Enhanced Motion-Text Alignment for Image-to-Video Transfer Learning

Supplementary Material

Overview

In the supplementary material, we first provide more ablation studies (in Sec. A.1) and visualization results (in Sec. A.2). The implementation details for the experiments on SSv2 and K400 are presented in Sec. A.3.

A. Appendix

A.1. Ablation Studies

Hyper-parameter α. To validate the flexibility of input frames in the temporal pathway, we conduct the experiments of varying the upsampling ratio α in Tab. 1a. It's obvious that the higher α can consistently improve the performances, especially on temporally-heavy dataset SSv2 [4]. We assume this is because the more input frames contain richer temporal dynamic information, facilitating the alignment with motion-enhanced descriptions. However, the higher α can also bring up more computational cost (*i.e.*, FLOPs). To achieve a better trade-off between computation efficiency and prediction accuracy, we set $\alpha = 2$ by default.

Top-k similar categories. The experiments in Tab. 1b explore the impact of different number of similar categories in generating discriminative motion descriptions. It is observed that, when applying more discriminative motion descriptions ($k = 1 \rightarrow 5$), we can achieve noticeable performance improvement on both datasets, especially on SSv2 (+3.5%), with negligible additional computational costs. However, comparing the performance with $k = 5$ and $k = 10$, the performance gains are relatively minor. In this work, we choose $k = 5$ for simplicity.

Motion encoder backbones. We apply the motion encoder in the temporal pathway to extract temporal dynamic information within the dense frames. In Tab. 1c, we conduct experiments on different designs of motion encoder, including the convolution-based R(2+1)D [13], cross-frame attention-based X-CLIP [7], and joint spatio-temporal attention-based VideoMAE [12]. The results reveal that: i) our framework is flexible and capable of integrating with different temporal networks; ii) our customized motion encoder achieves a relatively better performance. X-CLIP mainly focuses on the global long-term temporal information between frames, while VideoMAE mainly learns the dependencies between tokenized 3D cubes [12]. In contrast, our motion encoder has the capacity to simultaneously learn the global cross-frame dependencies between image features and the local key region interactions.

A.2. Visualization Cases

In Fig. 1, we show the confusion matrices for SSv2 classification using the models trained *with*/*without* the motion-enhanced descriptions. We select the categories containing similar sentence semantics, which are started with "Pushing something" or "Pulling something". Specifically, without motion-enhanced descriptions, the model is confused to differentiate the classes with fine-grained motions, two of them correspond to "Pulling something from behind of something" and "Pulling something out of something". These two categories have the same action of "pulling", but differ in the moving directions. In contrast, the model discriminates these two classes correctly, by integrating the motion-enhanced descriptions. This phenomenon reveals that our proposed motion-enhanced descriptions can contribute to stronger discrimination between easy-confusing categories.

Fig. 2 investigates the learned patterns of the spatial and temporal pathways, based on the reasoning tool[1] [3]. An intriguing finding is that the image encoder mainly focuses on the static visual contents (*e.g.*, "pieces", "paper"), while the motion encoder is capable of perceiving and tracking the moving objects corresponding to the motion-related words (*e.g.*, "tearing", "falling"). This phenomenon reveals the image stream and motion stream learn different patterns and complement each other to generate the integrated visual representations for each video clip.

A.3. Implementation Details

As shown in Tab. 2, we present the training hyperparameters for the experiments in the main manuscript on SSv2 and K400. The data augmentations (*e.g.*, ColorJitter, GrayScale) are available in PyTorch [8] torchvision package. In most of the various experimental settings, the shared configurations illustrate the remarkable adaptability of our proposed MoTED.

Supervised experiments. We conduct the fully-supervised experiments on K400 and SSv2. The complete training and validation sets are utilized for training and inference, respectively. Following prior works [14], we perform uniform sampling to obtain each temporal clip. For K400 dataset, we scale the shorter side of each frame in spatial resolution to 256 and take a 224×224 center crop. Following [1, 9], we adopt the multi-view inference with 1 spatial crop and 3 temporal clips.

Zero-shot experiments. Following the recipes in [7], we train MoTED (ViT-B/16) with 32 frames on K400 and adopt the single-view inference. We apply the following two

[1] https://github.com/hila-chefer/Transformer-MM-Explainability

Spat.	Temp.	α	SSv2	K400	GFLOPs
	8f	1	68.6	84.5	**176**
8f	16f	2	70.1	**85.1**	184
	32f	4	70.4	84.9	193
	64f	8	**70.5**	85.0	212

(a) Varying values of α, *i.e.*, the number of input frames in the temporal pathway.

