Supplementary Material for *Open Vocabulary Multi-Label Video Classification*

Overview

This supplementary material is organized into the following sections:

- Section A: Comparison of our results with Supervised State of the Art
- Section B: Comparison of our results with Single Label Open Vocabulary Baselines
- Section C: Comparison of our results with a Multi-Modal LLM
- Section D: Evaluation of our approach on Single Label Classification tasks
- Section E: Evaluation of our approach on EgoCentric tasks
- Section F: Additional Ablations for Label Encoder
- Section G: Additional Ablations for Temporal Encoder
- Section H: Inference and training costs of our approach
- Section I: Further details about the Synthetic Label Pipeline
- Section J: Further details about all the Baselines reported
- Section K: Implementation details
- Section L: Qualitative Results

A Comparison with Supervised SOTA

In order to provide some additional context for our results, we also evaluate some existing state of the art baselines on our downstream datasets.

The best ActivityNet trained model with public weights is ASM-Loc (He et al. [11], CVPR 2022). Our open vocabulary classifier comes within 10% Peak F1-Score of this supervised model despite not being trained on any ActivityNet data. We also provide results finetuning ViFi CLIP on downstream datasets, which diminishes open vocabulary generalization capabilities. Our single model is competitive across both datasets.

Table 5: Comparing with Supervised Results (Peak F1-Score)

Method	TAO	ActivityNet	Geometric Mean
Ours (Zero-Shot)	56.6	53.8	**55.2**
Supervised Methods			
ASM-Loc CVPR 2022	-	63.1	-
ViFi-CLIP (TAO-FineTuned)	60.2	12.6	27.5
ViFi-CLIP (ActivityNet-FT)	32.7	58.2	30.4

B Comparison with Single Label Open Vocabulary Baselines

STAN [30] is intended for fully fine-tuned setting; it doesn't report any zero-shot results. Open V-CLIP [51] is trained to solve single label classification. In contrast, our goal is to perform multi-label classification in the zero-shot setting. For comparison we provide zero-shot results for both. As STAN doesn't provide pretrained weights, we train it on our training dataset. For Open V-CLIP, we use author provided pretrained weights.

Table 6: Open Vocabulary baselines (Peak F1)

Method	TAO	ActivityNet
`STAN` (K400+YT8M, ours)	58.1	27.6
`Open V-CLIP` (original)	43.9	50.2
Ours	**59.6**	**52.6**

C Comparison with Multi-Modal LLMs

Recently in the literature [58,60], general purpose multi-modal LLMs have been demonstrated to achieve competitive performance across a range of video understanding tasks. They are not practical for our setting, since they impose a significant computation cost, however to demonstrate the advantage of our solution over multi-modal LLMs, we construct two LLaVA-based inference baselines. For the first, we prompt LLaVA regarding the presence of a class label in video frames. For second, we closely follow our synthetic label generation pipeline (see Section 1) and generate frame captions using LLaVA. The captions are then classified using CLIP's text encoder. The results in Table 7 show that LLaVA performs significantly worse than our method, even when no synthetic labels are used for training.

Both LLaVA based approaches require running the Multi-Modal LLM for every video at inference. Additionally, for the first approach, we need to run it for every label in the validation vocabulary. In contrast, for our method the LLM is not used during inference, but only when a new label is added to the classification vocabulary.

D Zero-Shot Single Label Action Classification

Our open vocabulary model though trained for multi-label classification is also competitive (see row **(a)** in Table 8) on the zero-shot single label action classification task.

Table 7: LLaVA Inference baselines (Peak F1-Score)

Method	TAO	ActivityNet	Inference Time (1× A100)
LLaVA Classification (yes/no polling every class)	27.8	11.5	10 min+
LLaVA Captioning + CLIP Classification	47.2	34.7	10s
Ours	**59.6**	52.6	0.25s
Ours w/ added synthetic labels during training	56.6	**53.8**	0.25s

A key difference between our approach and prior single label classification works tailored to this problem is our use of binary classification losses, which is essential for multi-label classification but is not optimal for single label classification, which only requires ranking the labels. In order to match the setting of prior works, we also train our model on only Kinetics-400 using Cross-Entropy loss and provide results in row **(b)** to show that it can exceed prior work such as ViFi-CLIP [39].

Model	**UCF101**	**HMDB51**	**Kinetics600**
No video data used in training			
CLIP	61.7	37.5	63.5
CLIP + LLM	73.8	46.1	64.8
Trained on *YouTube8M + Kinetics400*			
(a) Ours	74.1	53.2	67.7
Trained on *Kinetics400*			
Vi-Fi CLIP *	77.5	51.8	71.2
(b) Ours (Cross-Entropy Loss)	**79.0**	**54.5**	**72.8**

Table 8: Results on single label action classification datasets. Top-1 Accuracy is reported for all datasets. * Results reported in ViFi CLIP [39]

E Zero-Shot Evaluation on EgoCentric tasks

We provide results for scenario classification on Ego4d and verb & noun identification for Epic-Kitchens (unseen kitchens).

