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

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

Abstract

Multi-channel video-language retrieval require models to understand information from different channels (e.g. video+question, video+speech) to correctly link a video with a textual response or query. Fortunately, contrastive multimodal models are shown to be highly effective at aligning entities in images/videos and text, e.g., CLIP [20]; text contrastive models are extensively studied recently for their strong ability of producing discriminative sentence embeddings, e.g., SimCSE [5]. However, there is not a clear way to quickly adapt these two lines to multi-channel video-language retrieval with limited data and resources. In this paper, we identify a principled model design space with two axes: how to represent videos and how to fuse video and text information. Based on categorization of recent methods, we investigate the options of representing videos using continuous feature vectors or discrete text tokens; for the fusion method, we explore the use of a multimodal transformer or a pretrained contrastive text model. We extensively evaluate the four combinations on five video-language datasets. We surprisingly find that discrete text tokens coupled with a pretrained contrastive text model yields the best performance, which can even outperform state-of-the-art on the iVQA and How2QA datasets without additional training on millions of video-text data. Further analysis shows that this is because representing videos as text tokens captures the key visual information and text tokens are naturally aligned with text models that are strong retrievers after the contrastive pretraining process. All the empirical analysis establishes a solid foundation for future research on affordable and upgradable multimodal intelligence.

1. Introduction

From retrieving a trending video on TikTok with natural language descriptions to asking a bot to solve your technical problem with the question and a descriptive video, AI agents handling multi-channel video-language retrieval-style tasks have been increasingly demanded in this post-social-media era. These tasks require the agent to fuse information from multiple channels, i.e., video and text to retrieve a text response or return a multi-channel sample for a text query. To power such agents, a popular approach [9, 10, 15, 34, 40] consists of two rounds of pretraining: (1) The 1st round is to obtain unimodal pretrained models, such as visual-only encoders [3, 6, 17, 20] (e.g., S3D,CLIP) and text-only encoders [4, 13, 22, 24] (e.g., BERT) (2) The 2nd round aims at pretraining on visual-text dataset - specifically, researchers leverage techniques like masked token modeling [9, 40] or contrastive learning [10, 15, 33, 34] to align and fuse unimodal features from model pretrained in the 1st round.

Such methods achieve good performance on multi-channel retrieval-style tasks but they suffer from two major limitations: 1) huge amounts of data and computational resources are required for the second-round "pretraining", which significantly limits the research exploration without such resources; 2) the domain of video data used in the second round "pretraining" has to be strongly correlated with downstream tasks [9], which may restrict such methods from being generally applicable.

To alleviate such limitations, we study a novel problem: fast adaptation of pretrained contrastive models on multi-channel video-language retrieval under limited resources. Specifically, we propose to adapt both contrastive multimodal models [16, 20] and contrastive text models [5, 22] to enjoy their strong encoding ability and discriminative embedding space. There has been tremendous progress recently on large-scale contrastive multimodal models [16, 17, 20]. Through pretraining on millions of images/videos, these models are highly effective at *encoding visual inputs and linking entities across modalities*. Meanwhile, contrastive text models [5, 22] have been also densely studied to obtain *discriminative sentence embeddings*. These models are shown to perform well on challenging text retrieval tasks such as semantic search [19], which requires the model to *understand both language and real-world knowledge* [5, 22]. Such an ability allows the model to retrieve a text response or encode a text query in multi-channel video-language retrieval-style tasks. Thereby, once the video information is effectively incorporated, it eliminates the necessity of second-round "pretraining" on large scale multimodal datasets, enabling fast adaptation to any video-language retrieval-style tasks.

We first conduct a systematic analysis on potential model designs, as shown in Figure 1. We identify the model design space with two design principles: how to represent the video information and how to fuse this video information with questions or other text such as speech. To represent video information, we could either adopt **Continuous Features** which is commonly used in existing work [18, 34], or project videos into unified **Text Tokens** [11] from various modalities. To fuse information from multiple channels, i.e., video and question/speech, there are two potential options, namely, a **Multimodal Transformer** [34] or a **Text Transformer** [18]. Hence, there are four combinations derived from this model design space, namely, **Continuous Features + Multimodal Transformer** [34], **Continuous Features + Text Transformer**, **Text Tokens + Multimodal Transformer**, **Text Tokens + Text Transformer**.

Our exploration of this model design space results into a simple yet effective approach that allows fast adaptation of pretrained contrastive models, which first leverages contrastive multimodal models to retrieve a sequence of **Text Tokens** for the visual input and then feeds these tokens together with other text to contrastive **Text Transformer** for answer retrieval. Its fast adaptation ability not only comes from the ability of linking entities across modalities of the contrastive multimodal model, but it also enjoys the natural alignment with contrastive text models to produce discriminative embeddings. To the best of our knowledge, this is the first proposal to adapt pretrained contrastive models in this manner, although each individual design choice may have been adopted previously. We further conduct in-depth analysis to understand 1) the trade-off between data efficiency and accuracy, 2) the impact of pretrained contrastive text model, and 3) the possible limitation of this framework.

