MIST : Multi-modal Iterative Spatial-Temporal Transformer for Long-form Video Question Answering

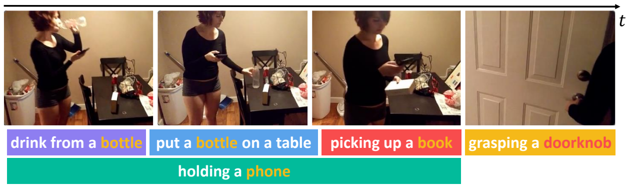
Difei Gao[1], Luowei Zhou[2*], Lei Ji[3], Linchao Zhu[4], Yi Yang[4], Mike Zheng Shou[1†]
[1]Show Lab, National University of Singapore, [2]Microsoft,
[3]Microsoft Research Asia, [4]Zhejiang University

Abstract

To build Video Question Answering (VideoQA) systems capable of assisting humans in daily activities, seeking answers from long-form videos with diverse and complex events is a must. Existing multi-modal VQA models achieve promising performance on images or short video clips, especially with the recent success of large-scale multi-modal pre-training. However, when extending these methods to long-form videos, new challenges arise. On the one hand, using a dense video sampling strategy is computationally prohibitive. On the other hand, methods relying on sparse sampling struggle in scenarios where multi-event and multi-granularity visual reasoning are required. In this work, we introduce a new model named Multi-modal Iterative Spatial-temporal Transformer (MIST) to better adapt pre-trained models for long-form VideoQA. Specifically, MIST decomposes traditional dense spatial-temporal self-attention into cascaded segment and region selection modules that adaptively select frames and image regions that are closely relevant to the question itself. Visual concepts at different granularities are then processed efficiently through an attention module. In addition, MIST iteratively conducts selection and attention over multiple layers to support reasoning over multiple events. The experimental results on four VideoQA datasets, including AGQA, NExT-QA, STAR, and Env-QA, show that MIST achieves state-of-the-art performance and is superior at efficiency. The code is available at `github.com/showlab/mist`.

1. Introduction

One of the ultimate goals of Video Question Answering (VideoQA) systems is to assist people in solving problems in everyday life [13, 27, 41], e.g., helping users find something, reminding them what they did, and assisting them while accomplishing complex tasks, etc. To achieve such functions, the systems should be able to understand and seek the answer from long-form videos with diverse events about users' activities.



Multi-event:
Q1. Did the person interact with a doorknob before or after putting something on a table? **Answer:** After

Multi-grained Visual Concepts:
Q2. Which object were they touching between drinking from a bottle and picking up a book? **Answer:** Phone

Causality:
Q3. Why does the person put the bottle on a table?
Answer: She has finished drinking

Figure 1. **Main challenges of long-form VideoQA.** The questions for long-form VideoQA usually involve multi-event, multi-grained, and causality reasoning.

Compared to understanding and reasoning over short videos, many unique challenges arise when the duration of the video increases, as shown in Fig. 1: 1) Multi-event reasoning. The long-form videos usually record much more events. The questions about these videos thus naturally require the systems to perform complex temporal reasoning, e.g., multi-event reasoning (Q1 in Fig. 1), causality (Q3), etc. 2) Interactions among different granularities of visual concepts. The questions of short-clip videos usually involve the interactions of objects or actions that happened simultaneously, while questions for long-form videos could involve more complex interactions of objects, relations, and events across different events, e.g., Q2 in Fig. 1.

Current vision-language methods [2, 7, 10, 24, 29, 31, 32, 51, 52] excel at QA over images or short clips spanning several seconds. In other words, they excel at learning multi-modal correspondences between a single caption with one or few events. Their tremendous progress over these years is fueled by 1) pre-training on large-scale image-language [22, 37, 38] and short-clip-language

[*]Currently at Google Brain.
[†]Corresponding author.

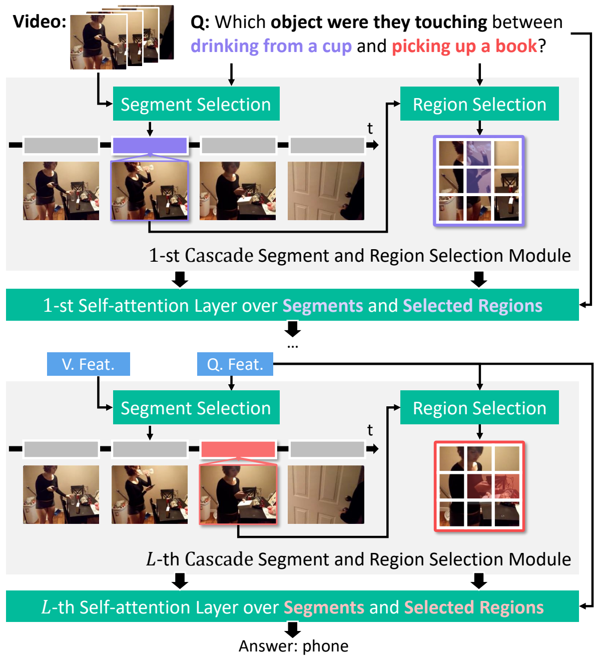

Video:
Q: Which object were they touching between drinking from a cup and picking up a book?
Segment Selection
Region Selection
t
1-st Cascade Segment and Region Selection Module
1-st Self-attention Layer over Segments and Selected Regions
...
V. Feat.
Q. Feat.
Segment Selection
Region Selection
t
L-th Cascade Segment and Region Selection Module
L-th Self-attention Layer over Segments and Selected Regions
Answer: phone

Figure 2. **Diagrammatic illustration of MIST .** It revises a standard spatial-temporal self-attention layer into two modules: a cascade selection module that dynamically eliminates question-irrelevant image regions, and a self-attention layer reasoning over multi-modal multi-grained visual concepts. The proposed modules further iterate multiple times to reason over different events.

datasets [2, 33], and 2) end-to-end multi-modal Transformers [1–3, 10, 37, 40], which is superior at learning the alignments between images with texts.

