## A. Limitations and future work

Although we have improved upon the state-of-the-art, there is still a large room for improvement on datasets other than Kinetics. Furthermore, we have relied on models pretrained on large image- or video-datasets for initialization. Reducing this dependence on supervised pretraining is a clear avenue of future research. We have conducted thorough ablations on standard transformer architectures [1, 4], and will investigate if our approach is complementary to recent, spatial-pyramid based multiscale transformer encoders such as MViT [5] and Swin [7].

**Societal impact**. Video classification models can be used in a wide range of applications. We are unaware of all potential applications, but are mindful that each application has its own merits, and that also depends on the intentions of the individuals building and using these systems. We also note that training datasets may contain biases that models trained on them are unsuitable for certain applications.

## B. Additional experiments

In this supplementary, we provide additional experimental details. Section B.1 provides accuracy-FLOPs and accuracy-throughput comparison between two model variants of ViViT and MTV. Section B.2 provides the effect of spatial resolution of tubelets. Section B.3 and Section B.4 provides details of our training hyperparameters and model configurations used in our experiments.

### B.1. Changing transformer encoder architecture

We present additional results by changing the transformer architecture used within our multiview encoder. Specifically, we use the unfactorized ViViT transformer encoder (Model 1 of [1]). In this variant, each transformer encoder layer computes self-attention over all spatio-temporal tokens. This makes our multiview transformer encoder cover a wide range of spatial and temporal dimensions across different views. A one-layer MLP with hidden dimension of 3072 is used as the global encoder for our unfactorized MTV model.

As shown in Fig. 1, MTV (unfactorized) consistently outperforms its single-view counterpart (*i.e.* ViViT unfactorized) for every scale (see Fig. 1a) and corresponds to a better accuracy-throughput curve as shown in Fig. 1b. Note how MTV can more than double the throughput of ViViT unfactorized, whilst still improving its accuracy, for each model scale. Specifically, MTV (unfactorized) H/4+B/8+S/16+Ti/32 model leads to a significant speed-up by $172\%$ while still keeping a higher accuracy of $0.4\%$ improvement compared to ViViT-H.

Moreover, we report the accuracy-throughput comparison between MTV and ViViT factorized model (ViViT-FE) in Fig. 1d. Note that the accuracy-FLOPs comparison is already reported in paper Section 4.3. The improvements in accuracy-throughput, and accuracy-FLOPs remain significant in this setting.

Table 1. Effect of spatial resolution of tubelets. All experiments are conducted on Kinetics 400 using the model variant B/4+Ti/16. Accuracies are for $4 \times 3$ crops.

| Tubelet spatial size | | GFLOPs | Top-1 |
|---|---|---|---|
| B | Ti | | |
| $24 \times 24$ | $16 \times 16$ | 68 | 78.1 |
| $16 \times 16$ | $24 \times 24$ | 165 | 80.5 |
| $16 \times 16$ | $16 \times 16$ | 168 | 80.5 |
| $16 \times 16$ | $12 \times 12$ | 169 | 80.6 |
| $12 \times 12$ | $16 \times 16$ | 295 | 81.0 |

Note that the unfactorized ViViT transformer encoder, which attends to all spatio-temporal tokens, is less efficient than the Factorized Encoder architecture that we used in the main paper. However, we achieve larger relative improvements in accuracy/computation trade-offs compared to the corrsponding single-view ViViT baseline when using this encoder architecture.

### B.2. Spatial resolution of tubelets

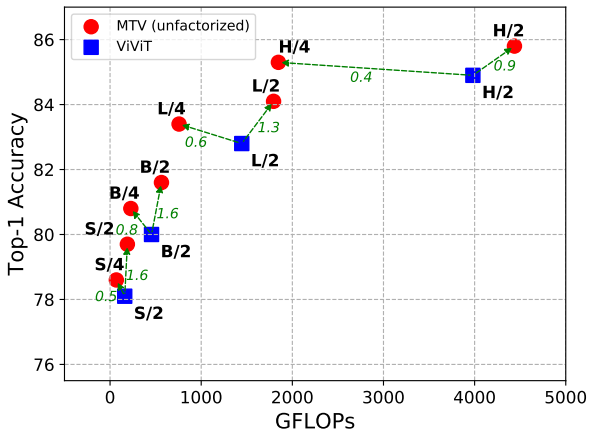
We study the effect of the spatial resolution of tubelets in Tab. 1. We use our B/4 + Ti/16 model variant, and vary the spatial resolution of the tubelets. Our results indicate that the accuracy is primarily impacted by the spatial resolution of the large encoder. We also note that processing more tokens, and thus using more computation, typically results in higher accuracies.

