Learning Fine-Grained Visual Understanding for Video Question Answering via Decoupling Spatial-Temporal Modeling

Hsin-Ying Lee[1]
shinlee@cmlab.csie.ntu.edu.tw
Hung-Ting Su[1]
Bing-Chen Tsai[1]
Tsung-Han Wu[1]
Jia-Fong Yeh[1]
Winston H. Hsu[1,2]

[1] National Taiwan University
[2] Mobile Drive Technology

Abstract

While recent large-scale video-language pre-training made great progress in video question answering, the design of spatial modeling of video-language models is less fine-grained than that of image-language models; existing practices of temporal modeling also suffer from weak and noisy alignment between modalities. To learn fine-grained visual understanding, we decouple spatial-temporal modeling and propose a hybrid pipeline, Decoupled Spatial-Temporal Encoders, integrating an image- and a video-language encoder. The former encodes spatial semantics from larger but sparsely sampled frames independently of time, while the latter models temporal dynamics at lower spatial but higher temporal resolution. To help the video-language model learn temporal relations for video QA, we propose a novel pre-training objective, Temporal Referring Modeling, which requires the model to identify temporal positions of events in video sequences. Extensive experiments demonstrate that our model outperforms previous work pre-trained on orders of magnitude larger datasets.

1 Introduction

Videos are the complex composition of human actions, objects, scenes, and their interactions over time. To examine the capability of machines for video understanding, video question answering (video QA), a task of answering questions about videos, is proposed and requires machines to associate questions in natural languages with visual contents, including scenes [62, 69], dialogues [5, 27], temporal relationships [16, 21, 60, 70], and higher-order cognition [28, 60, 67]. Recent breakthroughs were achieved by pre-training a deep multi-modality encoder, mostly Transformer [56], with large-scale video-language datasets [3, 40, 64]. Models first learned semantic connections between visual and linguistic contents and then were fine-tuned on downstream video-language tasks [33, 49, 64, 71, 76].

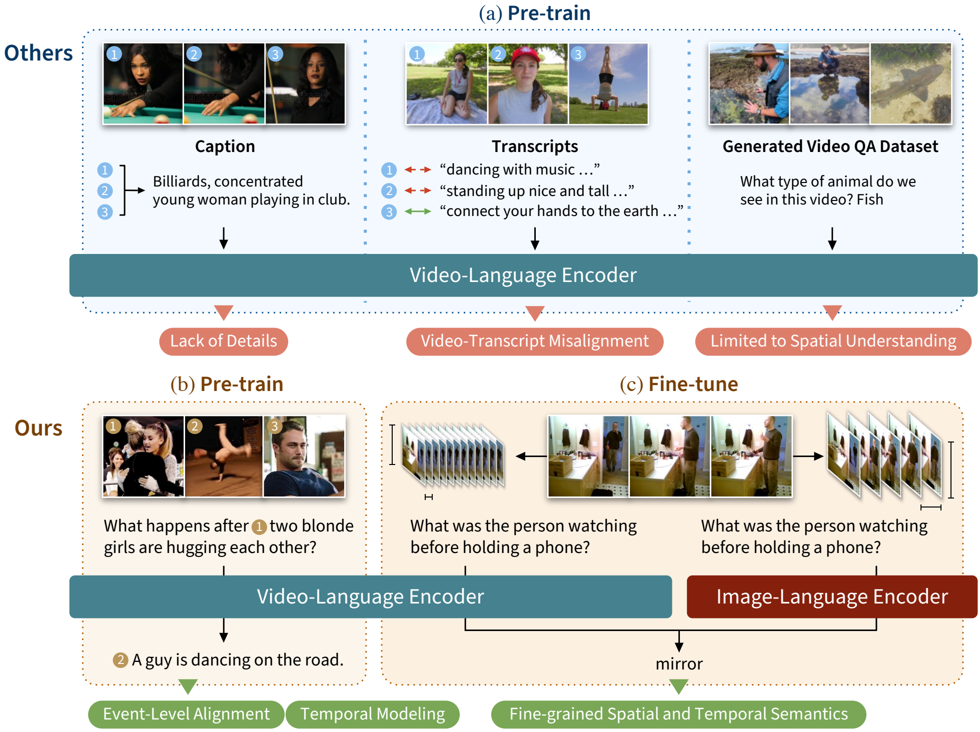

Figure 1: Comparison between (a) previous and (b)(c) our approaches for video QA. (a) Prior work solved video QA by video-language pre-training but might suffer from lack of event details, video-transcript misalignment or limited diversity of pre-training questions. (b) We pre-train a video-language encoder to learn event representations and temporal relations between them by asking the model to identify specific events in synthesized video sequences. (c) We integrate the video-language model with a pre-trained image-language model to encode fine-grained spatial and temporal semantics at different spatial-temporal resolutions.

