ATM: Action Temporality Modeling for Video Question Answering

Junwen Chen
chenjunw@msu.edu
Michigan State University

Jie Zhu
zhujie4@msu.edu
Michigan State University

Yu Kong
yukong@msu.edu
Michigan State University

ABSTRACT

Despite significant progress in video question answering (VideoQA), existing methods fall short of questions that require causal/temporal reasoning across frames. This can be attributed to imprecise motion representations. We introduce Action Temporality Modeling (ATM) for temporality reasoning via three-fold uniqueness: (1) rethinking the optical flow and realizing that optical flow is effective in capturing the long horizon temporality reasoning; (2) training the visual-text embedding by contrastive learning in an action-centric manner, leading to better action representations in both vision and text modalities; and (3) preventing the model from answering the question given the shuffled video in the fine-tuning stage, to avoid spurious correlation between appearance and motion and hence ensure faithful temporality reasoning. In the experiments, we show that ATM outperforms previous approaches in terms of the accuracy on multiple VideoQAs and exhibits better true temporality reasoning ability.

CCS CONCEPTS

• **Computational methodologies-Artificial intelligence-Computer Vision**;

KEYWORDS

Video Question Answering, Action, Static Bias

ACM Reference Format:
Junwen Chen, Jie Zhu, and Yu Kong. 2023. ATM: Action Temporality Modeling for Video Question Answering. In *Proceedings of the 31st ACM International Conference on Multimedia (MM '23), October 29-November 3, 2023, Ottawa, ON, Canada.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3581783.3612509

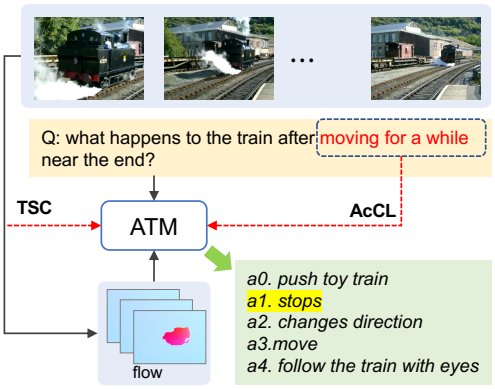

Figure 1: ATM addresses VideoQA featuring temporal reasoning by (1) an appearance-free stream *i.e.* optical flow to extract precise motion cues, (2) action-centric contrastive learning (AcCL) for action-plentiful cross-modal representation, and (3) a temporal sensitivity-aware confusion (TSC) loss to avoid learning a shortcut between temporality-critical motion and appearance.

1 INTRODUCTION

Video question answering (VideoQA) is an interactive AI task, which enables many downstream applications e.g. vision-language navigation and communication systems. It aims to answer the natural language question given the video content. Recent VideoQA benchmark [52] has gone beyond the understanding of descriptive content like "*A baby is crying*" and started to provide effective diagnostics for the models on solving temporal reasoning and causal reflection, *e.g.* "*The train stops after moving for a while*". To accurately answer the question, a VideoQA model needs to detect the object "train", recognize the "railway" scene, more importantly, ground the action "move" and "stop" and understand their temporal relations. The questions are unconstrained and complex, and thus, it is necessary to have a visual-text model that has the reasoning capability toward all aforementioned contents.

Recent advanced VideoQA models have shown the capability of learning from the descriptive contents [29, 30], thanks to the success of cross-modal transformers [28, 32]. However, the temporality reasoning in videos remains a great challenge, since these VideoQAs are only capable of holistic recognition of static content in a video. Recent work attempts to solve this issue by (1) enhancing the video representation with fine-grained dynamics [53, 54] and (2) answering by grounding to question-critical visual evidence [35, 36]. But it is hard to achieve a precise grounding, without the ground-truth of temporal boundaries for training. The state-of-the-art method VGT [54] proposes to model the atomic actions across frames from the spatio-temporal dynamics of objects. In this way, the fine-grained dynamics can be captured. But their model may rely on the static bias *i.e.* object appearance, as shortcuts from videos while the causal factors *i.e.* the dynamics are overlooked in training. In this paper, we address the importance of precise and faithful modeling of actions for the VideoQA task.

MM '23, October 29-November 3, 2023, Ottawa, ON, Canada

ACM ISBN 979-8-4007-0108-5/23/10…$15.00
https://doi.org/10.1145/3581783.3612509

We propose Action Temporality Modeling (ATM) to address the challenging temporality VideoQA (as shown in Fig. 1). A promise of VideoQA compared to ImageQA is to examine the temporal relation reasoning regarding motion information. As the targeted video is continuous, actions across a long video usually share the same scene in short moments. We realize that (1) leveraging an appearance-free stream *e.g.* optical flow as input, though the flow stream may become less considered in recent action recognition methods [4, 16], is still important in VideoQA. Because flow can capture the subtle transition in long horizon and aid the temporality reasoning. (2) ATM trains the visual-text encoding in a contrastive manner. Questions are usually unconstrained in the real world. Action may be only a small portion of the question, which is easily overwhelmed by other information such as objects. To learn an action-plentiful cross-modal embedding, we develop a novel action-centric contrastive learning (AcCL) before fine-tuning VideoQA. Specifically, it parses an action phrase from a question and encourages a feature alignment between the video and the parsed action phrase alone, discarding other textual information. The merit of the AcCL is that both video and text encoders are trained to focus on actions, mitigating the backbone's representation bias towards the static visual appearance in videos.

Based on the learned representations, we further introduce a novel temporal sensitivity-aware confusion loss (TSC) in VideoQA finetuning. It prevents a model from answering a temporality question if the corresponding video is shuffled in the temporal domain, thus avoiding simply learning the shortcut correlation to the static content. Note that VideoQA contains a lot of descriptive questions that can be answered invariant to temporal change. Thus, we only apply the confusion loss to temporal-sensitive questions that contain temporal keywords.

Thanks to these components, the proposed ATM outperforms all of the existing methods on three commonly used VideoQA datasets. It is worth noting that our method without external vision-language pretraining can surpass the existing method that relies on large-scale pre-training by a clear margin. Moreover, we devise a new metric that quantifies the accuracy difference between conditioned on a full video and conditioned on a single frame, which reveals the VideoQA's true temporality reasoning ability. Results show that our model experiences a larger performance escalation from a single frame to a full video, which demonstrates ours relies on less appearance bias and handles temporal reasoning in a faithful manner. To summarize, our main contributions are as follows:

- We propose the ATM to address VideoQA featuring temporal dynamic reasoning by a faithful action modeling. Our action-centric contrastive learning learns action-aware representations from both vision and text modalities. We realize an appearance-free stream is effective in the multi-event temporality understanding across frames.
- We fine-tune the model with a newly developed temporal sensitivity-aware confusion loss that mitigates static bias in temporality reasoning.
- Our method is more accurate than all existing methods on three widely used VideoQA datasets. By a new metric, we also indicate that our method addresses temporality reasoning more faithfully.

