Abstract

When people observe events, they are able to abstract key information and build concise summaries of what is happening. These summaries include contextual and semantic information describing the important high-level details (what, where, who and how) of the observed event and exclude background information that is deemed unimportant to the observer. With this in mind, the descriptions people generate for videos of different dynamic events can greatly improve our understanding of the key information of interest in each video. These descriptions can be captured in captions that provide expanded attributes for video labeling (e.g. actions/objects/scenes/sentiment/etc.) while allowing us to gain new insight into what people find important or necessary to summarize specific events. Existing caption datasets for video understanding are either small in scale or restricted to a specific domain. To address this, we present the Spoken Moments (S-MiT) dataset of 500k spoken captions each attributed to a unique short video depicting a broad range of different events. We collect our descriptions using audio recordings to ensure that they remain as natural and concise as possible while allowing us to scale the size of a large classification dataset. In order to utilize our proposed dataset, we present a novel Adaptive Mean Margin (AMM) approach to contrastive learning and evaluate our models on video/caption retrieval on multiple datasets. We show that our AMM approach consistently improves our results and that models trained on our Spoken Moments dataset generalize better than those trained on other video-captions datasets.

http://moments.csail.mit.edu/spoken.html

1. Introduction

Video understanding has typically been focused on action recognition and object tracking as the temporal aspect of videos lends itself strongly to the task of representing motion, a key component of an action. Breaking down video analysis to simple tasks, such as action recognition, allows for efficient data annotation for building large datasets to train deep learning models [31, 46, 21] which has been extremely successful for images with object annotations [35]. A main difficulty is that, in contrast to an image, a video often captures an interaction between agents and objects that evolves over time. These interactions can be as simple as “a person picking up a glass of water”, but even in this case three different objects (“person”, “glass” and “water”) are included in the interaction. Additionally, the video may also continue to depict the “person drinking from a glass” and the “person putting the glass back down on the table”. These sequential events present additional challenges for video datasets where single annotations may not be sufficient to explain the events depicted. Multi-label approaches to video annotation have attempted to address this problem by labeling multiple actions in a video [47, 22, 73]. However, these methods focus on single domain annotations, such as actions or objects, and do not capture additional contextual information, such as “person angrily putting down the dirty glass on a rusted table”, which can change the interpretation of an event and how it fits into a sequence of observations.

A solution for capturing more fully the content of video is to annotate multiple actions or objects in each video [22, 72, 47, 50]. However labels like “drinking”, “glass”, only provide a portion of the information needed to interpret the veracity of the event. Additional narratives may include intuitive descriptions and intentions, such as “an exhausted man picks up a dirty glass of water and drinks from
it before angrily putting it down on a table" which would dramatically change the event interpretation. The full lingual description combines these actions with adjectives and nouns (objects) that contextualize the events depicted leading to a better understanding of the video. This is our goal in providing a new large scale dataset for training models for full video understanding.

We introduce a large scale video caption dataset, Spoken Moments in Time (S-MiT), to allow deep learning models for video understanding to learn contextual information. Most existing video description datasets [71, 60, 33, 20, 80] are limited in size when compared to the large datasets for action recognition [31, 46, 21]. A likely cause is the increased cost of collecting full text descriptions for videos compared to single label annotations. Recent work in image captioning [25] addressed this problem by collecting audio descriptions for a large set of images from the Places dataset [77]. Collecting spoken captions is faster and more efficient due to the low overhead of speaking compared to typing. In addition, recording of spontaneous speech rather than typed text can produce more natural descriptions of an event. An automatic speech recognition (ASR) system was then used to transcribe the spoken descriptions to text captions. In this work, both audio, text and video models were jointly trained via contrastive learning to align the visual and caption representations learned via contrastive learning to align the visual and caption representations. We evaluate our models on multiple datasets for action recognition such as UCF101 [62] and HMDB [36]. With the increased availability of large scale video datasets, many different models have been proposed to improve performance on a number of video understanding tasks. Two-stream convolutional neural networks (CNNs) combine optical flow with RGB frames to capture both temporal and spatial information [61]. 3D models [8] combine 3D CNNs [65], which use a 3D kernel to learn temporal information from a frame sequence, with optical flow to form a two-stream 3D network “inflated” from 2D filters pre-trained on ImageNet [16]. More recently a temporal shift module has been proposed to integrate temporal information into 2D models by shifting frame representations across the temporal dimension [40].

Recently multi-modal visual understanding methods have received significant attention [25, 64, 47, 29, 4]. The DAVEnet model [25] has been proposed for jointly learning aligned representations between images and spoken captions, and has been extended to align frame-wise video representations with synchronized audio narration for cross-modal audio-visual concept learning [4]. Here, we build on the motivation from this paper and learn aligned representations between videos and unsynchronized spoken descriptions using the S-MiT Dataset.

1. The large-scale Spoken Moments in Time dataset (S-MiT) which includes 500k pairs of video clips and corresponding audio descriptions. This new dataset represents the largest video description dataset available and will serve as a new benchmark for the community.

2. Benchmark models with aligned spoken caption and video representations learned via contrastive learning. We compare approaches that learn directly from the spoken captions to models that include a trained ASR model which feeds generated text transcriptions into an NLP language model. We then jointly train caption and visual models (based on concatenated video and image features) using a novel Adaptive Mean Margin (AMM) approach to contrastive learning to align the visual and caption representations. We evaluate our models on multiple datasets for video/caption retrieval and show that a model trained using AMM on S-MiT achieves the best general performance across four datasets.

Altogether, our novel contributions include:

1. The large-scale Spoken Moments in Time dataset (S-MiT) which includes 500k pairs of video clips and corresponding audio descriptions. This new dataset represents the largest video description dataset available and will serve as a new benchmark for the community.

2. Benchmark models with aligned spoken caption and video representations learned via contrastive learning. We compare approaches that learn directly from the spoken captions as well as approaches that include ASR transcriptions that feed into different language models to generate caption representations.

3. An Adaptive Mean Margin (AMM) approach to cross-entropy based contrastive learning.

2. Related work

2.1. Video Understanding

The field of video understanding has recently seen fast progress partly due to the availability of large scale video datasets including ActivityNet [6], Kinetics [31], Moments in Time [46, 47] and YouTube-8M [1]. These large datasets are used to pretrain models that are fine-tuned on smaller action recognition datasets such as UCF101 [62] and HMDB [36]. With the increased availability of large scale video datasets, many different models have been proposed to improve performance on a number of video understanding tasks. Two-stream convolutional neural networks (CNNs) combine optical flow with RGB frames to capture both temporal and spatial information [61]. 3D models [8] combine 3D CNNs [65], which use a 3D kernel to learn temporal information from a frame sequence, with optical flow to form a two-stream 3D network “inflated” from 2D filters pre-trained on ImageNet [16]. More recently a temporal shift module has been proposed to integrate temporal information into 2D models by shifting frame representations across the temporal dimension [40].