Despite the advance of this framework in video QA, the spatial semantics encoding of video-language (VL) models is not as fine-grained as the sophisticated design for image-language (IL) models [2, 48, 73]. A preliminary analysis shows that on video QA benchmarks entailing spatial and temporal knowledge, simply averaging frame-by-frame predictions of an IL model can sometimes outperform state-of-the-art VL models. Though the VL models exhibit a slight advantage in questions involving temporal information, the IL model greatly excels in capturing spatial clues (improvement by 7% accuracy; see the full results in Section 4.1.1). The positive performance of IL models could also be attributed to the nature of video QA: the answers to the questions pertaining to only spatial semantics, without specifying time, are usually consistent across all related frames. This property suggests the potential of encoding fine-grained spatial semantics with only IL models.

In addition to spatial modeling, prior work modeled only coarse-grained temporal relations. A question involving temporal relations in video QA often refers to specific events happening in periods of time and inquires about the order of events [16, 21, 60, 70]. It is thus essential to model events in videos and associate the sequence with time conjunctions in questions, such as *before* and *after*. However, as the examples in Figure 1 (a), prior approaches [13, 48, 49, 57, 76] aligning a video with a sentence might lose details of sequential events (what happens after the woman hit the ball), while matching short clips with transcripts [33, 71] may suffer from noise as spoken words often contain something not related to scenes [41]. Others [64, 65] pre-training on generated video QA datasets were mostly limited to spatial understanding. In fact, another examination reveals that the performance with shuffled frame inputs of some of these approaches is similar to that with normal inputs on video QA benchmarks requiring temporal modeling (see more details in Section 4.1.2). The result suggests developing a more effective strategy for modeling temporal relations.

To obtain fine-grained encoding of spatial and temporal semantics for video QA, we propose a novel pipeline, Decoupled Spatial-Temporal Encoders (DeST), decoupling spatial-temporal modeling into IL and VL encoders, illustrated in Figure 1 (c). With IL models well-versed in fine-grained spatial modeling, we incorporate a pre-trained IL model to encode static spatial information independent of time from sparsely sampled frames at high spatial resolution. For questions requiring temporal relations, we train a VL encoder to model temporal dynamics, operating at high temporal but low spatial resolution. These two streams complement each other by paying attention to disparate aspects of videos.

To effectively model temporal relations for video QA, the VL encoder has to recognize events in videos, build their temporal relations, and associate such relations with languages containing temporal information. Thus, we introduce a novel pre-training objective, Temporal Referring Modeling (TRM). Depicted in Figure 1 (b), TRM queries absolute and relative positions of events in videos synthesized by concatenating clips sampled from video captioning datasets [35, 59]. The concatenation simulates transitions of scenes and events in videos. Answering such queries requires a model to aggregate contiguous frames into events and distinguish adjacent events from distant ones. These operations help a model learn both short- and long-term temporal dynamics.

We validate our model on two video QA benchmarks, ActivityNet-QA [70] and AGQA 2.0 [17]. The former contains diverse question types requiring spatial or temporal semantics, and the latter weaves spatial and temporal information together in each question to evaluate compositional reasoning. DeST outperforms the previous state-of-the-art. The ablation studies also demonstrate the efficacy of the proposed pipeline and pre-training objective.

In summary, we make the following key contributions. (*i*) With IL and VL models demonstrating complementary advantages, we decouple spatial and temporal modeling into a hybrid pipeline composed of both models to encode fine-grained visual semantics. (*ii*) We present a novel pre-training objective, Temporal Referring Modeling, to learn temporal relations between events by requiring models to identify specific events in video sequences. (*iii*) We outperform previous VL state-of-the-art methods on two benchmarks with orders of magnitudes less data for pre-training.

2 Related Work

2.1 Video Question Answering

To encode, accumulate and build relationships between visual contents and between modalities for video QA, conventional approaches adopted Recurrent Neural Networks with attention [21, 62, 72, 74, 75], Memory Networks [8, 14, 24, 42, 54], Graph Neural Networks [18, 23, 36, 44, 45, 61], Modular Networks [26], and self-attention [22, 34, 55]. By pre-training large-scale VL datasets, Transformers [56] have further improved the interaction between modalities and made great progress in video QA [13, 33, 48, 57, 64, 65, 71, 76]. Our approach is built on the benefit of modeling relationships with pre-trained Transformers. In contrast to prior work, we carefully examine and take the individual advantage of IL and VL pre-training to encode spatial and temporal semantics.

2.2 Pre-training for Temporal Relation Modeling

VL pre-training learns to model temporal relationships via different approaches.
Learning from Global Alignment. [13, 39, 48, 52, 57, 76] pre-trained models on datasets where a sentence delineates a single event of the entire corresponding video. With features of two modalities being aligned globally, events happening sequentially in a video are compressed, and details of events not mentioned in descriptions are likely lost. Such representations are not fine-grained enough for questions referring to specific moments.
Learning from Local Alignment and Frame Ordering. [33, 71] pre-trained models over datasets with dense annotations such as video transcripts [40]. They matched segmented visual features with utterances and required models to order shuffled or any two frames. With this approach, models learn event-level but weak alignment between videos and languages as spoken words do not always correspond to visual contents [41]. Besides, ordering frames without grounding in languages makes models learn, instead of temporal relations, rational predictions of what is likely to happen before and after an event, which is more related to visual common sense [1, 19, 43].
Learning from Large-Scale Video Question Answering Datasets. [64, 65] pre-trained VL models over large-scale video QA datasets. The diversity of pre-training questions thus determines the effectiveness and capacity of transferred knowledge, but generated questions in [64] and [65] mainly pertain to scene and dialogue understanding, leaving temporal relationship modeling unsolved.