2 RELATED WORK

Video Question Answering. Escalating ImageQA [2], VideoQA [29, 31, 34, 52, 57, 62] is enriched with reasoning about temporal nature. Prior arts [26, 43, 53] on VideoQA focus on learning an informative video content representation and a cross-modal fusion model to answer the question. An informative video representation is usually hierarchical, fusing object-, frame- and clip-level representations, which are extracted by graph neural network [23, 36, 43], relation learning or transformers. While those VideoQA methods achieve compelling results on VideoQA benchmarks, they mainly answer descriptive questions for the video content, such as questions that holistic recognize the main actions/objects across frames.

Recent benchmark [52] begins to challenge the temporal relationship reasoning ability, as actions in videos are diverse and causally dependent. Those methods that are only capable of descriptive content recognition cannot perform well, because they hardly capture the subtle transitions in the same scene in long-horizon. To this end, recent work [53, 54] proposes to encode video as a local-to-global dynamic graph of spatiotemporal objects, so that the interaction relations can be encoded. However, the VideoQA models [17, 53, 54] built upon the dynamic graph of patches may easily be distracted by the object's appearance and capture limited motion information. We alleviate the distraction by a novel two-stage training to ensure a faithful representation of motions that are critical for temporality reasoning. Specifically, we propose a novel contrastive learning in which the objectives that are the parsed action phrases in questions and a novel confusion loss to prevent question answering if the video is temporally shuffled.

Concurrent work HiTEA [60] also introduces temporal shuffling, but the shuffling is used as evaluation test to metric the temporal reliance of datasets and VLP models, while our method leverages this in training a temporal reliable VideoQA. Concurrent work CoVGT [55] also investigates the contrastive learning in VideoQA. But as our action-centric contrastive learning aims at learning a faithful action representation, our contrastive objectives are action phrase in question, different from CoVGT's question and QA pairs. Concurrent work Verbs in Action [41] also proposes to improve verb understanding in Video-language task. Verbs in Action focuses on training by the generated the new captions with hard mined verb based on large language model, while our AcCL extracts the action phrase and encourages learning motion representations agnostic to the appearance information.

Static Bias in Video. The uniqueness of video lies in the potential to go beyond image-level understanding of the static content *i.e.* scenes, objects, and people to evaluate the temporality reasoning ability of multiple events. However, for many video(+language) tasks and datasets, given just a single frame of video, an existing image-centric model can achieve surprisingly high performance, comparable to the model using multiple frames. The strong single-frame performance suggests that the video representation is biased towards the static appearance information, namely "static appearance bias". Existing work [5, 8, 27, 33] reveals this kind of bias in action recognition dataset [6, 49] and retrieval dataset [39, 40]. Circling around the fundamental video task action recognition, [8, 33] analyze the role of temporality in action recognition and inspires the

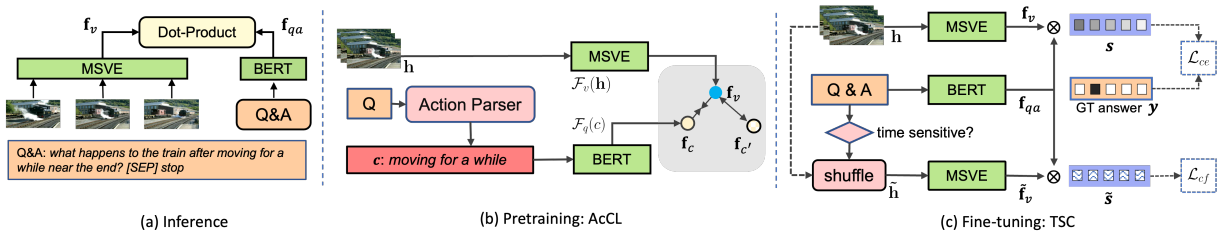

Figure 2: Framework Overview. Following the recent VQAs [54, 58], we solve VideoQA by a similarity comparison between video and text (a). To achieve this, we formulate the training procedure into two stages. Before finetuning, we present a novel action-centric contrastive learning (AcCL) to guide the visual and text representation expressive for action information (b). After that, we fine-tune the VideoQA (c) by a newly developed temporal sensitivity-aware confusion loss (TSC) to prevent leveraging static bias in temporality reasoning.

subsequent development of profound faithful evaluations [33, 48] and model structures [13, 15, 16, 37].

To address the challenging temporality reasoning in multi-modal scenarios *i.e.* VideoQA, motion representations, unbiased toward appearance, are necessary. As VideoQA requires a deep understanding of open-vocabulary action semantics, existing VideoQAs [26, 53] extract the motion features based on backbones pre-trained on a large-scale action recognition dataset [6]. As mentioned, static bias exists in action recognition, which makes the motion representations not the causal factors of actions, thus useless to temporality reasoning. Existing methods[8, 33] mitigate the static bias in action recognition and is evaluated on fine-grained action recognition[33, 48], where the scene context is the same across the different actions. However, in fine-grained action recognition, motion is the most critical information, which is different from VideoQA where object/entity appearance is inevitable.

To mitigate static bias in VideoQA, IGV [36] and EIGV [35] are proposed to ground the question-critical scenes across frames as the evidence of yielding the answers. However, the dominant content of a question is appearance information *e.g.* people, objects, and locations. The grounding may pay less attention to the actions that are critical for temporality understanding and be not precise as no ground truth boundaries are provided. Our method designs two simple yet effective schemes that learn faithful visual and text representations informative for action and temporality. We also revisit the early action recognition work [6, 50] and enhance the motion representation with an appearance-free stream.

3 METHODOLOGY

Figure 2 gives an overview of ATM framework. Our framework addresses the VideoQA task that challenges the temporal reasoning of dynamics in a video. Following the recent VQAs [54, 58], we solve VideoQA by a similarity comparison between video and QA pair (Figure 2-a). To achieve this, we formulate the training procedure into two stages. In the first stage (Figure 2-b), we present a novel action-centric contrastive learning (AcCL, Sec. 3.3), which makes the visual and text representation expressive for action information. After that, we finetune the VideoQA (Figure 2-c) by a newly developed temporal sensitivity-aware confusion loss (TSC, Sec. 3.4) to prevent leveraging static bias in temporality reasoning. We detailed the video and text encoding in Sec. 3.2.

3.1 Preliminaries

Given a video $\mathbf{h}$ and a question q, VideoQA task aims at combining the two modalities $\mathbf{h}$ and q to predict the answer a. Following existing VideoQA work [36, 53, 54], we predict the answer by selecting the best matched a^* from many candidates $\mathcal{A}$ of a question q, given the corresponding video $\mathbf{h}$:

$$a^* = \arg\max_{a\in\mathcal{A}} \mathcal{F}_W(a|q, \mathbf{h}, \mathcal{A}), \tag{1}$$

where $\mathcal{F}_W$ denotes the mapping function with learnable parameters W. The candidates $\mathcal{A}$ are multi-choices in multi-choiceQA or a global answer list in open-ended QA.