Table 9: Egocentric (Peak F1)

Method	Ego4D	EK-unseen (Verbs)	EK-unseen (Nouns)
CLIP	45.3	16.5	32.8
CLIP+LLM	48.7	20.3	39.1
Ours	**51.9**	**22.1**	**40.5**

Table 10: LLM Adaptation Ablations (Peak F1-Score, 10k training steps)

Steps	LLM Adapter	LLM-VLM Connector	TAO	ActivityNet
10k	LoRA (r=2)	Prompting Transformer	57.9	49.4
10k	LoRA (r=4)	Prompting Transformer	58.0	46.9
10k	LoRA (r=2)	Linear	46.2	38.9
10k	Prompts	Linear	48.4	35.2
10k	Prompts	MLP	52.9	44.7
10k	**Prompts**	**Prompting Transformer**	**58.3**	**50.8**
50k	Prompts	Prompting Transformer	59.6	52.6

F Additional Ablations for Label Encoder

We conducted additional ablations for the LLM adapter and LLM-VLM connector, with results shown in Table 10. Due to time constraints, we trained for only 10k steps, nearing convergence. We find that LoRA saturates after rank 2 and performs worse than prompt learning for Zero-Shot Generalization. Our prompting transformer outperforms MLP & Linear connectors.

G Additional Ablations for Temporal Encoder

We designed an alternative version of our architecture with serial blocks instead of parallel and train the model again. The results (Table 11) indicate that parallel blocks outperform serial blocks. As our main goal is open vocabulary classification, our temporal ablations (Table 4) are focused on regularization and related aspects. Exhaustive temporal architecture ablations are beyond the scope of a single paper, and different aspects of temporal modeling have been studied previously ([3], [27], [40]).

Table 11: Temporal (Peak F1)

Temporal Adapter	TAO	ActivityNet
Serial (n=4)	53.2	41.5
Parallel (n=4)	**59.6**	**52.6**

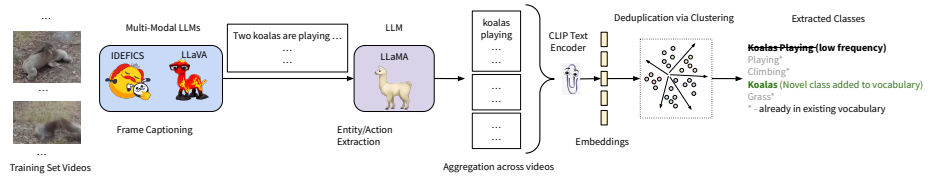
H Comparison of Computational Costs

Table 12: Computational Costs

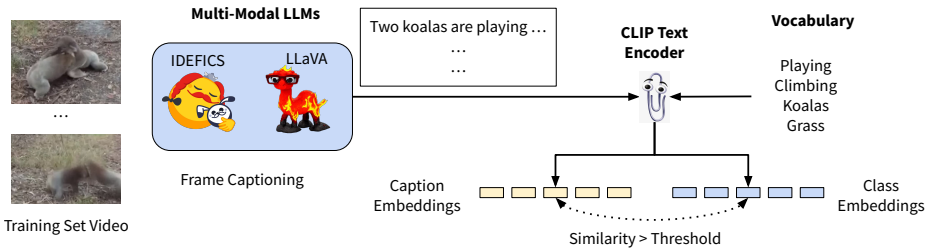
Method	**Training (YT8M+K400)**		**Inference time**
	Time	Mem/GPU	(batch size = 32)
ViFi-CLIP	36 Hrs	11.0 GB	338ms
Ours	40 Hrs	16.5 GB	393ms

Training time (on 16 × A100 GPUs) on YT-8M + K400 for our method is about 10% higher than ViFi-CLIP baseline. Inference on 1 RTX8000 is about 16% slower (batch size=32, using `torchinfo` package). Text embeddings for class labels can be pre-computed, only video features need computing on the fly during inference.

I Synthetic Label Generation Pipeline



(a) Expanding the label vocabulary.



(b) Assigning labels to videos.

Fig. 7: Incorporating synthetic labels into our training sets enhances our open vocabulary performance further. **(a) Vocabulary Expansion:** We have developed a pipeline to automatically extract action and object labels from a vast video dataset utilizing foundation models. For captioning video frames, we employ Multi-Modal Large Language Models (LLMs), specifically IDEFICS and LLaVA. Subsequently, LLaMA is prompted to distill object and action class labels from these captions. We aggregate these labels across videos and remove duplicates through clustering, forming a classification vocabulary. **(b) Label Assignment:** Labels from the vocabulary are aligned with the generated video captions using the text encoder from CLIP.