However, these multi-modal Transformers rely on the dense self-attention with the computation cost increasing exponentially over time especially when adapting to long-form videos. To make the dense self-attention computationally feasible in processing videos, almost all current state-of-the-art pre-trained Transformers are sparse sample-based methods, e.g., [2, 40] only sample 3 or 4 frames per video regardless of its length. If we simply adapt these pre-trained models to long-form videos with the same sampling strategy, there will be a domain gap between the pre-training and downstream VideoQA tasks. In pre-training, the sparsely sampled frames of a short video depict a coherent action, while they are likely to be random shots for part of events in a long video. Recently, some early attempts process the video hierarchically [5], which splits the video into several segments and performs QA only on aggregated segment-level features. It can ease the efficiency issue, but is still hard to capture complex interactions among multi-grained concepts. Thus, leveraging the advantages of models pre-trained from images or short videos and addressing the challenges of long-form VideoQA is worth exploring.

In this paper, we propose a new model, named Multi-modal Iterative Spatial-temporal Transformer (MIST), as shown in Fig. 2. MIST comes from a simple finding that for long-form VideoQA, it is not necessary to consider the details of all events in a video, like what dense self-attention over all patches do. The model only needs to consider the general content of all events and focuses on the details of a few question-related events. Thus, MIST decomposes dense joint spatial-temporal self-attention into a question-conditioned cascade segment and region selection module along with a spatial-temporal self-attention over multi-modal multi-grained features. The cascade selection reduces the computation cost and benefits the performance by focusing on the question-related segments and regions. The self-attention over segments and image patches, better captures interactions among different granularities of visual concepts. In addition, through iteratively conducting selection and self-attention, MIST can reason over multiple events and better perform temporal and causal reasoning.

We conduct experiments on several VideoQA datasets with relatively longer videos, AGQA [14], NExT-QA [44], STAR [42], and Env-QA [11], with an average video duration varies from 12s to 44s. The experimental results show that our approach achieves state-of-the-art performance. Further ablation studies verify the effectiveness of the key components. Moreover, quantitative and qualitative results also show that our method provides higher efficiency and reasonable evidence for answering questions.

2. Related Work

Video question answering. Video Question Answering is one typical type of vision-language task studied for many years. Some datasets [20, 47] focus on short clips about daily human activities, e.g., sports, household work, etc. Some others, such as TVQA [25], MovieQA [39], and Social-IQ [50], mainly focus on long videos cropped from movies or TV series for evaluating the understanding of the plot and social interactions, where subtitles play an essential role. Recently, [11, 14, 42, 44] aim to evaluate more complex spatial-temporal reasoning over long-form videos, e.g., causality, sequential order, etc. Current works achieve promising results on the first two types of benchmarks, while struggle on the last one, which is our focus.

In terms of methodology, early-stage works proposed various LSTM or Graph Neural Network-based models to capture cross-modal [28, 35, 54] or motion-appearance interaction [12, 23]. One recent work [45] integrates graph modeling into Transformers to explicitly capture the objects and their relations in videos. In addition, with the great success of pre-trained vision-language Transformers, many works [2, 10, 40] directly fine-tune the pre-trained model on downstream VideoQA tasks. [5] proposes a simple yet effective fine-tuning strategy to hierarchically process videos with pre-trained Transformers.

Compared to previous works, this paper is an early attempt to specifically focus on the challenges of long-form VideoQA for Transformers-based methods. Specifically, we revise the self-attention mechanism to better perform multi-event, multi-grained visual concepts reasoning.

Transferring pre-trained models to downstream tasks. Many works try to transfer pre-trained vision-language Transformers, such as CLIP [37], into downstream tasks, e.g., object detection [15], image generation [36], and video-text retrieval [9, 32, 48, 53]. CLIP4Clip [32] proposes various aggregation methods for CLIP features, e.g., mean pooling, Transformer, to better represent a video. CLIP2Video [9] proposes a temporal difference block to better capture motion information. Similar to the above methods, we preserve the strengths of pre-trained models and improve their weaknesses on downstream tasks, but this works focus on another one, long-form VideoQA, where the main focus is on multi-event and multi-granularity reasoning.

Long-form video modeling. With the great success of short-term video understanding in recent years, some pioneer works [8,43] have started to focus on long-form video modeling for action recognition or localization tasks. They mainly focus on increasing the efficiency of processing long video features. [8] proposes short-term feature extraction and long-term memory mechanisms that can eliminate the need for processing redundant video frames during training. [30] proposes to replace parts of the video with compact audio cues to succinctly summarize dynamic audio events and are cheap to process. [16] introduces structured multi-scale temporal decoder for self-attention to improve efficiency. The above methods utilize the natural characteristics of videos to reduce the computation. In contrast, this paper considers the characteristics of QA tasks to use the question as a guide to reduce computation.

Iterative Attention. Many existing works [4, 6, 17, 34] are for improving computation efficiency. Some of them propose similar iterative attention mechanisms to ours. [34] proposes a recurrent image classification model to iteratively attending on a sequence of regions at high resolution. Perceiver [17] revises self-attention in Transformer to an asymmetric attention to iteratively distill inputs into a tight feature, allowing it to handle large inputs. TimeSformer [4] proposes various self-attention schemes for video classification models to separately apply temporal and spatial attention. Our model differs in utilizing multi-modal correspondence (i.e., vision and question) to guide iterative attention.

3. Method

The goal of a VideoQA task is to predict the answer y for a given video $\mathcal{V}$ and a question q, formulated as follows:

$$\widetilde{y} = \arg\max_{y\in\mathcal{A}} \mathcal{F}_\theta(y|q,\mathcal{V},\mathcal{A}), \tag{1}$$

where $\widetilde{y}$ is the predicted answer chosen from the candidate answers (i.e., answer vocabulary or provides choices), denoted as $\mathcal{A}$, and θ is the set of trainable parameters of a VideoQA model $\mathcal{F}$.

In Fig. 3, we present the pipeline of our proposed Multi-Modal Iterative Spatial-temporal Transformer, MIST. MIST answers the question in three steps: 1) utilize a pre-trained model to extract the input features, 2) iteratively perform self-attention over a selected set of features to perform multi-event reasoning, 3) predict the answer based on the obtained video, question, and answer features.

3.1. Input Representation

Existing vision-language Transformers are good at representing images or short clips. To adapt them to handle the long-form video, we first split the video into K uniform length segments, where each segment contains T frames. In addition, each frame is divide into N patches. Note that, for the simplicity of notation, the [CLS] token for image patches and frames are counted in N and T.

