Everything at Once – Multi-modal Fusion Transformer for Video Retrieval

Nina Shvetsova [1] Brian Chen [2] Andrew Rouditchenko[3] Samuel Thomas[4,5]
Brian Kingsbury[4,5] Rogerio Feris[4,5] David Harwath[6] James Glass[3] Hilde Kuehne[1,5]
[1]Goethe University Frankfurt, [2]Columbia University, [3]MIT CSAIL [4]IBM Research AI, [5]MIT-IBM Watson AI Lab, [6] UT Austin
shvetsov@uni-frankfurt.de

Abstract

Multi-modal learning from video data has seen increased attention recently as it allows training of semantically meaningful embeddings without human annotation, enabling tasks like zero-shot retrieval and action localization. In this work, we present a multi-modal, modality agnostic fusion transformer that learns to exchange information between multiple modalities, such as video, audio, and text, and integrate them into a fused representation in a joined multi-modal embedding space. We propose to train the system with a combinatorial loss on everything at once – any combination of input modalities, such as single modalities as well as pairs of modalities, explicitly leaving out any add-ons such as position or modality encoding. At test time, the resulting model can process and fuse any number of input modalities. Moreover, the implicit properties of the transformer allow to process inputs of different lengths. To evaluate the proposed approach, we train the model on the large scale HowTo100M dataset and evaluate the resulting embedding space on four challenging benchmark datasets obtaining state-of-the-art results in zero-shot video retrieval and zero-shot video action localization. Our code for this work is also available.[1]

1. Introduction

Humans capture their world in various ways, combining different sensory input modalities such as vision, sound, touch, and more, to make sense of their environment. Video data approximates this type of input by combining visual and audio information as two coherent and complementary signals that can be further enhanced with a text description. Recent research has therefore started to explore how the information of those different modalities can be leveraged to learn meaningful representations from this kind of content. Such systems can be used for representation learning, for example, to learn multi-modal embedding spaces on video data [1, 2], where the input of one modality such as text, can be matched to one or more other modalities such as video and audio, enabling tasks such as nearest-neighbor based zero-shot classification or video retrieval [20, 37, 44]. Our work in this paper focuses on the later problem, namely the learning of meaningful multi-modal embedding spaces. Current approaches in this area usually learn encodings for different modalities by projecting inputs to a common space and applying contrastive loss to bring embeddings from co-occurred modalities together. Such approaches can be based on classical neural network elements to learn those encodings [4, 12, 35, 37, 44], i.e. convolutional neural networks backbones and non-linear projections [37], multiple instance learning [35], or clustering [12]. More recently transformer based methods have also been proposed [1, 10, 20, 32]. To generate the final embedding space, they use multiple independent single-modality self-attention transformer blocks [10,21,32], or a single transformer model for all modalities [20], or a single modality-agnostic transformer [1]. In the last approach, modalities are still processed independently and one-by-one forwarded to achieve a single-modality embedding. But so far, none of these transformers allow for adaption to any given number of input

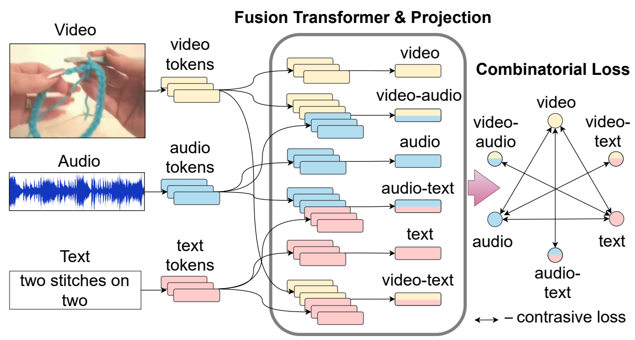

Figure 1. Overview of the proposed approach for self-supervised learning of multi-modal embedding space. The fusion transformer is able to process any combination of input modalities. Internally, the transformer allows each modality to attend to each other. The proposed architecture is trained with a combinatorial contrastive loss considering each possible combination of input modalities.

[1]https://github.com/ninatu/everything_at_once

modalities. Although modality-agnostic transformers that handle multiple input modalities such as PerceiverIO [26] have been proposed, they have been constructed for learning a latent space that can cover multiple tasks in different domains. Compared to our work, the latent space in such cases mainly serves the purpose of compressing multiple inputs and tasks in one model.

In this work, we propose an approach that leverages self-attention for multi-modal learning which jointly processes any number of modalities and allows modalities to attend to each other. A high level overview our architecture is shown in Figure 1. Input tokens from one or more modalities are passed through a fusion transformer that attends features relevant for a combined input, followed by a projection to a joint multi-modal embedding space. We design and train the fusion transformer to cover three aspects of multi-modal video learning: first, it should allow modalities to attend to each other and learn multi-modal correlations; second, it should be modality-agnostic and handle any possible modality input combination; and third, as different modalities and samples can vary in length, it should be able to process input of any length. To enable the fusion transformer to address all those tasks, we follow the idea of a universal self-attention in the transformer block and share key, query, and value weights to all tokens, agnostic of their input modality. In this way, self-attentions learns which input tokens to attend from single modalities as well as from any combination of modalities in a general way.

To train the model, we propose a combinatorial loss function which considers contrastive loss between all possible and available input combinations. For example, in the case of vision, text, and audio, the loss is based on each modality embedding alone as well as based on pairwise vision-text, audio-text, and text-audio combinations as shown in Figure 1. The resulting model is thus able to fuse any number of input modalities at test time. Compared to other universal self-attention methods, we omit any meta information encoding such as position or modality embedding. This further allows us to process any input of different lengths, as we are no longer bound to a maximum input size defined at training time. Note that while we refer to this transformer as a fusion transformer, we are not proposing a new transformer architecture, but rather refer to it as a transformer that is trained in a way that enables fusion without any need for changes to the self-attention mechanism. As a result, the final modal can be used for any type of input, single modalities or combinations of multiple ones, as well as for any input length.