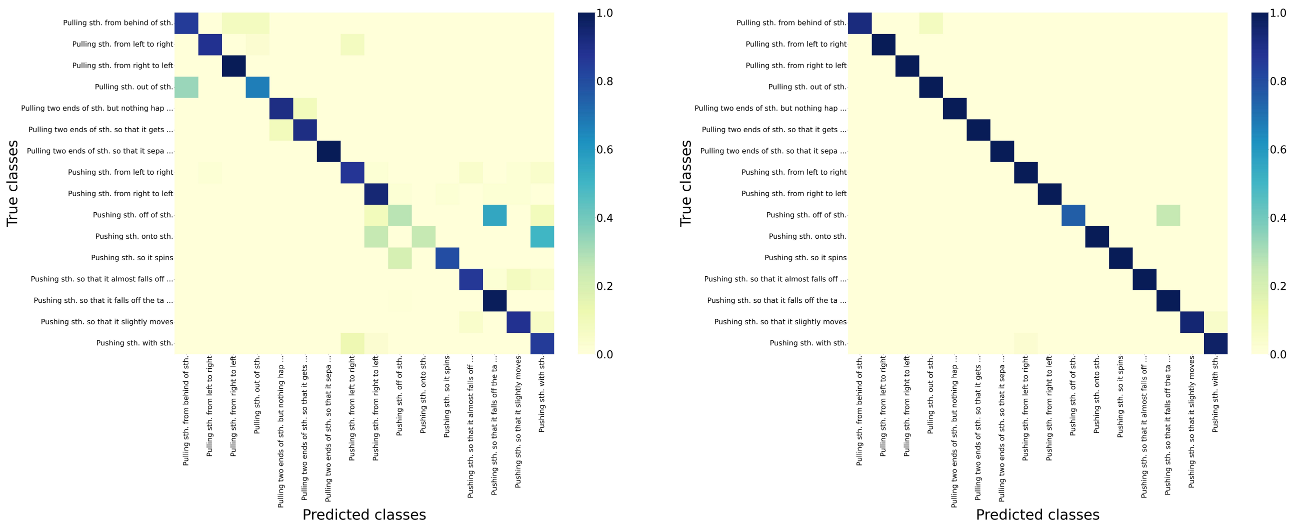
Top-k	SSv2	K400	GFLOPs
1	66.6	84.6	**180**
3	68.9	84.9	182
5	**70.1**	**85.1**	184
7	69.9	84.9	186
10	70.1	84.8	189

(b) Varying values of k, *i.e.*, the number of similar classes in generating descriptions.

Motion Encoder	SSv2	K400	GFLOPs
R(2+1)D [13]	68.3	83.4	**164**
X-CLIP [7]	68.6	84.6	167
VideoMAE [12]	69.7	83.4	170
Ours	**70.1**	**85.1**	184

(c) Alternative choices of the motion encoder.

Table 1. Ablations on **Something-Something V2** and **Kinetics-400**. The spatial encoder is a 8-frame vanilla ViT-B/16 pre-trained by CLIP [10]. The inference protocol of all models and datasets are 3 clips $\times$ 1 center crop.



(a) Confusion matrix **without** MoTED (b) Confusion matrix **with** MoTED

Figure 1. The comparison between the confusion matrices of the model trained *without*/*with* the motion-enhanced descriptions on SSv2 dataset [4]. We select the categories with similar semantics, starting with “Pushing something” or “Pulling something”.

