AVT: AUDIO-VIDEO TRANSFORMER FOR MULTIMODAL ACTION RECOGNITION

Anonymous authors
Paper under double-blind review

ABSTRACT

Action recognition is an essential field for video understanding. To learn from heterogeneous data sources effectively, in this work, we propose a novel multimodal action recognition approach termed Audio-Video Transformer (AVT). AVT uses a combination of video and audio signals to improve action recognition accuracy, leveraging the effective spatio-temporal representation by the *video* Transformer. For multimodal fusion, simply concatenating multimodal tokens in a cross-modal Transformer requires large computational and memory resources, instead we reduce the cross-modality complexity through an audio-video bottleneck Transformer. To improve the learning efficiency of multimodal Transformer, we integrate self-supervised objectives, *i.e.*, audio-video contrastive learning, audio-video matching, and masked audio and video learning, into AVT training, which maps diverse audio and video representations into a common multimodal representation space. We further propose a masked audio segment loss to learn semantic audio activities in AVT. Extensive experiments and ablation studies on three public datasets and two in-house datasets consistently demonstrate the effectiveness of the proposed AVT. Specifically, AVT outperforms its previous state-of-the-art counterparts on Kinetics-Sounds and Epic-Kitchens-100 datasets by 8% and 1%, respectively, without external training data. AVT also surpasses one of the previous state-of-the-art video Transformers (Li et al., 2022a) by 10% on the VGGSound dataset by leveraging the audio signal. Compared to one of the previous state-of-the-art multimodal Transformers (Nagrani et al., 2021), AVT is 1.3× more efficient in terms of FLOPs and improves the accuracy by 4.2% on Epic-Kitchens-100. Visualization results further demonstrate that the audio provides complementary and discriminative features, and our AVT can effectively understand the action from a combination of audio and video.

1 INTRODUCTION

Video understanding has many applications including automated event detection, autonomous robots, video ads, video compliance, *etc*. Deep learning based action recognition methods have been widely explored since the great success of AlexNet on image classification (Krizhevsky et al., 2012; Deng et al., 2009). Conventional deep learning based action recognition can be mainly divided into two aspects: deep ConvNets based methods (Qiu et al., 2019; Feichtenhofer, 2020; Feichtenhofer et al., 2019) and deep sequential learning based methods (Liu et al., 2016; 2017). Deep ConvNets based methods primarily adopt various factorization techniques (Xie et al., 2018; Qiu et al., 2017), or a *priori* (Feichtenhofer et al., 2019) for efficient video understanding (Feichtenhofer, 2020). Some works focus on extracting effective spatio-temporal features (Tran et al., 2015; Carreira & Zisserman, 2017) or capturing complicated long-range dependencies (Wang et al., 2018). Deep sequential and attention models (Liu et al., 2016; 2017) can also be used for spatial and temporal modeling.

Along with the recent advancement of Transformer, several attempts have been made to design video Transformer structures for action recognition (Arnab et al., 2021; Bertasius et al., 2021; Fan et al., 2021; Akbari et al., 2021). Simply applying a Transformer to 3D video domain is computationally expensive (Arnab et al., 2021). The Transformer based spatio-temporal learning methods primarily focus on designing efficient variants by factorization along spatial- and temporal-dimensions (Zha et al., 2021; Arnab et al., 2021; Bertasius et al., 2021), or employing a multiscale pyramid structure for a trade-off between the resolution and channel capacity while reducing the memory and computational

Figure 1: Visualization of one test case of "civil defense siren" in VGGSound. From top to bottom, we show raw video (the 1st row), GradCAM (Selvaraju et al., 2017) of video model, Uniformer (Li et al., 2022a) (the 2nd row), AVBottleneck in Sec. 3.2 (the 3rd row), AVT (ours) with advanced objectives (the 4th row). MViTV2 and AVBottleneck incorrectly predict it as "people whistling" and "planing timber", respectively. Understanding video requires an effective cross-modality learning.

cost (Fan et al., 2021; Li et al., 2022b;a). Multiview Transformer (Yan et al., 2022) employs multiple branches to efficiently learn from various granularities of video views.

Utilizing multimodal signals can help extract more representative and complementary feature representations compared to single modality. For instance, audio is extremely useful for recognizing some actions, *e.g.*, dancing, playing musical instruments. As shown in Table 1, audio modality can improve the action recognition accuracy for some classes compared to one of the state-of-the-art video-only models. Previous multimodal video Transformers generally employ simple *image* Transformers. We leverage the latest *video* Transformer to model complex spatio-temporal features with self-supervised learning, which can fully understand the action from a combination of video and audio input as shown in Fig. 1. An example of multimodal video Transformers, Merlot Reserve (Zellers et al., 2022), conducts audio-vision-language pretraining for holistic multimodal video understanding using an image encoder, word embedding and an audio encoder. MBT (Nagrani et al., 2021) constructs multimodal bottleneck tokens to learn video and audio features from *image* and audio Transformers.

In this work, we propose a novel multimodal video transformer, Audio-Video Transformer (AVT). As illustrated in Fig. 2, AVT effectively employs an audio spectrogram and a frame sequence as input. Then, a video encoder and an audio encoder are employed to extract video and audio representations, respectively. We expect the video encoder to extract complex spatio-temporal representation, which is important to action recognition. Next, we reduce the cross-modality self-attention complexity through training audio-video bottleneck tokens, which can efficiently learn the cross-modality fusion. To make the model fully understand the semantic content from multimodal signals, we design a novel structured masked audio-video reconstruction loss where we force the model to reconstruct a whole audio activity segment. Audio-video contrastive loss and audio-video matching loss are designed to reduce the discrepancy between audio and video representations from the same instance. Our contributions can be summarized as follows:

- We develop the first Audio-*Video* Transformer, AVT, which uses an audio-video bottleneck Transformer (Nagrani et al., 2021) to process the embeddings extracted by an audio Transformer and the latest video Transformer for multimodal action recognition.
- We propose a novel masked audio loss which fully exploits the structure of audio spectrogram and predicts a masked whole audio activity segment. Further, contrastive loss is constructed to align the audio and video embedding (Li et al., 2021), and audio-video matching loss is designed to align embeddings after cross-modality fusion.
- Extensive experiments on three public datasets and two in-house datasets consistently demonstrate that, AVT outperforms previous state-of-the-art video Transformers and simple model fusion across audio and video modalities. AVT achieves better than its previous state-of-the-art counterparts on Kinetics- Sounds (Arandjelovic & Zisserman, 2017) and Epic-Kitchens-100 (Damen et al., 2021a) without external training data.