2.2. Caption Datasets

There have been a number of different datasets released for providing language descriptions of visual information. Flickr8k [28] and Flickr30k [49] include 8k and 30k images respectively each sourced from Flickr. Each image is associated with 5 text captions describing what is in the image. An additional set of 5 audio captions per image in both sets was recently collected for learning joint embeddings between speech and images [25]. The Visual Genome dataset [34] includes captions for multiple regions of more than 180k images allowing for fine-grained descriptions of each image. The Places Audio Caption dataset [26] contains approximately 400k images from the Places 205 [78] image dataset with audio captions of people verbally describing each image. MS COCO [11] is a large image dataset for object recognition, segmentation, and captioning which includes roughly 1 million captions for 160k Flickr images. Conceptual Captions [59] contains 3.3M images with captions generated from HTML attributes associated with web based images. The Stock3M dataset [70] includes 3.2 mil-
Figure 1: Examples from the Spoken Moments Dataset: The dataset is composed of videos and the corresponding spoken captions. We show some examples of the text transcriptions, automatically generated using the public Google ASR engine.

Beyond the numerous datasets available or image captioning [28, 49, 34, 11, 59, 70], including those that provide spoken descriptions [26, 25], there is a variety of different video caption datasets available. A number of these datasets are related to cooking [51, 52, 55, 13, 12] including YouCook [14] and YouCook II [80] which include 2k videos from YouTube each with multiple captions annotated at different segments of each video. MPII-Movie Description Corpus [53] contains transcribed audio descriptions from 94 Hollywood movies split into 68k clips where each clip is paired with a sentence from the movie scripts and an audio description of the visual content in each clip. Similarly, Large Scale Movie Description Challenge (LSMDC) dataset [54] contains 200 movies with 120K sentence descriptions. VideoStory [20] contains 20k social media videos where each video contains a paragraph length description. The ActivityNet Captions dataset [33] has 20k videos with 100k text descriptions. The Microsoft Video Description (MSVD) dataset [9] contains 2k YouTube clips with a 10-25 second duration and an average of 41 single sentence descriptions per clip. MSR-Video to Text (MSR-VTT) [71] contains 7k videos split into 10k clips with 20 captions per video.

HowTo100M [45] contains 136 million clips sourced from 1.22 million instructional videos with narrations generated from subtitles associated with each video. However, the subtitles are not human verified captions and the content is constrained to instructional videos. Since the text associated with the clips in the HowTo100M dataset are transcriptions of a narrator completing a task in the video, the short text phrases from the subtitles occasionally share noisy associations with the reference clip. In Section 5, and Table 2, we decided to compare our contributions using strict caption datasets as we are proposing a large-scale human annotated caption dataset with full human generated descriptions for each video.

VaTeX [69] contains 41k videos sourced from the Kinetics-600 dataset [31, 7] annotated with 10 English captions and 10 Chinese captions for multilingual captioning. VaTeX is the most similar to our proposed dataset in that it is sourced from an existing video dataset for action recognition and the captions are directly annotated.

In this work, we present a new dataset, Spoken Moments in Time (S-MiT), which includes spoken audio captions for 500k unique three second clips each with different source videos from the Moments in Time dataset [46, 47]. In addition to vast increase in scale over other video-caption datasets, a major contribution is that we are using spoken descriptions rather than text. This allows us to train spoken caption models to directly align with video models. This is not possible with the other large video caption datasets and allows for spoken caption models to be analyzed with matching video information. We also show that models trained on our S-MiT dataset generalize much better in retrieval to the video-caption pairs in other datasets. This is due to the large coverage, diversity and scale of our proposed dataset.

2.3. Cross Modal Contrastive Learning

Cross modal learning has been used to jointly self-supervise audio-visual models [3, 48, 76] with synchronized information while NLP approaches have been leveraged to align joint representations for both visual and language modalities using spoken and text descriptions [2, 79]. This is typically done via Contrastive Learning where the alignment between positive pairs (language and visual input) is trained to be stronger than those of non-positive pairs [24]. For visual representations, a triplet based max-margin loss is commonly used to discriminate representations between positive and negative pairs [74, 75, 18]. Semi-hard negative mining [58] and a dot-product based similarity score have been used to jointly learn audio-visual embeddings between images and spoken captions [25] while batch-wise cross-entropy approaches to contrastive learning have been used to increase the amount of information utilized in learning by considering all negative examples in a mini-batch [66, 10]. Work on bidirectional speech/image retrieval using audio descriptions of images integrated ideas from max-margin contrastive learning and added a margin into the cross-entropy loss [29]. SimCLR [10] added a non-linear projection head that maps paired representations into a common space allowing for stronger representations.

A pretrained language model has recently been used to improve cross-modality learning with language and visual input pairs. ViLBERT [43] added a pretrained BERT
transformer to capture semantic language representations associated with object detection proposals from a pretrained faster RCNN network. VideoBERT [63] extended BERT to jointly learn the visual and linguistic domain by generating tokenized visual words. Inspired by this prior work, we propose adding a pretrained language model that maps word predictions from a trained ASR model to semantic language features in order to generate rich spoken caption representations. We then utilize an MLP to project these caption representations, and our video representations, to an aligned joint representation which can be used for video/caption retrieval (see Section 5).

2.3.1 Optimization Approaches

A common approach to optimization in contrastive learning settings is to use a similarity based loss function. We formulate the contrastive loss as, $\mathcal{L} = \mathcal{L}_{xc} + \mathcal{L}_{cv}$, where the goal is to maximize the discrimination between positive and negative paired captions $c$ and videos $v$. The loss is split into two tasks where $\mathcal{L}_{xc}$ forms pairs from a fixed video and each caption in a sampled mini-batch, while $\mathcal{L}_{cv}$ fixes the caption and forms pairs with each video in the mini-batch. Below we discuss different approaches of $\mathcal{L}_{xy}$, where $x$ and $y$ are interchangeable with $v$ and $c$.

Semi-hard negative mining (SHN) [58] has been used for learning aligned cross-modal embeddings using a triplet loss [25, 30]. This is an improvement over hard negative mining [19] since a sampled negative example is constrained to be less similar to the anchor than the positive sample while still being within the margin and thus contributing a loss at each step with the margin $M = 1$, $\mathcal{L}_{xy} = \max (S(x_i, y_j) - S(x_i, y_i) + M, 0)$, where $S(x_i, y_j)$ is a similarity score for the representations of $x_i$ and $y_j$, with $x_i$ and $y_j$ forming a positive pair.

Noise contrastive estimation (NCE) [23] has been applied to contrastive learning [10, 66] by using a log-likelihood based loss function that learns to discriminate between positive and negative pairs of feature embeddings,

$$\mathcal{L}_{xy} = -\frac{1}{B} \sum_{i=1}^{B} \log \frac{e^{S(x_i, y_i)}}{\sum_{j=1}^{B} I_{i \neq j} e^{S(x_i, y_j)}},$$  

where $I_{i \neq j}$ is an indicator function that we only considers negative pairs in the denominator. This has been shown to improve feature alignment compared to SHN [10].

