MERLOT:
Multimodal Neural Script Knowledge Models

Anonymous Author(s)
Affiliation
Address
email

Abstract

As humans, we understand events in the visual world contextually, performing multimodal reasoning across time to make inferences about the past, present, and future. We introduce MERLOT, a model that learns multimodal script knowledge by watching millions of YouTube videos with transcribed speech – in an entirely label-free, self-supervised manner. By pretraining with a mix of both frame-level (spatial) and video-level (temporal) objectives, our model not only learns to match images to temporally corresponding words, but also to contextualize what is happening globally over time. As a result, MERLOT exhibits strong out-of-the-box representations of temporal commonsense, and achieves state-of-the-art performance on 12 different video QA datasets when finetuned. It also transfers well to the world of static images, allowing models to reason about the dynamic context behind visual scenes. On Visual Commonsense Reasoning, MERLOT answers questions correctly with 80.6% accuracy, outperforming state-of-the-art models of similar size by over 3%, even those that make heavy use of auxiliary supervised data (like object bounding boxes).

Ablation analyses demonstrate the complementary importance of: 1) training on videos versus static images; 2) scaling the magnitude and diversity of the pretraining video corpus; and 3) using diverse objectives that encourage full-stack multimodal reasoning, from the recognition to cognition level.

1 Introduction

The human capacity for commonsense reasoning is shaped by how we experience causes and effects over time. Consider the still image of people dining at a restaurant in the bottom right of Figure 1: while a literal, concrete description like "people sitting at a table eating" might be technically correct for the static scene, it doesn't capture the richer temporal, commonsense inferences that are nonetheless obvious: *before* sitting down, the people had to meet up, agree where to go, and enter the

Submitted to 35th Conference on Neural Information Processing Systems (NeurIPS 2021). Do not distribute.

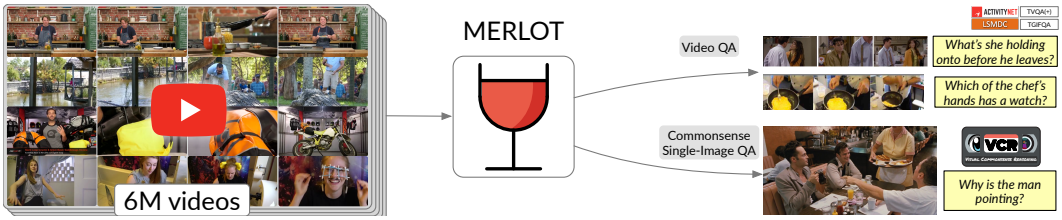

Figure 1: Multimodal Event Representation Learning Over Time. We pretrain representations of multimodal script knowledge over 6 million YouTube videos. These representations can then be applied to a variety of downstream tasks that require commonsense or temporal visual reasoning.

restaurant; *at present*, the man is pointing because the server just came to the table, and she might want to know whose food is whose; and *after*, it is likely the server will return to the kitchen to help another table.

Teaching machines this type of *script knowledge* [68] is a significant challenge in no small part because enumerating all facts, inferences, and counterfactuals is prohibitive. As a result, the highest performing models on vision-and-language tasks, including Visual Commonsense Reasoning (VCR) (where Figure 1's scene originates from), learn about the visual world exclusively through manually-annotated bounding boxes, along with image captions [79, 13, 46, 52, 88, 22, 63]. These captions tend to describe what is *literally* happening in the scene. Though some captions might hint at the past and future, it is not obvious that even training on, e.g., 400M literal image/text pairs [63] will result in models capable of temporal reasoning.

In this paper, we introduce MERLOT, short for Multimodal Event Representation Learning Over Time. MERLOT is a model that learns commonsense representations of multimodal events by self-supervised pretraining over 6M unlabelled YouTube videos. With the goal of learning multimodal reasoning capacity beyond static images/literal captions, we train MERLOT to **a)** match individual video frames with a contextualized representation of the video transcript, divided into segments, and to **b)**, contextualize that strong frame-level representation over time, and over spoken words, to both 'unmask' word-level corruptions [15] and reorder scrambled video frames. Our combination of pretraining objectives allow us to scale up the amount of language-and-vision context given to models.

We validate our model on a diverse suite of tasks, requiring both recognition- and cognition-level reasoning. When finetuned, MERLOT achieves a new state-of-the-art on 12 challenging video tasks, covering temporal reasoning at both long and short timescales. Additionally, we show that our script-knowledge representations transfer to the world of single images. On Visual Commonsense Reasoning (VCR; [92]), our model achieves particularly strong performance, outperforming models that require heavy visual supervision (in the form of object detection bounding boxes, or images paired with pristine captions).

In addition to being a strong model when finetuned for downstream tasks, we show both quantitatively and qualitatively that MERLOT has a strong out-of-the-box understanding of everyday events and situations. Given a scrambled visual story, [30, 1], MERLOT can order the frames to match the captions, to tell a globally coherent narrative. Despite considerable domain shift from videos to static images, MERLOT outperforms strong baselines like CLIP [63] and UNITER [13], which perform 'image-text' matching independently and cannot reason over long-term contexts. This capacity seems to emerge during pretraining: analysis of MERLOT's attention patterns (Figure 3) show that regions attend to captions that are distant in time (and vice versa), allowing it perform cross-modal coreference to piece together a holistic view of situations.

Finally, we present analyses and ablations to break down MERLOT's performance. We show that 1) pretraining works better when we train on videos rather than still images, aided crucially by our strategy of corrupting highly visual words in the masked language modeling task, 2) using a diverse set of videos covering many aspects of everyday situations improves downstream performance compared to curated instructional video corpora [78, 55] which both cover a smaller slice of the visual world (confirming hypotheses from past work [27]); and 3) MERLOT's performance does not saturate even after many epochs of training on YT-Temporal-180M, as it continues to improve performance simply with more pretraining. The combination of these results suggests that learning full-stack visual reasoning and multimodal world knowledge from video data is a promising path forward for future research.

