MAtch, eXpand and Improve: Unsupervised Finetuning for Zero-Shot Action Recognition with Language Knowledge
Supplementary

Wei Lin$^{\dagger 1}$ Leonid Karlinsky^{2} Nina Shvetsova^{3} Horst Possegger^{1}
Mateusz Kozinski^{1} Rameswar Panda^{2} Rogerio Feris^{2} Hilde Kuehne^{2,3}
Horst Bischof^{1}

^{1}Institute of Computer Graphics and Vision, Graz University of Technology, Austria
^{2}MIT-IBM Watson AI Lab, USA
^{3}Goethe University Frankfurt, Germany

For further insights into our approach MAXI, we introduce more dataset statistics (Sec. 1) and implementation details (Sec. 2) of MAXI.

In the additional results, we provide comparison of visualizations of attention heatmaps across several approaches in Sec. 3.1. Furthermore, we report more results of finetuning with noisy action dictionary (Sec. 3.2), and provide more examples of language sources used for training (Sec. 3.3). Lastly, we explore a cross-frame attention temporal module in Sec. 3.4.

1. Dataset Statistics

Kinetics-400 (K400) [8] is the most popular benchmark for action recognition tasks, containing around 240K training videos in 400 classes. The dataset consists of YouTube videos with an average length of 10 seconds. We use the training set of K400 for finetuning CLIP.

UCF101 [17] is collected from YouTube videos, consisting of 13K videos from 101 classes. There are three splits of training data (∼ 9.4K) and validation data (∼3.6K). Following XLCIP [13] and ViFi-CLIP [15], we report the average performance on the three validation splits.

HMDB51 [9] consists of 7K videos comprised of 51 action classes, collected from YouTube videos and movie clips. There are three splits of training data (∼ 3.5K, 70 videos per class) and validation data (∼1.5K, 30 videos per class). Following [13, 15], we report the average performance on the three validation splits.

Kinetics-600 (K600) [3] is an extension of K400, consisting of 650K videos in 600 classes. Following [5, 13, 15], we use the 220 classes[1] that are not included in K400 for zero-shot action recognition. There are three validation splits, each containing 160 classes randomly sampled from these 220 classes. We report the average performance on the three validation splits, each containing around 14K videos.

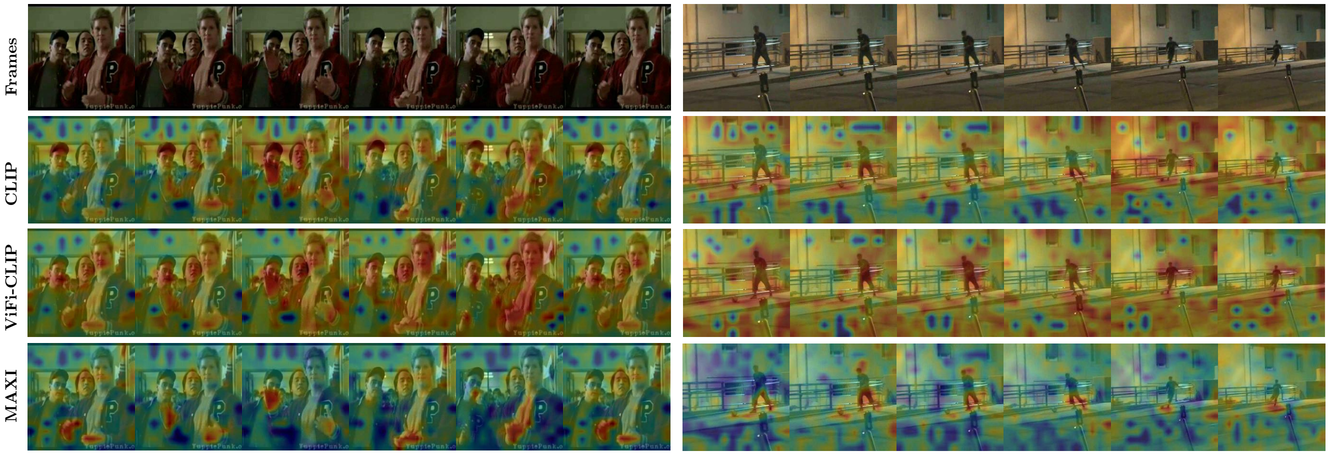
MiniSSv2 [4] (87 classes, 93K videos) is a subset of Something-Something v2 (SSv2) [7] (174 classes, 220K videos). SSv2 is an egocentric motion-based action dataset, which has a large visual domain shift to K400. Furthermore, the action classes are detailed descriptions of fine-grained movements, in a largely different language style than the K400 action dictionary, *e.g. Failing to put something into something because something does not fit*, and *Lifting a surface with something on it but not enough for it to slide down*. For zero-shot action recognition, we evaluate on the validation split of MiniSSv2 (12K videos). For few-shot action recognition, we follow [15] and evaluate on the validation split of SSv2 (25K videos).