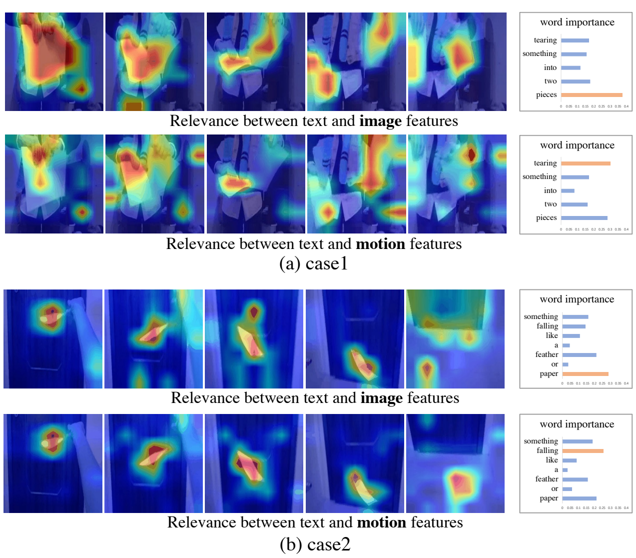


Figure 2. Two cases to visualize the relevance [3] between text and image/motion features to highlight the information relevant to the prediction. The different “regions of interest” and “words of importance” indicate that the motion and image features could be disentangled.

evaluation protocols in our zero-shot experiments. (1) For HMDB-51 and UCF-101, following [10], the prediction is conducted on the three splits of the test data, and we report the average top-1 accuracy and standard deviation. (2) For Kinetics-600, following [7], the 220 new categories outside K400 are used for evaluation. The evaluation is conducted three times. For each iteration, we randomly sampled 160 categories for evaluation from the 220 categories in Kinetics-600.

Motion description generation. To generate the required motion-enhanced descriptions, we first query large language models (LLMs) with one question: ‘`Q: What is the motion concept in a video of <category name>? A:`’, but it could result in the answers with duplicated sentences, such as: ‘`The motion concept of slapping involves striking someone or something with an open hand, usually in a quick and forceful manner. This motion involves a swift movement using the open hand or fingers that usually involves striking against something, primarily as a`

Table 2. The training hyperparameters on SSv2 and K400. Note that, "Lr." is the abbreviation of "learning rate".

settings	SSv2	K400
Optimization		
Optimizer [5]	AdamW	AdamW
Optimizer betas	(0.9, 0.98)	(0.9, 0.98)
Batch size	256	256
Lr. schedule [6]	cosine decay	cosine decay
Warmup schedule	linear	linear
Linear warmup	5	5
Base Lr.	1e-4	8e-5
Minimal Lr.	8e-7	8e-7
Weight decay	1e-3	1e-3
Epoches	40	40
Data augmentation		
RandomFlip	None	0.5
MultiScaleCrop [15]	None	(1, 0.875, 0.75, 0.66)
ColorJitter	0.8	0.8
GrayScale	0.2	0.2
Label smoothing [11]	0.1	0.1
Mixup [17]	0.8	0.8
Cutmix [16]	1.0	1.0

`way of getting attention or causing discomfort. It's worth noting that slapping can cause pain, discomfort, or injury, depending on the force and target area.`' The generated first two sentences share similar semantics and the last sentence describes the impact of the action. In this way, it could bring up additional costs and unnecessary noises. Thus, we take a two-shot prompt technique [2] to control the generated descriptions to be concise and motion-related.

References

[1] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, page 4, 2021. 1

[2] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *NeurIPS*, 2020. 3

[3] Hila Chefer, Shir Gur, and Lior Wolf. Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers. In *ICCV*, pages 387–396, 2021. 1, 2

[4] 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, 2

[5] Ilya Loshchilov and Frank Hutter. Fixing weight decay regularization in adam. *arXiv preprint arXiv:1711.05101*, 2017. 3

[6] Ilya Loshchilov and Frank Hutter. SGDR: stochastic gradient descent with warm restarts. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017. 3

[7] 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

[8] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Z. Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, pages 8024–8035, 2019. 1

[9] Zhiwu Qing, Shiwei Zhang, Ziyuan Huang, Yingya Zhang, Changxin Gao, Deli Zhao, and Nong Sang. Disentangling spatial and temporal learning for efficient image-to-video transfer learning. *ICCV*, 2023. 1

[10] 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 *ICML*, pages 8748–8763. PMLR, 2021. 2

[11] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, pages 2818–2826, 2016. 3

[12] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *NeurIPS*, 2022. 1, 2

[13] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *CVPR*, pages 6450–6459, 2018. 1, 2

[14] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks for action recognition in videos. *TPAMI*, 41(11): 2740–2755, 2018. 1

[15] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks for action recognition in videos. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(11):2740–2755, 2019. 3

[16] Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Seong Joon Oh, Youngjoon Yoo, and Junsuk Choe. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *ICCV*, pages 6022–6031, 2019. 3

[17] Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *ICLR*. OpenReview.net, 2018. 3