MERLOT RESERVE: Neural Script Knowledge through Vision and Language and Sound

Rowan Zellers Jiasen Lu Ximing Lu Youngjae Yu Yanpeng Zhao
Mohammadreza Salehi Aditya Kusupati Jack Hessel Ali Farhadi Yejin Choi

Paul G. Allen School of Computer Science & Engineering, University of Washington
Allen Institute for Artificial Intelligence University of Edinburgh
rowanzellers.com/merlotreserve

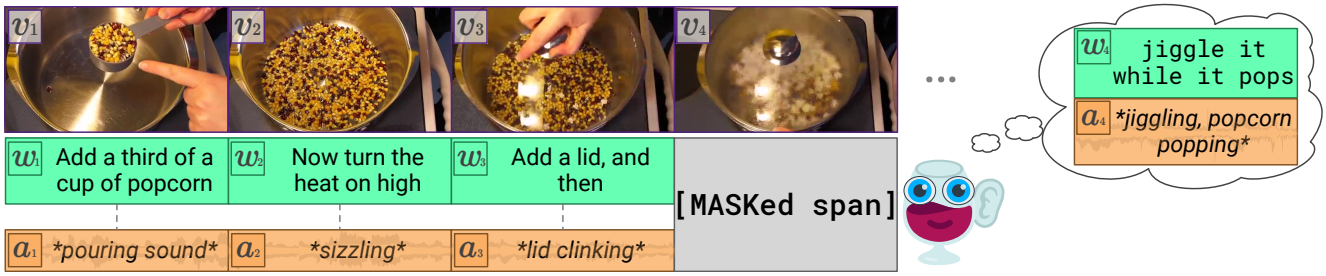

Figure 1: MERLOT RESERVE learns *multimodal neural script knowledge* representations of video – jointly reasoning over video frames, text, and audio. Our model is pretrained to predict which snippet of text (and audio) might be hidden by the MASK. This task enables it to perform well on a variety of vision-and-language tasks, in both zero-shot and finetuned settings.

Abstract

As humans, we navigate a multimodal world, building a holistic understanding from all our senses. We introduce MERLOT RESERVE, *a model that represents videos jointly over time – through a new training objective that learns from audio, subtitles, and video frames. Given a video, we replace snippets of text and audio with a* MASK *token; the model learns by choosing the correct masked-out snippet. Our objective learns faster than alternatives, and performs well at scale: we pretrain on 20 million YouTube videos.*

Empirical results show that MERLOT RESERVE *learns strong multimodal representations. When finetuned, it sets state-of-the-art on Visual Commonsense Reasoning (VCR), TVQA, and Kinetics-600; outperforming prior work by 5%, 7%, and 1.5% respectively. Ablations show that these tasks benefit from audio pretraining – even VCR, a QA task centered around images (without sound). Moreover, our objective enables out-of-the-box prediction, revealing strong multimodal commonsense understanding. In a fully zero-shot setting, our model obtains competitive results on four video tasks, even outperforming supervised approaches on the recently proposed Situated Reasoning (STAR) benchmark.*

We analyze why audio enables better vision-language representations, suggesting significant opportunities for future research. We conclude by discussing ethical and societal implications of multimodal pretraining.

1. Introduction

The world around us is dynamic. We experience and learn from it using all of our senses, reasoning over them temporally through *multimodal script knowledge* [99, 128]. Consider Figure 1, which depicts someone cooking popcorn. From the images and dialogue alone, we might be able to imagine what *sounds* of the scene are: the process might begin with raw kernels scattering in an empty, metallic pot, and end with the dynamic 'pops' of popcorn expanding, along with the jiggling of a metal around the stove.

Predicting this sound is an instance of *learning from reentry*: where time-locked correlations enable one modality to educate others. Reentry has been hypothesized by developmental psychologists to be crucial for how we as humans learn visual and world knowledge, much of it without need for an explicit teacher [89, 35, 20, 100]. Yet, we ask – can we build machines that likewise learn vision, language, and sound *together*? And can this paradigm enable learning *neural script knowledge*, that transfers to language-and-vision tasks, *even those without sound*?