As illustrated in Figure 7 our synthetic labeling pipeline consists of four steps: caption generation, concept extraction, vocabulary expansion and label assignment. The caption generation process employs off-the-shelf multi-modal LLMs and is straightforward.

For the second step, we prompt LLaMA 2-13B-Chat to extract concept labels for videos from captions generated in Stage 1.The LLM prompt for extracting labels from captions is provided in Listing I. We provide 3 in-context examples and the LLM is prompted to extract concept labels from the fourth video's captions.

As LLMs identify a large number of concepts, including many near-duplicates, a cleanup step is necessary to minimize these issues. We utilize the CLIP text encoder to obtain embeddings for all the identified concepts across the dataset. K-Means clustering is applied to the embeddings to cluster them into groups. For each group, we replace the labels with the most frequently observed concepts

from that group. This works reasonably well, and CLIP text encoder is excellent at detecting near-duplicate visual concepts. A random sample of identified clusters are shown in Table 13.

Finally, we reuse the CLIP text encoder to match labels from the deduplicated vocabulary back to videos, which are represented by their captions. In order to reduce domain shift between captions and the labels, standard CLIP prompt template `"a video of {label}"` is used.

For extracting extra action labels, we use IDEFICS-9b-Instruct model [19] and LLaVA [29] to caption the videos. Both models are based on LLaMA LLMs, with IDEFICS trained using the interleaved image text dataset OBELICS, while LLaVA is trained on a mix of image-caption data and instruction following data created using GPT-3 and image annotations. This stage is followed by LLaMA 2-13b-chat [49] to extract the labels from the captions. OpenAI CLIP B/32 is used to clean up label assignment to videos.

Airlines	'american airlines', 'delta air lines', 'southwest airlines', 'singapore airlines', 'air france', 'emirates (airline)', 'british airways', 'carnival cruise line'
Grilling	'barbequing', 'cooking on campfire', 'grilling', 'barbecue'
Lego	'legoland', 'lego star wars', 'lego minecraft', 'lego duplo', 'the lego group', 'lego friends', 'lego batman: the videogame', 'lego', 'lego batman 3: beyond gotham', 'lego ninjago', 'lego minifigure', 'playmobil', 'lego marvel super heroes', 'lego city', 'lego legends of chima'
Playground	'playing on a playground', 'playground', 'amusement park', 'amusement arcade', 'amusement ride', 'water park', 'ferris wheel'
Video Games	'gears of war (video game)', 'jill valentine', 'gears of war', 'silent hill 2', 'resident evil 2', 'hitman: absolution', 'gears of war 2', 'resident evil 5', 'resident evil 3: nemesis', 'resident evil', 'resident evil (1996 video game)', 'resident evil (2002 video game)'
Water Slide	'water sliding', 'riding water slide', 'water slide'

Table 13: Sampled clusters among concepts identified by the captioning + LLM steps of the label generation pipeline. CLIP Text encoder features were used to cluster the concepts for de-duplication. Cluster names assigned in the left column are only used for illustration.

Our LLM prompt for extracting action labels from video captions

```
Following is the description of a video. Output a numbered list of verbs
representing visual actions performed in the video. Do not add any explanation.

Video 1 description:
1. A group of people riding motorcycles at night.
2. A motorcycle is lit up with blue lights.
3. A person is riding a bike at night.
4. A motorcycle parked on the street at night.
5. A group of people are gathered in a dimly lit room.
6. A motorcycle parked in a dark room.
7. A motorcycle is parked in a dark room.
8. A person is riding a bike at night.

Verbs Found:
1. riding motorcycle
2. riding bike

Following is the description of a video. Output a numbered list of verbs
representing visual actions performed in the video. Do not add any explanation.

Video 2 description:
1. A man is performing on stage with a band.
2. A group of men are performing on a stage.
3. A man with a microphone is performing on stage.
4. A group of young men performing on stage.
5. A man is singing on a stage with a band.
6. A man is playing a guitar on a stage.
7. A man and a woman are performing on stage.
8. A dark room with a bright light shining on it.

Verbs Found:
1. performing on stage
2. singing on stage
3. playing guitar

Following is the description of a video. Output a numbered list of verbs
representing visual actions performed in the video. Do not add any explanation.


Video 3 description:
1. A person is putting lotion on another person's hand.
2. A person is putting nail polish on another person's nails.
3. A person is putting nail polish on their nails.
4. A person is holding a ball point pen.
5. A person is writing on a piece of paper.
6. A person is holding another person's hand.
7. A person is putting a ring on another person's finger.
8. A black screen with a white frame.

Verbs Found:
1. putting lotion
2. putting nail polish
3. writing
4. putting ring

Following is the description of a video. Output a numbered list of verbs
representing visual actions performed in the video. Do not add any explanation.


Video 4 description:
<output_captions>

Verbs Found:
```

J Baselines

J.1 CLIP + LLM Frozen Baseline

This baseline is an extension of and inspired by prior works [33,37] that utilize LLMs to prompt CLIP for image classification to our problem of video classification. The LLM is prompted to generate class descriptors utilizing its extensive world knowledge. CLIP can then be used to match these descriptors to video frames to classify them.