2 Related Work

Conventional deep learning based action recognition Conventional deep learning based action recognition mainly involves two aspects: deep sequential learning based methods (Liu et al., 2016; 2017) and deep ConvNet based methods (Qiu et al., 2019; Feichtenhofer, 2020; Feichtenhofer et al.,

	March	Waterfall	Tennis	Laugh	Engine	Clock	Speak	Cricket	Wind
Uniformer	**83.3**	46.7	**100.0**	14.0	39.1	57.1	26.8	34.1	10.4
AST	81.3	**60.0**	97.6	**30.2**	39.1	**66.7**	**39.0**	**40.9**	**31.3**

Table 1: Action recognition results using video-only model, Uniformer (Li et al., 2022a), and audio-only model, AST (Gong et al., 2021), on the first nine categories of VGGSound (Chen et al., 2020) demonstrate that video and audio are complimentary. Video does not always outperform audio.

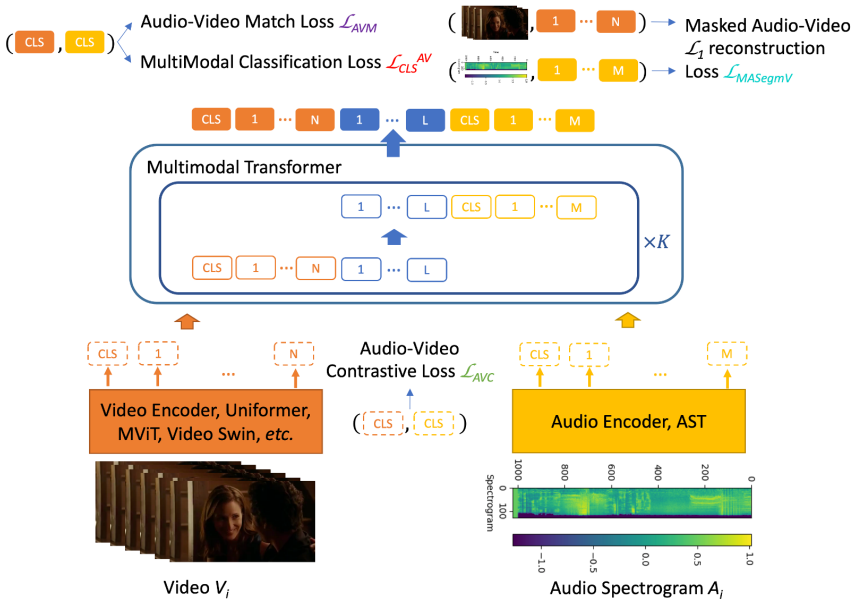

Figure 2: Framework of the Audio-Video Transformer (best viewed in color), AVT, where multimodal inputs are frame sequence V_i and audio spectrogram A_i from the i-th video. Then video encoder and audio encoder are employed to extract video embedding E^V (orange) and audio embedding E^A (yellow), respectively. Next we build audio-video bottleneck tokens, $\{E_1^F, \cdots, E_L^F\}$ (blue), to efficiently learn a cross-modality fusion leveraging self-supervised objectives.

2019). The recurrent networks can be extended to 3D spatio-temporal domain for action recognition (Liu et al., 2016). In deep ConvNet based methods, two-stream ConvNet employs two branches of 2D ConvNets and explicitly models motion by optical flow (Simonyan & Zisserman, 2014). The C3D (Tran et al., 2015) and I3D (Carreira & Zisserman, 2017) directly extend 2D ConvNets to 3D ConvNets, which is natural for 3D spatio-temporal representational learning (Christoph & Pinz, 2016). However, the 3D ConvNet requires significantly more computation and more training data to achieve a desired accuracy. Thus, P3D (Qiu et al., 2017) and S3D (Xie et al., 2018) attempt to factorize the 3D convolution into a 2D spatial convolution and a 1D temporal convolution. SlowFast network (Feichtenhofer et al., 2019) and X3D (Feichtenhofer, 2020) conduct trade-offs among resolution, temporal frame rate and the number of channels for the efficient video recognition. Non-local network (Wang et al., 2018) proposes to add non-local operations in deep network and captures long-range dependencies. The recent video Transformer enables longer dependency relationship modeling and further increases the accuracy (Arnab et al., 2021).

Transformer based action recognition Recently, several works have been conducted using pure-Transformer for spatio-temporal learning (Arnab et al., 2021; Fan et al., 2021; Bertasius et al., 2021; Akbari et al., 2021). Most of the efforts focus on designing efficient Transformer models to reduce computation and memory consumption. ViViT (Arnab et al., 2021) and TimeSformer (Bertasius et al., 2021) study various factorization methods along spatial- and temporal-dimensions. MViT (Fan et al., 2021; Li et al., 2022b) conducts a trade-off between resolution and the number of channels, and constructs a multiscale Transformer to learn a hierarchy from simple dense resolution and fine-grained features to complex coarse features. Multiview Transformer (Yan et al., 2022) further employs

multiple branches to efficiently learn from various granularities of video views. Uniformer (Li et al., 2022a) and DualFormer (Liang et al., 2022) instead modify the internal structures in video Transformer to achieve efficient local-global representation learning by leveraging 3D ConvNets and local-global stratified strategy, respectively. VATT (Akbari et al., 2021) conducts unsupervised multi-modality self-supervised pretraining with a pure-Transformer structure for video classification. MBT (Nagrani et al., 2021) further constructs multimodal bottleneck tokens to learn multimodal features from an image Transformer and an audio Transformer.

To fully exploit the recent powerful spatio-temporal video Transformer, we firstly construct an Audio-Video Transformer, AVT, employing an audio Transformer and the latest video Transformer instead of image Transformers in previous multimodal action recognition works. We further design a novel loss function to strengthen the representation learning of AVT using multimodal audio-video contrastive loss, audio-video matching loss, and masked audio and video objectives.

3 AVT: AUDIO-VIDEO TRANSFORMER

The framework of Audio-Video Transformer, AVT, is illustrated in Fig. 2, which includes modality encoders, audio-video contrastive loss $\mathcal{L}_{AVC}$ for modality encoders, cross-modality bottleneck fusion, audio-video matching loss $\mathcal{L}_{AVM}$ after multimodal fusion, masked audio-video loss $\mathcal{L}_{MASegmV}$ as illustrated in Fig. 3 (b), and multimodal classification loss $\mathcal{L}^{AV}_{CLS}$. The video Transformer extracts discriminative spatio-temporal representation, which is important for action recognition. Audio-video contrastive loss reduces the distribution discrepancy between audio and video, which benefits the modality fusion and cross-modality learning. Audio-video matching loss and structured masked audio loss considering voice activity structure enable the model to learn high-level semantic representation.