In summary, our main contributions are:

1. MERLOT a performant end-to-end vision and language model, that learns powerful multimodal world representations from videos and their transcripts – using no labeled data.
2. YT-Temporal-180M, a diverse corpus of frames/ASR derived from a filtered set of 6M diverse YouTube videos, which we show greatly aids performance, and
3. A set of experiments/ablations demonstrating the strong performance of MERLOT on a set of 14 tasks, spanning finetuning and zero-shot transfer, and images and videos.

We plan to release code and data for public research use.

2 Related Work

2.1 Joint representations of written text and images

There has been much past work on learning joint representations of written text and images. Most recently, A variety of papers have proposed "Visual BERT" style models [79, 13, 7, 46, 52, 88, 22]; these approaches train on datasets of image captions such as MSCOCO [48] and Visual Genome [37]. All but a few of these approaches extract features using Anderson et al. [9]'s frozen object detector, which was originally trained on Visual Genome. Two notable exceptions are Zhang et al. [94], who use an even larger object detector trained on more labeled data, and Kim et al. [35], who use a vision-transformer backbone that was pretrained on ImageNet. As such, they all rely on the significant human annotation to learn (static) visual representations. Instead, our approach learns *dynamic* visual representations purely from videos, with no human annotation. We do this through a contrastive loss that extends past work on images with independent captions [95, 63] to video frames with contextualized transcripts.

2.2 Learning from videos, with automatic speech recognition (ASR) transcripts

Prior works have used web videos+ASR to build weakly-supervised object detectors [61], action detectors/classifiers [89, 5, 38, 58], instruction aligners [53, 4, 12], video captioners [69, 26, 60, 73], and visual reference resolvers [29]. Of late, works have sought to learn multimodal representations transferable to many tasks from uncurated sets of (usually how-to) videos [55, 77, 78, 56, 96, 8, 6, 3]. Much of the design and evaluation of these models has been focused on video understanding tasks such as activity recognition. Recently, the proposed ClipBERT model from Lei et al. [44] learns vision-language representations from image captions, and was evaluated on video-language tasks, covering text-to-video retrieval, video captioning, and video QA. Our method, on the other hand, learns multimodal representations over time from video data and is evaluated on both static image and video tasks. We report results that suggest that visual commonsense may be more general, as our model transfers to both tasks.

2.3 Temporal ordering and forecasting

There has been a large body of work on analyzing 'what happens next' in videos [36]. Some modeling choices include using pixels [19, 82], graphs [10], euclidean distance using sensors [2], or studying cycle consistency across time Epstein et al. [17]. In addition to extrapolation, past work has studied deshuffling objectives in video frames [57, 83], though this has mostly been focused on the visual modality. In contrast to these papers, our goal is to learn *multimodal* script knowledge representations, including both image frames and transcribed speech. These representations cover what is happening in the world behind the pixels, rather than just what changes on-screen.

3 MERLOT: Multimodal Event Representation Learning Over Time

MERLOT is a Transformer-based [81] vision+language model. We now describe our pretraining dataset, architecture, and objectives.

3.1 YT-Temporal-180M

We collect YT-Temporal-180M, a dataset for learning multimodal script knowledge derived from 6 million YouTube videos. Our starting point is a list of 27 million candidate video IDs, aggregated from a variety of prior machine learning corpora and our own scraping efforts. For public videos, the YouTube API provides access to video itself, its ASR track (i.e., automatically transcribed speech tokens), and other metadata. Using this information, we discard videos 1) without an English ASR track; 2) that are too long (>20 minutes) or belong to visually "ungrounded" categories like video game commentaries [27]; and 3) that have thumbnails unlikely to contain objects, according to a lightweight image classifier. We add punctuation to the ASR by applying a sequence-to-sequence model trained to add punctuation to sentences/paragraphs from news articles. Additional details are provided in Appendix.

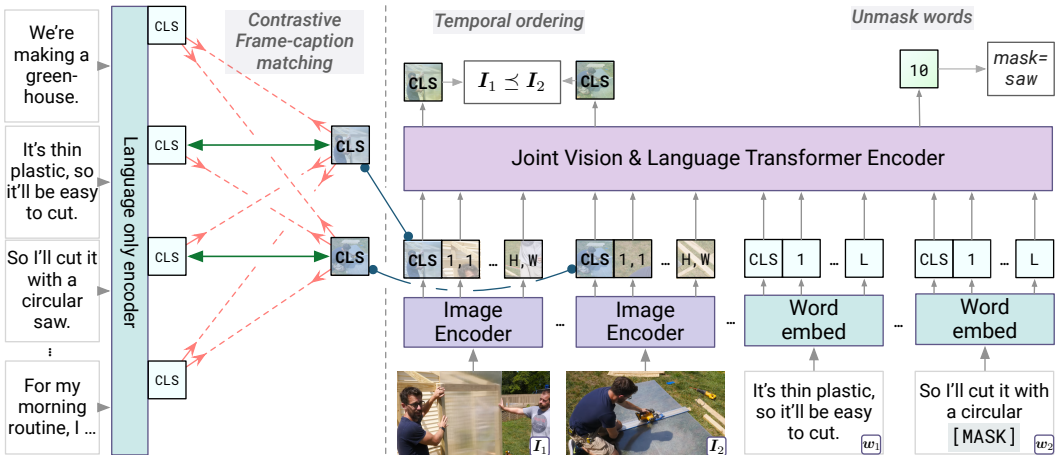

Figure 2: Left: MERLOT learns to match contextualized captions with their corresponding video frames. Right: the same image encoding is provided, along with (masked) word embeddings, into a joint vision-language Transformer model; it then unmasks ground words (like 'saw' in this example) and puts scrambled video frames into the correct order.

