Supplementary Materials for: Align before Adapt: Leveraging Entity-to-Region Alignments for Generalizable Video Action Recognition

Yifei Chen **Dapeng Chen** **Ruijin Liu** **Sai Zhou** **Wenyuan Xue** **Wei Peng**
IT Innovation and Research Center, Huawei Technologies
{chenyifei14, chendapeng8, liuruijin1}@huawei.com

A. Training configuration

In fully-supervised training, we set the batch size to 256 and adopt the AdamW optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.98$. The learning rate is 8×10^{-6} for the ViT backbone in the region-aware image encoder and 8×10^{-5} for the remaining learnable parts. In few-shot experiments, the learning rate of the video adapter is scaled up by ten times, and the batch size is reduced to 64. Regarding the text corpora, we initially constructed a text corpus from [8, 9, 19] with a total number of 913 text entities. All the text entities are embedded offline and fixed throughout experiments. It is worthwhile mentioning that when adapting to the motion-heavy dataset Something Something v2 [5], which predominantly uses action labels in the format of *"action on **something**"* without specifying instances, we directly collect the action labels as entities. For data augmentation, we utilize the technique including *RandomFlip, MultiScaleCrop, Mixup, and Label smoothing*, following the manner of X-CLIP [13].

B. Text corpus construction

In this paper, we propose an automatic pipeline for generating a text corpus. For each description, (1) we 'extract' the relevant action-related units in two approaches: noun and phrase entities extraction with NLTK & spaCy [3, 6] part of speech (POS) tools; And ChatGPT [14], where we design a prompt template *"What are identifying visual characteristics, such as object, body parts, scenes, and roles, of a/an {label} video action? List them concisely."* (2) We then use the WordNet [12] tool to generate a sequence of explanatory descriptions for each extracted single-word unit. For the extracted phrase unit, we prompt ChatGPT to generate explanatory descriptions with the following templates: *"Concisely describe what a/an {phrase unit} looks like"*, *"Concisely list potential explanations for {phrase unit}"*, and *"Concisely explain {phrase unit} in one sentence"*. (3) To determine the most appropriate description for each unit, we employ the Lesk algorithm [2] and T5-based word sense disambiguation [20] model according to the action labels. All of these procedures are automated through code.

Method	Pretrain	Top-1	GFLOPs	Views
MViT-B-24 [4]	-	83.8	1180	32×5×1
VideoSwin-L(384↑) [11]	IN-21k	85.9	25284	32×4×3
ViViT-H/16x2 320 [1]	JFT-300M	83.0	-	32×4×3
ViViT-H/16x2 [1]	JFT-300M	85.8	48916	32×4×3
TokenLearner-L/10 [18]	JFT-300M	86.3	48912	32×4×3
Florence(384↑) [23]	FLD-900M	87.8	-	32×4×3
CoVeR [24]	JFT-3B	87.9	-	96×1×3
MTV-L [22]	JFT-3B	85.4	18483	32×4×3
MTV-H [22]	WTS-17B	89.6	45537	32×4×3
X-CLIP-L/14 [13]	CLIP-400M	88.3	7896	8×4×3
ALT-B/16	CLIP-400M	86.1	1308	32×1×3
ALT-L/14	CLIP-400M	88.6	4947	32×1×3

Table C1. Fully-supervised comparison on Kinetics-600.

Method	Pretrain	Top-1	GFLOPs	Views
MViT-B-24 [4]	K-600	**69.7**	708	32×1×3
ViViT-L [1]	IN-21K+K-400	65.4	11892	32×1×3
MTV-B(384↑) [22]	IN-21K+K-400	68.5	11160	32×3×4
EVL-B/16 [10]	CLIP-400M	62.4	2047	32×1×3
ST-Adapter-B/16 [15]	CLIP-400M	<u>69.5</u>	1955	32×1×3
ALT-B/16	CLIP-400M	68.6	1308	32×1×3

Table C2. Fully-supervised comparison on SS-V2.

