Env-QA: A Video Question Answering Benchmark for Comprehensive Understanding of Dynamic Environments

Difei Gao[1,2], Ruiping Wang[1,2,3], Ziyi Bai[1,2], Xilin Chen[1,2]
[1]Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS),
Institute of Computing Technology, CAS, Beijing, 100190, China
[2]University of Chinese Academy of Sciences, Beijing, 100049, China
[3]Beijing Academy of Artificial Intelligence, Beijing, 100084, China
{difei.gao, ziyi.bai}@vipl.ict.ac.cn, {wangruiping, xlchen}@ict.ac.cn

Abstract

*Visual understanding goes well beyond the study of images or videos on the web. To achieve complex tasks in volatile situations, the human can deeply understand the environment, quickly perceive events happening around, and continuously track objects' state changes, which are still challenging for current AI systems. To equip AI system with the ability to understand dynamic **ENV**ironments, we build a video **Q**uestion **A**nswering dataset named Env-QA. Env-QA contains 23K egocentric videos, where each video is composed of a series of events about exploring and interacting in the environment. It also provides 85K questions to evaluate the ability of understanding the composition, layout, and state changes of the environment presented by the events in videos. Moreover, we propose a video QA model, Temporal Segmentation and Event Attention network (TSEA), which introduces event-level video representation and corresponding attention mechanisms to better extract environment information and answer questions. Comprehensive experiments demonstrate the effectiveness of our framework and show the formidable challenges of Env-QA in terms of long-term state tracking, multi-event temporal reasoning and event counting, etc.*

1. Introduction

In the last decades, tremendous works [9, 30, 45, 17, 58, 10, 49] have brought revolutionary advancements to computer vision systems for understanding web data, e.g., photos, videos, and movies, while for deploying machines in the human living environment (i.e., building embodied artificial intelligence), we will encounter brand new challenges on visual ability. 1) From "*broader*" to "*deeper*" visual understanding. The studies of internet AI focus on making the system recognize from dozens of object categories [29, 34] to thousands of categories [9, 43, 19] (*broader*). However, internet AI mainly pays attention to the salient objects shown in the images. For completing tasks in real world environment, such as cooking a meal, a system needs an in-depth understanding of every detail of the environment (*deeper*), e.g., knowing the positions of all utensils and ingredients in a kitchen. 2) From *static* to *dynamic* visual understanding. One of the essential characteristics of the real world environment lies in its dynamic nature. The interactions between a human and the environment will trigger the environment's state changes. A system must learn to sense and remember state changes to accomplish some long-term tasks, e.g., a housekeeping robot may need to continuously track the state of objects at home to plan the cleaning task.

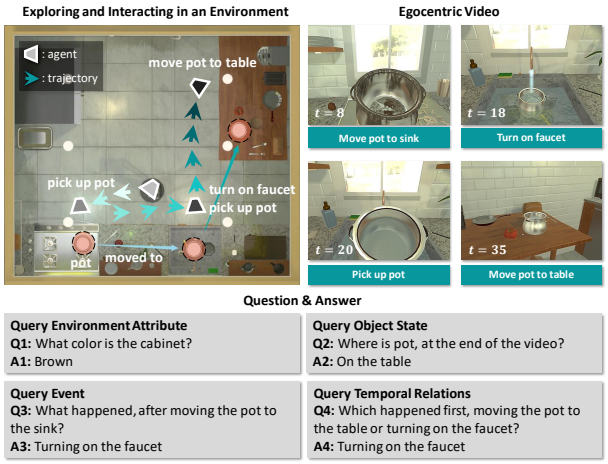

Figure 1. Env-QA dataset contains egocentric videos about exploring and interacting with environments, and diverse questions to evaluate the models of understanding dynamic environments from various perspectives.

However, few works purely study these visual abilities under embodied AI setting. Some of the video QA datasets, such as TVQA [32] and MovieQA [49], evaluate the model's understanding of movies, TV series or YouTube

Table 1. Comparison of Env-QA with other related video QA and embodied QA datasets. The table shows the basic information of the modalities involved in the existing datasets. The content in brackets shows the main characteristic of the visual material.

Dataset	Vision		Language	Action	#Clips	#QA
	Object-Centric Interaction	Human-Centric Action				
MovieQA [49]	-	Movie (Plot)	Question+Subtitle+Script	-	6.8K	6.5K
TGIF-QA [22]	-	Social Media (Action)	Question	-	56.7K	103.9K
TVQA [32]	-	TV Show (Plot)	Question+Subtitle	-	21.8K	152.5K
TVQA+ [33]	-	TV Show (Plot)	Question+Subtitle	-	4.2K	29.4K
Social-IQ [59]	-	YouTube (Social Situation)	Question+Transcript+Audio	-	1.2K	7.5K
CLEVERER [57]	Synthetic Video (Object Collision)	-	Question	-	10.0K	305.0K
Embodied QA [8]	AI Habitat (Static Env.)	-	Question	Navigation	-	1.1K
Interactive QA [14]	AI2-THOR (Nearly Static Env.)	-	Question	Navigation+Manipulation	-	75.0K
Env-QA (Ours)	AI2-THOR (**Dynamic Env.**)	-	Question	-	23.3K	85.1K

videos. In Table 1, we display the main characteristics of the related datasets. Although these tasks explore the dynamics of vision, they focus more on the dynamics introduced by human-centered actions, social activities, or plot development, rather than the interaction with environments. Correspondingly, these tasks mainly require the abilities of human posture recognition, dialogue understanding, and social knowledge understanding. Although some other related tasks, such as Visual Navigation and Manipulation [61, 48], and Embodied QA [8, 14], involve the understanding of the environment, they focus more on the comprehensive ability of how to plan actions in the environment. The visual ability of environment understanding is implicitly evaluated by the quality of performed actions. Besides, these tasks usually require models to perform in a nearly static environment, so they are also hard to investigate the dynamics of environments.

Thus, we propose to take the question answering as a proxy task to purely study the dynamic environments understanding. The task is required to watch an egocentric video composed of a series of events about exploring and interacting in the environment, e.g., move the pot, turn on the faucet, as shown in Figure 1. It must then answer a question that requires 1) understanding the environment's composition (like **Q1**), layout, trajectory of state changes (like **Q2**) presented by the events, or 2) performing temporal reasoning on events (like **Q3** and **Q4**).