In this work, we study these questions, and find that the answers are '**yes**.' We introduce a new model that learns self-supervised representations of videos, through all their modalities (audio, subtitles, vision). We dub our model MERLOT RESERVE[1], henceforth RESERVE for short.

[1]Short for **M**ultimodal **E**vent **R**epresentation **L**earning **O**ver **T**ime, with **RE**-entrant **S**up**ERV**ision of **E**vents.

Our model differs from past work that learns from audio-image pairs [54, 71], from subtitled videos [105, 128], or from static images with literal descriptions [106, 21, 92]. Instead, we learn joint representations from *all modalities of a video*, using each modality to teach others. We do this at scale, training on over 20 million YouTube videos.

We introduce a new *contrastive masked span* learning objective to learn script knowledge across modalities. It generalizes and outperforms a variety of previously proposed approaches (e.g. [29, 106, 92, 128]), while enabling audio to be used as signal. The idea is outlined in Figure 1: the model must figure out which span of text (or audio) was `MASK`ed out of a video sequence. We combine our objective with a second contrastive learning approach, tailored to learning *visual recognition* from scratch: the model must also match each video frame to a contextualized representation of the video's transcript [128]. Through ablations, we show that our framework enables rapid pretraining of a model and readily scales to 'large' transformer sizes (of 644M parameters).

Experimental results show that RESERVE learns powerful representations, useful even for tasks posed over only a few of the studied modalities. For example, when finetuned on Visual Commonsense Reasoning [126] (a vision+language task with no audio), it sets a new state-of-the-art, outperforming models trained on supervised image-caption pairs by **over 5%**. It does even better on video tasks: fine-tuning without audio, it outperforms prior work on TVQA [75] by a margin of **over 7%** (and given TVQA audio, performance increases even further). Finally, audio enables 91.1% accuracy on Kinetics-600 [19]. These performance improvements do not come at the expense of efficiency: our largest model uses one-fifths the FLOPs of a VisualBERT.

RESERVE also performs well in zero-shot settings. We evaluate on four diverse benchmarks: Situated Reasoning (STAR) [119], EPIC-Kitchens [26], LSMDC-FiB [96], and MSR-VTT QA [120]. These benchmarks require visual reasoning with respective emphasis on *temporality*, *future prediction*, and both *social* and *physical understanding*. With no fine-tuning or supervision, our model obtains competitive performance on each. Of note, it nearly doubles [123]'s SoTA zero-shot accuracy on MSR-VTT QA, and it outperforms supervised approaches (like ClipBERT [74]) on STAR.

Finally, we investigate *why*, and *on which training instances* audio-powered multimodal pretraining particularly helps. For instance, predicting audio rewards models for recognizing *dynamic state changes* (like cooked popcorn) and *human communication dynamics* (what are people's emotions and towards whom). Our model progressively learns these phenomena as pretraining progresses. These signals are often orthogonal to what snippets of text provide, which motivates learning from both modalities.

In summary, our key contributions are the following:

a. RESERVE, a model for multimodal script knowledge, fusing vision, audio, and text.

b. A new contrastive span matching objective, enabling our model to learn from text *and audio* self-supervision.

c. Experiments, ablations, and analysis, that demonstrate strong multimodal video representations.

Overall, the results suggest that learning representations from *all modalities* – in a time-locked, reentrant manner – is a promising direction, and one that has significant space for future work. We release code and model checkpoints at `rowanzellers.com/merlotreserve`.

2. Related Work

Our work brings together two active lines of research.

Joint representations of multiple modalities. Many language-and-vision tasks benefit from *early fusion* of the modalities [6]. A family of 'VisualBERT' models have been proposed for this: typically, these use a supervised object detector image encoder backbone, and pretrain on images paired with literal captions [106, 77, 81, 21, 124, 74]. Cross-modal interactions are learned in part through a *masked language modeling* (mask LM) objective [29], where subwords are replaced with '`MASK`', and models independently predict each subword conditioned on both images and unmasked tokens.[2]

Perhaps closest to our work is MERLOT [128], which learns a joint vision-text model from web videos with automatic speech recognition (ASR). Through a combination of objectives (including a variant of mask LM), MERLOT established strong results on a variety of video QA benchmarks when finetuned. However, it lacks audio: it is limited to representing (and learning from) video frames paired with subtitles. Our proposed RESERVE, which represents and learns from audio, outperforms MERLOT.