3.1 MODALITY ENCODERS

Leveraging the recent huge success of audio and video Transformers, we adopt AST (Gong et al., 2021) for audio modality encoder, and Uniformer (Li et al., 2022a), MViT (Fan et al., 2021) or Video Swin Transformer (Liu et al., 2022) for video modality encoder. Previous image Transformer based work is always constrained by the limited number of input frames, which cannot fully capture the fundamental temporal representation in action recognition, and we find that a clear accuracy gap exists between powerful image Transformer, CLIP ViT (Radford et al., 2021), and video Transformers, *i.e.*, MViT (Fan et al., 2021), Video Swin Transformer (Liu et al., 2022). More details can be found in appendix Table 9 and 10.

Let V_i be the input frame sequence for the i-th video, and A_i be the input audio spectrogram. We denote the video encoder as E^V and audio encoder as E^A. After the modality encoder, we obtain a video embedding $\{E^V_{CLS}, E^V_1, \cdots, E^V_N\}$ and audio embedding $\{E^A_{CLS}, E^A_1, \cdots, E^A_M\}$, where N is the total number of tokens in the final layer of video embedding and M is the total number of tokens in the final layer of audio embedding. The cross-entropy loss used to train the video only model is

$$\mathcal{L}^V_{CLS} = -\frac{1}{n}\sum_{i=1}^{n}\sum_{c=1}^{C}[y_i(c)\log p^V_i(c)], \tag{1}$$

where n is the batch size in the stochastic gradient descent, C denotes the total number of categories, y_i represents the one-hot ground truth label for the current i-th sample, and $p^V_i(c)$ is the video classification probability for label index c, which is implemented by a linear layer after E^V_{CLS} with a softmax activation function.

Audio-video contrastive loss Multimodal inputs can be considered as different views for the same instance in the contrastive learning. Previous image-text Transformer (Li et al., 2021) shows that the image-text contrastive loss yields better accuracy than its counterparts. The cross-modality contrastive learning aligns inter-modalities features, which benefits the following cross-modality fusion. Thus, we design an audio-video contrastive loss $\mathcal{L}_{AVC}$ to align the video and audio representation before cross-modality fusion Transformer

$$\mathcal{L}_{AVC} = -\mathbb{E}_{(A,V)\in D}[y_{AV}\log\frac{\exp((g_A(E^A_{CLS})^T g_V(E^V_{CLS}))/\tau)}{\sum_{(A,V)\in D}\exp((g_A(E^A_{CLS})^T g_V(E^V_{CLS}))/\tau)}], \tag{2}$$

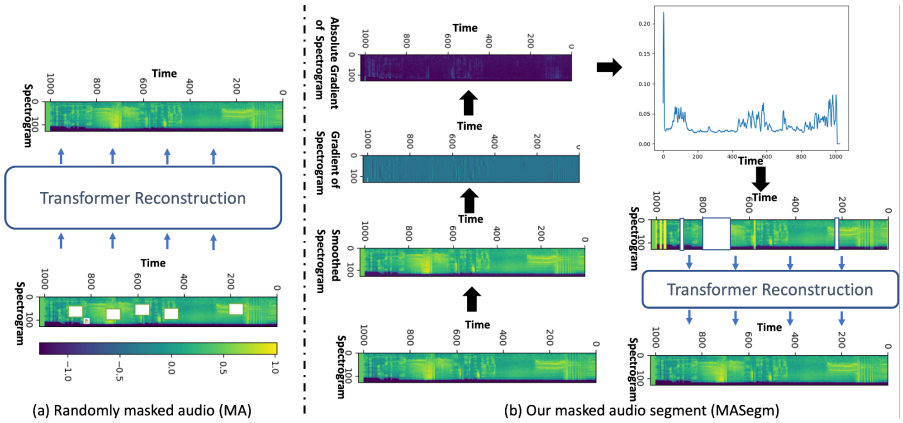

Figure 3: Masked audio models. (a) is a randomly masked audio model, and (b) is our masked audio segment model, which firstly detects the audio activities along the time dimension by smoothing, gradient calculation and averaging absolute gradients along the feature dimension, then apply masks to the whole complete activity along all the features.

where D is the multimodal input that consists of audio A and video V, y_{AV} is an indicator that the current A and V are from the same sample or not, τ is a temperature parameter, g_A and g_V are linear embedding layers for audio representation E^A_{CLS} and video representation E^V_{CLS}, respectively. The dot product $g_A(\cdot)^T g_V(\cdot)$ measures the similarity of audio and video embedding. Audio-video contrastive learning $\mathcal{L}_{AVC}$ penalizes the distribution divergence of audio and video representations for the same sample, which enhances the following cross-modality feature learning.

3.2 Cross-Modality Transformer

3.2.1 Audio-Video Cross-Modality Learning

AVBottleneck Previous cross-modality Transformers either simply concatenated multimodal representations (Akbari et al., 2021), or exchanged the key and value matrices between the two modalities (Hendricks et al., 2021). However, due to huge GPU memory consumption of existing video Transformer, we construct an audio-video bottleneck Transformer, AVBottleneck, which handles varied lengths of modality tokens efficiently as illustrated in Fig. 2 inspired by Nagrani et al. (2021). Let $\{E^F_1, \cdots, E^F_L\}$ be the initial multimodal tokens and L be the number of multimodal tokens. Without loss of generality, we omit the layer number in the denotation. One audio-video bottleneck Transformer block can be formulated as

$$
\begin{aligned}
E^{VF} &= [E^V_{CLS}, E^V_1, \cdots, E^V_N, E^F_1, \cdots, E^F_L], \quad \tilde{E}^{VF} = \text{MSA}(\text{LN}(E^{VF})) + E^{VF}, \\
\hat{E}^{VF} &= \text{MLP}(\text{LN}(\tilde{E}^{VF})) + \tilde{E}^{VF}, \quad E^{AF} = [E^A_{CLS}, E^A_1, \cdots, E^A_M, \hat{E}^F_1, \cdots, \hat{E}^F_L], \\
\tilde{E}^{AF} &= \text{MSA}(\text{LN}(E^{AF})) + E^{AF}, \quad \hat{E}^{AF} = \text{MLP}(\text{LN}(\tilde{E}^{AF})) + \tilde{E}^{AF},
\end{aligned} \tag{3}
$$

where multimodal tokens can be updated by averaging along all the AVBottleneck blocks. The multimodal bottleneck Transformer can be stacked into K blocks.

Computational complexity The multimodal bottleneck Transformer reduces the computing complexity from $O((M+N)^2)$ in merged concatenation based multimodal attention (Akbari et al., 2021) to $O((M+L)^2) + O((N+L)^2) \approx O(M^2) + O(N^2)$, which is the sum of complexity in one block of audio and video Transformers approximately, since the number of multimodal bottleneck tokens $L \ll M, N$. Here, $O(M^2)$ and $O(N^2)$ are the complexities of audio and video Transformers, where M and N are the numbers of tokens in the audio and video Transformers, respectively.