2.3 Encoding Motion and Appearance

Prior arts have explored two-stream networks to encode motion and appearance for action recognition [6, 9, 10, 11, 50, 58]. [7, 12] combined different spatial and temporal resolution to separately encode slow- and fast-changing scenes, and [46, 47] searched for multi-stream connectivity. Analogously, our two streams complement each other by focusing on disparate aspects of videos, but while their two streams both encode short-term actions, our IL stream aggregates scene information independent of time, and the VL stream encodes entire videos and constructs the temporal relationships between all actions and events.

Some recent work revealed that understanding temporality is not always necessary to solve VL tasks. [29, 30] taking sparsely sampled frames outperformed previous methods. [4] provided stronger baselines with single frame inputs. However, with new tasks requiring temporal modeling proposed, such conclusions are likely to be circumscribed. We thus take a further step by proposing an effective strategy to encode fine-grained temporal semantics.

3 Method

We introduce our video QA pipeline, Decoupled Spatial-Temporal Encoders (Section 3.1), and the pre-training objective, Temporal Referring Modeling (Section 3.2). Implementation details are described in the supplement (Section A).

3.1 Decoupled Spatial-Temporal Encoders

The coarse-grained spatial modeling of prior approaches motivates us to develop more effective architectures, and IL models have shown great potential. While most VL models take scene or multi-frame features pre-extracted by image or action recognition models [33, 39, 64, 71], region features [38, 51, 53, 73] and features processed by attention [2, 63] have been proved powerful for IL models. These features provide detailed information about visual elements along with their spatial relations. Since static scene information, if asked by questions without specifying time, are usually consistent across related frames, IL models should also be competent to encode fine-grained spatial relations for video QA.

Hence, we propose Decoupled Spatial-Temporal Encoders (DeST), a video QA pipeline decoupling spatial and temporal modeling into an IL and a VL encoder. The IL encoder takes unordered and sparsely sampled frames at high spatial resolution as input. Fine-grained spatial information of static scenes is obtained by building a consensus among these frames. The VL encoder with input action features at high temporal resolution recognizes and models the transitions of actions and events. These two streams of information are fused at the final stage to jointly form the prediction. We leave other ways of fusion for future exploration.

As illustrated in Figure 2, DeST consists of an image encoder, a video encoder, and a question encoder to process inputs, as well as an IL encoder and a VL encoder, both with cross-attention [20, 30, 31, 32], to perform multi-modality interaction. Another answer encoder encodes answer candidates, similar to [64]. To answer a question about a video, the question, video, and frames that are sparsely sampled from the video are encoded by their respective encoders. The question features then perform cross-attention to both frame and video features. The sum of two multi-modality representations is finally compared with encoded answer candidates to obtain the prediction. Formally, $\mathcal{Q}$ denotes the input question. $\{\mathcal{I}^1, ..., \mathcal{I}^T\}$ are T frames sampled from the input video $\mathcal{V}$, where $T \ll$ the length of $\mathcal{V}$. The question $\mathcal{Q}$ is first encoded into a sequence of embeddings $\mathbf{w} = \{w_{\texttt{cls}}, w_1, ..., w_L\}$, $w \in \mathbb{R}^D$, where $w_{\texttt{cls}}$ is the embedding of the `[CLS]` token, and L is the number of word tokens. Then $\mathbf{w}$ is fused with the frames and video as described below.

Image-Language Encoding. For each t from 1 to T, the image encoder transforms frame $\mathcal{I}^t$ into a sequence of patch embeddings $\mathbf{u} = \{u^t_{\texttt{cls}}, u^t_1, ..., u^t_N\}$, $u \in \mathbb{R}^D$, where N is the number of patches. Then the question feature $\mathbf{w}$ and frame feature $\mathbf{u}$ are fused by the IL encoder with cross-attention and transform into $\{x^t_{\texttt{cls}}, x^t_1, ..., x^t_L\}$, $x \in \mathbb{R}^D$. The multi-modality representation of the IL stream r is the average of `[CLS]` token embeddings $x^t_{\texttt{cls}}$ of all frames encoded by a final multi-layer perceptron (MLP):

$$r = \frac{1}{T}\sum_{t=1}^{T} \mathrm{MLP}(x^t_{\texttt{cls}}),\ r \in \mathbb{R}^D. \tag{1}$$