Masked Margin Softmax Loss (MMS) [29] and Large Margin Cosine Loss (LMCL) [68] incorporate a positive margin into the contrastive learning framework in order to improve feature discrimination among non-paired embeddings. MMS uses a monotonically increasing margin to allow for initial learning to begin to converge before a large alteration to the loss is added. LMCL proposes a theoretical limit on the maximum margin size of $1 - \cos \frac{\pi}{N}$ where $N$ refers to the number of classes being discriminated. For aligning captions to visual information, the class size can be considered unbounded as each caption represents a slightly different representation that we want to discriminate leading to a max margin size of 1. Concretely, MMS proposes adding a margin to Equation 1,

$$\mathcal{L}_{xy} = -\frac{1}{B} \sum_{i=1}^{B} \log \frac{e^{S(x_i, y_i) - M}}{\sum_{j=1}^{B} I_{i \neq j} e^{S(x_i, y_j)}},$$

where the margin, $M$, starts as 0.001 and is exponentially increased by a factor of 1.002 every 1000 training steps.

We propose extending the idea of an increasing margin in MMS to an adaptive setting that does not require setting the initial value of the margin or the growth rate. We refer to this approach as an Adaptive Mean Margin (AMM) where the margin is set as the mean distance between the positive pair and the set of negative pairs in a batch. We describe AMM in more detail in Section 4.3.

3. The Spoken Moments Dataset

We begin with the Moments in Time dataset [46] as it includes over 1 million videos sourced from a number of different video hosting sites with strong inter & intra-varietal variation in terms of the number of events depicted in each video. Further, the videos are all cut to 3 seconds allowing for a concise description to effectively capture the localized information of each event. Here we refer to concise descriptions as those that focus on key events depicted in the video and do not imply partial descriptions. In data collection, annotators may watch a video as many times as desired. During recording, we block the annotators from seeing/hearing the video to encourage descriptions of important memorable events rather than every specific detail. This approach does not preclude the annotators from describing sequential or simultaneous events as shown in our qualitative examples (see Figure 1). We describe our annotation approach in more detail in the supplementary material.

3.1. Dataset Statistics

Our proposed Spoken Moments dataset contains 500k videos randomly chosen from the Multi-Moments in Time (M-MiT) training set and all of the 10k videos from the validation set. Each video in the training set contains at least one audio description. We transcribed each audio recording using the public Google Automatic Speech Recognition (ASR) engine to generate text captions for each video. When analyzing these transcriptions, we build a picture of the coverage and diversity of our captions. Table 2 (left) shows that our captions have an average length of 18 words with a unique vocabulary of 50,570 words consisting of 20,645 nouns, 12,523 adjectives and 7,436 verbs with a...
total word count of 5.6 million. Table 2 (right) shows a comparison of our Spoken Moments dataset to other existing datasets for video captioning. Our dataset will be the largest public dataset in terms of video clips, source videos, total number of captions, total words in the captions and the vocabulary set of unique words occurring in the captions. The increase in vocabulary size is important as it shows that our increase in the number of videos over previous datasets does not simply include repeated events but covers a novel breadth of information. We can see the opposite effect of this in YouCook II [80] where the restricted domain of cooking videos results in a limited vocabulary used in the descriptions.

To understand how this vocabulary covers the class labels typically used for training computer vision models, we examined whether these labels exist in our vocabulary. Table 2 (right) shows that we have strong coverage of the two largest action recognition datasets for video understanding (Kinetics [31] and M-MiT [47]). We expected a large coverage of the events in M-MiT as we sourced our videos from this dataset and the action labels themselves are fairly general (e.g. “running” and “cooking”). For Kinetics, the labels are commonly tied to a noun preceded by a verb (e.g. “brushing hair”). For these labels we consider them to exist in our dataset if both the verb and noun are in the same caption. For example, “A boy is in a bathroom brushing his teeth” would cover the class “brushing teeth”. With this approach we see a 85.1% coverage of the classes in Kinetics and a 96.2% coverage of the classes in M-MiT showing a strong level of event diversity. Similarly we see a strong overlap of the object classes in MS-COCO [41] (100%) and ImageNet [16] (69.2%) in our captions. ImageNet coverage is likely lower due to the specific labels used for many of its classes (e.g. “coucal”). Still, 69.2% coverage means 692 ImageNet classes appear in our captions. Similarly, Places [78] scene labels are very specific and don’t necessarily match the language used in our descriptions. For example, an “abbey” will typically be described as a “church” or “monastery” in our captions. We did not account for all of the synonyms possible and are only considering direct matches in our captions. Even so we are able to find a 47.4% coverage of the scene labels in Places365 in our dataset.

Here we provide information on some additional characteristics of our data that may be of interest. While we do not release demographic info of our annotators or captions, about 57% of the spoken captions were recorded by male voices and 43% female. For the audio streams of the videos, roughly 51% include natural sound, 5% have music as the audio and 44% have no audio. This is consistent with the M-MiT dataset [47] from which we source our videos. Additionally, we found that less than 3% of the videos contain captions that describe non-visible events (e.g. a car horn when no car is visible in the video frames). For this reason we have chosen to focus our approach on learning a strong visual model in Section 4.

4. Learning Audio-Visual Representations

In order to learn from the large set of spoken captions in the proposed S-MiT dataset, we adopt a cross-modal architecture used in prior work [45, 25, 56] which is composed of a video model and a caption model as depicted in Figure 3. Specifically, we take N video-caption pairs as input and encode each modality into a 4096-D feature vector. We do this by adding a multilayer perceptron (MLP) as a projection head on top of both the video and the caption model. This projection head is composed of two linear layers followed by gated linear units (GLU) [15]. We then compute the dot product between the video and caption representa-
tions to produce an $N \times N$ similarity matrix, $S$, which is used to compute our contrastive loss for training. In Section 4.3, we describe our modified approach to margined contrastive learning which uses an Adaptive Mean Margin (AMM) which automatically adjusts itself during learning to improve the optimization signal during training.

### 4.1. Video Model

Following prior work [45], we use two encoders to represent input videos: image & video encoders. Specifically, we use a ResNet-152 [27] pretrained on ImageNet [35] and a temporal shift module (TSM) ResNet-50 model [40] pretrained on M-MiT [47]. Each encoder outputs a 2048-D feature vector after max-pooling over the temporal dimension (8 frames for the TSM (~3 fps) and 3 frames for the image model (1 fps)). We concatenate the two 2048-D vectors and feed the concatenated vector into an MLP projection head to get the final 4096-D visual representation. We examine the effect of using the image and video encoders as well as different pretrained models in the supplementary material.

