Learning Audio-Video Modalities from Image Captions

Arsha Nagrani, Paul Hongsuck Seo, Bryan Seybold, Anja Hauth, Santiago Manen, Chen Sun, and Cordelia Schmid

Google Research
https://a-nagrani.github.io/videocc.html

Abstract. There has been a recent explosion of large-scale image-text datasets, as images with alt-text captions can be easily obtained online. Obtaining large-scale, high quality data for video in the form of text-video and text-audio pairs however, is more challenging. To close this gap we propose a new video mining pipeline which involves transferring captions from image captioning datasets to video clips with no additional manual effort. Using this pipeline, we create a new large-scale, weakly labelled audio-video captioning dataset consisting of millions of paired clips and captions. We show that training a multimodal transformer based model on this data achieves competitive performance on video retrieval and video captioning, matching or even outperforming HowTo100M pretraining with 20x fewer clips. We also show that our mined clips are suitable for text-audio pretraining, and achieve state of the art results for the task of audio retrieval.

Keywords: data mining, video retrieval, captioning

1 Introduction

A key facet of human intelligence is the ability to effortlessly connect the visual and auditory world to natural language concepts. Bridging the gap between human perception (visual, auditory and tactile) and communication (via language) is hence becoming an increasingly important goal for artificial agents, enabling tasks such as text-to-visual retrieval [79,62,9], image and video captioning [77,84,44], and visual question answering [7,47]. In the image domain in particular, this has lead to an explosion of large scale image datasets with natural language descriptions, often by crawling alt text online [50,45,69,12,64]. In the video and audio domains, however, obtaining natural language descriptions is more challenging. Recent research has been either directed at modelling, for example in developing new architectures (eg. multimodal transformers [29,68,9]), or new training objectives (eg. those that can deal with misaligned [55] or overly specialised [63] inputs). Annotating videos manually with clean and diverse captions is often subjective, painstaking and expensive. This means that most video-captioning datasets (eg. MSR-VTT [82], LSMDC [66], CMD [8], ActivityNet [44] etc.) are small in size (order of magnitude 100K). Audio captioning datasets such

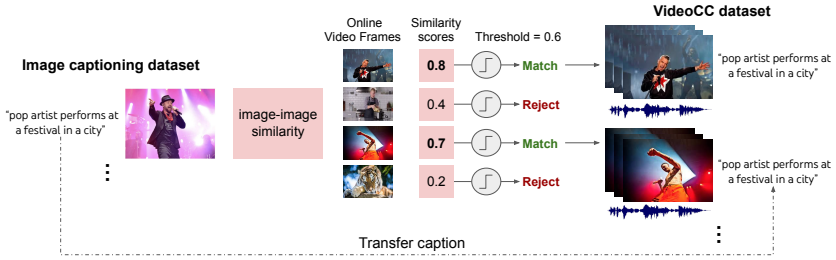

Fig. 1: **Mining audio-video clips automatically.** We use the images in image captioning datasets as 'seed' frames to mine related audio-visual clips. For each seed image-caption pair in a dataset, we find frames in videos with high similarity scores to the seed image. We then extract short video clips around the matching frames and transfer the caption to those clips. This gives us free captioning supervision for video and audio clips.

as AudioCaps [42] and Clotho [24], are even smaller. Given the well-known benefits of pretraining, many works have proposed creative but weak forms of supervision, such as hashtags [33], titles and descriptions [72], or Automatic Speech Recognition (ASR) in instructional videos [56]. The de facto standard for video-language pretraining [48,5,52,67,29,62] has become the large HowTo100M [56] dataset, pretraining on which gives a significant boost over training from scratch. The pitfalls of using ASR however are well known; (i) there is noise in imperfect ASR transcription, (ii) continuous narration may consist of incomplete or grammatically incorrect sentences, (iii) the domain is often limited to instructional videos to increase relevance between speech and video content and finally, and (iv) ASR may not be temporally aligned with the video, or indeed may not refer to the video at all [56]. Combined, this necessitates a huge amount of training data for good performance (100s of millions of samples), and consequently, a lot of compute.

Image annotation, on the other hand, is cheaper than video and easier to obtain from web pages [69,12], and large-scale image-text pretrained models such as CLIP [64] are available online. This has led to concurrent works [54,26,10] using image-text models for video-text tasks. While this is a valuable idea, using such models beyond weight initialization requires some additional complexity. If we treat videos as a bag of sparse frames [46], we lose all the benefits of video (modalities like audio and the chance to model low-level temporal information directly from the frames) or require complicated distillation procedures from image to video models [34]. Hence we believe there is still a necessity for large-scale *video*-text datasets.

Is there another way to leverage all the existing effort that has gone into image-captioning datasets? We propose a solution in the form of a new video mining method based on *cross-modal transfer*, where we use images from image captioning datasets as seeds to find similar clips in videos online (Fig. 1). We then transfer the image captions directly to these clips, obtaining weak, albeit free video and audio captioning supervision in the process. This can also provide us

with motion and audio supervision – for example, sometimes human-generated captions for images infer other modalities, eg. the caption 'Person throws a pitch during a game against university' from the CC3M dataset [69] was written for a single, still image, but is actually describing motion that would occur in a video. Similarly, the caption 'A person singing a song', is also inferring a potential audio track. We note that like HowTo100M, our dataset curation is entirely automatic, and requires no manual input at all. However, as we show in Sec. 3, our mined data samples are more diverse than HowTo100M, are matched to better-formed captions compared to ASR, and are likely to contain at least one frame that is aligned with the text caption.

In doing so we make the following contributions: (i) We propose a new, scalable video-mining pipeline which transfers captioning supervision from image datasets to video and audio. (ii) We use this pipeline to mine paired video and captions, using the Conceptual Captions3M [69] image dataset as a seed dataset. Our resulting dataset VideoCC3M consists of millions of weakly paired clips with text captions and will be released publicly. (iii) We propose a new audio-visual transformer model for the task of video retrieval, which when trained on this weakly paired data performs on par with or better than models pre-trained on HowTo100M for video retrieval and captioning, with 20x fewer clips and 100x fewer text sentences. In particular, we show a large performance boost in the zero-shot setting. (iv) Finally, we also show that our audio-visual transformer model seamlessly transfers to *text-audio* retrieval [60] benchmarks as well, achieving state of the art results on the AudioCaps [42] and Clotho [24] datasets.

2 Related work

Cross-modal supervision: Our key idea is to use labelled data in one modality (images) to aid learning in another modality (videos). A popular method for cross-modal transfer is knowledge *distillation* [37], which has shown great success for transferring supervision from RGB to depth [36], or faces to speech [4]. Another line of work enhances unimodal models via multimodal regularisations [2,3]. Ours is a related but tangential idea which involves mining new data and assigning labels to it (similar to video clips mined for action recognition using speech by [58,30]). This is particularly useful when there are large labelled datasets in one modality (here text-image retrieval [50,45,69]), but it is more challenging to obtain for a similar task in another modality (text-audio [60] or text-video [82,6,44,66,87,8,28] retrieval).