### B.3. Hyperparameters for each datasets

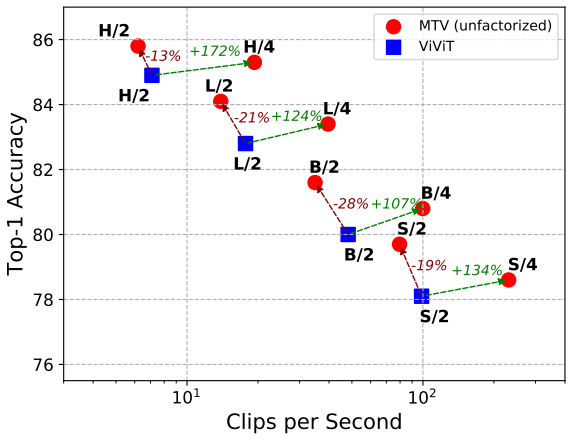
Table 2 details the hyperparamters used in all of our experiments. We use synchronous SGD with momentum, a cosine learning rate schedule with linear warmup, and a batch size of 64 for all experiments on the Kinetic datasets. We found that larger batch size and additional regularization are helpful when training on the smaller Epic Kitchens and Something-Something v2 datasets, as also noted by [1].

### B.4. Model configurations

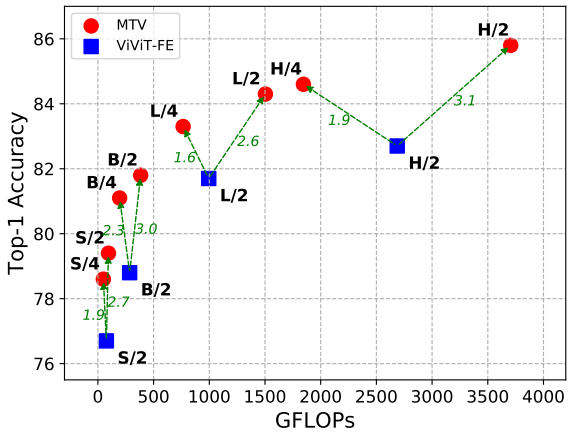
Table 3 summarizes our model configurations of each view for our multiview transformer encoder. For the backbone of each view, we consider five ViT variants, "Tiny", "Small", "Base", "Large", and "Huge". Their settings strictly follow the ones defined in BERT [3] and ViT [4, 8]. For the global encoder, all model variants of MTV use the same global encoder which follows the "Base" architecture, except that the number of heads is set to 8 instead of 12. The reason is that the hidden dimension of the tokens should be divisible by the number of heads for multi-head attention, and the number of hidden dimensions across all backbone sizes is divisible by 8 (as shown in Tab. 3). All model variants of MTV (unfactorized) use a one-layer MLP with the same hidden dimension as the "Base" architecture.



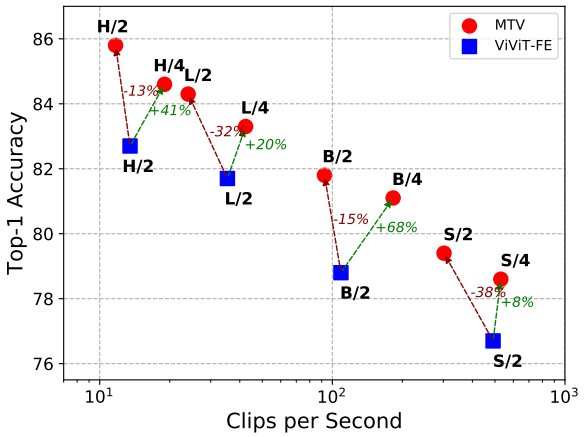

(a) Accuracy[%] - GFLOPs comparison between MTV (unfactorized) and ViViT.




(b) Accuracy[%] - Throughput comparison between MTV (unfactorized) and ViViT.




(c) Accuracy[%] - GFLOPs comparison between MTV and ViViT-FE.




(d) Accuracy[%] - Throughput comparison between MTV and ViViT-FE.