C. Additional experiments

C.1. Fully-supervised experiments on Kinetics-600

Tab. C1 presents the results on Kinetics-600. Our ALT-B/16 outperforms MTV-L [4] by *0.7%* by using 32 frames per video with three views. Equipping with a larger backbone, ALT-L/14 achieves *88.6%* top-1 accuracy with computation consumption of only 4947 GFLOPs, which takes the lead among the methods that adopt similar-level pre-trained models and data.

Method	K=2	K=4	K=8	K=16
Vanilla CLIP [16]	2.7	2.7	2.7	2.7
ActionCLIP [21]	4.1	5.8	8.4	11.1
X-CLIP-B/16 [13]	3.9	4.5	6.8	10.0
A5 [7]	4.4	5.1	6.1	9.7
ViFi-CLIP [17]	6.2	7.4	8.5	12.4
ALT-B/16	**6.6**	**7.7**	**9.4**	**12.9**

Table C3. Few-shot comparison on Something Something V2. All the models are trained on Kinetics-400, with top-1 accuracies(%) reported under a single-view inference.

C.2. Fully-supervised experiments on Something-Something v2

The Something-Something V2 (SS-V2) dataset collects more than 220000 video clips that belong to 174 action categories, covering basic human actions with everyday objects. Compared to Kinetics-400, it requires more temporal reasoning. We evaluate our approach on Something-Something V2 under full supervision. The accuracies are reported in Tab. C2. Among the CLIP-based works, our method outperforms EVL [10], but it is inferior to ST-adapter [15], which utilizes interleaved heavier 3D Convolution modules. We attribute the key to handling such kind of motion-heavy datasets to elaborately designed temporal communication mechanisms, which inspire future directions of our work.

Text corpus	Fully.	2-shot	0-shot
$\varnothing$	81.7	52.8	71.6
$body$	81.9	56.4	73.6
$object$	82.5	62.3	75.8
$scene$	82.4	61.1	75.1
$motion$	81.8	58.7	74.4
All	**82.8**	**64.3**	**79.4**

Table C4. Effect of subcollections of the text corpus.

C.3. Few-shot experiments on Something-Something v2

In tab. C3, we compare ALT-B/16 in the SS-V2 few-shot setting along with methods that adapt the same CLIP-B/16 for videos. We note that ALT-B/16 consistently surpasses these methods across different shot settings, which demonstrates the effectiveness of our video adapter. However, We acknowledge that existing approaches, which utilize image-language models and primarily emphasize visual semantics, do not provide satisfactory solutions in **low-shot** scenarios for datasets that are motion-heavy or instance-agnostic, such as Something Something v2. Instead, we believe that well-designed temporal modules (including other modalities and motion cues extraction) are crucial for achieving higher performance levels, which are deserving of future exploration.

C.4. Investigation of types text corpus

To further validate the effect of text corpus, we set a baseline model by replacing aligned embeddings of text entities $\mathbf{Q}_0$ (Eq. 7) with random learnable queries. The result is reported in the first row of Tab. C4 (2-shot and 0-shot experiments take 32 frames per video as input). Moreover, we evaluate the effectiveness of each sub-collection of text corpus by categorizing the text entities into four groups: object, body parts, scenes, and primitive motion. We find that each category is helpful, and the models with all text entities further outperformed the baseline, especially in the 2-shot (+*11.5%*) and 0-shot (+*7.8%*) experiments. The results reveal that our categorized text entities are complementary to each other, and semantic alignments promise more robust visual representations when facing a severe lack of data.

r	GFLOPs	Fully.(%)
0	141	83.1
4	129	83.0
8	110	82.8
13	86	82.2

Table C5. Trade-off between efficiency & accuracy. r: the number of tokens to reduce in each transformer block. Results are reported in a single view.

C.5. Efficiency and accuracy trade-off

By default, we set the number of token reductions per block in the image encoder r to 8. Here we further investigate the performance of varying r. As shown in Tab. C5, our method achieves the highest accuracy in fully-supervised experiments without the token merging strategy (also no region-aware semantic alignment.) As r increases, the consumption of computing decreases gradually, but so does the accuracy. On balance, r=8 is a cost-effective choice, it is worthwhile to explore and validate additional configurations for low-shot scenarios in future studies.