Audio-video matching loss We design an audio-video matching loss $\mathcal{L}_{AVM}$, which can be applied to audio and video embeddings after AVBottleneck and forces the multimodal Transformer to learn high level semantic labels precisely.

$$
\mathcal{L}_{AVM} = -\mathbb{E}_{(A,V)\in D}[y_{AV} \log p_{AVM}(y_{AV}) + (1 - y_{AV}) \log(1 - p_{AVM}(y_{AV}))], \tag{4}
$$

where y_{AV} is the same as the audio-video contrastive loss $\mathcal{L}_{AVC}$ in equation 2, and $p_{AVM}(y_{AV})$ is implemented by concatenating the video and audio embedding $[E^V_{CLS}, E^A_{CLS}]$ followed by a binary classification to determine the sampled audio-video pair (A, V) from the same sample or not. Through this cross-modality matching loss $\mathcal{L}_{AVM}$, we expect the AVT can effectively learn discriminative features in audio and video cross-modality Transformer.

3.2.2 MASKED AUDIO AND VIDEO LOSS

To learn high-level audio representation, *e.g.*, features for audio activity segments, we further design a masked audio-video loss, $\mathcal{L}_{MASegmV}$, in the multimodal Transformer as illustrated in Fig. 3, which is much more effective than previous random mask mechanism, $\mathcal{L}_{MAV}$, as shown in the ablation study of Table 4 and 5. The audio activity segment can be detected by second-order smoothing (Savitzky & Golay, 1964) to remove noise, calculating gradient to detect signal change along the time dimension, absolute gradient to detect changes in two temporal directions, and averaging the absolute gradient along the feature dimensions with smoothing to avoid a trivial activity segment. We then choose the top large change points as the transition points to segment different audio activities. In the training, we randomly mask a proportion of the whole complete audio activity segments.

$$\mathcal{L}_{MASegmV} = \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}_1(V_i, \hat{V}_i|\mathring{V}_i) + \mathcal{L}_1(A_i, \hat{A}_i|\mathring{A}_i), \tag{5}$$

where $\mathring{V}_i$ is a randomly masked video input, $\mathring{A}_i$ is a structured (audio complete activity segment) masked audio input, $\hat{V}_i$ and $\hat{A}_i$ are reconstructions from the masked input through the multimodal model, and the decoder can be easily constructed by rearranging the tokens into two or three-dimensional matrix followed by one layer of transposed convolution (Zeiler et al., 2010) to match the dimension. If $\mathring{A}_i$ is a randomly masked audio input, the loss becomes a conventional masked audio and video model $\mathcal{L}_{MAV}$. For the randomly masked mechanism, we uniformly choose a proportion of tokens after patch embedding and set these tokens as zero.

3.3 LEARNING FROM MULTIMODAL VIDEO

The multimodal classification can be achieved by concatenating the video and audio embedding, $[E^V_{CLS}, E^A_{CLS}]$, and a fully connected layer is constructed to yield the final action classification logits. The supervised multimodal loss can be cross-entropy loss

$$\mathcal{L}^{AV}_{CLS} = -\frac{1}{n}\sum_{i=1}^{n}\sum_{c=1}^{C}[y_i(c)\log p^{AV}_i(c)], \tag{6}$$

where $p^{AV}_i(c)$ is the multimodal classification probability for the i-th video and label index c.

A hybrid loss considering multimodal video classification and various levels of self-supervised objectives forces the multimodal Transformer to learn effectively from the training data. These self-supervised losses introduce auxiliary objectives to train the multimodal Transformer and act as regularization in the overall supervised learning loss function.

$$\mathcal{L} = \mathcal{L}^{AV}_{CLS} + \lambda_1\mathcal{L}_{AVC} + \lambda_2\mathcal{L}_{AVM} + \lambda_3\mathcal{L}_{MASegmV}, \tag{7}$$

where λ_1, λ_2, and λ_3 are hyperparameters to balance the loss terms in the training. The inference is consistent with the training, and we use the multimodal prediction p^{AV} directly.

4 EXPERIMENTAL RESULTS

4.1 DATASETS

We experiment with three public video classification datasets – Kinetics-Sounds (Arandjelovic & Zisserman, 2017; Kay et al., 2017), Epic-Kitchens-100 (Damen et al., 2021a; 2018; 2021b), and VGGSound (Chen et al., 2020). Results on two additional in-house datasets are in the appendix.

Kinetics-Sounds is a commonly used subset of Kinetics (Kay et al., 2017), which consists of 10-second videos sampled at 25fps from YouTube. As Kinetics-400 is a dynamic dataset and videos

Models	Modalities	Kinetics-Sounds Top-1	Kinetics-Sounds Top-5	VGGSound Top-1	VGGSound Top-5
Arandjelovic & Zisserman (2017)	A, V	74.0	-	-	-
AVSlowFast, R101 (Xiao et al., 2020)	A, V	85.0	-	-	-
Chen et al. (2020)	A	-	-	48.8	76.5
AudioSlowFast (Kazakos et al., 2021)	A	-	-	50.1	77.9
MBT (Nagrani et al., 2021)	A	52.6	71.5	52.3	78.1
MBT (Nagrani et al., 2021)	V	80.7	94.9	51.2	72.6
MBT (Nagrani et al., 2021)	A, V	85.0	96.8	**64.1**	**85.6**
AVT	A, V	**93.0** (8%↑)	**99.3**	63.9	85.0

Table 2: Comparison to state-of-the-art on Kinetics-Sounds and VGGSound. We report top-1 and top-5 classification accuracy. A: Audio, V: Visual.

may be removed from YouTube, we follow the dataset collection protocol in Xiao et al. (2020), and we collect 22,914 valid training multimodal videos and 1,585 valid test multimodal videos.

Epic-Kitchens-100 consists of 90,000 variable length egocentric clips spanning 100 hours capturing daily kitchen activities. The dataset formulates each action into a verb and a noun. We employ two classification heads, one for verb classification and the other one for noun classification, in the AVT. Note that the clips in the dataset are mainly short-term with average length of 2.6 seconds.

VGGSound is a large scale action recognition dataset, which consists of about 200K 10-second clips and 309 categories ranging from human actions and sound-emitting objects to human-object interactions. Like other YouTube datasets, *e.g.*, K400 (Kay et al., 2017), some clips are no longer available. After removing invalid clips, we collect 159,223 valid training multimodal videos and 12,790 valid test multimodal videos.