Co-supervision between modalities. A common pitfall when training a joint multimodal model is that complex *inter-modal* interactions can be ignored during learning, in favor of simpler *intra-modal* interactions [51, 24, 59]. For example, when using the aforementioned mask LM objective, models can *ignore visual input completely* in favor of text-text interactions [13]; this issue is magnified when training on videos with noisy ASR text [128].

A line of recent work thus learns independent modality-specific encoders, using objectives that cannot be shortcutted with simple intra-modal patterns. Models like CLIP learn image classification by matching images with their captions, contrastively [132, 92, 63]. Recent work has explored this paradigm for matching video frames with their transcripts [121], with their audio signal [97, 114], or both [3, 2]; these

[2]Recent papers propose extensions, like generating masked-out spans [22] or text [78, 116], but it is unclear whether they can outperform the VisualBerts on vision-language tasks like VCR [126]. Another extension involves learning from text-to-speech audio in a captioning setting [62, 79] – yet this lacks key supervision for environmental sounds and emotive speech.

works likewise perform well on single-modality tasks like audio classification and activity recognition. These independent encoders can be combined through late fusion [97], yet late fusion is strictly less expressive than our proposed joint encoding (early fusion) approach.

Our work combines both lines of research. We learn a model for jointly representing videos, through all their modalities, and train it using a new learning objective that enables *co-supervision* between modalities.

3. Model: MERLOT RESERVE

In this section, we present MERLOT RESERVE, including: our model architecture (3.1), new pretraining objectives (3.2), and pretraining video dataset (3.3). At a high level, MERLOT RESERVE represents a video by fusing its constituent modalities (vision, audio, and text from transcribed speech) together, and over time. These representations enable both finetuned and zero-shot downstream applications.

More formally, we split a video $\mathcal{V}$ into a sequence of non-overlapping segments in time $\{s_t\}$. Each segment has:

a. A frame $\boldsymbol{v}_t$, from the middle of the segment,

b. The ASR tokens $\boldsymbol{w}_t$ spoken during the segment,

c. The audio $\boldsymbol{a}_t$ of the segment.

Segments default to 5 seconds in length; we discuss details of how we split videos into segments in Appendix C.

As the text $\boldsymbol{w}_t$ was automatically transcribed by a model given audio $\boldsymbol{a}_t$, it is reasonable to assume that it contains strictly less information content.[3] Thus, for each segment $\boldsymbol{s}_t$, we provide models with exactly one of text *or* audio. We will further *mask out* portions of the text and audio during pretraining, to challenge models to recover what is missing.

3.1. Model architecture

An overview of MERLOT RESERVE is shown in Figure 2. We first pre-encode each modality independently (using a Transformer [110] or images/audio; a BPE embedding table for text). We then learn a joint encoder to fuse all representations, together and over time.

Image encoder. We use a Vision Transformer (ViT; [34]) to encode each frame independently. We use a patch size of 16 and apply a 2x2 query-key-value attention pool after the Transformer, converting an image of size $H \times W$ into a $H/32 \times W/32$ feature map of dimension d_h.

Audio encoder. We split the audio in each segment $\boldsymbol{a}_t$ into three equal-sized *subsegments*, for compatibility with the lengths at which we mask text (Appendix C). We use an Audio Spectrogram Transformer to encode each subsegment independently [47]. The three feature maps are concatenated; the result is of size $18 \times d_h$ for every 5 seconds of audio.