Prior arts on VideoQA usually build $\mathcal{F}_W$ as a cross-attention transformer [28, 65], which takes a holistic token sequence containing video, question and each candidate answer as input and classifies the answers as output. Recent work VGT [54] and VQA-T [59] propose to design $\mathcal{F}_W$ as two unimodal transformers that encode video and question-answer pair respectively and compare the visual-text similarity for each answer as output:

$$s_a = \mathcal{F}_v(\mathbf{h})\, \mathcal{F}_q([q;a])^\top, \tag{2}$$

in which $\mathcal{F}_v$ denotes the video encoder and $\mathcal{F}_v(\mathbf{h}) \in \mathbb{R}^d$ is the video' global feature obtained by mean-pooling the features across T frames. Likewise, $\mathcal{F}_q$ denotes the text encoder and $\mathcal{F}_q([q;a]) \in \mathbb{R}^d$ is the feature vector of a question-answer pair, where $[;]$ indicates the concatenation of question and answer text. The visual-text similarity s_a is obtained via a dot-product of video and text features *w.r.t.* the answer a. The optimal answer is selected by maximizing the similarity score from the candidate in the pool $\mathcal{A}$:

$$a^* = \arg\max_{a\in\mathcal{A}}(s_a). \tag{3}$$

Following existing work [36, 54], we implement $\mathcal{F}_q$ by the BERT [11] to extract text features. For video modality, many existing methods [53, 54] extract features in multiple streams including object-level and frame-level. Following them, we also formulate $\mathcal{F}_v$ as a multi-stream video encoder (MSVE), by which object features are encoded as $f_o \in \mathbb{R}^{T\times d}$ and frame features are encoded as $f_r \in \mathbb{R}^{T\times d}$. The object/frame feature extraction and transformer-based encoding are exactly the same as state-of-the-art method VGT [54] for a fair comparison.

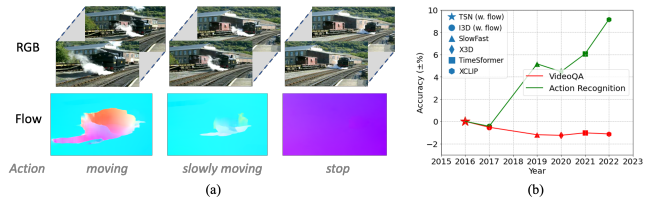

Figure 3: Motivation of using an appearance-free stream for motion representation in VideoQA task. The example in (a) shows the state transition on a train, from moving to stopping. We can see flow provides better cues for the actions than RGB. (b) summarizes the relative performance gain/loss of different video backbones pivot on TSN, for both action recognition (Kinetics [6]) and VideoQA (NextQA [52]), which shows appearance-free stream *i.e.* flow is necessary for VideoQA. The numbers for action recognition (green curves) are reported in their paper for Kinetics-400. The numbers for VideoQA are derived based on our implementation on Next-QA.

3.2 Rethinking motion representations in VideoQA

In video feature extraction of both the existing methods [26, 53, 54] and ours, frame-level features $f_r \in \mathbb{R}^{T\times d}$ and object features $f_o \in \mathbb{R}^{T\times d}$ both represent appearance. Optionally, they [26, 53] apply a pre-trained 3D Conv network [6] on the neighboring frames to capture motions. However, VideoQA studies the temporality of the actions in a video where multiple actions are performed across frames. As a video captures continuous information, these actions usually share the same scene context and are performed by the same people and on the entity. In this case, although 3D Conv can capture motions, neighboring RGB frames may be too redundant to precisely model the actions. For example, in Figure 3-a, it is hard to recognize "the train is stopping" in the last clip from RGB. This inspires us to enhance the video representation by a stream, where the appearance information is least and hence the motions are highlighted. To this end, we resort to optical flow that describes the apparent motion of individual pixels on the image. As shown in Figure 3-a's example, flow maps provide better cues to understand the state transition of objects *e.g.* "train" was moving (in the first and second clip) and stopped (in the third clip).

As VideoQA requires the open-vocabulary semantic understanding of motions, we use the backbone pretrained on a large-scale action recognition dataset Kinetics-400 [24] to extract flow features. Flow features are extracted as per appearance frame timestamps as $f_m \in \mathbb{R}^{T\times d}$. To fuse the object, appearance, and flow streams, our MSVE applies MLPs and a learnable multi-head self-attention layer MSA with position embedding to model the temporal interactions upon the multi-stream features and finally mean-pool the frames to obtain the global video representation f_v.

$$f_v = \text{Mean-Pool}(\text{MSA}(\text{MLP}([f_o; f_r; f_m]))) \tag{4}$$

Note that we should not ignore the appearance information in VideoQA task, as the questions are unconstrained and may contain characters, objects and locations that need to be grounded to videos. This is different from the action segmentation [12] or skeleton-based activity recognition [13, 64], where motion is the only critical information.

We revisit the fundamental video understanding task *i.e.* action recognition, in which the early methods *e.g.* TSN [6, 50] also utilized optical flow to capture motions. As shown in Figure 3-b, we observe that although the existing powerful RGB-based backbones e.g. SlowFast [16], X3D [15], TimeSformer [4] and XCLIP [42] achieve good performance w/o appearance-free stream i.e. optical flow in action recognition, they are less helpful in VideoQA compared to the early methods w/ appearance-free stream. This demonstrates that towards longer-horizon temporality understanding, a stream free of appearance is necessary. Detailed comparison will be discussed in Sec. 4.5.3.

3.3 Action-centric Contrastive Learning (AcCL)

As aforementioned, question-answer contains much information including characters, objects, and locations. Actions, the important reasoning objective in videos, may only occupy a small portion of QA text and be neglected in the cross-modal alignment. Since VideoQA takes the alignment of global video features and a full QA sequence features as the optimization objective, the precise motion information obtained from Sec. 3.2 may not be well exploited. But a VideoQA model, capable of answering temporal questions, should make good use of motion.