Text supervision for video: Existing manually annotated video captioning datasets [82,87,39] are orders of magnitude smaller than classification datasets [41]. This has led to a number of creative ideas for sourcing weakly paired text and video data. [74] use web images queried with sports activities to create temporal annotations for videos. WVT [72] mines videos from YouTube and their titles for action recognition starting from the Kinetics labels. Similarly [73] uses video level labels for the same task. Unlike these works where the labels are at a video level, our captions are localised to video clips, and are not limited to the domain

of action recognition only. [33] and [49] use hashtags and titles for supervision respectively, but only to learn a better video encoder. In the movie domain, [8] uses YouTube descriptions for movie clips while [66] uses audio description (AD) from movies. The recently released WebVid2M dataset [9] comprises manually annotated captions, but given the monetary incentive on stock sites, they often contain added metatags appended, and most lack audio. Another valuable recent dataset is Spoken Moments in Time [57], however this was created with significant manual effort. The largest video-text dataset by far is HowTo100M [56] generated from ASR in instructional videos; however, this data is particularly noisy, as discussed in the introduction.

Text supervision for audio: Textual supervision for audio is even scarcer than it is for video. Early works perform text-audio retrieval using single word audio tags as queries [13], or class labels as text labels [25]. Even earlier, [71] linked text to audio but only using 215 animal sounds from the BBC Sound Effects Library. Unlike these works, we study unconstrained caption-like descriptions as queries. While small, manually annotated datasets such as AudioCaps [42] and Clotho [24] do exist (and have been repurposed by [60,43] for audio-text retrieval), large-scale pretraining data for text-audio tasks is not available. Note that extracting audio from existing video-text datasets is difficult: WebVid2M [9] videos largely do not have audio, and HowTo100M captions are derived from the audio (training a model to predict HowTo100M captions from the audio might simply be learning how to do ASR). Hence we explore the link between audio and text transferred via image similarity to videos that all have audio, and show this improves text-audio retrieval. As far as we are aware, we are the first work to pre-train the same model for both *visual-focused* datasets such as MSR-VTT and *audio-focused* datasets such as AudioCaps and Clotho.

3 Text-video data

In this section we describe our automatic mining pipeline for obtaining video clips paired with captions. We then train text-video and text-audio models (described in Sec. 4) on this weakly paired data for 2 tasks, audiovideo retrieval and video captioning.

3.1 Mining pipeline

The core idea of our mining pipeline is to start with an image captioning dataset, and for each image-caption pair in a dataset, find frames in videos similar to the image. We then extract short video clips around the matching frames and transfer the caption to those clips. In detail, the steps are as follows:

1. Identify seed images: We begin by selecting an image-captioning dataset. The images in this dataset are henceforth referred to as 'seed' images (x_{seed}).

2. Feature Extraction: We then calculate a visual feature vector $f(x_{\text{seed}})$ for each seed image. Given our primary goal is to mine semantically similar images, we extract features using a deep model trained for image retrieval, the

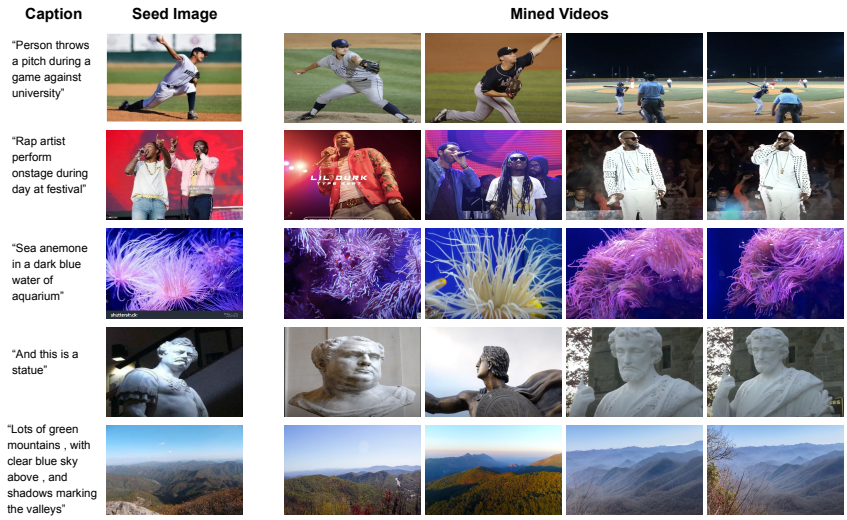

Fig. 2: **Examples of clips with captions that are mined automatically.** For each seed image, we show 3 'matched' clips obtained using our automatic video mining method. For the first 2 clips, we show only a single frame, but for the third clip we present 2 frames to show motion, either of the subjects in the video (first 3 rows) or small camera motion (last 2 rows). Note the diversity in the mined clips, for example the different pitching poses and angles (first row) and the different types of statues (fourth row). Clips in the second row also contain audio relevant to the caption. Note frames may have been cropped and resized for ease of visualisation. More qualitative results are provided in the supplementary.

Graph-Regularized Image Semantic Embedding (Graph-RISE) model [40]. We then extract the same visual features $f(x_v)$ for the frames x_v of a large corpus of videos. Because of frame redundancy, we can extract features at a reduced rate (1fps) relative to the original video frame rate.

3. Identify matches: Next, we calculate the dot product similarity between the feature vectors for each seed image in the caption data set and those for each video frame obtained from the video corpus. Pairs with a similarity above a threshold τ are deemed 'matches'. For each seed image, we keep the top 10 matches. For these top 10, we transfer the caption from the image to a short video clip extracted at a temporal span t around the matched image frame, and add it to our dataset. In Sec. 3.3, we provide brief ablations on the values of t and the threshold τ.

3.2 Video-Conceptual-Captions (VideoCC)

We ran our mining pipeline with the image captioning dataset - Conceptual Captions 3M [69] (CC3M). We only use images which are still publicly available online, which gives us 1.25 image-caption pairs. We apply our pipeline to videos online. We filter videos for viewcount > 1000, length < 20 minutes, uploaded within the last 10 years, but at least 90 days ago, and filter using content-appropriateness signals to get 150M videos. This gives us 10.3M clip-text pairs

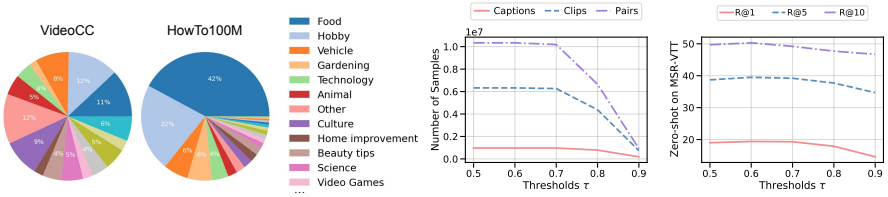

Fig. 3: **Domains in VideoCC3M vs HowTo100M (left), effect of match threshold τ on mining statistics (middle) and zero-shot performance on MSR-VTT (right).** VideoCC3M has a more diverse and balanced range of domains, 'Other' here includes a variety of content such as music videos, sports, politics, vlogs and so on. Note how almost half of HowTo100M videos are food-related (cooking videos). More details are provided in suppl. Effect of match threshold τ: Increasing the threshold τ beyond 0.6 decreases the size of the dataset, which leads to a corresponding performance drop on zero-shot retrieval. We use an optimal match threshold of 0.6.

