Learning text-to-video retrieval from image captioning

Lucas Ventura[1,2] Cordelia Schmid[2] Gül Varol[1]
[1] LIGM, École des Ponts, Univ Gustave Eiffel, CNRS, France
[2] Inria, ENS, CNRS, PSL Research University, France
lucas.ventura@enpc.fr

Abstract

We describe a protocol to study text-to-video retrieval training with unlabeled videos, where we assume (i) no access to labels for any videos, i.e., no access to the set of ground-truth captions, but (ii) access to labeled images in the form of text. Using image expert models is a realistic scenario given that annotating images is cheaper therefore scalable, in contrast to expensive video labeling schemes. Recently, zero-shot image experts such as CLIP have established a new strong baseline for video understanding tasks. In this paper, we make use of this progress and instantiate the image experts from two types of models: a text-to-image retrieval model to provide an initial backbone, and image captioning models to provide supervision signal into unlabeled videos. We show that automatically labeling video frames with image captioning allows text-to-video retrieval training, which adapts the features to the target domain at no manual annotation cost, consequently outperforming the strong zero-shot CLIP baseline. We extract captions from multiple video frames and use a scoring mechanism to filter out the captions that best match the visual content. We conduct ablations to provide insights and demonstrate the effectiveness of this simple framework by outperforming the CLIP zero-shot baseline on text-to-video retrieval on two standard datasets, namely MSR-VTT and MSVD. Code and models will be made publicly available.

1. Introduction

The research on automatic video understanding has witnessed a number of paradigm shifts recently. Following the rise of neural networks, the initial question was how to design an architecture to input spatio-temporal signals [18, 23]. Given the limited video training data, the focus then shifted to borrowing parameter initialization from image classification pretraining [3]. In an attempt to provide *video* pretraining, one line of work has put expensive efforts into annotating video classification datasets [11]. On the other hand, the research community is moving away from closed-vocabulary recognition training as the progress in language modeling inspired advances in retrieval of visual data given open-vocabulary textual input, bridging the gap between symbolic action categories and describing actions as text [10]. The latest shift was due to the huge scale of labeled image data, resulting in impressive zero-shot capability of image-text retrieval models on video action recognition tasks [19]. Now, the performance of the CLIP-initialized [19] image features (simply averaged over video frames) surpasses most previous works on a large number of video understanding tasks [16, 27]. This makes researchers question and rethink where to put their efforts to improve video modeling. In this study, we focus on enhancing the *zero-shot* text-to-video retrieval performance of CLIP by making a realistic assumption that we have access to *image experts*, more specifically an image captioning model.

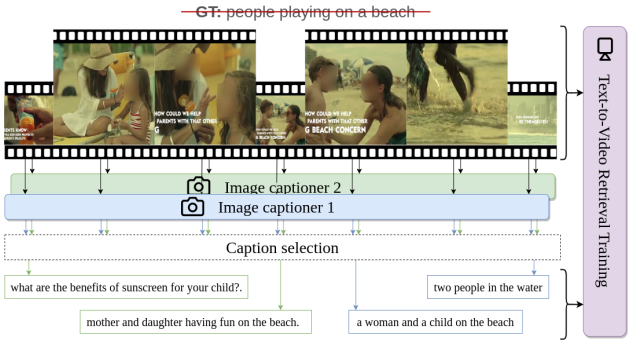

Figure 1. **Framework:** Instead of using the ground-truth video caption, we extract *image* captions to automatically label *unlabeled video* frames, which we filter to obtain high-quality captions. During the training process, we randomly select one of these captions to pair with each video and train for text-to-video retrieval.

Fully-supervised methods for video retrieval are limited due to the high cost of video annotation. Even training with the web-scale video-text pairs [1] do not outperform CLIP image-text pretraining [4], despite the rich descriptions typed manually by humans with the motivation to sell their videos on stock websites. On the other hand, meth-

ods learning from unlabeled videos often assume no access to *any* labels, even for images, with a particular focus on self-supervised training to use the structure of the data itself as the training signal [7, 8, 30]. In this paper, we ask the question whether an external off-the-shelf image expert can provide the supervision signal. We explore the usability of recently released robust image captioners, namely ClipCap [17] and BLIP [13], which benefit from training with large-scale image-text pairs. For example, ClipCap uses both CLIP visual pretraining and GPT-2 language model pretraining [20]. When applied on video frames, we observe that, while noisy, the output texts contain high-quality descriptions, which motivates this exploration.

