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

Wei Lin[†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

Abstract

Large scale Vision Language (VL) models have shown tremendous success in aligning representations between visual and text modalities. This enables remarkable progress in zero-shot recognition, image generation & editing, and many other exciting tasks. However, VL models tend to over-represent objects while paying much less attention to verbs, and require additional tuning on video data for best zero-shot action recognition performance. While previous work relied on large-scale, fully-annotated data, in this work we propose an unsupervised approach. We adapt a VL model for zero-shot and few-shot action recognition using a collection of unlabeled videos and an unpaired action dictionary. Based on that, we leverage Large Language Models and VL models to build a text bag for each unlabeled video via matching, text expansion and captioning. We use those bags in a Multiple Instance Learning setup to adapt an image-text backbone to video data. Although finetuned on unlabeled video data, our resulting models demonstrate high transferability to numerous unseen zero-shot downstream tasks, improving the base VL model performance by up to 14%, and even comparing favorably to fully-supervised baselines. The code will be released later at https://github.com/wlin-at/MAXI.

1. Introduction

Vision Language (VL) models [6, 10, 14] have met unprecedented success in unlocking many vision applications [14] to work with potentially unlimited open vocabularies, through the promise of zero-shot transfer [4,9,15,17,23,25–27]. This is empowered by the alignment between visual and language representation spaces, which is effectively attained by VL models leveraging huge amounts of paired image and text data. Incorporating a VL model as a source (base) model or as an architectural component has allowed scaling finetuning on relatively small datasets (e.g. limited in terms of the number of observed objects or other visual concepts compared to the vast VL pretraining) towards zero-shot transfer at inference time. Such zero-shot transfer includes recognizing [23, 25, 26], detecting [4, 17, 27], segmenting [9, 15], and even generating [18] objects unseen during the finetuning stage and only encountered for the first time at the inference stage.

However, despite the progress in zero-shot image tasks, VL models have been observed to underperform when applied to zero-shot action recognition on video data without any finetuning [2,7,12,16,20,21]. A possible reason, as extensively studied in several works [5, 19, 22, 24], is that VL models have a tendency to mostly represent objects (nouns) and not actions (verbs or verb phrases). Therefore, to deal with these shortcomings of VL models w.r.t. zero-shot action recognition, previous works [2, 7, 12, 16, 20, 21] have used datasets with full annotation (e.g. K400 [8]) to finetune VL models (e.g. the most popular CLIP [14]) towards improved video zero-shot recognition performance. The potential downsides of this approach are: (i) reliance on full annotation of large-scale action datasets that is time-consuming and cost-intensive, and (ii) the exposure of the model to only the limited action vocabulary during the supervised finetuning (e.g. 400 actions of K400 vs. over 8K possible single verb actions and much more possible general actions in English language) limiting the performance of zero-shot transfer to unseen action categories. In this context, we propose 'MAtch, eXpand and Improve' (MAXI) – to allow finetuning on completely unlabeled video data (e.g. unlabeled K400 [8]) and a set of language sources, such as unpaired action dictionaries, Large Language Models (LLM) (e.g. GPT-3 [1]), and VL models for matching (e.g. CLIP [14]) and captioning (e.g. BLIP [10]). To this end, MAXI relies on individual bags of potential texts, collected and refined based on the different language

† Correspondence: wei.lin@icg.tugraz.at

sources, that correspond to each video in the unlabeled set. It then applies Multiple Instance Learning (MIL) for finetuning the VL model using those bags. We extensively evaluate MAXI on seven downstream zero-shot and few-shot transfer action recognition benchmarks completely unseen during training. We show that MAXI is effective in leveraging unlabeled video data, not only significantly (up to 14%) improving the source VL model performance on all of those tasks, but also favorably competing with state-of-the-art supervised methods trained on fully supervised counterparts of the same finetuning data, and even improving upon them in some zero-shot and few-shot action recognition transfer tasks.

Our contributions are as follows: (i) we propose MAXI, an approach that leverages an unlabeled video collection and a set of language sources to improve downstream zero-shot action recognition; (ii) we propose to match each unlabeled video with *text bags* of knowledge mined from the language sources, and employ Multiple Instance Learning for finetuning a VL model using these text bags; (iii) we extensively evaluate our approach on seven unseen action recognition benchmarks, and demonstrate up to 14% absolute zero-shot performance improvements over the source VL model, and even outperform baseline models trained in a fully supervised manner on the same data.

