Scalable and Accurate Self-supervised Multimodal Representation Learning without Aligned Video and Text Data

Vladislav Lialin[1*] Stephen Rawls [2] David Chan [2,3]
Shalini Ghosh [2] Anna Rumshisky [1,2] Wael Hamza [2]
[1] UMass Lowell [2] Amazon [3] UC Berkeley

Abstract

Scaling up weakly-supervised datasets has shown to be highly effective in the image-text domain and has contributed to most of the recent state-of-the-art computer vision and multimodal neural networks. However, existing large-scale video-text datasets and mining techniques suffer from several limitations, such as the scarcity of aligned data, the lack of diversity in the data, and the difficulty of collecting aligned data. Currently popular video-text data mining approach via automatic speech recognition (ASR) used in HowTo100M provides low-quality captions that often do not refer to the video content. Other mining approaches do not provide proper language descriptions (video tags) and are biased toward short clips (alt text).

In this work, we show how recent advances in image captioning allow us to pre-train high-quality video models without any parallel video-text data. We pre-train several video captioning models that are based on an OPT language model and a TimeSformer visual backbone. We fine-tune these networks on several video captioning datasets. First, we demonstrate that image captioning pseudolabels work better for pre-training than the existing HowTo100M ASR captions. Second, we show that pre-training on both images and videos produces a significantly better network (+4 CIDER on MSR-VTT) than pre-training on a single modality. Our methods are complementary to the existing pre-training or data mining approaches and can be used in a variety of settings. Given the efficacy of the pseudolabeling method, we are planning to publicly release the generated captions.

1. Introduction

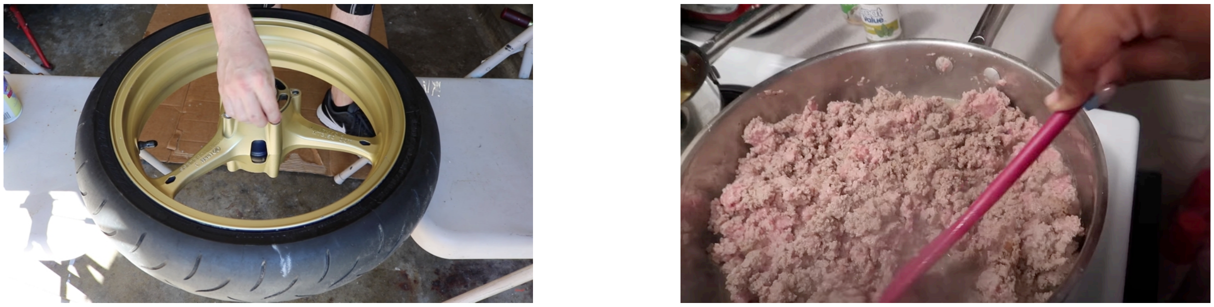
Large language models have revolutionized natural language processing [54, 6, 13] and are rapidly affecting adjacent fields such as computer vision [52, 27, 66, 67, 68]. For example, using **only** weakly-supervised image-text data CLIP [52] and CoCa [68] outperform ResNets [19] on ImageNet. Recent works also demonstrate that the flexibility of the language modeling approach allows us to apply it to any modality [2, 44]. Nevertheless, to pre-train such models, we need enormous amounts of aligned data, which is not yet easily available for all modalities. This holds true for most of the pre-training methods: either contrastive [52], discriminative [38, 43] or generative [11, 2, 37, 65]. Web-mining proved to be an invaluable source of image-caption pairs [58, 9, 55] due to its scalability, but the video domain still suffers from the scarcity of aligned video-text data.

While video data is abundant on the internet, it is hard to utilize it for pre-training. Existing large-scale video classification datasets like Kinetics [7] and YouTube8M [1] consist of 500K+ video clips. Still, they only provide class labels and can not be used for generative modeling. Mining videos with the alt text HTML attribute that provides a short description of the media content [4, 39] is a promising direction that has been immensely successful in the image captioning domain [52, 58, 9, 16, 55]. However, motivated by the VirTex finding [15] that dense image annotations (captions) work better for pre-training than sparse annotations (class labels), we speculate that video alt text does not describe long videos with enough detail and is not well-suited for pre-training.

Nagrani et al. [47] propose to align existing image captions with videos by searching for video frames similar to an annotated image. For example, if an image has the caption "pop artist performs at the festival," it is possible that this image is a part of a music video. Using an image as a proxy allows us to mine for aligned video-text data. Although this approach is interesting, it is limited to the videos that happen to have some of their frames labeled for image captioning. Additionally, producing such data requires an expensive pre-processing step that includes encoding multiple video frames of each video and all images, building a maximum inner-product search index, and performing the search.