YT-Temporal-180M spans multiple domains, datasets, and topics – including instructional videos from HowTo100M [55], lifestyle vlogs of everyday events from the VLOG dataset [20], and YouTube's auto-suggested videos for popular topics like 'science' or 'home improvement.' Our intent was explicitly to make the corpus as diverse as possible to encourage the model to learn about a broad range of objects, actions, and scenes [27]: a later ablation shows that limiting our pretraining to only instructional videos indeed hurts performance downstream.

Each video $\mathcal{V}$ might contain thousands of frames. In this work, we represent a video $\mathcal{V}$ as a sequence of consecutive **video segments** $\{\boldsymbol{s}_t\}$. Each segment $\boldsymbol{s}_t$ consists of:

a. an image frame $\boldsymbol{I}_t$, extracted from the middle timestep of the segment,

b. the words $\boldsymbol{w}_t$ spoken during the segment, with a total length of L tokens.

To split the videos into segments, we byte-pair-encode (BPE) each video transcript and align the tokens with YouTube's word-level timestamps [70, 62]. This enables us to split the videos into segments of L=32 BPE tokens each, which we discuss in Appendix A.5. Our final dataset contains 180 million segments of this form.

3.2 MERLOT Architecture

A diagram of MERLOT is given in Figure 2. MERLOT takes multiple unordered video frames as input $\{\boldsymbol{s}_t\}$. We encode each frame $\boldsymbol{I}_t$ using an image encoder, embed the words $\boldsymbol{w}_t$ using a learned embedding, and jointly encode both using a Transformer. Once the parameters of the model are fit, the architecture can then be easily applied to a variety of vision-and-language tasks with minimal modification. For video QA, for example, we pass several video frames to the image encoder, the question to the text encoder, and extract a single vector representation from the `CLS` token position. For each task, we learn a lightweight classification head mapping from this hidden state to the task's label space; specific modeling/optimization details are given in the Appendix.

Image encoder. We use a ResNet-50 backbone, followed by a 12-layer, 768-dimensional Vision Transformer [24, 81, 16]. The image encoder is trained end-to-end with the rest of the model, and is trained from random initialization – without any initialization on supervised data.

In addition to being unsupervised, our vision model is also faster than those typically used in top-performing vision and language models, which pre-extract features from a (supervised) object detector [79, 46, 52, 13, 45].[1] Most work uses the same object detector: [9]'s ResNet101 Faster-RCNN model

[1]Though recent work has proposed using grid-based features, on tasks like Visual Commonsense Reasoning, these approaches have so far underperformed the more expensive object detector backbones [92].

trained on Visual Genome [64, 24, 37].[2] These approaches "freeze" the detector during training, as otherwise it might unlearn how to localize objects, which reduces the scope of (temporal) visual knowledge available to the model. A few additional modeling decisions we made to balance efficiency and runtime include: 1) We train on smaller, widescreen images of size 192x352 (because most YouTube videos are widescreen) using a patch size of 16 pixels; 2) we mirror [16]'s alterations of removing the C5 block in ResNet-50; and 3) we save compute further by average-pooling the final-layer region cells using a kernel size of 2×2. With these modifications, our image encoder requires 40 gigaFLOPs for a forward pass, which is 2% of the 2 teraFLOPs required for the Faster-RCNN.

In summary: given an image of size $W \times H$, the image encoder outputs a sequence of $1 + W/32 \times H/32$ feature vectors, which includes a `CLS` token for image-level representations. These will next be contextualized with respect to other images, and tokens.

Joint Vision-Language Encoder. Given the feature maps computed by the image encoder over images $\{\boldsymbol{I}_t\}$, and the tokens $\{\boldsymbol{w}_t\}$, we will jointly represent all components using a 12-layer, 768-dimensional Transformer encoder [81]. We first embed the tokens, and add position embeddings to both language and vision components. The position embeddings differ between different segments, allowing the network to distinguish between images and captions at different timesteps. Finally, we encode both image regions and words using the Transformer, which attends globally over all segments.

The tokens $\boldsymbol{w}_t$ in each segment are prefixed with a `CLS` token, recall that the feature maps for each frame $\boldsymbol{I}_t$ start with one as well. At those positions, we will later pool final-layer hidden-state representations, for use in pretraining along with downstream tasks.

3.3 Pretraining Tasks and Objectives

We use the following three objectives to pretrain MERLOT, that cover 'full-stack' visual reasoning – from recognition subtasks (like object detection) that operate at the frame level, to more 'cognitive' tasks that operate at the video level.

1. **Contrastive frame-caption matching** [95, 63]. To ensure that the underlying image encoder produces helpful image representations, we add a contrastive loss to maximize their similarity to representations computed by a language-only representation of the video transcript. To conserve memory, we have our 'language-only encoder' for this subtask share architecture and parameters with the joint vision-and-language encoder.

   Unlike what is the case for many image captions, the words $\boldsymbol{w}_t$ in each segment are often not sufficient to describe the gist of $\boldsymbol{I}_t$, or even what the key objects might be – for that, video-level contextualization is often required. We thus pass the entire transcript into the Language-only Encoder, which then extracts hidden states for each segment at the segment-level `CLS` tokens.

   Given matching representations for each frame $\boldsymbol{I}_t$ and caption $\boldsymbol{w}_t$ as positive examples, the negative examples come from all other frame-caption pairs in the minibatch – whether or not they come from the same video. We project both of these representations into a size-768 hidden state which is then L2 normalized, and compute an all-pairs dot-product between all image and text representations. We divide these logits by a temperature of $\tau = 0.05$, and then apply a pairwise cross entropy loss to encourage matching captions and frames.
2. **(Attention) Masked Language Modeling** When providing words into the joint vision-and-language encoder, we randomly replace 20% with a `MASK` token, a random word, or the same word; MERLOT must then reconstruct the correct word with a cross-entropy loss, following [15].

