Towards Fast Adaptation of Pretrained Contrastive Models for Multi-channel Video-Language Retrieval
Supplementary Material

Xudong Lin[1] Simran Tiwari[1], Shiyuan Huang[1], Manling Li[2]
Mike Zheng Shou[3], Heng Ji[2], Shih-Fu Chang[1]
[1]Columbia University [2]UIUC [3]National University of Singapore
xudong.lin@columbia.edu

1. Discussion on Reasons Why Text + Text Performs Best

We think the main reason that the proposed Text + Text variant can outperform other variants is that it is hard and resource-intensive to learn the alignment between the text embedding space and the visual feature space, and train a good multimodal transformer even with strong pretrained text models as initialization. There are two lines of evidence we observed: first, as discussed in the next section, these two spaces are significantly different even in terms of statistics, which is verified by the fact that without the trick to initialize the projected visual features to have the same mean and variance with text embeddings, Conti. + Text would suffer from a huge performance drop of 4%; second, from results in the main paper, it is clear that Conti. + Multi. seriously rely on the huge amount of video-text samples to align the embedding space and train the multimodal transformer. The Text + Text variant exactly avoids these two processes in its design.

2. Additional Implementation Details

We use a learning rate of 0.00005, learning rate decay of 0.9 and batch size of 256. We trained our model for 20 epochs on iVQA and MSVD-QA and for 30 epochs on How2QA, ActivityNet-QA and MSRVTT-QA on 2 Nvidia V100 GPUs. On VATEX and YouCook II, the number epochs is 10 and 100, respectively. We find that the Masked Language Modeling objective is not helpful when we use MPNet as the language transformer, and thus, we only use the contrastive loss as described in the main paper. We use a gradient clipping of 1, following [11]. The other hyperparameters were directly borrowed from Just-ask [13].

In the implementation of Continuous Features + Text Transformer, we find it important to initialize the last LayerNorm [1] in the projector with parameters of the LayerNorm after the embedding layer of MPNet so that the projected video features are aligned with the textual embeddings. Without this initialization trick, the model only achieves 19.8% on iVQA (when with the initialization, it achieves 23.2%).

We use a public python library[1] to obtain the automatic speech transcripts from YouTube for the videos we used in iVQA, How2QA, ActivityNet-QA, YouCook II and VATEX. We also used subtitles from [7] for videos with speech but without ASR. Eventually, about 90%, 70%, 30%, 90%, and 50% of the videos have associated speech transcripts, respectively.

We fix the random seed in all the experiments and we do not observe significant change of accuracy ($< 0.5\%$) when changing the random seed. The code will be released at https://github.com/XudongLinthu/upgradable-multimodal-intelligence for any other details and results on additional datasets.

3. Details about Datasets and Evaluation Metric

iVQA [13]. It contains 10,000 instructional videos. Each video is annotated with one question and five corresponding answers. We follow the official split to use 6,000, 2,000, and 2,000 videos for training, validation, and testing, respectively. We follow [13] to calculate accuracy with five annotations per question.

How2QA [6]. The dataset contains 44,007 QA pairs that are annotated from 9,035 videos. We follow [13] to use the train and validation split for training and testing. Note that in this dataset, each question and answer pair are manually annotated with three negative answers so we actually don't need to retrieve from a huge answer set. The metric used for this dataset is accuracy.

ActivityNet-QA [14]. It contains 58,000 QA pairs manu-

[1] https://pypi.org/project/youtube-transcript-api/

Number of Tokens	60k-word	Answer-word
10	25.9	29.8
15	27.3	30.9
20	26.8	31.1
25	26.5	31.6
30	26.2	30.9

Table 1. Accuracy on varying the number of text tokens and 60k-word vocabulary and the answer-word vocabulary on the iVQA dataset

ally annotated from 5,800 videos from the ActivityNet [2] dataset depicting a wide range of complex human activities. The official split of 32,000, 18,000 and 8,000 QA pairs for training, validation and testing respectively is adopted.

YouCook II [15]. It is a instructional video dataset containing 2,000 long videos of 89 recipes. We follow [8] to use the temporal boundary of steps to formulate a pair as a step description and the corresponding video segment. The resulted number of training and testing pairs are 10,387 and 3,411.

VATEX [12]. We only take videos and English captions from it to evaluate retrieval performance. Due to the fact that only 50% of the videos have ASR and many ASRs only contain English stop words, we only keep the videos with at least 5 non-stop words to make sure the task is still multi-channel. The resulted number of training and testing pairs are 36,680 and 4,190. When comparing with HERO [6], we train and evaluate our model under its data split.

