MMTF: Multi-Modal Temporal Fusion for Commonsense Video Question Answering

Mobeen Ahmad
mobeen@pyler.tech

Geonwoo Park
rjsdn1120@g.skku.edu

Dongchan Park
cto@pyler.tech

Sanguk Park*
parksang1993@gmail.com

PYLER CO., LTD.
Seoul, South Korea

Abstract

Video question answering is a challenging task that requires understanding the video and question in the same context. This becomes even harder when the questions involve reasoning, such as predicting future events or explaining counterfactual events, because they need knowledge not explicitly shown. Existing methods use coarse-grained fusion of video and language features, ignoring temporal information. To address this, we propose a novel vision-text fusion module that learns the temporal context of the video and question. Our module expands question tokens along the video's temporal axis and fuses them with video features to generate new representations with local and global context. We evaluated our method on four VideoQA datasets, including MSVD-QA, NExT-QA, Causal-VidQA, and AGQA-2.0.

1. Introduction

Dominant multi-modal methods for understanding text and video are Multi-Head Attention (MHA) and Cross-Modal Attention (CMA) [19, 1, 28, 26, 16, 37, 30]. These techniques capture overall context but can overlook fine-grained information. For example, answering a question requiring strong temporal understanding can be challenging. Surprisingly, there has been little study of multi-modality at diverse granularity levels. Previous methods used mean-pooling or 3D-CNNs[29, 32, 13, 11] for temporal aggregations, but global representations do not provide fine-scale granularity needed for temporal or counterfactual questions.

In this work, we emphasize on two points.

1) Capturing fine-grained temporal context within two different modalities can enhance a model's ability to understand temporally specific information. Previous studies [34, 33, 13] have aggregated pre-extracted features to learn global representations of videos and fuse them with question features. However, these pre-extracted features are already highly encoded by strong backbone models. As a result, the model cannot learn fine-grained context using these implied features. We explicitly compared local-global context learning with only global-context learning methods. (see 4.2.1).

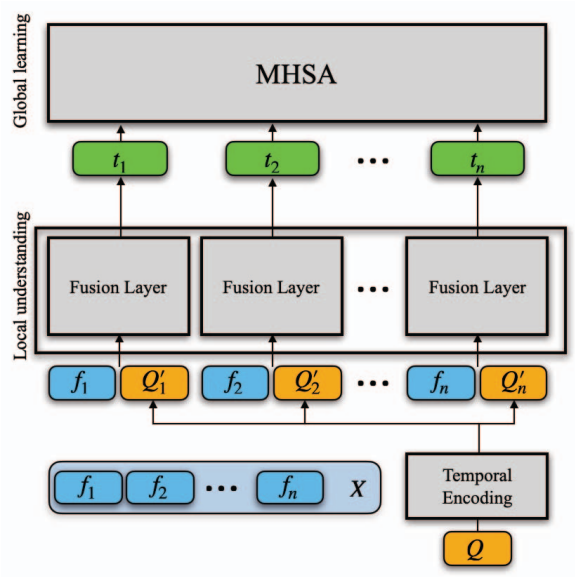

Figure 1. Illustration of overall architecture. Local contexts(fine-grained temporal contexts) are calculated by fusing scene features(frame or object-graph, or both) and temporally encoded question tokens.

2) Using a temporal fusion module at a fine-grained level can enhance a model's ability to capture local context. Previous studies have proposed replicating language features and concatenating them with video's temporal features to achieve performance gains through local interaction between video and text [36]. However, this method ignores the fact that videos have an explicit temporal dimension while text does not. To address this issue, we encode text representation along the temporal axis to facilitate capturing fine-grained temporal context. As revealed in section 4.2.2, temporally encoded texts can learn better-fused representations. To encode text in a temporal manner, we

* Corresponding author.

utilize simple linear layers. The Temporal Fusion Module is responsible for fusing expanded text and local video features (as described in section 3.2.2) to produce fine-grained temporal contexts. In summary, our contributions are as follows:

- We inject temporal information into the language to produce fine-grained temporal contexts.
- We propose multi-modal temporal fusion module that captures both local and global context to understand muti-level temporal granularity.
- Our method improves upon prior art and achieves state-of-the-art results on several benchmark question-answering datasets, including the newly introduced commonsense reasoning dataset Causal-VidQA and AGQA-2.0.