Another way to get aligned text-video pairs is automatic speech recognition. Unlike alt text, it produces long, dense captions for every video. It is used in the largest video

[*] Work done while at Amazon. Correspondance to Vlad Lialin vlialin@cs.uml.edu



HowTo100M: I was able to stuff it with random paper
Image Captioning: a man is adjusting the spoke on his motorcycle

HowTo100M: that's okay if it's a little pink in it,because it's still in the oven
Image Captioning: a pot with meat and a pink spoon sitting on top of it

Figure 1: Image captioning models provide better descriptions than the original HowTo100M labels.

description dataset currently available – HowTo100M [46]. This dataset contains 100 million instructional video clips with ASR captions. The authors of HowTo100M specifically selected the instructional domain to better align the ASR text and the video content. For example, in this domain, an espresso making tutorial can include a caption like *"grind 18 grams of coffee beans"*, providing a weak training signal for both action recognition *"grind"* and an object recognition *"coffee beans"*. However, this motivating example does not describe the usual case (Figure 1). Using a random sample of 100 HowTo100M clips, we estimate that only 45% of the captions refer to the video content in any ways (e.g., mention an object or an action). 13% are intro and outro-related speech *"hey guys we're back with another cooking video"* and 42% can be best described as chit-chat *"that's okay if it's a little pink in it, because it's still in the oven"*[1].

In this work, we propose to exploit recent advancements in image captioning [15, 35, 68] and large-scale image-text dataset mining [58, 55] for multimodal video pre-training. We explore how one can pre-train video captioning models without any aligned video-text data and show that pseudolabeling videos with image captioning models provides a strong baseline for building large-scale video-caption datasets. Unlike alt text mining, a pseudolabeling approach allows for the creation of dense labels for long videos via chunking them into small clips and labeling these individually. This approach is not limited to any particular video domain (unlike [46]) and is computationally cheaper than the approach from [47]. It only requires the generation of captions for several frames, without the need for additionally encoding large image-captioning datasets and large-scale search structures.

Our results can be summarized as follows. We utilize image captioning models to pseudolabel video pre-training data and show that it is possible to pre-train high-quality video models without any parallel video-text data. Further, we demonstrate that image captioning pseudolabels work better for pre-training than the original HowTo100M ASR captions. We investigate the importance of pre-training on both images and video and show that such a mix produces significantly better network (+4 CIDER on MSR-VTT) than pre-training on a single modality. We introduce a new *separable cross-attention* mechanism that allows to effectively attend to multidimensional data. Finally, we describe additional unexpected findings from training large multimodal models. They include tips on adapter gate implementation and initialization and the effect of ADAM's second momentum hyperparameter on training stability. Our methods are complementary to the existing pre-training or data mining approaches and can be used in a variety of settings.

2. Related work

Language-vision pre-training A pre-trained image encoder was used even in now-classical image captioning models [34, 28], but learning a joint visual-language representation became common after the success of pre-training in NLP [51, 23, 53, 17]. BERT and masked language modeling (MLM) inspired a new generation of models that use self-supervised objectives to connect language and vision [43, 62, 38, 10, 30, 37]. These methods allow learning from multimodal data using MLM-like and contrastive or optimal transport objectives. Now, generative language modeling approaches are becoming more common in both language [6, 13] and vision [67, 68, 2] thinning the lines between the multimodal NLP and computer vision fields.

Web-scale datasets Incredible results do not come from modeling alone. Increasing the training data amount is essential to achieve predictable improvement in performance [29]. Internet mining is a promising approach because unlabeled or weakly-labeled data is abundant on the Internet.

[1] The meat was not in the oven, this is an ASR error.

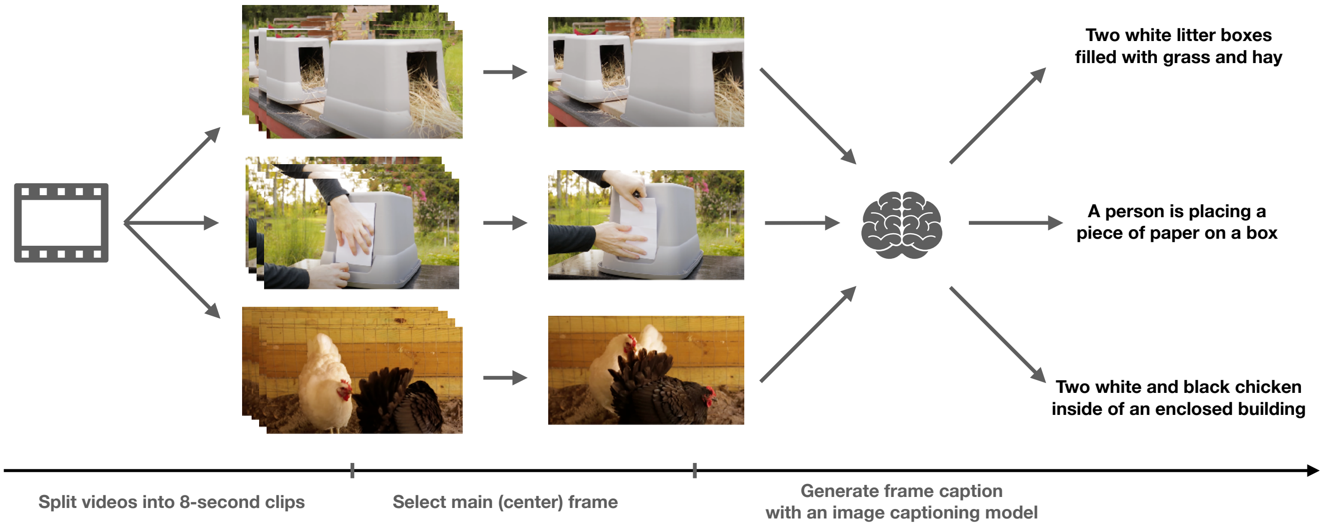

Two white litter boxes filled with grass and hay
A person is placing a piece of paper on a box
Two white and black chicken inside of an enclosed building
Split videos into 8-second clips
Select main (center) frame
Generate frame caption with an image captioning model

Figure 2: We use unlabeled videos and apply image captioning models to produce pre-training labels.