We evaluate the proposed approach by training the model on the HowTo100M dataset [37] and testing its zero-shot text-to-video retrieval and step action localization on four downstream datasets, namely YouCook2 [55], MSR-VTT [52], CrossTask [58] and Mining YouTube [29]. Our results show that the proposed combination of a fusion transformer together with a combinatorial loss function improves performance and leads to new state-of-the-art results. We summarize the contributions of the paper as follows:

- We propose a multi-modal fusion transformer that processes input of any combination of modalities and any length and attends relevant features with respect to cross-modal information.
- We propose a combinatorial contrastive loss that considers all possible combinations of input modalities at training time.
- We show that using such a multi-modal fusion transformer as an intermediate processing step can significantly improve performance for multi-modal embedding space learning.

2. Related Work

Multi-modal learning. The idea of learning from more than one modality can be seen as an integral part of machine learning research, comprising areas such as vision-language learning [42, 54], vision-audio learning [5–7, 13, 23, 47, 50], zero-shot learning [25, 34], cross-modal generation [33, 43, 56], as well as multi-modal multi-task learning [27]. Video naturally combines multiple modalities, while at the same time allowing to learn from large-scale data that would not be annotatable in a reasonable time. In this context, Miech et al. [37] proposed the HowTo100M dataset of narrated videos and presented a system showing the potential of multi-modal learning for learning a video-text embedding space via contrastive loss. The dataset contains YouTube instructional videos that come with audio and respective subtitles as a textual description obtained by Automatic Speech Recognition (ASR). As this data can be considered more noisy than curated vision-text datasets, Amrani et al. [4] proposed a noise estimation for multi-modal data via multi-modal density estimation. Miech et al. [35] proposed MIL-NCE, combining the idea of noise-contrastive estimation with a multiple instance learning formulation. Alwassel et al. [3] used the audio and video information only and proposed to leverage unsupervised clustering as a supervisory signal across modalities. While those works [3, 4, 35, 37] only use two modalities to train their models, others have focused on the problem of learning from vision, audio, and text at once [2, 8, 12, 19, 44]. As perhaps one of the first, Aytar et al. [8] proposed an architecture trained on image-text and image-audio pairs that allows to connect text and audio modalities. Later Alayrac et al. [2] followed the idea of different embedding spaces for different modality combinations and proposed Multi-Modal Versatile Networks. A shared embedding space was proposed by Rouditchenko et al. [44] mapping all three modalities in one joint space. This idea has recently been extended by additional clustering and reconstruction loss by Chen et al. [12].

Multi-modal learning with transformers. Architectures based on self-attention and transformers have been explored to learn from multi-modal video data. Cheng et al. [15] proposed a co-attention module to learn correspondences between audio and video samples. Luo et al. [31] pick up on that idea but proposed, similar to Uniter [14] for vision-language tasks, a joint cross-modal encoding for video-text pairs. Compared to that, Bain et al. [10] focused on the problem of how to attend to temporal as well as spatial information in the video backbone. They therefore processed both modalities, video and text, in two separate transformer backbones and only added a linear mapping layer on top of the backbones. In this context, recently, Nagrani et al. [39] proposed a multi-modal bottleneck transformer for effective audio-visual fusion trained in the supervised setting. A transformer-based approach that actually uses all three modalities, and can therefore be considered closest to our proposed work, has been proposed by Akbari et al. [1]. Here, a single backbone transformer is applied to any of the modalities separately, but with shared attention. For training, the model follows the idea of [2] and computes the matching of video-audio first, followed by video-text matching. It thus fuses those modalities in a pairwise way, which can be compared to a subset of our proposed loss function. Other approaches also leverage temporal aspects in context of multi-modal transformer learning. Gabeur et al. [20] used a combination of expert and temporal embeddings to train a multi-modal transformer while Wang et al. [49] proposed a local-global temporal alignment based on multi-modal experts to guide the training. The idea of simply using a pretrained vision-language transformer model has also been explored by Lou et al. [32], using the pretrained CLIP model [42] as a backbone with a transformer-based similarity encoder on top of a vision and text backbone and achieving good results on tasks such as video retrieval. As most transformer-based method use various and sometimes not-publicly available datasets for backbone pretraining or have a need for resources that make it hard to repeat experiments, it is difficult to directly compare performance across different architectures and pretraining set. We therefore decided to follow the setup used in majority of works here and rely on pre-extracted features that are then processed by the proposed architecture to allow a direct comparison with previous works.

3. Method

Our goal is to learn a projection function of single modalities or a set of modalities into a joint embedding space in a way that semantically similar inputs would be close to each other, e.g. the projection of the text description of a video scene should be close to the projection of the video-audio representation of this scene. In the following, we consider three modalities, video, audio, and text (corresponding ASR caption or linguistic narration); but the proposed method can be extended to more modalities.

3.1. Problem Statement

Given a set of text-video-audio triplets $\{(t_i, v_i, a_i)\}_{i=1}^N \in (\mathcal{T} \times \mathcal{V} \times \mathcal{A})^N$ of N video clips obtained from the data distribution, we are learning a projection $f(\cdot,\cdot,\cdot)$ that can take up to three inputs: text t, video v, and audio a and produce $d-$dimensional embedding representation of the input. For the simplicity of the notation, we will omit missing modalities, so that $f(t, v)$ will stand for projection $\mathcal{T} \times \mathcal{V} \to \mathbb{R}^d$ and represent the joint embedding of text t and video v. Our goal is to maximize the dot-product similarities between semantically related inputs $f(t)$, $f(v)$, $f(a)$, $f(t, a)$, $f(t, v)$, $f(v, a)$ (such as when t, v, and a are from the same video clip) and minimize otherwise.

3.2. Model Architecture

3.2.1 Token Creation

As illustrated in Figure 2, our architecture starts from features extracted from modality-specific backbones. We transform sets of extracted feature vectors into token space by learnable modality-specific projections and modality-specific normalization layers [9]. As a result, for the (t_i, v_i, a_i) input triplet, we obtain three sets of tokens: $[\tau_{i_1}, ..., \tau_{i_k}]$ from text t_i, $[\nu_{i_1}, ..., \nu_{i_m}]$ from video v_i, $[\alpha_{i_1}, ..., \alpha_{i_n}]$ from audio a_i. As the number of tokens may vary, e.g. depending on the length of the video clip, we normalize the length of inputs per batch to allow batch processing by padding and using attention masks [48]. Practically, for comparability, we follow the protocol of [37, 44] and train the model on fixed-length video clips. Technically, the model can handle clips of variable length, also with respect to different modalities, at training and at test time if needed.

