Supplementary Material for Vision Transformers are Parameter-Efficient Audio-Visual Learners

Our supplementary material consists of:

1. Implementation Details.
2. Additional Quantitative Results.

1. Implementation Details

For all of our experiments, we extract the visual frames at 1 fps. As our best performing model, we adopt a pre-trained Swin-V2-Large [8] with a 192×192 spatial resolution with all parameters frozen. For the audio-visual event localization task, we implement our LAVISH adapter with 2 latent tokens and the downsampling factor of 8 in the 2D group convolutional adapter layers, where the number of group convolutions is set to 2. Our group convolution adapter layers use only 0.5x parameters as the standard fully connected ones. For the audio-visual segmentation and audio-visual question-answering tasks on AVSBench-S4 and MUSIC-AVQA, we use 16 latent tokens and set the downsampling rate and the number of group convolutions to 4 and 2, respectively. For all of our experiments, we use Adam [5] optimizer to train our model. We set the learning rate of LAVISH adapter to $5e{-}6$ and $4e{-}6$ for the final prediction layer for audio-visual event localization, $1e-4$ for audio-visual segmentation, and $8e-5$ for LAVISH adapter and $3e{-}6$ for the grounding modules and the final prediction layer in audio-visual question answering. For audio preprocessing, we compute the audio spectrogram by PyTorch [10] kaldi fbank with 192 triangular mel-frequency bins and frameshift in 5.2 milliseconds. Then, we inflate the input channel of the audio spectrogram from 1 to 3 to match the dimensions of a linear patch projection layer in SwinV2.

Task	Batch Size	Num. Latent Tokens	Downsampling Factor
AVE	2	2	8
AVS	4	16	4
AVQA	1	16	4

2. Additional Quantitative Results

Comparing ViT and ResNet-152 Backbones. To investigate whether a visual transformer backbone is truly necessary for adapting a frozen visual model to an audio-visual task, we also conduct experiments with a ResNet-152 backbone [3]. We report that compared to a ViT-B [2] (86M params), using a ResNet-152 backbone (60M params) leads to a significant **18**% drop in accuracy. To make the comparison fairer in terms of a model's capacity, we also report the results using ViT-tiny (6M params) and ViT-small (23M params) architectures, which both have a smaller capacity than ResNet-152. We observe that in both of these cases, the ViT variants outperform ResNet-152 (by **5.4** % and **13.9**% respectively. These results demonstrate that the lack of inductive biases in visual transformer models enables more effective transfer between inputs across different modalities.

Table 1: **Comparison with Visual-only Variants.** We compare our audio-visual approach with visual-only variants on three audio-visual understanding tasks: audio-visual event localization (AVE), audio-visual segmentation (AVS), and audio-visual question answering (AVQA). As evaluation metrics, we use top-1 accuracy, mean intersection over union (mIoU), and top-1 accuracy for all three tasks respectively. Our results indicate that our model benefits significantly from jointly modeling audio and visual cues.

Task	Input Modality	Accuracy ↑
AVE [13]	Vision	75.3
	Audio+Vision	**81.1**
AVS [16]	Vision	72.1
	Audio+Vision	**80.1**
AVQA [7]	Vision	63.2
	Audio+Vision	**77.1**

Comparison with Visual-Only Baselines. To verify the importance of jointly considering audio-visual information in all three of our considered benchmarks/tasks (i.e., audio-visual event localization (AVE), audio-visual segmentation (AVS), and audio-visual question-answering (AVQA)), we compare our audio-visual approach with the visual-only

Table 2: **Audio-visual Action Recognition.** We evaluate our LAVISH approach on the UCF101 [12] dataset for audio-visual action recognition task. Compared to prior audio-visual approaches, LAVISH achieves the best action recognition accuracy while using the smallest number of trainable parameters.

Method	Visual Encoder	Audio Encoder	Pretrain Data	Trainable Params (M) ↓	Samples per Sec. ↑	Acc ↑
XDC [1]	R(2+1)D 🔥	ResNet-18 🔥	Kinetics-400 (A+V)	45	-	86.8
AVTS [6]	R(2+1)D 🔥	ResNet-18 🔥	Kinetics-400 (A+V)	45	-	86.2
GDT [11]	R(2+1)D 🔥	ResNet-9 🔥	Kinetics-400 (A+V)	39.2	-	89.3
MBT [9]	ViT-B 🔥	AST-B 🔥	Kinetics-400 (V) + AudioSet (A)	172	4.42	91.8
LAVISH	ViT-B❄ (shared)		Kinetics-400 (V)	**7.4**	**6.36**	**92.6**

Table 3: **Is LAVISH Complementary to Pretrained Audio Encoders?** We study whether our LAVISH approach can further benefit from audio features obtained using a VGGIsh [4] audio encoder pretrained on the large-scale AudioSet dataset. To do this, we concatenate the pretrained audio features with audio-visual features from our LAVISH approach. These results indicate that combining audio representations from these two sources leads to a slight boost in performance.

Method	Encoders	Visual Pretrain	Audio Pretrain	Acc
LAVISH	Swin-V2-L ❄	ImageNet	✘	81.1
LAVISH	Swin-V2-L ❄ + VGGish ❄	ImageNet	AudioSet	**82.4**

variants that only consider visual information without processing any audio cues. We present these results In Table 1, and report the audio-visual variant of our approach, which jointly considers audio and visual cues, consistently outperforms the visual-only baselines by **5.8%** top-1 acc., **8%** mIoU, and **13.9%** top-1 acc. for the AVE, AVS, and AVQA tasks respectively. These results indicate that our model benefits significantly from the joint modeling of audio and visual cues and also that visual information alone is not enough for achieving state-of-the-art results on these particular audio-visual tasks.