Using Wikipedia, Common Crawl, and other text sources for dataset mining was essential in the recent NLP progress [17, 6]. CLIP [52] and ALIGN [27] demonstrated the importance of the scale of weakly-supervised data for images.

Large-scale video datasets Similar to the image domain [14], until recently supervised pre-training on human-labeled datasets like Kinetics [7] dominated the field [70, 72, 40]. However, nowadays, generative approaches are becoming more widespread and successful [63, 37, 65]. These approaches require way more pre-training data than the supervised methods. Adapting image caption mining for videos is a promising direction to achieve this. Video metadata such as alt text or description of a YouTube video have been explored by Pan et al. [50] and Stroud et al. [60]. However, this unavoidably biases the dataset collection process towards short videos, as a single short caption cannot provide dense information about a video's visual content. For example, Auto-captions on GIF [50] limit the number of frames in a GIF video to 50 and WTS-70M [60] randomly select only 10 seconds of a video to download.

HowTo100M [46] takes a different approach. Each video is automatically captioned using an ASR system. This provides diverse, dense captions, yet it creates a number of method-specific problems. First, it restricts any video pre-training method from using both visual and audio modalities, as such models could ignore visual modality and focus on the speech alone. Second, ASR systems introduce multiple kinds of errors specific to these systems. Examples include mixing the speech of multiple people together, increased word error rate in noisy environments, and biasing the dataset towards people with particular accents [32], propagating ASR racial biases into new datasets. Finally, our analysis (Section 5.2) suggests that even for HowTo100M instructional videos, ASR captions *often do not describe neither the scene nor the actions*.

Utilizing image captioning for the benefits of video Several recent state-of-the-art video understanding models [2, 37, 65] use both image captioning and video captioning datasets during pre-training. However, the value of having both images and videos in pre-training has not yet been explicitly quantified.

Nagrani et al. [47] use image as a proxy between text, video, and audio. They use Conceptual Captions [58, 9] and 150M video clips and apply image search to align text to video clips and audio. Their approach yields a dataset of 6.3M video clips and 970K captions. One can see this approach as a k nearest neighbors video captioning model with $k = 1$ and a similarity metric defined by the image similarity model. It lacks caption diversity, as it cannot produce new captions unseen in the training data. In contrast to that, we propose to directly apply a high-quality image captioning model to a video frame. This significantly reduces encoding costs as we only need to process one frame per clip and removes image-video matching costs. Unlike the KNN method, frame captioning provides more diverse video captions and allows one to apply this method to a video of any domain. We demonstrate that this approach is simple and effective at producing large-scale video pre-training data.

3. Method

We propose to utilize unlabeled videos and weakly-labeled image-caption pairs to construct a large-scale video captioning dataset. In this section, we describe our method in detail. Section 4 describes an adapter-based video-conditioned language model and a novel separable cross-attention mechanism. To demonstrate the method's efficacy, we pre-train our model on different datasets and evaluate them in Section 5.

3.1. Pseudolabeling via image captioning

An image captioning model can only describe a still image. It cannot *explicitly* process any temporal information, such as how the objects move. However, we notice (Section 5.2) that in many cases images contain this information *implicitly*. For example, object orientation, placement, and blur allow inferring the movement. A type of object and its relative position to other objects allow inferring the action. In general, there is a lot of implicit information on the image that serves as a proxy allowing to describe a short clip with a single frame correctly.

We suggest that current image captioning models can utilize this information. For anecdotal evidence, we produce multiple captions[2] for three images with a moving race car, a standing race car, and a race car standing on a track (Figure 3). The model is able to distinguish between moving and standing based on the car's surroundings and only fails in the third case.

Therefore we propose to use image captioning to produce a large video-text dataset suitable for pre-training. Figure 2 describes our method. It consists of three steps: split videos into short clips, select the main frame and generate the frame caption. We split the videos into 8-second clips and feed the center frame into the video captioning model. We use publicly available state-of-the-art BLIP [35] model and nucleus sampling to generate captions. We do not use beam search as it tends to produce more bland and less diverse texts [21, 8], which would be undesirable for pre-training. Using a simple heuristic of selecting the center frame of a clip allows to minimize video decoding time by order of magnitude. This is important for high-resolution videos, where decoding can bottleneck the captioning pipeline. For a fair comparison, we use HoTo100M videos and compare our pseudolabeling method with HowTo100M ASR captions that are commonly used to pre-train video captioning networks [36, 45, 57, 61, 71, 56, 63].

4. Model

Our model architecture generally follows Flamingo [2], which was chosen for the sake of simplicity and flexibility; with several modifications specific to the video modality.

We utilize a pre-trained Transformer [64] language model and a pre-trained TimeSformer [5] network. We introduce multimodal adapters to some of the Transformer layers to condition the language generation on TimeSformer's last layer hidden states. The weights of the Transformer and the TimeSformer are kept frozen during both pre-training and fine-tuning, we only train the parameters of the adapters. Instead of using Perceiver [26] to resample videos we attend to a full video tensor with a novel separable cross-attention mechanism (Section 4.2). In our preliminary results we found it to be significantly faster and easier to train than Perceiver while maintaining the same language modeling performance.