While the idea of using automatic image captions is appealing, incorporating such *noisy* labels for training introduces additional challenges. To address this issue, we first employ a filtering approach where we select the captions that better describe the frame by computing the CLIPScore metric [9]. Measuring such cross-modal similarity between the visual frame and the output text is similar in spirit to the filtering step in [13]. Furthermore, we ensemble multiple image captioners to obtain a larger pool of labels. We experimentally validate the benefit of these steps in our ablations.

In this work, we test whether off-the-shelf image captioning models can serve as an automatic labeling strategy for video retrieval tasks. We propose a simple framework to answer this question. Our baseline, as well as our weight initialization, is CLIP [19]. We finetune this model such that video frame embeddings and the automatic captions map to the cross-modal joint space after a contrastive retrieval training. We demonstrate through experiments that our approach to pseudo-label unlabeled video frames with image captioning is a simple, yet effective strategy that boosts the performance over baselines.

Our contributions are three-fold: 1) We propose a new simple approach to train video retrieval models using automatic frame captions, which constitute free labels for supervision (see Figure 1). To the best of our knowledge, off-the-shelf captioning has not been used for such objective by prior work. 2) We outperform the zero-shot state-of-the-art CLIP model on three text-to-video retrieval benchmarks. 3) We provide ablations about the design choices on how to select high-quality captions. The code and models will be publicly available[1].

2. Training with automatic captions

In this section, we first describe how we obtain automatic captions for labeling videos, video retrieval training, and finally give implementation details for our experimental setup.

[1] http://imagine.enpc.fr/∼ventural/multicaps

The overview of our method is illustrated in Figure 2. In summary, we start by constructing a set of labels for each video, by applying image captioning models on video frames. Given these noisy frame-level captions (from multiple image captioners), we select the high-quality ones by sorting them according to their CLIPScore [9]. We adopt a contrastive video-text retrieval training using one of the selected captions.

Selecting high-quality captions. Given an unlabeled training video v consisting of F frames, we select M frames from the video ($M \leq F$) and extract captions using I image captioners to form an initial set of labels $\mathbb{C} = \{\mathbb{C}_i\}_{i=1}^{I}$, where $\mathbb{C}_i = \{c_{i1}, c_{i2}, \ldots, c_{iM}\}$. We then obtain I textual descriptions per frame, resulting in a total of $M \times I$ labels per video.

While we investigate several variants of label formation from captions in our experiments, our final strategy is the following. We select a subset of the initial labels, mainly to eliminate noisy captions that do not well represent the corresponding video frame. To this end, we employ CLIPScore [9] as a way to measure cross-modal similarity between a caption and its corresponding frame. For each captioner, we keep the top-K captions ($K < M$) with the highest CLIPScores, which gives us a remaining $L = K \times I$ labels per video. We refer to this subset as $\mathbb{C}'$. Note that some captions are repetitive across frames due to visual similarity within a video; we therefore conjecture that such a subset selection does not cause a significant loss in information.

Contrastive video retrieval objective. In this work, we employ a relatively standard vision-language cross-modal training, where the goal is to find a joint space between videos and automatic captions. Given a video v, we compute visual embeddings $\bar{\mathbb{V}} = \{\bar{v}_n\}_{n=1}^{N}$ on N video frames ($N \leq F$) using a visual encoder $f_v : \bar{v}_n \rightarrow \mathbb{R}^d$. Similarly, we compute textual embeddings with the text encoder f_t from the corresponding set of labels $\mathbb{C}'$ to obtain positive text representations $\bar{\mathbb{C}} = \{\bar{c}_l\}_{l=1}^{L}$, where $\bar{c}_l \in \mathbb{R}^d$ (with the same embedding dimension as $\bar{v}_n$). To obtain a single video embedding, we perform temporal pooling over video frame representations. During training, we randomly sample one positive label $\bar{c}_l$ from the set of candidate labels $\bar{\mathbb{C}}$ and we feed it to the text encoder to obtain the positive text representation. Finally, the pooled video embedding is compared against the text embedding with cosine similarity ϕ.