To support such a task, we construct a large-scale dataset, Env-QA, containing **23.3K** videos and **85.1K** questions. A critical challenge in building a dataset of this scale is how to control the distribution of samples. Most recent QA datasets with off-the-shelf visual materials from Internet contain unexpected biases [24, 1]. These biases could be more distinct for housework in natural scenes, like cooking, leading to high risks for models to guess the answer without even looking at the visual materials. To address this challenge, we resort to the recently proposed virtual simulator AI2-THOR [27] to generate videos with strictly controlled content by ourselves. Specifically, we design a semi-automatic data collection method. Our designed algorithm is responsible for controlling the sample distribution and automatically generating natural language guidance information. Then, annotators follow the guidance to manipulate in the simulator to generate videos and collect question-answer pairs.

Understanding the dynamic environments from a sequence of interaction events requires extracting key environmental information from the events and performing temporal reasoning to capture state changes. And the foundation of both abilities is to represent the video at the level of events, that is dividing the video into clips according to its content to let the model locate key events easier. However, the previous video QA methods [21, 33] mainly use the grid-level video features with a preset interval extracted by temporal CNN [50, 23]. To address this problem, we introduce Temporal Segmentation and Event Attention network (TSEA), which will first segment the video to flexible duration clips based on the content, then perform multi-step temporal reasoning to locate the key events for a given question and output the answer. Experiments on Env-QA demonstrate the effectiveness of our proposed method and reveal that Env-QA is challenging in terms of capturing the long-term state change, multi-event temporal reasoning, and event counting, etc.

2. Related Work

2.1. Embodied AI Tasks

In the 1990s, [51, 11] put forward the concept of embodied cognition and embodied artificial intelligence, and emphasize the importance of a body in cognitive learning. In recent years, many pioneer works spend great efforts in building powerful virtual environment simulators that can be explored and interacted for embodied AI researches, e.g., Matterport 3D [6], AI2-THOR [27], VirtualHome [40], AI Habitat [46] and UnrealCV [41, 42]. Researchers also propose corresponding embodied AI tasks, e.g., object navigation, vision-language navigation, vision-language manipulation, embodied QA, and rearrangement. Object navigation task [61] requires exploring in the environment to find the specified object, which evaluates the ability to make decisions based on the egocentric visual observations [55, 54, 47]. Vision-language navigation

task [3, 7, 52] requires models to act in the environment according to natural language instructions. This type of works [12, 55, 25, 16, 28, 52, 53, 18, 18, 37] mainly studies the comprehensive ability of vision-language-action, e.g., interpreting natural language instructions into specific move actions [12]. [48] proposes a more challenging task, ALFRED, which additionally requires the model to achieve complex manipulations based on detailed instructions. Embodied question answering [8, 14] requires models to explore in the environment [8] and manipulate objects [14], e.g., open a refrigerator to find the object asked by the question, and answer the question about object attribute. Rearrangment [5] is a recently proposed high-level cognition task that requires manipulating objects to make a given physical environment into a specified state.

Previous embodied AI tasks mainly evaluate the comprehensive capabilities of vision and action, so it is hard to purely diagnose the vision ability. Besides, these planning tasks usually only require to make action decision based on the *current environment state*, so it is hard to investigate the dynamics of environments. Thus, Env-QA collects egocentric videos with diverse events to purely study the visual challenges, and introduces new types of questions to evaluate the understanding of the whole *trajectory of environment state changes*. More illustrations are in Supplementary.

2.2. Video QA Tasks

With the development of image question answering [4, 15, 35], many works in recent years begin to study video question answering tasks [49, 22, 36, 26, 38, 59, 57]. One of the early tasks, TGIF-QA [22] proposes to answer questions about short videos (e.g., GIF images). This task examines the model's understanding of short-term actions, such as recognizing actions, count actions. Another type of tasks [33, 32] queries the content of movies or television series. The main characteristic of these tasks is the requirement of understanding human-centric plots, e.g., understanding subtitles and more advanced background common sense. The recent CLEVERER dataset [57] uses the renderer to construct videos containing a series of object collision events. This task's core difficulty lies in the reasoning of causality, e.g., which collision event caused another collision event.

Although the visual samples in existing video QA datasets, e.g., film, TV series, involve many scenes, most of the periods in these videos do not focus on the environments. Besides, most of the questions examine the understanding of human actions, dialogues, and social conventions. In contrast, our collected videos are all about exploring and interacting in the environments, and proposed questions evaluate the ability of dynamic environment understanding from diverse perspectives.

2.3. Video Representation and Temporal Reasoning

Early works [50, 23] extend 2D convolution to 3D convolution, which uses similar mechanisms to deal with temporal and spatial dimensions. [10] argues that the processing mechanism of temporal and spatial dimensions should be different. Thus, it proposes a two-way mechanism to capture appearance information and motion information separately. These video representation methods are mainly for single-action video recognition tasks. For tasks that require temporal reasoning, e.g., action localization, video caption, or video QA, the model requires a spatio-temporal attention mechanism to represent multi-action videos' key content. [60, 33] propose frame-level temporal attention and region-level spatial attention mechanism to locate the video's key content according to a natural language query and then achieve video captioning or question answering. Another video QA work [13] proposes to use memory networks to conduct spatio-temporal attention mechanisms. To achieve better event reasoning capabilities, [31] proposes to divide a video into equal length clips, then hierarchically extract the features of frames, clips, and entire video.

The existing video representation methods mainly extract grid-level features with a preset interval in the temporal dimension, e.g., the features of each frame [60, 33], or the clip features [31]. In contrast, for better performing multi-event temporal reasoning in Env-QA, we propose an event-level video representation that segments the video into clips according to its content.

3. Dataset Construction

This section describes how we collect videos and question-answer pairs for Env-QA dataset with both diverse content and controllable distribution. We design a semi-automatic construction method to collect samples with AI2-THOR [27] simulator, as shown in Figure 2. The key issue is to control the sample distribution and automatically generate natural language guidance; then, annotators follow the guidance to manipulate the simulator to generate videos and collect QA pairs.