We optimize conditional language modeling objective. During caption generation at any timestamp, the model can access all video frames V.

$$\mathcal{L} = -\sum_{i=1}^{N} \log P(w_t | w_{<t}, V),\ V \in \mathbb{R}^{t \times h \times w}$$

4.1. Multimodal adapters

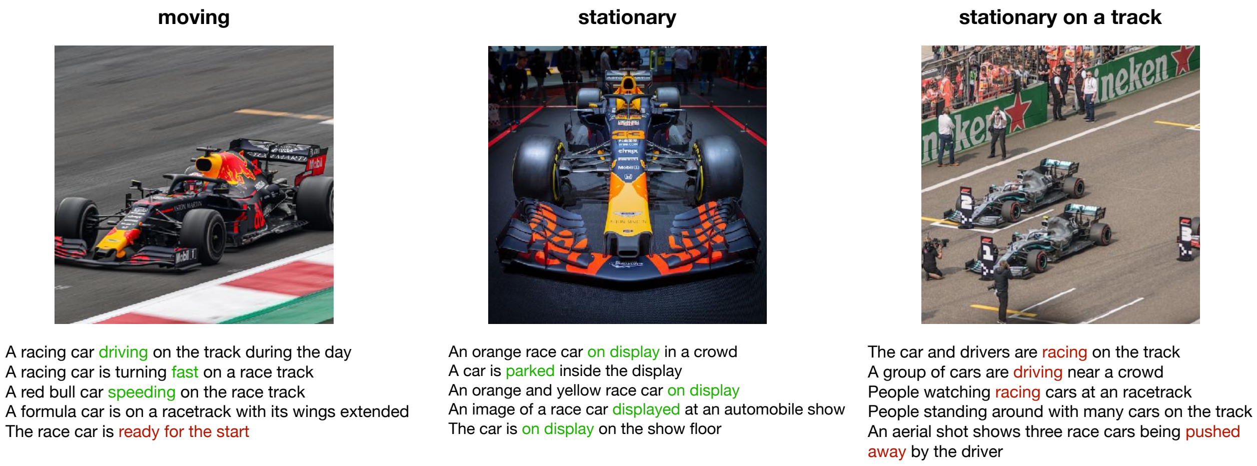
Connecting frozen models with adapters [22, 2] is beneficial for several reasons. First, adapters allow to achieve the same levels of performance as full fine-tuning of the language model [22]. Second, having less trainable parameters reduces memory requirements for optimizer states and communication volume between the GPUs in a distributed setup allowing them to scale more efficiently. Third, the frozen visual encoder does not require backpropagating to it, which saves both memory and time. This also allows us to cache visual features during training. Finally, using pre-trained language and vision models reduces training times. It maximizes the amount of pre-training data the model saw as a whole, which means that both vision and language models were already trained on vast amounts of unimodal data.

Each adapter consists of a separable cross-attention layer and a fully-connected network similar to Flamingo [2] . We use a pre-norm architecture and introduce $tanh$ gates to the residual connections to give the network a simple control mechanism of how much visual and text information to pass to the next layer. Unlike Flamingo, which uses a single scalar to weight the output of a sub-layer, we find that per-dimension (vector) gates improve training stability and allow to use higher learning rates. Figure 4 summarizes the architecture.

4.2. Separable cross-attention

Video is a challenging modality for many reasons, but the most straightforward one is the size. A video consists of many visual frames and naively could be represented as a tensor of shape `[3, t, h, w]`. Visual transformers [18], while being very successful at processing visual data, are extremely memory-hungry. The complexity of the self-attention operation is $O((ts)^2)$ where t is the number of video frames, and s is the number of spatial tokens. Such computation is both costly and vastly inefficient. Multiple solutions were proposed to mitigate this problem, and one

[2] top-p sampling with $p = 0.9$



A racing car driving on the track during the day
A racing car is turning fast on a race track
A red bull car speeding on the race track
A formula car is on a racetrack with its wings extended
The race car is ready for the start

An orange race car on display in a crowd
A car is parked inside the display
An orange and yellow race car on display
An image of a race car displayed at an automobile show
The car is on display on the show floor

The car and drivers are racing on the track
A group of cars are driving near a crowd
People watching racing cars at an racetrack
People standing around with many cars on the track
An aerial shot shows three race cars being pushed away by the driver

Figure 3: Image captions produced by BLIP [35]. In many cases, the model can correctly recognize the action and distinguish between moving vs. stationary objects on a still image using the surrounding context.

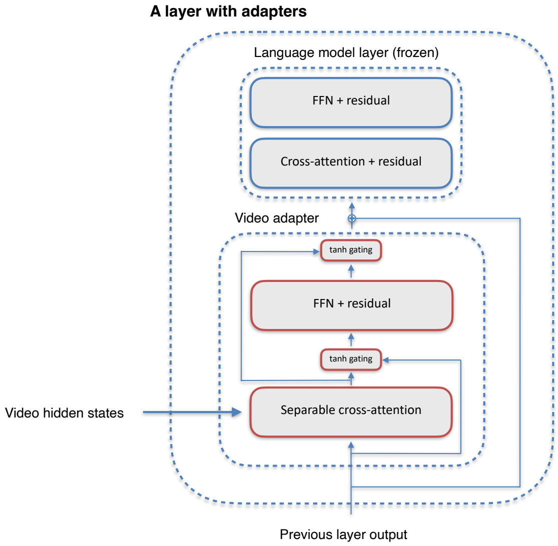


Figure 4: We insert adapters into some Transformer Layers to allow them to cross-attend to video representations using separable cross-attention.

of them is separable (axial) attention [20, 5]. Separable attention similarly to separable convolution [12] decomposes attention into multiple attentions of different axes. For example, time axis and space (height and width) axis. This reduces $O((ts)^2)$ to $O(t^2s + ts^2)$ or $40M$ operations to only $2.6M$ in a 16-patch 16-frame video.