Charades [16] is a long-range activity dataset recorded by people in their homes based on provided scripts for home activities. There are ∼10K videos in 157 classes. The average video length is 30 seconds. Each video has annotations of an average of 6.8 action instances, often in complex co-occurring cases. The validation split consists of 1.8K videos. We report the mean Average Precision (mAP) for the multi-label classification task.

Moments-in-Time (MiT) [12] is a large-scale action dataset of 3-second YouTube video clips, which cover actions in 305 classes, performed by both humans and animals. The validation split consists of 30K videos.

UAV Human (UAV) [11] is an action dataset recorded with an Unmanned Aerial Vehicle in unique camera viewpoints. There are 155 action classes. Actions in different categories are performed by a fixed group of subjects in the same background scenes. This leads to an extremely low

[1] In the evolution from K400 to K600, there are renamed, removed and split classes. See details in Appendix B in [5].



(a) clap

(b) kick ball

Figure 3: Attention heatmaps on actions which have a verb form (lemma or gerund) directly included in the K400 dictionary. We compare among CLIP (2nd row), ViFi-CLIP (3rd row) and our MAXI (4th row). Warm and cold colors indicate high and low attention. MAXI has more focused attention on hands (for *clap*) and legs (for *kick ball*).

object-scene bias and a large shift to the domain of K400 and CLIP. We evaluate on the RGB videos and report the average performance on the two official validation splits, each consisting of $\sim$ 6.2K videos.

2. Implementation Details

In addition to the details mentioned in the main manuscript, we cover more implementation specifics here.

CLIP matching. The CLIP matching step is for consuming the language source of the predefined action dictionary D. We use CLIP[2] [14] with the ViT-B/16 visual encoder [6] to match each video with texts in the predefined action dictionary. To improve the matching quality for Text Bag Construction, we perform prompt ensembling over the 28 prompt templates[3] which are proposed by CLIP for Kinetics videos. Important to note, during inference we follow the exact protocol of ViFi-CLIP [15] and use only a single prompt.

GPT-3 text expansion. We employ the GPT-3 `text-davinci-003` model [2]. We set the temperature to 0.4. We generate 5 verb phrases using the input instruction - *Generate 5 phrases to describe the action of <action> in simple words*. Here for a video x_i, *<action>* is the best matched text $\hat{t}_i$ from the predefined action dictionary.

BLIP captioning. We use BLIP model [4] [10] with ViT-L/16 as the image captioner. For each video, the image captioning is performed on 8 uniformly sampled frames. The frames are resized into 384.

Text augmentation. We use the natural language processing tool spaCy[5] to parse the verbs and verb phrases from the descriptions. We perform augmentation by converting the verbs into forms of lemma and gerund (present participle) and include results in the text bag.

Training. We employ CLIP with the ViT-B/16 visual encoder. We follow the full-finetuning configuration of [15] to finetune both the visual and text encoder. During training, we follow the configuration of [15, 13] for visual augmentation of multi scale crop, random flipping, color jitering and gray scaling. We do not perform augmentations of MixUp or CutMix.