### 4.2. Caption Model

#### 4.2.1 Language Caption Model

Prior work in learning joint representations between audio captions and visual models has shown that utilizing ASR transcriptions greatly improves results [25]. We build on this idea and use the predicted words from a pretrained ASR model (e.g. Google’s public ASR engine) to train our models. Concretely, we examine the effect of using different pretrained language models stacked on top of the ASR model predictions. We begin by comparing the results of using Fasttext [5], BART [37] and BERT [17] models to generate semantic and contextual word representations for our captions. During training, we randomly select 10 words from each caption to be included in training. In the case of the BART and BERT models, this selection happens after the full transformer model has been applied to avoid altering the results from the self-attention mechanisms. If less than 10 words occur in a caption then we allow words to be sampled multiple times in the random selection. This training augmentation allows different words in each caption to be represented differently at different training iterations. We examine the effect of this approach in the supplementary material. In test, we use the full transcription as input into the language model. We average the word representations from the output of the language model to generate a single representation for each caption which we align to the video representations described in the previous section.

#### 4.2.2 Spoken Caption Model

We also train caption models with raw spoken captions instead of the corresponding transcription. For each caption, we randomly sample 10 seconds of speech for training and compute the 40-dimensional log Mel spectrogram to serve as the input of spoken caption model. The input is fed into a spoken caption model where we consider ResDavenet [25] (which is designed specifically for speech) and two ImageNet ResNet [27] models (ResNet-34, ResNet-50). For the ResNet models, we modify the first convolutional layer to take the 1-channel input so that spectrogram can be processed. In addition, the wav2vec [57] model, which takes raw waveform as the input, is also involved in our experiments. Spoken captions are first fed into the pre-trained wav2vec model, which produces 512-D vectors per 210 ms. We then feed them into a learnable ResStack, taken from ResDavenet, to learn representations of spoken captions.

### 4.3. Adaptive Mean Margin

We train our model using the contrastive loss with a similar setting to MMS (Equation 2). The only difference is that we replace the margin, $M$, with an adaptive margin based on the difference between the similarity of the positive pair and the set of negative pairs in each batch.

The challenge in using the MMS margin for mini-batch sampled contrastive learning is that the initial margin and growth schedule are difficult to tune for a specific dataset and similarity metric. Additionally, depending on the sampled pairs in a mini-batch, the margin calculated may be too weak if the positive pair is much more similar than the sampled negative pairs and too strong if it is very similar to the negative pairs. The approach to monotonically increase the margin during training is meant to address this as the positive and negative pairs will share similar alignment early in training and begin to diverge closer to convergence. However, variable rates of convergence of different models on different datasets make this growth rate difficult to tune and this approach does not account for differences in the negative samples that appear in different mini-batches. To address this, we propose an adaptive margin based on relative batch-wise similarity scores.

Class labels have been proposed to be used for generating adaptive margins based on class similarity between positive and negative pairs [38, 42]. Likewise, prior work explored a non-class dependant approach for an adaptive similarity-based margin for human pose estimation [39] where the mean joint error between a positive pose and a hard sampled negative pose was used as a margin with the triplet loss. This adaptively increases the margin when the sampled negative pair is dissimilar to the positive pair in order to maximize the learning signal on less aligned negative samples. We follow a similar intuition and simply replace $M$ in Equation 2 with

$$M_{xy} = \alpha(S(x_i, y_i) - \frac{1}{B-1} \sum_{j=1}^{B} I_{i \neq j} S(x_i, y_j)),$$  \hspace{1cm} (3)
Table 1: Language Caption Model Comparison on Video/Caption Retrieval: Here we compare the video/caption retrieval results on the test set of the Spoken Moments dataset using models trained with three different language models.

| Language Caption Model | R@1 | R@5 | R@10 | mAP  |
|------------------------|-----|-----|------|------|
| Fasttext [3]           | 17.1 | 44.6 | 60.6 | 36.6 |
| BERT [17]              | 25.0 | 56.0 | 68.4 | 39.1 |
| BART [77]              | 32.3 | 66.2 | 79.3 | 43.8 |

| ActivityNet [33] | R@1 | R@5 | R@10 | mAP  |
|------------------|-----|-----|------|------|
| NCE              | 31.1 | 66.5 | 79.7 | 47.2 |
| SHN              | 33.1 | 68.1 | 80.9 | 56.4 |
| AMM              | 40.4 | 68.3 | 78.7 | 52.3 |

| ActivityNet Mean | R@1 | R@5 | R@10 | mAP  |
|------------------|-----|-----|------|------|
| NCE              | 19.1 | 48.1 | 61.2 | 32.5 |
| SHN              | 22.9 | 50.9 | 63.8 | 37.0 |
| AMM              | 31.6 | 62.1 | 70.0 | 46.5 |

| ActivityNet Loss | R@1 | R@5 | R@10 | mAP  |
|------------------|-----|-----|------|------|
| NCE              | 22.6 | 49.9 | 60.3 | 35.8 |
| SHN              | 28.1 | 51.0 | 62.5 | 39.7 |
| AMM              | 33.1 | 64.8 | 77.4 | 44.8 |

**Table 2: Loss Function Comparison for Video/Caption Retrieval:** Models trained on four datasets with different loss functions are compared. The proposed AMM loss function consistently achieves the best performance.

| Spoken Caption Model | Loss | R@1 | R@5 | R@10 | mAP  |
|----------------------|------|-----|-----|------|------|
| ResDavnet [25]       | NCE  | 30.7 | 57.9 | 68.7 | 42.9 |
| SHN                  | 30.7 | 57.9 | 68.7 | 42.9 |
| AMM                  | 32.9 | 60.2 | 70.7 | 46.5 |

**Table 3: Spoken Caption Model Comparison:** Models trained with different spoken caption architectures and different loss functions are compared for video/caption retrieval on the S-MiT test set. The proposed AMM loss function consistently achieves the highest performance while ResNet-50 is found to be significantly stronger than the other architectures.

| Train On | R@1 | R@5 | R@10 | mAP  |
|----------|-----|-----|------|------|
| ResNet-44 | NCE  | 22.6 | 49.8 | 60.3 | 35.8 |
| SHN      | 28.1 | 51.0 | 62.5 | 39.7 |
| AMM      | 33.1 | 64.8 | 77.4 | 44.8 |

**Table 4: Cross Dataset Evaluation on Video/Caption Retrieval:** Here we compare the generalization performance of models trained on four different datasets for video/caption retrieval. Each model is trained on a single dataset and we average the evaluation on five 1k video-captionsamples from the test set of each other dataset. We additionally show the mean performance across datasets. The S-MiT model shows it generalizes very strongly to the other datasets even beating the MSR-VTT model on its own test set.

| Evaluated On | R@1 | R@5 | R@10 | mAP  |
|--------------|-----|-----|------|------|
| MSR-VTT     | 22.6 | 49.8 | 60.3 | 35.8 |
| S-MiT        | 33.1 | 64.8 | 77.4 | 44.8 |

where α is a damping parameter to weight the strength of the margin. When \( M_{xy} \) in Equation 3 is applied to Equation 2 with \( α = 1 \), the margin removes the positive pair similarity from the optimization. Ablation studies on different alpha values can be found in the supplementary material. In practice we use \( α = 0.5 \) in our experiments. This has the effect of increasing the margin as the difference between the true pair similarity and the similarity of the negative pairs increases. As the training progresses, and the learning approaches convergence, the margin generally increases with the increased separation between positive and negative pair-wise similarities. This also removes the need to tune the margin and growth rate which may have different optimal values for different similarity metrics.
rics, batch sizes and datasets.