The vision-language Transformer, like CLIP, All-in-one, with frozen parameters, extracts patch-level features of all segments, $\boldsymbol{x} = \{x^1, x^2, ..., x^K\}$, where $x^k \in \mathbb{R}^{T\times N\times D}$ is the feature of k-th segment, where D is the dimension of each patch-level feature. The patch-level visual token features will be used to obtain frame and segment features in the following modules. Since the segment features are separately extracted, to indicate their temporal positions in the whole video, we add position embedding $P_t \in \{\phi_t(i)|i \in [0, K\cdot T]\}$ for each token with their frame index.

For the text part, the question is tokenized as a sequence of words, and then fed into the vision-language Transformer to get word-level features $\mathbf{X_w} = \{w_1, ..., w_M\}$, where w_1 corresponds to [CLS] and $w_2, ..., w_M$ are words in question.

3.2. Iterative Spatial-Temporal Attention Layer

The Iterative Spatial-Temporal Attention layer (ISTA) aims to iteratively select the segments and regions among a long video conditioned on questions and then perform multi-event reasoning over selected ones. Specifically, ISTA contains three steps: segment selection, region selection, and spatial-temporal self-attention, as shown in Fig. 4.

Segment Selection. Given a set of image patch features $\boldsymbol{x}$, we calculate the features of segments and the question, then select the patch features of Top_k segments by performing cross-modal temporal attention and differentiable top-k selection.

Specifically, to perform temporal attention, the frame features are first obtained by pooling the features in spatial dimension: the t-th frame feature in k-th segment is calculated as $f^k_t = \text{pool}(x^k_{t,1}, x^k_{t,2}, ..., x^k_{t,N})$, where $x^k_{t,n}$ indicates n-th patch at t-th frame of k-th segment. Then, the segment features are obtained by pooling frames features

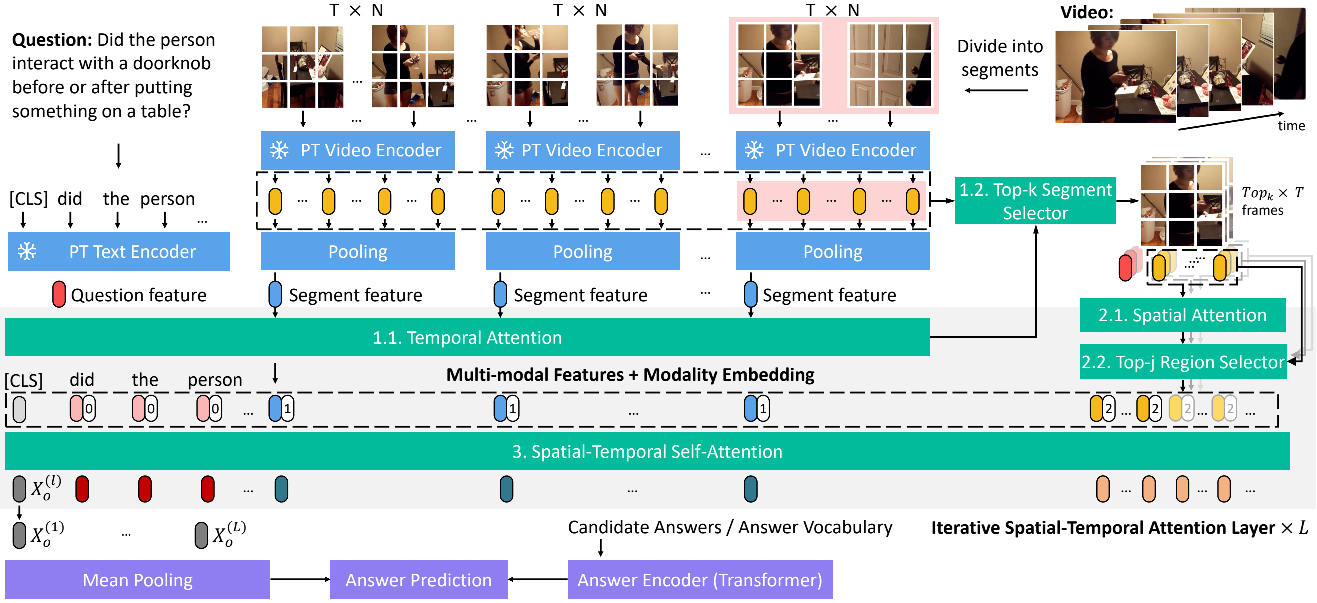


Figure 3. **Architecture of MIST .** MIST first divides video into several segments and utilizes the pre-trained (PT) video encoder to extract the feature of each one. Then, MIST iteratively performs self-attention over a selected set of features to reason over multiple events. Finally, it predicts the answer by comparing the combination of video and question features with answer candidate features. Note that the "PT Video Encoder" in the figure can also be image-based encoders.

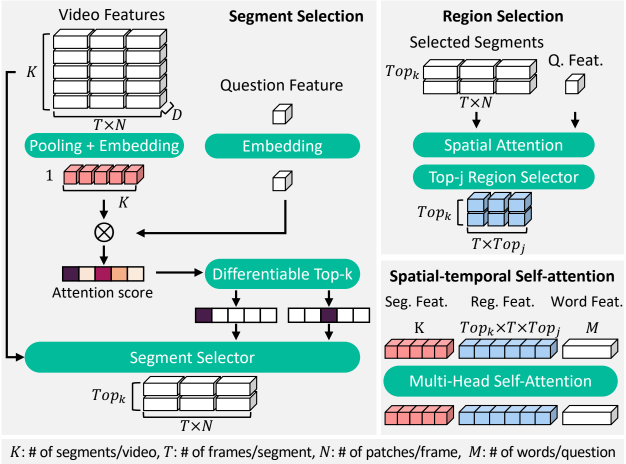


Figure 4. **Key components of Iterative Spatial-Temporal Attention Layer.** Since region selection follows the same architecture as segment selection, we only show its inputs and outputs.

along the temporal dimension: $s^k = \text{pool}(f_1^k, f_2^k, ..., f_T^k)$. The question feature is similarly obtained by pooling the word features, $\boldsymbol{q} = \text{pool}(w_1, ..., w_M)$. The pooling functions can be chosen from mean pooling, first token pooling, simple MLP layer, etc., according to the specific type of used vision-language Transformer. For example, for image-language Transformers, like CLIP, the first token pooling can be used for extracting frame and question features and mean pooling over frames for obtaining segment features.