To this end, we propose a novel training scheme that conducts contrastive learning for visual-language matching before finetuning VideoQA objective. Different from conventional VL contrastive learning, the contrastive learning in our method is action-centric. It encourages the video representation to be aligned with the representation of **action phrase** that is parsed from the question. That is to say, other information such as entity, location, objects are not present in the text for matching. For example, in the question "what happens to the train after moving for a while?", the action phrase to be aligned with the whole video clip is "moving for a while". Under this matching objective, the video representation has to focus on precise motions, leading to a deep understanding of temporality. In specific, we propose a contrastive loss $\mathcal{L}_{pt}$ to update the encoders $\mathbb{F}_v, \mathbb{F}_q$:

$$\mathcal{L}_{pt} = \sum_i \log\left(\frac{\exp(s_c)}{\exp(s_c) + \sum_{c'\in\mathcal{N}_i}\exp(s'_c)}\right), \tag{5}$$

where $\mathcal{N}_i$ denotes the negative pool of action phrase for the i-th sample, i.e., action phrases from the questions that are unpaired to the video $\mathbf{h}$. $s_c = \mathcal{F}_v(\mathbf{h})\mathcal{F}_q(c)^\top$ is the similarity between the action phrase c and video $\mathbf{h}$ of the i-th sample. It encourages the video representation closer to its paired action phrase c and far away from the unpaired c' that are randomly sampled into the mini-batch. Thus, by contrastive to many other action phrases $c' \in \mathbb{N}_i$ in the dataset, the motion in vision and the textual action are better mined and aligned. The motion-plentiful features and model provide a good starting point for VideoQA finetuning.

Many VideoQA task benefits from contrastive learning based video language pretraining [28, 63] from large-scale video-language data [3], which is also reflected in the SoTA model in our task [54].

However, our AcCL is just conducted on our task datasets themselves, without resorting to any of the external training data, and has already been more effective than VGT [54] with external data pretraining, while taking much less training resources.

3.4 Temporal Sensitivity-aware Confusion Loss

At the end of Sec. 3.2, we mention that although we have an appearance-free stream to extract precise motions, the appearance stream is indispensable. Unfortunately, the appearance stream, even fused with an appearance-free stream, provides the possibility to model action biased towards scene/object context [8]. To mitigate this issue, we propose to prevent the model from answering a question if the corresponding video is randomly ordered in the temporal domain. Our motivation is that the temporality reasoning requires the model to infer the inter-action relations across temporal, such as "stop (action 1) **after** moving for a while (action 2)". Thus, if the video is randomly shuffled, the "after" relation no longer exists. In this case, a reliable network should be unable to answer the "stop" to the question like "What is the train doing after moving for a while?".

Motivated by this, we design a confusion loss that takes as input the shuffled video $\tilde{\mathbf{h}}$ and question-answer $[q;a]$:

$$\begin{aligned}\mathcal{L}_{cf}^{(n)}(\hat{\mathbf{p}},\hat{\mathbf{p}}) &= -\sum_{j=1}^{|\mathcal{A}|}\hat{p}^{(j)}\log\hat{p}^{(j)},\\ \hat{p}^{(j)} &= \frac{\exp(\hat{s}^{(j)})}{\sum_{k=1}^{|\mathcal{A}|}\exp(\hat{s}^{(k)})},\end{aligned} \tag{6}$$

where $\hat{s}^{(j)} = \tilde{\mathbf{f}}_v\mathbf{f}_q^\top$ (denote the $\hat{s}^{(j)} \in [\hat{s}^{(1)},\dots,\hat{s}^{(|\mathcal{A}|)}]^\top$) is the inner-product similarity score for the j-th answer features $\mathbf{f}_q = \mathcal{F}_q([q;a])$ *w.r.t.* its shuffled video feature vector $\tilde{\mathbf{f}}_v = \mathcal{F}_v(\tilde{\mathbf{h}})$. The confusion loss is applied to encourage the maximization of the entropy of the predicted answer distributions over the multiple choices, given the shuffled video. This guides the model to produce confusing classification, so that the scene context invariant to temporal order change will be ignored in action relation modeling.

Many questions *e.g.* "Where is the video taken?" simply rely on descriptive content and can be answered even with the shuffled videos. Thus, the confusion loss is only applied to temporal-sensitive questions, *e.g.* the "after" question: "what does A do after raising her hand?" The temporal-sensitive questions contain specific English syntax, e.g. "before", "after", "when". We filter out the temporal-insensitive questions based on the existence of the syntaxes. The overall optimization objective is as follows.

$$\min\mathbb{E}_{q^{(n)}\sim Q_\tau}\left[\mathcal{L}_{ce}^{(n)}(\mathbf{y},\mathbf{p}) - \mathcal{L}_{cf}^{(n)}(\hat{\mathbf{p}},\hat{\mathbf{p}})\right], \tag{7}$$

where Q_τ denotes the set of questions that are temporally sensitive. $\mathcal{L}_{ce}^{(n)}$ is the cross entropy loss to metric if the probability over the candidates answers is $\mathbf{p} = [p^{(1)},...,p^{(|\mathcal{A}|)}]$ follows ground-truth answer y. $\mathcal{L}_{ce}^{(n)}$ is applied to all of the samples including the temporal-insensitive one, which is to optimize:

$$\min\mathbb{E}_{q^{(n)}\sim Q_{\setminus\tau}}\left[\mathcal{L}_{ce}^{(n)}(\mathbf{y},\mathbf{p})\right] \tag{8}$$

where $Q_{\setminus\tau}$ denotes the set of remaining temporally insensitive samples. The two loss are used for fine-tune the VideoQA after AcCL (see Sec. 3.3).

4 EXPERIMENTS

4.1 Datasets

NExT-QA [52] consists of 47.7K questions with answers in the form of multiple choices, which is annotated from 5.4K videos. It pinpoints the causal and temporal reasoning over the object interaction. Note that the causal questions *e.g.* "How", "Why" require the corresponding answers visible in the video and thus the causal questions also assess the multi-event temporality understanding.

TGIF-QA [22] contains 134.7K questions about repeated actions, state transitions and a certain frame, which is annotated from 91.8K GIFs. **MSRVTT-QA** [56] challenges a holistic visual recognition or description, which includes 10K annotated videos and 244K open-ended question-answer pairs.

4.2 Implementation Details

Appearance Features Following [53, 54], we decode the video into frames and sparsely sample 16 clips where each clip is in the length of 4 frames. To make a fair comparison with state-of-the-art VGT [54], we also the RoI aligned features as the object appearance features $f_o \in \mathbb{R}^{16*2048}$ pretrained by [1].

Motion Features We use denseflow [51] to extract the optical flow maps using videos' original FPS. Then, we use mmaction2 [9]-based ResNet from TSN [50] pre-trained on Kinetics-400 [6] to extract optical flow features. We uniformly distribute the flow maps into $K = 16$ clips per video and sample 5 frames as per each clip and obtain a 2048-d feature vector for a clip. Thus, motion features f_m for a video are $\mathbb{R}^{16*2048}$.

Action Phrase We parse the action phrases from questions using SpaCy parser [20] Specifically, we use dependency parsing to get the phrases in a question and use the pos-tag to find the verb in the question. Then we filter the phrases that contain the verb and select the shortest one as the action phrase. For example, for the question "what happens to the train after moving for a while near the end?", the action phrase is "moving for a while".

Action-centric Contrastive Learning We parse the action phrases from questions using SpaCy parser [20]. We use Adam optimizer [25] with cosine annealing learning schedule of PyTorch initialized at $1e-5$ on NVIDIA RTX A6000 at the maximum epoch of 10 among all of the datasets. Each batch contains 64 video-action pairs and forms 64 pairs in total for the contrastive learning.