3.1. Video Collection

Env-QA uses the recently released AI2-THOR simulator to collect egocentric videos about exploring and interacting in the environment. AI2-THOR provides four categories, a total of 120 indoor simulation environments, including kitchen, living room, bedroom, and bathroom. These environments contain 115 types of objects and support multiple types of interactive operations, such as turning on, throwing, etc.

For Env-QA dataset, a total of 15 types of basic actions are defined, as shown in Figure 2 (Note that, Tidying one object is defined as moving all objects out of this object.).

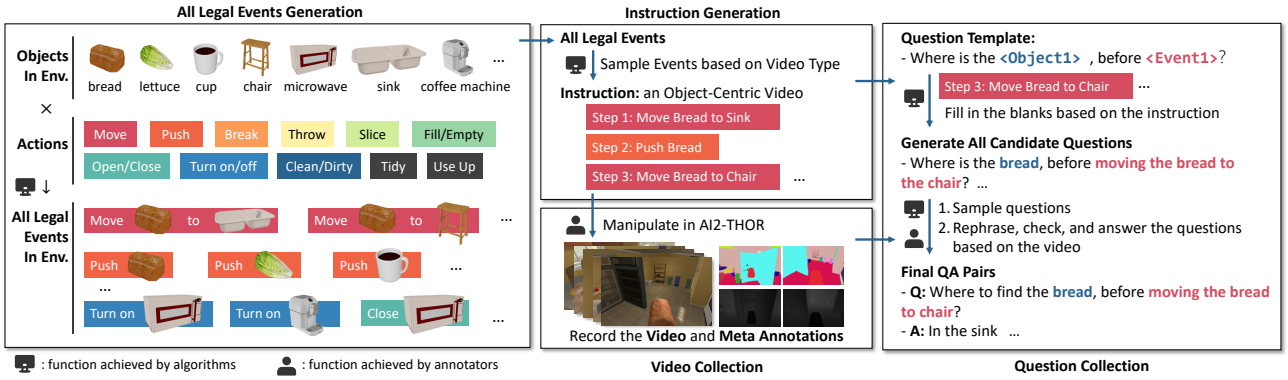

Figure 2. Pipeline of proposed semi-automatic construction of Env-QA dataset. The algorithms in the pipeline are responsible for generating auxiliary annotations, e.g., instructions and candidate questions, for guiding annotators to control the distribution and difficulty of samples. The annotators are responsible for collecting the videos and QA pairs to ensure the samples' naturalness and correctness.

Given objects that exist in a virtual environment, the algorithm will generate all executable legal events in the environment. Then, we design five types of videos, exploring, random, object-centric, action-centric, and comprehensive task, to evaluate different abilities of models, as shown in Figure 3 (a). Collecting exploring-type videos require annotators to walk in the environment to find some specified objects. This type of videos aims to examine the model's ability to understand static environments. Random type videos contain a series of completely random events in the environment, mainly examining the model's ability to recognize events and temporal reasoning. Object-centric videos contain a series of events surrounding some selected objects. This type of videos mainly examines the model's ability to track the objects' state. The action-centric videos contain events with similar actions for measuring the ability of event counting. The comprehensive task videos are about accomplishing complex daily life tasks, e.g., heating potatoes, washing a pot. This type of videos examines the understanding of complex events in daily human life. For collecting one specific type of videos, we design a sampler to automatically sample some actions from all legal events under specific constraints to generate the instruction, as shown in Figure 2 and 3 (b). Finally, the annotators manipulate the environments in our developed web-based AI2-THOR annotation platform according to the provided instructions. The platform will record the video and environment metadata, including the depth map, the instance segmentation map, and the environment metadata (the objects' pose and state). Through the above method, we collect 4,720 long-time videos, each of which mainly contains about 5 to 10 events. The video type distribution of these videos is shown in Figure 3 (a). These long-time videos can evaluate the model's ability to track volatile environment state changes. To also provide some simpler samples to test the understanding of short-term changes, we split part of the videos into shorter videos mainly containing about 1 to 4 events. Finally, we collect 23,261 videos of varying lengths in total, which evenly cover the four categories of environments.

3.2. Question Collection

After collecting the videos, we first design a template-based question generator to output the balanced candidate questions according to the instructions. Specifically, Env-QA defines five types of questions to evaluate dynamic environment understanding from different aspects, including querying object *Attribute*, object *State*, *Event*, temporal *Order* of events, and counting *Number* of events or objects. In Figure 3 (c) and (d), we display each type of questions and examples. Then, for each type of questions, we collect a set of question templates, e.g., *where is the* <Object1>, *before* <Event1>*?*, *is the* <Object1> *closed, at the end of video?*. The generator will automatically fill in the blanks according to the instructions. The generated questions then feed into a filter to balance the answer distribution. Finally, the annotators rephrase, modify and check the auto-generated questions based on the video content to ensure questions' diversity and accuracy, and annotate the answers.

3.3. Dataset Statistics

Env-QA collects a total of **23,261** egocentric videos, **85,072** question-answer pairs and rich annotations of the videos, such as instance segmentation map, depth map, environment metadata[1]. These samples are divided into three splits, train (70% of samples), validation (15%), and test (15%) split, where 60% of videos in validation and test split are recorded in new environments which do not appear in the train split to evaluate the cross-environment generalization performance of models. The average video duration is about 20 seconds. Moreover, we downsample the videos by extracting the video frames at 4 FPS (4 frames per second, as similarly done in [33]). In Figure 3 (e), we display the event number distribution of videos in Env-QA.

[1]The dataset is available at http://vipl.ict.ac.cn/resources/envqa.