Implementation details We investigate two variants of video Transformers, *i.e.*, Uniformer-B 16×4 (#frames×#views) and Uniformer-B 32×4. On Epic-Kitchens-100, we only employ Uniformer-B 16×4 because of the short clip length. We use batch size of 40 and 32 in 8 40 GB NVIDIA A100 GPUs for Uniformer-B with 16 and 32 frames, respectively. The numbers of AVBottleneck blocks K and tokens L are all 4, which follows Nagrani et al. (2021). τ is fixed as 0.07, and the dimensions of g_A and g_V are set as 256 following Li et al. (2021). AdamW (Loshchilov & Hutter, 2018) is used in the backpropagation and the learning rate is 1×10^{-4}. The number of epochs is 100. The λ_1, λ_2 and λ_3 are $0.5, 0.1, 0.01$ for the first 10 epochs, $0.2, 0.05, 0.005$ from the 11- to 20-th epochs, $0.1, 0.01, 0.001$ from the 21- to 30-th epochs, and 0 for the rest epochs. These hyperparameters are generally set to tune the losses into the same scale. The number of audio segments is 50, and the masked probability is 4% for both masked audio and video models because the baseline model has already achieved a competitive accuracy. Other hyperparameters follow the recipe in Li et al. (2022a).

4.2 RESULTS

Comparison to state of the art AVT surpasses its previous multimodal state-of-the-art counterpart, MBT (Nagrani et al., 2021), by 8% and 4.2% on the Kinetics-Sounds and Epic-Kitchens-100 (16 frames) in Table 2 and 3, which demonstrates the video Transformer based multimodal Transformer is better for action recognition than its image Transformer based counterpart. On VGGSound, AVT achieves comparable accuracy with the previous state-of-the-art approach, MBT, and is 1.3× more efficient based on the number of FLOPs than MBT (Nagrani et al., 2021) from the FLOPs comparison in Table 3 because of the advantage of multiscale mechanism in the video Transformer (Li et al., 2022a). Note that MBT uses 10% more training samples than AVT.

The ablation results $w.r.t.$ audio only, video only, simple averaging audio and video only predictions (Avg), the number of frames, $\mathcal{L}_{AVC}$, $\mathcal{L}_{AVM}$, random masked model $\mathcal{L}_{MAV}$, masked audio segment and video model $\mathcal{L}_{MASegmV}$, denoted as MASegmV, are concluded in Table 4 and 5. We find that models with 32 frames generally perform better than models with 16 frames, which validates that more frames enable more powerful spatio-temporal representation learning. AVT surpasses the video only model by 10% on VGGSound, which demonstrates that audio and video provide complementary

Models	Modalities	Verb	Noun	Action	FLOPs (G)
Damen et al. (2021a)	A	42.1	21.5	14.8	-
AudioSlowFast (Kazakos et al., 2021)	A	46.5	22.8	15.4	-
TSN (Wang et al., 2016)	V, F	60.2	46.0	33.2	-
TRN (Zhou et al., 2018)	V, F	65.9	45.4	35.3	-
TBN (Kazakos et al., 2019)	A, V, F	66.0	47.2	36.7	-
TSM (Lin et al., 2019)	V, F	67.9	49.0	38.3	-
SlowFast (Feichtenhofer et al., 2019)	V	65.6	50.0	38.5	-
MBT (Nagrani et al., 2021)	A	44.3	22.4	13.0	131
MBT (Nagrani et al., 2021)	V	62.0	56.4	40.7	140
MBT (Nagrani et al., 2021)	A, V	64.8	58.0	43.4	348
ViViT-L/16×2 (Arnab et al., 2021)	V	66.4	56.8	44.0	3410
MFormer-HR (Patrick et al., 2021)	V	67.0	58.5	44.5	959
MeMViT, 16×4 (Wu et al., 2022)	V	70.6	58.5	46.2	59
AVT (16 frames)	A, V	70.4	**59.3**	**47.2** (1.0%↑)	269

Table 3: Comparison to previous related work on Epic-Kitchens-100 (16 frames). F: Optical flow.

Models	Top-1	Top-5	Models	Top-1	Top-5
Audio Only	66.1	88.2	Audio Only	54.4	76.8
Video Only (16 f)	89.5	98.9	Video Only (16 f)	52.6	75.3
Avg (16 f)	89.6	98.9	Avg (16 f)	59.0	82.1
AVBottleneck (16 f)	91.7	99.0	AVBottleneck (16 f)	59.2	81.9
+AVC (16 frames)	92.5	99.4	+AVC (16 f)	61.2	82.9
+AVC+AVM (16 f)	92.4	99.5	+AVC+AVM (16 f)	61.6	83.4
+AVC+AVM+MAV (16 f)	92.5	99.6	+AVC+AVM+MAV (16 f)	61.8	83.7
+AVC+AVM+MASegmV (16 f)	93.0	99.2	+AVC+AVM+MASegmV (16 f)	62.7	84.9
Video Only (32 f)	90.7	99.1	Video Only (32 f)	53.2	74.8
Avg (32 f)	90.9	98.4	Avg (32 f)	58.6	82.0
AVBottleneck (32 f)	91.8	99.3	AVBottleneck (32 f)	58.2	80.5
+AVC (32 f)	91.8	99.5	+AVC (32 f)	60.7	82.2
+AVC+AVM (32 f)	92.0	99.4	+AVC+AVM (32 f)	61.0	83.0
+AVC+AVM+MAV (32 f)	92.4	99.3	+AVC+AVM+MAV (32 f)	62.4	84.8
+AVC+AVM+MASegmV (32 f)	93.0	99.3	+AVC+AVM+MASegmV (32 f)	63.9	85.0

Table 4: Ablation study on Kinetics-Sounds (left) and VGGSound (right). f denotes frames.

features and combining them together improves the action recognition accuracy. The audio only model achieves slightly better accuracy than that in MBT, because we conduct a hyperparameter tuning, use pretrained audio models, and tune the sampling frequency to cover a longer audio signal in appendix. Combining the advanced self-supervised objectives further improves the action recognition accuracy, which shows that these loss functions boost the multimodal feature learning, ranging from reducing the inter-modality discrepancy to forcing the model to learn semantic representation. More comparison with image Transformers and hyperparameter ablations in AST can be found in appendix.