The contribution could be summarized three-fold:

- We identified the principled model design space for fast adaption of pretrained contrastive multimodal models and pretrained contrastive text models.
- We conducted extensive experiments on five video-language datasets, observed a consistent trend across these four variants, and even obtained state-of-the-art performance (e.g., **6.5%** improvement on How2QA) with the proposed Text Tokens + Text Transformer variant **without using millions of extra multimodal data samples**, which is essential to democratize the community of video+language.
- The proposed Text Tokens + Text Transformer variant scales significantly better than the other variants, w.r.t. the quality of pretrained text models. The code will be released at `https://github.com/XudongLinthu/upgradable-multimodal-intelligence` to facilitate future research.

2. Related Work

Pretrained Contrastive Models. In this paper, we mainly consider two types of contrastive models: 1). contrastive multimodal models [10, 12, 16, 20, 29], which typically consist of a visual encoder and a text encoder, and learn to map visual and text embeddings into a common space. They sample positive/negative pairs from aligned/unaligned image/video and text, and train the visual and text encoders with a contrastive objective in a self-supervised manner. With access to large-scale multimodal data (e.g., 400 million web image-text pairs [20]), they are shown superior in linking entities across modalities; 2) Unlike general language models [4, 21], contrastive text models [5, 22], which learn discriminative sentence embeddings from large-scale text data (e.g., 2 billion sentence pairs [23]). They construct positive and negative sentence pairs and train the text encoder in a either supervised or unsupervised manner.

Multi-channel Vision-Language Learning. Interests have been raised in multi-channel VL applications, where the model is asked to understand from both visual and language inputs to output a response (e.g., a phrase). Several works seek to get the response in a generative way [8, 11, 41], i.e., using a generative model to predict a sequence. On the other hand, retrieval-style [1, 2, 14, 37] accomplishes the tasks in a discriminative way, where discriminative embeddings are produced to select the best answer from an answer candidate pool. In the paper, we focus on retrieval-style tasks, e.g., video question answering [9, 32, 34, 39] and multi-channel text-to-video retrieval [31, 42].

Adapting Pretrained Models to VL Applications. Existing works [18, 27, 34, 35] explored different ways to adapt pretrained models to various vision-language tasks. For example, Shen *et al.* [25] explored directly taking visual features from pretrained CLIP [20] image encoders and feed them to existing vision-language models [27] and then retrain the whole model using the new CLIP visual features. This empirical analysis shows that visual features learned by contrastive multimodal models are better for vision-language tasks. Just-ask [34] is a recent method for open-ended video question answering. The authors found the multimodal transformer gets significantly improved after a second-round pretraining on 69M video-question-answer triplets. All these methods require second-round pretraining on another large task-specific dataset, which requires