2. Related Work

Current methods can be categorized into 1) Canonical [10, 38, 11, 15, 38], and 2) Graph-based methods [34, 33, 9]. Canonical methods leverage the sequential nature of video frames and question words but fail to capture fine-grained relations among objects. Object-oriented graph-based approaches use extracted object regions as nodes and learn several atomic representations and later aggregate them with text features.

Graph Representation for VideoQA. There has been growing interest in graph representation of visual and text data for VideoQA [9, 11, 22, 18, 33]. These graph methods are either static or fail to disambiguate temporal and spatial context. VGT [34] proposed dynamic graphs to capture temporal and spatial dynamics based on object regions.

Cross-Modal Representation Learning. After the success of transformer architecture [30], it has been widely adapted to map individual modalities into a common latent space [20, 27, 21]. Further transformer-based methods aim to learn joint image-text or video-text representations by utilizing cross-modal attention or co-attention between input modalities [19, 1, 28, 26, 16, 37]. Attempts have been made to learn video-text representation for VideoQA based on pre-computed object features [39] or UNITER based cross-modal fusion layer [3, 14]. We propose to fuse video and text modalities by encoding question tokens over the temporal axis of the video to obtain representative features for each frame.

Commonsense VideoQA. To develop commonsense reasoning ability, it is necessary to understand temporal information. There have been efforts to separate related scenes from non-related scenes using invariant learning-based grounding approaches [31, 12, 2, 17]. However, they neglect that multi-level temporal granularity resides inside causal scenes. A more generic way to comprehend information within both global and local contexts is to process the video and language in a local-global manner.

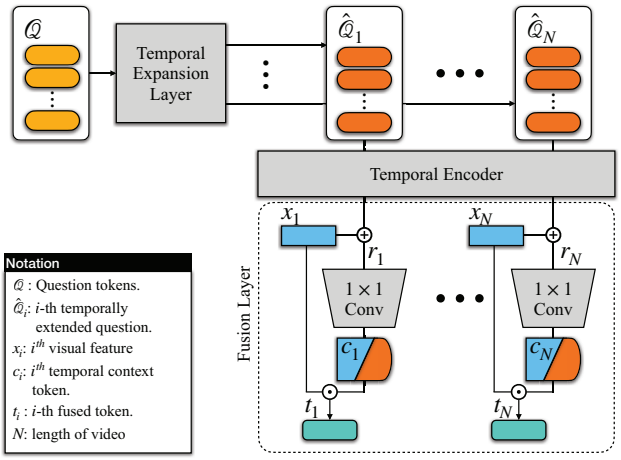

Figure 2. Detailed structure of Temporal Fusion Layers and Temporal Encoding Layers illustrating local context learning. As [24] utilize N independent layers to produce the N tokens, we also use N number of 1×1 convolution layers to make N temporal context tokens. The importance score c_i reveals the current frame's relationship with the expanded question. By dot product of vector x_i with matrix c_i, we get temporally encoded frame feature t_i.

3. Methodology

Using fusion layers, we propose a method having 3 major components: 1) question temporal encoding layers, 2) local context learning layers and 3) global context learning layers.

