reasoning steps (like comparison and arithmetic) in both spatial and temporal domains. Since the exact train-val-test subsets of the SVQA dataset are not readily available, we randomly sample a new split similar to Table 1 of Song *et al.* [21]. In comparison to TGIF-QA, it contains more complex questions requiring more elaborate spatio-temporal reasoning. However, unlike real-world GIFs in TGIF-QA, it contains perceptually-simpler scenes consisting of a few synthetic objects. These two datasets are well suited for our task because they contain well formed questions that require complex spatio-temporal reasoning.

4.2 Experiments

This section outlines the details of our experiments. First, we compare our method with the state-of-the-art baselines on both the TGIF-QA and the SVQA dataset. Then, we show an ablation study which demonstrates the effectiveness of joint spatio-temporal relational reasoning. Following previous work [2, 5, 8, 21], we use classification accuracy (ACC) as an evaluation metric for all tasks of the SVQA dataset and also the Trans, Action and FrameQA tasks of the TGIF-QA dataset. For the Count task of the TGIF-QA dataset, we use Mean Squared Error (MSE) between the predicted value and the ground truth value as an evaluation metric.

4.2.1 Comparison with the state-of-the-art methods

TGIF-QA Dataset: We summarize the results in Table 1. At the very top, *Random* and *Text* correspond to selecting an answer randomly and learning a model without any visual input, respectively. In the next four lines, we show results obtained using the image-VQA based baselines, which either mean-pool the video features (aggr) or average the results (avg). The rest of the results correspond to videoQA methods (refer to Section 2.1 for a comparison). The letters inside the brackets correspond to the features used to train the model: R means ResNet, C means C3D and F means Flow. The last three rows show the result of our models. Our STRN model outperforms all other approaches on all tasks which require spatio-temporal reasoning: **Action (2.74%), Trans (2.4%)** and **Count (4.63%)** ($\frac{4.10-3.91}{4.10}$) by a significant margin. On the other task, FrameQA, which can be answered using a single frame, we outperform all but one approach (PSAC). We gain this increase in performance despite not taking advantage of Flow features/ complex memory-networks (used in Co-memory), or co-attention mechanisms (used in PSAC). In the STRN-GC variants, we do not use the Global Context Encoder LSTM. As shown in the table, we get good results even without using action-dynamics with global-context.

Model	Action ↑	Trans ↑	FrameQA ↑	Count ↓
Random [5]	20.00	20.00	0.06	6.92
Text [5]	47.91	56.93	39.26	5.01
VIS+LSTM(aggr) [5]	46.80	56.90	34.60	5.09
VIS+LSTM(avg) [5]	48.80	34.80	35.00	4.80
VQA-MCB(aggr) [5]	58.90	24.30	25.70	5.17
VQA-MCB(avg) [5]	29.10	33.00	15.50	5.54
ST(R+C) [5]	60.10	65.70	48.20	4.38
ST-SP(R+C) [5]	57.30	63.70	45.50	4.28
ST-TP(R+C) [5]	60.80	67.10	49.30	4.40
ST-SP-TP(R+C) [5]	57.00	59.60	47.80	4.56
Co-memory (R+F) [2]	68.20	74.30	51.50	4.10
PSAC (R) [8]	70.40	76.90	**55.70**	4.27
STRN-GC (R) [ours]	72.16	79.18	52.90	4.42
STRN-GC (C) [ours]	71.42	78.85	50.04	4.10
STRN (R+C) [ours]	**73.14**	**79.30**	52.96	**3.91**

Table 1: Comparison with state-of-the-art on TGIF-QA dataset. ↑ means higher numbers correspond to better performance (ACC) and ↓ means lower numbers correspond to better performance (MSE).