Temporal Sensitivity-aware Confusion Loss English questions typically follow a syntactic structure. Temporal-sensitive questions contain specific syntax, *e.g.* "after", "before", "... when...", "...while..." and *etc.* . The remaining is descriptive questions, *e.g.* a count question, *"How many people are involved in the video?"*, which is insensitive to time. We detect the existence of the syntaxes and filter out the temporal-insensitive questions. For Next-QA, we have 17,681 temporal-sensitive questions and 16,451 temporal-insensitive questions in the training set. For T-Gif [34], as its "action" and "transition" splits focus on repeated actions and transitions respectively, all of the questions in those splits are temporally sensitive. For the open-ended QAs including TGif-FrameQA and

2*Methods	NExT-QA Val				NExT-QA Test			
	Acc@C	Acc@T	Acc@D	Acc@All	Acc@C	Acc@T	Acc@D	Acc@All
EVQA [2]	42.46	46.34	45.82	44.24	43.27	46.93	45.62	44.92
STVQA [22]	44.76	49.26	55.86	47.94	45.51	47.57	54.59	47.64
CoMem [18]	45.22	49.07	55.34	48.04	45.85	50.02	54.38	48.54
HCRN* [26]	45.91	49.26	53.67	48.20	47.07	49.27	54.02	48.89
HME [14]	46.18	48.20	58.30	48.72	46.76	48.89	57.37	49.16
HGA [23]	46.26	50.74	59.33	49.74	48.13	49.08	57.79	50.01
HQGA [53]	48.48	51.24	61.65	51.42	49.04	52.28	59.43	51.75
P3D-G [7]	51.33	52.30	62.58	53.40	-	-	-	-
IGV [36]	-	-	-	-	48.56	51.67	59.64	51.34
EIGV [35]	-	-	-	-	-	-	-	53.70
ATP[5]	53.1	50.2	66.8	54.30	-	-	-	-
VGT [54]	52.28	55.09	64.09	55.02	51.62	51.94	63.65	53.68
VGT* [54]	53.43	56.39	69.50	56.89	52.78	54.54	67.26	55.70
Ours	**56.04**	**58.44**	**65.38**	**58.27**	**55.31**	**55.55**	**65.34**	**57.03**

Table 1: Results of multi-choice QA on validation set and test set of NextQA [52] dataset. The best results are bolded. Note that the greyed out VGT* uses 0.18 million videos from webvid dataset [3] as pretraining, while the remaining include ATM do not pretrain on the external large-scale data. All of numbers for existing work are recorded from their papers. "-" indicates the missing results. Acc_C, Acc_T, Acc_D denote the accuracy for causality, temporality and descriptive questions.

MSRVTT [57], we do not apply the confusion loss as they focus more on the descriptive content.

Please refer to appendix for additional details.

4.3 Comparison with State-of-the-Art

Table 1 compares our method with existing state-of-the-art (SoTA) VideoQA methods on the widely used Next-QA dataset that feature the temporality reasoning. To ensure a fair comparison, ATM follows SoTA VGT [54] and uses the exactly same appearance feature extraction and applies DGT [54] to model the object features. From the table, we can observe that ATM outperforms all existing methods without external data pretraining, by at least 3.85% and 3.35% on val. and test splits respectively. The outperformance is across causal, temporal, and descriptive splits of the Next-QA dataset, which demonstrate that ATM is effective in various question types that span from short segment to full video, from causal to temporal, and from single to multiple event execution.

Moreover, ATM which comes without external large-scale pretraining, even surpasses the existing method that used large-scale pretraining on more than 0.18 million videos [3], by a clear margin of 1.38% and 1.33% on validation and test splits respectively. This demonstrates that ATM comprises of appearance-free motion features Sec. 3.2, action-centric contrastive learning Sec. 3.3 and temporal sensitive-aware confusion objective Sec. 3.4, which holistically models action temporality, is more effective than the global video-text matching while uses less training computation resources.

In ATP [5], the temporal modeling is performed on frames that are representative for single events and are encoded with CLIP model [46]. Our method also exceeds ATP [5] by a large margin of 3.97%. This shows in temporality-heavy tasks, precise and faithful motion modeling is more effective than selecting the informative single frame for an event. This validates that ATM to precisely model and reason about motion, sets the new SoTA on Next-QA [52] benchmark.

Furthermore, we compare ATM with SoTA on TGIF-QA in Table 2. Following the protocol, we use the same appearance features extracted by VGT [54] and extract the motion stream features. We observe that ATM set new SoTA for repeated actions, and transition in TGIF-QA, which shows ATM as a whole is also effective in the repeated action and object transition scenarios.

2*Models	TGIF-QA					2*MSRVTT-QA
	Action	Transition	Frame-QA	Action†	Transition†	
LGCN [21]	74.3	81.1	56.3	-	-	-
HGA [23]	75.4	81.0	55.1	-	-	35.5
HCRN [26]	75.0	81.4	55.9	55.7	63.9	35.6
B2A [43]	75.9	82.6	57.5	-	-	36.9
HOSTR [10]	75.0	83.0	58.0	-	-	35.9
HAIR [38]	77.8	82.3	60.2	-	-	36.9
MASN [47]	84.4	87.4	59.5	-	-	35.2
PGAT [44]	80.6	85.7	61.1	58.7	65.9	38.1
MHN [45]	83.5	90.8	58.1	-	-	38.6
ClipBERT* [28]	82.8	87.8	60.3	-	-	37.4
SiaSRea* [61]	79.7	85.3	60.2	-	-	41.6
MERLOT* [63]	94.0	96.2	69.5	-	-	**43.1**
VGT [54]	95.0	**97.6**	61.6	59.9	70.5	39.7
Ours [63]	**96.0**	97.3	61.6	**65.7**	**71.0**	40.3

Table 2: Results on TGIF-QA and MSVTT-QA. † denotes TGIF-QA-R [44] whose multiple choices for repeated action and state transition are more challenging. * denotes the models pretrained with large-scale external data.

For MSRVTT-QA in Table 2, our performance (free-of pretraining) is better than pretraining-free SoTA VGT but is inferior to the large-scale pre-trained methods MERLOT [63] and SiaSRea* [61]. Our method on TGIF-Frame-QA performs close to the pretraining-free sota VGT [54]. This is because pretraining help model the descriptive content, while our work focuses on action temporality.