Figure 1. Accuracy/complexity trade-off between ViViT / ViViT-FE [1] (blue) and our MTV (unfactorized) / MTV (red). MTV (unfactorized) is consistently better and requires less FLOPs (see Fig. 1a) than ViViT to achieve higher accuracy across different model scales (indicated by the dotted green arrows pointing upper-left). With additional FLOPs, MTV shows larger accuracy gains (shown by the dotted green arrows pointing upper-right). The lower number of FLOPs is translated to higher throughput (clips per second), as indicated by the green arrows in Fig. 1b. Note how MTV can more than double the throughput of ViViT unfactorized, whilst still improving its accuracy, across all model scales. Similar findings are also observed by the comparison between ViViT-FE and MTV model in Fig. 1c and Fig. 1d. Note that Fig. 1c appeared as Figure 3 in the main paper, and is included here for clarity and consistency. All speed comparisons are measured with the same hardware (Cloud TPU-v4). The complexity is for a single $32 \times 224 \times 224 \times 3$ input video (denoted as $T \times H \times W \times C$), and the accuracy is obtained by $4 \times 3$ view testing.

Table 2. Training hyperparamters for experiments in the main paper. "–" indicates that the regularisation method was not used at all. Values which are constant across all columns are listed once. Datasets are denoted as follows: K400: Kinetics 400. K600: Kinetics 600. K700: Kinetics 700. MiT: Moments in Time. EK: Epic Kitchens. SSv2: Something-Something v2.

| | K400 | K600 | K700 | MiT | EK | SSv2 |
|---|---|---|---|---|---|---|
| *Optimization* | | | | | | |
| Optimizer | Synchronous SGD | | | | | |
| Momentum | 0.9 | | | | | |
| Batch size | 64 | 64 | 64 | 256 | 128 | 512 |
| Learning rate schedule | cosine with linear warmup | | | | | |
| Linear warmup epochs | 2.5 | | | | | |
| Base learning rate | 0.1 | 0.1 | 0.1 | 0.1 | 0.2 | 0.5 |
| Epochs | 30 | 30 | 30 | 30 | 80 | 100 |
| *Data augmentation* | | | | | | |
| Random crop probability | 1.0 | | | | | |
| Random flip probability | 0.5 | 0.5 | 0.5 | 0.5 | 0.5 | – |
| Scale jitter probability | 1.0 | | | | | |
| Maximum scale | 1.33 | | | | | |
| Minimum scale | 0.9 | | | | | |
| Colour jitter probability | 0.8 | 0.8 | 0.8 | 0.8 | – | – |
| Rand augment number of layers [2] | – | – | – | – | 3 | 1 |
| Rand augment magnitude [2] | – | – | – | – | 10 | 15 |
| *Other regularisation* | | | | | | |
| Stochastic droplayer rate [6] | 0.1 | 0.1 | 0.1 | 0.1 | 0.1 | 0.3 |
| Label smoothing [9] | – | – | – | – | 0.2 | 0.2 |
| Mixup [10] | – | – | – | – | 0.1 | 0.3 |

Table 3. Model configurations for each view of MTV.

| Model name | Hidden size | MLP dimension | Number of attention heads | Number of encoder layers | Tubelet spatial size |
|---|---|---|---|---|---|
| Tiny | 192 | 768 | 3 | 12 | 16 |
| Small | 384 | 1536 | 6 | 12 | 16 |
| Base | 768 | 3072 | 12 | 12 | 16 |
| Large | 1024 | 4096 | 16 | 24 | 16 |
| Huge | 1280 | 5120 | 16 | 32 | 14 |

## References


[1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. ViViT: A video vision transformer. In *ICCV*, 2021. 1, 2

[2] Ekin D. Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V. Le. Randaugment: Practical automated data augmentation with a reduced search space. In *NeurIPS*, 2020. 3

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*, 2019. 1

[4] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021. 1

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

[6] Gao Huang, Yu Sun, Zhuang Liu, Daniel Sedra, and Kilian Weinberger. Deep networks with stochastic depth. In *ECCV*, 2016. 3

[7] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 1

[8] Andreas Steiner, Alexander Kolesnikov, Xiaohua Zhai, Ross Wightman, Jakob Uszkoreit, and Lucas Beyer. How to train your ViT? Data, augmentation, and regularization in vision transformers. In *arXiv preprint arXiv:2106.10270*, 2021. 1

[9] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, 2016. 3

[10] Hongyi Zhang, Moustapha Cisse, Yann N. Dauphin, and David Lopez-Paz. Mixup: Beyond empirical risk minimization. In *ICLR*, 2018. 3