	Exist	Count	Integer Comparison			Attribute Comparison					Query					All
			More	Equal	Less	Color	Size	Type	Dir	Shape	Color	Size	Type	Dir	Shape	
Random [21]	50.00	22.00	50.00	50.00	50.00	50.00	50.00	50.00	50.00	50.00	12.50	50.00	50.00	25.00	33.33	33.33
Text [21]	52.92	32.41	75.14	56.39	57.81	47.73	52.56	53.12	53.55	51.56	12.27	51.07	48.65	25.23	32.70	39.95
GRU+AVG [21]	51.77	33.18	59.66	54.12	59.38	52.27	50.00	51.13	53.27	47.58	19.78	51.91	53.33	28.26	38.29	41.43
2GRU [21]	53.54	35.02	68.18	53.70	56.10	54.12	51.28	51.70	52.70	47.86	19.59	53.50	58.38	34.79	38.34	41.85
ST-TP [21]	51.46	32.54	58.46	50.39	53.52	49.74	54.56	53.12	51.95	50.39	21.23	53.81	55.70	36.08	40.60	40.47
SVQA [21]	52.03	38.20	**74.28**	57.67	61.60	55.96	**55.90**	**53.40**	**57.50**	**52.98**	23.39	**63.30**	**62.90**	**43.20**	41.69	44.90
STRN [ours]	**54.01**	**44.67**	72.22	**57.78**	**62.92**	**56.39**	55.28	50.69	50.14	50.00	**24.31**	59.68	59.32	28.24	**44.49**	**47.58**

Table 2: Comparison with the state-of-the-art on different categories of the SVQA dataset. Since everything is accuracy (ACC), higher numbers correspond to better performance.

SVQA Dataset: Results are summarized in Table 2. Similar to Table 1, the two lines at the top correspond to random and text-only models; *GRU+AVG* is an image-VQA based approach; *2GRU* is similar to ST-based methods [5] and SP-TP is the same as Jang *et al.* [5]. As shown in the last column of Table 2, we outperform all methods by a margin of **2.68%**. We perform better in Exist, Count and five out of thirteen sub-categories of Integer Comparison, Attribute Comparison and Query. We do competitively in three and worse in five sub-categories. However, we would also like to highlight a substantial improvement of **6.47% in the Count category**, which unlike sub-categories, forms a significant portion (23%) of the total dataset (see Fig 3 of [21]). This result is in consonance with TGIF-QA dataset (Table 1), where we gain a substantial improvement of **4.63% in the Count** task. Since counting is considered a complex task requiring elaborate spatio-temporal reasoning, we believe this improvement conclusively demonstrates the effectiveness of our approach.

4.2.2 Ablations

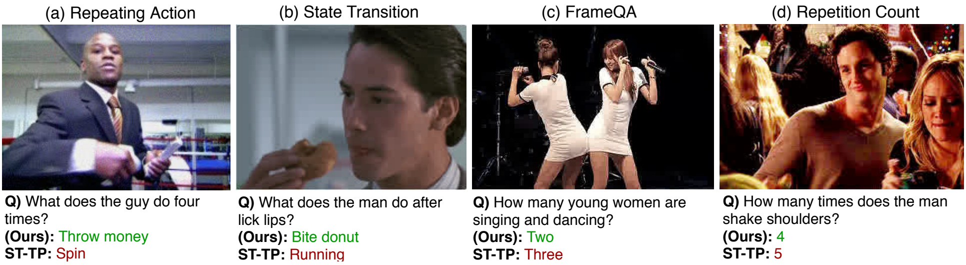
In this experiment, conducted on TGIF-QA dataset, we show the effectiveness of joint spatio-temporal relational reasoning, as opposed to individual spatial or temporal relational reasoning. To show that our experiment generalizes over different modalities, we consider separate models trained individually on both ResNet (ResNet-res5c) and C3D (C3D-conv5b). In order to avoid interference, we do not use Global Context Encoder LSTM and we call the resultant model as STRN-GC. We summarize the results in Table 3. We consider two baselines. In STRN-GC-TRM, we replace the Temporal Relations Module (TRM) with a two-layer LSTM as a baseline to model temporal relations. We initialize the hidden state of the LSTM using the last hidden state of the text-encoder LSTM, following the ST models of Jang *et al.* [5]. In STRN-GC-SRM, we replace the Spatial Relations Module (SRM) using an expressive CNN. As shown in the table, STRN-GC significantly outperforms both baselines in all four tasks, which shows the effectiveness of joint spatio-temporal relational reasoning. We believe the reason is that LSTMs and CNNs, despite being effective to model temporal and spatial data, need to learn relational reasoning from scratch which is inefficient.