Visualizations We randomly pick four test clips with category names of "train whistling", "chopping food", "playing acoustic guitar", and "people shuffling" from VGGSound test set, and visualize 16 frames of raw video, GradCAM (Selvaraju et al., 2017) of video only model, AVBottleneck, and the fully trained AVT sequentially. From the first test case (the 1-4th rows), we can find the video only model focuses on general scene and incorrectly predicts this clip as "subway, metro, underground". AVBottleneck incorrectly predicts the clip as "railroad car, train wagon", and the full AVT model with audio and video aligned focuses on different parts of the train and obtains the correct prediction. From the second test case (the 5-8th rows), we find that AVBottleneck in the 7th row cannot capture the knife in the corner and incorrectly predicts the clip as "arc welding". From the third case (the 9-12th

Models	Verb	Noun	Action
Audio Only	45.6	22.2	15.3
Video Only (16 frames)	69.0	58.1	45.8
Avg (16 frames)	65.0	52.7	37.5
AVBottleneck (16 frames)	69.9	59.1	46.6
+AVC (16 frames)	70.2	58.6	46.8
+AVC+AVM (16 frames)	70.3	58.9	46.9
+AVC+AVM+MAV (16 frames)	70.4	58.7	46.7
+AVC+AVM+MASegmV (16 frames)	70.4	59.3	47.2

Table 5: Ablation study on Epic-Kitchens-100 (Damen et al., 2021a).

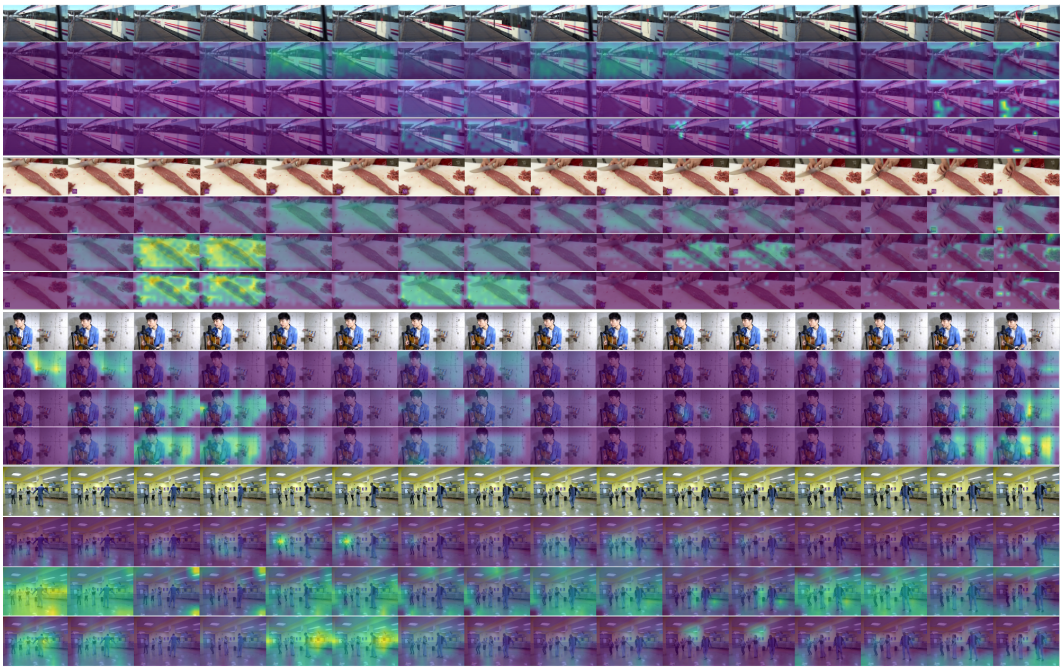

Figure 4: Visualization of four test cases in VGGSound. From top to bottom, we show 16 frames of the raw video, GradCAM (Selvaraju et al., 2017) of video only model, AVBottleneck, AVT. With well-designed strategies to learn audio and video fusion, AVT can effectively understand the clip.

rows), we find that the video only model pays attention to the background object and incorrectly predicts the clip as "metronome" and AVT can fully understand the scene. For the last test case (the 13-16th rows), without considering the audio signal, the video only model incorrectly predicts the clip as "tap dancing", which can be easily distinguished from the rhyme of the music.

5 Conclusion

In this work, we have presented an effective multimodal Transformer, AVT, which firstly leverages advanced video Transformer, audio-video contrastive loss function, audio-video matching loss and a novel masked audio model for multimodal action recognition. These self-supervised objectives penalize different aspects of multimodal Transformer, from reducing the feature divergence before multimodal fusion to forcing to learn high-level semantic representation. AVT surpasses its previous state-of-the-art counterparts by 8% and 1% on Kinetics-Sounds and Epic-Kitechens-100 without external training data. On VGGSound, AVT surpasses one of the current state-of-the-art video Transformers by 10%. Compared to MBT (Nagrani et al., 2021), AVT is 1.3× more efficient in terms of FLOPs and improves the accuracy by 4.2% on Epic-Kitchens-100.

REFERENCES

Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text. *Advances in Neural Information Processing Systems*, 34:24206–24221, 2021.

Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 609–617, 2017.

Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6836–6846, 2021.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, volume 2, pp. 4, 2021.

Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6299–6308, 2017.

Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A short note about kinetics-600. *arXiv preprint arXiv:1808.01340*, 2018.

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 721–725. IEEE, 2020.

R Christoph and Feichtenhofer Axel Pinz. Spatiotemporal residual networks for video action recognition. *Advances in Neural Information Processing Systems*, pp. 3468–3476, 2016.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The epic-kitchens dataset. In *European Conference on Computer Vision (ECCV)*, 2018.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, , Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision (IJCV)*, 2021a. URL `https://doi.org/10.1007/s11263-021-01531-2`.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. The epic-kitchens dataset: Collection, challenges and baselines. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, 43(11):4125–4141, 2021b. doi: 10.1109/TPAMI.2020.2991965.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 248–255. Ieee, 2009.

Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6824–6835, 2021.

Christoph Feichtenhofer. X3d: Expanding architectures for efficient video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 203–213, 2020.

Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 6202–6211, 2019.

Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 776–780. IEEE, 2017.

Yuan Gong, Yu-An Chung, and James Glass. Ast: Audio spectrogram transformer. In *Proceedings of Interspeech*, 2021.

Lisa Anne Hendricks, John Mellor, Rosalia Schneider, Jean-Baptiste Alayrac, and Aida Nematzadeh. Decoupling the role of data, attention, and losses in multimodal transformers. *Transactions of the Association for Computational Linguistics*, 9:570–585, 2021.

Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017.

Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Epic-fusion: Audio-visual temporal binding for egocentric action recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5492–5501, 2019.

Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Slow-fast auditory streams for audio recognition. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 855–859. IEEE, 2021.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 25:1097–1105, 2012.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in Neural Information Processing Systems*, 34, 2021.

Kunchang Li, Yali Wang, Peng Gao, Guanglu Song, Yu Liu, Hongsheng Li, and Yu Qiao. Uniformer: Unified transformer for efficient spatiotemporal representation learning. In *International Conference on Learning Representations*, 2022a.

Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Improved multiscale vision transformers for classification and detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022b.