Nevertheless, a direct adaptation of separable self-attention to the cross-attention case is just as ineffective as full cross-attention. Say the query size is q. Naïvely cross-attending from this query to every patch of the video would give us the complexity of $O(qst)$. Straightforward application of axial attention via a reshape of the time and space tensor gives us t attentions across space and s attentions across time. This yields a complexity of $O(qs \cdot t + qt \cdot s) = O(2qst)$. Such an operation is twice as expensive as a vanilla cross-attention because one has to compute multiple time and space cross-attentions.

This is why we propose to modify the separable attention mechanism for the cross-attention case. Before computing time attention, we maxpool the video tensor over the space dimension. The same operation is performed for the space attention, but we maxpool time dimension before computing it. This gives the complexity of $O(qs + qt)$. After that, both attention layer outputs are concatenated across the hidden dimension and downsampled using a linear layer. We additionally layer-normalize hidden states after maxpooling.

5. Experimental setup

5.1. Dataset generation

We apply our video pseudolabeling method described in Section 2 to HowTo100M videos. All the videos are chunked into 8 seconds, and their center frame is captioned with a BLIP-Large[3] model. In order to get high-quality captions, we use an image resolution of 320x320. Generation process took one day on 64 V100 GPUs. The resulting dataset consists of almost 50 million 8-second clips with average caption length of 10.0 words. To compare it to the original HowTo100M ASR captions, we manually evaluate 100 clips and find that in 65% of the cases, pseudolabels provide better video descriptions than ASR. We also find that ASR captions relate to the video in only 45% of the cases, while image captions correctly refer to the object and actions in the video in 88% of the cases.

[3] `BLIP/models/model_large_caption.pth`

In the following sections, we compare this dataset to the original HowTo100M captions produced by an ASR system. We pre-train our model on several data variations, including the original HowTo100M dataset and LAION-5B images, and compare them on video captioning tasks.

5.2. Pre-training

General model architecture is described in Section 4. In this section we describe the particular pre-training experiments and some hyperparameter choices.

ASR Captions vs Image Captions First, we want to learn if pseudogenerated captions are better suited for pre-training than commonly used HowTo100M ASR captions. We pre-train two models: one on 8-second clips[4] labeled with image captioning and one on the same clips but with the original ASR captions. Both models are pre-trained for 4K iterations with batch size 1.2K (about 1 epoch) and then fine-tuned on MSR-VTT.

Pre-training on image-only data Given that image captioning models are trained on weakly-supervised data and then applied to videos, a natural question arises: can we pre-train a video captioning model on images only? We treat images as one-frame videos and pre-train another model on a random subset of LAION-5B (only English captions) for the same number of steps as our video models.

Pre-training on a mix of data For this experiment, we sample examples from LAION with a probability of 0.95 and from HowTo100M with a probability of 0.05. Because preparing a video batch is usually about 20 times slower than preparing an image batch, this achieves high training throughput while having significant exposure to video data. This model is trained for the same amount of data to make it comparable to other models.

Training efficiency tricks For our pre-training experiments, we use OPT-1.3B and TimeSformer pre-trained on Kinetics and HowTo100M[5]. We insert six adapters to the OPT network, specifically to layers 12, 14, 16, 18, 20, and 22. Starting from the middle of the network allows to only compute backpropagation to the last 12 layers of the network, saving on computation and following best practices of vision-language fusion [62, 43, 48]. We use float16 precision for all parameters and Deepspeed Stage 2 data-parallel for distributed training. To maximize GPU utilization, we use roughly eight times larger batch size for images than for videos. This accomplishes two things: maximizing the batch size, thus increasing throughput and improving stochastic gradient estimate, and evening out the time for a forward-backward pass across the GPUs, minimizing wait times before synchronization.

Image Captions	LAION-5B	ASR	MSR-VTT
		✓	49.0
✓			49.7
	✓		49.6
✓	✓		**54.0**

Table 1: Ablation studies. MSR-VTT validation set, no beam search, all models are pre-trained on 500K examples (videos + images if any). Pseudo-labeling is significantly more effective than original HowTo100M ASR captions. Training on only images or only videos is significantly less effective than training on both with 95% images 5% videos mixture.

Using a distributed training setup simplifies data mixing and batching. Every batch contains either images or videos, but different GPUs sample if they need to process images or videos independently. This means that the stochastic gradient estimate across all GPUs includes both images and videos. At the same time, each particular GPU processes the same kind of data without needing padding or complex batching rules. We also test fully-synchronized modality selection when all GPUs process either videos or images simultaneously, but we find fully-synchronized training very unstable compared to mixed training.