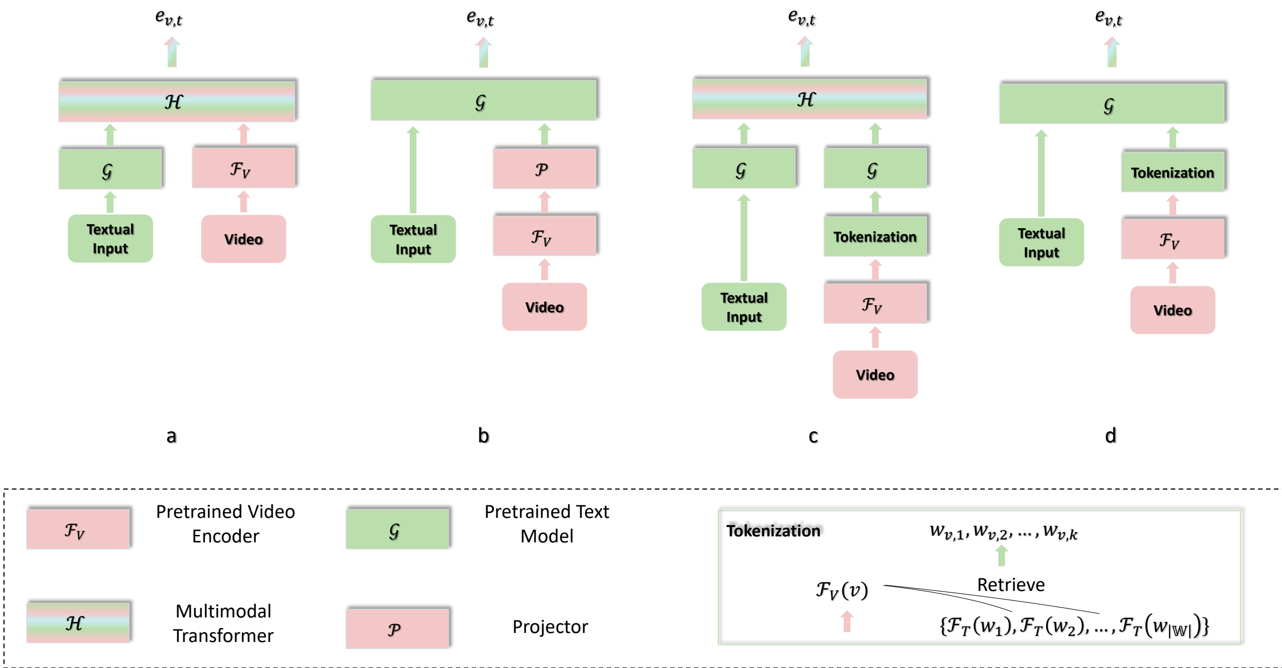


Figure 1. a. **Continuous Features + Multimodal Transformer**. This variant uses the pretrained video encoder to represent input video segment as continuous feature vectors. A randomly initialized multimodal transformer is used to fuse information of the video and textual input. b. **Continuous Features + Text Transformer**. Instead of a multimodal transformer, we use the pretained contrastive text model to fuse projected video features and textual input. c. **Text Tokens + Multimodal Transformer**. It uses the pretrained video-text encoders to retrieve top-k text tokens from a predefined vocabulary to describe the video content. Then we feed the text tokens separately to the pretrained text model, to obtain contextualized embeddings for a multimodal transformer to fuse with contextualized text embeddings of other textual input. d. **Text Tokens + Text Transformer**. We further remove the multimodal transformer and directly use the pretrained text model to fuse video and question.

additional engineering process and high demand of computational resources; on the other hand, we eliminate this process by efficiently leveraging the pretrained contrastive text models.

3. Technical Approach

In this section, we first include some preliminaries, and then introduce the four models in the model design space in detail. In the end, we describe the training technique and the implementation details.

3.1. Preliminaries

Multi-channel Video-language Retrieval-style Tasks. These tasks require the model to fuse information from multiple channels, i.e., video and text to retrieve a text response or return a multi-channel sample for a text query. Specifically, we mainly consider open-ended video question answering and multi-channel text-video retrieval. In open-ended video question answering, given a video v and question t as input, the model is required to retrieve the correct answer a_k from a large answer corpus $\mathbb{A} = \{a_1, .., a_{|\mathbb{A}|}\}$. In multi-channel text-video retrieval, given a text query a, the model is required to retrieve the most relevant video v_i with associated speech text t_i, which are from a corpus of multi-channel videos $\mathbb{V} = \{v_1, t_1, .., v_{|\mathbb{V}|}, t_{|\mathbb{V}|}\}$. For simplicity, in the model descriptions, we will adopt open-ended video question answering for illustration.

Pretrained Contrastive Multimodal Models. We mainly leverage pretrained video-text contrastive models. It consists of a video encoder $\mathcal{F}_V : \mathbb{R}^{H\times W\times 3\times F} \longrightarrow \mathbb{R}^D$ and a text encoder $\mathcal{F}_T : \mathbb{W}^L \longrightarrow \mathbb{R}^D$, where H, W, F are the height, width and number of frames of the video, L is the length of the sentence, D is the dimension of the common embedding space and $\mathbb{W}$ is the set of all the words. Note that in all of the models, the pretrained contrastive multimodal models are **frozen** for fast adaptation.

Pretrained Contrastive Text Models. The idea is to train a text encoder $\mathcal{G} : \mathbb{W}^L \longrightarrow \mathbb{R}^D$ to push the output embeddings of two relevant sentences to be similar and the irrelevant to be dissimilar. The pretrained contrastive text models will be **updated** in all of the models. Since all the actual instances of text models are transformer-based, text models and text transformers are used interchangeably.

Multimodal Transformer and Projector. It is a shallow transformer model [28] $\mathcal{H}:\mathbb{R}^{L\times D}\longrightarrow\mathbb{R}^{L\times D}$. We direct the audience for details of the transformer model to [28]. Inspired by prefix tuning [18], we use the same transformer architecture for the projector $\mathcal{P}:\mathbb{R}^{L\times D}\longrightarrow\mathbb{R}^{L\times D}$, which is used to project video feature vectors to the input space of the pretrained text model.

3.2. Model Variants in the Design Space

For all the four variants, there is a separate encoder $\mathcal{G}_A$ to encode answers, which is initialized from pretrained text models. In the following, we mainly focus on the difference of these four variants, i.e., the video representation and the multimodal fusion design.

a. Continuous Features + Multimodal Transformer

As shown in Figure 1, in this variant, we first directly use the pretrained video encoder $\mathcal{F}_V$ to represent input video segment v as continuous feature vectors. The pretrained contrastive text model $\mathcal{G}$ is used to extract contextualized text embeddings from the question t. A randomly initialized multimodal transformer $\mathcal{H}$ is used to fuse information of the video and the question. Overall the model could be expressed as follows:

$$e_{v,t}=\mathcal{H}(\mathcal{G}(t),\mathcal{F}_V(v)), \tag{1}$$

where $\mathcal{G}(t)$ are actually concatenated with $\mathcal{F}_V(v)$ along the length axis before they are fed into the multimodal transformer.

b. Continuous Features + Text Transformer

Inspired by prefix tuning work [18], in this variant, we leverage a transformer projector $\mathcal{P}$ to project the continuous features from the video encoder $\mathcal{F}_V$ into the input space of the text transformer. Then the text transformer will fuse the information from both the video v and the question t,

$$e_{v,t}=\mathcal{G}(t,\mathcal{P}(\mathcal{F}_V(v))), \tag{2}$$

where the question t actually first passes the embedding layer of $\mathcal{G}$ and then is concatenated with projected video features $\mathcal{P}(\mathcal{F}_V(v))$.