Given the segment features $\mathbf{S} = \{s^k\}_{k=1}^K$, patch features $\mathbf{X} = \{x^k\}_{k=1}^K$, and question features $\boldsymbol{q}$, we first perform cross-modal temporal attention among $\mathbf{S}$ given $\mathbf{q}$, and then conduct top-k feature selection over $\mathbf{X}$, as formulated:

$$\mathbf{Q} = g_q(\boldsymbol{q}), \mathbf{K} = g_s(\mathbf{S}), \mathbf{V} = \mathbf{X}, \tag{2}$$

$$\mathbf{X}_t = \underset{Top_k}{\text{selector}}(\text{softmax}(\frac{\mathbf{QK}^T}{\sqrt{d_k}}), \mathbf{V}), \tag{3}$$

where g_q and g_s are linear projection layers for different types of features, selector is a differentiable top-k selection function to choose the spatial features of Top_k segments. The top-k selection can be implemented by expanding the Gumbel-Softmax trick [18] or based on optimal-transport formulations [46] for ranking and sorting. In this paper, we simply conduct Gumbel-Softmax sampling Top_k times with replacement to achieve top-k selection. Note that we sample the segments with replacement because, in some cases, the question could only involve one segment. We hope the model learns to enhance the most related segment in such cases by re-sampling it, instead of forcing it to select an irrelevant segment, as sampling without replacement will do. See supplementfor more discussion about Top-k selection. The output of the module is $\mathbf{X}_t = \{x^k | k \in \mathcal{B}\} \in \mathbb{R}^{Top_k \times T \times N \times D}$, where $\mathcal{B}$ is the set of selected Top_k segments' indexes.

Region Selection. For the τ-th sampled frame, we want to select its most relevant patches with the question. Given its region feature of one frame $\mathbf{X}_\tau = \{x_{\tau,n}^k | n \in [1, N], k \in \mathcal{B}\}$ along with question $\mathbf{q}$, we perform cross-model attention over spatial patches of the τ-th sampled frame and select the Top_j most related patches. This can be formulated as:

$$\mathbf{Q} = h_q(q), \mathbf{K} = h_x(\mathbf{X}_\tau), \mathbf{V} = \mathbf{X}_\tau, \tag{4}$$

$$\mathbf{X}'_\tau = \underset{Top_j}{\text{selector}}(\text{softmax}(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}), \mathbf{V}), \tag{5}$$

where h_q and h_x are embedding layers for linear feature projection. The output of the given each frame is $\mathbf{X}'_\tau \in \mathbb{R}^{Top_j \times D}$. Finally, we stack the selected patch features of all selected frames to obtain $\mathbf{X}_{st} = \{\mathbf{X}'_\tau | \tau \in [1, Top_k \times T]\}$.

Spatial-Temporal Self-Attention. Given the selected frames and selected regions, along with the question, we aim to employ a self-attention layer to reason out a fused feature vector to jointly represent the question and video.

Regarding the inputs of self-attention, since the main computation cost comes from too many patches ($K \times T \times N$, about thousands of patches), we only keep the selected ones. While for temporal information, we keep all segments as the total number is only K (usually less than 10), which doesn't bring heavy cost and can benefit more comprehensive multi-event reasoning.

Specifically, we first add type embedding to indicate the types feature, e.g., image region, segment or word. The type embedding is formulated $P_h \in \{\phi_h(h) | h \in [1, 3]\}$ to each feature for indicating where ϕ_h is a trainable embedding layer. Then, a standard multi-head attention is performed to obtain the contextual features of all input tokens, formulated as:

$$\mathbf{X}_o^{(i)} = \text{MultiHead}([\phi_s(\mathbf{S}); \phi_x(\mathbf{X}_{st}); \phi_w(\mathbf{X}_w)]), \tag{6}$$

where ϕ_s, ϕ_x, and ϕ_w are linear transformation.

Iterative Execution of ISTA. A stack of L ISTA layers is used for modelling multi-event interactions between a given question and video, where the updated segment features and word features are fed into next layer. The output of each layer $\{\mathbf{X}_o^{(l)}\}_{l=1}^L$ is used for answer prediction.

3.3. Answer Prediction

Specifically, we mean pool the token features of all ISTA layers, $\mathbf{X}_o = \text{MeanPool}(\mathbf{X}_o^{(1)}, ..., \mathbf{X}_o^{(L)})$. In addition, following the work [49], we calculate the similarity between the $\mathbf{X}_o$ and the feature of all candidate answers $\mathbf{X}_A = \{\mathbf{x}_a | a \in \mathcal{A}\}$ obtained by using the pre-trained model. Finally, the candidate answer with the maximal similarity is considered as the final prediction $\widetilde{y}$.

$$\widetilde{y} = \arg\max_{y \in \mathcal{A}} (\mathbf{X}_o (\mathbf{X}_A)^T). \tag{7}$$

During training, we optimize the softmax cross-entropy loss between the predicted similarity scores and ground truth.

4. Experiments

4.1. Datasets

We evaluate our model on four recently proposed challenging datasets for the long-form VideoQA, namely AGQA [14], NExT-QA [44], STAR [42] and Env-QA [11].

AGQA is an open-ended VideoQA benchmark for compositional spatio-temporal reasoning. We use its v2 version, which has a more balanced distribution, as the dataset creator recommended. It provides 2.27M QA pairs over 9.7K videos with an average length of 30 seconds. **NExT-QA** is a multi-choice VideoQA benchmark for causal and temporal reasoning. It contains a total of 5.4K videos with an average length of 44s and about 52K questions. **STAR** is another multi-choice VideoQA benchmark for Situated Reasoning. STAR contains 22K video clips with an average length of 12s along with 60K questions. **Env-QA** is an open-ended VideoQA benchmark for dynamic environment understanding. It contains 23K egocentric videos with an average length of 20 seconds collected on virtual environment AI2THOR [21] along with 85K questions.