From a batch of B visual-texts pair samples, $\{(\bar{\mathbb{V}}_1, \bar{\mathbb{C}}_1), (\bar{\mathbb{V}}_2, \bar{\mathbb{C}}_2), ..., (\bar{\mathbb{V}}_B, \bar{\mathbb{C}}_B)\}$, we train with a symmetric contrastive loss using InfoNCE [24], i.e., treating all other samples in the batch as negatives:

$$\mathcal{L}_{v2c} = -\frac{1}{B}\sum_{b\in B}\log\frac{\exp(\phi(\mathbb{V}_b, \bar{c}_{b,l}))}{\sum_{j\in B}\exp(\phi(\mathbb{V}_b, \bar{c}_{j,l}))} \quad (1)$$

$$\mathcal{L}_{c2v} = -\frac{1}{B}\sum_{b\in B}\log\frac{\exp(\phi(\mathbb{V}_b, \bar{c}_{b,l}))}{\sum_{j\in B}\exp(\phi(\mathbb{V}_j, \bar{c}_{b,l}))} \quad (2)$$

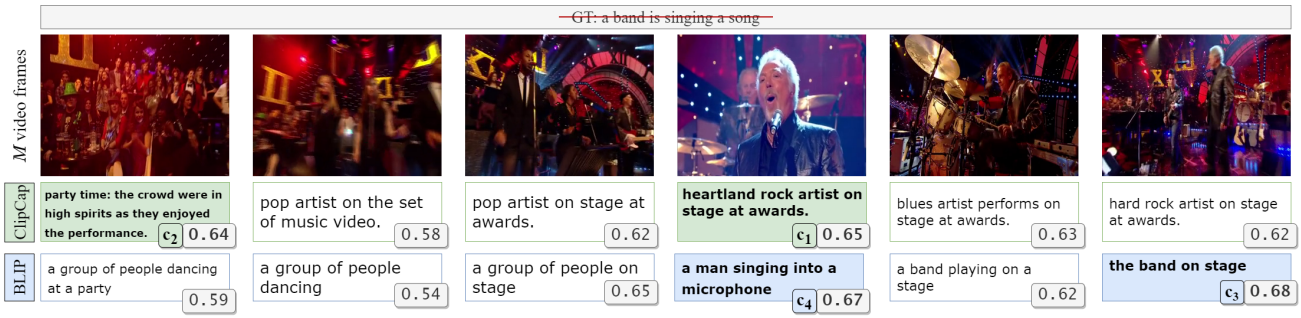

Figure 2. **Caption selection:** To select the best captions for a given video, we first extract image captions from both ClipCap [17] and BLIP [13] models for M number of frames. We then compute the CLIPScore [2] (gray box), and finally select Top $K = 2$ captions for each captioner: c_1 and c_2 for ClipCap (highlighted in green), and c_3 and c_4 for BLIP (highlighted in blue).

$$\mathcal{L} = \mathcal{L}_{c2v} + \mathcal{L}_{v2c}, \tag{3}$$

The final loss is the sum of video-to-captions ($\mathcal{L}_{v2c}$) and captions-to-video ($\mathcal{L}_{c2v}$) retrieval loss terms. Next, we detail the optimization procedure.

Implementation details. We instantiate two image captioners ($I = 2$) from ClipCap [17] and BLIP [13] models. ClipCap model is pretrained on the 3M images of the Google Conceptual Captions image-text dataset [22], using a MLP mapping between CLIP [19] image backbone and GPT-2 [20] text generation models. BLIP jointly trains for retrieval and captioning using 129M images (including a subset of LAION [21]) using a bootstrapping approach. We use the publicly available model, which is further finetuned on the COCO dataset [14]. Given one captioner, we extract $M = 10$ captions per video from equally spaced frames. We empirically set the number of high-quality captions to top $K = 2$ per captioner (i.e., $L = K \times I = 4$).

We minimize the loss function in Eq. 3 using Adam [12] optimizer and a learning rate schedule with a cosine decay [15] as described in [16]. For both datasets, we train on 4 NVIDIA GeForce GTX 1080 for 10 epochs, with initial learning rate 10^{-4} and mini-batch size $B = 16$.

The weights of our dual encoder model are initialized from CLIP [19] pretraining both for the image (f_v) and the text (f_t) encoders. The image encoder architecture follows ViT-B/16 [6] in all experiments. The text encoder architecture follows GPT-2 [20]. Both encoders are Transformer-based [25].

We resize the frames to 224×224 resolution before inputting to the model. We use $N = 10$ random frame sampling during training based on segments as in [1, 26] (note that these do not necessarily match the $M = 10$ captions). At test time, we compute the visual embeddings on the center spatial crop over 10 equally spaced frames.

3. Experiments

We start with Section 3.1 by describing the datasets and evaluation metrics used to report the results of our experiments. We then present our ablations in Section 3.2, quantifying the effects of the the captioning model and caption selection.