Video-Language Encoding. The video feature extractor first encodes the input video $\mathcal{V}$ into a sequence of features $\mathbf{e} = \{e_1, ..., e_M\}$, $e \in \mathbb{R}^H$, where M is the length of the feature sequence. To indicate the beginning and the end of the video, we add two learnable tokens before and after the feature sequence. Temporal position encoding is also added to each feature to indicate the temporal order. Next, the feature sequence $\mathbf{e}$ are contextualized and transformed into $\mathbf{v} = \{v_{\texttt{bos}}, v_1, ..., v_M, v_{\texttt{eos}}\}$, $v \in \mathbb{R}^D$, where $v_{\texttt{bos}}$ and $v_{\texttt{eos}}$ are the beginning and the end token after contextualization. The question feature $\mathbf{w}$ then performs cross attention to the video feature $\mathbf{v}$ through the VL encoder and transforms into $\{y_{\texttt{cls}}, y_1, ..., y_L\}$, $y \in \mathbb{R}^D$. The multi-modality representation of the VL stream $s \in \mathbb{R}^D$ is the output of the first token $y_{\texttt{cls}}$ transformed by a final MLP.

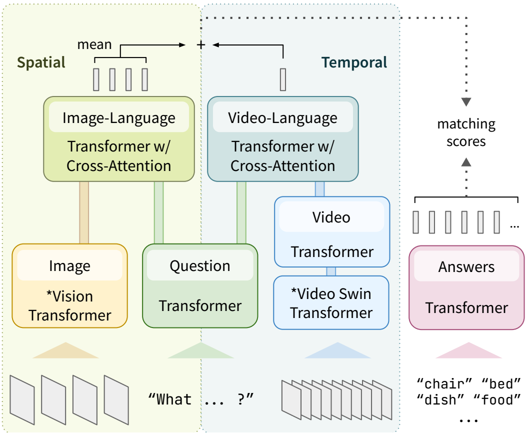

Figure 2: Decoupled Spatial-Temporal Encoders. Encoded questions are fused with frames and videos to gather spatial and temporal information. Their representations are then compared with answer candidates to obtain the final predictions. (* marks the frozen modules.)

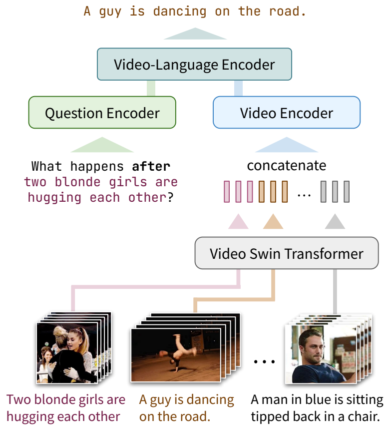

Figure 3: Temporal Referring Modeling, which associates visual events and their temporal relationships with languages by asking absolute and relative positions of events in concatenated video features sampled from video captioning data.

Answer Selection. Following [64], another text encoder encodes the answer candidates (collected from all answers in training data with frequency > 1 for open-ended QA). The prediction of each candidate is the dot product between each encoded candidate and the sum of two multi-modality representations. Formally, $\mathcal{A}$ denotes the answer set. For all $a \in \mathcal{A}$, we take the `[CLS]` token $z^a_{\texttt{cls}} \in \mathbb{R}^D$ of a's feature. Then the logit of a is obtained via:

$$p_a = (r+s)^\mathsf{T} z^a_{\texttt{cls}}, \; p \in \mathbb{R}. \tag{2}$$

3.2 Temporal Referring Modeling

To pre-train the multi-modality encoders with affordable computation resources, we adopt an IL encoder pre-trained with image question answering (image QA), specifically VQA [15], and train the VL encoder for fine-grained temporal modeling with a novel objective.

Modeling fine-grained temporal relations for video QA requires the encoder to understand videos as event sequences and to associate the temporal relations of events with descriptions containing time conjunctions. Therefore, we develop Temporal Referring Modeling (TRM), which, in the form of video QA, inquires about absolute and relative temporal positions of events in videos. As depicted in Figure 3, given a video composed of multiple events, TRM asks the model four questions: what happens at the beginning, at the end, before an event, or after an event? The model then selects an event description as the answer. To accomplish this task requires the model to identify events and manage the order.

TRM needs VL data that offers (1) event-level annotations that delineate scenes and events for segments of videos and (2) descriptions that explain the temporal dynamics of these segments. Dense video captioning [25] should be ideally suited for our needs, but unfortunately, many of its time segments overlap, making the temporal relations ambiguous, and labeling cost also hinders scalability. To satisfy the two conditions, we thus develop a

Type	Just-Ask	VIOLET	ALBEF	UB
Motion	28.00	18.25	32.50	70.63
Spatial Rel.	17.50	15.00	24.38	75.63
Temporal Rel.	4.88	2.12	3.75	32.88
Yes / No	66.28	71.87	79.75	100.00
Color	34.29	31.28	57.39	98.99
Object	26.73	22.33	31.45	70.13
Location	35.75	30.57	36.01	86.79
Number	50.17	50.33	55.61	99.83
Other	36.82	33.02	40.16	71.98
Overall	38.86	37.44	46.66	80.74

Table 1: Comparison between prior methods and our upper bound of ActivityNet-QA by question type. ALBEF exhibits advantages on the questions involving spatial reasoning. (Rel. is short for Relationships, and UB is the abbreviation for upper bound.)