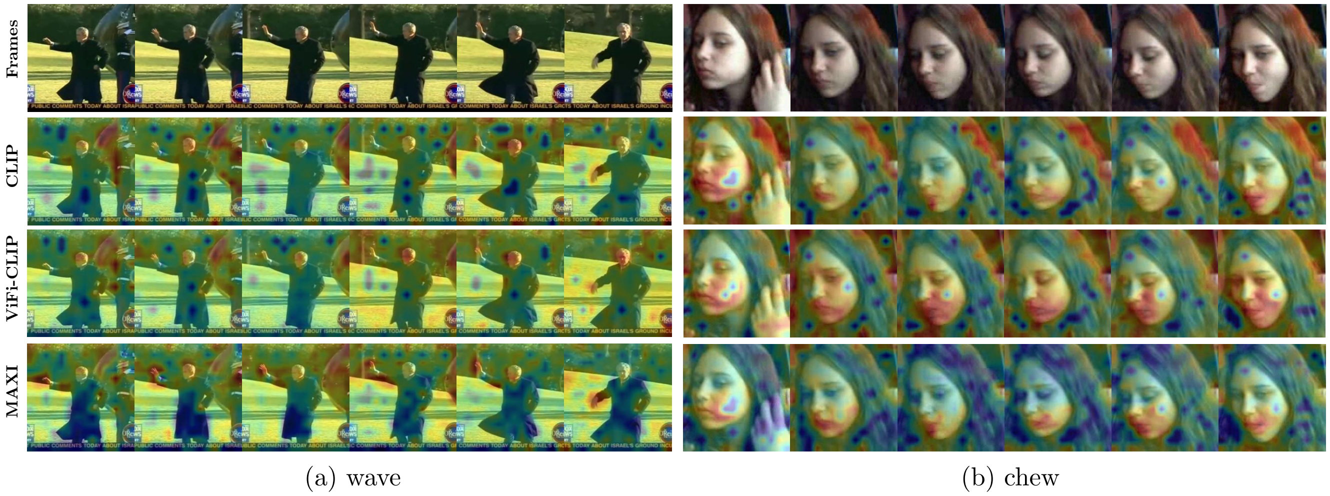
As different videos have varying numbers of texts in their bags, we randomly sample N_{bag} texts from the originally constructed bag in each training iteration. For multiple instance learning, we use all the N_{bag} words in a text bag to form N_{bag} text prompts for each video. The text prompt is in the format of *<text1>* + *<text2>*. The first part *<text1>* is uniform for all the N_{bag} text prompts. Specifically, we use a hand-crafted prompt template *a photo of <action>*, where *<action>* is the best-matched text $\hat{t}_i$ from the predefined action dictionary (see Eq. 1 in the main manuscript). *<text2>* is an individual text from the text bag. To avoid duplication, we do not use $\hat{t}_i$ as *<text2>*.

Inference. We follow [13, 15] and sample a single view via sparse temporal sampling and spatial center crop. The same single prompt template is used in inference.

[2] CLIP model source

[3] https://github.com/openai/CLIP/blob/main/data/prompts.md

[4] BLIP model source



(a) wave (b) chew

Figure 4: Attention heatmaps on novel actions which do not have any verb form included in the K400 dictionary. We compare among CLIP (2nd row), ViFi-CLIP (3rd row) and our MAXI (4th row). Warm and cold colors indicate high and low attention. MAXI has more focused attention on hand and arm for *wave*, and on the area of mouth for *chew*.

3. Additional Results

3.1. Attention Heatmaps

To gain more insights into the performance improvement of MAXI, we compare the visualizations of attention heatmaps across several approaches in Fig. 3, Fig. 4 and Fig. 5. *CLIP* is the original CLIP [14] without any finetuning. *ViFi-CLIP* [15] finetunes CLIP via supervised classification on K400 with ground truth annotations. *MAXI* is our approach of unsupervised finetuning with language knowledge.

We obtain the attention maps by computing the cosine similarity between the patch token features from the visual encoder and the text feature from the text encoder. We visualize the attention maps in several action classes from the downstream datasets used for the zero-shot action recognition task. Based on the relationship between the zero-shot action class and the K400 action dictionary used for training, we categorize the visualizations into 3 groups: (1) In-dictionary action classes which have a verb form (lemma or gerund) directly included in the K400 action dictionary, *e.g*. *clap* and *kick ball* in Fig. 3; (2) Novel actions classes which do not have any verb form included in the K400 action dictionary, *e.g*. *wave* and *chew* in Fig. 4; (3) General actions whose verb form is a basic component of several actions in the K400 action dictionary, *e.g*. *catch*, *hit*, *jump* and *run* in Fig. 5.