We refer to this as an Adaptive Mean Margin (AMM) for contrastive learning and show in Section 5 the effect of applying this adaptive margin.

5. Results

5.1. Video/Caption Retrieval

In Tables 1, 2 and 3 we show results of R@k recall scores (for $k = 1, 5, 10$) and mean average precision (mAP) on both caption to video and video to caption retrieval. Results are averaged over five random sets of 1k video-caption pairs from the test set. Each model in Tables 1 and 2 uses the output of a pretrained ASR model, the Google Cloud ASR engine, as input into a trained language model to generate a feature representation for each caption. Alternatively, the spoken caption models align visual representations directly from the audio signal without pretrained modules.

Table 1 shows the result of using different language models to generate our caption representations from ASR text transcriptions. Each of these models was trained using the proposed AMM loss function described in Section 4.3. We evaluate the AMM loss in Table 2 where we compare the results on the NCE, SHN, MMS and AMM loss functions described in Sections 2.3.1 and 4.3 on four different datasets (the proposed Spoken Moments in Time dataset (S-MiT) as well as Vatex-en [69], MSR-VTT [71] and ActivityNet Captions [33]). The proposed AMM loss function consistently achieves the best results across each dataset in Table 2 and the BART language model provides the strongest representations for the retrieval task in Table 1.

Table 2 shows a comparison of our AMM approach to other methods for cross-modal contrastive learning. We use the BART language model [37] to generate representations of words transcribed from the audio captions via a pretrained ASR model. Replacing the monotonically increasing margin used in MMS [29] with an adaptive margin that scales with the samples in a batch achieves the strongest results. We observed that as training continues and the margin in MMS continues to grow the training performance begins to degrade. This is likely due to the margin becoming too large for stable training as described in prior work [68].

In Table 3, we show a comparison of different spoken caption models with different loss functions. The proposed AMM approach beats the other loss functions consistently.

5.2. Cross Dataset Evaluation

To further examine the strength of our proposed Spoken Moments in Time (S-MiT) dataset, we compare the generalization performance of models trained on four different datasets (S-MiT as well as Vatex-en [69], MSR-VTT [71] and ActivityNet Captions [33]) for video/caption retrieval.

7. Acknowledgment

This work was supported by the MIT-IBM Watson AI Lab as well as the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior/Interior Business Center (DOI/IBC) contract number D17PC00341.
References

[1] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Apostol (Paul) Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. arXiv preprint arXiv:1609.08675, 2016. 2

[2] Jean-Baptiste Alayrac, Piotr Bojanowski, Nishant Agrawal, Josef Sivic, Ivan Laptev, and Simon Lacoste-Julien. Unsupervised learning from narrated instruction videos. In IEEE Conf. Comput. Vis. Pattern Recogn., June 2016. 3

[3] Yusuf Aytar, Carl Vondrick, and Antonio Torralba. Soundnet: Learning sound representations from unlabeled video. In Adv. Neural Inform. Process. Syst., 2016. 3

[4] Angie Boggust, Kartik Audhkhasi, Dhiraj Joshi, David Harwath, Samuel Thomas, Rogerio Fersis, Dan Gutfreund, Yang Zhang, Antonio Torralba, Michael Picheny, and James Glass. Grounding spoken words in unlabeled video. In CVPR Sight and Sound Workshop, 2019. 2

[5] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. Transactions of the Association for Computational Linguistics, 5:135–146, 2017. 6, 7

[6] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In IEEE Conf. Comput. Vis. Pattern Recogn., 2015. 2

[7] Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A short note about kinetics-600. arXiv preprint arXiv:1808.01340, 2018. 3

[8] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In Int. Conf. Comput. Vis., 2017. 2

[9] David Chen and William B Dolan. Collecting highly parallel data for paraphrase evaluation. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pages 190–200, 2011. 3, 5

[10] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. arXiv preprint arXiv:2002.05709, 2020. 3, 4

[11] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. arXiv preprint arXiv:1504.00325, 2015. 2, 3

[12] 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. CoRR, abs/2006.13256, 2020. 3, 5

[13] 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 Eur. Conf. Comput. Vis., 2018. 3, 5

[14] Pradipto Das, Chenliang Xu, Richard F Doell, and Jason J Corso. A thousand frames in just a few words: Linguistic description of videos through latent topics and sparse object stitching. In IEEE Conf. Comput. Vis. Pattern Recogn., pages 2634–2641, 2013. 3

[15] Yann N Dauphin, Angela Fan, Michael Auli, and David Grangier. Language modeling with gated convolutional networks. In International conference on machine learning, pages 933–941, 2017. 5

[16] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In IEEE Conf. Comput. Vis. Pattern Recogn., 2009. 2, 5

[17] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In North American Chapter of the Association for Computational Linguistics, pages 4171–4186, 2019. 4, 6, 7

[18] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In Int. Conf. Comput. Vis., pages 1422–1430, 2015. 3

[19] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. VSE++: Improving visual-semantic embeddings with hard negatives. arXiv preprint arXiv:1707.05612, 2017. 4

[20] Spandana Gella, Mike Lewis, and Marcus Rohrbach. A dataset for telling the stories of social media videos. In Empirical Methods in Natural Language Processing, pages 968–974, Oct.-Nov. 2018. 2, 3, 5

[21] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. The “something something” video database for learning and evaluating visual common sense. In Int. Conf. Comput. Vis., Oct 2017. 1, 2

[22] Chunhui Gu, Chen Sun, David A. Ross, Carl Vondrick, Caroline Pantofaru, Yeqing Li, Sudheendra Vijayanarasimhan, George Toderici, Susanna Ricco, Rahul Sukthankar, Cordelia Schmid, and Jitendra Malik. Ava: A video dataset of spatio-temporally localized atomic visual actions. In IEEE Conf. Comput. Vis. Pattern Recogn., June 2018. 1

[23] Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, pages 297–304. JMLR Workshop and Conference Proceedings, 2010. 4

[24] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In IEEE Conf. Comput. Vis. Pattern Recogn., volume 2, pages 1735–1742, 2006. 3

[25] David Harwath, Adria Recasens, Didac Suris, Galen Chuang, Antonio Torralba, and James Glass. Jointly discovering visual objects and spoken words from raw sensory input. Int. J. Comput. Vis., (128):620–641, 2020. 2, 3, 4, 5, 6, 7, 12

[26] David Harwath, Antonio Torralba, and James Glass. Unsupervised learning of spoken language with visual context. In Adv. Neural Inform. Process. Syst., 2016. 2, 3, 12
[27] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conf. Comput. Vis. Pattern Recogn.*, pages 770–778, 2016. 6