The overall process is illustrated in Figure 8a. Unlike the image classification setting it includes a mean pooling operation across frames to get the video level feature. Note that the LLM prompt is designed to elicit output in the form of a list. This simplifies the post-processing of the text output to generate CLIP prompts (Figure 8c). Firstly, the text is split into each item of the list, followed by removal of repetitions (common for this generation of LLMs). Finally we use a standard CLIP prompting template to incorporate both the class label and the descriptor. Sample descriptors for some classes from our downstream datasets are provided in Figure 8b. CLIP + LLM is a reasonably and consistently strong baseline across datasets, as it inherits CLIP's robustness.

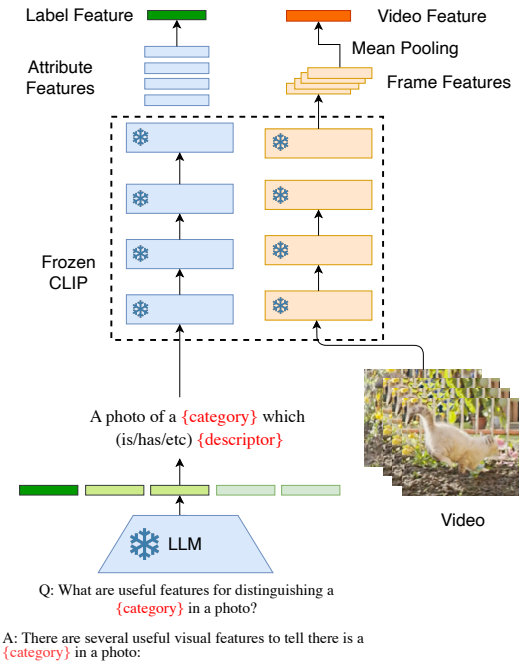
A key limitation of this baseline is the frequency of LLM failures. Different classes of failures such as getting trapped in a repetition loop, generating descriptors which are not visual, and semantic confusion are common. An end-to-end trainable approach could potentially alleviate some of these issues.

J.2 CoOp: Context Optimization

CoOp [62] learns prompts for CLIP's text encoder to adapt it for image classification. This is a parameter efficient adaptation method since it has very few learnable parameters. We extend it to the video setting by utilizing mean pooling across frame in the vision encoder and learnable prompts in the text encoder. (See Figure 9a)

J.3 DualCoOp

DualCoOp [47] refines prompt learning for the multi-label setting, with both positive and negative learnable prompts. A label is matched to a video if the similarity score of the video features with the features for the positive prompts is higher than for the negative prompts. A soft prediction score can be obtained by taking a softmax across the two similarity values. (See Figure 9b)



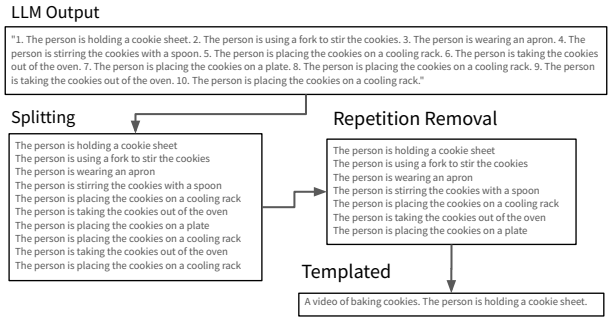
(a) CLIP + LLM Baseline Architecture.

{acorn} **[TAO]**
- The acorn has a hard shell
- The acorn is brown in color
- The acorn is round in shape
- The acorn is attached to a tree
- The acorn is on a tree branch
- The acorn is on a tree trunk

{baking cookies} **[ActivityNet]**
- The person is holding a cookie sheet
- The person is using a fork to stir the cookies
- The person is wearing an apron
- The person is stirring the cookies with a spoon
- The person is placing the cookies on a cooling rack
- The person is taking the cookies out of the oven
- The person is placing the cookies on a plate