3.2.2 No Positional Embeddings

Unlike other transformer-based methods [1, 10, 14, 30, 46], we omit adding any positional or type embedding information to the tokens. The reason for this is three-fold. With respect to type embedding, it can be assumed that tokens already encode this information based on the fact that they are generated by different backbones, and hence each come with their own "fingerprints." Positional information has been shown to be beneficial in the context of consistently structured data such as sentences. But in the case of multi-modal video learning, clips are sampled randomly from a larger video sequence at training time, usually without considering shot boundaries or speech pauses. We therefore do not expect a consistent temporal pattern in the sense that a clip always starts at the beginning of an action. Thus leaving

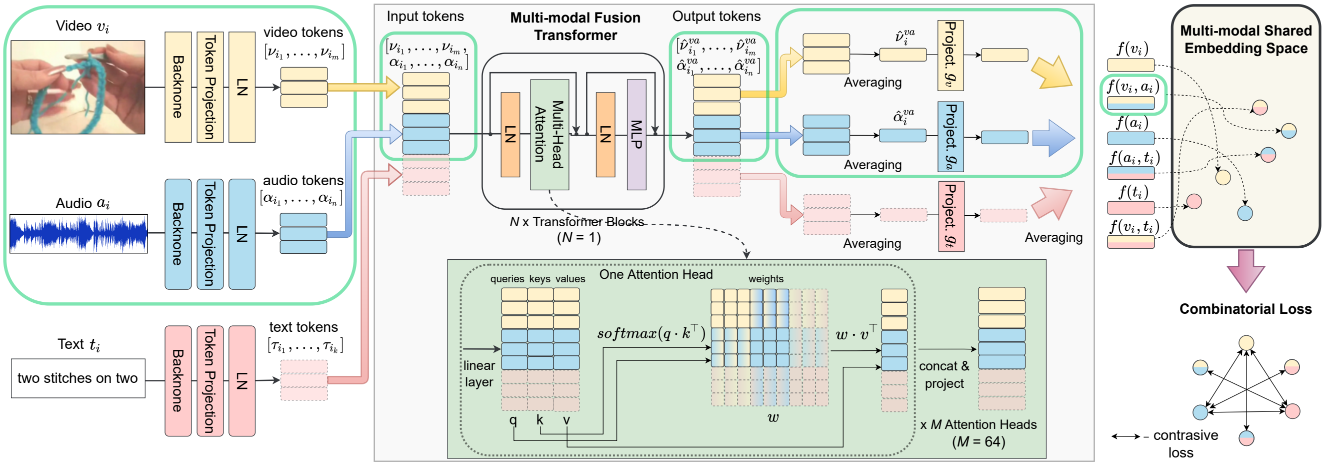

Figure 2. Schematic visualization the proposed method. While tokens from different modalities are processed in all possible combinations, we exemplary consider the video-audio pair marked with green rectangles here. The input tokens are forwarded together through the fusion transformer layer and attended by the respective weights, which are based on the combinations of keys and queries of input tokens in various modalities. The resulting outputs of multiple heads are then concatenated and projected to the final token space, which is then used to project each modality separately into the joint embedding space. During training, we apply the model six times to obtain six embeddings corresponding to text, video, audio, text-video, text-audio, and video-audio modalities to compute the combinatorial loss.

out positional embedding might prevent adding noise during training. At inference time, avoiding the positional embedding allows us to process sequences longer than used in the training.

3.2.3 Multi-modal Fusion Transformer

As our goal is to learn the representations of any number and combination of input modalities, we want the projection f to learn how to fuse information from multiple modalities to enhance the joint embedding representation. For this purpose, we propose a multi-modal, modality agnostic transformer, where the keys, queries, and values of the input tokens and all further transformations are computed independently from the modality. To create our multi-modal fusion transformer, we adopt regular transformer blocks [48]. Each transformer block consist of a multiheaded self-attention and a multilayer perceptron (MLP) with two LayerNorm (LN) transforms before them along with two residual connections, as illustrated in Figure 2. Note that the difference compared to other methods is not in the architecture itself, but in the way it is trained and the fact that the resulting fusion can actually be learned by a vanilla transformer block, if it is specifically trained for this task. Fusion transformer thus refers to the way a transformer block can be used rather than to a new architecture.

We train the system with a combinatorial input. Namely, we apply it to joint sets of input tokens from all possible combinations of modalities: singles - t, v, a and pairs - (t, v), (v, a), (t, a), allowing tokens from one modality to attend to tokens of other modalities. In this way, we can obtain a *fused* representation from multiple modalities: the combination (t, v) will result in a fused representation of text and video modalities denoted as tv, resp. for va - video and audio, and ta - text and audio. Note that e.g. in case of four modalities, we would consider all combinations up to a triplet (t, v, a) during training. As more modalities would be added, the number of combinations would grow to the point where it might be infeasible to consider all configurations. In this case random modality dropout could be used during training as done in AVSlowfast [50] or Perceiver [26].

Since we want the fusion transformer to be modality agnostic, in each training iteration, we apply it six times to obtain six representations for each sample i: t_i, v_i, a_i, $t_i v_i$, $v_i a_i$, $t_i a_i$. To obtain each representation, we create a joint list of tokens, e.g. for $v_i a_i$: $[\nu_{i_1}, ..., \nu_{i_m}, \alpha_{i_1}, ..., \alpha_{i_n}]$. We apply the transformer to this input and obtain output tokens as e.g. $[\hat{\nu}^{va}_{i_1}, ..., \hat{\nu}^{va}_{i_m} \hat{\alpha}^{va}_{i_1}, ..., \hat{\alpha}^{va}_{i_n}]$ for $v_i a_i$ (with superscript va denoting that tokens were attended to both v and a modalities), where each token was attended with information from another tokens. Note that, unlike the ViT model [18], we do not prepend a learnable $[cls]$ token, which usually serves as a joint representation of all tokens. In our ablation studies we show that this is beneficial for the model (Sec. 4.4).