For each benchmark, we follow standard protocols outlined by prior works [1, 5, 11, 14] for dataset processing, metrics, and settings. Please see supplementfor details.

4.2. Implementation Details

Our proposed method can be built upon most of the pre-trained multi-modal Transformers. In our experiments, we try two typical types of pre-trained models, CLIP (ViT-B/32) [37] for image-language pre-training models and All-in-One-Base [40] for video-language pre-training model, denoted as MIST -CLIP and MIST -AIO respectively. In MIST , we set $Top_k = 2$ and $Top_j = 12$ in cascade selection module and the layer of ISTA $L = 2$. For all videos, we sample 32 frames per video, and split them into $K = 8$ segments. AdamW is utilized to optimize model training. Our model is trained on NVIDIA RTX A5000 GPUs and implemented in PyTorch.

4.3. Comparison with State-of-the-arts

We compare our model with the state-of-the-art (SOTA) methods on four VideoQA datasets (i.e., AGQA v2, NExT, STAR, and Env-QA), as shown in Tab. 1, 2, 3, and 4 respectively. We can see that our proposed method achieves state-of-the-art performances and outperforms the existing methods on all datasets. The performance gain is relatively limited on Env-QA, because its videos are recorded in a virtual environment, AI2THOR. There is a domain gap for CLIP feature, while previous SOTA uses the features pre-trained on virtual environment data.

Notably, among SOTAs, TEMP[ATP] [5] uses the same feature, CLIP (ViT-B/32), as MIST -CLIP. And All-in-one [40] and MIST -AIO also use the same feature, All-

Question Types	Most Likely	PSAC	HME	HCRN [23]	AIO [40]	Temp[ATP] [5]	MIST - AIO	MIST - CLIP
Object-relation	9.39	37.84	37.42	40.33	48.34	50.15	51.43	**51.68**
Relation-action	50.00	49.95	49.90	49.86	48.99	49.76	54.67	**67.18**
Object-action	50.00	50.00	49.97	49.85	49.66	46.25	55.37	**68.99**
Superlative	21.01	33.20	33.21	33.55	37.53	39.78	41.34	**42.05**
Sequencing	49.78	49.78	49.77	49.70	49.61	48.25	53.14	**67.24**
Exists	50.00	49.94	49.96	50.01	50.81	51.79	53.49	**60.33**
Duration comparison	24.27	45.21	47.03	43.84	45.36	49.59	47.48	**54.62**
Activity recognition	5.52	4.14	5.43	5.52	18.97	18.96	**20.18**	19.69
All	10.99	40.18	39.89	42.11	48.59	49.79	50.96	**54.39**

Table 1. QA accuracies of state-of-the-art (SOTA) methods on AGQA v2 test set.

Method	Causal	Temporal	Descriptive	All
HGA	44.22	52.49	44.07	49.74
CLIP (single frame)	46.3	39.0	53.1	43.7
VQA-T [49]	49.60	51.49	63.19	52.32
AIO [40]	48.04	48.63	63.24	50.60
Temp[ATP] [5]	48.6	49.3	65.0	51.5
Temp[ATP]+ATP [5]	53.1	50.2	66.8	54.3
VGT [45]	52.28	55.09	64.09	55.02
MIST - AIO	51.54	51.63	64.16	53.54
MIST - CLIP	**54.62**	**56.64**	**66.92**	**57.18**

Table 2. QA accuracies of SOTA methods on NExT-QA val set.

Method	Interaction	Sequence	Prediction	Feasibility	Mean
ClipBERT [24]	39.81	43.59	32.34	31.42	36.7
CLIP [37]	39.8	40.5	35.5	36.0	38.0
RESERVE-B [51]	44.8	42.4	38.8	36.2	40.5
Flamingo-9B [1]	-	-	-	-	43.4
AIO [40]	47.53	50.81	47.75	44.08	47.54
Temp[ATP] [5]	50.63	52.87	49.36	40.61	48.37
MIST - AIO	53.00	52.37	49.52	43.87	49.69
MIST - CLIP	**55.59**	**54.23**	**54.24**	**44.48**	**51.13**

Table 3. QA accuracies of SOTA methods on STAR val set.

Method	Attribute	State	Event	Order	Number	All
CNN-LSTM	38.21	42.26	29.94	53.37	38.12	38.05
ST-VQA [19]	41.66	48.98	33.87	54.09	38.54	41.97
STAGE [26]	39.49	49.93	34.52	55.32	37.98	42.53
AIO [40]	41.78	52.98	37.57	55.16	38.50	44.86
Temp[ATP] [5]	42.87	53.49	38.35	55.25	38.65	45.43
TSEA [11]	42.96	56.73	39.84	55.53	39.35	47.06
MIST -AIO	43.63	55.17	40.99	55.44	39.54	47.19
MIST -CLIP	**44.05**	**58.13**	**42.54**	**56.83**	**40.32**	**48.97**

Table 4. QA accuracies of SOTA methods on Env-QA test set.

Method	AGQA v2			NExT-QA			
	Binary	Open	All	C.	T.	D.	All
MeanPool	49.26	34.01	41.58	47.87	45.22	58.01	48.59
Trans.-Frame	54.03	45.66	49.66	50.77	49.96	65.27	52.76
Trans.-Patch	55.09	47.08	51.05	52.58	50.42	64.55	53.74
Divided STA	55.93	46.88	51.37	52.03	50.24	64.31	53.36
MIST - CLIP	58.28	50.56	54.39	54.62	56.64	66.92	57.18

Table 5. QA accuracies of variants of MIST on AGQA v2 and NExT-QA.

in-One-Base. Compared to these methods, it can be found that our two versions of models, which build upon different types of pre-trained models, achieve substantial performance gains on all datasets.

Moreover, from the question type breakdown of each dataset, we can see that compared with AIO and Temp[ATP], our model obtains a much more significant performance boost on questions that require multi-grained visual concepts reasoning (i.e., *Rel.-act.*, *Obj.-act.* on AGQA v2) than those which mainly require information within one frame (i.e., *Obj.-rel.* on AGQA v2 and *Descriptive* on NExT-QA). In addition, we can see that our model surpasses these models with large margin on questions requiring causality or multi-event reasoning, e.g., *Sequencing* in AGQA v2, *Causal* & *Temporal* in NExT-QA, *Interaction* & *Prediction* in STAR, and *Event* in Env-QA. These results demonstrate that our proposed model can effectively address the unique challenges of long-form video QA.