4. Experiments on *k* and the Vocabulary.

In this section, we report the results when varying k and the vocabulary for retrieving text tokens, as shown in Table 1. 60k-word vocabulary contains all 65,000 the words that are used in the language model of [8]. The answer-word vocabulary is constructed by collecting all the unique verbs and nouns from parsing all the query/answer sentences in the downstream datasets with spaCy [5].

We observe that when larger than 15 words are retrieved for each segment in the video, the benefit of retrieving more tokens starts to be marginal. Therefore, we use $k = 15$ for all the other datasets as less tokens also help to further accelerate the training process. But we use $k = 25$ for the iVQA dataset as this turns out to be the optimal value when using the answer-word vocabulary on the iVQA dataset. This also indicates that the performance on other downstream datasets could be further improved if we optimize the number of tokens.

We also observe a consistent improvement for all the number of tokens when changing from the 60k-word vocabulary to the answer-word vocabulary. Therefore, we construct the answer-word vocabulary for each dataset separately, which have a size from 3K to 25K words, depending on the dataset. Note that for retrieval datasets, we use the words from all the text queries.

5. Additional Comparison with the State-of-the-art

As shown in Table 2, we provide additional comparison with state-of-the-art on the two retrieval datasets. We observe that our proposed **Text Tokens + Text Transformer** performs slightly better than HERO, which requires a lot more data and computational resources to train the model properly. Note that our model only uses pretrained S3D [8] (on YouCook II) or CLIP [10] (on VATEX) to retrieve text tokens as additional representation of the video but HERO that uses features from both 3D extractors [3, 8] and 2D extractors [10]. Note that we found that on YouCook II, using $K = 25$ helps to improve the results.

AT-ST HERO is annotated as gray because it further leverages about 730K extra **annotated** multimodal samples to perform multi-task training and then it is fine-tuned for specific datasets, which is not directly comparable to our setting. Interestingly, the additional annotated data helps AT-ST HERO to significantly improves the performance on VATEX but the performance on YouCook II is actually even lower. Overall, when without such multi-task learning, our proposed method still achieves comparable results with state-of-the-art on these datasets.

6. Additional Discussion on MSRVTT-QA and MSVD-QA

We list a few question-answer samples from the test set of MSRVTT-QA in Table 3. The questions are not natural sentences and the answers are not informative. To quantify the imperfection of question generation, we carefully convey a manual study on 50 randomly sampled question-answer-video triplets in the MSRVTT-QA dataset. We found that the 6% of the questions are not exactly aligned with the video, e.g., wrong entity descriptions/wrong action descriptions. 24% of the questions have grammatical errors. 10% of the questions are ambiguous. For the answers, we found that 12% of them are either not aligned with the video or too ambiguous to determine correctness. 32% of the answers are not informative enough but they are roughly aligned with the question and the video.

A side evidence is that the Table 7 in [13] compares between the method used to generate MSRVTT-QA and MSVD-QA and the method used in Just-ask for question-answer generation. The resulted model pretrained with generation method of [13] significantly outperforms the model pretrained with data generated from method [4], by a large

Model	Extra MM Samples	Δ GPU hours	YouCook II	VATEX
AT-ST HERO [6, 7]	7.6M + 700K	-	45.3	80.0
HERO [6, 7]	7.6M	8,000	49.5	**63.4**
Text + Text (Ours)	0	0	**50.6**	60.1

Table 2. Comparison with the state-of-the-art on YouCook II and VATEX in terms of accuracy and efficiency. Extra MM Samples indicate the number of video-text samples that are needed in the second-round pretraining. Δ GPU hours refer to the additional computation required for the second-round pretraining. Note that our variant typically requires 1 GPU hour for training. Average recall@$\{1,5,10\}$ (%) is reported. AT-ST HERO is annotated as gray because it further leverages about 730K extra **annotated** multimodal samples to perform multi-task training, which is not directly comparable to our setting. Note that even HERO is not completely comparable with our model, as it enjoys both video-level and frame-level feature extractors.

relative margin of 20% to 1000% of zero-shot accuracy on three manually annotated datasets, which indicates the unsatisfactory quality of MSRVTT-QA and MSVD-QA generated by [4].

Both the qualitative and quantitative results motivate us to not use them as the datasets to assess the performance of the four variants.

Question	Answer
what is a clip doing?	show
who is showing something in a computer?	someone
what can singing?	pain
what explains colors?	video

Table 3. Question-answer samples from the test set of MSRVTT-QA.