3.2.4 Projection to Shared Embedding Space

With the resulting output tokens, we create the final embeddings for each modality. For each training sample, we get six output sets of tokens and thus embeddings. As an example, we consider the case of creating the representation for $v_i a_i$. We divide output tokens $[\hat{\nu}^{va}_{i_1}, ..., \hat{\nu}^{va}_{i_m} \hat{\alpha}^{va}_{i_1}, ..., \hat{\alpha}^{va}_{i_n}]$ into groups based on modality: $[\hat{\nu}^{va}_{i_1}, ..., \hat{\nu}^{va}_{i_m}]$ and $[\hat{\alpha}^{va}_{i_1}, ..., \hat{\alpha}^{va}_{i_n}]$ and then average them: $\hat{\nu}^{va}_i = \sum_{j=1}^{m} \hat{\nu}^{va}_{ij}$, $\hat{\alpha}^{va}_i =$

$\sum_{j=1}^{n} \hat{\alpha}_{i_j}^{va}$. As a result, we obtain a vector representation for each modality included in this computation. But since modalities, even when enhanced with other modalities, are still very different, we project them into the shared embedding space by the learnable modality-specific projections g_t, g_v, or g_a for projections for t, v, a respectively, normalize them, and then combine into *a final embedding vector*:

$$f(v_i, a_i) = \text{norm}(\text{norm}(g_v(\hat{\nu}_i^{va})) + \text{norm}(g_a(\hat{\alpha}_i^{va}))). \quad (1)$$

The normalization ("norm") is used to align the magnitude of vectors. When computing dot product similarity, we take into account only the angle between vectors.

3.3. Combinatorial Loss

Contrastive loss can be used to learn representations such that semantically similar inputs are mapped close to each other. Unlike other methods [1, 2, 12, 44] that learn how to bring modalities together by training with three pairwise single-modality contrastive losses, L_{t_v} between (t, v), L_{t_a} between (t, a), and L_{v_a} between (v, a), we force tokens to exchange information between modalities while enabling additional contrastive losses: L_{t_va} between (t, va), L_{v_ta} between (v, ta), and L_{a_tv} between (a, tv), and introduce our *combinatorial loss*:

$$\begin{aligned} L = &\lambda_{t\_v} L_{t\_v} + \lambda_{v\_a} L_{v\_a} + \lambda_{t\_a} L_{t\_a} + \\ &+ \lambda_{t\_va} L_{t\_va} + \lambda_{v\_ta} L_{v\_ta} + \lambda_{a\_tv} L_{a\_tv}, \end{aligned} \quad (2)$$

where $\lambda_{m\_\hat{m}}$ denotes a weighting coefficient of $(m, \hat{m})$.

Our combinatorial loss considers all possible and available modality combinations and can be generalized to any set of modalities $\mathcal{M} = \{m_1, ..., m_N\}$ as follows:

$$L = \sum_{\mathcal{X}, \mathcal{Y} \subset \mathcal{M}; \mathcal{X} \cap \mathcal{Y} = \varnothing} \lambda_{\mathcal{X}\mathcal{Y}} L_{\mathcal{X}\mathcal{Y}}. \quad (3)$$

where $L_{\mathcal{X}\mathcal{Y}}$ is a contrastive loss between the fused representations of subsets $\mathcal{X}$ and $\mathcal{Y}$, $\lambda_{\mathcal{X}\mathcal{Y}}$ is a weighting coefficient.

To compute the contrastive losses for all combinations, we use Noise Contrastive Estimation [40] with temperature τ and batch size B:

$$\text{NCE}(x, y) = -\log \left(\frac{\exp(x^\top y / \tau)}{\sum_{i=1}^{B} \exp({x_i}^\top y_i / \tau)} \right). \quad (4)$$

By combining both aspects, the processing of all possible modality combinations and the training of the system with the proposed combinatorial loss, we obtain a multi-modal fusion transformer that learns how to attend tokens from one modality to the tokens from all other modalities.

4. Experimental Evaluation

4.1. Experimental Setup

If not stated otherwise, we use the following experimental setup in all our experiments and ablation studies.

Backbones. To ensure comparability, we follow the setup of previous works [4, 12, 37, 44] which is as follows: as visual backbone, we use a combination of ResNet-152 [24], pretrained on Imagenet [16] and compute one 2D-feature (2048-dimensional vector) per second, as well as ResNeXt-101 [22] pretrained on Kinetics [11] to get 1.5 3D-feature (2048 dim.) per second. We temporally upsample 2D-features with nearest neighbors to have the same number of features as 3D-features and concatenate them to obtain 4096-dimensional vectors. As a text backbone, GoogleNews pretrained Word2vec model [38] is used with 300-dimensional embedding per word. These backbones are fixed and not finetuned during training. Following [12, 44], we use a trainable CNN with residual layers as an audio backbone and adapt the last two residual blocks to extract 1.5 4096-dimensional features per second (see Supplementary Material).

Data Sampling. We use a batch of 224 videos and randomly sample ten 8-second clips per video. If the sampled clip contains narration (95% of all clips), we use ASR time stamps to select clip borders. To disentangle the very high text-audio correlation in HowTo100M, and to avoid text being learned just as an audio narration, we shift the audio clip randomly by 4 seconds with respect to the video and text boundaries.

Projections. Following [12, 37, 44], we use a gated linear projection [36] to project features into common token space, as well as to project resulting tokens into shared embedding space. We set the dimension of the common token space to 4096 and of the shared embedding space to 6144.

Transformer architecture. As a multi-modal fusion transformer, we use one transformer block with a hidden size of 4096, 64 heads, and an MLP size of 4096.

Loss computation. We use a temperature of 0.05 in NCE and normalize vectors prior to computing dot products. Since not every video clip has all three modalities, we computed NCE only over non-empty embeddings. Following [2], we set a larger weight for a text-visual loss in Equation 2, since it was beneficial for training on HowTo100M: $\lambda_{t\_v} = 1$, $\lambda_{v\_a} = \lambda_{t\_a} = \lambda_{t\_va} = \lambda_{v\_ta} = \lambda_{a\_tv} = 0.1$.

Optimization. We train all models for 15 epochs using an Adam optimizer [28] with a learning rate of 5e-5 and exponential decay of 0.9.

4.2. Datasets, Tasks, and Metrics

Pretraining Dataset. We train our model on the HowTo100M dataset [37], which contains over 1 million instructional videos with automatically generated text narrations. The text narrations can be assumed to be noisy and to not always describe the video scene [37].