   This approach is commonly used by 'visual BERT' models in the image captioning domain, where captions are concise, and thus the identity of masked concrete words is difficult for models to recover given language context alone. However, we observed qualitatively that videos break these assumptions: people tend to ramble, and often mention key objects multiple times. Thus, applying vanilla BERT-style masking often causes ungrounded fillers like 'umm' or 'yeah' to get masked, while the names of important objects would often be repeated, likely penalizing the learning of multimodal representations.

[2] This is why we chose the ResNet-Vision Transformer hybrid: like commonly used backbones, ours is a ResNet with much of the computation happening higher up in the network.

	Q→A	QA→R	Q→AR
UNITER-base [13]	75.0	77.2	58.2
VILLA-base [22]	76.4	79.1	60.6
ERNIE-ViL-base [88]	77.0	80.3	62.1
MERLOT (base-sized)	**80.6**	**80.4**	**65.1**

(a) Results on VCR [92]. We compare against SOTA models of the same size as ours (12-layer vision-and-language Transformers).

	Spearman (↑)	Pairwise acc (↑)	Distance (↓)
CLIP [63]	60.9	78.7	63.8
UNITER [13]	54.5	75.2	74.5
MERLOT	**85.5**	**89.4**	**39.6**

(b) Results unscrambling visual stories, on VIST [30, 1]. The captions are provided in the correct order, models must arrange the images temporally. MERLOT performs best on all metrics by reasoning over sequences of events, versus single image-caption pairs.

Table 1: Results on image-based tasks.

We introduce a simple solution to this problem, that we call **attention masking**: we use attention weights from language-only transformer (introduced in the previous objective) as a heuristic for which words are grounded. 50% of the time, we mask out a random token; the other 50% of the time, we mask out one of the top 20% most-attended-to-tokens. We then apply SpanBERT masking [33], randomly corrupting the following or preceding tokens with an average length of 0.5 tokens in each direction; this makes it harder for models to over-rely on BPE artifacts. We show in ablations that this improves performance.

3. **Temporal Reordering**. We experiment with having the model order image frames in a video, forcing it to learn temporal reasoning. 40% of the time, we pick a random integer i between 2 and N, where N is the number of segments provided to the joint vision-language encoder. Then we randomly scramble i frames chosen at random, by replacing the segment-level position embeddings (e.g. `[image_t]`) for that frame with a random and unique position embedding, e.g. `[image_unk_0]`). These random position embeddings are also learned, and are separate from the 'unshuffled' position embeddings. This allows the model to both distinguish between different pieces, and unshuffle each 'shuffled' frame or caption conditioned on the correctly-ordered ones (if there are any).

   To compute the loss from this objective, we extract hidden states h_{t_i} and h_{t_j} from the `CLS` token positions of each frame-frame pair. The hidden states are concatenated and passed through a two-layer multilayer perceptron, to predict whether $t_i < t_j$ or $t_i > t_j$. We optimize this loss using softmax cross entropy.

3.4 Pretraining MERLOT

We pretrain our model over our dataset of videos for 40 epochs. We preprocess the dataset into examples with sequences of N=16 video segments each, each containing up to L=32 BPE tokens.[3] The language-only encoder computes contrastive representations given this entire sequence, its total length is thus 512 tokens. To save memory for the joint vision-language encoder, we provide it 4 groups of $N = 4$ segments each. At an image training resolution of 192×352, the model's sequence length is 396 tokens. To combine the losses, we multiply the contrastive loss (**1.**) with a coefficient of 0.25, which we found put its gradient magnitudes at roughly the same scale as the Mask LM loss.

We train the model using a v3-1024 TPU pod, at a batch size of 1024 sequences (or 16k segments) in total. This pretraining process on our hardware takes 30 hours. We provide additional information about hyperparameters and experimental setup in the Supp.

4 Experiments: Transferring MERLOT to Downstream Tasks

In this section, we explore MERLOT on 14 different end tasks, covering vision-language reasoning over both images as well as videos.

4.1 Image tasks

VCR. We consider VCR [92], a task and dataset where models must answer commonsense visual questions about images. These questions test the ability of our model to transfer its video-level

[3] To train the model on as much data as possible, we merged together the segments of short videos.

Tasks	Split	Video Length	ActBERT [96]	ClipBERT$_{8x2}$ [44]	SOTA	MERLOT
MSRVTT-QA	Test	Short	-	37.4	41.5 [87]	**43.1**
MSR-VTT-MC	Test	Short	88.2	-	88.2 [96]	**90.9**
ActivityNetQA	Test	Short	-	-	38.9 [87]	**41.4**
TGIF-Action	Test	Short	-	82.8	82.8 [44]	**94.0**
TGIF-Transition	Test	Short	-	87.8	87.8 [44]	**96.2**
TGIF-Frame QA	Test	Short	-	60.3	60.3 [44]	**69.5**
LSMDC-FiB QA	Test	Short	48.6	-	48.6 [96]	**52.9**
LSMDC-MC	Test	Short	-	-	73.5 [90]	**81.7**
Drama-QA	Val	Long	-	-	81.0 [34]	**81.4**
TVQA	Test	Long	-	-	76.2 [34]	**78.7**
TVQA+	Test	Long	-	-	76.2 [34]	**80.9**
VLEP	Test	Long	-	-	67.5 [43]	**68.4**

Table 2: Comparison with state-of-the-art methods on video reasoning tasks. MERLOT outperforms state of the art methods in **12** downstream tasks that involve short and long videos.

understanding to the world of single images, as while the questions often cover 'what might happen next' or 'what are people's goals and motivations,' only a single frame is provided, and models (like humans) must read between the lines to infer the right answer.

VCR provides additional information to models, in the form of bounding boxes around entities, and explicit groundings between those entities and references in questions, e.g., if `PERSON-1` is referenced in the question, the location of `PERSON-1` is also given in the image. Different from prior works that arrange the neural architecture to be aware of this cross-reference, we incorporate this information *directly in the image input* by drawing (in pixel space) a colored highlight around the referenced entity, and making reference to the color, e.g., in red; we show examples of this in the Supplemental.

