Appendix for "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

1. Experiment and Data Details

Settings: We compare our proposed method with several state-of-the-art methods on 4 VideoQA datasets: **Causal-VidQA** [4] which features scene description, evidence reasoning, and commonsense reasoning questions with multiple choice Question-Answer setting. The questions are categorized into 4 categories: 1) Descriptive, 2) Explanatory, 3) Predictive, and 4) Counterfactual. The "predictive" and "counterfactual" questions are the most challenging because the model should output the correct reason along with the correct answer. It is based on the Kinetics-700 action recognition dataset, including 666 action categories. It contains 26,900 videos which are split into train, validation, and test set having 18,776, 2,695, and 5,429 videos, respectively. **NExT-QA** [6] which is another causal VideoQA dataset, contains causal, and temporal interactions among multiple objects. It consists of approximately 47,700 manually annotated questions in the multiple-choice Question-Answer setting. Aside from causal-based datasets, we present results on a common VideoQA dataset namely, **MSVD-QA** [9]. It mainly focuses on the descriptive question types with a total of 50,000 Question-Answer pairs with open-ended answer settings having a vocabulary of over 1,600 words. **AGQA-2.0** [2] is based on Action Genome Question Answering dataset, and provides diverse question types such as Reasoning, Semantic and Structure type questions with a total of 2.27 million QA pairs.

Implementation details: We follow the previous works and compute appearance features with pre-trained ResNet-101 [3] and for object features, we use pre-trained Faster-RCNN [1] with ResNet-50 backbone [5]. Further, we uniformly sample each video into 8 clips with 4 frames each. Within each frame, we pick 5 object regions with top scores and extract their features. The frame features have a dimension of 2048, whereas the object features have a dimension of 1024 for Causal-VidQA and MSVD-QA For other datasets, the dimension of object features is 2048. Alongside object features, we also use the coordinates of the regions to find the same objects in different frames following the work of [8]. Similar to [6], we use a pre-trained BERT model to obtain text representation having the dimension of 768. The dimension in our proposed models is set to 512 and the number of layers for the node transformer, edge transformer, and global transformer is set to 3. The number of heads for the node transformer and global transfer is set to 8, and for the edge transformer, it is set to 5 to match the number of object regions. We set the batch size to 64 and the max epochs to 30.

* Corresponding author.

1.1. Graph Representation

We use a graph representation for modeling the video, following [8]. Object features $\mathcal{O}$ are encoded using convolution layers and fed to the Node Transformer $\mathrm{NT}(\cdot)$ to get self-attended object nodes $\hat{\mathcal{O}}$. These nodes are forwarded to the graph builder $\Gamma(\cdot)$ to obtain relation matrix R. The relation matrices are fed into the edge transformer $\mathrm{ET}(\cdot)$ to reason about inter-object relations. The self-attended relation matrix and object nodes are fed to the graph convolution module $\mathrm{GC}(\cdot)$ to generate spatial relations between object nodes in a frame.

$$\begin{aligned} &\hat{\mathcal{O}} = \mathrm{NT}(\mathrm{E}_{\mathcal{O}}(\mathcal{O})),\ R = \mathrm{ET}(\Gamma(\hat{\mathcal{O}})) \\ &G = \mathrm{GC}(\hat{\mathcal{O}}, R) \end{aligned} \tag{1}$$

1.2. Parallel Streams for Visual Features

Both the appearance features F and the object graph G are parallelly fed into two Temporal Context Fusion Modules as shown in Figure 1. The outputs of two Temporal Context Fusion Modules are $T_{F,\mathcal{Q}}$ and $T_{G,\mathcal{Q}}$ when fed with appearance features and graphs respectively. These fused features from separate streams have captured the frame-level information from the object features, appearance features, and question tokens. The outputs of the individual fusion modules are concatenated along the token axis and fed to a projection layer to aggregate the local context fea-