2. Method

In this work, we propose an approach that effectively leverages a collection of unlabeled videos and a predefined action dictionary (a potentially noisy collection of possible action text labels) to finetune the CLIP model without any ground truth annotations. The purpose of finetuning is to adapt CLIP to video data and to facilitate subsequent Zero-Shot (ZS) transfer to video recognition tasks on novel video categories which are not seen during training. We denote the predefined action dictionary as D, and the unlabeled video collection as $V = \{x_j | j \in I\}$, with an index set $I = \{1, ..., N_V\}$.

Our pipeline is illustrated in Fig. 1. We first adapt the CLIP image encoder to a video encoder for deployment on video data (Sec. 2.1). Second, given the unlabeled video collection V and a predefined action dictionary D, we use different language sources to construct a *text bag* for each video (Sec. 2.2). The text bag is a (noisy) collection of texts that potentially correspond to the video contents. Third, we perform Multiple Instance Learning (MIL) to learn from the unlabeled videos and noisy text bags (Sec. 2.3), which allows to robustly finetune CLIP in an unsupervised manner.

2.1. CLIP on Video Data

CLIP [14] consists of a visual encoder $\phi_v(\cdot;\theta_v)$ and a text encoder $\phi_t(\cdot;\theta_t)$. We aim to adapt the CLIP image encoder for processing videos. It is demonstrated in [16] that frame-level processing on CLIP image encoder with feature pooling helps in implicitly modeling the temporal cues. This also leads to improved performance over related approaches that additionally incorporate learnable spatio-temporal components. Therefore, following [16], given a video x, we pass M frames into the visual encoder and compute the average of frame features as the video representation, *i.e.* $z_v = \sum_m \phi_v(x_m^F;\theta_v)/M$. An advantage of this paradigm is that the network can be initialized directly from a large-scale pretrained VL model (e.g. CLIP pretrained on 400M web image-text pairs [14]) without adding any randomly initialized parameters. This provides a good starting point with reasonable initial performance before finetuning. We also explore extending a non-randomly-initialized-parameters paradigm to include, e.g., a parameter-free temporal-aware module (see supplementary), confirming [16] that a sophisticated temporal module does not lead to better video adaptation from CLIP.

During inference, given a set of class prompts $C = \{t_c|_{c=1}^{N_C}\}$, the text feature is computed as $z_{t_c} = \phi_t(t_c;\theta_t)$. For simplicity, we denote the L2-normalized video feature and text feature as $z_v = \bar{\phi}_v(x)$ and $z_t = \bar{\phi}_t(t)$. The zero-shot classification is performed by selecting the class prompt with the maximum similarity to the video representation, *i.e.*, $\hat{c} = \arg\max_c \bar{\phi}_v(x)^\top \bar{\phi}_t(t_c)$.

2.2. Text Bag Construction

Given an unlabeled video collection V and a predefined action dictionary D (where each item is a short sentence or a verb phrase describing an action, see Fig. 1), we construct a text bag T_i for each video $x_i \in V$, *i.e.* a noisy collection of text prompts describing the video contents.

Predefined action dictionary. In a practical scenario, we usually expect to have coarse prior knowledge of the potential action types in an unannotated video collection. The prior knowledge defines the action dictionary. To have a reasonable action dictionary, we include category names of the action dataset we use for finetuning CLIP. However, the prior knowledge we could obtain in a practical case might not be completely accurate. Therefore, we also explore two cases of noisy action dictionary: a) an under-specified dictionary comprised of only part of possible actions in the set, and b) an over-specified dictionary - adding noisy verbs and verb phrases randomly collected from another text corpus. An evaluation of these settings is given in supplementary.

CLIP matching. For a video $x_i \in V$, we use the original CLIP to match x_i with texts in D w.r.t the cosine similarity. We denote the Top-1 matched text as

$$\hat{t}_i = \arg\max_{t \in D} \text{sim}(\phi_v(x_i), \phi_t(t)) \quad (1)$$

where $\text{sim}(u, v) = u^\top v/(\|u\|\|v\|)$ is the cosine similarity. We include $\hat{t}_i$ in the text bag T_i.