References

[1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lucic, and Cordelia Schmid. Vivit: A video vision transformer. In *ICCV*, pages 6816–6826, 2021. 1

[2] Pierpaolo Basile, Annalina Caputo, and Giovanni Semeraro. An enhanced lesk word sense disambiguation algorithm through a distributional semantic model. In *COLING*, pages 1591–1600. ACL, 2014. 1

[3] Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python*. O'Reilly, 2009. 1

[4] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *ICCV*, pages 6804–6815, 2021. 1

[5] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fründ, Peter Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. The "something something" video database for learning and evaluating visual common sense. In *ICCV*, pages 5843–5851. IEEE Computer Society, 2017. 1

[6] Matthew Honnibal and Ines Montani. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear, 2017. 1

[7] Chen Ju, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie. Prompting visual-language models for efficient video understanding, 2022. 2

[8] Will Kay, João Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *CoRR*, abs/1705.06950, 2017. 1

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

[10] Ziyi Lin, Shijie Geng, Renrui Zhang, Peng Gao, Gerard de Melo, Xiaogang Wang, Jifeng Dai, Yu Qiao, and Hongsheng Li. Frozen CLIP models are efficient video learners, 2022. 1, 2

[11] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *CVPR*, pages 3192–3201, 2022. 1

[12] George A. Miller. Wordnet: A lexical database for english. *Commun. ACM*, pages 39–41, 1995. 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 (4)*, pages 1–18, 2022. 1, 2

[14] OpenAI. Chatgpt: Optimizing language models for dialogue. https://chat.openai.com/, 2022. 1

[15] Junting Pan, Ziyi Lin, Xiatian Zhu, Jing Shao, and Hongsheng Li. St-adapter: Parameter-efficient image-to-video transfer learning for action recognition. In *NeurIPS*, 2022. 1, 2

[16] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021. 2

[17] Hanoona Abdul Rasheed, Muhammad Uzair Khattak, Muhammad Maaz, Salman H. Khan, and Fahad Shahbaz Khan. Fine-tuned CLIP models are efficient video learners. In *CVPR*, pages 6545–6554. IEEE, 2023. 2

[18] Michael S. Ryoo, A. J. Piergiovanni, Anurag Arnab, Mostafa Dehghani, and Anelia Angelova. Tokenlearner: Adaptive space-time tokenization for videos. In *NeurIPS*, pages 12786–12797, 2021. 1

[19] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *CoRR*, abs/1212.0402, 2012. 1

[20] Jan Philip Wahle, Terry Ruas, Norman Meuschke, and Bela Gipp. Incorporating word sense disambiguation in neural language models. *arXiv preprint arXiv:2106.07967*, 2021. 1

[21] Mengmeng Wang, Jiazheng Xing, and Yong Liu. Actionclip: A new paradigm for video action recognition. *CoRR*, abs/2109.08472, 2021. 2

[22] Shen Yan, Xuehan Xiong, Anurag Arnab, Zhichao Lu, Mi Zhang, Chen Sun, and Cordelia Schmid. Multiview transformers for video recognition. In *CVPR*, pages 3323–3333, 2022. 1

[23] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, Ce Liu, Mengchen Liu, Zicheng Liu, Yumao Lu, Yu Shi, Lijuan Wang, Jianfeng Wang, Bin Xiao, Zhen Xiao, Jianwei Yang, Michael Zeng, Luowei Zhou, and Pengchuan Zhang. Florence: A new foundation model for computer vision. *CoRR*, abs/2111.11432, 2021. 1

[24] Bowen Zhang, Jiahui Yu, Christopher Fifty, Wei Han, Andrew M. Dai, Ruoming Pang, and Fei Sha. Co-training transformer with videos and images improves action recognition. *CoRR*, abs/2112.07175, 2021. 1