	ResNet-res5c				C3D-conv5b			
Model	Action	Trans	Frame	Count	Action	Trans	Frame	Count
STRN-GC-TRM	64.95	71.25	44.86	4.50	63.10	71.26	44.63	4.18
STRN-GC-SRM	66.09	77.36	49.57	4.54	67.72	77.70	47.53	4.40
STRN-GC	**72.16**	**79.18**	**52.90**	**4.42**	**71.42**	**78.85**	**50.04**	**4.10**

Table 3: Effectiveness of joint spatio-temporal relational reasoning as opposed to individual.

5 Conclusion

In this paper, we propose a novel neural network module which provides an effective prior to capture spatio-temporal relations (object-interactions and action-dynamics). We achieve state-of-the-art performance on a real-world (TGIF-QA) and a synthetic (SVQA) videoQA datasets. Additionally, we achieve substantial improvement in the challenging counting task, which requires capturing spatio-temporal dynamics in different parts of a video.



(a) Repeating Action

Q) What does the guy do four times?
(Ours): Throw money
ST-TP: Spin

(b) State Transition

Q) What does the man do after lick lips?
(Ours): Bite donut
ST-TP: Running

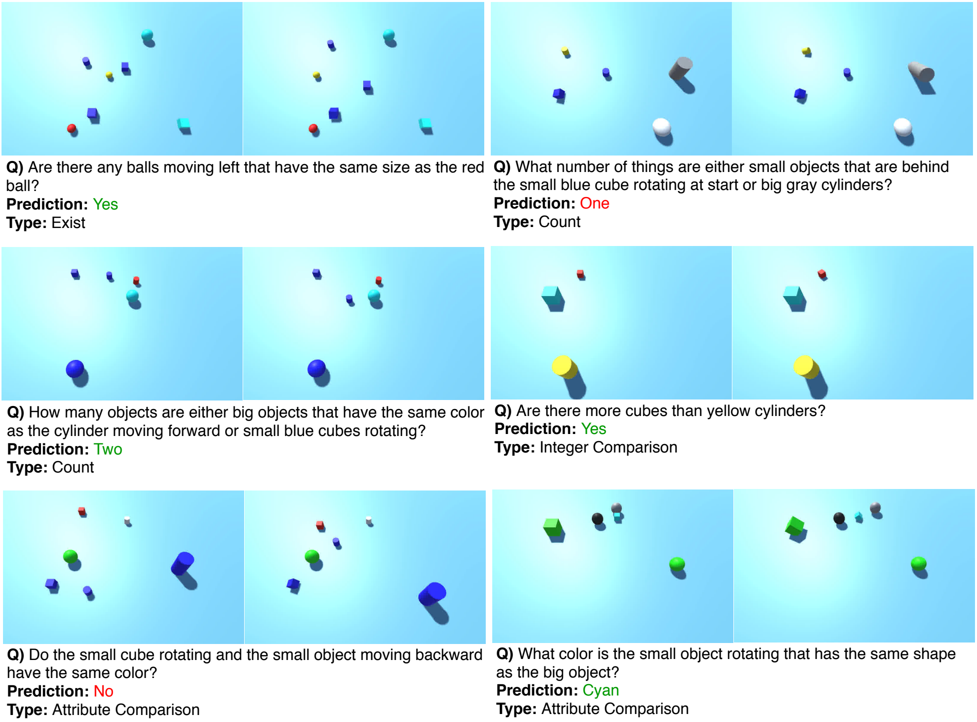
(c) FrameQA

Q) How many young women are singing and dancing?
(Ours): Two
ST-TP: Three

(d) Repetition Count

Q) How many times does the man shake shoulders?
(Ours): 4
ST-TP: 5

Figure 3: [Best viewed in color] A comparison of the qualitative results of ST-TP [5] and STRN (Ours). Green and Red refers to correct and incorrect predictions, respectively.



Q) Are there any balls moving left that have the same size as the red ball?
Prediction: Yes
Type: Exist

Q) What number of things are either small objects that are behind the small blue cube rotating at start or big gray cylinders?
Prediction: One
Type: Count

Q) How many objects are either big objects that have the same color as the cylinder moving forward or small blue cubes rotating?
Prediction: Two
Type: Count

Q) Are there more cubes than yellow cylinders?
Prediction: Yes
Type: Integer Comparison

Q) Do the small cube rotating and the small object moving backward have the same color?
Prediction: No
Type: Attribute Comparison

Q) What color is the small object rotating that has the same shape as the big object?
Prediction: Cyan
Type: Attribute Comparison

Figure 4: [Best viewed in color] Qualitative results of our approach (STRN) for different categories of the SVQA dataset. Green and Red refers to correct and incorrect predictions, respectively.