c. Text Tokens + Multimodal Transformer

We propose the Text Token retrieval process as follows. First, we construct the word vectors from a predefined word vocabulary $\mathbb{W}$; specifically, for each word w_i, we use the text encoder to obtain its word vector $\mathcal{F}_T(w_i)$. For the input video v, we encode it with the pretrained video encoder $\mathcal{F}_V(v)$. Then we compare the similarity between the video and each of the words to retrieve k most similar words,

$$w_{v,1},..,w_{v,k}=\arg\max_i^k \mathcal{F}_T(w_i)^\top\mathcal{F}_V(v). \tag{3}$$

Then the retrieved words are fed into the pretrained text model with the question in parallel to obtain the contextualized embeddings, which are finally concatenated to be fed into the multimodal transformer,

$$e_{v,t}=\mathcal{H}(\mathcal{G}(t),\mathcal{G}(w_{v,1},..,w_{v,k})). \tag{4}$$

Note that to maintain similar parameter usage, we share the weights between the text model $\mathcal{G}$ taking question as input and the text model $\mathcal{G}$ taking video words as input.

d. Text Tokens + Text Transformer

In this variant, we follow the same text token retrieval process and then further simplify the model to use pretrained text model for modality fusion. Formally, it can be described as,

$$e_{v,t}=\mathcal{G}(t,w_{v,1},..,w_{v,k}), \tag{5}$$

where the question t and all the retrieved words are concatenated as a whole text sequence and then input to the pretrained text model. The removal of the multimodal transformer enables this variant to directly benefit from the pretrained discriminative sentence embedding space of $\mathcal{G}$.

3.3. Training

During training, we leverage the NCE loss [34] to optimize the parameters θ of $\mathcal{G}$, $\mathcal{G}_\mathcal{A}$ $\mathcal{H}$, and $\mathcal{P}$ (if included in the model). In the actual training process, we calculate loss in a mini-batch manner. For simplicity, we omit average over a batch in the following equations.

For open-ended video question answering, we optimize the following objective

$$\min_\theta -\log\frac{\exp\left(e_{v,t}^\top\mathcal{G}_A(a_l)\right)}{\sum_i\exp\left(e_{v,t}^\top\mathcal{G}_A(a_i)\right)}, \tag{6}$$

where l is the ground-truth answer index.

For multi-channel text-video retrieval, we adopt the symmetric version of the loss [20], Specifically,

$$\begin{aligned}\min_\theta -\frac{1}{2}(\log&\frac{\exp\left(e_{v,t}^\top\mathcal{G}_A(a_l)\right)}{\sum_i\exp\left(e_{v,t}^\top\mathcal{G}_A(a_i)\right)}\\ &+\log\frac{\exp\left(e_{v,t}^\top\mathcal{G}_A(a_l)\right)}{\sum_i\exp\left(e_{v_i,t_i}^\top\mathcal{G}_A(a_l)\right)}),\end{aligned} \tag{7}$$

where $e_{v,t}$ and a_l are a pair of embeddings of multi-channel video and text query, and e_{v_i,t_i} is the representation of the i-th multi-channel video in the mini-batch. Note that although $\mathcal{G}_\mathcal{A}$ is also initialized from the pretrained text model, it does not share parameters with $\mathcal{G}$ during training, following the setting of [34].

Model	Open-ended Acc(%)		Multi-Choice Acc(%)	Retrieval AveR (%)	
	iVQA	ActivityNet-QA	How2QA	YouCook II	VATEX
Conti. + Multi.	22.4	36.9	79.2	41.9	69.4
Conti. + Text	23.2	37.3	80.4	46.2	72.7
Text + Multi.	23.4	37.1	79.4	40.4	67.5
Text + Text	**31.6**	**38.7**	**82.9**	**49.7**	**74.8**

Table 1. Comparison of the four model variants on various multi-channel video-text retrieval-style tasks. Acc is short for accuracy. AveR is short for average recall@{1,5,10}.

3.4. Implementation Details

We use MPNet (all-mpnet-base-v2) [23, 26] as $\mathcal{G}$ when not specified. It is ranked first by Sentence Transformers [23] at the time of writing (November, 2022). We use the pretrained models from [16] for $\mathcal{F}_T$ and $\mathcal{F}_V$ when not specified. When using CLIP [20], we follow FrozenBiLM [36]. The resulted video feature vector is at 1 vector per 1.5 seconds. For $\mathcal{H}$ and $\mathcal{P}$, we use the same shallow multimodal transformer as in [34], which has 2 transformer blocks with 512/768 as the embedding dimension. We follow the training procedure and hyper-parameter settings for all the variants to conduct fair comparisons. Details are provided in the supplementary material[1].

We find that the vocabulary $\mathbb{W}$ can be effectively constructed by parsing all the query/answer sentences in the downstream datasets with spaCy [7] to obtain a set of unique verbs and nouns. We use $k = 15$ words retrieved from the vocabulary for each feature vector in the video. We follow [34] to sub-sample video feature vectors when the video is too long to fit the memory. We further apply a max pooling with kernel size as 5 to sub-sample the retrieved words to avoid too many repetitions of words of neighbouring segmenting.