Zero-shot Text-to-video Retrieval. We use MSR-VTT [52] and YouCook2 [55] datasets to evaluate the zero-shot text-to-video retrieval capability of our model. The YouCook2 dataset contains instructional cooking videos from YouTube with human-annotated clips ($\sim 2 - 200$ seconds). For eval-

Method	Train. Mod.	Retrieval	Train. Dataset	Visual BB	Trainable BB t	v	a	YouCook2 R@1↑	R@5↑	R@10↑	MedR↓	MSR-VTT R@1↑	R@5↑	R@10↑	MedR↓
ActBERT [57]	tv	$t \to v$	HT100M	Res3D+Faster R-CNN				9.6	26.7	38.0	19	8.6	23.4	33.1	36
Support Set [41]	tv	$t \to v$	HT100M	R152 + R(2+1)D-34	✓			-	-	-	-	8.7	23.0	31.1	31
HT100M [37]	tv	$t \to v$	HT100M	R152 + RX101				6.1	17.3	24.8	46	7.5	21.2	29.6	38
NoiseEstim. [4]	tv	$t \to v$	HT100M	R152 + RX101				-	-	-	-	8.4	22.0	30.4	36
Ours	tv	$t \to v$	HT100M	R152 + RX101				11.2	28.5	39.7	19	9.6	**26.1**	**36.1**	**23**
Ours	tva	$t \to v$	HT100M	R152 + RX101			✓	10.7	27.9	38.9	19	10.3	24.6	35.3	25
MMT [20]	tva	$t \to va$	HT100M	7 experts	✓			-	-	-	-	-	14.4	-	66
AVLNet [44]	tva	$t \to v+a$	HT100M	R152+RX101			✓	19.9	36.1	44.3	16	8.3	19.2	27.4	47
MCN [12]	tva	$t \to v+a$	HT100M	R152+RX101			✓	18.1	35.5	45.2	-	**10.5**	25.2	33.8	-
Ours	tva	$t \to va$	HT100M	R152+RX101			✓	**20.0**	**40.7**	**51.3**	**10**	8.9	23.8	31.8	30
Models with a stronger visual backbone:															
MMV [2]	tva	$t \to v$	HT100M+AudioSet	TSM-50x2		✓	✓	11.7	33.4	45.4	13	9.3	23.0	31.1	38
VATT [12]	tva	$t \to v$	AudioSet	Transformer	✓	✓	✓	-	-	45.5	13	-	-	29.7	49
MIL-NCE [35]	tv	$t \to v$	HT100M	S3D		✓		15.1	38.0	51.2	10	**9.9**	**24.0**	32.4	29.5
Ours	tva	$t \to v$	HT100M	S3D†			✓	19.8	42.9	55.1	8	**9.9**	**24.0**	**32.6**	**28**
Ours	tva	$t \to va$	HT100M	S3D†			✓	**24.6**	**48.3**	**60.4**	**6**	9.3	22.9	31.2	35
FrozenInTime [10]	tv	$t \to v$	CC+WV+COCO	Transformer	✓	✓		-	-	-	-	24.7	46.9	57.2	7
CLIP4Clip [4]	tv	$t \to v$	WiT + HT100M	CLIP	✓	✓		-	-	-	-	32.0	57.0	66.9	4

Table 1. Zero-shot text-to-video retrieval results on YouCook2/MSR-VTT. In "Retrieval" column: $v + a$ stands for averaging video and audio embeddings for a video representation, va - our joint video-audio embedding where modalities attend to each other during embedding computation, t and v are single-modality embeddings. S3D† is the S3D pretrained by MIL-NCE [35]. We include CLIP4CLIP and FrozenInTime for completeness, but do directly compare because of different pre-training setups. Train Mod.=Training Modalities, BB=Backbone, CC=Conceptual Captions [45], WV=WedVid-2M [10].

Method	Train. Mod.	Retrieval	Pre-train. Dataset	Visual BB	Trainable BB t	v	a	YouCook2 R@1↑	R@5↑	R@10↑	MedR↓	MSR-VTT R@1↑	R@5↑	R@10↑	MedR↓
ActBERT [57]	tv	$t \to v$	HT100M	Res3D+Faster R-CNN				-	-	-	-	16.3	42.8	56.9	10
HT100M [37]	tv	$t \to v$	HT100M	R152 + RX101				8.2	24.5	35.3	24	14.9	40.2	52.8	9
NoiseEstim. [4]	tv	$t \to v$	HT100M	R152 + RX101				-	-	-	-	17.4	41.6	53.6	8
Ours	tv	$t \to v$	HT100M	R152 + RX101				13.7	35.3	48.4	12	21.0	49.3	60.1	5
Ours	tva	$t \to v$	HT100M	R152 + RX101			✓	12.7	33.9	45.8	13	20.4	47.7	59.3	6
AVLNet [44]	tva	$t \to v+a$	HT100M	R152 + RX101			✓	30.2	55.5	66.5	4	22.5	50.5	**64.1**	5
MCN [12]	tva	$t \to v+a$	HT100M	R152 + RX101			✓	28.2	53.0	63.7	5	-	-	-	-
Ours	tva	$t \to va$	HT100M	R152 + RX101			✓	**32.1**	**59.1**	**70.9**	**3**	**23.7**	**52.1**	63.7	**4**

Table 2. Text-to-video retrieval on YouCook2/MSR-VTT in the fine-tune setting. In "Retrieval" column: $v + a$ stands for averaging video and audio embeddings for a video representation, va - our joint video-audio embedding where modalities attend each other during embedding computation, t and v are single-modality embeddings. Train Mod.=Training Modalities, BB=Backbone.

uation we use at maximum first 48 seconds of video, since most video are shorter than that. The MSR-VTT dataset contains human-annotated video clips ($\sim 10-30$ seconds) on various topics and provides captions with natural language sentences. Following [12, 35, 37, 44], to evaluate our model on MSR-VTT, we use the 1k set of test clips [53], and for YouCook2, we use 3,350 validation clips [37]. To perform retrieval, we compute similarities by dot product between a text query t and all videos in the dataset using a fused va representation for each video. We report standard recall metrics R@1, R@5, R@10 and the median rank (MedR).

Text-to-video Retrieval after Fine-tuning. We additionally evaluate the retrieval performance of the models fine-tuned on downstream tasks. Following [44], we used 9,586 training clips to fine-tune the model on the YouCook2 dataset, and 6,783 training clips that contain the audio (out of 7,000 proposed by [37]) to fine-tune model on the MSR-VTT.