with 6.3M video clips (total 17.5K hours of video) and 970K unique captions. We call the resulting dataset VideoCC3M. We also run our pipeline on a more recently released seed dataset extension, called Conceptual Captions 12M [12] (CC12M). Note that while CC3M consists of higher quality captions [69], CC12M was created by relaxing the data collection pipeline used in CC3M, and hence the captions are far noisier. Results on this dataset are provided in the supplementary material. Some examples of the matched video frames to captions for VideoCC3M are provided in Figure 2. A preliminary fairness analysis on the data is provided in the supplementary material. The mined video clips have the following properties:

(i) Diversity: We compare the domains in our dataset to HowTo100M in Figure 3 (left). Note that because VideoCC3M is mined from a general corpus of videos online (unlike HowTo100M, which is restricted to instructional videos), our dataset is more balanced. A more comprehensive bar chart is provided in the supplementary. Some of the 'Other' categories are technology, team sports, family, medicine, beauty, history, religion, gardening, music, politics – while HowTo100M videos are largely dominated by the 'Food' and 'Hobby' domains (almost half are 'cooking videos'). This is unsurprising given that HowTo100M is limited to instructional videos.

(ii) Alignment: We mine frames that have high visual similarity to the seed image. If this seed has a relevant caption (largely the case for the high quality CC3M dataset), it is likely that at least one frame in the mined clip is aligned with the caption. A manual check of a small subset of clips found this to be the case in 91% (see suppl). This is a stricter constraint than ASR based datasets, which have occasional misalignment between speech and frames. As an additional quantitative metric, we also run a commercial image classification system on the frames in both the VideoCC3M and the HowTo100M datasets. We then compute the proportion of captions for which a word in the caption exactly matches a label from the image classification system. We find the proportion to be 69.6% for VideoCC3M, whereas HowTo100M only has 19.7%.

(iii) Caption Style: The quality of the captions is transferred directly from the seed dataset. Most of the captions in CC3M are fully formed, grammatically correct sentences, unlike the distribution of sentences obtained from ASR. Each caption is matched to a mean of 10.6 clips, with some captions matched to more than 10 clips. This is possible because, while we limit the clip mining to 10 clips per seed image, the original CC3M dataset has multiple seed images with the same caption, eg 'an image of digital art', leading to more than 10 mined clips for these captions.[1] Having multiple pairs from the same set of captions and video clips also helps ensure that learnt video and text representations are not overly-specialised to individual samples (which can be a problem for existing datasets, as noted by [63]).

Cross-modal transfer from the image domain Interestingly, this mining method provides us with *captioning* supervision for modalities such as video and audio that are difficult to annotate. Note that we use two existing sources of image supervision, the first is the seed image captioning dataset, and the second is the image similarity model $f(\cdot)$ which we use to mine related frames. This is not the same as simply applying a text-image model (even though that is a complementary idea) to different frames in a video for text-video retrieval. For example, our method provides some valuable supervision for new clips with motion (see the last column of retrieved clips in Fig. 2, first two rows). Many image captions in CC3M describe actions/motion, eg. *human-human interactions* ('baby smiling down at dad while being thrown in the air'), *interactions with objects/body parts* ('person shaves hair on neck', 'rugby player fields a punt'), *movement in an environment* ('elderly couple walking on a deserted beach').[2] Our mining method, since it retrieves videos, can actually find examples of these described motions. We also obtain some free supervision for the audio stream (Fig. 2, second row and Fig. 4, right). These weakly labelled audio samples can be used for pretraining text-audio models, as we show in the results.

3.3 Data mining ablations

In this section, we ablate the time span t and threshold τ, using zero-shot performance on the MSR-VTT test set (protocol described in Sec. 5.3).

Time span t: We try extracting different length clip segments t between 5 and 30 seconds, and found that performance increases up until 10 seconds, but decreases after that (results and discussion in the suppl. material). Hence we extract 10 second clips for our dataset.

Match threshold τ: We experiment with different match thresholds τ for the similarity in the range $\{0.5, 0.6, 0.7, 0.8.0.9\}$ and present the effect of this on mining statistics in Figure 3 (middle) and zero-shot performance (right). The higher the match threshold, the stricter the similarity requirement on the matched

[1] Full distribution of clips per caption in VideoCC3M is provided in suppl. material.

[2] We find that interestingly, 83% of the 7.9K verbs (extracted using spacy package) in MSR-VTT (video annotated dataset), are present in CC3M.

frames to the caption. We note that upto a match threshold of 0.6, performance increases slightly, and there is no steep reduction in dataset size. After 0.7 however, the number of matches falls steeply as the match threshold is increased, leading to fewer videos and clips in the dataset, and a corresponding drop in downstream performance. We hence use a match threshold of 0.6 to mine clips.

4 Method

We focus on two different tasks in this paper that rely on video and text annotation - video retrieval and video captioning. We implement state of the art multimodal transformer models for each – architectures and training objectives are defined in the next two sections.

4.1 Audiovisual Video Retrieval (AVR)

For retrieval, we use a dual-stream model (one stream being an audio-video encoder and one stream being a text encoder for the caption), which when trained with a contrastive loss allows for efficient text-video retrieval. Note that the efficient dual stream approach has also been used by MIL-NCE [55] and FIT [9], but unlike these works, our video encoder is multimodal (Fig. 4, left), and utilises the audio as well. Our model is flexible, and can be used for audio-only, video-only and audio-visual retrieval.

Multimodal Video Encoder: Our encoder is inspired by the recently proposed MBT [59], which operates on RGB frames extracted at a fixed sampling rate from each video, and log-mel spectrograms used to represent audio. We first extract N non-overlapping patches from the RGB image (or the audio spectrogram), similar to the way done by ViT [23] and AST [35] respectively. The model consists of a number of transformer layers for each modality, with separate weights for each modality and fusion done via bottleneck tokens. Unlike MBT, we use frames extracted at a larger stride (an ablation is provided in the experiments), to cover the longer videos in retrieval datasets. We implement both RGB-only, audio-only and RGB-audio fusion models.

Text encoder: The text encoder architecture is the BERT model [21]. For the final text encoding, we use the [CLS] token output of the final layer.

Joint embedding: For the final video encoding, we average the [CLS] tokens from both audio and RGB modalities. Both text and video encodings are then projected to a common dimension $D = 256$ via a single linear layer each. We then compute the dot product similarity between the two projected embeddings after normalisation.