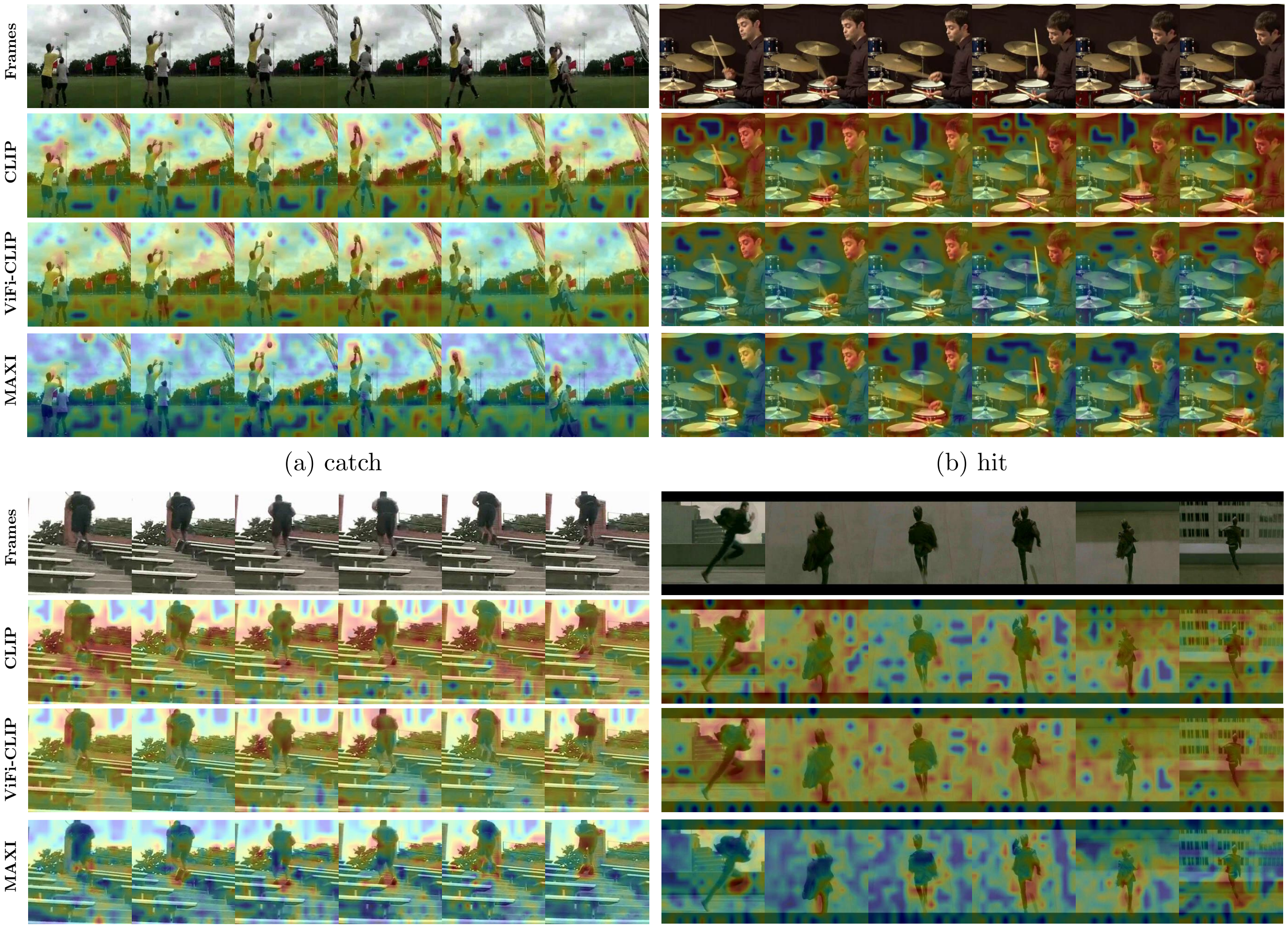
In-dictionary action classes. In Fig. 3, we visualize two samples of the action *clap* and *kick ball*. *clap* has the same lemma as *clapping* in the K400 dictionary, while *kick ball* has related actions of *kicking field goal* and *kicking soccer ball* in the K400 dictionary. We see that CLIP has incorrectly high attention on object (Fig. 3(a), 2nd row) or background scene (Fig. 3(b), 2nd row). ViFi-CLIP has cluttered high attention on both the subjects and the background scenes. On the contrary, MAXI has more focused attention on the hands (for *clap*) and legs (for *kick ball*).