Yuxuan Liang, Pan Zhou, Roger Zimmermann, and Shuicheng Yan. Dualformer: Local-global stratified transformer for efficient video recognition. In *European Conference on Computer Vision*, 2022.

Ji Lin, Chuang Gan, and Song Han. Tsm: Temporal shift module for efficient video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 7083–7093, 2019.

Jun Liu, Amir Shahroudy, Dong Xu, and Gang Wang. Spatio-temporal lstm with trust gates for 3d human action recognition. In *European Conference on Computer Vision*, pp. 816–833. Springer, 2016.

Jun Liu, Gang Wang, Ping Hu, Ling-Yu Duan, and Alex C Kot. Global context-aware attention lstm networks for 3d action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1647–1656, 2017.

Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3202–3211, 2022.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2018.

Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. *Advances in Neural Information Processing Systems*, 34, 2021.

Mandela Patrick, Dylan Campbell, Yuki Asano, Ishan Misra, Florian Metze, Christoph Feichtenhofer, Andrea Vedaldi, and João F Henriques. Keeping your eye on the ball: Trajectory attention in video transformers. *Advances in Neural Information Processing Systems*, 34:12493–12506, 2021.

Zhaofan Qiu, Ting Yao, and Tao Mei. Learning spatio-temporal representation with pseudo-3d residual networks. In *proceedings of the IEEE International Conference on Computer Vision*, pp. 5533–5541, 2017.

Zhaofan Qiu, Ting Yao, Chong-Wah Ngo, Xinmei Tian, and Tao Mei. Learning spatio-temporal representation with local and global diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12056–12065, 2019.

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*, pp. 8748–8763. PMLR, 2021.

Abraham Savitzky and Marcel JE Golay. Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, 36(8):1627–1639, 1964.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 618–626, 2017.

Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *Proceedings of the 27th International Conference on Neural Information Processing Systems-Volume 1*, pp. 568–576, 2014.

Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4489–4497, 2015.

Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European Conference on Computer Vision*, pp. 20–36. Springer, 2016.

Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 7794–7803, 2018.

Chao-Yuan Wu, Yanghao Li, Karttikeya Mangalam, Haoqi Fan, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. Memvit: Memory-augmented multiscale vision transformer for efficient long-term video recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13587–13597, 2022.

Fanyi Xiao, Yong Jae Lee, Kristen Grauman, Jitendra Malik, and Christoph Feichtenhofer. Audiovisual slowfast networks for video recognition. *arXiv preprint arXiv:2001.08740*, 2020.

Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 305–321, 2018.

Shen Yan, Xuehan Xiong, Anurag Arnab, Zhichao Lu, Mi Zhang, Chen Sun, and Cordelia Schmid. Multiview transformers for video recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022.

Matthew D Zeiler, Dilip Krishnan, Graham W Taylor, and Rob Fergus. Deconvolutional networks. In *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 2528–2535. IEEE, 2010.

Rowan Zellers, Jiasen Lu, Ximing Lu, Youngjae Yu, Yanpeng Zhao, Mohammadreza Salehi, Aditya Kusupati, Jack Hessel, Ali Farhadi, and Yejin Choi. Merlot reserve: Neural script knowledge through vision and language and sound. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16375–16387, 2022.

Xuefan Zha, Wentao Zhu, Lv Xun, Sen Yang, and Ji Liu. Shifted chunk transformer for spatio-temporal representational learning. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal relational reasoning in videos. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 803–818, 2018.

	Violence	Fight	Push	Hit	Gunshot	Vehi. acci.	Stab	Blast	Drink alco.
#Train /Test	320/142	3078/790	683/246	207/73	162/52	244/81	51/25	69/33	53/36
MViT	**91.2**	50.2	**51.1**	**60.9**	50.7	**77.2**	37.0	75.0	100.0
AST	75.5	**75.5**	43.5	28.0	**82.5**	61.1	**57.7**	**82.8**	100.0

Table 6: Top-1 action recognition accuracy using video-only model, MViT (Fan et al., 2021), and audio-only model, AST (Gong et al., 2021), on an in-house with nine attributes dataset demonstrate that video and audio are complimentary. Video does not always outperform audio. Dataset properties are also reported.

Attr.	Fantasy Violence	Injury	Murder	Fight	Push	Hit	Gunshot	Surgery
#Data	487/142	309/101	294/85	1133/256	352/76	242/55	335/85	56/23
Vehi. acci.	Stab	Blast	Kiss	Sex Act	Dance	Smoke	Drink alco.	Subst Use
78/26	102/33	83/39	1231/289	1183/309	229/69	8313/1092	3627/799	1193/341

Table 7: Dataset property, the number of training and test samples, of the 17 attributes dataset

A Appendix

A.1 Results on Internal Datasets

We collect two more datasets from our in-house asset, where each clip length varies from 5 seconds to 10 seconds. We collect 47,021 video-audio aligned clips to construct a 17 attributes dataset and the data statistics are shown in Table 7, where 9 attributes are chosen to validate the effectiveness of audio signal in Table 6. On 9 attributes dataset, we find that audio modality can improve the action recognition accuracy for some classes compared to one of the state-of-the-art video-only models, MViT (Fan et al., 2021). We further collect a 28 attributes dataset, which consists of 104,429 clips, and the dataset property is shown in Table 8.

There is a missing label issue in the annotation, and one clip can belong to multiple attributes. Therefore, we employ a masked multilabel loss to train the model as

$$L_{CLS}^{AV} = -\frac{1}{n}\sum_{i=1}^{n}\frac{1}{|C_i|}\sum_{c\in C_i}[c\log p^{AV}(c) + (1-c)\log(1-p^{AV}(c))], \tag{8}$$

where n is the batch size in the stochastic gradient descent, C_i is the annotated label set for the current i-th sample, and $p^{AV}(c)$ is the video classification probability for label c, which is implemented by a linear layer after E_{CLS}^{AV} with a sigmoid activation function. We report the top-1 accuracy in the experiments.

We mainly investigate two video Transformers, MViT (Fan et al., 2021) and Video Swin Transformer (Liu et al., 2022). We employ K600 (Carreira et al., 2018) pretrained weights in the training. We use batch size of 16 in 8 NVIDIA 16 GB V100 GPUs and 24 in 8 32 GB V100 GPUs for MViT and Video Swin Transformer, respectively. The number of frames used is 32 and the sampling rate is 3 for MViT, Video Swin Transformer and AVT with MViT. We use 24 frames and sampling rate of 4 for AVT with Video Swin Transformer. The numbers of AVBottleneck blocks are 4 and 2 for 17 and 28 attributes datasets, respectively.