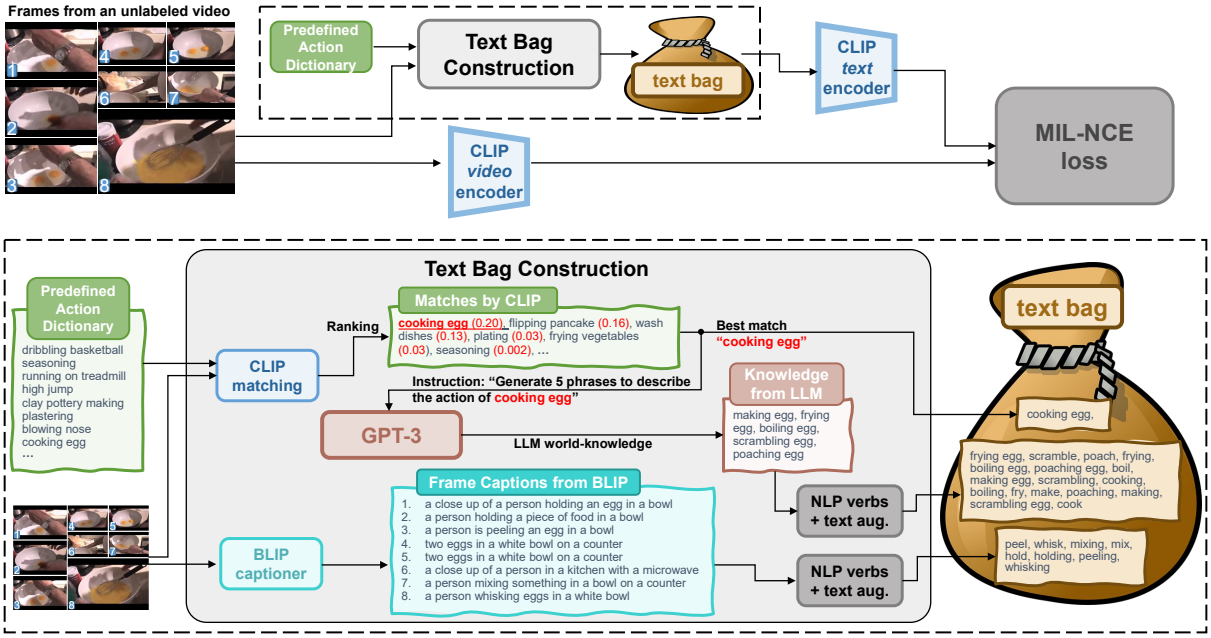

Figure 1. Pipeline of MAXI. Given an unlabeled video collection and a predefined action dictionary, we construct a text bag for each video. We finetune CLIP by passing the video and text bag through the adapted CLIP video encoder (Sec. 2.1) and CLIP text encoder, and optimizing with the Multiple-Instance Learning objective (Sec. 2.3). The text bag construction (Sec. 2.2) for an unlabeled video consists of (1) CLIP matching (2) GPT-3 text expansion and (3) BLIP captioning for video to text expansion.