Our results on the three VCR settings, in comparison to (non-ensembled) methods, are given in Table 1. Our model, despite having only a base transformer, and with visual features that are much more inexpensive than the bigger models, achieves strong performance.

Unsupervised SortStory. To probe our model's ability to do out-of-the-box commonsense reasoning over events in images, we next consider the SortStory task [1, 51]. In our formulation, a model is given five images and captions; the captions are in sorted order, and the model must order the images to match the captions and tell a coherent narrative.[4]

We compare MERLOT with two strong out-of-the-box baselines for text-image matching: CLIP [63], which encodes each caption and image separately and computes similarity through a dot product, and UNITER [13] which jointly represents each image/caption pair, and is trained in part using a 'text-image matching' objective. We use our temporal reordering loss to find the most probable ordering of the video frames; for CLIP and UNITER we compute a maximum-weight matching [39] over the pairwise image-text similarity scores. Results over 5K stories are given in Table 1b. MERLOT's performance in comparison to the algorithms trained from image-literal caption pairs suggests that, with no fine-tuning, our model has greater capability to reason about past and future events.

4.2 Video Reasoning

We also report results on 10 video reasoning tasks. Following the findings of [44], instead of modeling visual content using video-specific, 3D convolutional neural networks (e.g., S3D [85]) we adopt a sparse sampling approach, collecting a sequence of still frames from each video clip instead (we generally use 5, but details per-task are in Appendix C).

The tasks we report on are: TVQA [41], TVQA(+) [42], VLEP [43], MSRVTT-QA [86], MSRVTT-Multichoice [90], LSMDC-Multichoice, LSMDC fill-in-the-blank QA [80, 65], ActivityNetQA

[4] We set the task up this way to avoid unimodal biases. In [1]'s original formulation, models have to unshuffle (image,caption) pairs, but past work has shown that large language models can often rely on discursive clues to do this task [15, 74].

Training setup	VCR	TVQA+
One segment (N=1)	73.8	75.2
One segment, attention masking	73.5	74.5
Four segments	74.1	73.3
Four segments, atttention masking (Ours)	**75.2**	**75.8**
Four segments, no contrastive V-L loss	57.5	67.6

(a) **Context helps, but only with attention masking and contrastive loss.** Increasing N given to our model improves performance, but more context can encourage language-only representation learning. Attention masking seems to counteract this, giving an additional 1 point boost. All this relies on the contrastive loss bridging frame representations with those from captions; removing this loss hurts performance.

Dataset	VCR	TVQA+
Conceptual ∪ COCO	58.9	44.3
HowTo100M	66.3	72.1
YT-Temporal-180M	**75.2**	**75.8**

(b) **Diverse (video) data is important.** Applying our architecture to caption data leads to bad results. Our model performs better on HowTo100M, yet still below our dataset of diverse videos spanning many grounded domains.

# epochs	VCR
5 epochs	75.1
10 epochs	75.9
20 epochs	77.0
30 epochs	78.5
40 epochs	**79.4**

(c) **Training for longer helps,** with performance increasing monotonically over training iterations.

Table 3: Ablation study on the validation set of VCR question answering ($Q \rightarrow A$) and TVQA+

[91, 25], TGIFQA [31], and DramaQA [14][5]. We apply MERLOT to these tasks in the same way – initializing new parameters only to map the model's pooled CLS hidden state into the output labels, and using a softmax cross entropy loss; see the supplemental for details.

Table 2 contains the performance results of MERLOT when applied to these corpora. For all of these datasets, MERLOT represents a new state-of-the-art. Given the diversity of tasks and the strengths of the models used for comparison, these results provide strong evidence that MERLOT has learned strong representations, useful for a variety of temporal, multimodal reasoning tasks.

5 Analysis and Ablations

5.1 Ablations

We present ablations over VCR and TVQA+ to study the effect of several modeling decisions.

Context size. Table 3a shows the effect of varying the number of segments N given to the joint vision-and-language encoder during pretraining. In the first two rows, we consider passing only N=1 segments to the model; we keep the effective batch size the same, so that we use $4\times$ the number of sequences at $\frac{1}{4}$th the length. In this regime, we find that our proposed 'attention masking' approach – where we preferentially mask out tokens that are highly attended to by the contrastive language-only encoder – does not help over a strong baseline of masking purely randomly (with spans, as in [33]). We suspect that our targeted masking and random masking are comparable when only L=32 tokens available in the single segment case: both are likely to mask out a visually-groundable word (e.g., an object reference), forcing the model to make cross-modal inferences. However, when we scale the model to N=4 segments/128 tokens, our masking becomes more effective, improving by 1 point over the baseline. We suspect that random masking under-performs for 128 tokens because, unlike our targeted approach, random selection is unlikely to obscure *most* visual references in the transcript, allowing the model to attend less to text only patterns instead of cross-modal factors.

Additionally, we perform qualitative analyses in the supplemental material of the attention patterns produced by the language-only encoder. We find that highly attended-to tokens tend to be more 'visual' (e.g., the 'toys' the children are playing with in the supp. figure) and thus we suspect are good candidates to mask out if the focus is on learning grounded event representations.

Dataset source. In Table 3b, we investigate pretraining MERLOT on other two additional datasets beyond YT-Temporal-180M. First, we try training on 3 million static image-caption pairs from Conceptual Captions [72] combined with MSCOCO [48]; for fair comparison, we train for the same number of steps as 5 epochs on our dataset. The resulting model achieves 58.9% accuracy on VCR. We suspect this might be due to 1) a smaller context window (Table 3a), and 2) overfitting (5 epochs

[5] We attempted to submit to the DramaQA test set leaderboard, and found it was closed to new submissions. The organizers kindly offered to re-open the leaderboard, and we will provide leaderboard test set predictions. But, we weren't able to coordinate in time for submission, and thus only report validation results for now.

on YT-Temporal-180M corresponds to 300 epochs on the caption data). Because our vision pipeline is trained from scratch, the scale of the curated/supervised image pairing corpora is a concern.