In our GPT-3 text bag of *clapping*, related words such as *clap*, *smacking hands*, *slapping palms* and *clapping hands* are included. This strengthens the association between the action *clap* and the body part of hands, and leads to more accurate attention. Furthermore, in BLIP caption verb text bags, the verb *clap* appears several times in frame captions of K400 videos of *clapping*, *giving or receiving award* and *applauding*. This further improves the understanding of *clap*. Similarly, in BLIP frame captions, *kick* is an even more basic verb with large amount of occurrences.

Novel action classes. In Fig. 4, we compare the attention maps for the novel verbs *wave* and *chew* that do not appear in the K400 action dictionary. We see that for *wave*, CLIP and ViFi-CLIP have attention on the background scene or on the head, while MAXI has correct attention on the hand and arm. For *chew*, CLIP has more attention on the hair and ViFi-CLIP has attention on a large area of the face. On the contrary, MAXI has consistent focused attentions on the area of the mouth where the action *chew* happens.

The verb *wave* appears in BLIP caption verb text bags of several K400 videos of *clapping*, *applauding*, *celebrating*. The verb *chew* appears in captions of K400 videos of *eating carrots*, *eating spaghetti*, *eating watermelon* and *baby waking up*. The additional language source improves the knowledge of actions that never appear in the K400 action dictionary.

[5] spaCy https://spacy.io/



(a) catch

(b) hit

(c) jump

(d) run

Figure 5: Attention heatmaps on actions which have a verb form (lemma or gerund) directly included in the K400 dictionary. We compare among CLIP (2nd row), ViFi-CLIP (3rd row) and our MAXI (4th row). Warm and cold colors indicate high and low attention. MAXI has more concentrated attention on the part where the action happens, *e.g*. catching ball with hands (Fig. 5(a), 4th row), hitting drum with stick (Fig. 5(b), 4th row), legs and feet jump on stairs (Fig. 5(c), 4th row), and attention on the running body (Fig. 5(d), 4th row).

General actions. In Fig. 5, we illustrate the attention maps for four general verbs *catch*, *hit*, *jump* and *run*. These verbs are basic components of several actions in the K400 dictionary, *e.g*. *catching fish*, *catching or throwing frisbee*, *hitting baseball*, *jumping into pool* and *running on treadmill*. In these samples, CLIP and ViFi-CLIP have cluttered attention on the background scene or objects. MAXI has more concentrated attention on the part where the action happens, *e.g*. catching ball with hands (Fig. 5(a), last row), hitting drum with stick (Fig. 5(b), last row), legs and feet jump on stairs (Fig. 5(c), last row), and attention on the running body (Fig. 5(d), last row).