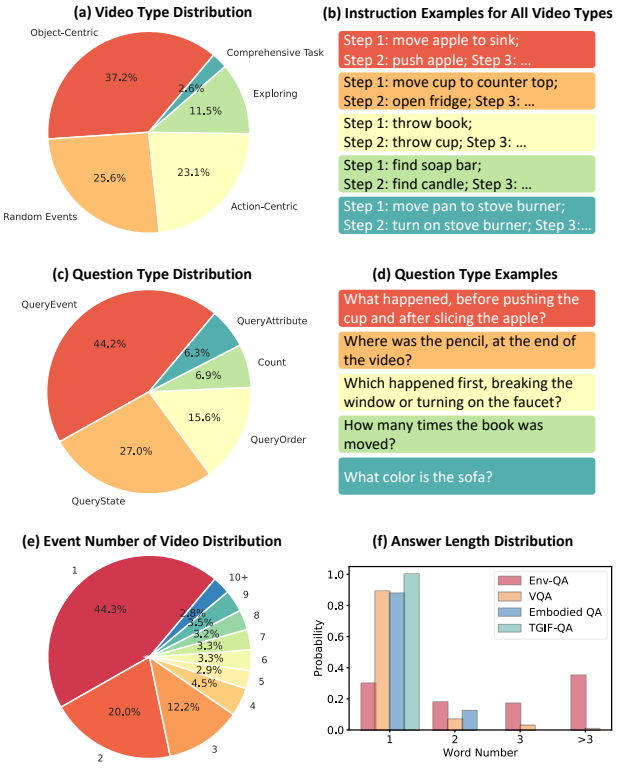

Figure 3. Statistics of Env-QA dataset.

The videos cover a wide range of difficulties, which contain 1 to 10 events. For the questions, 42% of them are about videos containing 1 to 4 events, and the rest are about more complex videos. In Figure 3 (f), we also compare the answer length distributions of Env-QA and other datasets which contain open-ended questions. It can be seen that a large part of the answers in Env-QA contain multiple words. These answers are mainly describing an event. In Figure 6, we show some samples in our dataset. More samples and statistics are shown in Supplementary.

3.4. Evaluation Metrics

As shown in Figure 3 (f), answers in the form of phrases are common in Env-QA dataset. If using conventional metric [4] to evaluate the answers, it is difficult to precisely measure the similarity between the ground truth and predicted answers. For example, "move the pan to countertop" is closer to the answer "move the plate to countertop" than "slice the apple", and it deserves a higher score. Therefore, inspired by the work of situation recognition [56], Env-QA evaluates the answers in the role-value format. Specifically, the answers of Env-QA will involve the following seven roles: *Action* (e.g., move, open, etc.), *Object1* (the object being manipulated), *Prep.* (indicates the position of the object, e.g., on, near, etc.), *Object2* (Some actions may involve two objects, e.g., move egg to plate. This role represents the second object.), *Adjective* (indicates the attributes or non-location type states of the objects, e.g., broken, sliced. Abbreviated as Adj.), *Number*, and *Yes/No*. Every answer in Env-QA can be mapped into the role-value format, as shown in Figure 6. Note that some roles could be empty for a specific answer. This format provides a better evaluation which part of the answer is wrong. Besides, the accuracy of the predicted answer can be calculated as an IoU-like score of the predicted values and the ground-truth values:

$$s = \frac{|C|}{|P \cup G|} \tag{1}$$

where $|P \cup G|$ represents the number of roles that is non-empty in the predicted role set P or ground-truth role set G, and $|C|$ represents the number of roles that are non-empty and have equal values in both P set and G set.

4. Method

This section presents the details of TSEA for dynamic environment understanding. To better extract the environment information from the events in videos, this model presents an event-level video representation and multi-step temporal attention mechanism. Specifically, TSEA is composed of three modules: 1) event-level video feature extraction module, 2) multi-step temporal attention module, and 3) answer prediction module, as shown in Figure 4.

Event-Level Video Feature Extraction Module. This module splits the video $\boldsymbol{v}$ into clips and extracts the feature of each video clip. Specifically, we use Faster R-CNN model [44] to extract the region features, then feed them into a temporal CNN [33] to encode the short-time temporal information into the feature of each frame. The predicted object names and bounding boxes are also appended to the corresponding object features, and we obtain the final object features $\{\boldsymbol{o}_{t1}, ..., \boldsymbol{o}_{tN}\}$ of each t-th frame, where N indicates the number of objects in the frame. Besides, for the egocentric videos, the distance between the object i's center $\boldsymbol{c}_{ti}$ (the coordinate of bounding box center) and the image center $\boldsymbol{c}_t$ naturally expresses its importance. Thus, to make the model focus more on the key objects, we design a window function to calculate the attention value α_{ti} on each object $\boldsymbol{o}_{ti}$ in the frame (denoted as focus attention):

$$\alpha_{ti} = \begin{cases} \epsilon, & d(\boldsymbol{c}_{ti}, \boldsymbol{c}_t) < \tau \\ 1-\epsilon, & otherwise \end{cases}, \tag{2}$$

where d indicates the Euclidean distance function, ϵ is a hyper-parameter indicating the attention value, and τ is a hyper-parameter indicating the size of the focused area in the image. Then, the frame feature $\boldsymbol{v}_t$ is the weighted sum of object features $\boldsymbol{v}_t = (\alpha_{ti}/\Sigma_{i=1}^{N}\alpha_{ti})\boldsymbol{o}_{ti}$,

After obtaining the frame features, we design a heuristic algorithm to segment the video to generate event-level video features without using the additional segmentation annotations. This algorithm is designed based on the assumption

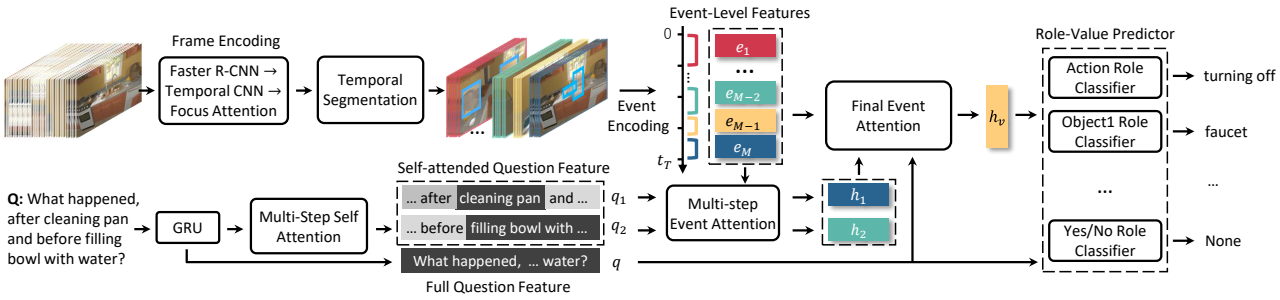

Figure 4. Pipeline of our proposed TSEA model. TSEA first extracts the event-level video features, then performs a multi-step temporal attention on the events, and finally predicts the role-value format answer.