	MSR-VTT		MSVD	
	R@1	R@5	R@1	R@5
CLIP baseline [19]	32.80	55.73	39.39	64.55
Ours w/ OFA [28]	33.60	59.15	41.06	67.42
Ours w/ ClipCap [17]	34.71	59.76	40.61	68.94
Ours w/ BLIP [13]	**35.81**	**60.56**	**41.11**	**69.09**

Table 1. **Captioning models:** Training with automatic captions obtained with OFA [28], ClipCap [17], and BLIP [13] all improve over the zero-shot CLIP baseline [19] on all three text-to-video retrieval benchmarks. BLIP captions result in best performances.

3.1. Datasets and evaluation metrics

We conduct experiments on two established benchmarks for text-to-video retrieval, namely MSR-VTT [29] and MSVD [5] datasets.

As previously explained, even though these datasets contain ground-truth captions, we do *not* use them during training. We report the standard evaluation protocols: text-to-video (T2V) recall at rank 1 and 5 for all experiments. Recall at rank k (R@k) quantifies the number of times the correct video is among the top k results. Higher recall means better performance.

3.2. Ablation study

This work constitutes an exploratory study to test whether captions can provide training signal for unlabeled videos. The answer is yes; however, there are certain design choices we make. Here, we provide ablations to measure the sensitivity to these decisions. More specifically, we investigate the effects of the captioning model, the quality of the captions provided to the model, and combining captions from different captioners.

(i) Captioning models. The first design choice is on the image captioning model to use. In Table 1, we present

Captioner	Caption selection	MSR-VTT R@1	MSR-VTT R@5	MSVD R@1	MSVD R@5
ClipCap	Rand(10)	31.79	55.23	39.75	68.48
	Middle 1	34.10	56.94	38.89	66.97
	Top 1	34.31	57.95	40.45	68.64
	Rand(Top 2)	**34.71**	**59.76**	**40.61**	**68.94**
	Rand(Top 3)	33.10	58.95	40.45	68.38
BLIP	Rand(10)	34.61	60.46	40.45	68.74
	Middle 1	33.20	57.75	40.10	69.85
	Top 1	34.91	60.26	**41.82**	68.33
	Rand(Top 2)	**35.81**	**60.56**	41.11	**69.09**
	Rand(Top 3)	35.61	59.46	40.91	68.18

Table 2. **Caption selection:** For both captioners, we compare training with a random caption at each epoch, training with only the middle frame caption, and training with different number of Top K captions (best CLIPScore [9]). Using CLIPScore filtering improves over using all the 10 captions or only using the middle one on both datasets. Selecting the Top 2 captions results in overall best performance.

a comparative study experimenting with three recent captioning models: OFA [28], ClipCap [17] and BLIP [13]. More specifically, we use the best available model checkpoints: OFA-huge trained with 20M publicly available image-text pairs, ClipCap trained with Conceptual Captions, and BLIP-Large trained with 129M images, finetuned on COCO. Best results are obtained with BLIP, potentially due to the large amount of pretraining compared to the other two models. The results also demonstrate the effectiveness of using captions to improve over the strong CLIP baseline [19], where we average video frame embeddings using the frozen CLIP. In this experiment, we randomly select one caption out of the two best captions during training. We next assess the influence of this selection.

(ii) Caption selection. Automatically generated captions vary in quality. We select the captions with high image-text compatibility to eliminate potential noise in our training. The above image captioning models do not output a confidence score; therefore, we use CLIPScore [9] between the generated caption and the corresponding input video frame as a caption quality measure.

In Table 2, we evaluate whether such filtering is beneficial. In this experimental setup, we train with one caption as the video label. We experiment with five different variants per captioner: (a) randomly selecting one of the 10 extracted captions at each epoch, (b) using only the caption corresponding to the middle frame (i.e., same label in all epochs), (c) using only the best caption (i.e., top 1 based on the CLIPscore metric), (d) randomly selecting one of the 2 best captions at every epoch, (e) randomly selecting one of the 3 best captions at every epoch. The results support the idea that CLIPScore is an effective filtering method to keep the highest quality captions. On both datasets, and on both captioners (ClipCap and BLIP), using the best caption(s) improves over using all the captions or the middle one. There exists a trade-off between the number of captions and their quality. With more captions per video we avoid overfitting as this may serve as data augmentation. On the other hand, the variance among the caption qualities starts to increase. We empirically find that taking the best two captions constitutes a good compromise, yielding the best performance overall.

	MSR-VTT R@1	MSR-VTT R@5	MSVD R@1	MSVD R@5
C	34.71	59.76	40.61	68.94
B	35.81	60.56	41.11	69.09
C+B	**36.52**	**61.47**	**41.72**	**70.00**

Table 3. **Combining two captioners:** We observe slight improvements when using captions from both ClipCap (C) and BLIP (B) over using them individually.

