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

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

Overview

In the supplementary material, we provide additional details for the main paper:

- More discussions of top-k selector in our proposed model, MIST , in Sec. A.
- More details of experimental settings in Sec. B.
- More experimental results in Sec. C.
- More visualizations of prediction results in Sec. D.

A. More Discussions of Top-k Selector

As illustrated in the main paper, MIST calculates multi-modal attention between segment/patch features and question features, then performs top-k hard selection over segment/patch features. Commonly used hard selection, such as argmax or top-k selection functions, will stop the back-propagation of gradients and affect the training of the attention module. Thus, we use the Gumbel-Softmax trick to perform the differentiable hard selection. Note that the standard Gumbel-Softmax trick is designed for top-1 selection, and extending it to top-k selection is still an open problem.

The core of extending Gumbel-softmax to top-k selection is multiple sampling. The main difference between different implementations is whether sample with replacement. Many previous works [6, 7, 12] for sequence generation tasks choose to sample without replacement to generate diverse sequences, i.e., sampling the first element, then renormalizing the remaining probabilities to sample the next element, etcetera. But as also mentioned in their papers, sampling without replacement means that the inclusion probability of element i is not proportional to p_i. For an extreme example, if we sample $k = n$ elements, all elements are included with probability 1. It may affect their word sequence generation tasks a little, because the number of candidate words in each step is usually much larger than top-k, so all top-k choices could be plausible. However, for our targeted long-form VideoQA task, the model needs only to select the segment related to the question. And in some cases, the question could only involve one segment. Thus, we expect the model learns to enhance the most related segment in such cases by re-sampling it, instead of forcing it to select an irrelevant segment.

B. More Details about Experiment Setting

Evaluation Metrics. In the main paper, we evaluate our method on AGQA v2 [5], NExT-QA [15], STAR [14], and Env-QA [3]. In the supplement, we additionally evaluate MIST on AGQA v1 [4]. For the AGQA v1/v2, NExT-QA, and STAR, we follow their paper using the QA accuracy as the metric, i.e., if the answer is the same as the ground truth, then the model gets 1 score; otherwise gets 0. For Env-QA, we follow the dataset paper that first decomposes the phrase answer into several parts, called role-value format, then uses an IOU-like score to evaluate the similarity between the prediction and ground truth.

Implementation Details. For the Image Encoder in MIST, due to the slightly insufficient alignment ability between the patch feature of the original CLIP and text features, in all experiments, the patch features were obtained by dividing the image into 4×4 patches and then sending them separately to CLIP. As illustrated in Sec. 3.3 of the main paper, we calculate the similarity between the answer candidates and the combining feature of video and question to predict the answer. For AGQA v1/v2, NExT-QA, and STAR, the answer prediction follows the above-mentioned method. But for Env-QA, since the model needs to predict a set of role-value answers, we slightly modify the answer prediction module, i.e., plug a set of classifiers to predict the answer in each role. Then, we sum the losses of all classifiers to train the whole model.

C. Experimental Results

Performances on AGQA v1. Since the dataset creators [4] of AGQA recommend using its v2 version, we evaluated our method and conducted ablation studies on AGQA v2 in the main paper. Here, we report the results on AGQA v1 to compare with some recent works [9, 11, 16] which didn't report their results on AGQA v2.

Method	Binary	Open	All
PSAC [10]	54.19	27.20	40.40
HCRN [8]	58.11	37.18	47.42
HME [2]	59.77	36.23	47.74
DualVGR [13]	55.48	40.75	47.80
HQGA [16]	56.15	39.49	47.48
DSTN-E2E [11]	57.38	42.43	49.60
Temp[ATP] [1]	59.60	49.16	54.17
MIST - CLIP	63.36	53.51	58.26

Table 1. QA accuracies of state-of-the-art methods on AGQA v1 test set.

From the results in Table 1, we can see that MIST outperforms state-of-the-art methods by a large margin. Compared to the previous SOTA, which uses the same feature, i.e., Temp[ATP], MIST obtains about 4% performance boost. In addition, [9,11] proposed specific designs, such as modular networks and dependency attention modules, tailored to compositional questions in AGQA. MIST achieves better compositional reasoning ability with a more general design for long-form video QA.