6. Results

We fine-tune pre-trained models on MSR-VTT and MSVD video captioning datasets. Comparison of different pre-training datasets is presented in Table 1. Using HowTo100M ASR captions produces the worst results of all, demonstrating the low quality of video-text alignment that ASR provides. On the other hand, our pseudolabeled dataset performs on par with a similar amount of image captioning data. Combining videos and images outperforms the rest by more than 4 CIDER-D points, showing that having both modalities in the data is the most effective.

To push the results further, we train our model for 40K steps on the mixture of videos and texts. Our results compared to state of the art are presented in Table 2. Our model underperforms the current state of the art that we attribute to the frozen visual network. Unlike GIT [65], we do not modify visual representations inside a transformer but only attend to them. This, together with separable cross-attention (Section 4.2), allows our model to scale linearly with the number of video frames and attend to significantly longer videos than GIT, which scales quadratically.

[4] If needed we concatenate multiple HowTo100M clips together to form videos of roughly 8 seconds.

[5] `TimeSformer_divST_96x32_224_HowTo100M`

	Pre-training data	Input features	MSVD	MSR-VTT
O2NA [42]	-	video frames	96.4	51.1
DECEMBERT [63]	HowTo100M	video frames, ASR, image captions	-	52.3
MV-GPT [56]	HowTo100M	video frames, ASR	-	60.0
LAVENDER [37]	LAVENDER mixture	video frames	150.7	60.1
GIT [65]	GIT mixture	video frames	180.2	73.9
FrozenCaptioner	HowTo100Mblip + LAION	video frames	128.8	54.6

Table 2: Comparison with the state of the art models. HTM stands for HowTo100M. LAVENDER mixture is Vid2.5M [4] + CC3M [58] + CC12M [9] + COCO [41] + Visual Genome [33] + SBU Captions [49] + 12M crawled video text pairs. GIT mixture is similar to LAVENDER, but also includes ALT200M [24].

Additional findings

Vector gates improve training stabiliy We observed that using scalar gates similar to Flamingo [2] causes loss divergences at high learning rates (greater than 10^{-3}). One simple and effective way of mitigating this problem was using vector (per-dimension) gates that allowed us to use a very high learning rate $7 \cdot 10^{-3}$. We hypothesize that per-dimension gates serve as a kind of a normalization layer [25, 3], and they can cut off some large value dimensions in the adapter outputs.

Effect of Adam second momentum Starting with GPT-3 [6] several large models [69, 59] used Adam's [31] $\beta_2 = 0.95$ which is significantly smaller than the default $\beta_2 = 0.999$. In our experiments we found that a small β_2 value stabilizes the training with minimal effect on convergence speed. However, several of our experiments suggest that small values of β_2 can negatively impact generalization and fine-tuning capabilities. We found that models trained with $\beta_2 = 0.95$ for 10K+ steps can significantly underperform models trained with $\beta_2 = 0.999$ trained for 4K steps on downstream tasks. This happens even though the pre-training loss of 320px models is lower, suggesting that $\beta_2 = 0.95$ hurts generalization capability.

Unreasonable effectiveness of tanh(1) initialization Flamingo initializes gates at $tanh(0) = 0$. This achieves two things: first, it maximizes the gradient through the $tanh$ nonlinearity allowing to learn optimal gate values faster. Second, it allows each layer to smoothly learn how much visual information it should contribute to the language model. However, adapter values change very slowly during training, requiring many thousands of iterations to converge. This is usually way past the point of overfitting and losing generalization capabilities on relatively small downstream datasets like MSR-VTT.

For a more fair comparison between pre-trained and non-pre-trained networks, we evaluate non-pre-trained in two scenarios. The adapter gates of the first network are initialized at $tanh(0) = 0$ and for the second they are initialized at $tanh(1) \approx 0.8$. Initializing $tanh$ values closer to 1 skyrockets a non-pre-trained models performance and reduces the pre-trained model's gap from 20.8 to 5.3 CIDER-D points (Figure 5). Diving deeper into networks fine-tuned on MSR-VTT shows almost no change with less than 0.05 absolute change in gate values during fine-tuning for either model explaining the drastic difference in performance.

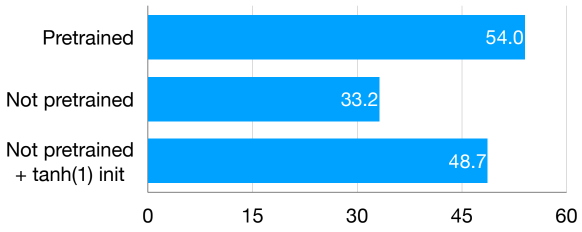


Figure 5: Initialization of visual gates closer to 1 boost non-pretrained network performance. MSR-VTT CIDEr-D scores after fine-tuning.

Effect of crop size during pre-training and fine-tuning We pre-train two networks: with 224x224 crop and with 320x320 crop. Both of them are trained with the same batch size and the number of update steps[6]. After pre-training we observe that the 320px crop network has lower training loss, but underperforms the 224px network on MSR-VTT consistently throughout the training.

An intuitive explanation could be that the TimeSformer vision encoder was pre-trained on 224x224 videos, and using a larger crop introduces a distribution shift. However, after fine-tuning, we see a different picture. Using larger video sizes during fine-tuning consistently improves captioning quality, plateauing at 320-380 pixels. Using a smaller resolution of 224x224 during pre-training, thus, seems like a great way to improve both training speed and quality, but the reason why it is so effective requires further investigation.