4.4 True Temporality Metric

ATP [5] evaluated the upper bound performance of a single-frame model on a video dataset and pointed out that even though NextQA dataset focuses on temporality reasoning, the dataset still contains

	Next-QA (%)				TGIF-QA (%)			
	val-Acc	val-δ	test-Acc	test-δ	act-Acc	act-δ	trans-Acc	trans-δ
Ours	**58.27**	**+5.51**	**57.03**	**+5.13**	**96.0**	**+1.2**	97.3	**+1.3**
w/o AcCL	56.87	+2.71	55.02	+2.30	93.5	+0.5	97.1	+0.7
w/o TSC	57.99	+2.98	56.24	+3.25	95.6	+0.8	96.9	+0.2
w/o motion stream	56.57	+3.02	55.78	+2.80	95.2	+0.9	96.7	+0.8
VGT w/o pretrained	55.02	+2.91	53.68	+2.15	95.0	+0.6	**97.6**	+0.3
VGT w/ pretrained	56.89	+1.02	55.70	+0.84	-	-	-	-

Table 3: True temporality evaluation: Study of model components and comparison with SoTA.

static appearance bias. A small portion of questions can be correctly answered exclusively from a single frame without temporal information. To this end, we propose to measure the temporality faithfulness of VideoQA methods, *i.e.* revealing if a VideoQA method learns true temporality to answering questions, instead of learning the spurious correlation between the static appearance and the answer. In specific, the proposed true temporality metric measures the difference of QA accuracy between given the full video and given the middle frame respectively, as δ. The middle-frame setting is that only the middle clip (7th only among the 16 clips) is taken as vision input for QA, so MSA in Eq. 4 is applied on a single token sequence.

Table 3 shows that ATM better learns the true temporality compared to SoTA VGT, w/ w/o pretraining, on both Next-QA and TGif-QA. We observe that the external large-scale data for pretraining VGT guides the model to leverage more static information in temporality reasoning (only +0.84% on Next-QA test) since the pre-training helps more on the descriptive content that is static. Each of our component *i.e.* AcCL, TSC, and appearance-free motion stream, helps to learn the true temporality. TSC mitigates the static bias by preventing answering temporality question if the temporal relations are destroyed. AcCL encourages learning motion representation agonistic to the entity or other appearance information. Appearance-free motion streams extract motion-plentiful representations that are necessary to understand the true temporality.

4.5 Ablation Studies

In addition to the study of each component's individual contribution, we conduct further ablation studies on NextQA [52] dataset.

4.5.1 Impact of Action-centric Contrastive Learning. We test different variants of the text in Action-centric Contrastive Learning (AcCL). Table 4-a summarizes the results of the ablations. AcCL aims at learning action features by aligning the video with the action phrase from the question. The variants replace the action phrase by (1) the correct answer text *w.r.t.* to video-question, denoted as "Answer", (2) the concatenation of the entire question and the correct answer text, denoted as "Question+Answer", (3) the entire question text, denoted as "Question", (4) the verb in question.

Table 4-a shows that our AcCL outperforms all of the other variants. We observe that the "Question" variant performs 0.65% worse than our "action in question" on test split since the full question text contains entity, scene, and other appearance information in addition to the action phrase. Contrasting with full questions will distract the representation from the motion information to the dominant and easily learned appearance features, which is less effective than action-centric version. Using "Answer", "Question+Answer" also performs worse than ours. This demonstrates that the action phrases in questions are the information that the randomly initialized model parameters easily overlook but are important for temporality. Using "verb from question" is also less effective, as the action cannot be described by a single word in many cases, e.g. verb "get" is not informative enough for the action phrase "get up".

Variants	val (%)	test (%)
Action phrase (ours)	**58.27**	**57.03**
Answer	55.92	53.60
Question+Answer	56.83	55.16
Question	57.51	56.38
Verb in Question	57.07	56.57
w/o AcCL	56.87	55.02

(a)

Variants	test (%)	test-δ (%)
TS-aware (Ours)	**57.03**	**+5.13**
TS-unaware	56.89	+3.67
w/o TSC	56.24	+3.25

(b)

Variants	val (%)	test (%)
TSN (ours)	**58.27**	**57.03**
I3D	**57.71**	**56.40**
3D ResNext101	57.01	55.30
SlowFast	56.97	55.83
X3D	56.27	55.78
Timesformer	56.99	56.00
XCLIP	56.08	55.90
I3D-RGB only	57.35	55.63
TSN-RGB only	56.94	55.42
TSN-Flow only	56.89	55.85
I3D-Flow only	56.76	55.73
w/o motion stream	56.57	55.78
$K = 8$	57.63	56.66
$K = 24$	57.82	56.35

(c)

Table 4: Ablation study on different variants of (a) AcCL (b) TSC and (c) motion representations.

4.5.2 Impact of TSC Loss. We compare our Temporal Sensitivity-aware Confusion loss (TSC) in Table 4-b, with variants (1) removing the TSC and only training with cross-entropy, as "w/o TSC". (2) applying the confusion loss to all samples regardless of time-sensitivity, as "TS-unaware". Our method is slightly better than these two variants in VideoQA accuracy and much higher on the proposed true temporality reasoning metric. This validates that alleviating static bias by TSC helps a faithful temporal reasoning model, which in turn improves the event temporality understanding.

4.5.3 Impact of Appearance-free stream. Table 4-c shows the ablations on motion features f_m and analyzes the effectiveness of incorporating an appearance-free stream. In the table, TSN and I3D extract motion features with an appearance-free stream *i.e.* flow maps, while the remaining extract motions only from the appearance-included input *i.e.* RGB. These RGB-only methods SlowFast [16], X3D [15], TimeSformer [4] and XCLIP [42] show superb performance on action recognition, as shown in Fig. 3-b. But they fall behind of the methods with the optical flow on motion representations for VideoQA, though TSN and I3D are relatively early work without fancy network structures. RGB frames may be enough for characterizing limited sets of atomic actions that are dominant for action recognition, but it is less effective in modeling events with long-horizon temporality. 3D ResNext101 [19] has been used for motion feature extraction in existing VideoQA [26, 53], but it is also RGB-only and 1.73% worse than TSN where flow is used.

In addition, Table 4-c also shows that the flow maps are helpful when accompanied by the corresponding RGB frames. Motions in VideoQA cannot be extracted purely from an appearance-free stream, since appearance also provides important cues. The table also shows that with the number of clips as per video $K = 16$, we achieve the best accuracy which is 57.03% on test split. The accuracy slightly drops if we distributed the videos into clips that are more e.g. $K = 24$ or less e.g. $K = 8$. This shows that sampling at