[28] Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899, 2013. 2, 3

[29] Gabriel Ilharco, Yuan Zhang, and Jason Baldridge. Large-scale representation learning from visually grounded untranscribed speech. In *Conference on Computational Natural Language Learning*, pages 55–65, Nov. 2019. 2, 3, 4, 8

[30] Aren Jansen, Manoj Plakal, Ratheesh Pandya, Daniel PW Ellis, Shawna Hershley, Jiayang Liu, R Channing Moore, and Rif A Saurous. Unsupervised learning of semantic audio representations. In *IEEE international conference on acoustics, speech and signal processing*, pages 126–130, 2018. 4

[31] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 1, 2, 3, 5, 12

[32] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 12

[33] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Int. Conf. Comput. Vis.*, 2017. 2, 3, 5, 7, 8, 12, 14

[34] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vis.*, 123(1):32–73, 2017. 2, 3

[35] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Adv. Neural Inform. Process. Syst.*, 25:1097–1105, 2012. 1, 6, 12

[36] Hilde Kuehne, Hueihan Jhuang, Rainer Stiefelhagen, and Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Moments in time dataset: one million videos for event understanding. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(2):502–508, 2019. 1, 2, 3, 5, 6

[37] Mathew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfreund, Carl Vondrick, and Aude Oliva. Moments in time dataset: one million videos for event understanding. *IEEE Trans. Pattern Anal. Mach. Intell.*, 42(2):502–508, 2019. 1, 2, 3, 5, 6

[38] Mathew Monfort, Kandan Ramakrishnan, Alex Andonian, Barry A McNamara, Alex Lascelles, Bowen Pan, Quanfu Fan, Dan Gutfreund, Rogerio Feris, and Aude Oliva. Multi-moments in time: Learning and interpreting models for multi-action video understanding. *arXiv preprint arXiv:1911.00232*, 2019. 1, 2, 3, 5, 6, 12

[39] Andrew Owens, Jiajun Wu, Josh McDermott, William Freeborn, and Antonio Torralba. Ambient sound provides supervision for visual learning. In *Eur. Conf. Comput. Vis.*, 2016. 3

[40] Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Int. Conf. Comput. Vis.*, December 2015. 2, 3

[41] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In *Eur. Conf. Comput. Vis.*, pages 740–755, 2014. 5

[42] Hao Liu, Xiangyu Zhu, Zhen Lei, and Stan Z. Li. Adaptive-face: Adaptive margin and sampling for face recognition. In *IEEE Conf. Comput. Vis. Pattern Recogn.*, June 2019. 6

[43] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Adv. Neural Inform. Process. Syst.*, pages 13–23, 2019. 3

[44] Danny Merckx, Stefan L. Frank, and Mirjam Ernestus. Language learning using speech to image retrieval. In *International Speech Communication Association*, September 2019. 2

[45] Antoine Miche, 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 *Int. Conf. Comput. Vis.*, pages 2630–2640, 2019. 3, 5, 6

[46] Matthew Monfort, Alex Andonian, Bolei Zhou, Kandan Ramakrishnan, Sarah Adel Bargal, Tom Yan, Lisa Brown, Quanfu Fan, Dan Gutfreund, Carl Vondrick, and Aude Oliva. Multi-moments in time: Learning and interpreting models for multi-action video understanding. *arXiv preprint arXiv:1911.00232*, 2019. 1, 2, 3, 5, 6, 12

[47] Andrew Owens, Jiajun Wu, Josh McDermott, William Freeborn, and Antonio Torralba. Ambient sound provides supervision for visual learning. In *Eur. Conf. Comput. Vis.*, 2016. 3

[48] Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Int. Conf. Comput. Vis.*, December 2015. 2, 3

[49] Esteban Real, Jonathon Shlens, Stefano Mazzoccchi, Xin Pan, and Vincent Vanhoucke. Youtube-boundingboxes: A large high-precision human-annotated data set for object detection in video. In *IEEE Conf. Comput. Vis. Pattern Recogn.*, pages 5296–5305, 2017. 1

[50] Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics*, 1:25–36, 2013. 3, 5

[51] Anna Rohrbach, Marcus Rohrbach, Wei Qiu, Annemarie Rif A Saurous. Unsupervised learning of semantic audio representations. In *IEEE Conf. Comput. Vis. Pattern Recogn.*, pages 55–65, Nov. 2019. 2, 3, 4, 8

[52] Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics*, 1:25–36, 2013. 3, 5

[53] Mathew Monfort, Alex Andonian, Barry A McNamara, Alex Lascelles, Bowen Pan, Quanfu Fan, Dan Gutfreund, Rogerio Feris, and Aude Oliva. Multi-moments in time: Learning and interpreting models for multi-action video understanding. *arXiv preprint arXiv:1911.00232*, 2019. 1, 2, 3, 5, 6, 12

[54] Andrew Owens, Jiajun Wu, Josh McDermott, William Freeborn, and Antonio Torralba. Ambient sound provides supervision for visual learning. In *Eur. Conf. Comput. Vis.*, 2016. 3

[55] Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Int. Conf. Comput. Vis.*, December 2015. 2, 3

[56] Esteban Real, Jonathon Shlens, Stefano Mazzoccchi, Xin Pan, and Vincent Vanhoucke. Youtube-boundingboxes: A large high-precision human-annotated data set for object detection in video. In *IEEE Conf. Comput. Vis. Pattern Recogn.*, pages 5296–5305, 2017. 1

[57] Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. Grounding action descriptions in videos. *Transactions of the Association for Computational Linguistics*, 1:25–36, 2013. 3, 5

[58] Anna Rohrbach, Marcus Rohrbach, Wei Qiu, Annemarie Friedrich, Manfred Pinkal, and Bernt Schiele. Coherent multi-sentence video description with variable level of detail. In *German conference on pattern recognition*, pages 184–195, 2014. 3
A. Annotation

We follow the approach used to collect the Places Audio Caption dataset [26, 25] and collect audio descriptions of each video in the dataset using Amazon Mechanical Turk (AMT). In order to ensure that we have a large and diverse dataset, we collect an audio description using AMT for each video in a set of 500k randomly selected videos from the training set and at least two unique descriptions for each video in the 10k videos used for both the validation and test sets. Each AMT worker is presented with a task of recording themselves describing 10 different videos. Each video is shown on the left of the screen while a video with an example text description is shown on the right. This example helps to show the workers the types of descriptions we are looking for and the amount of detail we expect from them. This example stays on the right side of the screen throughout the task while the target videos on the left cycle as the worker completes each description. Figure 4 shows an example of this interface with an example video and caption on the right and a target video on the left. Below each target description is a button that allows the worker to start recording their voice as they describe the video. Once they press this button, the video is removed from the screen and the recording is started. We block the worker from seeing the video while recording the description to ensure that the recordings are concise and pertain only to the important events highlighted in their memory. We use the Google Cloud ASR engine to verify the quality of each recorded description and flag AMT workers for poor performance. This is done by checking that the generated text has at least five words, is unique (some bots repeat pre-recorded audio to trick the system) and that the audio is at least three seconds long. If any of these checks fail we don’t let the worker continue to the next video until they record a new description that passes our checks. Once the descriptions are recorded, we periodically sample videos to check the quality of the audio paired with the ASR to ensure they match the videos and have an appropriate level of detail. If these checks fail, we flag the workers that recorded the descriptions, don’t allow them to record in the future and recheck all of their recorded data. This process allows us to ensure a strong level of quality in our collected spoken captions. Examples of some of the videos and corresponding text transcriptions of the descriptions we collected can be seen in Figure 1.