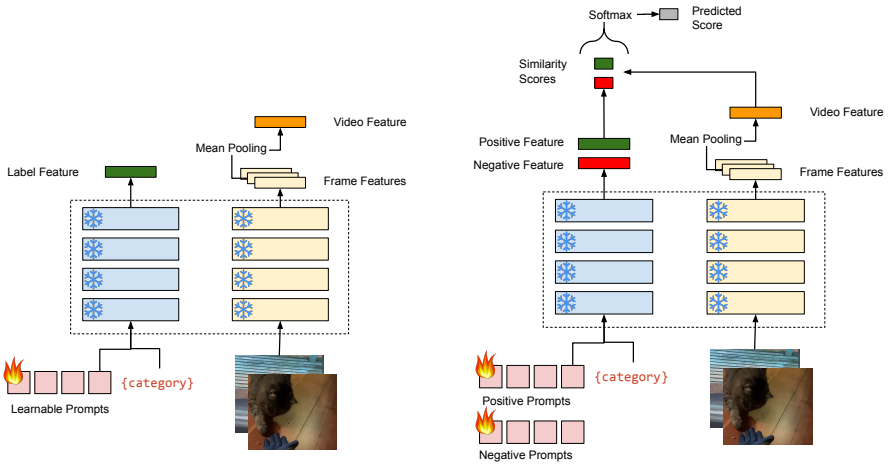
{measure carrots} **[RareAct]**
- The person is holding a tape measure
- They are holding the tape measure with one hand and the carrots with the other hand
- The carrots are on a cutting board
- The person has a knife in their hand
- They are slicing carrots
- They are cutting carrots into pieces
- They are using a grater to grate the carrot pieces
- The carrot pieces are in a bowl
- The bowl has carrots in it
- The bowl is empty

(b) Sample LLM generated Class Descriptors.



(c) Process for cleaning up LLM output to generate CLIP Prompts.

Fig. 8: Here we illustrate the CLIP+LLM baseline discussed in the main paper. **(a)** Architecture consist of frozen CLIP and LLM model. The LLM is prompted to generate class descriptors to assisst CLIP. **(b)** Some sample class descriptors generated by the LLM. **(c)** Process for converting the raw LLM output text to attribute prompts for CLIP. Firstly, the raw text is split into separate list items, then repetitions (which LLMs are prone to) are removed and finally standard CLIP prompting templates are used to combine the class name with the descriptor.



(a) CoOp. **(b)** Dual CoOp.

Fig. 9: Trainable CLIP based baselines without an LLM **(a) CoOp** utilizes learnable prompts on the text encoder side to guide the model towards classification task. **(b) Dual CoOp** is designed for multi-label classification and utilizes learnable positive and negative prompts to generate a probability score for each label.

K Implementation Details

We use Open AI CLIP-B/32 [38] as the backbone VLM and Flan-T5-XL [4] as the promptable LLM. We sample 8 frames per clip during training, and during evaluation we use 4 clips per video (total of 32 frames). Our best model uses 4 learnable LLM Prompts, and 4 layers of temporal modeling. (Following the notation from the method section, $N = 4$, $K = 5$, $L = 5$ and $T = 4$). Following the Flan-T5 text generation instructions we use the following prompt:

```
Q: What are useful features for distinguishing a {label} in a photo?
A: There are several useful visual features to tell about a {label}
in a photo:
1. <extra_id_0>
```

Where {label} represents a given class label and <extra_id_0> is T5 decoder's start token ID. In case of the frozen baseline, when presented with this prompt, the LLM produces text that consists of a list of label attributes, which can be parsed and separated into different prompts for CLIP. We mimic this approach in our learnable version by chunking LLM output into K groups of L tokens each.

We use a batch size of 12 videos/GPU across 32 A100 GPUs (Total Batch Size=384) and train all models for 30000 training steps and evaluate at 10000, 20000 and 30000 steps, providing the best results for each methods across these 3 checkpoints. We do this to ensure a fair comparison as our different baselines and methods have widely varying number of trainable parameters, it would not be a fair comparison to use the same number of steps for each.

K.1 Datasets Used

For training, we use YouTube-8M and Kinetics datasets. For evaluation we use TAO (Tracking Any Object) dataset for Object Classification and ActivityNet for action classification. We also leverage the RareAct dataset in a novel way by using their noun and verb labels to generate 3 labels for each clip, noun, verb and noun-verb combination. For YouTube-8M, we use the human verified validation set for reporting results. Overall these evaluation datasets cover a wide range of entities and actions, providing a comprehensive evaluation of open vocabulary multi-label video classification capabilities.

Dataset	**# Videos**	**# Classes**	**# Labels/Video**
Training Datasets			
YouTube-8M	2,285,432	2429	2.9
+ Generated Labels	2,285,432	3281	6.7
Kinetics 400	246,245	400	1
+ Generated Labels	246,245	1355	4.5
Test Datasets			
YT-8M Segments Val	42,407	1000	1.05
TAO	655	1230	1.44
ActivityNet	4,593	200	1.01
RareAct	905	214	3.02

Table 14: Details about datasets used for training. For YouTube-8M and Kinetics, we also generate additional labels for training using our synthetic labelling pipeline.