Method	Benchmark	Accuracy
VIOLET	AGQA	49.15
	AGQA*	49.22±.02
Just-Ask	AGQA	51.27
	AGQA*	47.73±.06
HERO	VIOLIN	69.01
	VIOLIN*	68.71±.08

Table 2: Results of prior work taking shuffled frames as input. The little performance drop indicates that some methods are not sensitive to the order of frames. (* signifies that input frames are shuffled. We report the average of three results for the shuffle experiment.)

simple yet effective way to generate data. As the example in Figure 3, we concatenate videos sampled from video captioning datasets to create videos with scene and event transitions. Then we generate questions by completing templates with captions of these videos. Incorrect answers are the other captions in the same video sequences, making the task more difficult.

Take, as an example, generating a video and a question that asks which event happens after an event. We first sample K pairs from a video captioning dataset, with each pair k composed of a video $\mathcal{V}_k$ and a caption $\mathcal{C}_k$. The videos are encoded by the feature extractor into feature sequences $\{e_1^k,\ldots,e_{M_k}^k\}$ for all k from 1 to K, where M_k is the length of features of $\mathcal{V}_k$. These sequences are then concatenated and form $\mathbf{e} = \{e_1^1,\ldots,e_{M_1}^1,e_1^2,\ldots,e_{M_K}^K\}$. To generate the question, we first sample a captions $\mathcal{C}_i$ where $1 \leq i < K,\ i \in \mathbb{N}$. Then the question $\mathcal{Q}$ is "What happens after $\mathcal{C}_i$?" with the choices $\mathcal{A} = \{\mathcal{C}_k \mid 1 \leq k \leq K,\ k \neq i,\ k \in \mathbb{N}\}$ and the correct answer $\mathcal{C}_{i+1}$. Other questions are constructed similarly, where the answers to the questions about the beginning and the end are $\mathcal{C}_1$ and $\mathcal{C}_K$ respectively. With all input the same as general video QA, the encoded feature $\mathbf{w}$ of question $\mathcal{Q}$ and the video feature $\mathbf{e}$ are input to the VL encoder, going through the encoding and contextualizing process described in Section 3.1. The final objective is to minimize a standard cross-entropy loss.

4 Experiments

We elaborate on the preliminary analysis of spatial and temporal reasoning capability of prior work (Section 4.1). Then we demonstrate the improvement in two video QA benchmarks with DeST and TRM (Section 4.2). The ablation studies are lastly presented evaluating the efficacy of each component. (Section 4.3).

4.1 Preliminary Analysis

Baselines. We take ALBEF [31] as an example of IL models. For VL models, we study VIOLET [13], HERO [33], and Just-Ask [64], which respectively instantiate three approaches discussed in Section 2.2. These are state-of-the-art of each approach with public code bases.

Method	Pre-training Data	Acc
CoMVT [48]	100M vid	38.8
Just-Ask [64]	69M vid	38.9
MV-GPT [49]	100M vid	39.1
SiaSamRea [68]	5.6M img	39.8
MERLOT [71]	180M vid	41.4
VIOLET [13]	180M vid + 2.5M vid + 3M img	37.5
FrozenBiLM [66]	10M vid	43.2
Singularity [30]	14M img + 2.5M vid	44.1
DeST (ours)	14M img + 120K VQA + 14K vid	**46.8**

Table 3: Comparison with previous methods on ActivityNet-QA. We outperform all methods with significantly less pre-training data. The dataset names are provided in the supplement Section B.2. (img: images. vid: videos.)

Type	Best	DeST	Diff (%)
Motion	32.50	35.75	**10.00**
Spatial Rel.	24.38	23.88	-2.05
Temporal Rel.	4.88	5.25	**7.58**
Yes / No	79.75	78.61	-1.43
Color	57.39	59.11	3.00
Object	31.45	30.50	-3.02
Location	36.01	36.27	0.72
Number	55.61	55.28	-0.59
Other	40.16	39.63	-1.32
Overall	46.66	**46.79**	0.28

Table 4: Comparison with prior methods on AcivityNet-QA by question type. We perform comparably in question types of spatial information and improve temporal modeling.

4.1.1 Encoding Spatial Semantics

We first assess the ability of encoding spatial semantics of IL models and VL models[1]. ALBEF is run as image QA by sampling frames from a video and averaging frame predictions.
Benchmark. We conduct the analysis on ActivietNet-QA [70], which contains 5.8K videos of human activities in daily life and 58K question-answer pairs spanning diverse categories across spatial and temporal semantics offering comprehensive evaluations.
Results. Table 1 contrasts the accuracy (acc) by question type of the IL model with other VL models. ALBEF, though without temporal modeling, is adept at spatial reasoning, such as *Spatial Relationships* and *Color*, while Just-Ask demonstrates a slight advantage in *Temporal Relationships*. Due to the removal of rare answers following [64], we report our performance upper bound of each type, which is the proportion of questions in the test set whose answers appeared in the training set. The tiny number of *Temporal Relationships* reveals the long-tailed distribution of its answers, which partially explains the poor performance.