4. Experimental Results

In this section, we will first introduce the datasets and evaluation metric, then we organize the following subsections by answering a set of important questions of the four model variants. Extra ablations are provided in the supplementary material for hyper-parameter selection.

4.1. Dataset and Evaluation Metric

We select 5 commonly used multi-channel retrieval-style datasets: **iVQA [34]**, **How2QA [9]**, **ActivityNet-QA [39]**, **YouCook II [43]**, and **VATEX [31]** as the main evaluation datasets. For the first three datasets, accuracy is used as the evaluation metric and for the other two, an average of recall@{1,5,10} is used. Details are provided in the supplementary material.

MSRVTT-QA and MSVD-QA [32]. These two datasets are automatically generated from video-caption pairs. The data is too noisy compared to the aforementioned manually annotated datasets but this actually could serve as a testbed to see under what circumstances our variants won't work well. We follow [34] for experimental settings and evaluation metric on these two datasets.

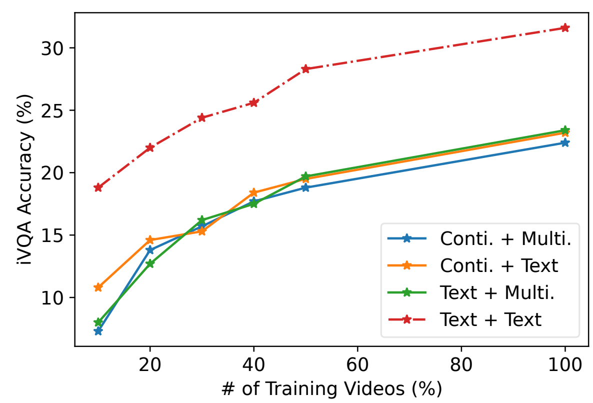


Figure 2. Comparison of the four variants under few-shot setting on the iVQA dataset.

4.2. Which of the variants performs the best?

As shown in Table 1, the **Text Tokens + Text Transformer** consistently performs better than the other three variants when we directly tune these models without second-round large-scale multimodal "pretraining". We suppose this is because of this variant can easily benefit from both pretrained contrastive multimodal models and contrastive text models without struggling to align the space of them or training a multimodal transformer from scratch. We also observe that on downstream task from similar domains with that of the pretrained contrastive multimodal model [16], our proposed variant usually enjoys higher improvement, e.g., on iVQA and YouCook II.

To further understand the behavior of these four variants with limited training data for downstream adaptation, we explore the few-shot setting. We sample a subset of iVQA and train the four variants with the same sub-sampled set for the same number of iterations with that of the full-shot setting. In Figure 2, we consistently observe a large margin between **Text Tokens + Text Transformer** and the other

[1] https://arxiv.org/abs/2206.02082

Model	$\mathcal{F}_V, \mathcal{F}_T$	Extra MM Samples	Δ GPU hours	iVQA	ActivityNet	How2QA
MERLOT [40]	-	180M	-	-	41.4	-
SiaSamRea [38]	-	5.6M + 80K	-	-	39.8	84.1
VQA-T [35]	S3D [16]	69M + 3M	350 + 30	35.2	**39.0**	**85.3**
Conti. + Multi.	S3D [16]	69M	400	35.4	38.9	84.4
Conti. + Multi. (+ *ASR*)	S3D [16]	69M	400	36.0	38.9	84.8
Text + Text (Ours)	S3D [16]	0	0	31.6	38.7	82.9
Text + Text (+ *ASR*, Ours)	S3D [16]	0	0	**36.8**	38.8	84.6
FrozenBiLM [36]	CLIP [20]	10M	160	39.7	43.2	81.5
FrozenBiLM [36] (+ *ASR*)	CLIP [20]	10M	160	39.6	**43.2**	86.7
Text + Text (Ours)	CLIP [20]	0	0	36.9	41.4	92.4
Text + Text (+ *ASR*, Ours)	CLIP [20]	0	0	**40.2**	41.4	**93.2**

Table 2. Comparison with the state-of-the-art on iVQA, ActivityNet and How2VQA 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. + *ASR* indicates the use of ASR texts of the video as additional inputs. Note that our variant typically requires 0.5 GPU hours for training. Methods in gray enjoy a more costly end-to-end training process.

variants. Continuous Features + Multimodal Transformer generally performs the worst on the iVQA dataset and also other datasets, which implicitly verifies our hypothesis that it is crucial to directly leverage the aligned representation space and the discriminative output space.

4.3. Comparison with state-of-the-art on accuracy-efficiency trade-off

We aim to examine the trade-off between downstream performance and the additional resource requirements, through comparison between our best variant with state-of-the-art models that leverage huge amounts of data and computational resources for pretraining the multimodal transformer [35, 36]. As shown in Table 2, with **three orders of magnitude smaller amount GPU training time**, our model can actually achieve comparable or even better results than the state-of-the-art. When using pretrained $\mathcal{F}_V, \mathcal{F}_T$ from [16], our model can even outperform the state-of-the-art by 0.8% when with ASR as input.

