Supplementary Material of "FitCLIP: Refining Large-Scale Pretrained Image-Text Models for Zero-Shot Video Understanding Tasks"

Santiago Castro*[1]
https://santi.uy/
Fabian Caba Heilbron[2]
https://fabiancaba.com/

[1] University of Michigan
[2] Adobe Research

1 Pretraining Datasets

Table 1a summarizes existing datasets for pretraining visual-language models. CC3M [8] is one of the first datasets to bridge images with natural language supervision leveraging the internet (HTML image alt texts). This dataset collects about 3M clean images through a pipeline that warranty clean supervision signal. The MS COCO Captions [3] (COCO) dataset contains 500k human-curated caption-image pairs. The images come from the MS COCO [4] dataset, which in turn were collected from Flickr. WIT [10] contains 37.5M image-caption pairs obtained from the Wikipedia. CLIP authors [7] constructed dataset that contains more than 400M text-image pairs scrapped from the internet. The dataset contains images retrieved from queries formed with the 1000 most common visual concepts in Wikipedia. While the dataset does not rely on manual cleaning to verify the image-text pairs, it is assumed that a person provided a good enough image caption before uploading the image to the internet. In the same spirit, the WebVid-2.5M dataset [2] crawls 2.5M text-video pairs leveraging manually-curated titles from Stock footage. Differently, the HowTo100M (HT100M) dataset [5] contains 100M pairs of noisy aligned video-text pairs. In this dataset, the video-text pairs come from long YouTube videos and their automatically transcribed speech.

2 FitCLIP *vs*. CLIP per-class performance

Previous experiments showed that FitCLIP offers a simple strategy to boost zero-shot performance in video understanding tasks; however, where are those improvements emerging from? To understand better the differences between FitCLIP and CLIP (which is also our teacher), we compute the performance difference per class, between both models, in the Moments in Time dataset. Figure 1 summarizes the results by plotting the largest and smallest

*Work done as an intern at Adobe Research.

Dataset	Domain	Supervision	Size
COCO [8]	Images	Clean	600k
CC3M [8]	Images	Clean	3M
WIT [8]	Images	Clean	37.5M
CLIP [8]	Images	Weak	400M
WebVid [8]	Videos	Weak	2.5M
HT100M [8]	Videos	Noisy	100M

(a) **Pretraining datasets**

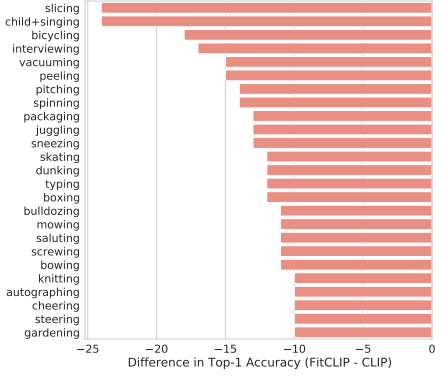
Dataset	# Classes	# Samples
MiT [8]	339	33,900
UCF101 [8]	101	1,794

(b) **ZS action recognition**

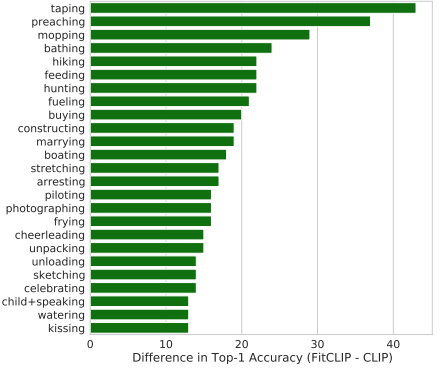
Dataset	# Samples	Genre
MSR-VTT [8]	1000	UGC
YouCook2 [8]	3305	Cooking
DiDeMo [8]	4021	UGC

(c) **ZS text-to-video retrieval**

Table 1: **Pretraining and Zero-shot Datasets.** (a) Diverse image and video datasets are available for pretraining visual-language models. (b) We benchmark zero-shot (ZS) action recognition in two popular datasets. MiT denotes Moments in Time [8]. (c) To benchmark zero-shot (ZS) text-to-video retrieval, we rely on three well-established datasets. UGC stands for user-generated content, and Genre refers to the type of videos in the dataset.