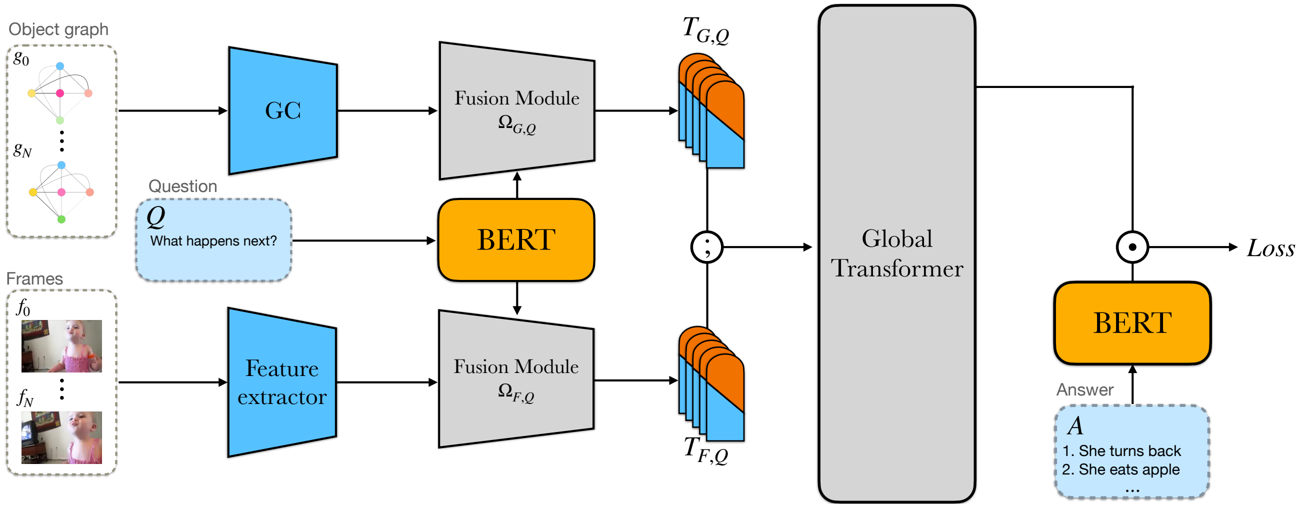


Figure 1: Overview of the proposed approach based on Multi-Modal Temporal Fusion (MMTF) module.

tures within a clip to obtain N tokens.

$$T_{C,Q} = \{W_g([T_{g_i,Q}; T_{f_i,Q}])\}_{i=1}^{N} \tag{2}$$

These fused features have captured the fine-grained (local) temporal context from both the language and video. Now to learn the coarse-grained (global) temporal context, these features are fed to the global transformer GT(.) to obtain a global context P.

1.3. Answer Prediction

We follow previous works [8, 7]. According to Equation 3, we adopt the mean pooling to generate a single representative video embedding vector.

$$v = \frac{1}{N}\sum_{i=1}^{N} P_i, \text{ where } P = \text{GT}(T_{C,Q}) \tag{3}$$

For multiple-choice tasks, we calculate the similarity between v and answer candidates $\mathcal{A}^*$. As formulated in Equation 4, we use cross-entropy loss for the multiple answer choices, where $\odot$ denotes element-wise dot product.

$$\mathcal{L} = -\sum_{i=1}^{|\mathcal{A}^*|} y\log(s_i), \text{ for } s_i = \frac{v \odot \mathcal{A}_i^*}{\sum_{j=1}^{|\mathcal{A}^*|} v \odot \mathcal{A}_j^*} \tag{4}$$

However, in open-ended tasks such as MSVD-QA, we leverage a dictionary as answer candidates. Therefore, the same loss function is utilized for open-ended tasks. Namely, $\mathcal{A}^*$ consists of a set of candidate words.

2. Results Analysis

2.1. Temporal Similarity with Answer

In this section, we present the analysis of the temporal fusion module to verify its effectiveness in understanding the temporal context of the question and the video. We formulate this analysis such that if the proposed module learns the temporal features effectively, then the correct answer should have a higher similarity with the fused features at that specific temporal region.

For computing the similarity between the answer embedding and the temporally fused frame features we use the cosine-similarity metric. In MMTF, a video's frames are split into multiple clips. The features corresponding to the frames within a clip are aggregated after the temporal fusion is performed. We used these aggregated fused features to perform the similarity analysis with the answer. As the motivation for this analysis is to verify the ability of MMTF to attend to the multi-modal temporal information, we choose the temporal questions for this analysis.

As shown in Figure 2a, the question "*which object were they behind before standing up but after lying on a bed?*" is pinpointing towards a very specific temporal event. Upon close inspection, it can be observed that the cosine-similarity value is the highest for that specific moment (*just before standing*). It is to be noted that the queried event consists of a narrow range of frames yet MMTF is able to learn the fused features which are relevant to the correct answer.

In Figure 2b, results are shown for another question with the same video. This question refers to the overlap of two events specifically i.e., "*while putting a shoe*", and "*what object did the person close?*". The highest cosine-similarity values are found at the exact location on the temporal axis where the person is interacting with the shoes and closing the *laptop*. Interestingly, the similarity score is the second highest right before *wearing shoes* as the *person* starts moving while *closing the laptop*.