When switching the $\mathcal{F}_V, \mathcal{F}_T$ to CLIP [20], our **Text Tokens + Text Transformer** even significantly outperforms than models that enjoys a more costly end-to-end second-round "pretraining" process like MERLOT [40] or SiaSamRea [38]. Most encouragingly, our model also outperforms the state-of-the-art FrozenBiLM on How2QA (**+6.5%**) and iVQA (**+0.5%**), which uses 10 million samples for second-round "pretraining" and uses a much larger text model.

We also would like to highlight that our **Text Tokens + Text Transformer** *easily benefits from a better pretrained* $\mathcal{F}_V, \mathcal{F}_T$ as we observe significant improvement for our model, which is encouraging for future development of our proposed model when better pretrained multimodal contrastive models are available. Comparisons on the other two datasets are separately provided in the supplementary material as the compared methods in Table 2 are not state-of-the-art on them.

4.4. How well does Text Tokens + Text Transformer benefit from pretrained text models?

To comprehensively understand the effect of pretrained contrastive text models, we investigate two sources of models, as shown in Figure 3 and Figure 4. SBERT [23] is a widely-used library with models trained on paired sentences. We measure the quality of pretrained text models with the average performance over 20 sentence embedding/semantic search tasks. SimCSE [5] explores unsupervised contrastive learning of transformer models and provides various models. We measure the quality of pretrained text models with the average performance over 7 semantic textual similarity (STS) tasks.

In Figure 3, we first study the four variants trained on the same data with different model sizes. MiniLM [30] (paraphrase-MiniLM [23]) with $\{3, 6, 12\}$ layers is a suitable testbed as the model architecture is identical but the number of layers changes. We obtain a clear positive correlation between the quality of text models and the accuracy on iVQA. We then study the same model trained with different size of data: paraphrase-mpnet-base-v2, multi-qa-mpnet-base-dot-v1, and all-mpnet-base-v2, which are trained on 83 million, 215 million, and 2 billion sentence pairs, respectively. Again our proposed variant scales well with the performance of the text models.

The scalability w.r.t. the quality of pretrained text model of the proposed **Text Tokens + Text Transformer** is surprisingly superior compared the other variants. As shown in Figure 3, among all the four variants, the proposed **Text Tokens + Text Transformer** has the most positive correlation between the multimodal task performance on iVQA and the quality of the pretrained text models. We also observe that both models using pretrained text transformer for mul-

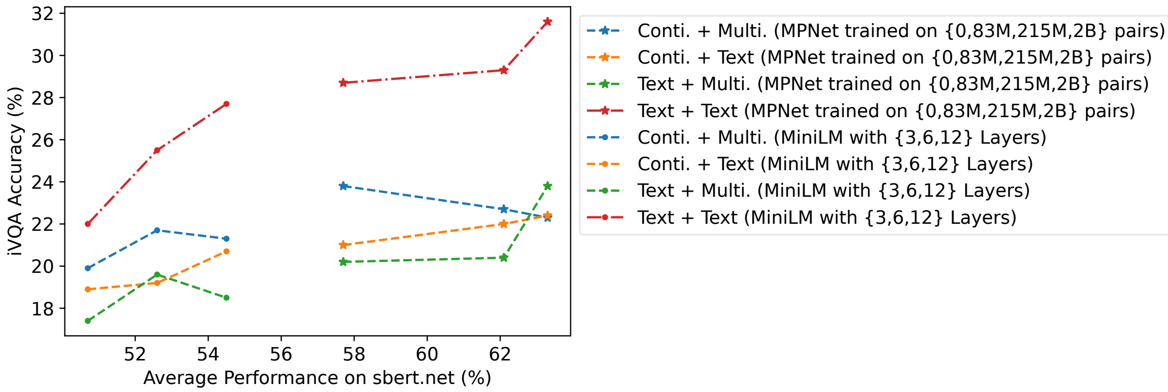


Figure 3. Results of varying the pretrained text model in the four variants on the iVQA dataset. The performance of all SBERT models are obtained from [23].

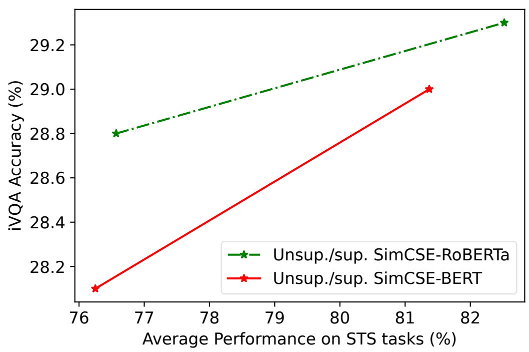


Figure 4. Results of varying the pretrained text model in the **Text Tokens + Text Transformer** on the iVQA dataset. The semantic textual similarity (STS) performance of SimCSE models is obtained from [5].

timodal fusion have a positive correlation and both models using multimodal transformer for multimodal fusion don't benefit from the improvement of the pretrained text models. From this comparison, we conclude that only the proposed **Text Tokens + Text Transformer** can well enjoy the improvement of pretrained text models.

We also verify the necessity of using pretrained contrastive text models by directly tuning the MPNet from [26] without contrastive pretraining. It is not shown in Figure 3 as its text performance is not evaluated by sbert.net. This model only achieves 22.0%, which is significantly lower than the three contrastively pretrained versions.