4.1.2 Modeling Temporal Relationships

We evaluate the capability of modeling temporal relationships by shuffling input frames and measuring the performance drop. Models are first trained with normal input and tested their performance with shuffled input. Intuitively, taking shuffled frames as input should be detrimental to the performance of the questions requiring temporal modeling, such as those inquiring about the order of actions or events in videos.
Benchmarks. For VIOLET and Just-Ask, we conduct the study on AGQA 2.0 [17], a large-scale open-ended video QA benchmark where spatial and temporal information is required in each question for evaluating compositional reasoning. It contains 2.27M question-answer pairs and 9.6K videos. For HERO, we consider VIOLIN [37], a task of judging hypotheses from visual premises, which has been officially tested in their experiments.
Result. In Table 2, Just-Ask demonstrates the slight capability of temporal modeling, while VIOLET and HERO are not sensitive to the order of input frames, and their performances of

[1] Just-Ask and VIOLET as HERO does not support open-ended QA

	Type	Best w/o PT	Best w/ PT	DeST
Reasoning	Object-Relationship	40.33	48.91	**59.66**
	Relationship-Action	49.95	66.55	**72.98**
	Object-Action	50.00	68.78	**75.20**
	Superlative	33.55	39.83	**48.94**
	Sequencing	49.78	67.01	**73.53**
	Exists	50.01	59.35	**63.21**
	Duration Comparison	47.03	50.49	**60.39**
	Activity Recognition	5.52	21.53	**27.78**
Semantic	Object	40.40	49.31	**61.27**
	Relationship	49.99	59.60	**63.93**
	Action	47.58	58.03	**65.96**
Structure	Query	36.34	47.98	**61.22**
	Compare	49.71	65.11	**72.04**
	Choose	46.56	46.90	**53.01**
	Logic	50.02	56.20	**59.18**
	Verify	50.01	58.13	**63.02**
Overall	Binary	48.91	55.35	**62.61**
	Open	36.34	47.98	**61.22**
	All	42.11	51.27	**61.91**

Table 5: Comparison with prior work on AGQA 2.0. We list the best performance among methods without (Best w/o PT) and with pre-training (Best w/ PT) for each question type. DeST exceeds all methods in all question types.

Question	Frames	Video	Acc
✓			41.32
✓	✓		50.07
✓	VQA		51.00
✓		✓	51.08
✓		TRM	55.62
✓	VQA	✓	56.61
✓	VQA	TRM*	56.97
✓	VQA	TRM	61.91

Table 6: Ablation study of input modalities and pre-training strategies on AGQA 2.0. The results favor our hybrid pipeline and TRM. (✓ means the modality is presented. VQA: pretrained on VQA. TRM: pre-trained with TRM. *: shuffled input.)

Training Stream	Acc
Image-Language	49.91
Video-Language	16.56
Both	61.91

Table 7: Ablation study of two encoding streams on AGQA 2.0.

taking normal and shuffled input frames are similar. The result suggests clear insufficiency for temporal relationship modeling.

4.2 Video Question Answering

DeST takes frames and videos as input. Frames are extracted at 3 FPS, following [27]; then we sample T frames randomly during training and uniformly during inference, similar to the strategy for action recognition. Video features are also pre-extracted by the video encoder and excluded from the optimization. More details are left in the supplement (Section A.1).

Table 3 compares DeST with prior work on ActivityNet-QA. We outperform all previous methods with orders of magnitudes less pre-training data. The performance of each question type is listed in Table 4, where *Best* shows the highest scores among the three methods in Table 1. This rigorous comparison leads to a more comprehensive analysis in terms of both spatial and temporal modeling. *Diff* lists the difference between *Best* and our performance in proportion to *Best*. Our hybrid model performs, as expected, comparably with the IL model in spatial modeling since we are not improving IL processing. On the other hand, the performance of categories such as *Motion* and *Temporal Relationships* are boosted, verifying the efficacy of TRM.

Table 5 presents the performance on AGQA 2.0, which offers extensive annotation of multiple abilities necessary to answer each question. We list the highest accuracy among the methods without pre-training reported by [17] (Best w/o PT) and the higher scores between our implementation of Just-Ask and VIOLET (Best w/ PT). DeST surpasses all prior work in all question types. Besides, while TRM is similar to only the questions of *Sequencing*, which accounts for about 7% of the dataset, TRM can serve as an abstraction of temporal modeling and generalize to other question types, such as *Relationship-Action* and *Object-Action*, which inquire about the temporal relationship between human actions and their interactions with objects. The full table and detailed analysis are provided in the supplement (Section B.3).