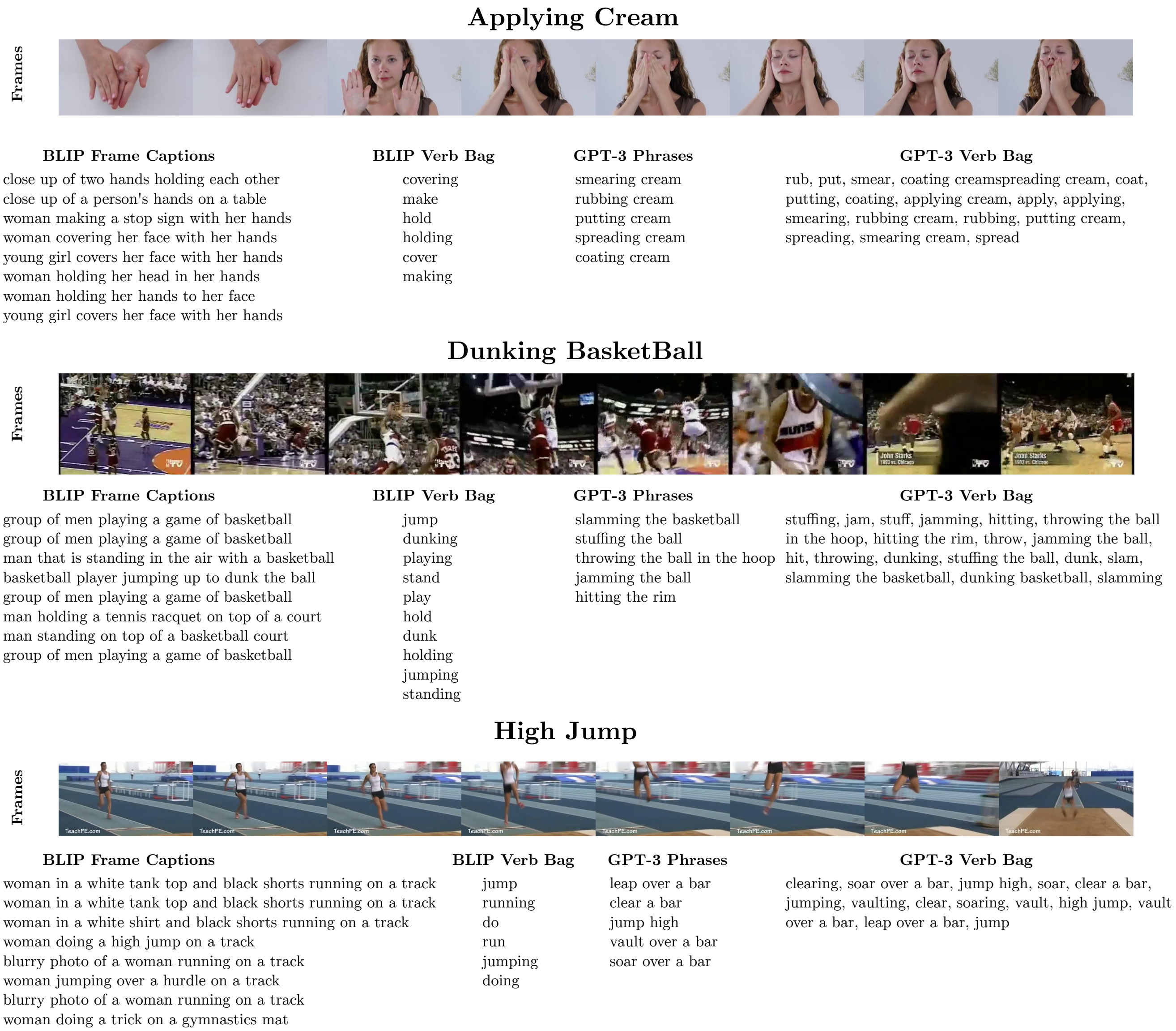
These verbs are very general and could have highly diverse instantiations. *E.g*. *hit* (drum) in Fig. 5(b) is not close to *hitting baseball* on K400. *jump* (on stairs) in Fig. 5(c) is not close to *jumping into pool* or *bungee jumping* on K400, even if they share the same verb. In our GPT-3 verb bag and BLIP caption verb bag, there is a large amount of these verb instances that facilitate the comprehensive understanding of these general verbs. This leads to better focus even in unusual complex scenes, *e.g*. jumping on stairs (Fig. 5(c)).

3.2. Robustness Against Noisy Action Dictionary

In Table 5 in the main manuscript, we explored the robustness of our finetuning pipeline against noisy action dictionaries. In case of an over-specified dictionary, we added noisy verbs and verb phrases into the original K400 action dictionary. The noisy verbs are parsed from the captions in the WebVid2.5M dataset [1]. Here we further increase the ratio of noisy verbs, and add 800 and 1200 verbs into the

