Video Recognition in Portrait Mode

Supplementary Material

1. Implementation details

1.1. Training recipes

In addition to the frame resizing and input cropping strategy, we provide additional training details. Uniformer-S [4] is trained for 100 epochs with a learning rate of 0.0002, using a batch size of 96×4 (the number of clips per GPU and the number of GPUs used for training). X3D-M [2] is trained for 300 epochs with a learning rate of 0.2 and a batch size of 64×4. MViTv2-S [1] is trained for 200 epochs with a learning rate of 0.0001 and a batch size of 32×4. We keep the other settings, such as repeated augmentation [3], and learning rate decay identical to those mentioned in their original papers.

1.2. Testing recipes

To ensure a fair comparison of all results presented in our paper, we conducted experiments using a resolution of 224×224, unless otherwise stated. For example, we used a rectangular crop input in Table 4 and conducted spatial prior experiments in Section 4.2.

We apply identical test-time temporal and spatial augmentation methods as described in the original papers. Specifically, for X3D-M, we use 10 temporal views with 3 spatial crops. For Uniformer-S, we use 4 temporal views with 1 spatial crop, and for MViTv2-S, we use 5 temporal views with 1 spatial crop. Readers who seek more detailed information on the augmentation techniques used can refer to the original papers.

2. PortraitMode-400

In addition to the dataset statistics presented in the main submission, we offer further insights into our dataset through visualizations. Figure S.1 displays the duration distribution of videos, while Figure S.2 depicts accuracy per category. These visualizations demonstrate that our dataset is well-balanced and suitable for training models that can handle a wide range of scenarios. Furthermore, we have created a webpage in *demo* to highlight the diversity of our taxonomy and the unique characteristics of our videos. The *demo* is included in the supplemental zip file.

2.1. Duration per category

Understanding the duration distribution of a video dataset is critical in developing models for video recognition tasks. The distribution of durations can guide the selection of appropriate temporal scales for video analysis, including window sizes or sampling rates. For our dataset, Figure S.1 illustrates a balanced distribution of video durations, providing valuable insights for optimizing video recognition models in portrait mode.

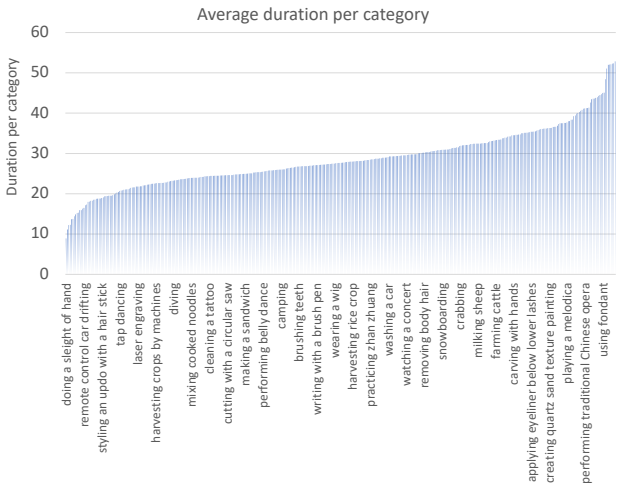

Figure S.1. Average duration per category of PortraitMode-400. The horizontal axis indicates the category name, and the vertical axis represents the average duration per category.

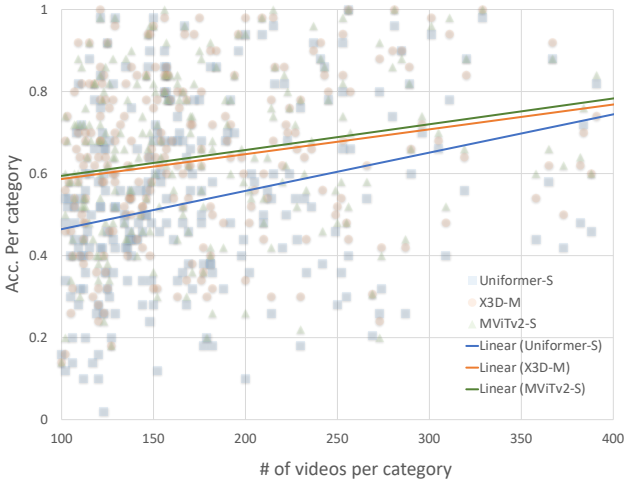

Figure S.2. Accuracy per category of PortraitMode-400 with Uniformer-S, X3D-M and MViTv2-S. The horizontal axis indicates the number of videos per category, and the vertical axis represents the accuracy of each respective category. The graph includes individual data points and a linear trend line.