Algorithm 1 Temporal Segmentation of Video

Input: array $S[1, ..., T]$, where each $S[t]$ is a set whose members are the names of objects $\boldsymbol{o}_{ti}, i \in \{1, ..., N\}$ satisfying $d(\boldsymbol{c}_{ti}, \boldsymbol{c}_t) < \tau$;

Output: array p, of which each element is a segment point of the video;

initial $s = 1$; $p = []$;
for $t = 1$ to T **do**
  if $S[s] \cap S[s+1] \cap ... \cap S[t] \in \varnothing$ **then**
    p.append(t);
    $s = t$;
  end if
end for
return p;

that when doing one action, the objects in the center of visual observation are consistent. Specifically, the algorithm iterates from the start to the end of the frames to find each longest video clip that the intersection of the object sets in the image center is not empty, as shown in Algorithm 1. The algorithm outputs the segment points $\boldsymbol{p}$ of the video. Then, we convert it to a matrix, $A \in \mathbb{R}^{T \times M}$, to represent the video segments, where T indicates the frame number and M indicates the number of events in a video. The element a_{ij} indicates if the i-th frame belongs to j-th event. Specifically, $a_{ij} = 1$ when $i \in [p_j, p_{j+1}]$, otherwise the $a_{ij} = 0$. Then, the feature of each event is calculated as:

$$\boldsymbol{e}_j = \frac{\Sigma_{t=1}^{T} a_{tj} \boldsymbol{v}_t}{\Sigma_{t=1}^{T} a_{tj}}. \quad (3)$$

Multi-Step Temporal Attention Module. After obtaining the event-level video representation, we design a multi-step temporal attention mechanism to attend the events based on the key parts of the question. To encode the text input, we first use GloVe embedding [39] along with GRU to obtain the features of each word w_i in the question. Due to that the questions sometimes mention multiple events, we design a two-step self-attention mechanism [20] to obtain the features of key parts of the question, $\boldsymbol{q}_1$ and $\boldsymbol{q}_2$. Then, a soft attention mechanism [2] is performed to use $\boldsymbol{q}_1$ and $\boldsymbol{q}_2$ to locate the event mentioned in the question. The attended event features are denoted as $\boldsymbol{h}_1$ and $\boldsymbol{h}_2$. Finally, we concatenate the $\boldsymbol{h}_1$, $\boldsymbol{h}_2$, and full question feature $\boldsymbol{q}$, and use the concatenated feature to attend the event most related to answering the question. The finally attended video feature is denoted as $\boldsymbol{h}_v$.

Answer Prediction Module. We design seven classification heads to predict the values of seven roles, which are in the same architecture but different label sizes and parameters. They take the attended video feature $\boldsymbol{h}_v$ and full question feature $\boldsymbol{q}$ as the inputs, then predict the values, formalized as:

$$P(y_i|\boldsymbol{h}_v, \boldsymbol{q}) = \text{Softmax}(W_i(W_v \boldsymbol{h}_v \odot W_q \boldsymbol{q})) \quad (4)$$

where W_i, W_v, and W_q are trainable parameters, y_i indicates the values for the i-th role, and $\odot$ indicates element-wise multiplication. Finally, we calculate the cross entropy loss of all predicted values of roles to train the whole model. More details of the model are shown in the Supplementary.

5. Experiments

5.1. Baselines

- Q-ONLY: This baseline only uses the question feature as input to predict the answer.
- I-ONLY: This baseline only uses the visual feature as input to predict the answer.
- CNN-LSTM: This baseline simply concatenates the video and question feature to generate the answer.
- ST-VQA: ST-VQA [21] is the state-of-the-art model on TGIF-QA task. ST-VQA introduces a dual-LSTM based spatio-temporal mechanism to better represent the video content. Because the model does not support the region features and the output mechanism is not compatible with the Env-QA task, we keep the spirit of its attention mechanism and modify its input processing and output module.
- STAGE: STAGE [33] is the state-of-the-art model on TVQA+ task. STAGE proposes a CNN-based frame-level spatio-temporal attention mechanism on video content and subtitle. We also make some modifications, e.g.,

Table 2. Comparison with baseline methods on Env-QA test split. The table shows the accuracy on each type of question and each role.

Model	Question Type Accuracy (%)						Role Accuracy (%)					
	Attribute	State	Event	Order	Number	Overall	Action	Object1&2	Prep.	Adj.	Yes/No	Number
Q-ONLY	37.29	32.17	24.26	51.79	37.84	32.48	42.05	36.48	51.03	34.15	50.53	37.83
I-ONLY	3.51	3.76	3.56	0.57	2.12	3.05	4.60	7.87	19.42	2.84	0.01	1.08
CNN-LSTM	38.21	42.26	29.94	53.37	38.12	38.05	45.89	43.07	54.15	37.90	43.27	38.07
ST-VQA [21]	41.66	48.98	33.87	54.09	38.54	41.97	45.08	45.06	54.50	41.07	55.44	38.51
STAGE [33]	39.49	49.93	34.52	55.32	37.98	42.53	45.69	47.24	54.35	42.71	52.07	37.66
TSEA	**42.96**	**56.73**	**39.84**	**55.53**	**39.35**	**47.06**	**47.61**	**50.51**	**55.33**	**44.93**	**57.56**	**39.35**

removing the subtitle processing branch, modifying the output module to adapt to Env-QA.

5.2. Results and Ablations

In Table 2, we show the results of baseline methods and our proposed model. The Q-ONLY method only reaches a low-level accuracy, 32.48%, indicating that Env-QA dataset is relatively balanced and contains limited language priors. The I-ONLY method is much lower, showing that the visual content is diverse. The previous state-of-the-art models on other video question answering tasks achieve at most 42.53% overall accuracy. Our proposed TSEA obtains a boost over the best of previous methods by 4.5% accuracy. Still, the overall accuracies of all models are far from satisfactory, and there is significant headroom remaining.