(iii) Combining captioners. One way to increase the amount of captions per video without decreasing the quality of the captions is to use the best K captions from each captioner to form the label set. In Table 3, we test this hypothesis by taking two captioners ClipCap and BLIP, to then ensemble their labels. The results are slightly better than the performance of individual captioners. One can potentially further extend to more captioners $I > 2$.

Note that we could also select the top K from all the captions combined from both captioners. This would be equivalent to taking the best 2 captions out of the 20 (10 per captioner). However, this leads to poorer results, perhaps due to the different CLIPScore distributions (slight preference for ClipCap potentially because of the CLIP backbone), and tendency to output repetitive captions across frames for a given captioner.

4. Conclusion

We showed a simple yet effective framework to utilize an image captioning model as a source of supervision for text-to-video retrieval datasets. We demonstrated significant improvements over the strong zero-shot CLIP baseline with a comprehensive set of experiments. Our method comes with limitations. First, we note that image captioning does not necessarily capture the dynamic content of videos. Similarly, our temporal pooling approach remains simple, ignoring the order of frames. Future work will explore upgrading the image captioning model with a video captioning model or using multiple captions during training.

Acknowledgements. This work was granted access to the HPC resources of IDRIS under the allocation 2022-AD011013060 made by GENCI. The authors would like to acknowledge the research gift from Google, the ANR project CorVis ANR-21-CE23-0003-01, and thank Elliot Vincent, Charles Raude, Georgy Ponimatkin, and Andrea Blazquez for their feedback.

References

[1] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *ICCV*, 2021. 1, 3

[2] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. A CLIP-hitchhiker's guide to long video retrieval. *arXiv*, 2022. 3

[3] João Carreira and Andrew Zisserman. Quo vadis, action recognition? A new model and the Kinetics dataset. In *CVPR*, 2017. 1

[4] Santiago Castro and Fabian Caba Heilbron. FitCLIP: Refining large-scale pretrained image-text models for zero-shot video understanding tasks. *arXiv*, 2022. 1

[5] David Chen and William Dolan. Collecting highly parallel data for paraphrase evaluation. In *Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 2011. 3

[6] 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. In *ICLR*, 2021. 3

[7] Christoph Feichtenhofer, Haoqi Fan, Bo Xiong, Ross B. Girshick, and Kaiming He. A large-scale study on unsupervised spatiotemporal representation learning. In *CVPR*, 2021. 2

[8] Daniel Gordon, Kiana Ehsani, Dieter Fox, and Ali Farhadi. Watching the world go by: Representation learning from unlabeled videos. *arXiv:2003.07990*, 2020. 2

[9] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In *EMNLP*, 2021. 2, 4

[10] Chen Ju, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie. Prompting visual-language models for efficient video understanding. *arXiv:2112.04478*, 2021. 1

[11] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The Kinetics human action video dataset. *arXiv:1705.06950*, 2017. 1

[12] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 3

[13] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022. 2, 3, 4

[14] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context. In *ECCV*, 2014. 3

[15] Ilya Loshchilov and Frank Hutter. Sgdr: Stochastic gradient descent with warm restarts. In *ICLR*, 2017. 3

[16] 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:2104.08860*, 2021. 1, 3

[17] Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021. 2, 3, 4

[18] Joe Yue-Hei Ng, Matthew Hausknecht, Sudheendra Vijayanarasimhan, Oriol Vinyals, Rajat Monga, and George Toderici. Beyond short snippets: Deep networks for video classification. In *CVPR*, 2015. 1

[19] 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. In *ICML*, 2021. 1, 2, 3, 4

[20] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI blog*, 2019. 2, 3

[21] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. LAION-400M: open dataset of clip-filtered 400 million image-text pairs. In *Data Centric AI NeurIPS Workshop*, 2021. 3

[22] 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 the 56th Annual Meeting of the Association for Computational Linguistics*, 2018. 3

[23] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3D convolutional networks. In *ICCV*, 2015. 1

[24] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv:1807.03748*, 2018. 2

[25] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 3

[26] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks for action recognition in videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(11):2740–2755, 2019. 3

[27] Mengmeng Wang, Jiazheng Xing, and Yong Liu. Actionclip: A new paradigm for video action recognition. *arXiv:2109.08472*, 2021. 1

[28] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. OFA: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *ICML*, 2022. 3, 4

[29] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. MSR-VTT: A large video description dataset for bridging video and language. In *CVPR*, 2016. 3

[30] Ceyuan Yang, Yinghao Xu, Bo Dai, and Bolei Zhou. Video representation learning with visual tempo consistency. *arXiv:2006.15489*, 2020. 2