4.4. Comparison with Baselines

Here we devise several alternative solutions for long-form video modeling to replace our proposed ISTA. Specifically, in our CLIP-based MIST framework, we compare ISTA against other solutions, by fine-tuning the same pre-training input representation on AGQA v2 dataset.

- MeanPool: It simply takes the average of frame features as the representation of the whole video.
- Trans.-Frame: We follow the seqTransf type in CLIP4Clip, utilizing a Transformer to perform self-attention over frame features to represent the video.
- Trans.-Patch: This model is similar to Trans.-Frame, but it performs self-attention over all patch tokens.
- Divided STA: We follow TimeSformer [4] in the video classification model to perform uni-modal two-step Space-Time Attention over image patches.

From the results in Tab. 5, we can see that ISTA achieves substantial improvement over other variants with larger than 3% improvement on the overall accuracy. In addition, we find that for long-form VideoQA, the Transformer-based answer prediction models are much better than the MeanPool method, while in the video-text retrieval field, sometimes mean pooling is even better. The reason could be that the content of a long-form video is often complex and di-

Method	AGQA v2			NExT-QA			
	Binary	Open	All	C.	T.	D.	All
MIST w/o. SS	55.37	47.50	51.40	51.24	51.39	65.43	53.49
MIST w/o. RS	58.18	50.14	54.13	54.32	56.14	66.56	56.81
MIST w/o. STA	50.93	36.75	43.79	48.99	43.92	60.37	49.12
MIST - CLIP	58.28	50.56	54.39	54.62	56.64	66.92	57.18

Table 6. Ablations results of ISTA on AGQA v2 and NExT-QA.

verse, and a simple method for aggregating all frame features, such as mean pooling, may cause information loss. And long-form video QA requires more powerful temporal and spatial reasoning ability to focus on some details of a video, while mean pooling only performs well on capturing overall content.

Moreover, we can see that it is helpful to consider region information in long-form QA (Divided STA and Trans.-Path outperform Trans.-Frame). But, neither dense self-attention nor divided STA considers the interaction among multi-grained concepts; thus, the performance improvement is limited. And after integrating different granularities of visual concepts during reasoning, our method benefits the performance. All the above findings show that our method is effective, and transferring pre-trained transformers to long-form video QA is a challenging topic worth exploring.

4.5. Ablation Study

In this section, we propose several sets of variants of MIST to show the effectiveness of its key components.

Effect of each component in ISTA. We ablate key modules in ISTA layer, i.e., Segment Selection, Region Selection, or Self-attention layer, denoted as MIST w/o. SS/RS/STA, respectively:

- MIST w/o. SS: It removes the Segment Selection module, and only performs region selection. Patch features with word features are fed into the self-attention module.
- MIST w/o. RS: It removes Segment Selection module. All region features within selected segments are fed into self-attention layer.
- MIST w/o. STA: The segment features and selected region features are mean pooled as the output of ISTA.

The results of these variants on AGQA v2 and NExT-QA are shown in Tab. 6. We can see that removing Segment Selection causes a larger than 3% accuracy drop. The reason could be that removing it will introduce a lot of irrelevant region information when predicting the answer and thus hurt the performance. Tab. 6 also shows that Segment Selection is important for multi-event reasoning because removing it hurts the performances on questions requiring temporal reasoning, i.e., *Causal* and *Temporal*.

In addition, the performance drop on both datasets is significant when removing Spatial-temporal self-attention. The reason may be similar to MeanPool. We need a powerful model to capture multi-grained reasoning.

Moreover, we can see that removing spatial attention doesn't hurt performance too much. The number of objects in the video frames is relatively small (compared with natural scene images in image QA), and after temporal attention, the patch number has already been greatly reduced. So, the existing model is able to effectively focus on the appropriate objects. But, It is worth mentioning that we can reduce the computation cost by using a spatial selection module. It may be useful when we face high-resolution or extremely complex videos in the future.

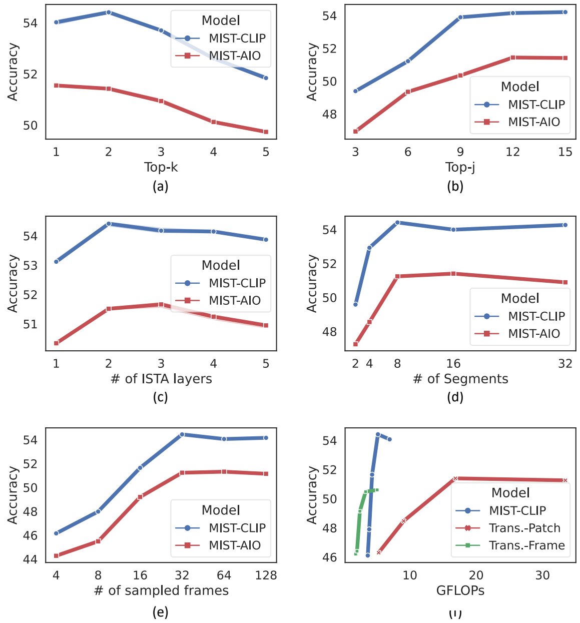
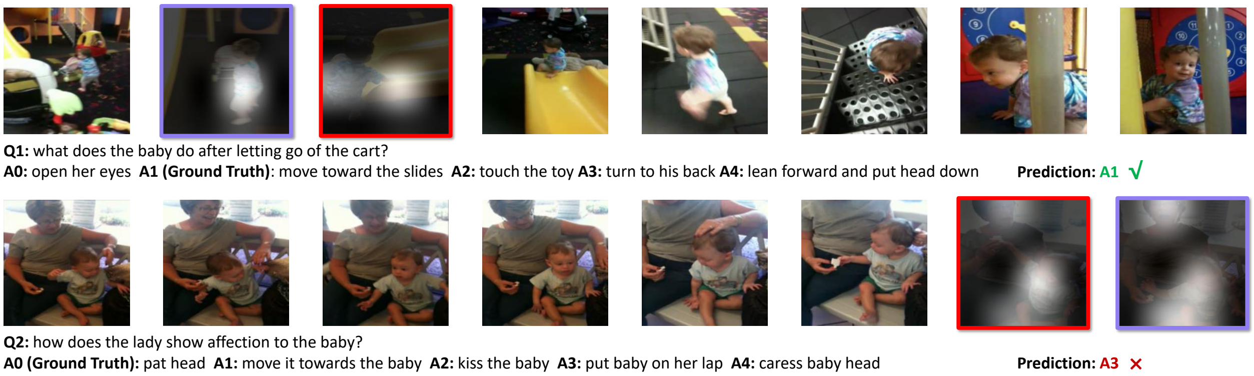


Figure 5. **Performances of MIST with different settings.** (a-e) Performances of MIST with different hyper-parameters on AGQA v2. (f) Performance of variants of MIST under different GFLOPs on AGQA v2, where GFLOPs rise with the number of sampled frames increase.

