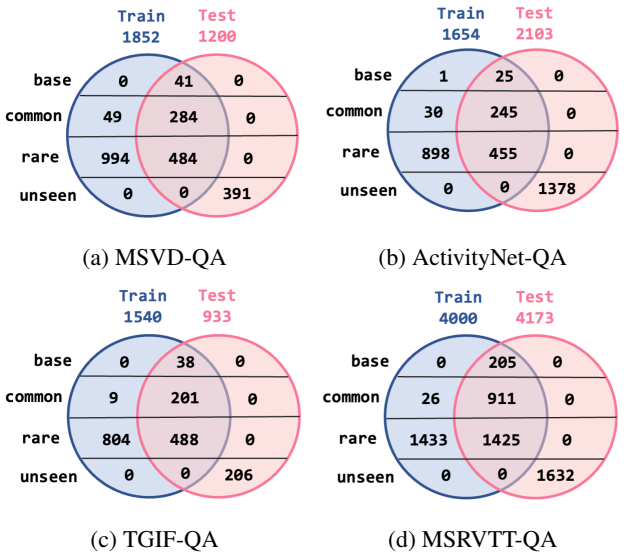
Open-Vocabulary Video Question Answering: A New Benchmark for Evaluating the Generalizability of Video Question Answering Models (Supplementary Materials)

Dohwan Ko Ji Soo Lee Miso Choi
Jaewon Chu Jihwan Park Hyunwoo J. Kim*

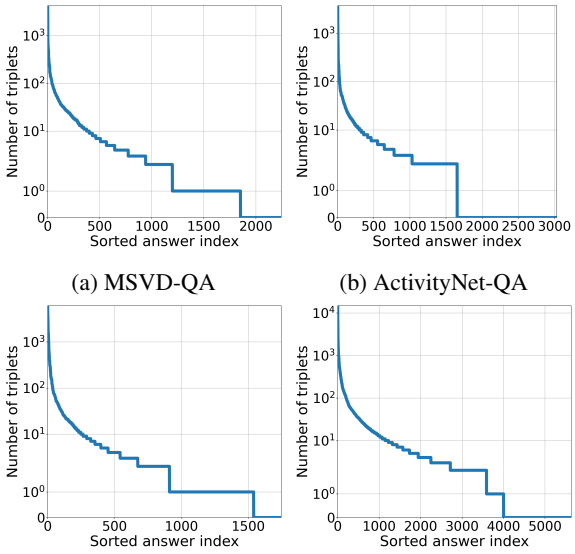
Department of Computer Science and Engineering, Korea University
{ikodoh, simplewhite9, miso8070, allonsy07, jseven7071, hyunwoojkim}@korea.ac.kr



(a) MSVD-QA (b) ActivityNet-QA

(c) TGIF-QA (d) MSRVTT-QA

Figure A1: **Dataset Venn diagram.** The distribution of rare, common, and frequent categories in train and test sets for four benchmark datasets. The total number of vocabularies for each set is specified under the corresponding title.



(a) MSVD-QA (b) ActivityNet-QA

(c) TGIF-QA (d) MSRVTT-QA

Figure A2: **Dataset Statistics.** Sorted frequency statistics for each answer candidate reveal long tail distribution for all datasets.

A. Dataset details

Fig. A1 presents the distribution of answer candidates for the base, common, rare, and unseen answer categories in MSVD-QA, ActivityNet-QA, TGIF-QA, and MSRVTT-QA respectively. Note that the test answer candidates are composed mostly of rare and unseen answers, *e.g.*, the number of rare and unseen answers (488 + 206) possess about 74% of the test answer candidates (933) in TGIF. In terms of base and common answers, most of them also appear in the test set. Yet interestingly, for each dataset, more than half of the rare answers do not appear in the test set. Furthermore, as depicted in Fig. A2, four datasets exhibit a long-tail answer distribution. Therefore, due to such imbalanced distribution, it is necessary to design the model under the open-vocabulary setting instead of the closed-vocabulary.