(a) **Bottom-25 Classes**



(b) **Top-25 Classes**

Figure 1: **FitCLIP *vs.* Teacher per-class improvements.** The plots show the per-class difference between FitCLIP and CLIP performances (Top-1) on the Moments in Time (MiT) dataset. Noticeably, the performance difference varies significantly across various action classes, which reinforce our intuition that FitCLIP encodes complementary video information compared to CLIP. Interestingly, FitCLIP improves performance for abstract action classes such as *preaching and tapping*, while CLIP does so for actions involving common actions like *cycling, boxing, or skating*.

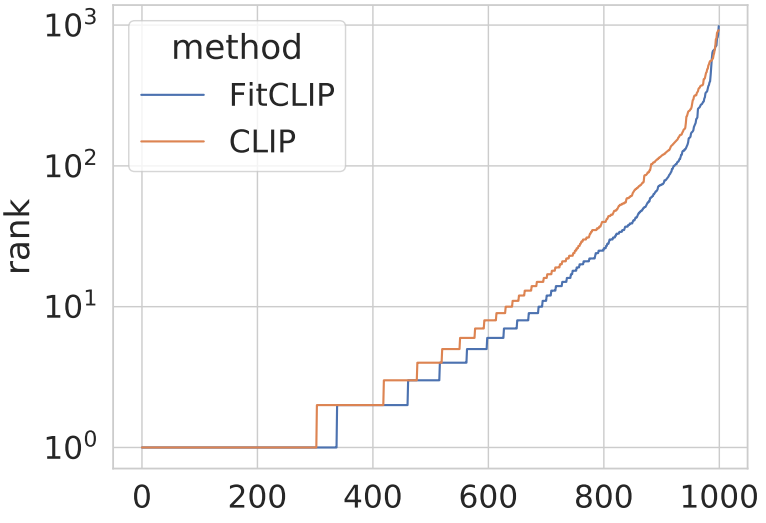


Figure 2: **FitCLIP vs. CLIP distribution of Text-to-Video Retrieval rankings.** The x-axis represents each text in the MSR-VTT validation set (1K-A split) and the y-axis (in log scale) represents the rank each model gave to the corresponding video. The x-axis is sorted by rank (the values increase).

(including actions with worst performance) 25 changes in performance. First, we observe that the performance accuracy (Top-1) of several classes changes drastically. This validates our hypothesis that the Student provides FitCLIP with complementary information concerning the knowledge CLIP (the Teacher) already provides. Interestingly, FitCLIP obtains, overall, better performance for abstract action classes such as *preaching and taping*. On the contrary, CLIP tends to do better for common actions often captured in photographs such as *skating, or boxing*.

3 FitCLIP *vs*. CLIP ranking distributions

The Text-to-Video Retrieval results show that FitCLIP outperforms CLIP at multiple points of this zero-shot setting. However, it is not clear how the methods behave for the rest of them. Figure 2 shows the distribution of rankings for the validation set of MSR-VTT for both methods. We can see that FitCLIP is under the CLIP curve for virtually all points. FitCLIP ranks the videos better for this dataset, regardless of the cutting point.

4 Frozen in Time Variants

Tables 2 and 3 show the results on zero-shot action recognition and text-to-video retrieval for Frozen in Time [2] on different pre-training datasets. These pre-trained checkpoints are provided by the authors[1]. They use different combinations of Conceptual Captions [3]

[1] https://github.com/m-bain/frozen-in-time#-pretrained-weights

Dataset	Top 1	Top 5
WebVid	11.4	27.2
CC3M+WebVid	13.2	29.3
CC3M+WebVid+COCO	**14.0**	**31.8**

(a) **Moments in Time (MiT)**

Dataset	Top 1	Top 5
WebVid	36.9	61.1
CC3M+WebVid	49.2	61.1
CC3M+WebVid+COCO	**51.9**	**76.1**

(b) **UCF101**

Table 2: Zero-shot action recognition results of Frozen in Time [2] pre-trained on different datasets.

(CC3M), WebVid [2], and Microsoft COCO Captions [3] (COCO). Combining the three of them presents the best results. However, note COCO Captions were obtained using an expensive data collection procedure and are richly annotated while the other two datasets were obtained from data available on the internet and thus have weaker annotations.