Besides, from the question type accuracy in Table 2, it can be seen that compared with the Q-ONLY method, the state-of-the-art methods and TSEA mainly achieve performance improvements in Query Attribute, Event, and State questions. In contrast, the performance improvements in Number and Order questions are limited. This shows that the multi-event reasoning is still quite challenging for the existing methods. Researchers need to further design more sophisticated symbolic reasoning mechanisms, e.g., modular networks, to effectively tackle this difficulty. From the role-value accuracy in Table 2, we can see that the performance gains from introducing visual features are relatively small on Action role recognition, compared to Object and Adj. roles. Action role recognition requires the understanding of longer video clips, while Object and Adj. roles may only need to find the key frames. It shows that understanding concepts related to long-term clips is also difficult.

The Effect of Video Length. In Figure 5, we display the performances of various methods on Query Event Questions in different video lengths. We show this type of questions because it is involved in all lengths of video, and its answer space is larger, which can better reflect the model's performance difference. The video length is measured by the number of events contained in a video. The performance of the Q-ONLY reflects the extent of language priors that participates in answering questions of different video lengths. As the video length increases, the language prior exploited almost continually increases. This is because the questions for short videos usually contain limited information that could be used to guess the answer, e.g., "What happened in the video?". When the length is too large (video length > 5), there is a slight decrease. The possible reason is that when the content of a video is much richer, the answers are more uncertain and harder to guess.

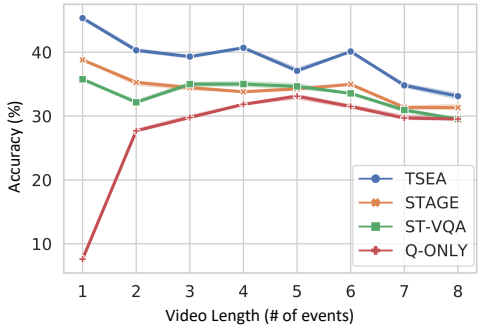

Figure 5. The performances of Query Event Questions on different lengths of video. The length of a video is measured by the number of events in it.

Moreover, it can be seen that as the length of the video increases, the performance of all vision-based methods decreases, and the performance gaps between vision-based methods and Q-ONLY are shrinking significantly. For the longest video, almost all models' performances are close to the Q-ONLY model. In Figure 6, we display the qualitative results of TSEA model. It can be seen that TSEA fails at both answering and event attention for videos requiring long-term tracking (**Q5** and **Q6**). These results show that existing video QA methods all struggle to extract useful information from long-time videos. We may need a more structural video representation to record rich video content.

Ablation. We present an extensive experiment that compares our model TSEA as presented above with its variants that remove some core modules to determine which components are most important. Concretely, we cumulatively ablate some parts of TSEA, and evaluate them on Env-QA test split. In Table 3, we display the performances of these variants. From the results, it can be seen that all components obtain desired gains, demonstrating they tailor to the challenges of Env-QA. Besides, the results show that the event feature, the core of TSEA, contributes the most to perfor-

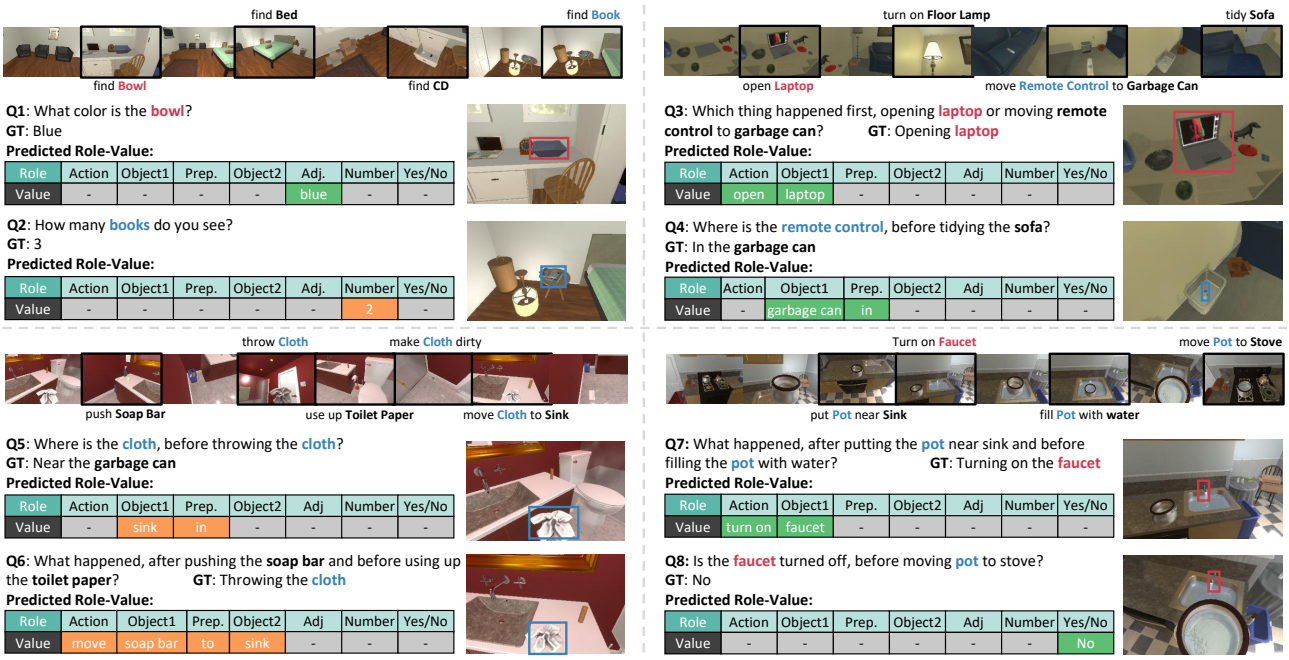

Figure 6. Example predictions from TSEA. We display some key frames in example videos and provide corresponding questions, ground-truth answers (denoted as **GT**), and predicted role-value answers. The larger images on the right side are the frames in the attended events predicted by TSEA.

mance.