K.2 Training Details

We use Open AI CLIP-B/32 as the Vision-Language model and Google Flan-T5 XL as the promptable LLM.

We use `AdamW` optimizer for training with a base learning rate of 0.00001. Weight decay for newly initialized layers is set to 0.0000001 and 0.0 for CLIP initialized layers. Weight regularization loss weight for STAN's spatial attention layers is set to $\lambda = 0.000001$. Cosine decay learning rate scheduler with warmup is used. Total training length is $30,000$ steps including $2,000$ steps of warmup.

L Qualitative Results

References

1. Abu-El-Haija, S., Kothari, N., Lee, J., Natsev, P., Toderici, G., Varadarajan, B., Vijayanarasimhan, S.: Youtube-8m: A large-scale video classification benchmark (2016)
2. Bertasius, G., Wang, H., Torresani, L.: Is space-time attention all you need for video understanding? In: Proceedings of the International Conference on Machine Learning (ICML) (July 2021)
3. Cheng, et al.: VindLU: A recipe for effective video-and-language pretraining. In: CVPR (2023)
4. Chung, H.W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., Webson, A., Gu, S.S., Dai, Z., Suzgun, M., Chen, X., Chowdhery, A., Castro-Ros, A., Pellat, M., Robinson, K., Valter, D., Narang, S., Mishra, G., Yu, A., Zhao, V., Huang, Y., Dai, A., Yu, H., Petrov, S., Chi, E.H., Dean, J., Devlin, J., Roberts, A., Zhou, D., Le, Q.V., Wei, J.: Scaling instruction-finetuned language models (2022)
5. Dave, A., Khurana, T., Tokmakov, P., Schmid, C., Ramanan, D.: Tao: A large-scale benchmark for tracking any object. In: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16. pp. 436–454. Springer (2020)
6. Desai, K., Kaul, G., Aysola, Z., Johnson, J.: Redcaps: Web-curated image-text data created by the people, for the people. arXiv preprint arXiv:2111.11431 (2021)
7. Fan, L., Krishnan, D., Isola, P., Katabi, D., Tian, Y.: Improving clip training with language rewrites. In: NeurIPS (2023)
8. Fang, H., Xiong, P., Xu, L., Chen, Y.: Clip2video: Mastering video-text retrieval via image clip. arXiv preprint arXiv:2106.11097 (2021)

9. Gorti, S.K., Vouitsis, N., Ma, J., Golestan, K., Volkovs, M., Garg, A., Yu, G.: X-pool: Cross-modal language-video attention for text-video retrieval. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 5006–5015 (2022)
10. Gupta, R., Roy, A., Christensen, C., Kim, S., Gerard, S., Cincebeaux, M., Divakaran, A., Grindal, T., Shah, M.: Class prototypes based contrastive learning for classifying multi-label and fine-grained educational videos. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 19923–19933 (June 2023)
11. He, B., Yang, X., Kang, L., Cheng, Z., Zhou, X., Shrivastava, A.: Asm-loc: Action-aware segment modeling for weakly-supervised temporal action localization. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 13925–13935 (2022)
12. Heilbron, F.C., Niebles, J.C.: Collecting and annotating human activities in web videos. In: Proceedings of International Conference on Multimedia Retrieval. p. 377–384. ICMR '14, Association for Computing Machinery, New York, NY, USA (2014). https://doi.org/10.1145/2578726.2578775, https://doi.org/10.1145/2578726.2578775
13. Hu, E.J., yelong shen, Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2022), https://openreview.net/forum?id=nZeVKeeFYf9
14. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 4904–4916. PMLR (18–24 Jul 2021), https://proceedings.mlr.press/v139/jia21b.html
15. Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International conference on machine learning. pp. 4904–4916. PMLR (2021)
16. Kaul, P., Xie, W., Zisserman, A.: Multi-modal classifiers for open-vocabulary object detection. In: International Conference on Machine Learning (2023)
17. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., Suleyman, M., Zisserman, A.: The kinetics human action video dataset (2017)
18. Kojima, T., Gu, S.S., Reid, M., Matsuo, Y., Iwasawa, Y.: Large language models are zero-shot reasoners. In: Oh, A.H., Agarwal, A., Belgrave, D., Cho, K. (eds.) Advances in Neural Information Processing Systems (2022), https://openreview.net/forum?id=e2TBb5y0yFf
19. Laurençon, H., Saulnier, L., Tronchon, L., Bekman, S., Singh, A., Lozhkov, A., Wang, T., Karamcheti, S., Rush, A.M., Kiela, D., Cord, M., Sanh, V.: Obelics: An open web-scale filtered dataset of interleaved image-text documents (2023)
20. Lester, B., Al-Rfou, R., Constant, N.: The power of scale for parameter-efficient prompt tuning. In: Moens, M.F., Huang, X., Specia, L., Yih, S.W.t. (eds.) Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. pp. 3045–3059. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic (Nov 2021). https://doi.org/10.18653/v1/2021.emnlp-main.243, https://aclanthology.org/2021.emnlp-main.243