5 Impact of Fusing the Teacher-Student Knowledge

Tables 4 and 5 present all the metrics for the results on the impact of our method on zero-shot action recognition and zero-shot text-to-video retrieval. Overall, FitCLIP presents the best results. We highlight the importance of fusing the knowledge of the teacher and the student as they individually perform worse than in combination.

6 Alpha Value

We analyze the effect of changing the value of α necessary for the weight-space ensembling step when fusing the teacher and student knowledge in our method. Figure 3 shows the effect of this hyperparameter by varying it from 0 to 1, with increments of size 0.1, where 0 is only the teacher and 1 only the student. We show the results on a different split from the training distribution (Figure 3a) and on the other datasets we have reported throughout this paper (Figure 3b). For WebVid, the best value we obtain is when $\alpha = 0.3$. Still, we decided to use $\alpha = 0.4$, which is close enough and the best value obtained by [11]. For the other datasets, the best value we obtain is when $\alpha = 0.2$. For $\alpha = 0.4$ the score is still high.

7 Impact of the Labeled Data Size

The more labeled data for training typically implies the better results. However, more training implies the obtained checkpoint in the weight landscape to be further away from the point of origin and thus harder for weight-ensembling to work well. We study the impact of the labeled data size and try to find a good trade-off point. Figure 4 show the results of preliminary experiments which are performed by fine-tuning on different subset sizes of the training set from WebVid and applying weight-space ensembling (without distillation).

Dataset	R@1	R@5	R@10	MdR
WebVid	12.9	31.0	41.2	16
CC3M+WebVid	17.1	39.1	49.6	11
CC3M+WebVid+COCO	**21.3**	**43.6**	**55.9**	**7**

(a) **MSR-VTT**

Dataset	R@1	R@5	R@10	MdR
WebVid	1.1	4.2	6.8	329
CC3M+WebVid	2.7	9.5	14.2	162
CC3M+WebVid+COCO	**3.2**	**10.1**	**16.2**	**135**

(b) **YouCook2**

Dataset	R@1	R@5	R@10	MdR
WebVid	14.5	34.9	45.4	14
CC3M+WebVid	20.3	42.7	53.5	9
CC3M+WebVid+COCO	**23.2**	**45.8**	**56.8**	**7**

(c) **DiDeMo**

Table 3: Zero-shot text-to-video retrieval results of Frozen in Time [2] pre-trained on different datasets.

Dataset	Top 1	Top 5
Teacher (CLIP)	19.9	40.3
Student	17.7	39.1
FitCLIP	**21.8**	**44.6**
△	↑ 1.9	↑ 4.3
Error rate reduction	↑ 2.4	↑ 7.2

(a) **Moments in Time (MiT)**

Dataset	Top 1	Top 5
Teacher (CLIP)	**74.5**	94.3
Student	64.7	90.4
FitCLIP	73.3	**95.3**
△	↓ 1.2	↑ 1.0
Error rate reduction	↓ 4.7	↑ 17.5

(b) **UCF101**

Table 4: **Impact of fusing teacher-student knowledge on zero-shot action recognition.** △ denotes the absolute difference in performance between FitCLIP and the Teacher model.

Dataset	R@1	R@5	R@10	MdR
Teacher (CLIP)	30.4	55.1	64.1	4
Student	28.1	52.6	63.7	4
FitCLIP	**33.8**	**59.8**	**69.4**	**3**
△	↑ 3.4	↑ 4.7	↑ 5.3	↑ 1
Error rate reduction	↑ 4.9	↑ 10.5	↑ 14.8	↑ 25.0%