Generalization on Unseen Environments. As illustrated in Section 3.3, the test split videos of Env-QA have two parts, one part is collected in the same environments as the training set, and the other part is collected in unseen environments. In Table 3, we display the performances of methods on seen environments and unseen environments separately to analyze the generalization ability on unseen environments. It can be found that the performances of questions about unseen environments are quite similar to that of the questions about seen environments. This shows that at the level of visual understanding, the generalization performance of current framework is relatively promising. It may be because the current feature extractor, which is pre-trained on diverse large-scale web data, is strong enough for handling the cross-environment generalization in Env-QA task. More results are shown in Supplementary.

Table 3. Cumulative ablation of TSEA network on test split. The ablations of table rows are cumulative from top to bottom.

Model	Accuracy (%)		
	Seen Env.	Unseen Env.	Overall
TSEA	**46.85**	**47.20**	**47.06**
– Multi-Step Attention	45.41	45.60	45.46
– Event Feature	43.05	43.61	43.39
– Focus Attention	41.84	41.79	41.80
– Object Name Feature	41.05	40.32	40.60

6. Conclusion

In this paper, we propose a new video question answering task for the understanding of the dynamic environments and correspondingly construct a large-scale dataset. The proposed task requires intelligent systems to watch a video about exploring and interacting in an environment, then extract useful information and perform temporal reasoning to answer the questions. We further propose a novel video QA method with event-level video representation, TSEA, to deal with above task. Comprehensive experiments demonstrate the effectiveness of the TSEA. Besides, the results reveal the main challenges of Env-QA task for current models: 1) The lower accuracies on Query Event, Number questions show the limitations of models on performing temporal reasoning on multiple events. 2) Unsatisfactory results on long videos indicate the formidable challenge of tracking objects' states for a long time. All these imply that some innovative ideas need to be explored, e.g., environment-level representation (like 3D scene graphs), a more powerful event feature extractor (like video Transformer), or a symbolic temporal reasoning mechanism (like modular network). We hope Env-QA can empower the researches of understanding dynamic environments and help to move the fields of video analysis, QA, and embodied AI forward.

Acknowledgements. This work is partially supported by National Key R&D Program of China (2020AAA0105200), Natural Science Foundation of China under contracts Nos. U19B2036, 61922080, 61772500, and 61390510. Besides, we sincerely thank the anonymous reviewers and area chairs for their valuable comments.

References

[1] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4971–4980, 2018. 2

[2] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and vqa. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6077–6086, 2018. 6

[3] Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton Van Den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3674–3683, 2018. 3

[4] 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 (ICCV)*, pages 2425–2433, 2015. 3, 5

[5] Dhruv Batra, Angel X Chang, Sonia Chernova, Andrew J Davison, Jia Deng, Vladlen Koltun, Sergey Levine, Jitendra Malik, Igor Mordatch, Roozbeh Mottaghi, et al. Rearrangement: A challenge for embodied ai. *arXiv preprint arXiv:2011.01975*, 2020. 3

[6] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niebner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. In *International Conference on 3D Vision (3DV)*, pages 667–676, 2017. 2

[7] Howard Chen, Alane Suhr, Dipendra Misra, Noah Snavely, and Yoav Artzi. Touchdown: Natural language navigation and spatial reasoning in visual street environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12538–12547, 2019. 3

[8] Abhishek Das, Samyak Datta, Georgia Gkioxari, Stefan Lee, Devi Parikh, and Dhruv Batra. Embodied question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1–10, 2018. 2, 3

[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009. 1

[10] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 6202–6211, 2019. 1, 3

[11] Stan Franklin. Autonomous agents as embodied ai. *Cybernetics & Systems*, 28(6):499–520, 1997. 2

[12] Daniel Fried, Ronghang Hu, Volkan Cirik, Anna Rohrbach, Jacob Andreas, Louis-Philippe Morency, Taylor Berg-Kirkpatrick, Kate Saenko, Dan Klein, and Trevor Darrell. Speaker-follower models for vision-and-language navigation. In *Advances In Neural Information Processing Systems (NeurIPS)*, pages 3318–3329, 2018. 3

[13] 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 (CVPR)*, pages 6576–6585, 2018. 3

[14] Daniel Gordon, Aniruddha Kembhavi, Mohammad Rastegari, Joseph Redmon, Dieter Fox, and Ali Farhadi. Iqa: Visual question answering in interactive environments. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4089–4098, 2018. 2, 3

[15] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6904–6913, 2017. 3

[16] Weituo Hao, Chunyuan Li, Xiujun Li, Lawrence Carin, and Jianfeng Gao. Towards learning a generic agent for vision-and-language navigation via pre-training. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13137–13146, 2020. 3

[17] 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 (CVPR)*, pages 770–778, 2016. 1

[18] Sachithra Hemachandra, Felix Duvallet, Thomas M Howard, Nicholas Roy, Anthony Stentz, and Matthew R Walter. Learning models for following natural language directions in unknown environments. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 5608–5615, 2015. 3

[19] Ronghang Hu, Piotr Dollár, Kaiming He, Trevor Darrell, and Ross Girshick. Learning to segment every thing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4233–4241, 2018. 1

[20] Ronghang Hu, Anna Rohrbach, Trevor Darrell, and Kate Saenko. Language-conditioned graph networks for relational reasoning. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 10294–10303, 2019. 6

[21] Yunseok Jang, Yale Song, Chris Dongjoo Kim, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Video question answering with spatio-temporal reasoning. *International Journal of Computer Vision*, 127(10):1385–1412, 2019. 2, 6, 7

[22] 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 (CVPR)*, pages 2758–2766, 2017. 2, 3

[23] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):221–231, 2012. 2, 3

[24] 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 (CVPR)*, pages 1988–1997, 2017. 2

[25] Liyiming Ke, Xiujun Li, Yonatan Bisk, Ari Holtzman, Zhe Gan, Jingjing Liu, Jianfeng Gao, Yejin Choi, and Siddhartha Srinivasa. Tactical rewind: Self-correction via backtracking in vision-and-language navigation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6741–6749, 2019. 3

[26] Kyung-Min Kim, Min-Oh Heo, Seong-Ho Choi, and Byoung-Tak Zhang. Deepstory: Video story qa by deep embedded memory networks. In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2016–2022, 2017. 3