3.1. Overview

Given a video $\mathcal{V} \in \mathbb{R}^{N \times D_{\text{frame}}}$ and a question $\mathcal{Q} \in \mathbb{R}^{S_{\mathcal{Q}} \times D_{\text{text}}}$, the goal is to find the correct answer $\mathcal{A} \in \mathbb{R}^{S_{\mathcal{A}} \times D_{\text{text}}}$. $S_{\mathcal{Q}}$, and $S_{\mathcal{A}}$ denote the sequence lengths of question and answer tokens, respectively. D_{frame} and D_{text} are the dimensions of frame and text features, respectively. Formally, the VideoQA task can be formulated as Equation 1, where $\Psi(\cdot)$ is the VideoQA model and l is the number of provided answer choices.

$$\mathcal{A} = \Psi(\mathcal{V}, \mathcal{Q}, \mathcal{A}^*) : \mathcal{A}^* = \{a_1, a_2, ..., a_l\} \quad (1)$$

Our VideoQA model is built upon an object encoder $\text{E}_{\mathcal{O}}(\cdot)$, a language model $\text{E}_{\mathcal{L}}(\cdot)$ for interpreting $\{\mathcal{Q}, \mathcal{A}\}$ pairs, a node transformer $\text{NT}(\cdot)$ for graphs nodes, an edge transformer $\text{ET}(\cdot)$ for edges of the graph. The constructed graph is fed into object-question Fusion Layer $\Omega_{\mathcal{O}}(\cdot)$ to produce locally temporal context tokens with $\mathcal{Q}$. Simultaneously, another Fusion Layer $\Omega_F(\cdot)$ whose inputs are F and $\mathcal{Q}$ generates frame-question temporal context tokens. At this moment, we generate fine-grained temporal tokens which capture temporally local contexts. Finally, local context tokens are fed into a global transformer $\text{GT}(\cdot)$ to learn

the global contexts. We use the Multi-Head Self Attention layer [30] for global context learning. The details are explained in the below sections.

3.2. Multi-level Granularity Context Learning

Inspired by TokenLearner [25], we have built a Temporal Expansion Layer and Temporal Encoder followed by Multi-modal Fusion Layers. Detailed illustration can be found in Figure 2.

3.2.1 Temporal Expansion and Encoder

The goal of temporal expansion is to project language features along the temporal axis of local video features to capture the temporally local context. This is important for commonsense and temporal reasoning-based questions. The Temporal Expansion Layer is built using linear layers, as shown in Figure 2. N mapping layers $\{W_i\}_{i=1}^{N}$ are used to expand question tokens $\mathcal{Q}$ along the temporal axis. The expanded question for the i^{th} frame is formed as shown in Equation 2, where q represents a question token. The output of the Temporal Expansion Layer will have the same length as N, which is the length of the video.

$$\hat{\mathcal{Q}}_i = \{W_i \mathcal{Q}\}, \text{where } \mathcal{Q} = \{q_1, q_2, \ldots, q_{S_\mathcal{Q}}\} \quad (2)$$

The expanded questions $\hat{\mathcal{Q}} \in \mathbb{R}^{N \times S_\mathcal{Q} \times D_{\text{text}}}$ are forwarded to the Temporal Encoder, which projects a set of expanded question tokens into the local visual feature space for fine-grained fusion. Through Temporal Encoding, $\hat{\mathcal{Q}}$ becomes $\bar{\mathcal{Q}}$, which represents questions encoded with temporal information, where $\bar{\mathcal{Q}} \in \mathbb{R}^{N \times D_{\text{frame}}}$.

3.2.2 Temporal Fusion

Encoded $\bar{\mathcal{Q}}_i$ are reshaped and added with the visual features x_i to get fused representation r_i for the i^{th} frame. The local temporal context token c_i for the i^{th} context token can be obtained by Equation 3:

$$\mathrm{c}_i = \sigma(h_i^{1\times1}([r_i^{avg}; r_i^{max}])) \quad (3)$$

where $[;]$, and σ represent the concatenation operation and the sigmoid function, respectively. r_i represents the fused representation $\bar{\mathcal{Q}}_i + x_i$, and $h^{1\times1}$ represents the 1×1 convolution. r_i^{avg} and r_i^{max} represent the channel-wise average and max pooling operations, respectively. The choice of 1×1 convolution layers is motivated by their ability to learn the channel-wise importance scores for the individual frames in relevance with the text features while being lightweight. $X = \{x_i\}_{i=1}^{N}$ can be any visual features such as an object graph G or appearance features F. At this point, the text and visual features are inter-mixed, so the produced context tokens have captured a unified context from both modalities regarding local context. A dot product is computed between the context tokens $\{c_i(r_i)\}_{i=1}^{N}$ and the visual features $\{x_i\}_{i=1}^{N}$ to obtain N number of vision-text fused tokens $T_{X,Q}$. In Equation 4, $\odot$ denotes dot product.

$$T_{X,Q} = \{x_i \odot \mathrm{c}_i\}_{i=1}^{N} \quad (4)$$

Causal-VidQA (Test)

Method	ACC_D	ACC_E	ACC_P	ACC_C	ACC_A
HME [4]	63.36	61.44	28.91	30.92	46.16
CoMem [5]	64.08	62.79	31.40	32.54	47.70
HCRN [13]	65.35	61.61	32.56	32.65	48.04
HGA [11]	65.66	63.51	32.21	34.27	48.91
B2A [22]	66.21	62.92	31.14	35.16	48.86
Ours	**72.02**	**71.59**	**39.63**	**42.78**	**56.51**

Table 1. Results comparing to previous approaches on Causal-VidQA. ACC_D, ACC_E, ACC_P, and ACC_C represent accuracy on descriptive, explanatory, predictive, and counterfactual question types, respectively. ACC_A represents the overall accuracy. These results are obtained from the official leader board at `https://codalab.lisn.upsaclay.fr/competitions/6269#results` where detailed results can be seen.

Method	NExT-QA				MSVD-QA
	Test-C	Test-T	Test-D	Test-A	Test
HME [4]	46.76	48.89	57.37	49.16	-
IGV [17]	48.56	51.67	59.64	51.34	40.8
MHN [23]	-	-	-	-	40.4
VGT [34]	51.62	51.94	**63.65**	53.68	40.3*
HQGA [33]	49.04	**52.28**	59.43	51.75	41.2
Ours	**51.78**	52.05	63.63	**53.81**	**41.35**

Table 2. The Results of NExT-QA. Test-C represents causal questions while Test-T means Temporal questions. Test-D denotes Descriptive and Test-A shows overall accuracy. The best results are highlighted in bold. *Reproduced with official code.

4. Experiments

Setup. We use BERT [19] model for the Transformer encoder, Faster-RCNN [6] for object features and ResNet-101 [8] for frame features. We evaluate our approach on four VideoQA datasets i.e., Causal-VidQA [31], NExT-QA [32], MSVD-QA [35] and AGQA-2.0 [7]. Further details about experimental setup and data are provided in the Appendix sec. 1.

4.1. Results

4.1.1 Comparison with SoTA Methods

Causal and Commonsense Reasoning. Our method outperforms the state-of-the-art on Causal-VidQA, especially on predictive and counterfactual question types as shown in Table 1. To investigate the reasoning capability of the proposed method, further experiments were conducted by iso-

NExT-QA

Method	Test-C	Test-T	Test-D	Test-A
Ours w/ O	49.76	51.04	61.07	52.01
Ours w/ A	50.49	51.34	62.56	52.73
Ours w/ O + A	**51.78**	52.05	63.63	**53.81**
Ours w/ O + A + M	51.13	**52.31**	**63.91**	53.60

Table 3. Study of different features with the proposed multi-modal fusion module on NExT-QA. O, A, and M represents Object graph, Appearance, and Motion features.