Loss: We use the NCE loss [85] to learn a video and text embedding space, where matching text-video pairs in the batch are treated as positives, and all other pairwise combinations in the batch are treated as negatives. We minimise the sum of two losses, video-to-text and text-to-video [9]. At test time, inspired by FILIP [83], we sample K clips equally spaced from the video, compare each one to the text embedding, and average the similarity scores.

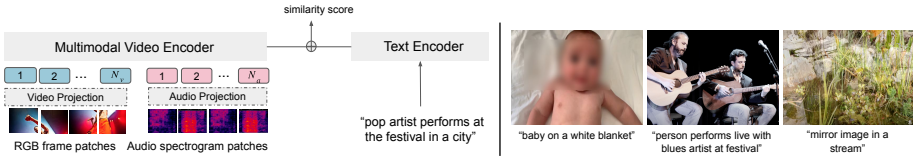

Fig. 4: **(Left) Our audiovisual dual stream retrieval model (AVR)**, which works for both image and audio focused retrieval datasets. **(Right) Examples from VideoCC3M of automatically mined clips with relevant audio to the caption.** We show a single relevant frame from each clip as a proxy for visualising the audio. The accompanying audio contains (left to right) the sounds of a baby gurgling, music and water flowing sounds (left image intentionally blurred).

4.2 Video Captioning

For video captioning, we use an encoder-decoder style generative model. Our video encoder is the same as the one used above for retrieval.

Decoder: To generate a text caption, we adapt the autoregressive GPT-2 (117M) decoder [65], however we condition each predicted text token on video features from the video encoder as well as previously generated text tokens. More formally, given video features C as context, to generate the next token y_i in our caption Y, we first encode the previous generated tokens $Y_i = \{y_0, \ldots, y_{i-1}\}$ with a look-up table and a positional embedding to produce $H_i = \{h_0, \ldots, h_{i-1}\}$. We then encode the context C and the previous embedded tokens H_i using a single transformer. The outputs of this transformer are $\tilde{C} \cup \tilde{H}_i$, where $\tilde{H}_i = \{\tilde{h}_0, \ldots, \tilde{h}_{i-1}\}$. We then predict the next token y_i from $\tilde{h}_{i-1}$ using a linear projection with a softmax: $y_i = \text{argmax}(\text{softmax}(\Phi \tilde{h}_{i-1}))$ where $\Phi \in \mathbb{R}^{\nu \times d}$ is the linear projection matrix and ν is the vocabulary size. As is standard, the first word h_0 is set using a special `BOS` (beginning of sentence) token, and tokens are generated until a special `EOS` (end of sentence) token is generated.

Loss: We minimise the negative log-likelihood of generating the ground-truth caption [17].

5 Experiments

We evaluate our text-video models on the following two tasks - (i) text-video retrieval (Sec. 5.3), which includes video retrieval on primarily *visual focused* datasets, as well as text-audio retrieval, where captions are primarily focused on *audio sounds*; and (ii) video captioning (Sec. 5.4). We use the common protocol of pretraining our models on a large dataset first, either VideoCC3M or HowTo100M, and then finetune on the target downstream dataset. Note that unlike other works, we apply the same pretrained models for both visual-focused datasets such as MSR-VTT and audio-focused datasets such as AudioCaps and Clotho. We also investigate zero-shot performance, where we apply pretrained models directly to the target task, without any finetuning at all. In this case, no supervised video-text data is used at all. We first describe datasets and metrics, then the implementation details, before finally discussing the results for each task.

5.1 Datasets and Metrics

VideoCC3M: We use the VideoCC3M dataset created using our automatic mining method described in Sec. 3.
HowTo100M [56]: consists of 1.2M instructional videos. Weak captions are in the form of transcribed speech, which we obtain using the YouTube ASR API [1].
MSR-VTT [82] contains 10K videos with 200K descriptions. For retrieval, we follow other works [51], and train on 9K train+val videos, reporting results on the 1K-A test set. For captioning, we use the standard splits proposed in [82].
AudioCaps [42] contains video clips from the AudioSet dataset [32] with captions for the task of audio captioning. This dataset was then repurposed by [60] for text-audio retrieval, by taking a subset that does not overlap with the VGGSound [16] dataset. After filtering out the videos no longer available on the web, we have 47,107 training, 403 val and 778 test samples.
Clotho [24] is an audio-only dataset of described sounds from Freesound [27]. During labelling, annotators only had access to audio (no meta tags or visual information). The data consists of a dev set and eval set of 2893 and 1045 audio samples respectively. Every audio sample is accompanied by 5 captions. We follow [60] and treat each of the 5 captions per test audio as a separate query.
Metrics As is standard for retrieval, we report recall@K, $K \in \{1, 5, 10\}$. For captioning, we use the established metrics Bleu-4 (B-4) [61], CIDEr (C) [76], and Meteor (M) [11].

5.2 Implementation details

In this section we describe implementation details for our models as well as certain design choices for sampling and initalisation. More details are provided in the supplementary material.
Audio-visual encoder: We use the ViT-Base (ViT-B, $L = 12$, $N_H = 12$, $d = 3072$), as a backbone with $B = 4$ fusion tokens and fusion layer $l_f = 8$. We sample 32 RGB frames for MSR-VTT, and 8 RGB frames for AudioCaps. For audio we extract spectrograms of size 800×128 spanning 24 seconds.
Text encoder: We use the BERT-Base architecture ($L = 12, N_H = 12, d = 768$) with uncased wordpiece tokenization [22]. We use a total number of 32 tokens per caption during training – cropping and padding for sentences longer and shorter respectively. No text augmentation is applied.
Clip coverage: A single segment per clip is randomly sampled at training time. We experiment with the length of this segment, controlled by the stride of the frames (32 frames at a stride of 2 frames at 25fps indicates an effective segment length of 2.5 seconds). We experiment with stride = 2, 6, 10, 14, 18, and find optimal performance with stride = 14 frames (effective coverage of 18s). At test time, we sample $K = 4$ clips equally spaced from the video, compare them to the text embedding, and average the similarity scores. More details are provided in the supplementary material.
Video encoder initialisation: Unless otherwise specified, we use Kinetics-400 [41] initialisation for both video retrieval and captioning. For audio retrieval we initialise the model with VGGSound [16] (see supplementary).

Init.	Modality	R@1	R@5	R@10
Scratch	V	9.4	22.5	31.7
ImNet21K [20]	V	30.2	59.7	71.3
K400 [41]	V	30.2	60.7	71.1
ImNet21k [20]	V+A	32.2	62.7	74.4
K400 [41]	V+A	**32.3**	**64.1**	**74.6**

PT Data	Modality	#Caps	R@1	R@5	R@10	R@1	R@5	R@10
			Finetuned			*Zero-shot*		
-	V	-	30.2	60.7	71.1	-	-	-
HowTo100M [56]	V	130M	33.1	62.3	72.3	8.6	16.9	25.8
VideoCC3M	V	970K	35.0	63.1	75.1	18.9	37.5	47.1
VideoCC3M	A+V	**970K**	**35.8**	**65.1**	**76.9**	**20.4**	**39.5**	**50.3**

Table 1: **Ablations with different initializations of the video encoder and the modalities (left) and effect of pretraining data (right) for text-video retrieval on the MSR-VTT dataset. Init.** Initialisation of *video encoder only*. Modalities are **V:** RGB, **A:** Audio spectrograms. **#Caps:** Number of unique captions. (left) No VideoCC data is used in the left and we do not show audio-only results as some videos in the MSR-VTT dataset are missing audio. (right) Training on VideoCC3M provides much better performance than Howto100M, with a fraction of the dataset size (VideoCC3M has only 970K captions and 6.3M clips compared to the 130M clips in HowTo100M). The performance boost is particularly large for the zero-shot setting.