Method	gt	language	vis.encoder	frames	UCF101	HMDB51	K600 Top1	K600 Top5
ER-ZSAR [3]	yes	Manual description	TSM	16	51.8 ± 2.9	35.3 ± 4.6	42.1 ± 1.4	73.1 ± 0.3
JigsawNet [13]	yes	Manual description	R(2+1)D	16	56.0 ± 3.1	38.7 ± 3.7	-	-
ActionCLIP [20]	yes	K400 dict.	ViT-B/16	32	58.3 ± 3.4	40.8 ± 5.4	66.7 ± 1.1	91.6 ± 0.3
XCLIP [12]	yes	K400 dict.	ViT-B/16	32	72.0 ± 2.3	44.6 ± 5.2	65.2 ± 0.4	86.1 ± 0.8
A5 [7]	yes	K400 dict.	ViT-B/16	32	69.3 ± 4.2	44.3 ± 2.2	55.8 ± 0.7	81.4 ± 0.3
ViFi-CLIP [16]*	yes	K400 dict.	ViT-B/16	16	74.9 ± 0.6	50.9 ± 0.7	67.7 ± 1.1	90.8 ± 0.3
ViFi-CLIP [16]	yes	K400 dict.	ViT-B/16	32	76.8 ± 0.7	51.3 ± 0.6	71.2 ± 1.0	92.2 ± 0.3
Text4Vis [21]	yes	K400 dict.	ViT-L/14	16	-	-	68.9 ± 1.0	-
CLIP [14]	no	-	ViT-B/16	16	69.9 ± 1.3	38.0 ± 1.7	63.5 ± 0.4	86.8 ± 0.4
MAXI	no	K400 dict.	ViT-B/16	16	76.6 ± 0.9	50.5 ± 0.9	70.4 ± 0.8	91.5 ± 0.3
MAXI	no	K400 dict, GPT3 verbs	ViT-B/16	16	77.8 ± 0.3	51.6 ± 0.9	**71.6** ± 1.0	92.3 ± 0.3
MAXI	no	K400 dict, GPT3 verbs	ViT-B/16	16/32	77.8 ± 0.5	51.9 ± 1.1	**71.6** ± 1.0	92.4 ± 0.3
MAXI	no	K400 dict, GPT3 verbs, BLIP verbs	ViT-B/16	16	**78.2** ± 0.8	52.2 ± 0.6	71.4 ± 0.9	**92.5** ± 0.3
MAXI	no	K400 dict, GPT3 verbs, BLIP verbs	ViT-B/16	16/32	**78.2** ± 0.8	**52.3** ± 0.7	71.5 ± 0.8	**92.5** ± 0.4

Table 1. Zero-shot action recognition on UCF101, HMDB51 and K600. We report mean and standard deviation of results on three official validation splits. All models (except for the original CLIP) are trained on K400. We set the text bag filtering ratio p to 90%. We train with 16 frames per video and report single-view inference results with 16 and 32 frames here. *denotes our re-evaluation.

Method	gt	language	Charades	MiT	MiniSSv2	UAV
ViFi-CLIP [16]	yes	K400 dict.	**25.77**	21.68 / 44.19	5.98 / **19.04**	**4.67** / **15.18**
CLIP [14]	no	-	19.80	20.11 / 40.81	3.96 / 14.42	1.79 / 7.05
MAXI	no	K400 dict.	23.47	21.94 / 45.68	5.19 / 17.71	2.42 / 8.39
MAXI	no	K400 dict., GPT3 verbs	23.74	22.11 / 45.79	5.60 / 16.73	2.77 / 9.07
MAXI	no	K400 dict., GPT3 verbs, BLIP verb	23.79	**22.91** / **46.38**	**6.37** / 18.73	2.72 / 9.00

Table 2. Zero-shot action recognition on Charades, MiT, MiniSSv2 and UAV. All models (except for CLIP) are trained on K400. We report the mAP of multi-label classification on Charades and Top-1/Top-5 single-label classification accuracy for MiT, MiniSSv2 and UAV. We set the text bag filtering ratio p to 90%.

The CLIP matching is a means of distilling knowledge from the original CLIP as the teacher. Common choices of unlabeled video collection V are usually of much smaller scale than the original CLIP domain and might be prone to overfitting. Using knowledge from the original CLIP prevents the model from overfitting to the smaller domain V, preserving the generalizability learned in the pretraining stage of CLIP.

GPT-3 text expansion. We expand the text bag by leveraging the large-scale language model (LLM) GPT-3 [1]. We build upon the fact that GPT-3 has high performance on language instruction tasks [1]. By providing the best-matched text $\hat{t}_i$ in the instruction for LLM requiring it to describe this text using its language (world) knowledge (see instruction example in Fig. 1), we obtain a collection of expanded alternative descriptions of the action. The descriptions contain

details hallucinated by the LLM leveraging its collective world knowledge. We collect the verbs and verb phrases extracted from the generated expanded action descriptions. Furthermore, we perform text augmentation by including both the lemma and gerund (present participle) forms of the verbs. We add the collection of words to the text bag T_i.