[6] We compensate for increased memory using gradient accumulation

7. Conclusion

In this paper, we show that it is possible to pre-train high-quality video models without any parallel video-text data. To do this, we employ a simple but effective video pseudolabeling technique: captioning individual frames with high-quality image description models. We demonstrate that current image captioning models can provide useful video captions that allow the network to learn both static (objects) and dynamic (actions) information about the video. To evaluate our pseudolabeling method, we pre-train several adapter-based captioning models and show that image captions provide better training signal than commonly used ASR captions. We additionally demonstrate the importance of using both images and videos in pre-training. Finally, we develop a new cross-attetion method that allows to effectively and efficiently attend to dense video representations.

8. Limitations

Using unaligned video data is a promising path toward high-quality video models. However, image captioning models are not perfect. They make factual mistakes and exhibit societal biases such as race, gender, cultural and more from the training data. Using them for large-scale dataset creation can amplify these biases and requires mitigation techniques. Image captioning models also suffer from hallucinations. For example, mentioning objects that are not on a picture. While a high-resolution frame that we use for captioning contains a lot of information that can be inferred, aligned video-text data mining is still an open question. A potential solution could be a combination of web-mining techniques that work well for short videos and dense pseudolabeling techniques for longer videos.

We also would like to highlight some of the domains where our methods can perform poorly. For example, video description for hearing-impaired people. This task requires full video understanding including visual and audio modalities. Image captioning pseudolabels do not utilize audio modality and cannot provide a training signal to describe what people say or what sounds the environment makes and should be augmented with aligned (or pseudoaligned) audio-text data as well.

References

[1] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Apostol (Paul) Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. In *arXiv:1609.08675*, 2016.

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. *ArXiv*, abs/2204.14198, 2022.

[3] Jimmy Ba, Jamie Ryan Kiros, and Geoffrey E. Hinton. Layer normalization. *ArXiv*, abs/1607.06450, 2016.

[4] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *IEEE International Conference on Computer Vision*, 2021.

[5] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, 2021.

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc., 2020.

[7] João Carreira, Eric Noland, Chloe Hillier, and Andrew Zisserman. A short note on the kinetics-700 human action dataset. *ArXiv*, abs/1907.06987, 2019.

[8] David M Chan, Yiming Ni, Austin Myers, Sudheendra Vijayanarasimhan, David A Ross, and John Canny. Distribution aware metrics for conditional natural language generation. *arXiv preprint arXiv:2209.07518*, 2022.

[9] Soravit Changpinyo, Piyush Kumar Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3557–3567, 2021.

[10] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Learning universal image-text representations. *ArXiv*, abs/1909.11740, 2019.

[11] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying vision-and-language tasks via text generation. In *International Conference on Machine Learning*, pages 1931–1942. PMLR, 2021.

[12] Francois Chollet. Xception: Deep learning with depthwise separable convolutions. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul 2017.

[13] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek B Rao, Parker Barnes, Yi Tay, Noam M. Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du,

Benton C. Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier García, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Oliveira Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Díaz, Orhan Firat, Michele Catasta, Jason Wei, Kathleen S. Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways. *ArXiv*, abs/2204.02311, 2022.

[14] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[15] Karan Desai and Justin Johnson. Virtex: Learning visual representations from textual annotations. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11157–11168, 2021.

[16] Karan Desai, Gaurav Kaul, Zubin Aysola, and Justin Johnson. RedCaps: Web-curated image-text data created by the people, for the people. In *NeurIPS Datasets and Benchmarks*, 2021.

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *ArXiv*, abs/1810.04805, 2019.

[18] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929, 2021.

[19] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[20] Jonathan Ho, Nal Kalchbrenner, Dirk Weissenborn, and Tim Salimans. Axial attention in multidimensional transformers. *ArXiv*, abs/1912.12180, 2019.

[21] Ari Holtzman, Jan Buys, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. *ArXiv*, abs/1904.09751, 2020.

[22] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR, 09–15 Jun 2019.

[23] Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. In *ACL*, 2018.

[24] Xiaowei Hu, Zhe Gan, Jianfeng Wang, Zhengyuan Yang, Zicheng Liu, Yumao Lu, and Lijuan Wang. Scaling up vision-language pretraining for image captioning. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 17959–17968, 2022.

[25] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *ArXiv*, abs/1502.03167, 2015.

[26] Andrew Jaegle, Felix Gimeno, Andrew Brock, Andrew Zisserman, Oriol Vinyals, and João Carreira. Perceiver: General perception with iterative attention. In *ICML*, 2021.

[27] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 2021.

[28] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4565–4574, 2016.

[29] Jared Kaplan, Sam McCandlish, T. J. Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeff Wu, and Dario Amodei. Scaling laws for neural language models. *ArXiv*, abs/2001.08361, 2020.

[30] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *ICML*, 2021.

[31] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2015.

[32] Allison Koenecke, Andrew Joo Hun Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky, and Sharad Goel. Racial disparities in automated speech recognition. *Proceedings of the National Academy of Sciences of the United States of America*, 117:7684 – 7689, 2020.

[33] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123:32–73, 2016.

[34] Jie Lei, Tamara L. Berg, and Mohit Bansal. Revealing single frame bias for video-and-language learning. *ArXiv*, abs/2206.03428, 2022.

[35] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022.

[36] 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*, abs/2005.00200, 2020.