Action Recognition on UCF101. In Table 2, we also test our model on UCF101 action recognition. We implement LAVISH using VideoMAE codebase [14] pretrained on videos only. Compared to XDC, AVTS, and GDT, all of which used large-scale audio-visual pretraining on Kinetics-400, LAVISH achieves better results (**92.6**% vs **86.8**%, **86.8**%, and **89.3**%) with fewer trainable parameters (**7.4M** vs **45M** and **39.2M**) and without any audio-visual pretraining. Our method also outperforms MBT, which uses ViT and AST pretrained on Kinetics and AudioSet, respectively.

Is LAVISH Complementary to Pretrained Audio Encoders? In Table 3, we also study whether our LAVISH approach can further benefit from audio features obtained using an external VGGish [4] audio encoder pretrained on the large-scale AudioSet dataset. To do this, we concatenate the features from the VGGish [4] audio encoder with the audio-visual features from our LAVISH approach and train a linear layer to predict the event category for the audio-visual event localization task. Based on the results in Table 3, we observe that using an external VGGish audio classifier leads to a 1.3% boost in performance. This indicates that our LAVISH adapters and VGGish encode complementary audio information, and combining audio representations from these two sources is beneficial.

Table 4: **Throughput Comparison.** We compare the throughput of our LAVISH with the state-of-the-art CMBS approach. The throughput is measured using the number of samples per second. In addition to achieving higher accuracy, our method is almost 2× faster than CMBS.

Method	Visual Encoder	Audio Encoder	Samples per Sec. ↑	Acc ↑
CMBS [15]	Swin-L❄	VGGish❄	0.72	80.4
LAVISH	Swin-L❄ (shared)		**1.40**	**81.1**

Throughput Comparisons. In Table 4, we also compare LAVISH to CMBS [79] on the same A6000 GPU. Despite using Swin-L for audio (compared to VGGish), LAVISH has better throughput (**1.40** vs. **0.72** samples/sec). This is because, unlike LAVISH, CMBS uses additional temporal modules.

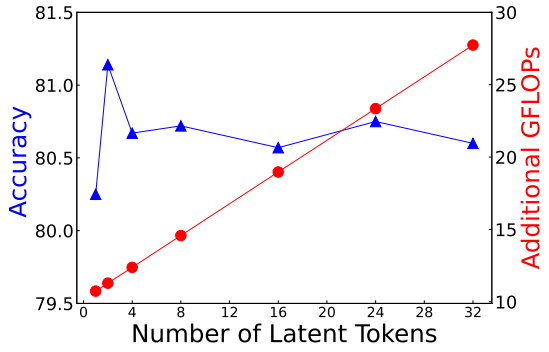

Figure 1: **Number of Latent Tokens.** We investigate the accuracy (in blue) and the computational cost (in GFLOPs) (in red) as a function of the number of latent tokens. LAVISH achieves the best accuracy with two latent tokens. Such a small number of latent tokens enables highly efficient implementation of our approach.

Number of Latent Tokens. Additionally, in Figure 1, we study the performance and computational cost as a function of the number of latent tokens. These results indicate that our model achieves the best accuracy with only two tokens (**81.1%**). Furthermore, we observe that using more latent tokens linearly increases the computational cost but does not yield better results. We conjecture that this happens because the AVE dataset is relatively small, and the model might overfit with more latent tokens. This hypothesis is supported by our results on the larger audio-visual segmentation and audio-visual question answering datasets, where the optimal number of latent tokens is 16. We note that a similar trend has also been reported in prior work [9]. Thus, these results suggest that LAVISH obtains a favorable trade-off between performance and efficiency as cross-modal fusion operation can be implemented very efficiently when few (i.e., 2) latent tokens are used.

References

[1] Humam Alwassel, Dhruv Mahajan, Bruno Korbar, Lorenzo Torresani, Bernard Ghanem, and Du Tran. Self-supervised learning by cross-modal audio-video clustering. In *NeurIPS*, 2020. 2

[2] 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

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 1

[4] Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron Weiss, and Kevin Wilson. Cnn architectures for large-scale audio classification. In *ICASSP*, 2017. 2

[5] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015. 1

[6] Bruno Korbar, Du Tran, and Lorenzo Torresani. Cooperative learning of audio and video models from self-supervised synchronization. In *NeurIPS*, 2018. 2

[7] Guangyao Li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. Learning to answer questions in dynamic audio-visual scenarios. In *CVPR*, 2022. 1

[8] Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In *CVPR*, 2022. 1

[9] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. In *NeurIPS*, 2021. 2, 3

[10] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019. 1

[11] Mandela Patrick, Yuki M Asano, Polina Kuznetsova, Ruth Fong, Joao F Henriques, Geoffrey Zweig, and Andrea Vedaldi. On compositions of transformations in contrastive self-supervised learning. In *ICCV*, 2021. 2

[12] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv Preprint*, 2012. 2

[13] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *ECCV*, 2018. 1

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

[15] Yan Xia and Zhou Zhao. Cross-modal background suppression for audio-visual event localization. In *CVPR*, 2022. 2

[16] Jinxing Zhou, Jianyuan Wang, Jiayi Zhang, Weixuan Sun, Jing Zhang, Stan Birchfield, Dan Guo, Lingpeng Kong, Meng Wang, and Yiran Zhong. Audio–visual segmentation. In *ECCV*, 2022. 1