(a) **MSR-VTT**

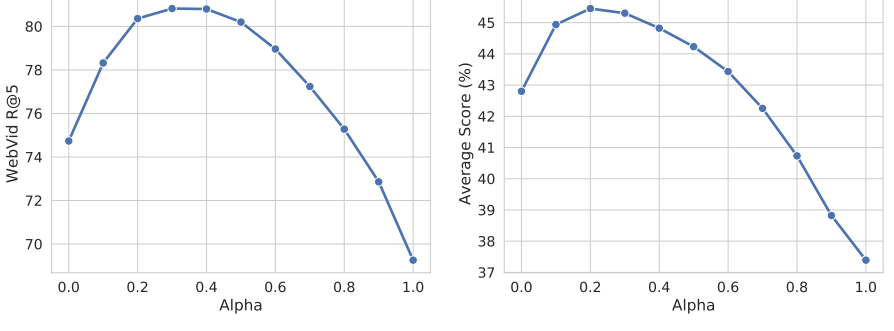
Dataset	R@1	R@5	R@10	MdR
Teacher (CLIP)	5.3	14.6	20.9	94
Student	2.9	9.7	14.1	159
FitCLIP	**5.8**	**15.5**	**22.1**	**75**
△	↑ 0.5	↑ 0.9	↑ 1.2	↑ 19
Error rate reduction	↑ 0.5	↑ 1.1	↑ 1.5	↑ 20.2%

(b) **YouCook2**

Dataset	R@1	R@5	R@10	MdR
Teacher (CLIP)	26.2	49.9	60.6	5
Student	20.7	42.4	54.0	8
FitCLIP	**28.5**	**53.7**	**64.0**	**4**
△	↑ 2.3	↑ 3.8	↑ 3.4	↑ 1
Error rate reduction	↑ 3.1	↑ 7.6	↑ 8.6	↑ 20.0%

(c) **DiDeMo**

Table 5: **Impact of fusing teacher-student knowledge on zero-shot text-to-video retrieval.** △ denotes the absolute difference in performance between FitCLIP and the Teacher model. To measure the error rate reduction for the median rank, we directly use its reduction rate.



(a) **Supervised WebVid** (b) **Zero-shot average across 5 datasets**

Figure 3: **Impact of changing the value of weight-ensembling α value when fusing the teacher and the student.** We report (a) supervised text-to-video retrieval WebVid R@5 (recall we trained on this domain) and (b) an average across 5 other datasets. The zero-shot text-to-video retrieval datasets used are DiDeMo, MSR-VTT, and YouCook2 (R@5). The zero-shot action recognition datasets used are Moments in Time and UCF-101 (top-1 accuracy). The average value across these datasets is shown.

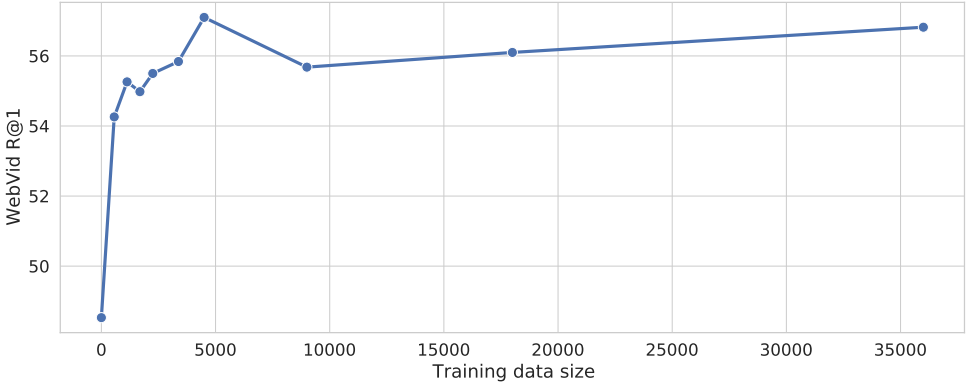


Figure 4: **Text-to-Video top-1 recall on WebVid-2.5M (supervised) of different training subset sizes when fine-tuning CLIP ViT-B/16 and then applying weight-space ensembling.** The evaluated subset sizes are: 0, 563, 1125, 1688, 2250, 3375, 4500, 9000, 18000, and 36000. The subset size 0 represents the evaluation of the pre-trained model without fine-tuning. We exclude large values as we have observed a great drop in performance. Note this experiment doesn't employ distillation.

Each of these subsets where sampled from the whole dataset (they are unlikely subsets of each other). We find the best value when the WebVid-2.5M training subset size is 4500. We recognize that we are indirectly using other parts of WebVid, which can boost the in-distribution performance of the selected subset. However, note this doesn't imply better out-of-distribution performance. We skip showing results for large values as we have observed a great drop in performance. In particular, we obtained results that are considerably worse than the pre-trained model when using the whole training set (2.5M).