Training for retrieval: The temperature hyperparameter σ for the NCE loss is set to 0.05, and the dimension of the common text-video projection space is set to 256. All models are trained with batch size 256, synchronous SGD with momentum 0.9, and a cosine learning rate schedule with warmup of 1.5 epochs on TPU accelerators. We pretrain for 4 epochs, and finetune for 5 epochs.
Training for captioning: We use the Adam optimizer with initial learning rate $1E-4$ and weight decay 0.01. For all models, we pretrain for 120K iterations with a batch size of 512. For finetuning, we train for 1K iterations.

5.3 Text-audiovisual Retrieval

Video encoder initialisation: We first experiment with initalising the video encoder *only* (Table 1, left), and find that while ImageNet initalisation provides a significant boost over training from scratch, using Kinetics-400 (K400) only provides a very marginal further gain. This suggests that at least for retrieval, the initialisation of the video encoder is not as important as joint text-video pretraining for the entire model (as demonstrated next).
Effect of pretraining data: We begin by analysing the results with fine-tuning for text-video retrieval on the MSR-VTT dataset, presented in Table 1(right). We note that pretraining on VideoCC3M provides a significant boost to performance over HowTo100M, with far less data, and for an RGB-only model, yields a 5% improvement over training from scratch on R@1. This effect is even more profound in the zero-shot case, where for an RGB-only model, using VideoCC3M more than doubles the R@1 performance compared to HowTo100M pretraining. This is done with 100x fewer captions and 20x less video data. We believe that this shows the value in high-quality video-captioning pairs. Regarding audio inputs, we note that MSR-VTT is a visual benchmark (unlike AudioCaps and Clotho), with some videos missing an audio track entirely. However we show that adding audio provides a modest performance boost. We then compare to previous works on this dataset in Table 2 (left), including recently released Frozen In Time (FIT) [9] and VideoCLIP [81]. We note that our model outperforms FIT which pretrains on 3 different datasets - CC3M, WebVid2M and COCO [18]. We

Method	V-T PT	#Caps	R@1	R@5	R@10
Finetuned					
HERO [48]	HT100M	136M	16.8	43.4	57.7
NoiseEst. [5]	HT100M	136M	17.4	41.6	53.6
CE [51]†	-		20.9	48.8	62.4
UniVL [52]	HT100M	136M	21.2	49.6	63.1
ClipBERT [46]	Coco, VGen	5.6M	22.0	46.8	59.9
AVLnet [67]	HT100M	136M	27.1	55.6	66.6
MMT [29]†	HT100M	136M	26.6	57.1	69.6
T2VLAD [80]†	-		29.5	59.0	70.1
SupportSet [62]	HT100M	136M	30.1	58.5	69.3
VideoCLIP [81]	HT100M	136M	30.9	55.4	66.8
FIT [9]	CC3M	3M	25.5	54.5	66.1
FIT [9]	Multiple‡	6.1M	32.5	61.5	71.2
Ours	VideoCC3M	**970K**	**35.8**	**65.1**	**76.9**
Zero-shot					
MIL-NCE [56]	HT100M	136M	7.5	21.2	29.6
SupportSet [62]	HT100M	136M	8.7	23.0	31.1
EAO [70]	HT100M	136M	9.9	24.0	32.6
VideoCLIP [81]	HT100M	136M	10.4	22.2	30.0
FIT [9]	WebVid2M*	2.5M	15.4	33.6	44.1
Ours	VideoCC3M	**970K**	**20.4**	**39.5**	**50.3**

Method	V-T PT	Modality	B-4	C	M
Finetuned					
POS+CG [78]	-	V	42.00	49	28.20
POS+VCT [38]	-	V	42.30	49	29.70
SAM-SS [15]	-	V	43.80	51	28.90
ORG-TRL [86]	-	V	43.60	51	28.80
VNS-GRU [14]	-	V	45.30	53	29.90
UniVL [53]	HT100M	V+T	41.79	50	28.94
DecemBT [75]	HT100M	V	45.20	52	29.70
Ours	HT100M	V	**47.33**	55	**37.11**
Ours	VCC3M	V	45.47	**55**	36.96
Zero-shot					
Ours	HT100M	V	7.5	0.5	8.23
Ours	VCC3M	V	**13.23**	**8.24**	**11.34**

Table 2: **Comparison to state-of-the-art results on MSR-VTT for text-to-video retrieval (left) and video captioning (right). V-T PT:** Visual-text pretraining data. **#Caps:** Number of unique captions used during pretraining. † These works use numerous experts, including Object, Motion, Face, Scene, Speech, OCR and Sound classification features. ‡ Pretrained on WebVid-2M, CC3M and COCO datasets. *Numbers obtained from the authors. Modalities: **V:** RGB frames. **T:** ASR in videos.

were unable to train on WebVid2M due to data restrictions but believe further performance gains could be achieved by training on VideoCC3M and WebVid jointly. We also note that by training on VideoCC3M, we outperform FIT trained only on the CC3M dataset by a big margin (R@1 25.5 to 35.3), even though the amount of manually annotated supervision is the same. This shows the benefit of mining extra video data using our data mining pipeline. On zero-shot performance, we outperform all previous works that pretrain on HowTo100M, and FIT [9] when it is trained only on video data (WebVid2M). We note that adding in various image datasets provides a huge boost to performance in FIT [9], and this complementary approach could be used with VideoCC. We could also use additional seed datasets such as COCO Captions [18] to mine more text-video clips, which we leave as future work.

Results using CLIP [64] Given the recent flurry of CLIP based [54,31,19,26], RGB-only works for video retrieval, in this section we show the complementarity of using CLIP [64] based models trained on the 400M pair WiT dataset such as Clip4Clip [54] finetuned on the VideoCC dataset. We reproduce Clip4Clip [54] with mean pooling in our framework (Table 3). Using CLIP (trained on 400M diverse image-caption pairs) leads to very strong zero-shot performance, however finetuning it on VideoCC *further* improves performance by over 3% R@1, showing the additional value of automatically mined *videos*. We also outperform the zero-shot SOTA from Clip4Clip which was post trained on a curated subset of HowTo100M and is the highest online number for this zero-shot benchmark (CaMoE [19] and Clip2TV [31] do not report zero-shot results). This shows the

Model	Pre-Training Data	*R*@1	*R*@5	*R*@10
Clip4Clip [54]	WiT [64]	30.6	54.4	64.3
Ours	WiT [64] + VideoCC	**33.7**	**57.9**	**67.9**

Table 3: Finetuning Clip4Clip on VideoCC for zero-shot performance on MSR-VTT.