[37] Linjie Li, Zhe Gan, Kevin Lin, Chung-Ching Lin, Zicheng Liu, Ce Liu, and Lijuan Wang. Lavender: Unifying video-language understanding as masked language modeling. *ArXiv*, abs/2206.07160, 2022.

[38] Xiujun Li, Xi Yin, Chunyuan Li, Xiaowei Hu, Pengchuan Zhang, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*, 2020.

[39] Yuncheng Li, Yale Song, Liangliang Cao, Joel R. Tetreault, Larry Goldberg, Alejandro Jaimes, and Jiebo Luo. Tgif: A new dataset and benchmark on animated gif description. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4641–4650, 2016.

[40] Kevin Lin, Linjie Li, Chung-Ching Lin, Faisal Ahmed, Zhe Gan, Zicheng Liu, Yumao Lu, and Lijuan Wang. Swinbert: End-to-end transformers with sparse attention for video captioning. *ArXiv*, abs/2111.13196, 2021.

[41] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. *ArXiv*, abs/1405.0312, 2014.

[42] Fenglin Liu, Xuancheng Ren, Xian Wu, Bang Yang, Shen Ge, and Xu Sun. O2na: An object-oriented non-autoregressive approach for controllable video captioning. In *FINDINGS*, 2021.

[43] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, 2019.

[44] Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-io: A unified model for vision, language, and multi-modal tasks. *ArXiv*, abs/2206.08916, 2022.

[45] 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. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9876–9886, 2020.

[46] 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 *ICCV*, 2019.

[47] Arsha Nagrani, Paul Hongsuck Seo, Bryan Seybold, Anja Hauth, Santiago Manén, Chen Sun, and Cordelia Schmid. Learning audio-video modalities from image captions. *ArXiv*, abs/2204.00679, 2022.

[48] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. In *NeurIPS*, 2021.

[49] Vicente Ordonez, Girish Kulkarni, and Tamara L Berg. Im2text: Describing images using 1 million captioned photographs. In *Proceedings of the 24th International Conference on Neural Information Processing Systems*, NIPS'11, page 1143–1151, Red Hook, NY, USA, 2011. Curran Associates Inc.

[50] Yingwei Pan, Yehao Li, Jian-Hao Luo, Jun Xu, Ting Yao, and Tao Mei. Auto-captions on gif: A large-scale video-sentence dataset for vision-language pre-training. *ArXiv*, abs/2007.02375, 2020.

[51] Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. Deep contextualized word representations. In *NAACL*, 2018.

[52] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

[53] Alec Radford and Karthik Narasimhan. Improving language understanding by generative pre-training. 2018.

[54] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.

[55] Christoph Schuhmann, Romain Beaumont, Cade W Gordon, Ross Wightman, Theo Coombes, Aarush Katta, Clayton Mullis, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, et al. Laion-5b: An open large-scale dataset for training next generation image-text models.

[56] Paul Hongsuck Seo, Arsha Nagrani, Anurag Arnab, and Cordelia Schmid. End-to-end generative pretraining for multimodal video captioning. *ArXiv*, abs/2201.08264, 2022.

[57] Paul Hongsuck Seo, Arsha Nagrani, and Cordelia Schmid. Look before you speak: Visually contextualized utterances. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16872–16882, 2021.

[58] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of ACL*, 2018.

[59] Shaden Smith, Mostofa Ali Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Anand Korthikanti, Elton Zhang, Rewon Child, Reza Yazdani Aminabadi, Julie Bernauer, Xia Song, Mohammad Shoeybi, Yuxiong He, Michael Houston, Saurabh Tiwary, and Bryan Catanzaro. Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model. *ArXiv*, abs/2201.11990, 2022.

[60] Jonathan C. Stroud, David A. Ross, Chen Sun, Jia Deng, Rahul Sukthankar, and Cordelia Schmid. Learning video representations from textual web supervision. *ArXiv*, abs/2007.14937, 2020.

[61] Chen Sun, Austin Myers, Carl Vondrick, Kevin P. Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7463–7472, 2019.

[62] Hao Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *ArXiv*, abs/1908.07490, 2019.

[63] Zineng Tang, Jie Lei, and Mohit Bansal. DeCEMBERT: Learning from noisy instructional videos via dense captions and entropy minimization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2415–2426, Online, June 2021. Association for Computational Linguistics.

[64] 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.

[65] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang.

Git: A generative image-to-text transformer for vision and language. *ArXiv*, abs/2205.14100, 2022.

[66] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *ICML*, 2022.

[67] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. *ArXiv*, abs/2108.10904, 2022.

[68] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *ArXiv*, abs/2205.01917, 2022.

[69] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. Opt: Open pre-trained transformer language models, 2022.

[70] Ziqi Zhang, Yaya Shi, Chunfen Yuan, Bing Li, Peijin Wang, Weiming Hu, and Zhengjun Zha. Object relational graph with teacher-recommended learning for video captioning. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13275–13285, 2020.

[71] Linchao Zhu and Yi Yang. Actbert: Learning global-local video-text representations. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8743–8752, 2020.

[72] Xinxin Zhu, Longteng Guo, Peng Yao, Jing Liu, Shichen Lu, Zheng Yu, Wei Liu, and Hanqing Lu. Multi-view features and hybrid reward strategies for vatex video captioning challenge 2019. *ArXiv*, abs/1910.11102, 2019.