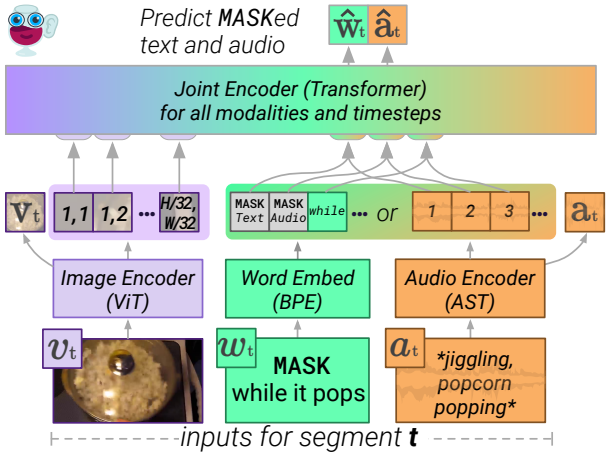

Figure 2: MERLOT RESERVE architecture. We provide sequence-level representations of video frames, and *either* words or audio, to a joint encoder. The joint encoder contextualizes over modalities and timesteps, to predict what is behind MASK for audio $\widehat{\mathbf{a}}_t$ and text $\widehat{\mathbf{w}}_t$. We supervise these predictions with independently encoded targets: $\mathbf{a}_t$ from the audio encoder, and $\mathbf{w}_t$ from a separate text encoder (not shown).

Joint encoder. Finally, we jointly encode all modalities (over all input video segments) using a bidirectional Transformer. We use a linear projection of the final layer's hidden states for all objectives (e.g. $\widehat{\mathbf{w}}_t$ and $\widehat{\mathbf{a}}_t$).

Independently-encoded targets. We will supervise the joint encoder by simultaneously learning independently-encoded 'target' representations for each modality. Doing this is straightforward for the image and audio encoders: we add a CLS to their respective inputs, and extract the final hidden state $\mathbf{v}_t$ or $\mathbf{a}_t$ at that position. For text, we learn a separate bidirectional Transformer *span encoder*, which computes targets $\mathbf{w}_t$ from a CLS and embedded tokens of a candidate text span. This enables zero-shot prediction (4.4).

Architecture sizes. We consider two model sizes in this work, which we pretrain from random initialization:

1. MERLOT RESERVE-B, with a hidden size of 768, a 12-layer ViT-B/16 image encoder, and a 12-layer joint encoder.
2. MERLOT RESERVE-L, with a hidden size of 1024, a 24-layer ViT-L/16 image encoder, and a 24-layer joint encoder.

We always use a 12-layer audio encoder, and a 4-layer text span encoder. Details are in Appendix B.

3.2. Contrastive Span Training

We introduce *contrastive span* training, which enables learning across and between the three modalities. As shown in Figure 3, the model is given a sequence of video segments.

[3] Despite being derived from the audio, pretraining with text is still paramount: 1) in §3.2 we discuss how jointly modeling audio+text prevents models from shortcutting pretraining objectives via surface correlations; 2) in §4.2 we show that incorporating both transcripts and audio during fine-tuning improves performance; and 3) a textual interface to the model is required for downstream vision+language with textual inputs.

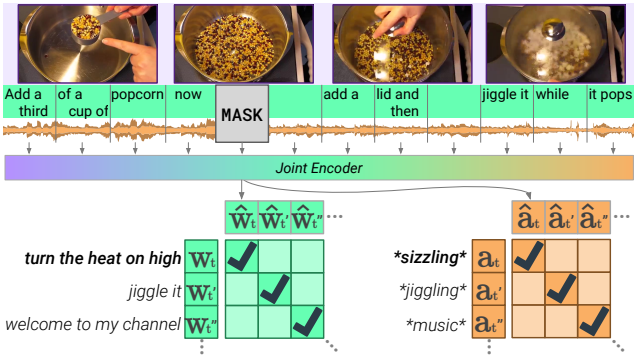

Figure 3: Contrastive span training. Given a video with all modalities temporally aligned, we MASK out a region of text and audio. The model must maximize its similarity *only to* an independent encoding of the text $\mathbf{w}_t$ and audio $\mathbf{a}_t$.

For each one, we include the video frame, and then three 'sub-segments' that are each either text *or* audio. The subdivided audio segments are encoded independently by the Audio Encoder, before being fused by the Joint Encoder. We train by replacing 25% of these text and audio subsegments with a special MASK token. The model must match the representation atop the MASK *only with* an independent encoding of its span.