8 Share of Pseudo-Labels/Labels

We are interested in comparing the effect of applying weight-ensembling to a distilled model with applying it to a model that has been trained only on labeled data. Figure 5 shows the effect of varying the proportion of the labeled loss in the final loss in our zero-shot benchmarks. The use of the distillation loss with $\lambda = 10^{-4}$ outperforms the usage of only the labeled loss in YouCook2 and UCF101 and shows similar performance on MSR-VTT. In contrast, The performance on DiDeMo and Moments in Time seems to be better with using only the labeled loss. We hypothesize our method is especially useful on datasets whose distribution (*e.g.*, YouCook2) is more distant from the training-time dataset (WebVid-2.5M).

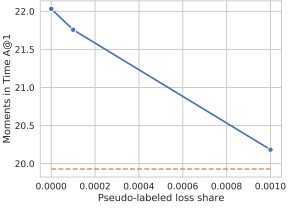
References

[1] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017.

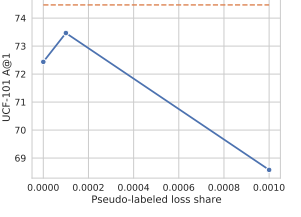
[2] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in Time: A joint

	Action Recognition		Text-to-video Retrieval		
	UCF101	MiT	MSR-VTT	YouCook2	DiDeMo
CLIP	74.5	19.9	55.1	14.6	49.9
WiSE-FT	72.5	**22.0**	**59.9**	15.1	**55.4**
FitCLIP	**73.3**	21.8	59.8	**15.5**	53.7

Table 6: **Importance of the Pseudo-Labels.** We report the top-1 accuracy for the zero-shot action recognition datasets, and the top-5 recall for the zero-shot text-to-video retrieval ones. We show in bold the best results between WiSE-FT and FitCLIP for each dataset.

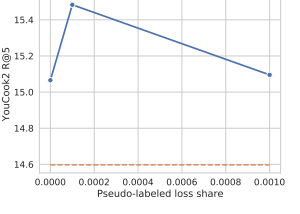


(a) **Moments in Time**

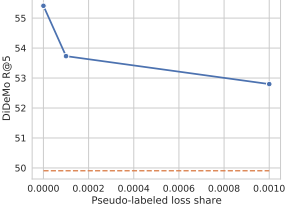


(b) **UCF101**

(c) **MSR-VTT**



(d) **YouCook2**



(e) **DiDeMo**

Figure 5: **The effect on the zero-shot performance of the share of the pseudo-labeled and labeled losses in FitCLIP.** Each plot shows how the proportion of the pseudo-labeled loss (x-axis) affects the zero-shot performance on a given dataset. The dashed orange line shows the performance of CLIP, as a reference. We skip the sampled values greater than 0.01 to better visualize the plots since they tend to bring a worse performance.

video and image encoder for end-to-end retrieval. In *IEEE International Conference on Computer Vision*, 2021.

[3] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

[4] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing. ISBN 978-3-319-10602-1.

[5] 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 (ICCV)*, October 2019.

[6] Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfreund, Carl Vondrick, and Aude Oliva. Moments in time dataset: One million videos for event understanding. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 42(2):502–508, 2020. doi: 10.1109/TPAMI.2019.2901464.

[7] 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 Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 18–24 Jul 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

[8] 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 (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1238. URL https://aclanthology.org/P18-1238.

[9] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *CRCV-TR-12-01*, November 2012.

[10] Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. *WIT: Wikipedia-Based Image Text Dataset for Multimodal Multilingual Machine Learning*, page 2443–2449. Association for Computing Machinery, New York, NY, USA, 2021. ISBN 9781450380379. URL https://doi.org/10.1145/3404835.3463257.

[11] Mitchell Wortsman, Gabriel Ilharco, Mike Li, Jong Wook Kim, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. Robust fine-tuning of zero-shot models. *arXiv preprint arXiv:2109.01903*, 2021.

[12] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. MSR-VTT: A large video description dataset for bridging video and language. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[13] 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. URL https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17344.