4.3 Ablation Studies

We present the influence of input modalities and pre-training over AGQA 2.0 to study the effect of modeling decisions. As presented in Table 6, question-only input reveals the language bias, which serves as a baseline. The boost in performance with frames and videos suggests successful encoding. Pretraining the IL encoder with VQA and the VL encoder with TRM both enhance the modeling capacity further. The performance drop due to shuffling videos verifies the efficacy of TRM. The full results are included in the supplement (Section B.3).

In Table 7, we ablate the IL or VL stream. A model is trained with both streams and tested on AGQA 2.0 with a single stream. The performance drastically drops in both settings, proving that our hybrid model is not a trivial ensemble. It might also be noted that the overwhelming advantage of the IL stream over its VL counterpart cannot conclude the utility of any stream, for each stream can be trained to perform better than the question-only baseline. We hypothesize that temporal information can be seen as the complex evolution of spatial information, and thus when both streams collaborate in spatial-temporal modeling, the IL stream offers an overall understanding of visual elements and scenes, while the VL stream assists it and models the detailed changes.

5 Conclusion

In this work, considering the complementary advantage of image- and video-language models, we decouple spatial-temporal modeling and propose a hybrid pipeline for video QA, where an image-language encoder encodes spatial information and a video-language encoder models temporal dynamics. To capture event-level temporal relations, the video-language encoder is pre-trained with an objective to identify events in videos by their temporal positions. With the collaboration between image- and video-language models as well as fine-grained temporal modeling, we advance the visual understanding for video QA.

Acknowledgement

This work was supported in part by the National Science and Technology Council under Grant MOST 110-2634-F-002-051, Mobile Drive Technology Co., Ltd (MobileDrive), and NOVATEK fellowship. We are also grateful to the National Center for High-performance Computing.

References

[1] Harsh Agrawal, Arjun Chandrasekaran, Dhruv Batra, Devi Parikh, and Mohit Bansal. Sort story: Sorting jumbled images and captions into stories. In *Proceedings of the*

2016 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2016.

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

[3] 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 (ICCV)*, 2021.

[4] 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 (CVPR)*, 2022.

[5] Seongho Choi, Kyoung-Woon On, Yu-Jung Heo, Ahjeong Seo, Youwon Jang, Minsu Lee, and Byoung-Tak Zhang. DramaQA: Character-centered video story understanding with hierarchical QA. *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2021.

[6] Ali Diba, Vivek Sharma, and Luc Van Gool. Deep temporal linear encoding networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[7] Ali Diba, Mohsen Fayyaz, Vivek Sharma, Manohar Paluri, Jürgen Gall, Rainer Stiefelhagen, and Luc Van Gool. Large scale holistic video understanding. In *European Conference on Computer Vision (ECCV)*, 2020.

[8] Chenyou Fan, Xiaofan Zhang, Shu Zhang, Wensheng Wang, Chi Zhang, and Heng Huang. Heterogeneous memory enhanced multimodal attention model for video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[9] Christoph Feichtenhofer, Axel Pinz, and Richard Wildes. Spatiotemporal residual networks for video action recognition. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

[10] Christoph Feichtenhofer, Axel Pinz, and Andrew Zisserman. Convolutional two-stream network fusion for video action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[11] Christoph Feichtenhofer, Axel Pinz, and Richard P. Wildes. Spatiotemporal multiplier networks for video action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[12] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[13] 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*, 2022.

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

[15] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[16] 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 (CVPR)*, 2021.

[17] Madeleine Grunde-McLaughlin, Ranjay Krishna, and Maneesh Agrawala. AGQA 2.0: An updated benchmark for compositional spatio-temporal reasoning. *arXiv preprint arXiv:2204.06105*, 2022.

[18] Deng Huang, Peihao Chen, Runhao Zeng, Qing Du, Mingkui Tan, and Chuang Gan. Location-aware graph convolutional networks for video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2020.

[19] Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. COMET-ATOMIC 2020: On symbolic and neural commonsense knowledge graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2021.

[20] 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 (ICML)*, 2021.

[21] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. TGIF-QA: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[22] Jianwen Jiang, Ziqiang Chen, Haojie Lin, Xibin Zhao, and Yue Gao. Divide and conquer: Question-guided spatio-temporal contextual attention for video question answering. *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2020.

[23] Pin Jiang and Yahong Han. Reasoning with heterogeneous graph alignment for video question answering. *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2020.

[24] Junyeong Kim, Minuk Ma, Kyungsu Kim, Sungjin Kim, and Chang D. Yoo. Progressive attention memory network for movie story question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

[25] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2017.

[26] Thao Minh Le, Vuong Le, Svetha Venkatesh, and Truyen Tran. Hierarchical conditional relation networks for multimodal video question answering. *International Journal of Computer Vision (IJCV)*, 2021.

[27] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. TVQA: Localized, compositional video question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018.

[28] Jie Lei, Licheng Yu, Tamara Berg, and Mohit Bansal. What is more likely to happen next? video-and-language future event prediction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

[29] 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 (CVPR)*, 2021.

[30] Jie Lei, Tamara L. Berg, and Mohit Bansal. Revealing single frame bias for video-and-language learning. *arXiv preprint arXiv:2206.03428*, 2022.