Finally, Figure 2c is taken from the most challenging type of questions. These questions most require temporal understanding, as they query about the sequence in which certain events have occurred. In the given example, the question text queries the sequence in which the *person* is *tidying* the *objects*. It can be seen that the *thing they stood on* is *floor*. If the video frames are observed, the *person* is *tidying the floor* towards the end of the video and is *tidying* the *object they were behind* at the beginning of the video. As the question is about the sequence, the highest similarity is found at the boundary of the two events, i.e., the exact point in time, where the person shifts from one activity to the other.

2.2. Qualitative Results

2.2.1 Causal-VidQA

Causal-VidQA features counterfactual and predictive question types which are the foremost representative question types for commonsense reasoning. Moreover, for a prediction to be evaluated as correct, the model must predict the right reason along with the answer prediction. Therefore, this evaluation setting helps in better evaluation of a model's commonsensical question-answering ability.

In this section, we demonstrate the effectiveness of the proposed MMTF over the previous state-of-the-art VGT on two video samples. As can be seen in Figure 3, MMTF's strengths lie in the reasoning-based questions, whereas VGT's strength lies in the descriptive, or explanatory questions. In the first video's **predictive** question, it can be seen that VGT captured some atomic representations i.e., "*[person_3] is going to continue moving back and forth in combination with hands*", but failed to recognize it as an action class. It is further highlighted by wrong **reason** prediction i.e., "*[person_3] walked away*", which is irrelevant in the given visual context. On the contrary, MMTF recognizes the action of "*moving hands*" as "*polishing shoes*". It is further supported by the **reason** prediction that the model understands that "*holding the blacking brush*" and a certain hand movement represents the action "*polishing shoes*".

As both methods use the same object graphs, and frame features the predictive strengths can be associated with the proposed temporal fusion module. MMTF utilizes both fine-grained and coarse-grained temporal context fusion, therefore it has superior performance on temporal context-sensitive questions as compared to VGT (which only uses coarse-grained fusion).

In the second video's **explanatory** question "*why is [person_1] sitting on the [chair_1]*?", VGT predicts wrong answer i.e., "*[person_1] is playing with [chair_1]*". In our understanding, this is due to VGT's inability to find a fine-grained relationship and therefore is prone to language bias as the only answer choice with the word "chair" is **A1**. Similarly, in the **counterfactual** question, it is also evident that VGT misunderstood the scene and predicts "*[person_1] may not keep playing.*"

In summary, the answer and reason predictions of the proposed MMTF model are well-aligned whereas, VGT's answer and reasons do not depict proper alignment as in some scenarios, the answer is correctly predicted but the predicted reason has no relation with the question, and video's context and vice versa.

2.2.2 AGQA-2.0

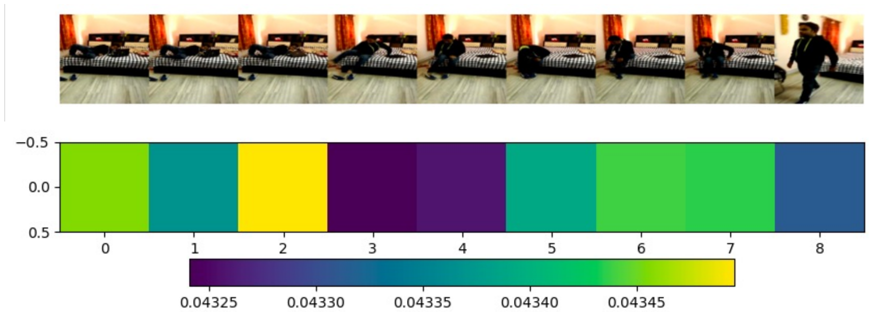
In Figure 4, we present qualitative results for three different question types from the AGQA-2.0 dataset. Each category is further divided into subcategories. For example, under the "reasoning" category, questions of the "object-relationship" type require reasoning capabilities to assess subject-object relationships. In the given example, the question asks about a person's relationship with an object. To answer such a complex question, the model must be able to recognize and temporally localize at least two actions ("watching" and "holding") and have object recognition capabilities to determine whether these interactions involve one or two different objects. The "exists" category asks about the existence of an interaction between a specific subject and object. In this case, although there is a laptop in the video, there is no interaction between it and the person; however, the model outputs an incorrect answer. It should be noted that in many cases, the text of these questions can be confusing and may contribute to lower overall results for some question types such as those in the "semantic" category.