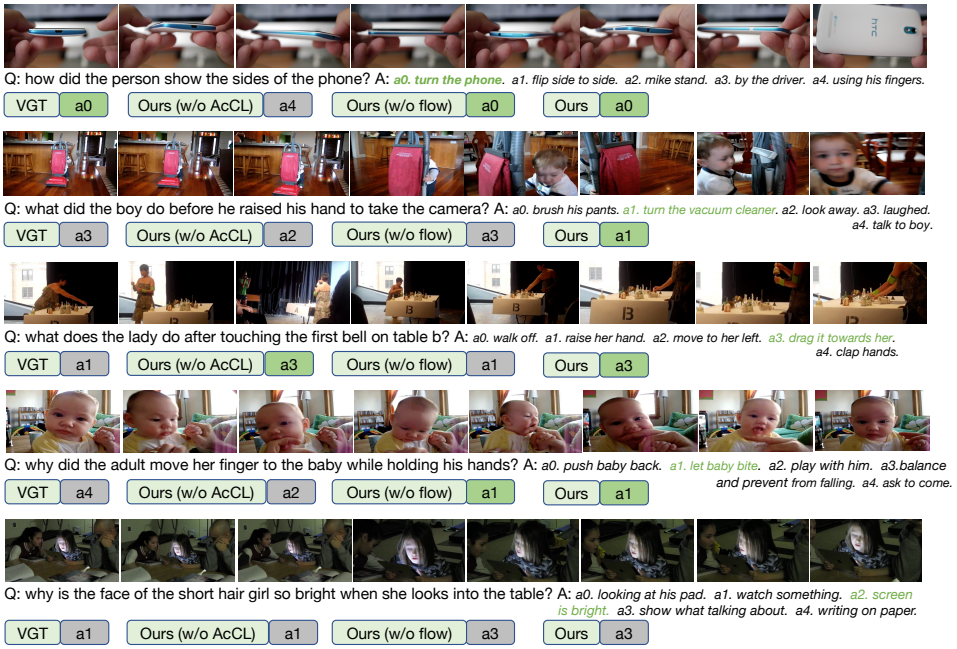

Figure 4: Visualization. The ground-truth are marked in green. We display the results of ATM , ATM w/o AcCL (as "Ours w/o AcCL"), ATM w/o optical flow (as "Ours w/o flow") and the existing SoTA method VGT [54]. The samples span across causality (1,4,5) and temporality (2, 3) reasoning.

a certain rate can encode the informative features across multiple visual modalities. But beyond a certain extent of sampling rate, the model may perform worse due to overfitting.

4.6 Qualitative Analysis

In Fig. 4, we qualitatively evaluate the improvement of the ATM by visualizing the results of representative samples in val. split. We can observe that the AcCL scheme helps to learn the discriminative representations for actions *e.g.* "turn" in (1), while the variant w/o AcCL may learn the superficial correlations between appearance *e.g.* "his fingers" and the answers. Moreover, the appearance-free stream also helps in extracting precise and useful motions. Since the scene and actor do not change in (3), the optical flow stream is informative for recognizing the "drag towards" action. We observe that ATM also avoids over-exploiting language bias, as the proposed AcCL helps ground the action text to the visual evidence e.g. "baby bite" in (4), while others may rely on the question-answer shortcuts between "move finger to baby" and "a2. play with him".

ATM focus on action modeling and it may fall short in reasoning about the object characteristics, *e.g.* the "light screen" causes the girl in a "bright face" in (5). Large-scale vision-language pre-training augmented with knowledge could be helpful. We leave the knowledge-driven action modeling for future work.

5 CONCLUSION

In this paper, we propose a novel framework to solve the VideoQA featuring temporality reasoning. To this end, we realize that it is worth revisiting optical flow, as flow may become less considered in atomic action recognition but is still effective in long-horizon temporality. Then, we propose an action-centric contrastive learning that makes both video and text representations informative for action. Finally, we fine-tune the VideoQA via a novel temporal sensitivity-aware confusion loss to mitigate the potential static bias. Our ATM method is demonstrated to be superior to all existing VideoQA methods on multiple benchmarks and shows a faithful temporality reasoning via a new metric.

Limitations: While ATM outperforms the existing work, there is ample room for further research. Although ATM can deal with arbitrary-length video, it divides the video into a finite number of clips and extracts features per clip. This may not be adequate to capture enough action information when the action occurs in a very short time over a long duration video. Another challenge is time complexity of optical flow computation. It would be worthwhile to study the efficient ways to extract the appearance-free stream.

6 ACKNOWLEDGEMENTS

Junwen Chen and Yu Kong are supported in part by NSF SaTC award 1949694, and the Office of Naval Research under grant number N00014-23-1-2046. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Office of Naval Research or the U.S. Government.

REFERENCES

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

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *CVPR*. 2425–2433.

[3] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. 2021. 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 (ICCV)*. 1728–1738.

[4] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. 2021. Is Space-Time Attention All You Need for Video Understanding?. In *ICML*. PMLR, 813–824.

[5] Shyamal Buch, Cristóbal Eyzaguirre, Adrien Gaidon, Jiajun Wu, Li Fei-Fei, and Juan Carlos Niebles. 2022. Revisiting the" Video" in Video-Language Understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2917–2927.

[6] Joao Carreira and Andrew Zisserman. 2017. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 6299–6308.

[7] Anoop Cherian, Chiori Hori, Tim K Marks, and Jonathan Le Roux. 2022. (2.5+ 1) D Spatio-Temporal Scene Graphs for Video Question Answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 444–453.

[8] Jinwoo Choi, Chen Gao, Joseph CE Messou, and Jia-Bin Huang. 2019. Why can't i dance in the mall? learning to mitigate scene bias in action recognition. *Advances in Neural Information Processing Systems* 32 (2019).

[9] MMAction2 Contributors. 2020. OpenMMLab's Next Generation Video Understanding Toolbox and Benchmark. https://github.com/open-mmlab/mmaction2.

[10] Long Hoang Dang, Thao Minh Le, Vuong Le, and Truyen Tran. 2021. Hierarchical Object-oriented Spatio-Temporal Reasoning for Video Question Answering. In *IJCAI*.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

[12] Guodong Ding and Angela Yao. 2021. Temporal Action Segmentation with High-level Complex Activity Labels. *arXiv preprint arXiv:2108.06706* (2021).

[13] Haodong Duan, Yue Zhao, Kai Chen, Dahua Lin, and Bo Dai. 2022. Revisiting skeleton-based action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2969–2978.

[14] Chenyou Fan, Xiaofan Zhang, Shu Zhang, Wensheng Wang, Chi Zhang, and Heng Huang. 2019. Heterogeneous Memory Enhanced Multimodal Attention Model for Video Question Answering. In *CVPR*. 1999–2007.

[15] Christoph Feichtenhofer. 2020. X3d: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 203–213.

[16] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. 2019. Slowfast networks for video recognition. In *CVPR*. 6202–6211.

[17] Difei Gao, Luowei Zhou, Lei Ji, Linchao Zhu, Yi Yang, and Mike Zheng Shou. 2023. MIST: Multi-modal Iterative Spatial-Temporal Transformer for Long-form Video Question Answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14773–14783.

[18] Jiyang Gao, Runzhou Ge, Kan Chen, and Ram Nevatia. 2018. Motion-appearance co-memory networks for video question answering. In *CVPR*. 6576–6585.

[19] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. 2017. Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet? *arXiv preprint* arXiv:1711.09577 (2017).

[20] Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. (2017). To appear.