BLIP captioning for video to text expansion. We employ the vision-language model BLIP [10] for generating captions of individual frames on a video. Note that this image captioning model is not pretrained on any video domain. The frame captions provide instance-level descriptions that are dependent on the visual content of frames of the unlabeled videos. Similar to the case of GPT-3 text expansion, we collect verbs and verb phrases from these descriptions, and perform text augmentation (as stated above), adding the resulting texts to the text bag T_i.

Filtering text bags. To improve the quality of the text bags, we set a threshold δ_p on the similarity score from CLIP matching. We determine δ_p such that $p \times 100\%$ of videos (or text bags) remain after thresholding. For video $x_i \in V$, we keep the corresponding text bag T_i if the best matched text $\hat{t}_i$ has a similarity above the threshold, *i.e.* $\text{sim}(\phi_v(x_i), \phi_t(\hat{t}_i)) \geq \delta_p$. The filtering results in a sampled index set $I_p = \{i \,|\, \text{sim}(\phi_v(x_i), \phi_t(\hat{t}_i)) \geq \delta_p, \forall i \in I\}$ and video set $V_p = \{x_i \,|\, i \in I_p\}$.

2.3. Multiple Instance Learning

We employ Multiple Instance Learning (MIL) to learn from the unlabeled videos and noisy text bags collected above. The MIL-NCE loss proposed in [11] combines Multiple Instance Learning and Noise Contrastive Estimation. Following MIL-NCE, instead of enforcing the match of one specific positive text to each video, we softly associate a text bag T_i with each video $x_i \in V$, in which one or multiple texts could be a positive match to the video. As different videos have varying numbers of texts in bag, we randomly sample N_{bag} texts from the original bag in each training iteration. We refine the definition of the sampled text bag T_i as $T_i = \{t_{i,n}|_{n=1}^{N_{\text{bag}}}\}$, where N_{bag} is the constant bag size.

The original MIL-NCE loss encourages the instance-level match between each video and its corresponding text bag. In this work, we further propose to encourage the videos and text bags, which have the same best matched text, to be close to each other. Noting that each video x_i has a best matched text $\hat{t}_i$ in the dictionary from CLIP matching step, than our proposed loss is

$$\mathcal{L} = -\frac{1}{|I_B|}\sum_i \log \frac{\sum_j \sum_n \exp(\bar{\phi}_v(x_i)^\top \bar{\phi}_t(t_{j,n})/\sigma) \cdot \mathbb{1}(\hat{t}_i = \hat{t}_j)}{\sum_k \sum_n \exp(\bar{\phi}_v(x_i)^\top \bar{\phi}_t(t_{k,n})/\sigma)} \tag{2}$$

where $i, j, k \in I_B$ and $n \in \{1, ..., N_{\text{bag}}\}$. $I_B \subset I_p$ is a sampled batch of indices. $t_{j,n} \in T_j$ is text in a text bag, and σ is a temperature parameter for contrastive learning. $\mathbb{1}(\hat{t}_i = \hat{t}_j)$ is an indicator that x_i and x_j have the same best matched text.

3. Results of Zero-Shot Action Recognition

We finetune CLIP on the large-scale K400 dataset stripped of the original ground truth labels. We perform zero-shot action recognition on seven different datasets to verify that cross-dataset model generalizability transfer after the finetuning. In zero-shot setting, the model is evaluated directly on downstream datasets with unseen classes, without being trained on any samples of these datasets.

In Table 1, we first compare to other state-of-the-art methods, all of which use K400 to adapt CLIP models for zero-shot recognition tasks on UCF, HMDB and K600. Following [3, 12, 16], we report the mean and standard deviation of results on three official validation sets. ER-ZSAR [3] and JigsawNet [13] are zero-shot action recognition approaches that train with K400 ground truth annotations. They leverage crawled descriptions of action classes with manual correction, which requires efforts from human annotators. Afterwards, the class descriptions are assigned to videos based on ground truth annotations. We see that the original CLIP has good direct zero-shot performance across the three datasets, which performs better or on par with ER-ZSAR [3] and JigsawNet [13]. The rest of the compared approaches all adapt CLIP models on video-text pairs with the K400 ground truth class labels as texts. Among them, the most recent ViFi-CLIP [16] achieves the best result, outperforming all the other approaches, without adding any learnable spatio-temporal modules (as done by other approaches such as [7, 12, 20]).