References

[1] 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*, pages 2425–2433, 2015.

[2] Jiyang Gao, Runzhou Ge, Kan Chen, and Ram Nevatia. Motion-appearance co-memory networks for video question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6576–6585, 2018.

[3] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE International Conference on Computer Vision*, volume 1, page 3, 2017.

[4] 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*, pages 770–778, 2016.

[5] 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*, pages 2758–2766, 2017.

[6] 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*, pages 2901–2910, 2017.

[7] Kyung-Min Kim, Min-Oh Heo, Seong-Ho Choi, and Byoung-Tak Zhang. Deepstory: video story qa by deep embedded memory networks. *arXiv preprint arXiv:1707.00836*, 2017.

[8] Xiangpeng Li, Jingkuan Song, Lianli Gao, Xianglong Liu, Wen bing Huang, Xiangnan He, and Chuang Gan. Beyond rnns: Positional self-attention with co-attention for video question answering. In *AAAI 2019*, 2019.

[9] Chih-Yao Ma, Asim Kadav, Iain Melvin, Zsolt Kira, Ghassan AlRegib, and Hans Peter Graf. Attend and interact: Higher-order object interactions for video understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6790–6800, 2018.

[10] Jonghwan Mun, Paul Hongsuck Seo, Ilchae Jung, and Bohyung Han. Marioqa: Answering questions by watching gameplay videos. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2867–2875, 2017.

[11] Seil Na, Sangho Lee, Jisung Kim, and Gunhee Kim. A read-write memory network for movie story understanding. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 677–685, 2017.

[12] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543, 2014.

[13] Zhaofan Qiu, Ting Yao, and Tao Mei. Learning spatio-temporal representation with pseudo-3d residual networks. In *proceedings of the IEEE International Conference on Computer Vision*, pages 5533–5541, 2017.

[14] David Raposo, Adam Santoro, David Barrett, Razvan Pascanu, Timothy Lillicrap, and Peter Battaglia. Discovering objects and their relations from entangled scene representations. *arXiv preprint arXiv:1702.05068*, 2017.

[15] Mengye Ren, Ryan Kiros, and Richard Zemel. Exploring models and data for image question answering. In *Advances in Neural Information Processing Systems*, pages 2953–2961, 2015.

[16] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, pages 91–99, 2015.

[17] 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*, pages 3202–3212, 2015.

[18] Adam Santoro, David Raposo, David G Barrett, Mateusz Malinowski, Razvan Pascanu, Peter Battaglia, and Timothy Lillicrap. A simple neural network module for relational reasoning. In *Advances in Neural Information Processing Systems*, pages 4967–4976, 2017.

[19] Franco Scarselli, Marco Gori, Ah Chung Tsoi, Markus Hagenbuchner, and Gabriele Monfardini. The graph neural network model. *IEEE Transactions on Neural Networks*, 20(1):61–80, 2009.

[20] Zheng Shou, Dongang Wang, and Shih-Fu Chang. Temporal action localization in untrimmed videos via multi-stage cnns. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1049–1058, 2016.

[21] Xiaomeng Song, Yucheng Shi, Xin Chen, and Yahong Han. Explore multi-step reasoning in video question answering. In *2018 ACM Multimedia Conference on Multimedia Conference*, pages 239–247. ACM, 2018.

[22] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4489–4497, 2015.

[23] Xiaolong Wang and Abhinav Gupta. Videos as space-time region graphs. In *Proceedings of the European Conference on Computer Vision*, pages 399–417, 2018.

[24] Philippe Weinzaepfel, Zaid Harchaoui, and Cordelia Schmid. Learning to track for spatio-temporal action localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3164–3172, 2015.

[25] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM International Conference on Multimedia*, pages 1645–1653. ACM, 2017.

[26] Youngjae Yu, Hyungjin Ko, Jongwook Choi, and Gunhee Kim. End-to-end concept word detection for video captioning, retrieval, and question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3165–3173, 2017.

[27] Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *Proceedings of the European Conference on Computer Vision*, pages 803–818, 2018.