In the category "structure", the question under "query" inquires about spatial information after a specific event has occurred. However, it is ambiguous as the model's prediction seems accurate as well as the ground truth because "bed" (prediction) and "laptop" are both on the side of the person. The questions in the "choose" category provides choices to select. However, it is different from typical multiple-choice question answering as the choices are part of the question text. Finally, the "compare" category inquires about the sequence of two events which is challenging to answer.

2.2.3 MSVD-QA

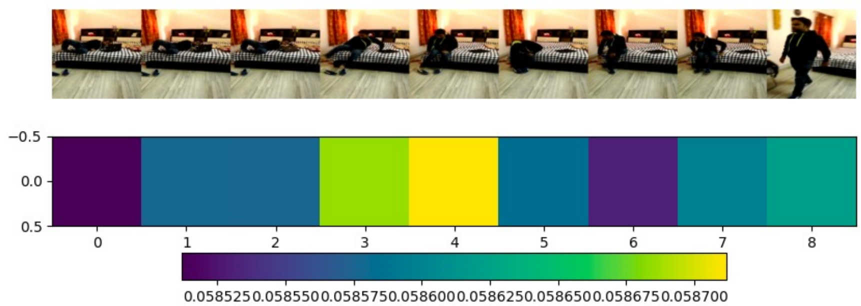
Apart from commonsense Video Question Answering, we evaluated the proposed method on MSVD-QA, which is a conventional VideoQA dataset. The question-answer pairs in this dataset are descriptive and explanatory types. We present qualitative results on two video samples from MSVD-QA in Figure 5 and compare it with the previous state-of-the-art on VideoQA i.e., VGT [8]. In Figure 5a it can be seen that the proposed method has the capability to answer descriptive-type questions as well. Further, we demonstrate the failure cases of our model in Figure 5b, where the model outputs the wrong answer. However, it is

Question: Which object were they behind before standing up but after lying on a bed?
Answer: shoe



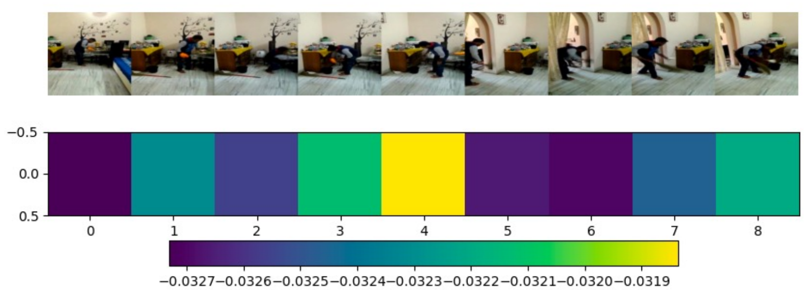
(a)

Question: While putting on a shoe, which object did the person close?
Answer: laptop



(b)

Question: Was the person tidying up the object they were behind first before or after tidying something on the thing they stood on?
Answer: before



(c)

Figure 2: Cosine similarity between answer embedding and temporally fused Question-Video features.

interesting to note that there is not enough information in the video to identify the gender of the person.

2.3. Effect of Visual Features

Moreover, we performed experiments with the proposed MMTF module while using different visual features i.e., appearance, object graphs, and motion features to demonstrate the feature-agnostic properties of MMTF. For these experiments, we use the NExT-QA dataset. Our method achieved comparable results on all visual features which show the generalization ability of MMTF. In our understanding, this is due to the joint temporal context fusion as its main purpose is to learn to represent text and vision features in a combined space while focusing on the temporal context. As this module learns to fuse very different modalities, it learns to generalize well so different visual features are trivial as compared to learning visual and text features. The results in Table 1 demonstrate that the proposed fusion module is