B. Implementation Details

We train each model on a server with 8 24GB Titan RTX cards using a mini-batch size of 2048 for 100 epochs. We examine the effect of the mini-batch size on learning in the next section. We take the best parameters as evaluated on the evaluation set of the training dataset after each epoch. We repeat this process for two phases of training. First we freeze the visual backbone models and train only the projection heads (including the full caption model for the spoken models) and then, in a second round, allow the full visual model to train as well. We keep the language and ASR components frozen for the language caption models and reserve fine-tuning these components for future work. For model training, we use an Adam [32] optimizer where a fixed learning rate of 0.001 and 0.00001 are set for the first and the second round model training, respectively.

C. Ablation Studies

In Tables 5, 6, 7, 8, and 9, we show several ablation studies. Unless otherwise listed in the table we use the proposed AMM loss function with the BART [37] language model as part of the language caption model described in Section 4.2.1 for each experiment. Results are averaged over five rounds with a single random batch of 1k caption-video pairs from the test set. Due to the increased computation demand of these studies we freeze the base models and train the projection heads for alignment. We use the best model settings found in this analysis to train the full models with results reported in Section 5.

Table 5 shows the effect of using two different pretrained temporal shift [40] video models on four different datasets in order to choose the most appropriate base models (Multi-Moments in Time (M-MiT) [47] or Kinetics [31]). Here we use the BART language model and the proposed AMM loss function as described in Section 4 as this combination gave us the best results on each dataset.

Table 6 compares the effect of the video model (TSM) trained for action recognition and the 2D model trained for object recognition. Most captions reference both objects and actions in a video with an average of 4.37 nouns used per caption compared to 1.58 verbs. The strength of the 2D object model makes sense when we consider this prevalence of nouns in the captions. The combination of the TSM model trained on M-MiT [47] and the 2D models trained on ImageNet [35] provided the best performance when used with the model described in Section 4.

In Tables 7 and 8 we compare the the effect of the batch size and projection size on the performance of the S-MiT model described in Section 4 in order to validate our choice of a 2048 batch and a 4096 projection. Similarly, Table 9 shows the effect of using the caption sampling approach for the transcription model as described in Section 4.2.1. In Table 10, we explore different dampening parameters.

D. Cross Dataset Generalization

In Table 11, we expand on Table 4 and compare the generalization performance of models trained on four different datasets (S-MiT as well as Vatex-en [69], MSR-VTT [71] and ActivityNet Captions [33]) for video/caption retrieval on their full test sets. In Table 4 we ran the comparison on
five samples of 1k video-caption pairs to be consistent on evaluating across different size test sets. Here we evaluate on the full test set of each dataset to provide a baseline for each test set. The strength of the model trained on S-MiT is even more evident here as it achieves higher results on the test sets of both ActivityNet and MSR-VTT than the models trained on those datasets. It even comes very close to the performance of the Vatex model on the Vatex test set. This shows that the scale and diversity of the S-MiT dataset is highly beneficial to training robust models.

E. Qualitative Results

In Tables 12 and 13, we show the top five retrieval results for some examples from the Spoken Moments dataset. For this analysis, we use the language caption model described in Section 4.2.1 with the BART [37] language model and the proposed AMM loss function. Table 12 shows the top five retrieved captions given a query video, while Table 13 shows the top five retrieved videos given a query caption. Blue boxes indicate the ground-truth results.

Our model retrieves results by recognizing key objects or environments in the videos. For example, in Table 12 (c), *lettuce* is distinguished from the other vegetables. In Table 13 (f), the model not only recognizes the planets in space but also understands that they are crashing into each other. Some of the examples show that the top retrieval result is not the ground-truth. However, as we can see, the top predictions are typically still a strong match for the queries, as in (e), (i) in Table 12 and (a), (b) in Table 13.

For this demonstration, we use transcribed words from the audio captions using a pretrained ASR model. Noise in these transcriptions may contribute to some errors. In the future, we plan to investigate jointly training a pre-trained ASR model, and language model, with the video model to improve our performance.

F. Captions in the Spoken Moments Dataset

Table 14 shows some captions in the Spoken Moments dataset that capture motion and sequential events which would be difficult to represent with a single image.
Table 5: Comparison of different pretrained TSM models on multiple datasets using AMM and Bart

| Dataset          | Pretrained TSM Dataset | R@1 | R@5 | R@10 | Mean |
|------------------|------------------------|-----|-----|------|------|
| Vatex [49]       | Kinetics (S-MIT)       | 28.1 ± 1.2 | 50.5 ± 1.4 | 61.8 ± 1.3 | 31.8 ± 1.2 |
|                  | MSR-VTT (M-MIT)       | 40.9 ± 1.6 | 73.2 ± 1.8 | 78.5 ± 1.1 | 45.3 ± 0.8 |

Table 6: Comparison of different visual base model combinations on S-MIT using AMM and Bart

| Batch Size | R@1 | R@5 | R@10 | mAP |
|------------|-----|-----|------|-----|
| 512        | 27.2 ± 1.6 | 57.4 ± 1.3 | 69.4 ± 1.0 | 41.0 ± 1.5 |
| 1024       | 27.8 ± 2.0 | 57.1 ± 1.4 | 68.9 ± 1.2 | 41.5 ± 1.9 |
| 2048       | 29.8 ± 2.5 | 60.6 ± 2.4 | 72.2 ± 2.9 | 44.0 ± 2.2 |
| 4096       | 29.2 ± 2.7 | 58.4 ± 1.6 | 70.8 ± 1.9 | 42.8 ± 2.3 |

Table 7: Comparison of different batch sizes on S-MIT using AMM and Bart

| Projection Size | R@1 | R@5 | R@10 | mAP |
|-----------------|-----|-----|------|-----|
| 1024            | 27.4 ± 1.8 | 56.6 ± 1.9 | 69.5 ± 1.9 | 41.1 ± 1.4 |
| 2048            | 27.8 ± 2.1 | 57.4 ± 2.0 | 69.2 ± 2.1 | 41.5 ± 1.8 |
| 4096            | 29.8 ± 2.5 | 60.6 ± 2.4 | 72.2 ± 2.9 | 44.0 ± 2.2 |
| 8192            | 29.4 ± 2.4 | 58.0 ± 2.3 | 70.3 ± 1.2 | 42.6 ± 2.1 |