	AGQA v2.0				
	Q Types	ML	PSAC[15]	HME[4]	HCRN[13]	**Ours**
Reasoning	obj-rel	9.39	37.84	37.45	40.33	**43.1**
	rel-act	**50**	49.95	49.9	49.86	49.78
	obj-act	50	50	49.97	49.85	**50.05**
	superlative	21.01	33.2	33.21	33.55	**35.4**
	sequencing	**49.78**	49.78	49.77	49.7	49.77
	exists	50	49.94	49.96	50.01	**50.42**
	dur comp	24.27	45.21	47.03	43.84	**47.54**
	act recog	5.52	4.14	5.43	5.52	**11.5**
Semantic	object	9.17	37.97	37.55	40.4	**43.04**
	relation	50	49.95	49.99	49.96	**50.37**
	action	30.11	46.85	47.58	46.41	**47.83**
Structure	query	13.05	31.63	31.01	36.34	**39.69**
	compare	50	49.49	49.71	49.22	**49.74**
	choose	**50**	46.56	46.42	43.42	46.53
	logic	50	49.96	49.87	50.02	**50.06**
	verify	50	49.9	49.96	50.01	**51.11**
	overall	10.99	40.18	39.89	42.11	**44.36**

Table 4. Comparison of the proposed method on the AGQA v2.0 dataset. The results are separated based on the question types. obj: object, rel: relationship, act: action, comp: comparison, seq: sequence

Causal-VidQA (Test)						
Method	I	ACC_D	ACC_E	ACC_P	ACC_C	ACC_A
G + Attn.	glob	**72.24**	70.95	37.77	42.49	55.86
G + Attn.	loc-glob	71.61	71.09	39.12	42.45	56.07
G + Ours	glob	70.16	68.98	34.81	40.83	53.69
G + Ours	loc-glob	72.02	**71.59**	**39.63**	**42.78**	**56.51**

Table 5. Study of different cross-modal fusion methods using the same graph representation G which is based on VGT. Attn. stands for Attention module. loc. and glob. stand for local and global, respectively. Column I represents the cross-modal fusion granularity level i.e., local or global.

Causal-VidQA (Test)						
Method	F	ACC_D	ACC_E	ACC_P	ACC_C	ACC_A
Replication	A	67.01	62.37	33.32	30.91	48.40
Expansion	A	**71.91**	**69.35**	**35.92**	**43.38**	**55.14**

Table 6. Results comparing Temporal Expansion with Token Replication without using object graphs, (i.e., ResNet-101 based frame encoding) on Causal-VidQA.

lating several modules as shown Table in 5. To our understanding, causality and commonsense reasoning is highly dependent on how scenes adapt as the time passes. Hence, if a method can leverage the temporal understanding it can have a better commonsense reasoning ability. This suggests the effectiveness of our temporal fusion module. Significant improvements are seen in predictive question accuracy ($ACCP_P$). A similar trend is observed in NExT-QA on causal questions in Table 2.

Temporal Understanding. The proposed method has strong reasoning, semantic and structural understanding on AGQA-2.0, outperforming with a significant margin on activity recognition-type questions as shown in Table 4. Some categories perform worse than random selection, such as relationship action. Except for 3 categories, our model shows the best performance with low variance between models, hinting at the challenging nature of these questions. Question-type examples with qualitative results as shown in the Appendix sec. 2.

Conventional VideoQA. We further demonstrate the performance of our model in the open-ended QA setting using MSVD-QA. As can be seen in Table 2, our model outperforms other baselines with a reasonable margin.

4.2. Ablation Studies

In this section, we present the analysis of our proposed fusion module with a wide range of configurations on the Causal-VidQA dataset and the NExT-QA dataset.

4.2.1 Multi-Level Temporal Granularity

As can be seen in Table 5, we perform a study to compare the effect of local and global temporal context learning. We used two different fusion methods i.e., the proposed MMTF, and attention-based fusion. Then we performed two different experiments for each of these modules to demonstrate the effectiveness of learning both fined-grained (local) and coarse-grained (global) representations over coarse-only representation. It is observed that in both types of fusion methods, the performance improves when local and global temporal contexts are learned for both modalities as compared to global-only fusion. Specifically, the ACC_P which is the predictive question accuracy gains significant improvement over the global-only variant with an absolute improvement of 4.82 for MMTF and 1.35 for attention-based fusion.