Our approach combines past success at matching images to their captions [92, 63] along with 'VisualBERT'-style prediction of independent tokens [106, 21] – though, crucially, we predict representations at a higher-level semantic unit than individual tokens. Our approach also enables the model to learn from both audio and text, while discouraging *memorization* of raw perceptual input, or tokens – which can harm representation quality [112].

Formally, we minimize the cross entropy between the MASKed prediction $\hat{\mathbf{w}}_t$ and its corresponding phrase representation $\mathbf{w}_t$, versus others in the batch $\mathcal{W}$:

$$\mathcal{L}_{\text{mask}\rightarrow\text{text}} = \frac{1}{|\mathcal{W}|}\sum_{\mathbf{w}_t\in\mathcal{W}}\left(\log\frac{\exp(\sigma\hat{\mathbf{w}}_t\cdot\mathbf{w}_t)}{\sum_{\mathbf{w}\in\mathcal{W}}\exp(\sigma\hat{\mathbf{w}}_t\cdot\mathbf{w})}\right). \quad (1)$$

We first L^2-normalize $\mathbf{w}$ and $\hat{\mathbf{w}}$, and scale their dot product with a parameter σ [92].[4] We then add this to its transposed version $\mathcal{L}_{\text{text}\rightarrow\text{mask}}$, giving us our text-based loss $\mathcal{L}_{\text{text}}$. Analogously, we define $\mathcal{L}_{\text{audio}}$ for audio, between the MASKed prediction $\hat{\mathbf{a}}_t$ and its target $\mathbf{a}_t$, versus others $\mathbf{a}$ in the batch.

In addition to these masked text and audio objectives, we simultaneously train the model to match video frames with a contextualized encoding of the transcript.[5] Here, the joint encoder encodes the entire video's transcript at once, extracting a single hidden representation per segment $\hat{\mathbf{v}}_t$. We use the same contrastive setup as Equation 1 to maximize the similarity of these vectors with the corresponding $\mathbf{v}_t$ vectors from the frames, giving us a symmetric frame-based loss $\mathcal{L}_{\text{frame}}$. The final loss is the sum of the component losses:

$$\mathcal{L} = \mathcal{L}_{\text{text}} + \mathcal{L}_{\text{audio}} + \mathcal{L}_{\text{frame}}. \quad (2)$$

Avoiding shortcut learning. Early on, we observed that training a model to predict a *perceptual* modality (like audio or vision) given input from *the same modality*, led to shortcut learning – a low training loss, but poor representations. We hypothesize that this setup encourages models to learn imperceptible features, like the exact model of the microphone, or the chromatic aberration of the camera lens [33]. We avoid this, while still using audio as a target, by simultaneously training on two kinds of masked videos:

i. **Audio only as target.** We provide only video frames and subtitles. The model produces representations of both *audio* and *text* that fill in MASKed blanks.

ii. **Audio as input.** We provide the model video frames, and subtitles *or audio* at each segment. Because the model is given audio as an input somewhere, the model only produces representations for MASKed *text*.

Another issue is that YouTube's captions are not perfectly time-aligned with the underlying audio. During our initial exploration, models took ready advantage of this shortcut: for instance, predicting an audio span based on what adjacent (overlapping) words sound like. We introduce a masking algorithm to resolve this; details in Appendix C.

Pretraining setup. We train on TPU v3-512 accelerators; training takes 5 days for RESERVE-B, and 16 days for RESERVE-L. We made pretraining more efficient through several algorithmic and implementation improvements. Of note, we simultaneously train on written (web) text, which enables more text candidates to be used. We use a batch size of 1024 videos, each with N=16 segments (split into two groups of 8 segments each). We use AdamW [69, 80] to minimize Equation 2. More details and hyperparameters are in Appendix B.

3.3. Pretraining Dataset

Recent prior work on static images that demonstrates empirical improvements by increasing dataset size – all the way up to JFT-3B [70, 34, 92, 130]. The same pattern emerges in videos: prior work that has shown promising empirical improvements not only by scaling to 6 million videos/180M frames [128], but also by collecting a diverse set (i.e., going beyond instructional videos [60]).