Comparison of Different Top-k Selectors. We try different top-k selectors to show the effect of different implementations:

- G-S w. Replacement: We sample the region features of Top_k segments by sampling segment indexes with replacement Top_k times. It is our default setting in the main paper.
- G-S w/o. Replacement: We sample the region features of Top_k segments by sampling segment indexes without replacement Top_k times. Specifically, after each selection, the probabilities of selected segments on next round sampling are set to 0.
- Non-params. Attention: Another naive solution for addressing training issue of attention module is to propose a non-parametric attention module. Specifically, the selector decides whether a certain segment should be selected based on the matching score of its pre-trained visual feature and the given question feature. The pre-trained feature already has a certain matching ability. Thus, we first normalize the pre-trained visual features and question features, then perform a dot product to obtain the attention score. Note that there is one limitation for this model, because the attention strategy cannot be trained, so no matter how many iterations, the selected segments are all the same. So, this variant has only one layer of ISTA.

Method	# of ISTA layer	Binary	Open	All
G-S w/o. Replacement	2	58.28	50.56	54.39
G-S w. Replacement	2	57.68	50.06	53.84
G-S w. Replacement	1	56.84	49.18	52.98
Non-params. Attention	1	56.34	49.52	52.91

Table 2. QA accuracies of MIST with different top-k selectors on AGQA v2 test set.

From the results in Table 2, we can see that sampling with replacement slightly outperforms the one without replacement. This may be because of the issue we mentioned in Sec A, i.e., sampling without replacement may introduce some irrelevant segments that disturb the reasoning. In addition, we can see that Non-params. Attention achieves similar performance with G-S w. Replacement with 1 ISTA layer, but is worse than G-S w. Replacement with 2 ISTA layers. It indicates that pre-trained features do already have a certain matching ability, but we need to modify it to support iterative attention.

D. More Visualizations of MIST

In the Figure 1, we show more visualizations of the prediction results of MIST . It can be seen that our model can select video clips and image regions relevant to the question. We also find the model to be wrong in the following cases: 1) The object is partially visible, and the model needs to infer what the object is from the video context. For example, in Q3, a partially visible lady puts the box aside at the beginning of the video. The lady is fully visible only in the latter frames. So, given a question requiring locating segments where a lady appears in the beginning, the model incorrectly finds the segment with a fully visible lady. 2) The video contains a large number of similar events. It is hard for our model to distinguish these subtle differences. In Q5, the boy first stirs the food in the pot, then the lady takes over the stirrer, and the boy pulls the lady's arm to want to get the stirrer back. All these events are quite similar, and it is still relatively hard to locate the correct segments with the given question.