B. Implementation details

All-in-one [4]. The model is fine-tuned on four datasets with a batch size of 512 for 20 epochs. The learning rate is 1e-4 with a warm up step of 10% of the total iterations. AdamW optimizer [5] is used. For video features, 3 video frames are randomly sampled and resized to 224×224.

*Corresponding author.

Models	MSVD-QA						ActivityNet-QA						TGIF-QA						MSRVTT-QA					
	B	C	R	U	T	M	B	C	R	U	T	M	B	C	R	U	T	M	B	C	R	U	T	M
CVQA																								
Random	-	-	-	-	0.1	-	-	-	-	-	0.1	-	-	-	-	-	0.1	-	-	-	-	-	0.1	-
CLIP [1]	-	-	-	-	7.2	-	-	-	-	-	1.2	-	-	-	-	-	3.6	-	-	-	-	-	2.1	-
JustAsk [2]	17.1	10.1	12.8	0.0	13.5	7.0	19.9	8.6	8.3	0.0	12.3	2.8	28.4	10.4	9.9	0.0	23.8	6.9	5.9	5.5	5.5	0.0	5.6	3.3
FrozenBiLM [3]	**46.4**	**26.6**	12.6	0.0	33.7	9.9	44.1	**17.9**	7.4	0.0	25.9	3.8	48.9	27.4	11.0	0.0	41.9	11.5	**19.3**	**13.9**	0.0	0.0	**16.7**	3.2
OVQA																								
JustAsk+	18.2	12.9	13.5	13.1	15.7	11.4	12.8	5.9	6.2	**6.7**	9.4	**6.3**	29.5	12.3	12.7	**13.2**	25.3	11.9	6.0	5.2	5.5	**4.6**	5.8	4.5
FrozenBiLM+	46.3	**26.6**	**16.5**	**13.2**	**34.9**	**13.7**	**45.3**	17.3	**8.9**	3.1	**27.3**	6.0	**49.1**	**27.6**	**14.7**	8.1	**42.5**	**15.4**	15.5	11.7	**9.3**	4.3	14.1	**6.0**

Table A1: **Comparison with zero-shot state-of-the-art models.**

Models	Answer encoder	MSVD-QA						ActivityNet-QA						TGIF-QA						MSRVTT-QA					
		B	C	R	U	T	M	B	C	R	U	T	M	B	C	R	U	T	M	B	C	R	U	T	M
All-in-one+	CLIP	62.4	24.3	0.5	0.1	40.1	5.3	64.4	25.9	0.6	0.2	36.7	2.6	77.3	29.7	2.0	0.0	63.0	8.0	49.3	7.8	0.2	**0.0**	37.9	2.8
	DeBERTa	**62.8**	**34.0**	**6.3**	**0.4**	**43.8**	**9.4**	**64.9**	**35.9**	**9.8**	**0.5**	**40.2**	**6.8**	**78.3**	**39.3**	**10.2**	**0.4**	**66.0**	**13.2**	**49.8**	**14.6**	**1.6**	**0.0**	**39.5**	**4.7**
VIOLET+	CLIP	68.0	31.0	1.5	**0.1**	45.5	7.4	**64.3**	33.8	2.6	0.1	38.6	3.9	76.3	29.4	2.5	0.0	62.4	8.8	52.7	7.4	0.4	**0.0**	40.3	3.0
	DeBERTa	**70.6**	**38.8**	**6.7**	**0.1**	**49.5**	**10.7**	63.4	**37.1**	**9.2**	**0.6**	**39.7**	**6.1**	**77.3**	**38.9**	**10.8**	**2.0**	**65.3**	**14.3**	**53.8**	**14.7**	**0.9**	**0.0**	**42.4**	**4.5**

Table A2: **Ablation study on the answer encoder type.**

Then each frame is split into patches of size 14×14. In the setting of CVQA, the number of training and test answers are identical to one another with MSVD 1000, MSRVTT is 1500, ActivityNet is 1000, and TGIF is 1540.