7. Additional Visualisation

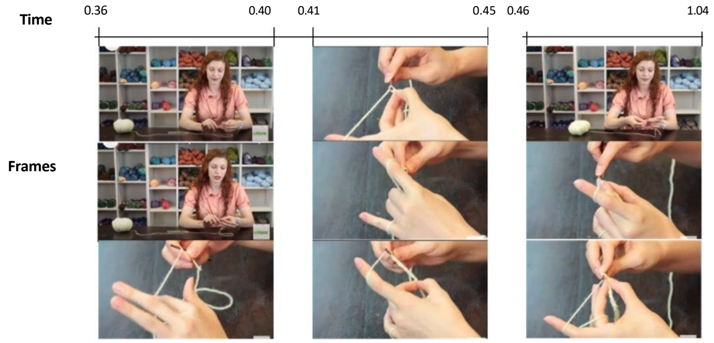
As shown in Figure 1, the answer words or some of the answer words are retrieved as one of the text tokens to describe the video. As discussed in the main paper, we find that 64% of the videos have at least one word overlap between the retrieved text tokens and the answers. This kind of examples show the high explainability of the proposed method as it is clear these important key words are the input to the model to make the final prediction.

8. Discussion on Social Impact, Limitation and Future Work

We first argue that the tasks we handle in this work are potentially helpful for visually-impaired people to better handle daily life as our model can be used to help them understand the ongoing events with text queries.

The datasets we used are mostly based on videos from YouTube. Therefore, they may contain personal information but our algorithm is not designed to specifically leverage certain private information. Overall, we do not expect negative societal impact from the designed algorithms but the dataset we use for training may lead the models to produce biased or undesired results.

One limitation is on the imperfection of the visual words retrieval process. We have already shown using CLIP [10] significantly helps to improve the performance in the main paper. We leave further using better pretrained multimodal contrastive models to future work. We also leave the utilization of recent larger pretrained contrastive text models [9] (with billions of parameters) to future research.



Question	What is the woman wrapping around a needle?
Text Tokens	First. crucifix. earbud. knitting. thread chain. pullover. crochet needle. yarn ball. knitted. designer chain. sweater. crochet hook. kerchief. yarn. crocheting. crochet. Then. knitting. pullover. yarn ball. knitting nettle. knitted. knitting needle. crocheting. designer chain. crochet hook. knitting pin. knit scarf. sweater. knitting stick. knitting needles. woolen thread. Then. earbud. yarn ball. crochet needle. thread chain. knitting. crucifix. knitted. crochet hook. left ring finger. needle. knitting needle. tapestry. index finger. crochet. crocheting.
Predicted Answer	yarn
Ground Truth Answers	yarn, yarn, yarn, yarn, yarn

Figure 1. Visualisation of a successful case for **Text Tokens + Text Transformer** on the iVQA dataset. In this example, the answer “yarn” is retrieved as one of the text tokens to describe the video.

References

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 1

[2] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970, 2015. 2

[3] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019. 2

[4] Michael Heilman and Noah A Smith. Good question! statistical ranking for question generation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 609–617, 2010. 2, 3

[5] Matthew Honnibal and Ines Montani. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*, 7(1):411–420, 2017. 2

[6] 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. 1, 2, 3

[7] Linjie Li, Jie Lei, Zhe Gan, Licheng Yu, Yen-Chun Chen, Rohit Pillai, Yu Cheng, Luowei Zhou, Xin Eric Wang, William Yang Wang, et al. Value: A multi-task benchmark for video-and-language understanding evaluation. *arXiv preprint arXiv:2106.04632*, 2021. 1, 3

[8] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9879–9889, 2020. 2

[9] Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernández Ábrego, Ji Ma, Vincent Y Zhao, Yi Luan, Keith B Hall, Ming-Wei Chang, et al. Large dual encoders are generalizable retrievers. *arXiv preprint arXiv:2112.07899*, 2021. 3

[10] 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 *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2, 3

[11] Nils Reimers and Iryna Gurevych. Sentence Transformers. `https://www.sbert.net/`, 2019. 1

[12] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 2

[13] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Just ask: Learning to answer questions from millions of narrated videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1686–1697, 2021. 1, 2

[14] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9127–9134, 2019. 1

[15] Luowei Zhou, Chenliang Xu, and Jason J Corso. Towards automatic learning of procedures from web instructional videos. In *AAAI Conference on Artificial Intelligence*, pages 7590–7598, 2018. 2