Model	Pretraining	Modality	R@1	R@10
SOTA [60]†	-	A	24.3	72.1
Ours	-	A	32.0	82.3
Ours	HowTo100M	A	33.7	83.2
Ours	VideoCC3M	A	35.5	84.5
Ours (ZS)	HowTo100M	A	1.4	6.5
Ours (ZS)	VideoCC3M	A	8.7	37.7
SOTA [60]†	-	A+V	28.1	79.0
Ours	-	A+V	41.4	85.3
Ours	VideoCC3M	A+V	**43.2**	**88.9**
Ours (ZS)	VideoCC3M	A+V	10.6	45.2

Model	Pretraining	R@1	R@10
SOTA [60]	-	6.7	33.3
Ours	-	7.8	35.4
Ours	VideoCC3M	**8.4**	**38.6**
Ours (ZS)	VideoCC3M	3.0	17.5
SOTA [60]	AudioCaps	9.6	40.1
Ours	AudioCaps	11.4	43.4
Ours	VideoCC3M+AudioCaps	**12.6**	**45.4**

Table 4: **Results on AudioCaps (left) and Clotho (right) for text-audio retrieval.** † Higher than reported in the paper, as these are provided by authors on our test set. Inputs refers to video inputs as follows: **A:** Audio spectrograms **V:** RGB video frames. Rows highlighted in light blue show Zero-shot (ZS) performance. Note the CLOTHO dataset contains audio only (no RGB) frames.

value of our automatic video mining pipeline.

Audio Retrieval: For text-audio retrieval we report results on two audio-centric datasets (i.e. datasets paired with natural language descriptions that focus explicitly on the content of the audio track) - AudioCaps [42] and Clotho [24]. The goal here is to retrieve the correct audio segment given a free form natural language query. While Clotho comes with only audio, AudioCaps has both audio and RGB frames. Results on the AudioCaps dataset are provided in Table 4 (left). We first show results for an audio-only encoder (we only feed spectrograms as input). We note that our model with no audio-text pretraining already outperforms the current state of the art [60] by a large margin (R@1: from 24.3 to 32.0), despite the fact that [60] uses features pretrained on VGGSound and VGG-ish features pretrained on YouTube8M. This could be because unlike their encoder, our encoder is trained end-to-end directly from spectrograms. We then show results with pretraining on the spectrograms from HowTo100M (no RGB frames are used here), and find that there is some improvement. Pretraining on the audio and captions from VideoCC3M however, gives substantial performance gains to R@1 by over 3%. This improvement is particularly impressive because the captions were transferred via visual similarity to still images and no additional manual audio-text supervision was used. We also report zero-shot results, and find that unsurprisingly, pretraining on HowTo100M results in poor performance, likely because the model has learned to focus on speech. VideoCC3M provides a large improvement, however there is still a distance to finetuning performance. Finally, we also show that using an audio-visual fusion encoder and training on VideoCC3M provides a further significant improvement demonstrating the complementarity of RGB information for this task. Results on Clotho are provided in Table 4 (right). Here we show a similar trend, but as Clotho is also

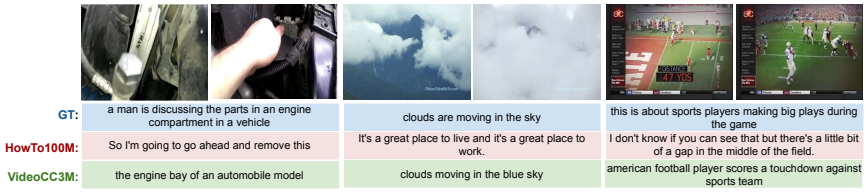

Fig. 5: **Zero-shot captioning results on MSR-VTT test set videos.** We show 2 frames per clip. As expected, the style of predicted captions from HowTo100M pretraining is similar to ASR, and concepts may be tenuously related (middle). Pretraining on VideoCC3M yields captions that are closer to the ground truth.

a much smaller dataset, we also show results with AudioCaps pre-training as is done by [60]. Combining AudioCaps supervised pretraining with VideoCC3M pretraining provides the best result.

5.4 Video Captioning

Results for video captioning are provided in Table 2 (right). For finetuning, our model pretrained on VideoCC3M outperforms previously published works. Unlike retrieval, pretraining on HowTo100M provides slight gains to the B-4 and M metrics, but VideoCC3M is still competitive with a fraction of the data size. We then compare zero-shot performance, and find that pretraining on HowTo100M performs poorly, potentially because of the large difference in style between instructional speech and human-generated captions. Training on VideoCC3M provides a substantial boost across all metrics, with a fraction of the data. Qualitative results are shown in Fig. 5.

6 Conclusion

We propose a new, automatic method for leveraging existing image datasets to mine video and audio data with captions. We apply it to the CC3M dataset [69] to mine millions of weakly labelled video-text pairs. Training a multimodal retrieval model on these clips leads to state of the art performance for video retrieval and captioning, and shows complementarity with existing image-text models such as CLIP. Future work can focus on augmenting these automatic captions with even more video related text, such as action labels, to overcome the image-centric bias in the mining pipeline. Societal impacts and fairness are discussed in supplementary.

References

1. YouTube Data API. https://developers.google.com/youtube/v3/docs/captions 10
2. Abavisani, M., Joze, H.R.V., Patel, V.M.: Improving the performance of unimodal dynamic hand-gesture recognition with multimodal training. In: CVPR (2019) 3
3. Aguilar, G., Rozgic, V., Wang, W., Wang, C.: Multimodal and multi-view models for emotion recognition. In: ACL (2019) 3
4. Albanie, S., Nagrani, A., Vedaldi, A., Zisserman, A.: Emotion recognition in speech using cross-modal transfer in the wild. In: Proceedings of the 26th ACM international conference on Multimedia. pp. 292–301 (2018) 3
5. Amrani, E., Ben-Ari, R., Rotman, D., Bronstein, A.: Noise estimation using density estimation for self-supervised multimodal learning. arXiv preprint arXiv:2003.03186 (2020) 2, 12
6. Anne Hendricks, L., Wang, O., Shechtman, E., Sivic, J., Darrell, T., Russell, B.: Localizing moments in video with natural language. In: ICCV (2017) 3
7. Antol, S., Agrawal, A., Lu, J., Mitchell, M., Batra, D., Lawrence Zitnick, C., Parikh, D.: VQA: Visual question answering. In: ICCV (2015) 1
8. Bain, M., Nagrani, A., Brown, A., Zisserman, A.: Condensed movies: Story based retrieval with contextual embeddings. In: ACCV (2020) 1, 3, 4
9. Bain, M., Nagrani, A., Varol, G., Zisserman, A.: Frozen in time: A joint video and image encoder for end-to-end retrieval. ICCV (2021) 1, 4, 8, 11, 12
10. Bain, M., Nagrani, A., Varol, G., Zisserman, A.: A clip-hitchhiker's guide to long video retrieval. arXiv preprint arXiv:2205.08508 (2022) 2
11. Banerjee, S., Lavie, A.: Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In: ACL workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization (2005) 10
12. Changpinyo, S., Sharma, P., Ding, N., Soricut, R.: Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3558–3568 (2021) 1, 2, 6
13. Chechik, G., Ie, E., Rehn, M., Bengio, S., Lyon, D.: Large-scale content-based audio retrieval from text queries. In: Proceedings of the 1st ACM international conference on Multimedia information retrieval. pp. 105–112 (2008) 4
14. Chen, H., Li, J., Hu, X.: Delving deeper into the decoder for video captioning. In: ECAI (2020) 12
15. Chen, H., Lin, K., Maye, A., Li, J., Hu, X.: A semantics-assisted video captioning model trained with scheduled sampling. Frontiers in Robotics and AI **7** (2020) 12
16. Chen, H., Xie, W., Vedaldi, A., Zisserman, A.: Vggsound: A large-scale audio-visual dataset. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 721–725. IEEE (2020) 10
17. Chen, S., Jiang, Y.G.: Motion guided spatial attention for video captioning. In: AAAI (2019) 9
18. Chen, X., Fang, H., Lin, T.Y., Vedantam, R., Gupta, S., Dollár, P., Zitnick, C.L.: Microsoft coco captions: Data collection and evaluation server. arXiv preprint arXiv:1504.00325 (2015) 11, 12
19. Cheng, X., Lin, H., Wu, X., Yang, F., Shen, D.: Improving video-text retrieval by multi-stream corpus alignment and dual softmax loss (2021) 12

20. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: CVPR (2009) 11
21. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: NAACL-HLT (2019) 8
22. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018) 10
23. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: ICLR (2021) 8
24. Drossos, K., Lipping, S., Virtanen, T.: Clotho: An audio captioning dataset. In: ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 736–740. IEEE (2020) 2, 3, 4, 10, 13
25. Elizalde, B., Zarar, S., Raj, B.: Cross modal audio search and retrieval with joint embeddings based on text and audio. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 4095–4099. IEEE (2019) 4
26. Fang, H., Xiong, P., Xu, L., Chen, Y.: Clip2video: Mastering video-text retrieval via image clip. arXiv preprint arXiv:2106.11097 (2021) 2, 12
27. Font, F., Roma, G., Serra, X.: Freesound technical demo. In: Proceedings of the 21st ACM international conference on Multimedia. pp. 411–412 (2013) 10
28. Gabeur, V., Nagrani, A., Sun, C., Alahari, K., Schmid, C.: Masking modalities for cross-modal video retrieval. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1766–1775 (2022) 3
29. Gabeur, V., Sun, C., Alahari, K., Schmid, C.: Multi-modal transformer for video retrieval. In: ECCV (2020) 1, 2, 12
30. Gao, R., Oh, T.H., Grauman, K., Torresani, L.: Listen to look: Action recognition by previewing audio. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10457–10467 (2020) 3
31. Gao, Z., Liu, J., Chen, S., Chang, D., Zhang, H., Yuan, J.: Clip2tv: An empirical study on transformer-based methods for video-text retrieval. arXiv preprint arXiv:2111.05610 (2021) 12
32. Gemmeke, J.F., Ellis, D.P., Freedman, D., Jansen, A., Lawrence, W., Moore, R.C., Plakal, M., Ritter, M.: Audio set: An ontology and human-labeled dataset for audio events. In: 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). pp. 776–780. IEEE (2017) 10
33. Ghadiyaram, D., Tran, D., Mahajan, D.: Large-scale weakly-supervised pre-training for video action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12046–12055 (2019) 2, 4
34. Girdhar, R., Tran, D., Torresani, L., Ramanan, D.: Distinit: Learning video representations without a single labeled video. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 852–861 (2019) 2
35. Gong, Y., Chung, Y.A., Glass, J.: Ast: Audio spectrogram transformer. arXiv preprint arXiv:2104.01778 (2021) 8
36. Gupta, S., Hoffman, J., Malik, J.: Cross modal distillation for supervision transfer. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (June 2016) 3
37. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531 (2015) 3

38. Hou, J., Wu, X., Zhao, W., Luo, J., Jia, Y.: Joint syntax representation learning and visual cue translation for video captioning. In: ICCV (2019) 12
39. Huang, G., Pang, B., Zhu, Z., Rivera, C., Soricut, R.: Multimodal pretraining for dense video captioning. In: AACL (2020) 3
40. Juan, D.C., Lu, C.T., Li, Z., Peng, F., Timofeev, A., Chen, Y.T., Gao, Y., Duerig, T., Tomkins, A., Ravi, S.: Graph-rise: Graph-regularized image semantic embedding. arXiv preprint arXiv:1902.10814 (2019) 5
41. Kay, W., Carreira, J., Simonyan, K., Zhang, B., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, P., et al.: The kinetics human action video dataset. arXiv preprint arXiv:1705.06950 (2017) 3, 10, 11
42. Kim, C.D., Kim, B., Lee, H., Kim, G.: Audiocaps: Generating captions for audios in the wild. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 119–132 (2019) 2, 3, 4, 10, 13
43. Koepke, A., Oncescu, A.M., Henriques, J.F., Akata, Z., Albanie, S.: Audio retrieval with natural language queries: A benchmark study. arXiv preprint arXiv:2112.09418 (2021) 4
44. Krishna, R., Hata, K., Ren, F., Fei-Fei, L., Carlos Niebles, J.: Dense-captioning events in videos. In: ICCV (2017) 1, 3
45. Krishna, R., Zhu, Y., Groth, O., Johnson, J., Hata, K., Kravitz, J., Chen, S., Kalantidis, Y., Li, L.J., Shamma, D.A., et al.: Visual genome: Connecting language and vision using crowdsourced dense image annotations. International journal of computer vision **123**(1), 32–73 (2017) 1, 3
46. Lei, J., Li, L., Zhou, L., Gan, Z., Berg, T.L., Bansal, M., Liu, J.: Less is more: Clipbert for video-and-language learning via sparse sampling. In: CVPR (2021) 2, 12
47. Lei, J., Yu, L., Bansal, M., Berg, T.L.: Tvqa: Localized, compositional video question answering. In: EMNLP (2018) 1
48. Li, L., Chen, Y.C., Cheng, Y., Gan, Z., Yu, L., Liu, J.: Hero: Hierarchical encoder for video+ language omni-representation pre-training. In: EMNLP (2020) 2, 12
49. Li, T., Wang, L.: Learning spatiotemporal features via video and text pair discrimination. arXiv preprint arXiv:2001.05691 (2020) 4
50. Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: ECCV (2014) 1, 3
51. Liu, Y., Albanie, S., Nagrani, A., Zisserman, A.: Use what you have: Video retrieval using representations from collaborative experts. In: BMVC (2019) 10, 12
52. Luo, H., Ji, L., Shi, B., Huang, H., Duan, N., Li, T., Chen, X., Zhou, M.: UniVL: A unified video and language pre-training model for multimodal understanding and generation. arXiv preprint arXiv:2002.06353 (2020) 2, 12
53. Luo, H., Ji, L., Shi, B., Huang, H., Duan, N., Li, T., Li, J., Bharti, T., Zhou, M.: UniVL: A unified video and language pre-training model for multimodal understanding and generation. arXiv e-prints (2020) 12
54. Luo, H., Ji, L., Zhong, M., Chen, Y., Lei, W., Duan, N., Li, T.: Clip4clip: An empirical study of clip for end to end video clip retrieval. arXiv preprint arXiv:2104.08860 (2021) 2, 12, 13
55. Miech, A., Alayrac, J.B., Smaira, L., Laptev, I., Sivic, J., Zisserman, A.: End-to-end learning of visual representations from uncurated instructional videos. In: CVPR (2020) 1, 8