To this end, we introduce a new training dataset of 20 million English-subtitled YouTube videos, and 1 billion frames, called YT-Temporal-1B. At the same time, we take steps to protect user privacy, directing scraping towards public, large, and monetized channels. We detail our collection, preprocessing, and release strategy in Appendix E.

[4] Following past work, we optimize σ and clip it at 100, which enables the model to 'warm-up' its emphasis placed on hard negatives [92, 113].

[5] In MERLOT [128], this objective was found to be critical for learning visual recognition from self-supervised videos.

4. Experiments

In this section, we present model ablations (4.1.1), and show that a finetuned RESERVE obtains state-of-the-art results on VCR (4.1.2), TVQA (4.2), and Kinetics-600 (4.3). We then show that our model has strong zero-shot capability, over four challenging zero-shot tasks (4.2).

4.1. Visual Commonsense Reasoning (VCR)

We evaluate RESERVE first through finetuning on VCR [126]. Most competitive models for VCR are pretrained exclusively on images paired with captions, often with supervised visual representations (e.g. from an object detector). To the best of our knowledge, the only exception is MERLOT [128], which uses YouTube video frames and text as part of pretraining; no VCR model to date was pretrained on audio.

VCR Task. A model is given an image from a movie, and a question. The model must choose the correct answer given four multiple choice options ($Q \rightarrow A$); it then is given four *rationales* justifying the answer, and it must choose the correct one ($QA \rightarrow R$). The results are combined with a $Q \rightarrow AR$ metric, where a model must choose the right answer *and then* the right rationale, to get the question 'correct.'

Finetuning approach. We follow [128]'s approach: 'drawing on' VCR's detection tags onto the image, and jointly finetuning on $Q \rightarrow A$ and $QA \rightarrow R$. For both subproblems, we learn by scoring each $Q \rightarrow A$ (or $QA \rightarrow R$) option independently. We pool a hidden representation from a `MASK` inserted after the text, and pass this through a newly-initialized linear layer to extract a logit, which we optimize through cross-entropy (details in Appendix D.1.1.)

4.1.1 Ablations: contrastive learning with audio helps.

While we present our final, state-of-the-art VCR performance in 4.1.2, we first use the corpus for an ablation study. We use the same architecture and data throughout, allowing apples-to-apples comparison between modeling decisions. We start with a similar configuration to MERLOT [128] and show that contrastive span training improves further, particularly when we add audio.

Contrastive Span helps for *V*ision+*T*ext modeling. We start by comparing pretraining objectives for learning from YouTube ASR and video alone:

a. Mask LM. This objective trains a bidirectional model by having it *independently* predict masked-out tokens. We make this baseline as strong as possible by using SpanBERT-style masking [64], where text spans are masked out (identical to our *contrastive spans*). Each span $\boldsymbol{w}$ is replaced by a `MASK` token, and we predict each of its subwords w_i independently.[6]

[6]Like [64], we concatenate the `MASK`'s hidden state with a position embedding for index i, pass the result through a two-layer MLP, and use tied embedding weights to predict w_i.

	Configuration *for one epoch of pretraining*	VCR Q→A	val (%)
V+T	Mask LM [29, 106, 128]	67.2	
V+T	VirTex-style [27]	67.8	
V+T	Contrastive Span	**69.7**	
V+T+A	Audio as target	70.4	
V+T+A	Audio as input and target	**70.7**	
V+T+A	Audio as input and target, w/o strict localization	70.6	
	RESERVE-B	**71.9**	

Table 1: **Ablation study** of our contrastive span objective. It outperforms prior work in a Vision+Text setting, with a 1% boost when audio is added. Our full setup, adding written text, improves another 1%. denotes part of our full model.

b. VirTex [27]. In this objective, we likewise mask text subsegments and extract their hidden states. The difference is that we sequentially predict tokens $w_i \in \boldsymbol{w}$, using a left-to-right language model (LM) with the same architecture details as our proposed span encoder.

Results are in Table 1. Versus these approaches, our contrastive span objective boosts performance by over 2%, after one epoch of pretraining only on vision and text. We hypothesize that its faster learning is caused by encouraging models to learn concept-level span representations; this might not happen when predicting tokens individually [23].