Table 8: Comparison of different projection sizes on S-MIT using AMM and Bart

| Sampling        | R@1 | R@5 | R@10 | mAP |
|-----------------|-----|-----|------|-----|
| N               | 29.8 ± 2.5 | 60.6 ± 2.4 | 72.2 ± 2.9 | 44.0 ± 2.2 |
| Y               | 31.5 ± 2.5 | 65.5 ± 2.7 | 72.7 ± 3.1 | 45.5 ± 2.2 |

Table 9: Comparison of sampling approach on S-MIT using AMM and Bart

| α              | R@1 | R@5 | R@10 | mAP |
|----------------|-----|-----|------|-----|
| 0.1            | 29.3 ± 1.4 | 60.0 ± 1.2 | 72.7 ± 1.4 | 43.4 ± 1.2 |
| 0.2            | 28.4 ± 1.2 | 58.1 ± 1.9 | 70.9 ± 1.5 | 42.3 ± 1.4 |
| 0.3            | 27.1 ± 2.3 | 58.9 ± 2.9 | 71.5 ± 2.2 | 41.6 ± 2.3 |
| 0.4            | 28.1 ± 1.1 | 58.1 ± 2.2 | 69.8 ± 2.2 | 41.9 ± 2.2 |
| 0.5            | 29.8 ± 2.5 | 60.6 ± 2.4 | 72.2 ± 2.9 | 44.0 ± 2.2 |
| 0.6            | 28.6 ± 2.1 | 59.3 ± 2.1 | 71.3 ± 2.1 | 42.2 ± 2.2 |
| 0.7            | 28.9 ± 1.5 | 59.2 ± 1.9 | 70.8 ± 1.3 | 42.8 ± 1.4 |
| 0.8            | 29.0 ± 1.9 | 59.2 ± 2.4 | 70.7 ± 1.4 | 42.8 ± 1.8 |
| 0.9            | 29.7 ± 1.2 | 57.0 ± 2.1 | 68.2 ± 2.0 | 41.1 ± 2.2 |

Table 10: Comparison of different damping multipliers, α, in AMM using S-MIT

| Trained On | Vatex (ActivityNet) | R@1 | R@5 | R@10 | mAP |
|------------|---------------------|-----|-----|------|-----|
| 19.8       | 48.4 ± 3.7          | 35.4 |     |      |     |
| 21.1       | 33.3 ± 3.3          | 46.8 | 23.0 | 7.3 | 12.0 |
| 21.4       | 44.6 ± 3.7          | 33.7 |     |      |     |
| 19.4       | 44.6 ± 3.7          | 33.7 |     |      |     |

Table 11: Cross Dataset Evaluation on Video/Caption Retrieval on Full Test Set

| Trained On | Vatex (ActivityNet) | R@1 | R@5 | R@10 | mAP |
|------------|---------------------|-----|-----|------|-----|
| 19.8       | 48.4 ± 3.7          | 35.4 |     |      |     |
| 21.1       | 33.3 ± 3.3          | 46.8 | 23.0 | 7.3 | 12.0 |
| 21.4       | 44.6 ± 3.7          | 33.7 |     |      |     |
| 19.4       | 44.6 ± 3.7          | 33.7 |     |      |     |
| Query                                                                 | R@1                                                                 | R@2                                                                 | R@3                                                                 | R@4                                                                 | R@5                                                                 |
|----------------------------------------------------------------------|----------------------------------------------------------------------|----------------------------------------------------------------------|----------------------------------------------------------------------|----------------------------------------------------------------------|----------------------------------------------------------------------|
| it is raining outside in between some houses there is a small stream of water running down in between them | ![Image](image1)                                                   | ![Image](image2)                                                   | ![Image](image3)                                                   | ![Image](image4)                                                   | ![Image](image5)                                                   |
| four people in a yellow raft white water rafting on the Gorge         | ![Image](image6)                                                   | ![Image](image7)                                                   | ![Image](image8)                                                   | ![Image](image9)                                                   | ![Image](image10)                                                  |
| someone uses a knife to cut lettuce on a cutting board with their hands | ![Image](image11)                                                  | ![Image](image12)                                                  | ![Image](image13)                                                  | ![Image](image14)                                                  | ![Image](image15)                                                  |
| a young boy with the blue shirt rides artificial waves with the surf  | ![Image](image16)                                                  | ![Image](image17)                                                  | ![Image](image18)                                                  | ![Image](image19)                                                  | ![Image](image20)                                                  |
| a man is standing on a stage in front of an audience telling jokes    | ![Image](image21)                                                  | ![Image](image22)                                                  | ![Image](image23)                                                  | ![Image](image24)                                                  | ![Image](image25)                                                  |
| a video showing several people seated in the audience all men all nicely dressed applauding | ![Image](image26)                                                  | ![Image](image27)                                                  | ![Image](image28)                                                  | ![Image](image29)                                                  | ![Image](image30)                                                  |
| a man plows the ground with his hands sitting in the hot sun with another man sitting in the background appeared to rest | ![Image](image31)                                                  | ![Image](image32)                                                  | ![Image](image33)                                                  | ![Image](image34)                                                  | ![Image](image35)                                                  |
| children are standing in the rain and holding their umbrellas upside down | ![Image](image36)                                                  | ![Image](image37)                                                  | ![Image](image38)                                                  | ![Image](image39)                                                  | ![Image](image40)                                                  |
| a very old picture of boxers boxing in a boxing competition            | ![Image](image41)                                                  | ![Image](image42)                                                  | ![Image](image43)                                                  | ![Image](image44)                                                  | ![Image](image45)                                                  |

Table 12: Spoken Moments Examples of Caption to Video Retrieval Results: Given a query caption, we show five top retrieved captions where words transcribed from the audio captions using a pretrained ASR model are used as a caption. We use a BART model trained with the AMM loss function on the S-MiT dataset. Blue indicates the ground-truth results.
Table 13: Spoken Moments Examples of Video to Caption Retrieval Results: Given a query video, we show five top retrieval captions where words transcribed from the audio captions using a pretrained ASR model are used as a caption. We use a BART model trained with the AMM loss function on the S-MiT dataset. Blue indicates the ground-truth results.
Table 14: **Spoken Moments Captions:** We show some examples of captions, and associated video frames, from the Spoken Moments dataset, where the captions describe a sequence of actions or motion.

| Caption                                                                 | Frames |
|------------------------------------------------------------------------|--------|
| (a) a boy and a red white and blue shirt is sitting on a couch he is holding an infant life vest and picks it up to blow through the two |        |
| (b) there’s a gauge or a lock thing turns from rides and then being turned to the left |        |
| (c) a picture of a man drinking coffee and play with a cell phone in fast motion |        |
| (d) in slow motion we see a collie jump into the air and catch a white frisbee in flight |        |
| (e) these are track and field runners and it’s a relay race and they take off when they are handed the batons |        |
| (f) there is water dripping off the edge of something all you can hear is the water dripping |        |