[21] Deng Huang, Peihao Chen, Runhao Zeng, Qing Du, Mingkui Tan, and Chuang Gan. 2020. Location-Aware Graph Convolutional Networks for Video Question Answering.. In *AAAI*. 11021–11028.

[22] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. 2017. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2758–2766.

[23] Pin Jiang and Yahong Han. 2020. Reasoning with Heterogeneous Graph Alignment for Video Question Answering. In *AAAI Conference on Artificial Intelligence (AAAI)*.

[24] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. 2017. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950* (2017).

[25] Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR*.

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

[27] Jie Lei, Tamara L Berg, and Mohit Bansal. 2022. Revealing Single Frame Bias for Video-and-Language Learning. *arXiv preprint arXiv:2206.03428* (2022).

[28] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. 2021. 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 (CVPR)*. 7331–7341.

[29] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. 2018. Tvqa: Localized, compositional video question answering. *EMNLP* (2018).

[30] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. 2020. TVQA+: Spatio-Temporal Grounding for Video Question Answering. *ACL* (2020).

[31] Jiangtong Li, Li Niu, and Liqing Zhang. 2022. From Representation to Reasoning: Towards both Evidence and Commonsense Reasoning for Video Question-Answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 21273–21282.

[32] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. 2020. HERO: Hierarchical Encoder for Video+ Language Omni-representation Pre-training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2046–2065.

[33] Yingwei Li, Yi Li, and Nuno Vasconcelos. 2018. Resound: Towards action recognition without representation bias. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 513–528.

[34] Yuncheng Li, Yale Song, Liangliang Cao, Joel Tetreault, Larry Goldberg, Alejandro Jaimes, and Jiebo Luo. 2016. TGIF: A new dataset and benchmark on animated GIF description. In *CVPR*. 4641–4650.

[35] Yicong Li, Xiang Wang, Junbin Xiao, and Tat-Seng Chua. 2022. Equivariant and Invariant Grounding for Video Question Answering. In *Proceedings of the 30th ACM International Conference on Multimedia*. 4714–4722.

[36] Yicong Li, Xiang Wang, Junbin Xiao, Wei Ji, and Tat-Seng Chua. 2022. Invariant grounding for video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2928–2937.

[37] Ji Lin, Chuang Gan, and Song Han. 2019. Tsm: Temporal shift module for efficient video understanding. In *ICCV*. 7083–7093.

[38] Fei Liu, Jing Liu, Weining Wang, and Hanqing Lu. 2021. HAIR: Hierarchical Visual-Semantic Relational Reasoning for Video Question Answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 1698–1707.

[39] Yang Liu, Samuel Albanie, Arsha Nagrani, and Andrew Zisserman. 2019. Use what you have: Video retrieval using representations from collaborative experts. *arXiv preprint arXiv:1907.13487* (2019).

[40] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. 2021. Clip4clip: An empirical study of clip for end to end video clip retrieval. In *arXiv preprint arXiv:2104.08860*.

[41] Liliane Momeni, Mathilde Caron, Arsha Nagrani, Andrew Zisserman, and Cordelia Schmid. 2023. Verbs in Action: Improving verb understanding in video-language models. *arXiv preprint arXiv:2304.06708* (2023).

[42] Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. 2022. Expanding language-image pretrained models for general video recognition. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IV*. Springer, 1–18.

[43] Jungin Park, Jiyoung Lee, and Kwanghoon Sohn. 2021. 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 (CVPR)*. 15526–15535.

[44] Liang Peng, Shuangji Yang, Yi Bin, and Guoqing Wang. 2021. Progressive Graph Attention Network for Video Question Answering. In *ACM MM*. 2871–2879.

[45] Min Peng, Chongyang Wang, Yuan Gao, Yu Shi, and Xiang-Dong Zhou. 2022. Multilevel Hierarchical Network with Multiscale Sampling for Video Question Answering. *IJCAI* (2022).

[46] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*. PMLR, 8748–8763.

[47] Ahjeong Seo, Gi-Cheon Kang, Joonhan Park, and Byoung-Tak Zhang. 2021. Attend What You Need: Motion-Appearance Synergistic Networks for Video Question Answering. In *ACL*. 6167–6177.

[48] Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. 2020. Finegym: A hierarchical video dataset for fine-grained action understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2616–2625.

[49] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. 2012. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402* (2012).

[50] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. 2016. Temporal segment networks: Towards good practices for deep action recognition. In *ECCV*.

[51] Shiguang* Wang, Zhizhong* Li, Yue Zhao, Yuanjun Xiong, Limin Wang, and Dahua Lin. 2020. denseflow. https://github.com/open-mmlab/denseflow.

[52] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. 2021. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of*

the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). 9777–9786.

[53] Junbin Xiao, Angela Yao, Zhiyuan Liu, Yicong Li, Wei Ji, and Tat-Seng Chua. 2022. Video as Conditional Graph Hierarchy for Multi-Granular Question Answering. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*. 2804–2812.

[54] Junbin Xiao, Pan Zhou, Tat-Seng Chua, and Shuicheng Yan. 2022. Video Graph Transformer for Video Question Answering. In *European Conference on Computer Vision*. Springer, 39–58.

[55] Junbin Xiao, Pan Zhou, Angela Yao, Yicong Li, Richang Hong, Shuicheng Yan, and Tat-Seng Chua. 2023. Contrastive Video Question Answering via Video Graph Transformer. *arXiv preprint arXiv:2302.13668* (2023).

[56] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. 2017. Video question answering via gradually refined attention over appearance and motion. In *ACM MM*. 1645–1653.

[57] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*. 5288–5296.

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

[59] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. 2021. Just Ask: Learning To Answer Questions From Millions of Narrated Videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 1686–1697.

[60] Qinghao Ye, Guohai Xu, Ming Yan, Haiyang Xu, Qi Qian, Ji Zhang, and Fei Huang. 2022. Hitea: Hierarchical temporal-aware video-language pre-training. *arXiv preprint arXiv:2212.14546* (2022).

[61] Weijiang Yu, Haoteng Zheng, Mengfei Li, Lei Ji, Lijun Wu, Nong Xiao, and Nan Duan. 2021. Learning from Inside: Self-driven Siamese Sampling and Reasoning for Video Question Answering. *Advances in neural information processing systems (NeurIPS)* 34 (2021).

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

[63] Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. 2021. Merlot: Multimodal neural script knowledge models. In *Advances in neural information processing systems (NeurIPS)*, Vol. 34.

[64] Honglu Zhou, Asim Kadav, Aviv Shamsian, Shijie Geng, Farley Lai, Long Zhao, Ting Liu, Mubbasir Kapadia, and Hans Peter Graf. 2021. COMPOSER: Compositional Reasoning of Group Activity in Videos with Keypoint-Only Modality. *arXiv preprint arXiv:2112.05892* (2021).

[65] Linchao Zhu and Yi Yang. 2020. Actbert: Learning global-local video-text representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 8746–8755.