[27] Eric Kolve, Roozbeh Mottaghi, Winson Han, Eli VanderBilt, Luca Weihs, Alvaro Herrasti, Daniel Gordon, Yuke Zhu, Abhinav Gupta, and Ali Farhadi. Ai2-thor: An interactive 3d environment for visual ai. *arXiv preprint arXiv:1712.05474*, 2017. 2, 3

[28] Jacob Krantz, Erik Wijmans, Arjun Majumdar, Dhruv Batra, and Stefan Lee. Beyond the nav-graph: Vision-and-language navigation in continuous environments. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 104–120, 2020. 3

[29] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. *Citeseer*, 2009. 1

[30] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 1097–1105, 2012. 1

[31] Thao Minh Le, Vuong Le, Svetha Venkatesh, and Truyen Tran. Hierarchical conditional relation networks for video question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9972–9981, 2020. 3

[32] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. Tvqa: Localized, compositional video question answering. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1369–1379, 2018. 1, 2, 3

[33] Jie Lei, Licheng Yu, Tamara Berg, and Mohit Bansal. Tvqa+: Spatio-temporal grounding for video question answering. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 8211–8225, 2020. 2, 3, 4, 5, 6, 7

[34] Bastian Leibe and Bernt Schiele. Analyzing appearance and contour based methods for object categorization. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages II–409, 2003. 1

[35] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 13–23, 2019. 3

[36] Tegan Maharaj, Nicolas Ballas, Anna Rohrbach, Aaron Courville, and Christopher Pal. A dataset and exploration of models for understanding video data through fill-in-the-blank question-answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6884–6893, 2017. 3

[37] Arsalan Mousavian, Alexander Toshev, Marek Fišer, Jana Košecká, Ayzaan Wahid, and James Davidson. Visual representations for semantic target driven navigation. In *International Conference on Robotics and Automation (ICRA)*, pages 8846–8852, 2019. 3

[38] 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 (ICCV)*, pages 2867–2875, 2017. 3

[39] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014. 6

[40] Xavier Puig, Kevin Ra, Marko Boben, Jiaman Li, Tingwu Wang, Sanja Fidler, and Antonio Torralba. Virtualhome: Simulating household activities via programs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8494–8502, 2018. 2

[41] Weichao Qiu and Alan Yuille. Unrealcv: Connecting computer vision to unreal engine. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 909–916, 2016. 2

[42] Weichao Qiu, Fangwei Zhong, Yi Zhang, Siyuan Qiao, Zihao Xiao, Tae Soo Kim, and Yizhou Wang. Unrealcv: Virtual worlds for computer vision. In *Proceedings of the ACM international conference on multimedia (ACM-MM)*, pages 1221–1224, 2017. 2

[43] Joseph Redmon and Ali Farhadi. Yolo9000: Better, faster, stronger. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6517–6525, 2017. 1

[44] 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 (NeurIPS)*, pages 91–99, 2015. 5

[45] Bryan C Russell, Antonio Torralba, Kevin P Murphy, and William T Freeman. Labelme: a database and web-based tool for image annotation. *International journal of computer vision*, 77(1-3):157–173, 2008. 1

[46] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, et al. Habitat: A platform for embodied ai research. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 9339–9347, 2019. 2

[47] Alexander Sax, Jeffrey O Zhang, Bradley Emi, Amir Zamir, Silvio Savarese, Leonidas Guibas, and Jitendra Malik. Learning to navigate using mid-level visual priors. In *International Conference on Learning Representations (ICLR)*, 2020. 2

[48] Mohit Shridhar, Jesse Thomason, Daniel Gordon, Yonatan Bisk, Winson Han, Roozbeh Mottaghi, Luke Zettlemoyer, and Dieter Fox. Alfred: A benchmark for interpreting grounded instructions for everyday tasks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10740–10749, 2020. 2, 3

[49] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4631–4640, 2016. 1, 2, 3

[50] 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 (ICCV)*, pages 4489–4497, 2015. 2, 3

[51] Francisco J Varela, Evan Thompson, and Eleanor Rosch. The embodied mind: Cognitive science and human experience. *The MIT Press*, 1991. 2

[52] Xin Wang, Qiuyuan Huang, Asli Celikyilmaz, Jianfeng Gao, Dinghan Shen, Yuan-Fang Wang, William Yang Wang, and Lei Zhang. Reinforced cross-modal matching and self-supervised imitation learning for vision-language navigation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6629–6638, 2019. 3

[53] Xin Wang, Wenhan Xiong, Hongmin Wang, and William Yang Wang. Look before you leap: Bridging model-free and model-based reinforcement learning for planned-ahead vision-and-language navigation. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 37–53, 2018. 3

[54] Mitchell Wortsman, Kiana Ehsani, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Learning to learn how to learn: Self-adaptive visual navigation using meta-learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6750–6759, 2019. 2

[55] Wei Yang, Xiaolong Wang, Ali Farhadi, Abhinav Gupta, and Roozbeh Mottaghi. Visual semantic navigation using scene priors. In *International Conference on Learning Representations (ICLR)*, 2018. 2, 3

[56] Mark Yatskar, Luke Zettlemoyer, and Ali Farhadi. Situation recognition: Visual semantic role labeling for image understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5534–5542, 2016. 5

[57] Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum. Clevrer: Collision events for video representation and reasoning. In *International Conference on Learning Representations (ICLR)*, 2020. 2, 3

[58] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 33, pages 9127–9134, 2019. 1

[59] Amir Zadeh, Michael Chan, Paul Pu Liang, Edmund Tong, and Louis-Philippe Morency. Social-iq: A question answering benchmark for artificial social intelligence. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8807–8817, 2019. 2, 3

[60] Luowei Zhou, Yannis Kalantidis, Xinlei Chen, Jason J Corso, and Marcus Rohrbach. Grounded video description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6578–6587, 2019. 3

[61] Yuke Zhu, Roozbeh Mottaghi, Eric Kolve, Joseph J Lim, Abhinav Gupta, Li Fei-Fei, and Ali Farhadi. Target-driven visual navigation in indoor scenes using deep reinforcement learning. In *IEEE international conference on robotics and automation (ICRA)*, pages 3357–3364, 2017. 2