21. Li, J., Li, D., Xiong, C., Hoi, S.: BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., Sabato, S. (eds.) Proceedings of the 39th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 162, pp. 12888–12900. PMLR (17–23 Jul 2022), https://proceedings.mlr.press/v162/li22n.html
22. Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., Hoi, S.C.H.: Align before fuse: Vision and language representation learning with momentum distillation. Advances in neural information processing systems **34**, 9694–9705 (2021)
23. Li, L.H., Zhang, P., Zhang, H., Yang, J., Li, C., Zhong, Y., Wang, L., Yuan, L., Zhang, L., Hwang, J.N., Chang, K.W., Gao, J.: Grounded language-image pre-training. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10965–10975 (June 2022)
24. Li, X.L., Liang, P.: Prefix-tuning: Optimizing continuous prompts for generation. In: Zong, C., Xia, F., Li, W., Navigli, R. (eds.) Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). pp. 4582–4597. Association for Computational Linguistics, Online (Aug 2021). https://doi.org/10.18653/v1/2021.acl-long.353, https://aclanthology.org/2021.acl-long.353
25. Lin, W., Karlinsky, L., Shvetsova, N., Possegger, H., Kozinski, M., Panda, R., Feris, R., Kuehne, H., Bischof, H.: Match, expand and improve: Unsupervised finetuning for zero-shot action recognition with language knowledge. In: ICCV (2023)
26. Lin, X., Petroni, F., Bertasius, G., Rohrbach, M., Chang, S.F., Torresani, L.: Learning to recognize procedural activities with distant supervision. arXiv preprint arXiv:2201.10990 (2022)
27. Liu, et al.: Mug-STAN: Adapting image-language pretrained models for general video. arXiv:2311.15075 (2023)
28. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: Advances in Neural Information Processing Systems. Curran Associates, Inc. (2023)
29. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: NeurIPS (2023)
30. Liu, R., Huang, J., Li, G., Feng, J., Wu, X., Li, T.H.: Revisiting temporal modeling for clip-based image-to-video knowledge transferring. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 6555–6564 (2023)
31. Liu, X., Zheng, Y., Du, Z., Ding, M., Qian, Y., Yang, Z., Tang, J.: Gpt understands, too. AI Open (2023). https://doi.org/https://doi.org/10.1016/j.aiopen.2023.08.012, https://www.sciencedirect.com/science/article/pii/S2666651023000141
32. Luo, H., Ji, L., Zhong, M., Chen, Y., Lei, W., Duan, N., Li, T.: Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. Neurocomputing **508**, 293–304 (2022)
33. Menon, S., Vondrick, C.: Visual classification via description from large language models. In: The Eleventh International Conference on Learning Representations (2023), https://openreview.net/forum?id=jlAjNL8z5cs
34. Miech, A., Alayrac, J.B., Laptev, I., Sivic, J., Zisserman, A.: Rareact: A video dataset of unusual interactions. arxiv:2008.01018 (2020)
35. Minderer, M., Gritsenko, A., Stone, A., Neumann, M., Weissenborn, D., Dosovitskiy, A., Mahendran, A., Arnab, A., Dehghani, M., Shen, Z., Wang, X., Zhai, X., Kipf, T., Houlsby, N.: Simple open-vocabulary object detection. In: Avidan,