56. Miech, A., Zhukov, D., Alayrac, J.B., Tapaswi, M., Laptev, I., Sivic, J.: HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In: ICCV (2019) 2, 4, 10, 11, 12
57. Monfort, M., Jin, S., Liu, A., Harwath, D., Feris, R., Glass, J., Oliva, A.: Spoken moments: Learning joint audio-visual representations from video descriptions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14871–14881 (2021) 4
58. Nagrani, A., Sun, C., Ross, D., Sukthankar, R., Schmid, C., Zisserman, A.: Speech2action: Cross-modal supervision for action recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10317–10326 (2020) 3
59. Nagrani, A., Yang, S., Arnab, A., Jansen, A., Schmid, C., Sun, C.: Attention bottlenecks for multimodal fusion. NeurIPS (2021) 8
60. Oncescu, A.M., Koepke, A., Henriques, J.F., Akata, Z., Albanie, S.: Audio retrieval with natural language queries. arXiv preprint arXiv:2105.02192 (2021) 3, 4, 10, 13, 14
61. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: ACL (2002) 10
62. Patrick, M., Huang, P.Y., Asano, Y., Metze, F., Hauptmann, A., Henriques, J., Vedaldi, A.: Support-set bottlenecks for video-text representation learning. arXiv preprint arXiv:2010.02824 (2020) 1, 2, 12
63. Patrick, M., Huang, P.Y., Asano, Y., Metze, F., Hauptmann, A.G., Henriques, J.F., Vedaldi, A.: Support-set bottlenecks for video-text representation learning. In: ICLR (2021) 1, 7
64. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision (2021) 1, 2, 12, 13
65. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners. Technical Report (2019) 9
66. Rohrbach, A., Torabi, A., Rohrbach, M., Tandon, N., Pal, C., Larochelle, H., Courville, A., Schiele, B.: Movie description. International Journal of Computer Vision **123**(1), 94–120 (2017) 1, 3, 4
67. Rouditchenko, A., Boggust, A., Harwath, D., Joshi, D., Thomas, S., Audhkhasi, K., Feris, R., Kingsbury, B., Picheny, M., Torralba, A., et al.: AVLnet: Learning audio-visual language representations from instructional videos. arXiv preprint arXiv:2006.09199 (2020) 2, 12
68. Seo, P.H., Nagrani, A., Schmid, C.: Look before you speak: Visually contextualized utterances. In: CVPR (2021) 1
69. Sharma, P., Ding, N., Goodman, S., Soricut, R.: Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In: ACL (2018) 1, 2, 3, 5, 6, 14
70. Shvetsova, N., Chen, B., Rouditchenko, A., Thomas, S., Kingsbury, B., Feris, R., Harwath, D., Glass, J., Kuehne, H.: Everything at once–multi-modal fusion transformer for video retrieval. CVPR (2022) 12
71. Slaney, M.: Semantic-audio retrieval. In: 2002 IEEE International Conference on Acoustics, Speech, and Signal Processing. vol. 4, pp. IV–4108. IEEE (2002) 4
72. Stroud, J.C., Lu, Z., Sun, C., Deng, J., Sukthankar, R., Schmid, C., Ross, D.A.: Learning video representations from textual web supervision. arXiv preprint arXiv:2007.14937 (2020) 2, 3

73. Sun, C., Shetty, S., Sukthankar, R., Nevatia, R.: Temporal localization of fine-grained actions in videos by domain transfer from web images. In: ACM Multimedia (2015) 3
74. Sun, C., Shetty, S., Sukthankar, R., Nevatia, R.: Temporal localization of fine-grained actions in videos by domain transfer from web images. In: ACM Multimedia (2015) 3
75. Tang, Z., Lei, J., Bansal, M.: Decembert: Learning from noisy instructional videos via dense captions and entropy minimization. In: NAACL (2021) 12
76. Vedantam, R., Lawrence Zitnick, C., Parikh, D.: Cider: Consensus-based image description evaluation. In: CVPR (2015) 10
77. Vinyals, O., Toshev, A., Bengio, S., Erhan, D.: Show and tell: Lessons learned from the 2015 MSCOCO image captioning challenge. IEEE transactions on pattern analysis and machine intelligence **39**(4), 652–663 (2016) 1
78. Wang, B., Ma, L., Zhang, W., Jiang, W., Wang, J., Liu, W.: Controllable video captioning with pos sequence guidance based on gated fusion network. In: ICCV (2019) 12
79. Wang, L., Li, Y., Lazebnik, S.: Learning deep structure-preserving image-text embeddings. In: CVPR (2016) 1
80. Wang, X., Zhu, L., Yang, Y.: T2vlad: Global-local sequence alignment for text-video retrieval (2021) 12
81. Xu, H., Ghosh, G., Huang, P.Y., Okhonko, D., Aghajanyan, A., Metze, F., Zettlemoyer, L., Feichtenhofer, C.: Videoclip: Contrastive pre-training for zero-shot video-text understanding. arXiv preprint arXiv:2109.14084 (2021) 11, 12
82. Xu, J., Mei, T., Yao, T., Rui, Y.: Msr-vtt: A large video description dataset for bridging video and language. In: CVPR (2016) 1, 3, 10
83. Yao, L., Huang, R., Hou, L., Lu, G., Niu, M., Xu, H., Liang, X., Li, Z., Jiang, X., Xu, C.: Filip: Fine-grained interactive language-image pre-training. arXiv preprint arXiv:2111.07783 (2021) 8
84. You, Q., Jin, H., Wang, Z., Fang, C., Luo, J.: Image captioning with semantic attention. In: CVPR (2016) 1
85. Zhai, A., Wu, H.Y.: Classification is a strong baseline for deep metric learning. In: BMVC (2019) 8
86. Zhang, Z., Shi, Y., Yuan, C., Li, B., Wang, P., Hu, W., Zha, Z.J.: Object relational graph with teacher-recommended learning for video captioning. In: CVPR (2020) 12
87. Zhou, L., Xu, C., Corso, J.: Towards automatic learning of procedures from web instructional videos. In: AAAI (2018) 3