[31] Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[32] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the 39th International Conference on Machine Learning (ICML)*, 2022.

[33] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. HERO: Hierarchical encoder for Video+Language omni-representation pre-training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2020.

[34] 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. *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2019.

[35] Yuncheng Li, Yale Song, Liangliang Cao, Joel Tetreault, Larry Goldberg, Alejandro Jaimes, and Jiebo Luo. TGIF: A New Dataset and Benchmark on Animated GIF Description. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[36] Fei Liu, Jing Liu, Weining Wang, and Hanqing Lu. HAIR: Hierarchical visual-semantic relational reasoning for video question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021.

[37] Jingzhou Liu, Wenhu Chen, Yu Cheng, Zhe Gan, Licheng Yu, Yiming Yang, and Jingjing Liu. Violin: A large-scale dataset for video-and-language inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

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

[39] Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou. UniVL: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353*, 2020.

[40] 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 (ICCV)*, 2019.

[41] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[42] Seil Na, Sangho Lee, Jisung Kim, and Gunhee Kim. A read-write memory network for movie story understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2017.

[43] Jae Sung Park, Chandra Bhagavatula, Roozbeh Mottaghi, Ali Farhadi, and Yejin Choi. VisualCOMET: Reasoning about the dynamic context of a still image. In *European Conference on Computer Vision (ECCV)*, 2020.

[44] 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 (CVPR)*, 2021.

[45] Liang Peng, Shuangji Yang, Yi Bin, and Guoqing Wang. Progressive graph attention network for video question answering. In *Proceedings of the 29th ACM International Conference on Multimedia (MM)*, 2021.

[46] Michael S. Ryoo, A. J. Piergiovanni, Juhana Kangaspunta, and Anelia Angelova. Assemblenet++: Assembling modality representations via attention connections. In *European Conference on Computer Vision (ECCV)*, 2020.

[47] Michael S. Ryoo, A. J. Piergiovanni, Mingxing Tan, and Anelia Angelova. Assemblenet: Searching for multi-stream neural connectivity in video architectures. In *International Conference on Learning Representations (ICLR)*, 2020.

[48] Paul Hongsuck Seo, Arsha Nagrani, and Cordelia Schmid. Look before you speak: Visually contextualized utterances. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[49] Paul Hongsuck Seo, Arsha Nagrani, Anurag Arnab, and Cordelia Schmid. End-to-end generative pretraining for multimodal video captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[50] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2014.

[51] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. VL-BERT: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations (ICLR)*, 2020.

[52] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. VideoBERT: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[53] Hao Tan and Mohit Bansal. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019.

[54] 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/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[55] Aisha Urooj, Amir Mazaheri, Niels Da vitoria lobo, and Mubarak Shah. MMFT-BERT: Multimodal Fusion Transformer with BERT Encodings for Visual Question Answering. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020.

[56] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[57] 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.

[58] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European Conference on Computer Vision (ECCV)*, 2016.

[59] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. VaTeX: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR)*, 2019.

[60] 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 (CVPR)*, 2021.

[61] Junbin Xiao, Angela Yao, Zhiyuan Liu, Yicong Li, Wei Ji, and Tat-Seng Chua. Video as conditional graph hierarchy for multi-granular question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2022.

[62] 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 (MM)*, 2017.

[63] Kelvin Xu, Jimmy Lei Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, 2015.

[64] 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 (ICCV)*, 2021.

[65] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Learning to answer visual questions from web videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 2022.

[66] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Zero-shot video question answering via frozen bidirectional language models. *arXiv preprint arXiv:2206.08155*, 2022.

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

[68] Weijiang Yu, Haoteng Zheng, Mengfei Li, Lei Ji, Lijun Wu, Nong Xiao, and Nan Duan. Learning from inside: Self-driven siamese sampling and reasoning for video question answering. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[69] Youngjae Yu, Jongseok Kim, and Gunhee Kim. A joint sequence fusion model for video question answering and retrieval. In *European Conference on Computer Vision (ECCV)*, 2018.

[70] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. ActivityNet-QA: A dataset for understanding complex web videos via question answering. *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2019.

[71] Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. MERLOT: Multimodal neural script knowledge models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[72] Kuo-Hao Zeng, Tseng-Hung Chen, Ching-Yao Chuang, Yuan-Hong Liao, Juan Carlos Niebles, and Min Sun. Leveraging video descriptions to learn video question answering. *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2017.

[73] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. VinVL: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

[74] Zhou Zhao, Jinghao Lin, Xinghua Jiang, Deng Cai, Xiaofei He, and Yueting Zhuang. Video question answering via hierarchical dual-level attention network learning. In *Proceedings of the 25th ACM international conference on Multimedia (MM)*, 2017.

[75] Zhou Zhao, Qifan Yang, Deng Cai, Xiaofei He, Yueting Zhuang, Zhou Zhao, Qifan Yang, Deng Cai, Xiaofei He, and Yueting Zhuang. Video question answering via hierarchical spatio-temporal attention networks. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI)*, 2017.

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