VIOLET [6]. For all experiments, we employ the AdamW with $\beta = (0.9, 0.98)$, and the initial learning rate is set to 1.2e-5. The weight decay is 1e-3. The number of video frames sampled is 5 with the size of 224×224 and are split into patch sizes of 32×32. The batch size used for MSVD, MSRVTT, TGIF, and ActivityNet is 10, 12, 10, and 8 per GPU respectively. For training the model in CVQA, the number of answers used for testing and training is consistent with MSVD 1000, MSRVTT 1500, TGIF 1540, and ActivityNet 1654.

JustAsk [2]. Fine-tuning for the model is implemented for 20 epochs and we use Adam [7] optimizer with a batch size of 256 and validation batch size of 2048. For the learning rate, we utilize the cosine annealing scheduler with an initial value of 1e-5. The video features are equally space sampled and padded up to a maximum of 20. The dimension of the video feature is 1024, the text is 768 and the final embedding is 512. The Dropout [8] probability is set to 0.1. The number of training and test answers for CVQA is MSVD 1852, MSRVTT 4000, TGIF 1540, and ActivityNet 1654.

FrozenBiLM [3]. For each video and text encder, we use $T = 10$ for the number of frames and $N = 256$ for the number of text tokens. Each frame is resized to the size of 224×224 and its feature is extracted by CLIP ViT-L/14 [1, 9]. We use a hidden dimension size of $D = 1536$. Learning rate is set to 5e-5 and linear warm up is applied for the first 10% of total iterations. After the warm up, a learning rate is decayed to 0 for the remaining iterations. We train the model with a batch size of 32 during 20 epochs for all the datasets. Dropout probability is 0.1 and Adam optimizer of $\beta = (0.9, 0.95)$ is adapted with no weight decay.

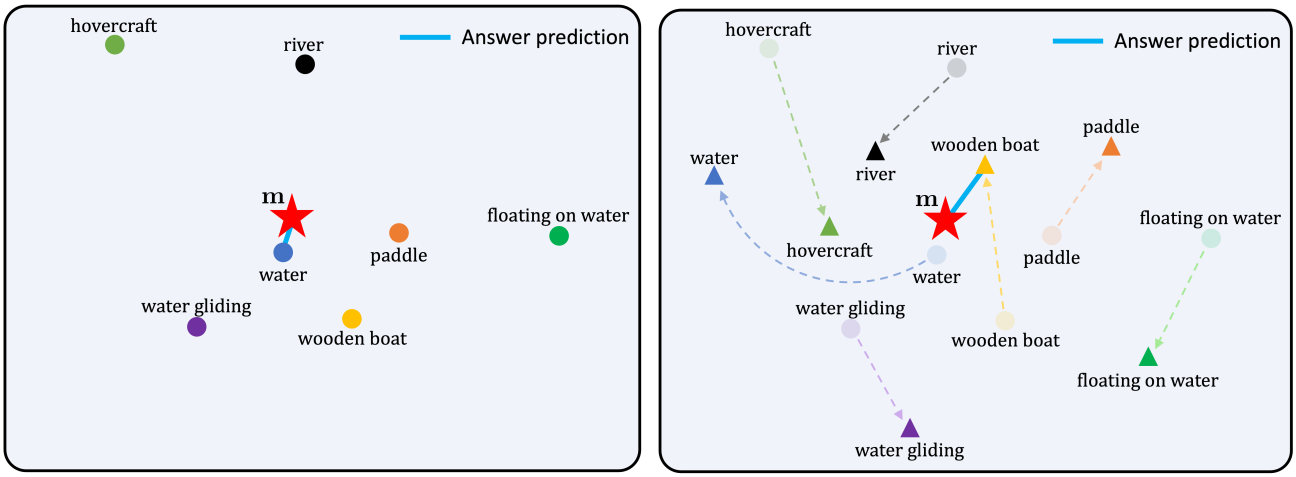
ε	ActivityNet					
	B	C	R	U	T	M
1.0	67.7	37.4	15.5	4.2	43.2	10.4
0.9	**68.7**	37.3	15.2	4.5	43.7	10.7
0.8	67.8	38.6	16.9	4.7	43.8	11.1
0.7	68.2	**39.9**	**18.5**	**5.8**	**44.6**	**11.9**
0.6	68.1	38.7	17.6	5.1	44.1	11.7
0.5	67.5	38.4	16.2	4.9	43.6	11.1
0.4	68.3	37.8	15.6	5.3	43.8	11.1
0.3	68.2	36.8	14.9	5.2	43.4	11.2
0.2	68.2	36.3	13.1	5.1	43.1	10.3
0.1	68.3	35.5	12.5	4.1	42.7	9.3
0.0	66.2	34.9	12.2	4.2	41.6	9.3