We next investigate the impact of video selection, comparing YT-Temporal-180M with HowTo100M [55]. The model performs better by roughly 9 points on VCR and 4 points on TVQA+. In this case, overfitting is less of a concern (5 epochs on YT-Temporal-180M corresponds to only 30 epochs on HowTo100M, while training past this level helps MERLOT (Table 3c)). Instead, we suspect that coverage of a broader range of genres of videos is the cause of improved performance [27]; e.g., for VCR, the objects/actions/scenes diverge from the how-to domain, and instead might be more similar to other types of YouTube videos (e.g., web dramas, VLOGs, etc.).

Pretraining longer. Last, in Table 3c, we investigate the effect of pretraining MERLOT for longer. The performance increases monotonically and doesn't begin to plateau, which suggests that had we pretrained MERLOT for *even* longer, its performance could improve even further.

5.2 Qualitative analysis: MERLOT learns representations spanning distant frames

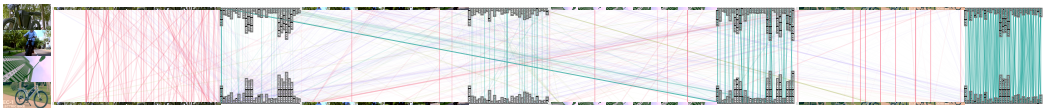

Figure 3: MERLOT attention patterns, averaged over all layers, show it learns to reason over distant frames and captions. The video is about an e-bike, which is absent from the third frame, the model nevertheless learns to unmask words like 'electric' by attending to context further away in time.

We briefly analyze the attention patterns produced by our model. In Figure 3, we show one such example, MERLOT addresses video tasks by merging attention across (distant) video segments, suggesting the importance of contextual reasoning across segments. A more detailed discussion (and additional examples) are given in the supplementary material.

6 Conclusion, Limitations, and Broader Impacts

We introduced Multimodal Event Representation Learning Over Time (MERLOT). We train our model through a combination of self-supervised objectives on 6M YouTube videos, in the process of learning powerful visual representations that go beyond single frames. The model achieves strong performance on tasks requiring event-level reasoning over videos and static images.

We hope that MERLOT can inspire future work for learning vision+language representations in a more human-like fashion compared to learning from literal captions and their corresponding images. There are several potential limitations of MERLOT that would make for promising avenues of future work, including: 1) exploring finer-grained temporal reasoning pretraining objectives vs. frame ordering e.g., a temporal frame *localization* within transcripts; and 2) investigating non-English videos+communities on YouTube; given the relative limited availability of non-English multimodal annotated resources, a multilingual version of our model offers a natural extension.

Our work in constructing MERLOT risks some potential negative impacts, including:

a. *Data collection+privacy.* As with other corpora gathered from the web used for pretraining data, YT-Temporal-180M contains publicly available content posted by users. While the YouTube API provides an interface to individual videos, our intent in *broadly* gathering data was to respect a user's capacity to be *in public* without *being public* (i.e., purposefully seeking exposure), a distinction noted by Marwick and boyd [54].

b. *Social biases resulting from curation.* It's possible that curation choices made could cause the model to exhibit undesirable social bias (and, thus, we do not advocate for deployed use-cases). A direct example is our decision to discard non-English ASR, which not only limits MERLOT to English-only representations, but also could result in, e.g., disparate capacity to accurately model the non-English speaking world.

c. *Energy Use.* While pretraining is expensive up-front [75], because our model uses fewer FLOPs vs. Faster-RCNN object-detection based methods in other V+L models, we hope the public release and adoption of our model can amortize this cost.

References

[1] Harsh Agrawal, Arjun Chandrasekaran, Dhruv Batra, Devi Parikh, and Mohit Bansal. Sort Story: Sorting Jumbled Images and Captions into Stories. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 925–931, 2016.

[2] Pulkit Agrawal, Ashvin Nair, Pieter Abbeel, Jitendra Malik, and Sergey Levine. Learning to poke by poking: experiential learning of intuitive physics. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 5092–5100, 2016.

[3] Hassan Akbari, Linagzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. VATT: transformers for multimodal self-supervised learning from raw video, audio and text. *arXiv preprint arXiv:2104.11178*, 2021.

[4] Jean-Baptiste Alayrac, Piotr Bojanowski, Nishant Agrawal, Josef Sivic, Ivan Laptev, and Simon Lacoste-Julien. Unsupervised learning from narrated instruction videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4575–4583, 2016.

[5] Jean-Baptiste Alayrac, Ivan Laptev, Josef Sivic, and Simon Lacoste-Julien. Joint discovery of object states and manipulation actions. In *ICCV*, 2017.

[6] Jean-Baptiste Alayrac, Adrià Recasens, Rosalia Schneider, Relja Arandjelović, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. Self-supervised multimodal versatile networks. *arXiv preprint arXiv:2006.16228*, 2020.

[7] Chris Alberti, Jeffrey Ling, Michael Collins, and David Reitter. Fusion of detected objects in text for visual question answering. *arXiv preprint arXiv:1908.05054*, 2019.

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

[9] Peter Anderson, Xiaodong He, Chris Buehler, Damien Teney, Mark Johnson, Stephen Gould, and Lei Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *CVPR*, 2018.

[10] Peter Battaglia, Razvan Pascanu, Matthew Lai, Danilo Jimenez Rezende, and Koray kavukcuoglu. Interaction networks for learning about objects, relations and physics. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 4509–4517, 2016.

[11] Björn Bebensee and Byoung-Tak Zhang. Co-attentional transformers for story-based video understanding. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4005–4009. IEEE, 2021.

[12] Chien-Yi Chang, De-An Huang, Yanan Sui, Li Fei-Fei, and Juan Carlos Niebles. D3tw: Discriminative differentiable dynamic time warping for weakly supervised action alignment and segmentation. In *CVPR*, 2019.