Effects of different ISTA configurations. In this part, we try different configurations of model architecture, including a number of selected segments Top_k, select patches Top_j, ISTA layers L, and the number of segments K. The results are shown in Fig. 5 (a-d).

First, Fig. 5 (a) shows that the performance is relatively good under the small Top_k. The performance slightly drops if Top_k further increases. The reason could be that large k will introduce either some incorrect segments or repeated segments. Incorrect segments will bring misleading information causing performance drops. Repeated segments lead to a larger number of repeated region features, causing it difficult for the model to focus on question and segment information. For the number of selected patches Top_j, as shown in Fig. 5 (b), we can see that with the increase of Top_j, the performance first increases and then reaches stability. The reason for this phenomenon could be that when selecting too few image regions, it may incorrectly filter some regions used for answering questions. And when the



Q1: what does the baby do after letting go of the cart?
A0: open her eyes **A1 (Ground Truth)**: move toward the slides **A2:** touch the toy **A3:** turn to his back **A4:** lean forward and put head down **Prediction:** A1 √

Q2: how does the lady show affection to the baby?
A0 (Ground Truth): pat head **A1:** move it towards the baby **A2:** kiss the baby **A3:** put baby on her lap **A4:** caress baby head **Prediction:** A3 ×

Figure 6. **Qualitative results of MIST on NExT-QA dataset.** We visualize its prediction results along with spatial-temporal attention, where the frames with purple and red outlines indicate the highest temporal attention score in the first and second ISTA layers, respectively.

selected regions increase, though it introduces some irrelevant regions, since the patch number after segment selection is already relatively small, the self-attention module can effectively attend to relevant regions.

For the number of ISTA layers, as shown in Fig. 5 (c), with the increase of L, the performance increases first and then reaches stability or slightly drops. It shows that stacking several layers of ISTA can benefit multi-event reasoning. In addition, the performance doesn't constantly increase with larger L. This is probably due to (1) the datasets are not large enough to train a deeper network and (2) the questions usually only involving two or three events, so considering more events may not bring more benefits. Fig. 5 (d) shows that when varying the number of video segments, performance tends to suffer when the videos are under-segmentation, because, in this case, each segment spans a relatively long duration, and hence the Segment Selection module is useless. More importantly, all those findings imply that MIST is effective in multi-event reasoning by attending to multiple segments.

4.6. Computation Efficiency

In Fig. 5 (e), we can see that the accuracy increases significantly when sampling more frames. It indicates that sampling more frames for long video QA tasks could be necessary. Though current datasets don't provide videos with several minutes or hours duration, such long videos are likely to be encountered in real application scenarios. Efficiency issues thus could be a more crucial consideration in such cases. In Fig. 5 (f), we compare GFLOPs vs. accuracy for ours against other long-form video QA methods. It can be seen that the standard Transformer over patches is computationally expensive. The frame-based method is lightweight in computation, but its performance is limited. Our method requires only a little extra computation but achieves much better performance. It is also worth mentioning that MIST doesn't enlarge model size for higher efficiency. Compared with other methods, it only contains some extra shallow networks for spatial-temporal attention.

4.7. Qualitative Results

We visualize some success and failure cases from the NExT-QA dataset in Fig. 6. It can be seen that our model can explicitly select video clips and image regions relevant to the question. We can also find that it is difficult for the model to correctly select segments and regions, when the question mainly involves some concepts related to social emotions. Existing pre-trained models may not well understand the correspondence between abstract concepts and videos. However, we believe that these issues can be alleviated by proposing better pre-trained models on short videos, and our method is easy to build upon the stronger ones.

5. Conclusion and Future Work

This paper introduces Multi-modal Iterative Spatial-temporal Transformer for long-form VideoQA, which decomposes dense self-attention into a cascade segment and region selection module to increase the computation efficiency along with a self-attention layer to reason over various grained visual concepts. In addition, by iteratively conducting selection and attention over layers, MIST better performs multi-event reasoning. Experimental results on four VideoQA datasets show its effectiveness and advantages in efficiency and interpretability. For future work, although MIST has increased the number of sample frames, the ability to capture high-frequency motion may still need to be improved. In addition, patch features naturally have some limitations in complex object-level reasoning. Recently, there have been some pre-trained models for specifically modeling actions and objects. It may be interesting to try more types of pre-trained models or even combine many of them to achieve more general reasoning.

Acknowledgements. This project is supported by the National Research Foundation, Singapore under its NRFF Award NRF-NRFF13-2021-0008. The computational work for this article was partially performed on resources of the National Supercomputing Centre, Singapore.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022. 2, 5, 6

[2] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1728–1738, 2021. 1, 2

[3] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv preprint arXiv:2106.08254*, 2021. 2

[4] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, volume 2, page 4, 2021. 3, 6

[5] Shyamal Buch, Cristóbal Eyzaguirre, Adrien Gaidon, Jiajun Wu, Li Fei-Fei, and Juan Carlos Niebles. Revisiting the" video" in video-language understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2917–2927, 2022. 2, 5, 6

[6] Joya Chen, Kai Xu, Yuhui Wang, Yifei Cheng, and Angela Yao. Dropit: Dropping intermediate tensors for memory-efficient dnn training. In *ICLR*, 2023. 3

[7] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 104–120, 2020. 1

[8] Feng Cheng and Gedas Bertasius. Tallformer: Temporal action localization with long-memory transformer. *arXiv preprint arXiv:2204.01680*, 2022. 3