Table A3: **Ablation study on ε.**

C. Additional quantitative results

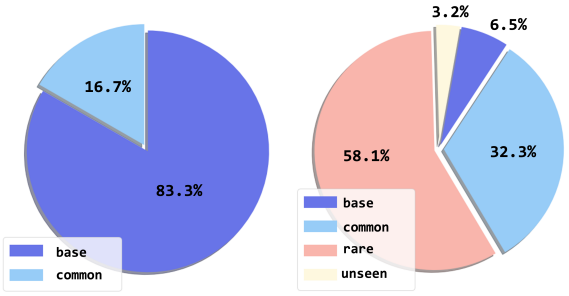
C.1. Zero-shot performance

We compare the zero-shot performances between the standard CVQA baselines and our developed OVQA baselines in Tab. A1. On MSVD, ActivityNet and TGIF, our FrozenBiLM+ outperforms the standard FrozenBiLM by 1.2%, 1.4%, and 0.6% on the total performance (**T**), achieving state-of-the-art results. Also for all the datasets, mAcc (**M**) on both JustAsk+ and FrozenBiLM+ are improved by a



(a) FrozenBiLM+ without GNN-based soft verbalizer (b) FrozenBiLM+ with GNN-based soft verbalizer

Figure A3: **TSNE of answer embeddings before/after adapting GNN-based soft verbalizer.** $\mathbf{m}$ is an output feature of the [MASK] token. The prediction of the model is changed from "water" in (a) to "wooden boat" in (b).



(a) VIOLET (b) VIOLET+

Figure A4: **Proportion of answer categories with an accuracy of 90%.** The portion of answer categories in TGIF that (a) VIOLET and (b) VIOLET+ achieve an accuracy of 90%.

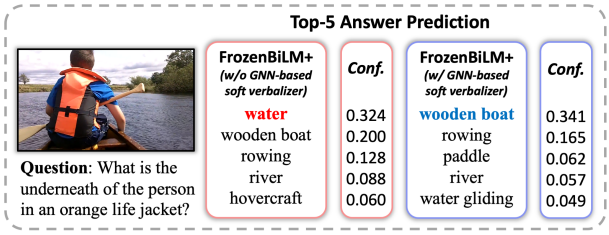


Figure A5: **Confidence scores of the top-5 predictions w/ and w/o GNN-based soft verbalizer on FrozenBiLM+.**

large margin. This implies that considering rare and unseen answers by fully leveraging the generalizability of backbone models pretrained on the large-scale dataset also improves the zero-shot performance.

C.2. Ablation studies

Answer encoder type. We conduct an ablation study on the answer encoder type by comparing CLIP [1] and DeBERTa [10] in Tab. A2. In general, adopting DeBERTa outperforms CLIP by a large margin especially on mAcc (**M**) for all datasets.

Effectiveness of ε. In Tab. A3, we also experiment by adjusting the ε in Eq. (7) of the main paper on FrozenBiLM+. Note that with a wide range of $\varepsilon \in [0.3, 0.9]$, our method equipped with the GNN-based soft verbalizer shows superior performance to the standard FronzeBiLM ($\varepsilon = 1.0$).