[13] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. UNITER: Learning universal image-text representations. *arXiv preprint arXiv:1909.11740*, 2019.

[14] Seongho Choi, Kyoung-Woon On, Yu-Jung Heo, Ahjeong Seo, Youwon Jang, Seungchan Lee, Minsu Lee, and Byoung-Tak Zhang. DramaQA: character-centered video story understanding with hierarchical qa. *arXiv preprint arXiv:2005.03356*, 2020.

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

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

[17] Dave Epstein, Jiajun Wu, Cordelia Schmid, and Chen Sun. Learning temporal dynamics from cycles in narrated video. *arXiv preprint arXiv:2101.02337*, 2021.

[18] Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, C Lawrence Zitnick, et al. Visual storytelling. *arXiv preprint arXiv:1604.03968*, 2016.

[19] Chelsea Finn, Ian Goodfellow, and Sergey Levine. Unsupervised learning for physical interaction through video prediction. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 64–72, 2016.

[20] David F Fouhey, Wei-cheng Kuo, Alexei A Efros, and Jitendra Malik. From lifestyle vlogs to everyday interactions. In *CVPR*, 2018.

[21] David F. Fouhey, Weicheng Kuo, Alexei A. Efros, and Jitendra Malik. From lifestyle vlogs to everyday interactions. In *CVPR*, 2018.

[22] Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. Large-scale adversarial training for vision-and-language representation learning. *arXiv preprint arXiv:2006.06195*, 2020.

[23] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the V in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*, volume 1, page 9, 2017.

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[25] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 961–970, 2015.

[26] Jack Hessel, Bo Pang, Zhenhai Zhu, and Radu Soricut. A case study on combining ASR and visual features for generating instructional video captions. In *CoNLL*, November 2019.

[27] Jack Hessel, Zhenhai Zhu, Bo Pang, and Radu Soricut. Beyond instructional videos: Probing for more diverse visual-textual grounding on youtube. In *EMNLP*, 2020.

[28] Ari Holtzman, Jan Buys, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. *arXiv preprint arXiv:1904.09751*, 2019.

[29] De-An Huang, Joseph J. Lim, Li Fei-Fei, and Juan Carlos Niebles. Unsupervised visual-linguistic reference resolution in instructional videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[30] Ting-Hao Kenneth Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, C. Lawrence Zitnick, Devi Parikh, Lucy Vanderwende, Michel Galley, and Margaret Mitchell. Visual storytelling. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1233–1239, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1147. URL https://www.aclweb.org/anthology/N16-1147.

[31] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017). Honolulu, Hawaii*, pages 2680–8, 2017.

[32] Jianwen Jiang, Ziqiang Chen, Haojie Lin, Xibin Zhao, and Yue Gao. Divide and conquer: Question-guided spatio-temporal contextual attention for video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11101–11108, 2020.

[33] Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77, 2020.

[34] Seonhoon Kim, Seohyeong Jeong, Eun-Byul Kim, Inho Kang, and Nojun Kwak. Self-supervised pre-training and contrastive representation learning for multiple-choice video qa. *ArXiv*, abs/2009.08043, 2020.

[35] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. *arXiv preprint arXiv:2102.03334*, 2021.

[36] Kris M Kitani, Brian D Ziebart, James Andrew Bagnell, and Martial Hebert. Activity forecasting. In *European Conference on Computer Vision*, pages 201–214. Springer, 2012.

[37] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73, 2017.

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

[39] Harold W Kuhn. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97, 1955.

[40] Thao Minh Le, Vuong Le, Svetha Venkatesh, and Truyen Tran. Hierarchical conditional relation networks for video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9972–9981, 2020.

[41] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. Tvqa: Localized, compositional video question answering. In *EMNLP*, 2018.

[42] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. Tvqa+: Spatio-temporal grounding for video question answering. In *Tech Report, arXiv*, 2019.

[43] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. What is more likely to happen next? video-and-language future event prediction. *arXiv preprint arXiv:2010.07999*, 2020.

[44] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. *arXiv preprint arXiv:2102.06183*, 2021.

[45] Gen Li, Nan Duan, Yuejian Fang, Ming Gong, Daxin Jiang, and Ming Zhou. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *AAAI*, pages 11336–11344, 2020.

[46] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557*, 2019.

[47] Wei Li, Can Gao, Guocheng Niu, Xinyan Xiao, Hao Liu, Jiachen Liu, Hua Wu, and Haifeng Wang. UNIMO: towards unified-modal understanding and generation via cross-modal contrastive learning. *arXiv preprint arXiv:2012.15409*, 2020.

[48] 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 *European conference on computer vision*, pages 740–755. Springer, 2014.

[49] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[50] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[51] Cewu Lu, Ranjay Krishna, Michael Bernstein, and Li Fei-Fei. Visual relationship detection with language priors. In *European Conference on Computer Vision*, 2016.

[52] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems*, pages 13–23, 2019.

[53] Jonathan Malmaud, Jonathan Huang, Vivek Rathod, Nick Johnston, Andrew Rabinovich, and Kevin Murphy. What's cookin'? interpreting cooking videos using text, speech and vision. In *NAACL*, 2015.

[54] Alice E Marwick and danah boyd. Networked privacy: How teenagers negotiate context in social media. *New media & society*, 16(7):1051–1067, 2014.

[55] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In *ICCV*, 2019.

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

[57] Ishan Misra, C Lawrence Zitnick, and Martial Hebert. Shuffle and learn: unsupervised learning using temporal order verification. In *European Conference on Computer Vision*, pages 527–544. Springer, 2016.

[58] Yasufumi Moriya, Ramon Sanabria, Florian Metze, and Gareth JF Jones. Grounding object detections with transcriptions. *arXiv preprint arXiv:1906.06147*, 2019.

[59] Meinard Müller. Dynamic time warping. *Information retrieval for music and motion*, pages 69–84, 2007.