Then we compare both the unsupervised and supervised BERT/RoBERTa models from SimCSE. As shown in Figure 4, as STS performance increases, the overall downstream performance on iVQA is improved, despite incoherence across different model architectures. As a summary, even with various model architectures, training methods, supervision sources and dataset sizes, we consistently observe positive correlation, which is highly encouraging for further upgradability of the proposed **Text Tokens + Text Transformer** in the future.

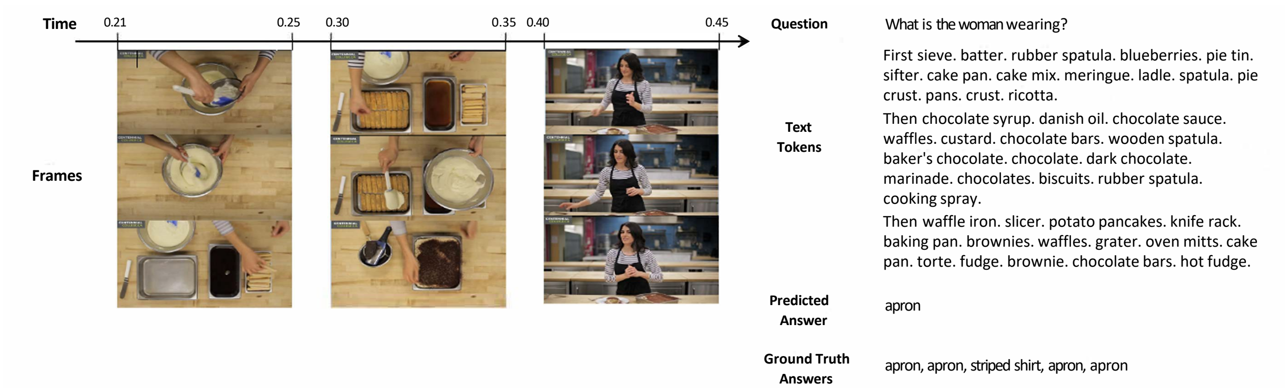
4.5. Model Interpretation

To understand why our proposed **Text Tokens + Text Transformer** performs well, we select iVQA to first check a simple statistic: the proportion of test samples that have at least one word overlapped between the answer and retrieved text tokens of the video. We find that the proportion is actually 66.4%, which partially explains why we obtain a huge improvement on the iVQA dataset.

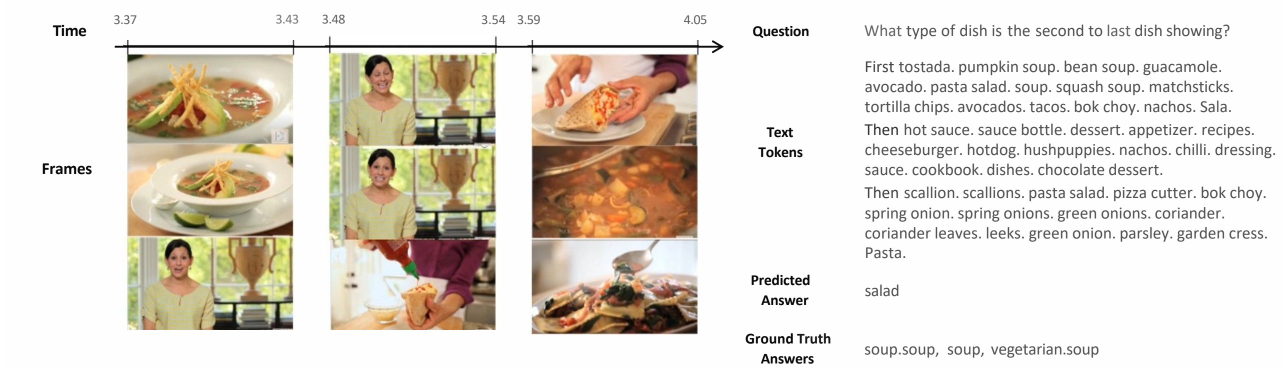
We visualize one successful case and one failure case in Figure 5. In Figure 5a, the model is not able to retrieve the answer word "apron" for the video but based on the rich kitchen-related context words, the text-transformer can still answer the question correctly. This indicates that our proposed model can accommodate imperfection of the tokenization process. In Figure 5b, although "soup" is retrieved as a text token for the video in the first few segments of the video, the crucial video segment required fine-grained temporal understanding of the video in the last segment and the text tokens after max pooling is dominated by other food-related words. But overall, according the aforementioned statistic, this Text Tokens + Text Transformer approach certainly enjoys the high explainability in its design, compared with other using Continuous Features.

4.6. Results on Automatically Generated Datasets

We evaluate the proposed **Text Tokens + Text Transformer** on the automatically generated MSRVTT-QA and MSVD-QA datasets. We find that Text Tokens + Text Transformer (39.7%, 40.9%) performs similarly with the Continuous Features + Multimodal Transformer baseline (39.6%, 41.2%). To quantify the imperfection of question generation, we carefully convey a manual study on



(a) Successful case.