Zero-shot Step Action Localization. We further evaluate our model on zero-shot step action localization tasks on two datasets: CrossTask [58] and Mining YouTube [29]. The CrossTask dataset consist of 2.7k instructional videos over 18 different tasks. The Mining YouTube dataset provides 250 testing cooking video equipped with an ordered list of action steps. To perform step localization, we use a sliding window and compute the similarity between the current video segment and all step names of the task. Following the inference procedure in [58], we obtain the final labeling by running dynamic programming to find the best labeling with respect to the given order of steps based on similarities. We report average recall over all tasks as defined in [58]. For both datasets, we use a 3-second sliding window with 1-second step and predict the action for the central time-stamp using a fused va representation.

4.3. Comparison with State-of-the-art

Zero-shot Text-to-video Retrieval. First, we assess the performance of the learned multi-modal representation in context of zero-shot text-to-video retrieval task on YouCook2 and MSR-VTT datasets as shown in Table 1. In case of the YouCook2 dataset, our method achieves state-of-the-art results over all baselines, including methods with trainable visual and text backbones or a stronger visual backbone as well as methods that do not train the visual backbone. Particularly, our approach improves the AVLnet [44] and MCN [12] baselines that use the same visual, text, and audio backbone and also train with three modalities by increasing R@10 from 45.2% to 51.3%. For MSR-VTT however, it shows that a fusion of video and audio modalities is not so beneficial and best performance is reached when considering only text to video retrieval and leaving out audio information. We attribute this behaviour to the domain shift between HowTo100M and MSR-VTT datasets as audio of the HowTo100M dataset mainly contains speech and text as a transcription of speech, while in MSR-VTT, audio can be much less related to the textual description. This assumption is supported by the fact that best-performing methods on MSR-VTT do not use HowTo100M for training at all, such as FrozenInTime [10] or CLIP4CLIP [32]. Notably, we can further strengthen our model on YouCook2 by leveraging a stronger backbone such as S3D [51], pre-trained on HowTo100M by MIL-NCE [35], reaching R@10 over 60%. Again, these results show that this better adaptation to HowTo100M does not necessarily translate to better results on MSR-VTT. In the supplement we additionally include experiments with a stronger CLIP [42] backbone.

Text-to-video Retrieval after Fine-tuning. We further evaluated the retrieval performance after fine-tuning on downstream tasks in Table 2. Note that, since several experimental splits were proposed for the MSR-VTT dataset, we report only baselines that used the same training split as us for a fair comparison. Results demonstrate that the proposed method clearly outperforms prior works on both datasets. Moreover, after finetuning on the MSR-VTT, the model greatly improves performance by utilizing an audio channel.

Zero-shot Step Action Localization. We further evaluate our methods on zero-shot step action localization on the CrossTask and the MiningYouTube (MYT) datasets in Table 3. As video representations we again use the fused video and audio modalities. These results show that the proposed approach clearly outperforms the directly comparable MCN approach on both datasets, as well as a fully supervised baseline CrossTask [58], HT100M [37] and MIL-NCE [35] with a trainable I3D visual backbone [35]. Moreover, with a stronger S3D backbone our model also gains improvement over MIL-NCE and is comparable to UniVL (with a trainable backbone) and ActBERT (with additional region-based features from Faster-R-CNN).

4.4. Ablation Studies

Impact of fusion components. We first address the question of how the proposed components: transformer layer, transformer fusion, and combinatorial loss, impact the overall performance of our system. To this end, in Table 4 we considered the following architectures: 1) *no transformers*: our architecture without transformer and with three pairwise contrastive losses; 2) *single modality transformer*: using three separate modality-specific transformer layers to learn three projection functions; 3) *fusion transformer*: using the proposed modality agnostic transformer but trained with three pairwise contrastive losses without fused modality components; 4) *fusion transformer + comb. loss*: using the proposed modality agnostic transformer with combinatorial input, trained with combinatorial loss to obtain the proposed method. Schematic visualization of these four setups is included in the supplement. We further consider two ways to forward two modalities, first by forwarding them separately and summing up both outputs $(v + a)$ and, second, by forwarding them together (va). Overall we observe that adding a simple transformer to process each modality separately already significantly improves performance compared to the baseline, especially for the YouCook2 dataset. We further observe that the overall performance depends on the combination of model, loss function and fusion strategy at test time. While token fusion with transformers is overall beneficial, the best performance is achieved in the *fusion transformer + comb. loss* setup. When using *fusion transformer* alone,

Method	Tr. Mod.	Tr. BB v	Visual Backbone	Recall↑ CrossTask	MYT
CrossTask [58]	tv		R152 + I3D	31.6	-
HT100M [37]	tv		R152 + RX101	33.6	15.0
MIL-NCE [35]	tv	✓	I3D	36.4	-
MCN [12]	tva		R152 + RX101	35.1	18.1
Ours	tva		R152 + RX101	**39.3**	**19.4**
Models with a stronger visual backbone:					
MIL-NCE [35]	tv	✓	S3D	40.5	-
ActBERT [57]	tv		Res3D+Faster R-CNN	41.4	-
UniVL [31]	tv	✓	S3D†	**42.0**	-
Ours	tva		S3D†	41.1	**19.7**

Table 3. Zero-shot action localization performance on CrossTask/MiningYouTube(MYT). S3D† is the S3D pre-trained by MIL-NCE [35]. Tr Mod=Training Modalities, Tr. BB v= Trainable Backbone for video modality.

Configuration	Retrieval	YouCook2 R@5↑	YouCook2 R@10↑	MSR-VTT R@5↑	MSR-VTT R@10↑
1) no transformers	$t \to v + a$	32.7	41.4	24.1	33.7
2) single-mod. transformer per mod.	$t \to v + a$	39.9	50.7	25.3	33.9
3) fusion transformer	$t \to v + a$	39.5	50.2	23.8	32.7
4) fusion transformer	$t \to va$	36.6	47.0	22.6	32.1
5) fusion transf. + comb. loss	$t \to v + a$	38.2	49.2	23.3	33.2
6) fusion transf. + comb. loss (ours)	$t \to va$	40.7	51.3	23.8	31.8

Table 4. Evaluation of the contribution of the proposed fusion transformer and the combinatorial loss. In "Retrieval" column: $v + a$ stands for independently extracting video and audio embeddings and summing up both outputs, while va for forwarding both modalities together allowing them to attend to each other.