D. Additional qualitative results

D.1. Comparison of answer category proportion

We analyze the answers that VIOLET and VIOLET+ correctly predict. Fig. A4 shows the proportion of answer categories that are predicted by VIOLET and VIOLET+ with an accuracy of 90% or higher. VIOLET in Fig. A4a focuses on base and common categories, and the portion of the base category answers is 83.3%. On the other hand, Fig. A4b shows that VIOLET+ accurately predicts the answers in the rare and unseen categories beyond base and common answers. The portion of rare and unseen categories significantly increased. This evidences that the bias of VIOLET toward frequent answers is alleviated in VIOLET+.

Models	MSVD		ActivityNet		TGIF		MSRVTT	
	BNG↓	M↑	BNG↓	M↑	BNG↓	M↑	BNG↓	M↑
All-in-one [4]	41.3	7.9	49.1	5.3	56.0	10.1	42.2	3.9
All-in-one+	**39.3**	**9.4**	**47.3**	**6.8**	**50.6**	**13.2**	**39.9**	**4.7**
VIOLET [6]	70.7	2.7	49.6	3.7	77.9	4.5	54.6	1.4
VIOLET+	**44.2**	**10.7**	**46.1**	**6.1**	**49.2**	**14.3**	**43.9**	**4.5**
JustAsk [2]	38.5	12.6	41.2	8.2	44.9	11.7	38.2	7.0
JustAsk+	**37.2**	**14.5**	**39.5**	**11.5**	**43.5**	**14.4**	**37.8**	**7.6**
FrozenBiLM [3]	37.4	17.2	47.3	7.9	37.8	23.5	40.2	6.7
FrozenBiLM+	**35.0**	**21.3**	**46.6**	**11.9**	**35.0**	**30.2**	**35.7**	**12.2**

Table A4: **Comparison of Base and Non-base performance gap (BNG).**

D.2. Answer embeddings visualization

Fig. A5 illustrates another qualitative example of the model with and without a GNN-based soft verbalizer on FrozenBiLM+. GNN-based soft verbalizer successfully corrects the prediction from "water" to "wooden boat". Also, in Fig. A3, we visualize TSNE of answer embedding changes before/after adapting GNN-based soft verbalizer in the above example. Fig. A3a shows that the model predicts "water", which is the closest answer to $\mathbf{m}$, as the answer without a GNN-based soft verbalizer. On the other hand, in Fig. A3b, GNN-based soft verbalizer effectively updates the answer embeddings by moving the embedding of "wooden boat" close to $\mathbf{m}$, and the prediction is corrected to "wooden boat".

E. A new metric to measure the model bias

We here introduce a new metric, Base and Non-base performance Gap (BNG). BNG evaluates how much the model is biased toward base answers, and is calculated as:

$$\text{BNG}\ (\%) = \text{Base}\ (\%) - \text{Non-base}\ (\%), \qquad \text{(A1)}$$

where Non-base consists of common, rare, and unseen answers. The lower BNG indicates that the model has less bias. In Tab. A4, our developed baselines outperforms previous CVQA baselines by a large margin in terms of BNG as well as mAcc (**M**). Especially, by comparing VIOLET and VIOLET+, the BNG is decreased by 26.5% and 28.7% on MSVD and TGIF respectively, and mAcc (**M**) is also improved by 8% and 9.8%. This implies that the model bias toward frequent answers is effectively alleviated on VIOLET+.

References

[1] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 2, 3

[2] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Just ask: Learning to answer questions from millions of narrated videos. In *ICCV*, 2021. 2, 4

[3] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Zero-shot video question answering via frozen bidirectional language models. In *NeurIPS*, 2022. 2, 4

[4] 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. 1, 4

[5] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 1

[6] 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*, 2021. 2, 4

[7] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 2

[8] Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: a simple way to prevent neural networks from overfitting. *JMLR*, 2014. 2

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 2

[10] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. Deberta: decoding-enhanced bert with disentangled attention. In *ICLR*, 2021. 3