(b) Failure case

Figure 5. Visualization of **Text Tokens + Text Transformer** on iVQA. "First" and "then" indicate the temporal order and they have minimal effect on performance ($< 0.5\%$).

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. These results indicate that when the text queries have very low-quality, the well-trained text embedding space may suffer from the style change of the input language. More discussion is included in the supplementary.

5. Conclusion

We aim at a novel and challenging problem: fast adaption of pretrained contrastive multimodal models and pretrained contrastive text models for multi-channel video-language retrieval under limited data and computation resources. We systematically evaluate four variants of models from our identified principled model design space. Through extensive analysis, we find the variant with text tokens as video representations and contrastive text model for multimodal fusion achieve the best performance. Without millions of multimodal data for pretraining and three orders of magnitude more training time, our model achieves comparable performance with the state-of-the-art model. We further show that this simple yet effective variant can be easily improved by using better pretrained contrastive multimodal models or pretrained contrastive text models, which uncovers the great potential of our proposed model to democratize the video-language research community from heavy dependence on huge data and computation resources towards upgradable multimodal intelligence.

6. Acknowledgement

We thank the anonymous reviewers, Richard Zemel, other DVMMers and colleagues for their feedback. This paper is based upon work supported by U.S. DARPA KAIROS Program No. FA8750-19-2-1004. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of DARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein. Mike Zheng Shou does not receive any funding for this work.

References

[1] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6077–6086, 2018. 2

[2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 2

[3] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017. 1

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1, 2

[5] Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*, 2021. 1, 2, 6, 7

[6] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1

[7] 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. 5

[8] Hung Le, Doyen Sahoo, Nancy Chen, and Steven Hoi. Multimodal transformer networks for end-to-end video-grounded dialogue systems. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5612–5623, 2019. 2

[9] 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, 5

[10] Manling Li, Ruochen Xu, Shuohang Wang, Luowei Zhou, Xudong Lin, Chenguang Zhu, Michael Zeng, Heng Ji, and Shih-Fu Chang. Clip-event: Connecting text and images with event structures. *arXiv preprint arXiv:2201.05078*, 2022. 1, 2

[11] Xudong Lin, Gedas Bertasius, Jue Wang, Shih-Fu Chang, Devi Parikh, and Lorenzo Torresani. Vx2text: End-to-end learning of video-based text generation from multimodal inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7005–7015, 2021. 2

[12] Xudong Lin, Fabio Petroni, Gedas Bertasius, Marcus Rohrbach, Shih-Fu Chang, and Lorenzo Torresani. Learning to recognize procedural activities with distant supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13853–13863, 2022. 2

[13] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. 1

[14] Jiasen Lu, Jianwei Yang, Dhruv Batra, and Devi Parikh. Hierarchical question-image co-attention for visual question answering. In *Advances in neural information processing systems*, pages 289–297, 2016. 2

[15] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval. *arXiv preprint arXiv:2104.08860*, 2021. 1

[16] 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. 1, 2, 5, 6

[17] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2630–2640, 2019. 1

[18] Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021. 2, 4

[19] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. Ms marco: A human generated machine reading comprehension dataset. In *CoCo@ NIPS*, 2016. 1

[20] 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. 1, 2, 4, 5, 6

[21] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*, 2019. 2

[22] Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. 1, 2

[23] Nils Reimers and Iryna Gurevych. Sentence Transformers. https://www.sbert.net/, 2019. 2, 5, 6, 7

[24] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019. 1

[25] Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. How much can clip benefit vision-and-language tasks? *arXiv preprint arXiv:2107.06383*, 2021. 2

[26] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language

understanding. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 16857–16867. Curran Associates, Inc., 2020. 5, 7

[27] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019. 2

[28] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 4

[29] Jinpeng Wang, Yixiao Ge, Guanyu Cai, Rui Yan, Xudong Lin, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Object-aware video-language pre-training for retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3313–3322, 2022. 2

[30] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems*, 33:5776–5788, 2020. 6

[31] 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, 5

[32] 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. 2, 5

[33] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084*, 2021. 1

[34] 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, 4, 5

[35] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Learning to answer visual questions from web videos. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (01):1–1, 2022. 2, 6

[36] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Zero-shot video question answering via frozen bidirectional language models. *arXiv preprint arXiv:2206.08155*, 2022. 5, 6

[37] Zekun Yang, Noa Garcia, Chenhui Chu, Mayu Otani, Yuta Nakashima, and Haruo Takemura. Bert representations for video question answering. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 1556–1565, 2020. 2

[38] Weijiang Yu, Haoteng Zheng, Mengfei Li, Lei Ji, Lijun Wu, Nong Xiao, and Nan Duan. Learning from inside: Self-driven siamese sampling and reasoning for video question answering. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 26462–26474. Curran Associates, Inc., 2021. 6

[39] 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. 2, 5

[40] Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. Merlot: Multimodal neural script knowledge models. *Advances in Neural Information Processing Systems*, 34, 2021. 1, 6

[41] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. *arXiv preprint arXiv:1909.11059*, 2019. 2

[42] Luowei Zhou, Chenliang Xu, and Jason J Corso. Towards automatic learning of procedures from web instructional videos. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018. 2

[43] 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. 5