Configuration	YouCook2 R@5↑	YouCook2 R@10↑	MSR-VTT R@5↑	MSR-VTT R@10↑
ours	40.7	51.3	23.8	31.8
ours + shared final proj.	37.2	48.7	23.1	29.4
ours + $[cls]$ token + shared final proj.	35.9	46.9	21.4	28.8

Table 5. Evaluation of different design choices for fusion transformer (with/without $[cls]$ token) and final projection (shared/modality specific) into multi-modal embedding space.

Configuration	YouCook2 R@5↑	YouCook2 R@10↑	MSR-VTT R@5↑	MSR-VTT R@10↑
positional emb. + uniform sampling	32.7	43.4	21.8	29.1
positional emb. + max polling	33.2	43.9	21.6	30.5
positional emb. + averaging over clips	35.6	46.9	23.0	29.2
no positional emb. + uniform sampling	34.5	45.1	22.1	30.2
no positional emb. + max polling	34.8	45.2	23.0	31.8
no positional emb. + averaging over clips	36.9	47.9	22.7	31.3
no positional emb. + video at once	40.7	51.3	23.8	31.8

Table 6. Evaluation of different strategies to process long videos (longer than clips used in the training), as well as impact of positional embeddings to encode positional information.

the performance drops a bit compared to the *single modality transformer*. However, utilizing shared transformer with combinatorial loss to fuse tokens from different modalities outperforms modality *summing* in the *single modality transformer* setup.

Token Aggregation and Projection. We further evaluate the impact of separate processing of output tokens compared to aggregating information in the $[cls]$ token. To this end, we compare our model architecture to a setup with a shared final projection, as well as a setup with an additional $[cls]$ input token (similarly to BERT [17]). In the last scenario, we use the output $[cls]$ of the multi-modal fusion transformer as an aggregated representation of input tokens and apply the final shared projection to map it into the shared embedding space. At shown in Table 5, our modality-specific projections with no $[cls]$ token benefit over others options on both datasets.

Positional Embedding and Testing on Longer Clips. Finally, we address the question how the option of having random length inputs impacts the overall performance of the model at test time. We consider four different scenarios for testing on longer clips as shown in Table 6: 1) *uniform sampling* - after obtaining initial local features from backbones, features are uniformly sampled to fit the maximum number of tokens; 2) *max-pooling* - local features are merged via adaptive max-pooling; 3) *averaging over clips* comprises slicing a longer clip into train-time-length clips and averaging obtained representations; 4) *video at once* considers processing all features at once as proposed. We further compare the first three settings in a scenario with positional embedding to no positional embedding. As positional embedding, we used vanilla trainable embeddings [17] that are summed up to the input tokens before being input to the transformer. It shows that setups with positional embedding perform almost consistently below the setups without positional embedding on both datasets. Looking at results of different processing strategies, we find that our model benefits from leveraging local temporal dependencies in data via slicing a video clip into shorter clips or *video at once* compared to *max-pooling*, *uniform sampling*. Moreover, utilizing all input data at *video at once* further significantly boosts performance.

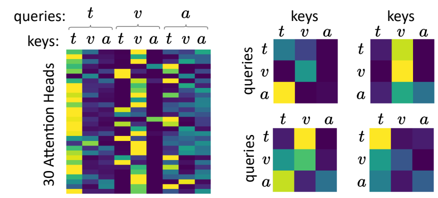

Figure 3. Attention Analysis. Left: Average attentions of 30 heads over 128 video clips for queries and keys from different modalities. Right: average attention for random 4 heads presented in 3x3 matrix.

Attention Analysis. Finally, we qualitatively analyze the fusion capability of our multi-modal transformer. In Figure 3 we show the average attention for query-key pairs of tokens from different modalities. We observe that some heads have a strong attention for single-modality fusion, mostly for t and v modalities, and in between, some heads are responsible for cross-modal attention.

5. Limitations and Conclusion

In this work, we propose a multi-modal, modality agnostic fusion transformer approach that learns to exchange information between multiple modalities, such as video, audio, and text, and to integrate them into a joined multi-modal representation. We show that training the system with a combinatorial loss on any possible combinations of modalities allows the fusion transformer to learn a strong multi-modal embedding space leaving out any add-ons such as position encoding. A clear limitation of the system becomes evident when looking at the performance difference on two downstream datasets, YouCook2 and MSR-VTT, showing that a better fusion can result in a loss in generalizability to multi-modal data that was acquired in a different way. A future research direction to mitigate those effects might be to consider techniques from domain adaptation or generalization in context of multi-modal zero-shot recognition. We hope that the proposed setup might inspire further research on this topic as well as on self-attention based multi-modal video processing in general.

6. Acknowledgments

This work was supported by the Goethe Center for Scientific Computing (G-CSC) at Goethe University Frankfurt, the MIT-IBM-Satori team, and the IBM-MIT Watson AI Lab.

References

[1] 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. *arXiv preprint arXiv:2104.11178*, 2021. 1, 3, 5

[2] Jean-Baptiste Alayrac, Adria Recasens, Rosalia Schneider, Relja Arandjelović, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. Self-supervised multimodal versatile networks. In *NeurIPS*, 2020. 1, 2, 3, 5, 6

[3] 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

[4] Elad Amrani, Rami Ben-Ari, Daniel Rotman, and Alex Bronstein. Noise estimation using density estimation for self-supervised multimodal learning. *arXiv preprint arXiv:2003.03186*, 2020. 1, 2, 5, 6

[5] Relja Arandjelovic and Andrew Zisserman. Objects that sound. In *ECCV*. 2

[6] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *ICCV*, pages 609–617, 2017. 2

[7] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. *NeurIPS*, 29, 2016. 2

[8] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. See, hear, and read: Deep aligned representations. *arXiv preprint arXiv:1706.00932*, 2017. 2

[9] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 3

[10] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *ICCV*, 2021. 1, 3, 6, 7

[11] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *CVPR*, 2017. 5

[12] Brian Chen, Andrew Rouditchenko, Kevin Duarte, Hilde Kuehne, Samuel Thomas, Angie Boggust, Rameswar Panda, Brian Kingsbury, Rogerio Feris, David Harwath, et al. Multimodal clustering networks for self-supervised learning from unlabeled videos. *ICCV*, 2021. 1, 2, 3, 5, 6, 7

