Perceive, Query & Reason: Enhancing Video QA with Question-Guided Temporal Queries Supplementary Material

Roberto Amoroso[3,*] Gengyuan Zhang[1,2,*] Rajat Koner[1]
Lorenzo Baraldi[3] Rita Cucchiara[3,4] Volker Tresp[1,2]
[1]LMU Munich [2]MCML [3]University of Modena and Reggio Emilia [4]IIT-CNR
roberto.amoroso@unimore.it zhang@dbs.ifi.lmu.de

In this supplementary material, we provide additional implementation details about our PQR model. The sections delve into critical aspects, including ablation studies, hyperparameter configurations, and a thorough analysis of the impact of linguistic bias on performance.

A. Ablation Studies

In this section, we present additional ablation studies to further analyze our model's behavior. Section A.1 investigates the effect of varying the number and dimensionality of the layers in T-Former. Section A.2 explores the extent of linguistic bias within benchmark datasets.

A.1. T-Former settings

We investigate the impact of layer number and intermediate dimensionality in the feed-forward layer of the T-Former, as shown in Tab. 1. Our experiments demonstrate that increasing the number of hidden layers improves model performance, while larger bottleneck dimensionality yields the opposite effect. Our findings suggest that a configuration of 2 hidden layers with a 768-dimensional feed-forward layer yields the best performance.

#Linear Layers	Bottleneck Dim	NExT-QA			
		Tem.	Cau.	Des.	Avg.
2	3,072	72.4	75.8	81.7	75.7
2	1,536	72.5	**77.0**	82.5	76.4
2	768	**72.8**	76.9	**84.7**	**76.7**
1	768	71.7	75.0	82.9	75.2

Table 1. Effect of different feed-forward bottleneck size. Increasing the number of linear layers improves the model performance, but larger bottleneck dimensionality affects the results.

Method	#Pre-train videos/images	TGIF		MSRVTT
		Act.	Trans.	MC
All-in-one	283M	95.5	94.7	92.3
VIOLET	186M	92.5	95.7	91.9
MERLOT	180M	94.0	96.2	90.2
Singularity	17M	-	-	92.1
Clover	5M	94.9	98.0	95.0
ClipBERT	200k	82.8	87.8	88.2
PQR (Ours)	**0**	96.1	98.4	96.2

Table 2. Extending comparison to additional datasets. Our PQR consistently outperforms other baseline models despite being trained solely on the target dataset.

A.2. Exploring Linguistic Bias

Linguistic bias in video question-answering datasets is a significant concern when using Large Language Models (LLMs). We conduct a comprehensive analysis to verify if the questions in the benchmarks contain biases that enable models to answer correctly without visual input.

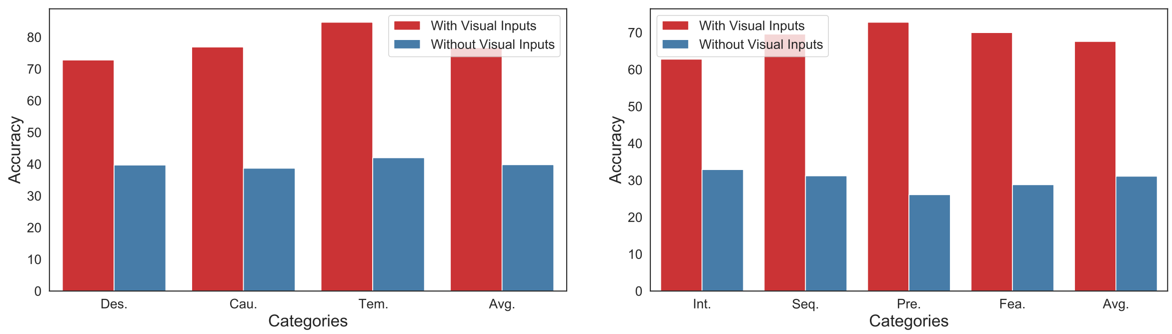
Fig. 1 presents the full performance results across different datasets and categories. Our observations indicate that in the absence of visual information, LLM reasoners exhibit modest performance, comparable to a "blind guess". This finding highlights the robustness of the video question-answering benchmarks, ensuring they are minimally influenced by linguistic bias.

Notably, the performance gap between our model and LLM reasoners is even more pronounced in categories such as causal and temporal reasoning. This underscores the effectiveness of our approach in leveraging visual information rather than being overly dependent on linguistic cues.

A.3. Extended Results

We further evaluate PQR on the TGIF-QA [1] and MSRVTT-MC [2] datasets, as shown in Tab. 2. Notably, PQR is trained solely on the target dataset without requiring extensive pre-training on millions of videos. Despite this, it consistently outperforms other baseline models.

*Equal contribution.



(a) NextQA

(b) STAR

Figure 1. **Exploring Linguistic Bias**: We observe that LLM reasoners can only achieve a modest performance, akin to a "blind guess" when visual inputs are absent.

Dataset	Batch Size	Epochs	Iterations per Epoch	Warmup Epochs	Cooldown Epochs	Initial LR	Warmup LR	Minimum LR
NExT-QA	2	10	2500	1	2	3e-5	8e-6	1e-6
STAR	2	10	5000	1	2	5e-5	1e-5	1e-6
How2QA	4	10	5000	1	5	3e-5	8e-6	1e-6
VLEP	4	10	1000	1	5	2e-5	7e-6	1e-6

Table 3. PQR training hyperparameters for different datasets.

B. Additional Implementation Details

B.1. Hyperparameters

In this section, we provide a detailed overview of the training hyperparameters used across all benchmark datasets to unsure reproducibility. Tab. 3 presents the optimal values for key parameters, including batch size, total epoch numbers, number of iteration steps per epoch, warmup and cooldown epochs, and learning rate.

References

[1] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *CVPR*, 2017. 1

[2] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*, 2016. 1