2.2. Number of videos per category

In general, having a larger training video set can improve performance in video recognition tasks due to increased prior knowledge. Our experiments with three different architectures on our dataset further validate this observation. Figure S.2 displays the accuracy and number of videos for each category, along with a linear trend line. From the trend lines of the three models, we can infer that increased training video volume leads to improved testing performance.

2.3. Dataset glance

To concisely present our dataset, we have selected one representative video from each category across four of our nine domains. Access to these videos is available through the *index.html* page in the *demo* directory, where clicking on the leaf node text redirects to a webpage for video playback.

It's important to note that for CVPR's reviewing process, we provide 40 downsampled video files in the *demo* directory, adhering to the conference's supplementary file size constraints. However, our dataset does not store or distribute raw videos. Instead, for public release, we will provide links to the video sources, requiring users to download the videos themselves. This approach ensures compliance with distribution guidelines while facilitating ease of access and review.

3. Zero-shot video recognition

In this section, we expand our PortraitMode-400 with a zero-shot portrait-mode video recognition dataset, introducing 100 newly curated categories specifically designed for zero-shot recognition. Our dataset embodies critical aspects of portrait mode video recognition, notably spatial priors, temporal information, and audio modality. This enrichment not only challenges but also broadens the horizons of zero-shot video recognition research.

Model	Pretrain	Accuracy	Views
CLIP [6]	CLIP400M	50.1	1×1
X-CLIP [5]	CLIP400M+K600	52.5	1×1
X-CLIP [5]	CLIP400M+K600	54.6	3×4

Table S.1. Performance on our reserved zero-shot subset. Views during inference are shown by the multiplication of # of spatial crops and # of temporal views.

As shown in Table S.1, to facilitate initial explorations, we establish baselines using the CLIP model [6] with a mean pooling temporal aggregation method. Additionally, we present a comparative analysis with XCLIP [5]'s performance. We aim for this dataset to drive forward research and innovation in this domain.

4. Broader impact

Data Limitations and Ethical Considerations. We do not store or distribute videos; users must obtain them from the original sources. Furthermore, our careful manual annotation process is designed to prevent any ethical or legal issues.

Human Rights in Annotation Process. We have carefully organized the annotation task to guarantee reasonable workloads and just remuneration for annotators, adhering to human rights principles.

Scope of Conclusions. It is essential to be aware that experiments and data, including ours, may only be a small part of the whole picture. Nevertheless, due to the extensive range of data, such as 3Massiv, S100 and PortraitMode-400, that we have used in our experiments, we are confident that our results offer a reliable comprehension that can be applied to portrait-mode video analysis. Although these discoveries are particular to the data we have examined, they offer considerable insight into the wider area of video analysis.

Future Research and Development. Aligned with our commitment to the research community and in adherence to CVPR guidelines, we will release both our code and dataset. This is intended to encourage further research and enable others to build upon our work.

References

[1] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. *arXiv preprint arXiv:2104.11227*, 2021. 1

[2] Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 203–213, 2020. 1

[3] Elad Hoffer, Tal Ben-Nun, Itay Hubara, Niv Giladi, Torsten Hoefler, and Daniel Soudry. Augment your batch: Improving generalization through instance repetition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8129–8138, 2020. 1

[4] Kunchang Li, Yali Wang, Junhao Zhang, Peng Gao, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Qiao. Uniformer: Unifying convolution and self-attention for visual recognition, 2022. 1

[5] 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. 2022. 2

[6] 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