4.2.2 Temporal Expansion and Fusion Module

We experiment using only frame features from ResNet-101 backbone with our temporal expansion and question token replication along the temporal axis of the video to demonstrate the effectiveness of the Temporal Fusion module. As shown in Table 6, there is significant gap between the overall accuracy (ACC_A). It is to be noted that the major contribution to this performance difference comes from the counterfactual questions (ACC_C) which hints towards the strong reasoning capabilities of MMTF.

5. Conclusion and Limitations

A novel vision-text fusion module is proposed, which jointly learns the multi-level granularity of the contexts. Moreover, we present a novel technique to expand text tokens along the video's temporal axis to learn the fine-grained fused contexts which can represent local events. We demonstrate that the proposed multi-level granularity context learning improves over the current state-of-the-art on four VideoQA datasets, especially on commonsense reasoning tasks. The temporal expansion of text tokens is a new concept, that aims to better align the temporal contexts of both videos and questions.

References

[1] Arjun Reddy Akula. *Gaining Justified Human Trust by Improving Explainability in Vision and Language Reasoning Models*. University of California, Los Angeles, 2021.

[2] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

[3] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020.

[4] 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*, pages 1999–2007, 2019.

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

[6] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015.

[7] 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.

[8] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[9] 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*, volume 34, pages 11021–11028, 2020.

[10] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2758–2766, 2017.

[11] Pin Jiang and Yahong Han. Reasoning with heterogeneous graph alignment for video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11109–11116, 2020.

[12] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning*, pages 5815–5826. PMLR, 2021.

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

[14] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. Hero: Hierarchical encoder for video+ language omni-representation pre-training. *arXiv preprint arXiv:2005.00200*, 2020.

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

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

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

[18] 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*, pages 1698–1707, 2021.

[19] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019.

[20] Antoine Miech, Jean-Baptiste Alayrac, Ivan Laptev, Josef Sivic, and Andrew Zisserman. Thinking fast and slow: Efficient text-to-visual retrieval with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9826–9836, 2021.

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

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

[23] Min Peng, Chongyang Wang, Yuan Gao, Yu Shi, and Xiang-Dong Zhou. Multilevel hierarchical network with multiscale sampling for video question answering. *arXiv preprint arXiv:2205.04061*, 2022.

[24] AJ Piergiovanni, Kairo Morton, Weicheng Kuo, Michael S Ryoo, and Anelia Angelova. Video question answering with iterative video-text co-tokenization. In *European Conference on Computer Vision*, pages 76–94. Springer, 2022.

[25] Michael Ryoo, AJ Piergiovanni, Anurag Arnab, Mostafa Dehghani, and Anelia Angelova. Tokenlearner: Adaptive space-time tokenization for videos. *Advances in Neural Information Processing Systems*, 34:12786–12797, 2021.

[26] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. *arXiv preprint arXiv:1908.08530*, 2019.

[27] 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*, pages 7464–7473, 2019.

[28] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019.

[29] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018.

[30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[31] Tan Wang, Chang Zhou, Qianru Sun, and Hanwang Zhang. Causal attention for unbiased visual recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3091–3100, 2021.

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

[33] Junbin Xiao, Angela Yao, Zhiyuan Liu, Yicong Li, Wei Ji, and Tat-Seng Chua. Video as conditional graph hierarchy for multi-granular question answering. AAAI, 2022.

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

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

[36] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Tubedetr: Spatio-temporal video grounding with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16442–16453, 2022.

[37] Zhou Yu, Jun Yu, Yuhao Cui, Dacheng Tao, and Qi Tian. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6281–6290, 2019.

[38] 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. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

[39] 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*, pages 8746–8755, 2020.