In a similar full finetuning paradigm to ViFi-CLIP, MAXI achieves favorable results without using any ground truth annotation. We report the performance of MAXI with different combinations of language sources. Simply with the original K400 action dictionary, we already outperform most of the related work across the three datasets. With the additional GPT-3 verbs and BLIP verbs in the text bag, we further boost the performance, achieving the state-of-the-art among the three datasets.

For a thorough analysis of the model generalizibility, we further report the performance of MAXI on four datasets (Charades, MiT, MiniSSv2 and UAV) with larger domain shift to K400 in Table 2. In comparison to the original CLIP, our finetuned model has improved zero-shot transfer on all datasets. With the additional language sources of GPT-3 and BLIP, we even outperform ViFi-CLIP trained with ground truth of K400, on the challenging MiT and MiniSSv2 datasets.

In supplementary, we provide dataset statistics, attention map visualizations, evaluations of few-shot action recognition, ablation studies of text bag filtering, noisy action dictionary, strategies of learning from words in a text bag.

References

[1] 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. 1, 3

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

[3] Shizhe Chen and Dong Huang. Elaborative rehearsal for zero-shot action recognition. In *ICCV*, pages 13638–13647, 2021. 3, 4

[4] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *ICLR*, 2022. 1

[5] Lisa Anne Hendricks and Aida Nematzadeh. Probing image-language transformers for verb understanding. *arXiv preprint arXiv:2106.09141*, 2021. 1

[6] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021. 1

[7] Chen Ju, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie. Prompting visual-language models for efficient video understanding. In *ECCV*, pages 105–124. Springer, 2022. 1, 3, 4

[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] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and René Ranftl. Language-driven semantic segmentation. In *ICLR*, 2022. 1

[10] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*, 2022. 1, 4

[11] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *CVPR*, pages 9879–9889, 2020. 4

[12] 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, 3, 4

[13] Yijun Qian, Lijun Yu, Wenhe Liu, and Alexander G Hauptmann. Rethinking zero-shot action recognition: Learning from latent atomic actions. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IV*, pages 104–120. Springer, 2022. 3, 4

[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. 1, 2, 3

[15] Yongming Rao, Wenliang Zhao, Guangyi Chen, Yansong Tang, Zheng Zhu, Guan Huang, Jie Zhou, and Jiwen Lu. Denseclip: Language-guided dense prediction with context-aware prompting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18082–18091, 2022. 1

[16] 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, 4

[17] Hanoona Rasheed, Muhammad Maaz, Muhammad Uzair Khattak, Salman Khan, and Fahad Shahbaz Khan. Bridging the gap between object and image-level representations for open-vocabulary detection. In *NeurIPS*, 2022. 1

[18] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation, Aug. 2022. arXiv:2208.12242 [cs]. 1

[19] Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248, 2022. 1

[20] Mengmeng Wang, Jiazheng Xing, and Yong Liu. Actionclip: A new paradigm for video action recognition. *arXiv preprint arXiv:2109.08472*, 2021. 1, 3, 4

[21] Wenhao Wu, Zhun Sun, and Wanli Ouyang. Revisiting classifier: Transferring vision-language models for video recognition. *Proceedings of the AAAI, Washington, DC, USA*, pages 7–8, 2023. 1, 3

[22] Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *The Eleventh International Conference on Learning Representations*, 2023. 1

[23] Renrui Zhang, Rongyao Fang, Wei Zhang, Peng Gao, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. Tip-adapter: Training-free clip-adapter for better vision-language modeling. In *ECCV*, 2022. 1

[24] Tiancheng Zhao, Tianqi Zhang, Mingwei Zhu, Haozhan Shen, Kyusong Lee, Xiaopeng Lu, and Jianwei Yin. Vl-checklist: Evaluating pre-trained vision-language models with objects, attributes and relations. *arXiv preprint arXiv:2207.00221*, 2022. 1

[25] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Conditional prompt learning for vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16816–16825, 2022. 1

[26] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *International Journal of Computer Vision*, 130(9):2337–2348, 2022. 1

[27] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IX*, pages 350–368. Springer, 2022. 1