Action dictionary	dictionary size	UCF101	HMDB51	K600	MiniSSv2	Charades	UAV Human	Moments-in-time
CLIP [14] (w/o finetune) Zero-Shot		69.93 / 92.7	38.02 / 66.34	63.48 / 86.80	3.96 / 14.42	19.80	1.79 / 7.05	20.11 / 40.81
K400	400	**78.18 / 96.03**	**50.35 / 77.10**	**70.78 / 92.17**	**5.74 / 17.70**	**23.89**	**3.06 / 9.46**	**22.41 / 45.83**
K400+WebVid2.5M	800	75.99 / 96.00	45.97 / 73.94	69.14 / 91.13	4.81 / 15.79	22.67	2.11 / 8.00	20.92 / 43.99
K400+WebVid2.5M	1200	75.72 / 96.02	45.51 / 73.97	69.36 / 91.11	4.21 / 15.15	22.35	2.39 / 7.98	21.29 / 44.33
K400+WebVid2.5M	1600	76.14 / 96.01	44.84 / 71.79	69.23 / 91.10	4.42 / 14.71	22.89	2.14 / 7.71	20.69 / 43.59

Table 9: Robustness of finetuning with noisy action dictionaries. We add noisy verbs parsed from the WebVid2.5M dataset into the original K400 action dictionary. We report the zero-shot transfer performance (mAP on Charades and Top1/Top5 accuracy on other datasets). We set the text bag filtering ratio $p = 50\%$ for improved text bag quality.



Applying Cream

BLIP Frame Captions

a close up of two hands holding each other
a close up of a person's hands on a table
a woman making a stop sign with her hands
a woman covering her face with her hands
a young girl covers her face with her hands
a woman holding her head in her hands
a woman holding her hands to her face
a young girl covers her face with her hands

BLIP Verb Bag

covering
make
hold
holding
cover
making

GPT-3 Phrases

smearing cream
rubbing cream
putting cream
spreading cream
coating cream

GPT-3 Verb Bag

rub, put, smear, coating creamspreading cream, coat, putting, coating, applying cream, apply, applying, smearing, rubbing cream, rubbing, putting cream, spreading, smearing cream, spread

Dunking BasketBall

BLIP Frame Captions

a group of men playing a game of basketball
a group of men playing a game of basketball
a man that is standing in the air with a basketball
a basketball player jumping up to dunk the ball
a group of men playing a game of basketball
a man holding a tennis racquet on top of a court
a man standing on top of a basketball court
a group of men playing a game of basketball

BLIP Verb Bag

jump
dunking
playing
stand
play
hold
dunk
holding
jumping
standing

GPT-3 Phrases

slamming the basketball
stuffing the ball
throwing the ball in the hoop
jamming the ball
hitting the rim

GPT-3 Verb Bag

stuffing, jam, stuff, jamming, hitting, throwing the ball in the hoop, hitting the rim, throw, jamming the ball, hit, throwing, dunking, stuffing the ball, dunk, slam, slamming the basketball, dunking basketball, slamming

High Jump

BLIP Frame Captions

a woman in a white tank top and black shorts running on a track
a woman in a white tank top and black shorts running on a track
a woman in a white shirt and black shorts running on a track
a woman doing a high jump on a track
a blurry photo of a woman running on a track
a woman jumping over a hurdle on a track
a blurry photo of a woman running on a track
a woman doing a trick on a gymnastics mat

BLIP Verb Bag

jump
running
do
run
jumping
doing

GPT-3 Phrases

leap over a bar
clear a bar
jump high
vault over a bar
soar over a bar

GPT-3 Verb Bag

clearing, soar over a bar, jump high, soar, clear a bar, jumping, vaulting, clear, soaring, vault, high jump, vault over a bar, leap over a bar, jump

Figure 6: Examples of video frames, BLIP frame captions, GPT-3 phrases, together with the derived BLIP verb bag and GPT-3 verb bag. The videos are from the K400 dataset.

dictionary, resulting in 1200-class and 1600-class spaces.

In Table 9, we report the zero-shot transfer performance of models finetuned with the resulted 1200-class and 1600-class space. We set the text bag filtering ratio $p = 50\%$ for improved text bag quality. We see that even with extremely noisy dictionary where 50% to 75% of words do

Temp. attention layers	UCF101	HMDB51	K600	MiniSSv2	Charades	UAV Human	Moments-in-time
None	**78.17**	**52.24**	**71.43**	**6.37**	**23.79**	2.72	**22.91**
2	77.38	51.83	70.41	5.98	22.87	**2.90**	22.51
6	75.91	51.92	69.23	6.09	21.78	2.52	22.52

Table 10: Cross-frame temporal attention modules. we report the zero-shot transfer performance after finetuning CLIP on K400. We train with text bags of GPT3 verbs and BLIP verbs. We set the text bag filtering ratio $p = 90\%$. Adding temporal attention module does not lead to performance improvement.

not match with the video data, our finetuning still results in a robust zero-shot transfer performance to unseen datasets. The robustness is the consequence of the fact that we collect knowledge from multiple language sources and learn from them via Multiple Instance Learning. Note that the zero-shot transfer does not have consistent change in performance across the downstream datasets, as different datasets have different language domain shift to the action dictionary used for training.