[9] Han Fang, Pengfei Xiong, Luhui Xu, and Yu Chen. Clip2video: Mastering video-text retrieval via image clip. *arXiv preprint arXiv:2106.11097*, 2021. 3

[10] Tsu-Jui Fu, Linjie Li, Zhe Gan, Kevin Lin, William Yang Wang, Lijuan Wang, and Zicheng Liu. Violet: End-to-end video-language transformers with masked visual-token modeling. *arXiv preprint arXiv:2111.12681*, 2021. 1, 2

[11] Difei Gao, Ruiping Wang, Ziyi Bai, and Xilin Chen. Env-qa: A video question answering benchmark for comprehensive understanding of dynamic environments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1675–1685, 2021. 2, 5, 6

[12] 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. 2

[13] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022. 1

[14] Madeleine Grunde-McLaughlin, Ranjay Krishna, and Maneesh Agrawala. Agqa: A benchmark for compositional spatio-temporal reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11287–11297, 2021. 2, 5

[15] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *International Conference on Learning Representations*, 2021. 3

[16] Md Mohaiminul Islam and Gedas Bertasius. Long movie clip classification with state-space video models. *arXiv preprint arXiv:2204.01692*, 2022. 3

[17] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pages 4651–4664. PMLR, 2021. 3

[18] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016. 4

[19] 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. 6

[20] 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. 2

[21] 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. 5

[22] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73, 2017. 1

[23] Thao Minh Le, Vuong Le, Svetha Venkatesh, and Truyen Tran. Hierarchical conditional relation networks for video question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9972–9981, 2020. 2, 6

[24] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7331–7341, 2021. 1, 6

[25] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. Tvqa: Localized, compositional video question answering. *arXiv preprint arXiv:1809.01696*, 2018. 2

[26] Jie Lei, Licheng Yu, Tamara Berg, and Mohit Bansal. Tvqa+: Spatio-temporal grounding for video question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8211–8225, 2020. 6

[27] Stan Weixian Lei, Yuxuan Wang, Dongxing Mao, Difei Gao, and Mike Zheng Shou. Assistsr: Affordance-centric question-driven video segment retrieval. *arXiv preprint arXiv:2111.15050*, 2021. 1

[28] Xiangpeng Li, Jingkuan Song, Lianli Gao, Xianglong Liu, Wenbing Huang, Xiangnan He, and Chuang Gan. Beyond rnns: Positional self-attention with co-attention for video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 8658–8665, 2019. 2

[29] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137, 2020. 1

[30] Yan-Bo Lin, Jie Lei, Mohit Bansal, and Gedas Bertasius. Eclipse: Efficient long-range video retrieval using sight and sound. *arXiv preprint arXiv:2204.02874*, 2022. 3

[31] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 12-in-1: Multi-task vision and language representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10437–10446, 2020. 1

[32] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 508:293–304, 2022. 1, 3

[33] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2630–2640, 2019. 2

[34] Volodymyr Mnih, Nicolas Heess, Alex Graves, et al. Recurrent models of visual attention. *Advances in neural information processing systems*, 27, 2014. 3

[35] Jungin Park, Jiyoung Lee, and Kwanghoon Sohn. Bridge to answer: Structure-aware graph interaction network for video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15526–15535, 2021. 2

[36] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2085–2094, 2021. 3

[37] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1, 2, 3, 5, 6

[38] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018. 1

[39] 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*, pages 4631–4640, 2016. 2

[40] Alex Jinpeng Wang, Yixiao Ge, Rui Yan, Yuying Ge, Xudong Lin, Guanyu Cai, Jianping Wu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. All in one: Exploring unified video-language pre-training. *arXiv preprint arXiv:2203.07303*, 2022. 2, 5, 6

[41] Benita Wong, Joya Chen, You Wu, Stan Weixian Lei, Dongxing Mao, Difei Gao, and Mike Zheng Shou. Assistq: Affordance-centric question-driven task completion for egocentric assistant. In *European Conference on Computer Vision*, pages 485–501. Springer, 2022. 1

[42] Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. Star: A benchmark for situated reasoning in real-world videos. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. 2, 5

[43] Chao-Yuan Wu and Philipp Krahenbuhl. Towards long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1884–1894, 2021. 3

[44] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9777–9786, 2021. 2, 5

[45] Junbin Xiao, Pan Zhou, Tat-Seng Chua, and Shuicheng Yan. Video graph transformer for video question answering. In *European Conference on Computer Vision*, pages 39–58. Springer, 2022. 2, 6

[46] Yujia Xie, Hanjun Dai, Minshuo Chen, Bo Dai, Tuo Zhao, Hongyuan Zha, Wei Wei, and Tomas Pfister. Differentiable top-k with optimal transport. *Advances in Neural Information Processing Systems*, 33:20520–20531, 2020. 4

[47] 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, 2017. 2

[48] Hongwei Xue, Yuchong Sun, Bei Liu, Jianlong Fu, Ruihua Song, Houqiang Li, and Jiebo Luo. Clip-vip: Adapting pre-trained image-text model to video-language representation alignment. *arXiv preprint arXiv:2209.06430*, 2022. 3

[49] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Just ask: Learning to answer questions from millions of narrated videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1686–1697, 2021. 5, 6

[50] 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/CVF Conference on Computer Vision and Pattern Recognition*, pages 8807–8817, 2019. 2

[51] Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. Merlot reserve: Neural script knowledge through vision and language and sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16375–16387, 2022. 1, 6

[52] Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. Merlot: Multimodal neural script knowledge models. *Advances in Neural Information Processing Systems*, 34:23634–23651, 2021. 1

[53] Shuai Zhao, Linchao Zhu, Xiaohan Wang, and Yi Yang. Centerclip: Token clustering for efficient text-video retrieval. *arXiv preprint arXiv:2205.00823*, 2022. 3

[54] Zhou Zhao, Zhu Zhang, Shuwen Xiao, Zhou Yu, Jun Yu, Deng Cai, Fei Wu, and Yueting Zhuang. Open-ended long-form video question answering via adaptive hierarchical reinforced networks. In *IJCAI*, volume 2, page 8, 2018. 2