***A*udio pretraining helps**, even for the audio-less VCR:

d. Audio as target. Here, the model is only given video frames and ASR text as input. In addition to performing contrastive-span pretraining over the missing text spans, it does the same over the (held-out) audio span (Equation 2. This boosts VCR accuracy by 0.7%.

e. Audio as input and target. The model does the above (for video+text input sequences), and simultaneously is given video+text+audio sequences, wherein it must predict missing text. This boosts accuracy by 1% in total.

f. Sans strict localization. We evaluate the importance of our strict localization in time. Here, in addition to correct subsegments at the *true* position t as a correct match, we count adjacent `MASK`ed out regions as well. An extreme version of this was proposed by [49], where a positive match can be of any two frames in a video. Yet even in our conservative implementation, performance drops slightly, suggesting localization helps.

Putting these all together, we find that contrastive span pretraining outperforms mask LM, with improved performance when audio is used **both as input and target**. For our flagship model, we report results in Table 1 on simultaneously training on web-text sequences as well (Appendix C.4), this improves performance by an additional 1%.

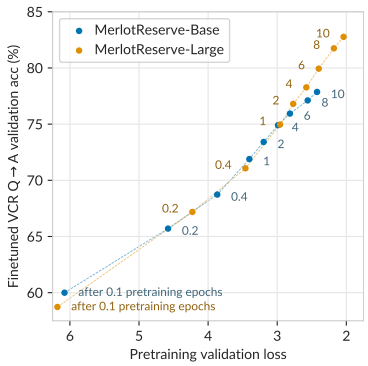

Figure 4: **Pretraining progress**: performance on contrastive-span pretraining, vs. fine-tuned VCR validation accuracy. Pretraining RESERVE-B for 9 more epochs boosts performance by 5%; L by 8%.

	Model	VCR test (acc; %) Q→A	QA→R	Q→AR
Caption/ObjDet-based	ERNIE-ViL-Large [124]	79.2	83.5	66.3
	Villa-Large [39]	78.9	83.8	65.7
	UNITER-Large [21]	77.3	80.8	62.8
	Villa-Base [39]	76.4	79.1	60.6
	VilBERT [81]	73.3	74.6	54.8
	B2T2 [4]	72.6	75.7	55.0
	VisualBERT [77]	71.6	73.2	52.4
Video-based	MERLOT [128]	80.6	80.4	65.1
	RESERVE-B	79.3	78.7	62.6
	RESERVE-L	**84.0**	**84.9**	**72.0**

Table 2: RESERVE gets **state-of-the-art leaderboard performance on VCR**. We compare it with the largest submitted single models, including image-caption models that utilize heavy manual supervision (e.g. object detections and captions).

	Model	TVQA (acc; %) Val	Test
	Human [75]	–	89.4
Subtitles	MERLOT [128]	78.7	78.4
	MMFT-BERT [109]	73.5	72.8
	Kim et al [68]	76.2	76.1
	RESERVE-B	82.5	–
	RESERVE-L	**85.9**	**85.6**
Audio	RESERVE-B	81.3	–
	RESERVE-L	**85.6**	**84.8**
Both	RESERVE-B	83.1	82.7
	RESERVE-L	**86.5**	**86.1**

Table 3: RESERVE gets state-of-the-art results on TVQA by **over 7%**, versus prior work (that cannot make use of audio).

4.1.2 VCR Results

Encouraged by these results, we train our models for 10 epochs on YT-Temporal-1B. Figure 4 demonstrates that fine-tuned VCR performance tracks with the number of pretraining epochs, as well as the validation loss.[7]

Finally, in Table 2, we compare RESERVE against the largest published models from the VCR leaderboard. Of note, RESERVE-L outperforms all prior work, by **over 5%** on Q→AR metric. It outperforms even large ensembles (e.g. 15 ERNIE-Large's) submitted by industry [124], though we do not show these on this table to focus on only single models.