[60] Shruti Palaskar, Jindrich Libovickỳ, Spandana Gella, and Florian Metze. Multimodal abstractive summarization for how2 videos. *arXiv preprint arXiv:1906.07901*, 2019.

[61] Alessandro Prest, Christian Leistner, Javier Civera, Cordelia Schmid, and Vittorio Ferrari. Learning object class detectors from weakly annotated video. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3282–3289. IEEE, 2012.

[62] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019.

[63] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

[64] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in neural information processing systems*, pages 91–99, 2015.

[65] Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Chris Pal, Hugo Larochelle, Aaron Courville, and Bernt Schiele. Movie description. *International Journal of Computer Vision*, 2017. URL http://link.springer.com/article/10.1007/s11263-016-0987-1?wt_mc=Internal.Event.1.SEM.ArticleAuthorOnlineFirst.

[66] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4510–4520, 2018.

[67] Oleksandr Savsunenko. How tensorflow's tf.image.resize stole 60 days of my life. Technical report, Hacker Noon.

[68] Roger C. Schank and Robert P. Abelson. Scripts, plans, and knowledge. In *Proceedings of the 4th International Joint Conference on Artificial Intelligence - Volume 1*, IJCAI'75, pages 151–157, San Francisco, CA, USA, 1975. Morgan Kaufmann Publishers Inc. URL http://dl.acm.org/citation.cfm?id=1624626.1624649.

[69] Ozan Sener, Amir R Zamir, Silvio Savarese, and Ashutosh Saxena. Unsupervised semantic parsing of video collections. In *ICCV*, 2015.

[70] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, 2016.

[71] Paul Hongsuck Seo, Arsha Nagrani, and Cordelia Schmid. Look before you speak: Visually contextualized utterances. *arXiv preprint arXiv:2012.05710*, 2020.

[72] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, 2018.

[73] Botian Shi, Lei Ji, Yaobo Liang, Nan Duan, Peng Chen, Zhendong Niu, and Ming Zhou. Dense procedure captioning in narrated instructional videos. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6382–6391, 2019.

[74] Wei Shi and Vera Demberg. Next sentence prediction helps implicit discourse relation classification within and across domains. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5794–5800, 2019.

[75] Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in nlp. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, 2019.

[76] Alane Suhr, Stephanie Zhou, Ally Zhang, Iris Zhang, Huajun Bai, and Yoav Artzi. A corpus for reasoning about natural language grounded in photographs. *arXiv preprint arXiv:1811.00491*, 2018.

[77] Chen Sun, Fabien Baradel, Kevin Murphy, and Cordelia Schmid. Contrastive bidirectional transformer for temporal representation learning. *arXiv preprint arXiv:1906.05743*, 2019.

[78] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. VideoBERT: A joint model for video and language representation learning. In *ICCV*, 2019.

[79] Hao Tan and Mohit Bansal. LXMERT: Learning cross-modality encoder representations from transformers. In *EMNLP*, 2019.

[80] Atousa Torabi, Niket Tandon, and Leon Sigal. Learning language-visual embedding for movie understanding with natural-language. *arXiv preprint*, 2016. URL http://arxiv.org/pdf/1609.08124v1.pdf.

[81] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[82] Jacob Walker, Carl Doersch, Abhinav Gupta, and Martial Hebert. An uncertain future: Forecasting from static images using variational autoencoders. In *European Conference on Computer Vision*, pages 835–851. Springer, 2016.

[83] Donglai Wei, Joseph J Lim, Andrew Zisserman, and William T Freeman. Learning and using the arrow of time. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8052–8060, 2018.

[84] Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706*, 2019.

[85] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning for video understanding. *arXiv preprint arXiv:1712.04851*, 1(2):5, 2017.

[86] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1645–1653, 2017.

[87] Antoine Yang, Antoine Miech, Josef Sivic, Ivan Laptev, and Cordelia Schmid. Just ask: Learning to answer questions from millions of narrated videos. *arXiv preprint arXiv:2012.00451*, 2020.

[88] Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. Ernie-vil: Knowledge enhanced vision-language representations through scene graph. *arXiv preprint arXiv:2006.16934*, 2020.

[89] Shoou-I Yu, Lu Jiang, and Alexander Hauptmann. Instructional videos for unsupervised harvesting and learning of action examples. In *ACM MM*, 2014.

[90] Youngjae Yu, Jongseok Kim, and Gunhee Kim. A joint sequence fusion model for video question answering and retrieval. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 471–487, 2018.

[91] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. ActivityNet-QA: a dataset for understanding complex web videos via question answering. In *AAAI*, pages 9127–9134, 2019.

[92] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6720–6731, 2019.

[93] Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi. Defending against neural fake news. In *Advances in Neural Information Processing Systems 32*, 2019.

[94] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. *arXiv preprint arXiv:2101.00529*, 2021.

[95] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. Contrastive learning of medical visual representations from paired images and text. *arXiv preprint arXiv:2010.00747*, 2020.

[96] Linchao Zhu and Yi Yang. ActBERT: Learning global-local video-text representations. In *CVPR*, 2020.

[97] Yueting Zhuang, Dejing Xu, Xin Yan, Wenzhuo Cheng, Zhou Zhao, Shiliang Pu, and Jun Xiao. Multichannel attention refinement for video question answering. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 16(1s):1–23, 2020.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Section 6.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] See Section 6.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We plan to release our model's code either via the supplementary material or on `github` after acceptance (but prior to publication).
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] We will be including per-dataset discussion of hyperparameters in Appendix.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] Many of the datasets we used adopt a leaderboard style format where multiple submissions are not possible, so we were unable to do this.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [No] All the datasets and code used are public and released under licenses managed by the original creators.
   (c) Did you include any new assets either in the supplemental material or as a URL? [No] But, we are planning to release our model upon (or before) acceptance.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [No] We discuss our perspective on this in Section 6.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A] We didn't do any new crowdworking/human participant experiments for the described work.
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]