A.1.1 Seventeen Attributes Action Dataset

The accuracy comparison on 17 attributes dataset is listed in Table 9. The ViT uses CLIP (Radford et al., 2021) pretrained model, which is a strong baseline for the image Transformer, and we fine-tune

Attr.	Fantasy Violence	Injury	Kick	Army Battle	Fall	Point	Punch
#Data	719/377	410/245	570/315	57/29	2061/506	3887/722	913/337
Push	Slap	Slit	Strangle	Throw	Murder	Dead body	Fight
1685/523	125/25	100/33	559/147	361/98	1149/390	2284/613	5416/1747
Gun Shot	Surgery	Vehi. Acci.	Stab	Blast	Riot	Suicide	Drink alco.
2769/723	1992/534	389/119	762/329	1706/415	71/33	113/17	4571/1022
Erotic Dance	Kiss	Sex act	Smoke	Subst use			
270/79	1526/354	1458/378	11072/1469	1411/404			

Table 8: Dataset property, the number of training and test samples, of the 28 attributes dataset

Model	ViT	MViT	AST	AVBottleneck	w/ AVC	w/ AVC +AVM	w/ AVC+ AVM+MAV	w/ AVC+AVM+ MASegmV (Ours)
Fantasy Violence	94.9	94.6	69.1	92.9	94.0	97.3	95.6	96.4
Injury	67.6	59.0	41.4	58.6	56.6	59.2	61.6	60.3
Murder	77.2	78.9	77.1	83.9	83.6	83.3	83.8	86.9
Fight	60.7	72.5	62.9	74.9	79.3	79.4	80.0	83.2
Push	69.2	81.9	65.3	82.4	85.2	92.5	91.5	88.9
Hit	59.0	56.7	47.7	61.9	74.3	65.1	65.9	76.2
Gunshot	84.8	76.9	91.3	87.3	97.2	97.1	100.0	97.2
Surgery	100.0	100.0	77.8	100.0	100.0	100.0	100.0	100.0
Vehicle accident	90.0	86.9	88.9	100.0	100.0	100.0	100.0	100.0
Stab	72.7	72.0	82.1	82.8	92.9	92.9	92.9	92.9
Blast	87.1	88.9	76.7	92.3	90.9	93.9	93.9	100.0
Kiss	98.4	97.3	65.6	96.9	96.8	97.3	97.6	97.7
Sex Act	95.1	96.1	81.5	96.1	96.1	95.0	95.4	95.9
Dance	91.7	95.2	81.7	100.0	96.7	100.0	98.4	100.0
Smoke	78.5	83.9	36.5	85.3	79.2	84.4	84.8	86.0
Drink alco.	90.8	92.0	51.5	90.1	88.9	92.1	92.7	94.0
Subst Use	85.2	78.8	50.6	80.4	77.2	83.7	84.6	85.1
Average	82.5	83.0	67.5	86.2	87.6	89	89.3	**90.6** (7.6↑)

Table 9: Accuracy (%) comparison on 17 attributes dataset. ViT is pretrained by CLIP (Radford et al., 2021). MViT (Fan et al., 2021) is pretrained in K600 (Carreira et al., 2018). AST (Gong et al., 2021) is pretrained in ImageNet (Deng et al., 2009).

Model	ViT	Video Swin	AST	Video Swin +AST (Avg)	AVBottleneck	w/ AVC	w/ AVC +MAV	w/ AVC+ MASegmV (Ours)
Avg Acc.	71.3	86.9	61.1	65.2	87.7	88.4	88.6	**89.3** (2.4↑)

Table 10: Accuracy (%) comparison on 28 attributes dataset. Video Swin+AST (Avg) is the average prediction from the two models.

Model	AST IN384	AST AS384	AST IN224	AST IN-S	AST IN-T	Video Swin 32×3	Video Swin 24×4	Video Swin 16×6
Avg Acc.	58.7	60.1	**61.1**	57.9	57.1	86.9	**86.9**	86.4

Table 11: Effect of the number of frames and different audio models based on top-1 accuracy (%) on 28 attributes dataset.

the last classification layer because of the GPU memory constraint. The number of frames in ViT is fixed as 16 and the batch size is fixed the same as MViT or Video Swin Transformer, and we use an average pooling along the predicted probability for each frame to generate the classification score for one clip. Our complete AVT outperforms MViT by 7.6% on the dataset. For the masked probability of 14%, the AVT achieves 89.7% accuracy, thus we decide to set the masked probability as low as 2% on the in-house datasets, which can be considered as a regularization for the multimodal Transformer classification model. Table 9 also demonstrates the effectiveness of audio-video contrastive loss and audio-video matching loss in AVT.

A.1.2 Twenty-Eight Action Dataset

We also conduct experiments on our recent large scale 28 attributes dataset. Table 10 shows our AVT surpasses Video Swin Transformer by 2.4%. The improvement is less than that in the 17 attributes dataset, because 1) Video Swin Transformer is the current one of the most competitive state-of-the-art approaches, and the base model we used only outperforms Video Swin-small by 0.4% on the 28 attributes dataset, 2) from the full accuracy comparison of each attribute, there is no attribute where audio model, AST, performs better than video model, Video Swin Transformer, because the data annotation uses video signal only, and 3) the current annotation quality of 28 attributes dataset is improving.

A.1.3 Effect of the number of frames in Video Swin Transformer

Because of huge GPU memory consumption, we investigate the effect of the number of frames in Video Swin Transformer on 28 attributes dataset in Table 11. We find Video Swin Transformer with 16 frames and sampling rate of 6 performs worse than the original Video Swin Transformer with 32 frames on our internal dataset. Video Swin Transformer with 24 frames is comparable, and we use 24 frames in our AVT to reduce the GPU memory. Note that this hyperparameter is only valid for the in-house datasets.

A.1.4 Effect of AST

On 28 attributes dataset, we find the video modality is much more important than audio modality from Table 10. We tradeoff the Video Swin Transformer to reduce the GPU memory firstly and find the accuracy drops significantly. We investigate different sized audio spectrogram Transformer (Gong et al., 2021) in Table 11. We find the AST with ViT for ImageNet (Deng et al., 2009) input of 224×224 pretrained model performs the best instead of original AST pretrained on ImageNet 384×384 and AudioSet (Gemmeke et al., 2017).

A.1.5 Visualizations

We upload the visualization and real demo on the test set of 17 attributes dataset in the supplementary, where we randomly choose one or two test clips for each attribute and display the predicted probabilities of our AVT, MViT and AST. From the visualization, it clearly reveals that, the audio provides a complementary and discriminative feature for attributes, *e.g.*, dance, shooting, bomb blast, murder, vehicle accident, drinking alcohol, hit, fight, stab, *etc.*, and AVT alleviates drawbacks of visual occlusion and non-transcribed text from audio.