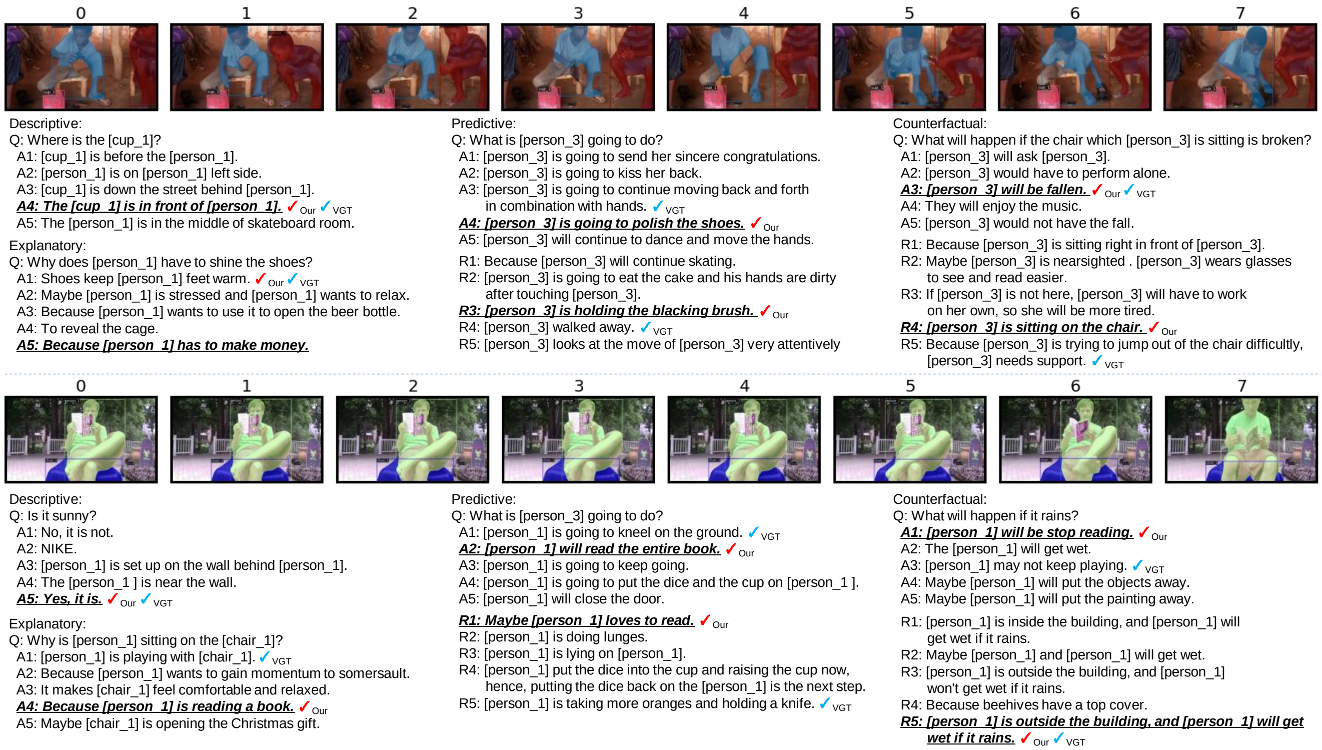


Figure 3: Visualization of VGT and our comparative results on Causal-VidQA. The correct answers (A) and reasons (R) are underlined.

Figure 4: Example of 9 different categories of AGQA-2.0 [2]. Green represents the correct questions and red represents the wrong questions. The above video contains a person taking off his/her shoes and leaving the bed.

feature-agnostic and achieves satisfactory results with several types of visual features such as object, appearance, and motion features.

References

[1] Ross Girshick. Fast r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 1440–1448, 2015. 1

[2] 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. 1, 5

[3] 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 recog-*

Questions	Answers	VGT	Ours
who is climbing up a rock wall facade?	woman	man	woman
who is climbing an artificial rock wall?	woman	man	woman
who is climbing up a rock wall?	woman	man	woman
who is engaged in rock climbing?	woman	man	woman
who climbs a rock wall?	woman	man	woman
who is climbing a climbing wall?	woman	man	woman
who is rock climbing?	woman	man	woman

(a)

Questions	Answers	VGT	Ours
who opens a box containing an assault rifle?	man	man	woman
who is opening a box that has a gun in it?	man	man	man
who is opening a box with a gun inside of it?	man	man	man
who opens up a box containing a gun?	man	man	woman
who is opening a box containing a gun?	man	man	woman
who opens a box containing a rifle?	man	man	woman
who is opining a box?	man	man	woman
who opens a box?	man	man	woman
who is showing a gun?	man	man	woman
who shows a rifle inside?	man	man	woman
who is opening a box up?	man	man	woman

(b)

Figure 5: Qualitative results from MSVD-QA, which is an open-ended Video Question Answering Task. (a) Correct prediction (b) Failure case in an incomprehensible scenario.

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 1: 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.

nition, pages 770–778, 2016. 1

[4] Jiangtong Li, Li Niu, and Liqing Zhang. 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*, pages 21273–21282, 2022. 1

[5] Yanghao Li, Saining Xie, Xinlei Chen, Piotr Dollar, Kaiming He, and Ross Girshick. Benchmarking detection transfer learning with vision transformers. *arXiv preprint arXiv:2111.11429*, 2021. 1

[6] 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. 1

[7] 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. 2

[8] 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. 1, 2, 3

[9] 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. 1