S., Brostow, G., Cissé, M., Farinella, G.M., Hassner, T. (eds.) Computer Vision – ECCV 2022. pp. 728–755. Springer Nature Switzerland, Cham (2022)
36. Ni, B., Peng, H., Chen, M., Zhang, S., Meng, G., Fu, J., Xiang, S., Ling, H.: Expanding language-image pretrained models for general video recognition. In: European Conference on Computer Vision. pp. 1–18. Springer (2022)
37. Pratt, S., Covert, I., Liu, R., Farhadi, A.: What does a platypus look like? generating customized prompts for zero-shot image classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15691–15701 (2023)
38. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
39. Rasheed, H., khattak, M.U., Maaz, M., Khan, S., Khan, F.S.: Finetuned clip models are efficient video learners. In: The IEEE/CVF Conference on Computer Vision and Pattern Recognition (2023)
40. Rizve, et al.: VidLA: Video-language alignment at scale. In: CVPR (2024)
41. Roth, K., Kim, J.M., Koepke, A.S., Vinyals, O., Schmid, C., Akata, Z.: Waffling around for performance: Visual classification with random words and broad concepts. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 15746–15757 (October 2023)
42. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., et al.: Laion-5b: An open large-scale dataset for training next generation image-text models. Advances in Neural Information Processing Systems **35**, 25278–25294 (2022)
43. Schuhmann, C., Vencu, R., Beaumont, R., Kaczmarczyk, R., Mullis, C., Katta, A., Coombes, T., Jitsev, J., Komatsuzaki, A.: Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. arXiv preprint arXiv:2111.02114 (2021)
44. Sharma, P., Ding, N., Goodman, S., Soricut, R.: 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). pp. 2556–2565 (2018)
45. Shvetsova, N., Kukleva, A., Hong, X., Rupprecht, C., Schiele, B., Kuehne, H.: Howtocaption: Prompting llms to transform video annotations at scale (2023)
46. Singh, A., Hu, R., Goswami, V., Couairon, G., Galuba, W., Rohrbach, M., Kiela, D.: Flava: A foundational language and vision alignment model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15638–15650 (2022)
47. Sun, X., Hu, P., Saenko, K.: Dualcoop: Fast adaptation to multi-label recognition with limited annotations. Advances in Neural Information Processing Systems **35**, 30569–30582 (2022)
48. Thomee, B., Shamma, D.A., Friedland, G., Elizalde, B., Ni, K., Poland, D., Borth, D., Li, L.J.: Yfcc100m: The new data in multimedia research. Commun. ACM **59**(2), 64–73 (jan 2016). https://doi.org/10.1145/2812802, https://doi.org/10.1145/2812802
49. Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al.: Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 (2023)
50. Wasim, S.T., Naseer, M., Khan, S., Khan, F.S., Shah, M.: Vita-clip: Video and text adaptive clip via multimodal prompting. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 23034–23044 (2023)

51. Weng, Z., Yang, X., Li, A., Wu, Z., Jiang, Y.G.: Open-VCLIP: Transforming CLIP to an open-vocabulary video model via interpolated weight optimization. In: Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., Scarlett, J. (eds.) Proceedings of the 40th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 202, pp. 36978–36989. PMLR (23–29 Jul 2023), https://proceedings.mlr.press/v202/weng23b.html
52. Weng, Z., Yang, X., Li, A., Wu, Z., Jiang, Y.G.: Open-vclip: Transforming clip to an open-vocabulary video model via interpolated weight optimization. In: ICML (2023)
53. Wortsman, M., Ilharco, G., Kim, J.W., Li, M., Kornblith, S., Roelofs, R., Lopes, R.G., Hajishirzi, H., Farhadi, A., Namkoong, H., et al.: Robust fine-tuning of zero-shot models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7959–7971 (2022)
54. Xu, Z., Zhu, Y., Deng, T., Mittal, A., Chen, Y., Wang, M., Favaro, P., Tighe, J., Modolo, D.: Challenges of zero-shot recognition with vision-language models: Granularity and correctness (2023)
55. Xue, H., Sun, Y., Liu, B., Fu, J., Song, R., Li, H., Luo, J.: Clip-vip: Adapting pre-trained image-text model to video-language representation alignment. arXiv preprint arXiv:2209.06430 (2022)
56. Yang, Y., Panagopoulou, A., Zhou, S., Jin, D., Callison-Burch, C., Yatskar, M.: Language in a bottle: Language model guided concept bottlenecks for interpretable image classification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 19187–19197 (June 2023)
57. Yao, L., Huang, R., Hou, L., Lu, G., Niu, M., Xu, H., Liang, X., Li, Z., Jiang, X., Xu, C.: Filip: Fine-grained interactive language-image pre-training. arXiv preprint arXiv:2111.07783 (2021)
58. Yousaf, A., Naseer, M., Khan, S., Khan, F., Shah, M.: VIDEOPROMPTER: AN ENSEMBLE OF FOUNDATIONAL MODELS FOR ZERO-SHOT VIDEO UNDERSTANDING (2024), https://openreview.net/forum?id=9F0xInGNBF
59. Yu, J., Wang, Z., Vasudevan, V., Yeung, L., Seyedhosseini, M., Wu, Y.: Coca: Contrastive captioners are image-text foundation models. arXiv preprint arXiv:2205.01917 (2022)
60. Zhang, C., Lu, T., Islam, M.M., Wang, Z., Yu, S., Bansal, M., Bertasius, G.: A simple llm framework for long-range video question-answering. CoRR **abs/2312.17235** (2023), https://doi.org/10.48550/arXiv.2312.17235
61. Zhao, Y., Misra, I., Krähenbühl, P., Girdhar, R.: Learning video representations from large language models. In: CVPR (2023)
62. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International Journal of Computer Vision **130**(9), 2337–2348 (2022)
63. Zhu, X., Zhang, R., He, B., Guo, Z., Zeng, Z., Qin, Z., Zhang, S., Gao, P.: Pointclip v2: Prompting clip and gpt for powerful 3d open-world learning. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 2639–2650 (October 2023)