3.3. Examples of Language Sources

Similar to the *cooking egg* example in Fig. 2 in the main manuscript, we illustrate more examples of video frames, BLIP frame captions, GPT-3 phrases, together with the derived BLIP verb bag and GPT-3 verb bag in Fig. 6. The videos are from the unlabeled K400 dataset which we use for training.

3.4. Parameter-Free Temporal Module

As mentioned in Sec. 3.1 in the main manuscript, we explore a parameter-free temporal-aware module on the CLIP model. We modify the multi head attention module [18] in the visual encoder of CLIP to be temporal aware. Originally, the attention on the frame t is computed via $A_t(Q_t, K_t, V_t) = \text{softmax}\frac{Q_t {K_t}^\top}{d_k} V_t$, where Q_t, K_t and V_t are the query, key and value from frame t.

We explore to compute the cross-frame attention via

$$A'_t(Q_t, K_{t+i}|_{i \in I}, V_t) = \text{softmax}\frac{\sum_{i \in I}(Q_t \cdot {K_{t+i}}^\top)/|I|}{d_k} V_t \quad (1)$$

where we set $I = \{-1, 0, 1\}$. In this case, we use the keys from the frame $t-1$, t and $t+1$ to compute the attention for frame t.

We apply the cross-frame attention on the last 2 and on the last 6 transformer layers in the visual encoder of CLIP. In Table 10, we report the zero-shot transfer performance. We see that in comparison to the variant without any temporal attention module, using cross-frame attention does not lead to performance improvement. K400 is of far smaller scale in comparison to the original CLIP domain. Finetuning from the CLIP model weights with a modified architecture could result in the case that the model drifts far away from the wise CLIP source domain. The results are consistent with the claims in [15] that a sophisticated temporal module does not necessarily lead to performance improvement.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1728–1738, 2021. 4

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *NeurIPS*, 33:1877–1901, 2020. 2

[3] Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A short note about kinetics-600. *arXiv preprint arXiv:1808.01340*, 2018. 1

[4] Chun-Fu Richard Chen, Rameswar Panda, Kandan Ramakrishnan, Rogerio Feris, John Cohn, Aude Oliva, and Quanfu Fan. Deep analysis of cnn-based spatio-temporal representations for action recognition. In *CVPR*, pages 6165–6175, 2021. 1

[5] Shizhe Chen and Dong Huang. Elaborative rehearsal for zero-shot action recognition. In *ICCV*, pages 13638–13647, 2021. 1

[6] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 2

[7] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *ICCV*, pages 5842–5850, 2017. 1

[8] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 1

[9] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *ICCV*, pages 2556–2563. IEEE, 2011. 1

[10] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for uni-

fied vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*, 2022. 2

[11] Tianjiao Li, Jun Liu, Wei Zhang, Yun Ni, Wenqian Wang, and Zhiheng Li. Uav-human: A large benchmark for human behavior understanding with unmanned aerial vehicles. In *CVPR*, pages 16266–16275, 2021. 1

[12] Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfreund, Carl Vondrick, et al. Moments in time dataset: one million videos for event understanding. *TPAMI*, 42(2):502–508, 2019. 1

[13] Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. Expanding language-image pretrained models for general video recognition. In *ECCV*, pages 1–18. Springer, 2022. 1, 2

[14] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 2, 3, 5

[15] Hanoona Rasheed, Muhammad Uzair Khattak, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. Fine-tuned clip models are efficient video learners. *arXiv preprint arXiv:2212.03640*, 2022. 1, 2, 3, 6

[16] Gunnar A Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *ECCV*, pages 510–526. Springer, 2016. 1

[17] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 1

[18] 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*, volume 30, 2017. 6