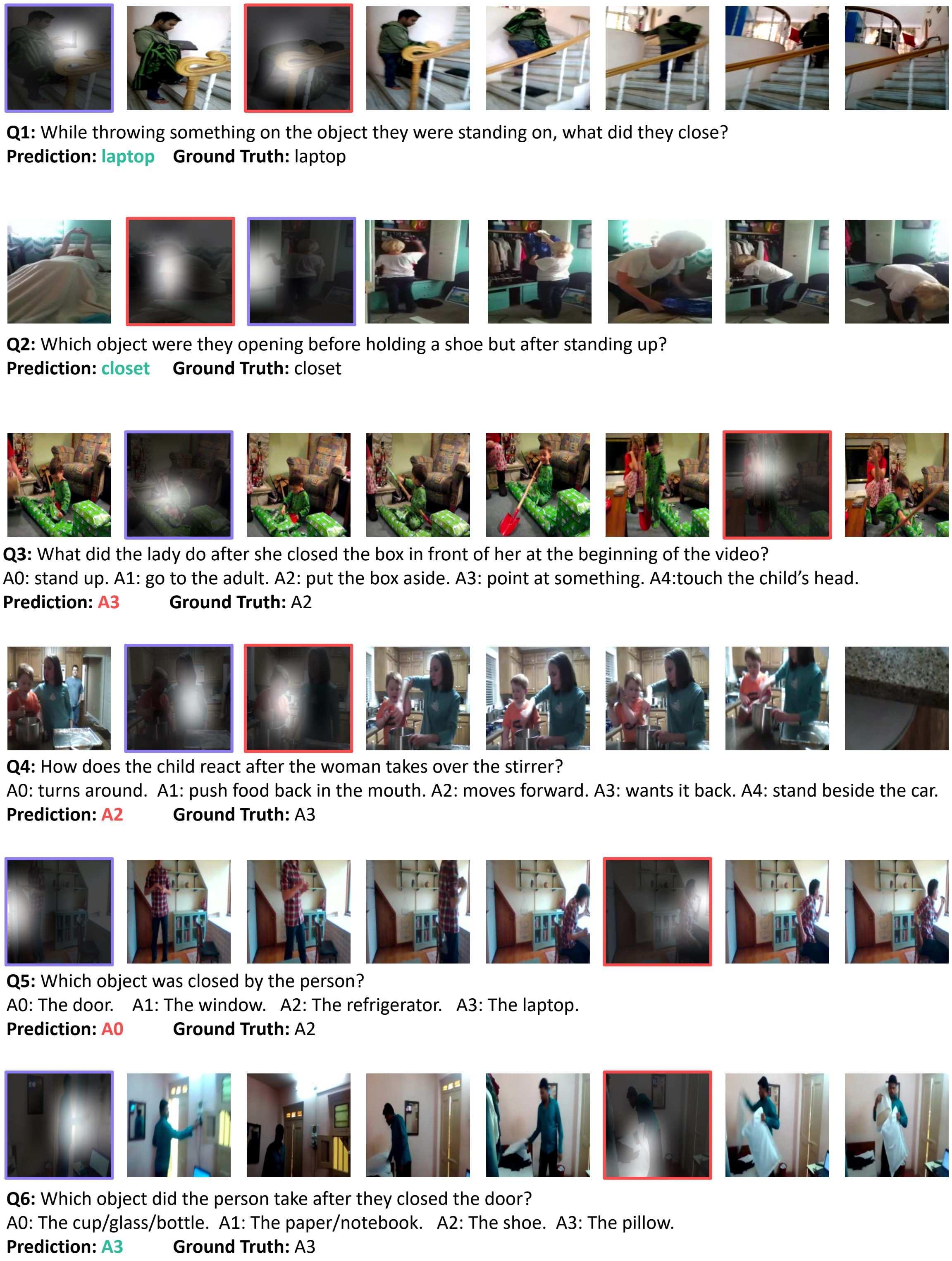
References

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

[2] 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. 2

[3] Difei Gao, Ruiping Wang, Ziyi Bai, and Xilin Chen. Env-qa: A video question answering benchmark for comprehensive understanding of dynamic environments. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1675–1685, 2021. 1

[4] Madeleine Grunde-McLaughlin, Ranjay Krishna, and Maneesh Agrawala. Agqa: A benchmark for compositional



Q1: While throwing something on the object they were standing on, what did they close?
Prediction: laptop **Ground Truth:** laptop

Q2: Which object were they opening before holding a shoe but after standing up?
Prediction: closet **Ground Truth:** closet

Q3: What did the lady do after she closed the box in front of her at the beginning of the video?
A0: stand up. A1: go to the adult. A2: put the box aside. A3: point at something. A4:touch the child's head.
Prediction: A3 **Ground Truth:** A2

Q4: How does the child react after the woman takes over the stirrer?
A0: turns around. A1: push food back in the mouth. A2: moves forward. A3: wants it back. A4: stand beside the car.
Prediction: A2 **Ground Truth:** A3

Q5: Which object was closed by the person?
A0: The door. A1: The window. A2: The refrigerator. A3: The laptop.
Prediction: A0 **Ground Truth:** A2

Q6: Which object did the person take after they closed the door?
A0: The cup/glass/bottle. A1: The paper/notebook. A2: The shoe. A3: The pillow.
Prediction: A3 **Ground Truth:** A3

Figure 1. **Qualitative results of MIST .** We visualize its prediction results along with spatial-temporal attention, where the frames with purple and red outlines indicate the highest temporal attention score in the first and second ISTA layers, respectively.

spatio-temporal reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11287–11297, 2021. 1

[5] 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

[6] Wouter Kool, Herke Van Hoof, and Max Welling. Stochastic beams and where to find them: The gumbel-top-k trick for sampling sequences without replacement. In *International Conference on Machine Learning*, pages 3499–3508. PMLR, 2019. 1

[7] Wouter Kool, Herke van Hoof, Max Welling, et al. Ancestral gumbel-top-k sampling for sampling without replacement. *J. Mach. Learn. Res.*, 21:47–1, 2020. 1

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

[9] Jihyeon Lee, Wooyoung Kang, and Eun-Sol Kim. Dense but efficient videoqa for intricate compositional reasoning. *arXiv preprint arXiv:2210.10300*, 2022. 1, 2

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

[11] Zi Qian, Xin Wang, Xuguang Duan, Hong Chen, and Wenwu Zhu. Dynamic spatio-temporal modular network for video question answering. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4466–4477, 2022. 1, 2

[12] Kirill Struminsky, Artyom Gadetsky, Denis Rakitin, Danil Karpushkin, and Dmitry P Vetrov. Leveraging recursive gumbel-max trick for approximate inference in combinatorial spaces. *Advances in Neural Information Processing Systems*, 34:10999–11011, 2021. 1

[13] Jianyu Wang, Bing-Kun Bao, and Changsheng Xu. Dualvgr: A dual-visual graph reasoning unit for video question answering. *IEEE Transactions on Multimedia*, 24:3369–3380, 2021. 2

[14] Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. Star: A benchmark for situated reasoning in real-world videos. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. 1

[15] 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

[16] 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