[13] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP*, 2020. 2

[14] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *ECCV*, 2020. 3

[15] Ying Cheng, Ruize Wang, Zhihao Pan, Rui Feng, and Yuejie Zhang. Look, listen, and attend: Co-attention network for self-supervised audio-visual representation learning. In *ACMMM*, 2020. 3

[16] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009. 5

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 8

[18] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020. 4

[19] Kevin Duarte, Brian Chen, Nina Shvetsova, Andrew Rouditchenko, Samuel Thomas, Alexander Liu, David Harwath, James Glass, Hilde Kuehne, and Mubarak Shah. Routing with self-attention for multimodal capsule networks. *arXiv preprint arXiv:2112.00775*, 2021. 2

[20] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *ECCV*, 2020. 1, 3, 6

[21] Andrey Guzhov, Federico Raue, Joern Hees, and Andreas Dengel. Audioclip: Extending clip to image, text and audio. In *GCPR*, 2021. 1

[22] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. Can spatiotemporal 3d cnns retrace the history of 2d cnns and imagenet? In *CVPR*, 2018. 5

[23] David Harwath, Adria Recasens, Dídac Surís, Galen Chuang, Antonio Torralba, and James Glass. Jointly discovering visual objects and spoken words from raw sensory input. In *ECCV*, 2018. 2

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 5

[25] Dat Huynh and Ehsan Elhamifar. A shared multi-attention framework for multi-label zero-shot learning. In *CVPR*, 2020. 2

[26] Andrew Jaegle, Sebastian Borgeaud, Jean-Baptiste Alayrac, Carl Doersch, Catalin Ionescu, David Ding, Skanda Koppula, Daniel Zoran, Andrew Brock, Evan Shelhamer, Olivier Hénaff, Matthew M. Botvinick, Andrew Zisserman, Oriol Vinyals, and João Carreira. Perceiver io: A general architecture for structured inputs & outputs. *arXiv preprint arXiv:2107.14795*, 2021. 2, 4

[27] Lukasz Kaiser, Aidan N Gomez, Noam Shazeer, Ashish Vaswani, Niki Parmar, Llion Jones, and Jakob Uszkoreit. One model to learn them all. *arXiv preprint arXiv:1706.05137*, 2017. 2

[28] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[29] Hilde Kuehne, Ahsan Iqbal, Alexander Richard, and Juergen Gall. Mining youtube - a dataset for learning fine-grained action concepts from webly supervised video data. *arXiv preprint arXiv:1906.01012*, 2019. 2, 6

[30] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. In *CVPR*, 2021. 3

[31] Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Xilin Chen, and Ming Zhou. Univl: A unified

video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353*, 2020. 3, 7

[32] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. CLIP4Clip: An empirical study of clip for end to end video clip retrieval. *arXiv preprint arXiv:2104.08860*, 2021. 1, 3, 7

[33] Shuang Ma, Daniel McDuff, and Yale Song. Unpaired image-to-speech synthesis with multimodal information bottleneck. In *CVPR*, pages 7598–7607, 2019. 2

[34] Massimiliano Mancini, Muhammad Ferjad Naeem, Yongqin Xian, and Zeynep Akata. Open world compositional zero-shot learning. In *CVPR*, 2021. 2

[35] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *CVPR*, 2020. 1, 2, 6, 7

[36] Antoine Miech, Ivan Laptev, and Josef Sivic. Learnable pooling with context gating for video classification. *arXiv preprint arXiv:1706.06905*, 2017. 5

[37] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *ICCV*, 2019. 1, 2, 3, 5, 6, 7

[38] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013. 5

[39] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion. *NeurIPS*, 34, 2021. 3

[40] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018. 5

[41] Mandela Patrick, Po-Yao Huang, Yuki Asano, Florian Metze, Alexander Hauptmann, João Henriques, and Andrea Vedaldi. Support-set bottlenecks for video-text representation learning. *arXiv preprint arXiv:2010.02824*, 2020. 6

[42] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021. 2, 3, 7

[43] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *Proceedings of the International Conference on Machine Learning*, pages 1060–1069. PMLR, 2016. 2

[44] Andrew Rouditchenko, Angie Boggust, David Harwath, Brian Chen, Dhiraj Joshi, Samuel Thomas, Kartik Audhkhasi, Hilde Kuehne, Rameswar Panda, Rogerio Feris, Brian Kingsbury, Michael Picheny, Antonio Torralba, and James Glass. Avlnet: Learning audio-visual language representations from instructional videos. In *Interspeech*, 2021. 1, 2, 3, 5, 6, 7

[45] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 2018. 6

[46] Zineng Tang, Jie Lei, and Mohit Bansal. Decembert: Learning from noisy instructional videos via dense captions and entropy minimization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021. 3

[47] Antigoni Tsiami, Petros Koutras, and Petros Maragos. Stavis: Spatio-temporal audiovisual saliency network. In *CVPR*, 2020. 2

[48] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 3, 4

[49] Xiaohan Wang, Linchao Zhu, and Yi Yang. T2vlad: Global-local sequence alignment for text-video retrieval. In *CVPR*, 2021. 3

[50] Fanyi Xiao, Yong Jae Lee, Kristen Grauman, Jitendra Malik, and Christoph Feichtenhofer. Audiovisual slowfast networks for video recognition. *arXiv preprint arXiv:2001.08740*, 2021. 2, 4

[51] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *ECCV*, 2018. 7

[52] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*, 2016. 2, 5

[53] Youngjae Yu, Jongseok Kim, and Gunhee Kim. A joint sequence fusion model for video question answering and retrieval. In *ECCV*, 2018. 6

[54] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *CVPR*, 2021. 2

[55] Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In *AAAI*, 2018. 2, 5

[56] Yipin Zhou, Zhaowen Wang, Chen Fang, Trung Bui, and Tamara L Berg. Visual to sound: Generating natural sound for videos in the wild. In *CVPR*, pages 3550–3558, 2018. 2

[57] Linchao Zhu and Yi Yang. Actbert: Learning global-local video-text representations. In *CVPR*, 2020. 6, 7

[58] Dimitri Zhukov, Jean-Baptiste Alayrac, Ramazan Gokberk Cinbis, David Fouhey, Ivan Laptev, and Josef Sivic. Cross-task weakly supervised learning from instructional videos. In *CVPR*, 2019. 2, 6, 7