Efficiency. The accuracy increase of RESERVE is not simply due to compute.[8] In fact, our RESERVE-L requires *one-fifth the FLOPs* of detector-based systems, like UNITER-Large [21] (Appendix B.3). Moreover, because RESERVE-L uses a pure ViT backbone versus MERLOT's ViT-ResNet hybrid, it uses fewer FLOPs than MERLOT, while scoring 7% higher. Meanwhile, RESERVE-B outperforms 'base' detector-based models, while using *less than one-tenth their FLOPs*.

In terms of parameter count, RESERVE-B is comparable to prior work. On VCR, including the vision stack, RESERVE-B has 200M finetunable parameters and performs similarly to the 378M parameter UNITER-Large. RESERVE-L has 644M parameters.

4.2. Finetuning on TVQA

Next, we use TVQA [75] to evaluate our model's capacity to transfer to multimodal video understanding tasks. In TVQA, models are given a video, a question, and five answer choices. The scenes come from American TV shows, and depict characters interacting with each other through dialogue – which past work represents through subtitles.

Audio-Subtitle Finetuning. To evaluate how much audio can help for TVQA, we finetune RESERVE jointly between the 'Subtitles' and 'Audio' settings. Like on VCR, we consider one sequence per candidate: each contains video frame features, the question, the answer candidate, and a `MASK` token (from where we pool a hidden representation). During training, each sequence is duplicated: we provide one sequence with *subtitles* from the video, and for the other, we use *audio*. This lets us train a single model, and then test how it will do *given subtitles*, *given audio*, or *given both* (by averaging the two softmax predictions).

Results. We show TVQA results in Table 3. With subtitles and video frames alone, our RESERVE-B outperforms all prior work by over 3%. Combining subtitle-only and audio-only predictions performs even better, improving over 4% versus the prior state-of-the-art, MERLOT (and in turn over other models). The same pattern holds (with additional performance gains) as model size increases: RESERVE-L improves over prior work by **7.6%**.

4.3. Finetuning on Kinetics-600 Activity Recognition

Next, we use Kinetics-600 [19] to compare our model's (finetuned) activity understanding versus prior work, including many top-scoring models that do not integrate audio. The task is to classify a 10-second video clip as one of 600 categories. We finetune RESERVE jointly over two settings: vision only, and vision+audio.

Results. We show Kinetics-600 results on the validation set, in Table 4. RESERVE improves by **1.7%** when it can jointly represent the video's frames with its sound. This enables it to outperform other large models, including VATT

[7] The plot suggests that if we pretrained longer, VCR performance might continue to increase, though a confounding factor might be the learning-rate schedule. With access to compute beyond our current capacity, future work would be well-suited to consider this and other pre-training modifications.

[8] Here, we use FLOPs as our key efficiency metric, as they are a critical bottleneck in model scaling [66, 34, 130]. On the other hand, we argue that parameter count can be misleading – for instance, many Transformer parameters can be tied together with minimal performance loss [72].

	Model	Kinetics-600 (%) Top-1	Top-5
Vision Only	VATT-Base[2]	80.5	95.5
	VATT-Large [2]	83.6	96.6
	TimeSFormer-L [9]	82.2	95.6
	Florence [125]	87.8	97.8
	MTV-Base [122]	83.6	96.1
	MTV-Large [122]	85.4	96.7
	MTV-Huge [122]	89.6	98.3
	RESERVE-B	88.1	95.8
	RESERVE-L	89.4	96.3
+Audio	RESERVE-B	89.7	96.6
	RESERVE-L	**91.1**	**97.1**

Table 4: RESERVE gets state-of-the-art results on Kinetics-600 by **1.5%** versus standard approaches (that cannot make use of audio).

	Model	Situated Reasoning (STAR) (test acc; %) Interaction	Sequence	Prediction	Feasibility	Overall	EPIC-Kitchens (val class-mean R@5; %) Verb	Noun	Action	LSMDC (FiB test %) Acc	MSR-VTT QA (test acc %) top1	top5
	Supervised SoTA	ClipBERT [74] 39.8	43.6	32.3	31.4	36.7	AVT+ [46] 28.2	32.0	15.9	52.9	MERLOT [128] 43.1	
zero-shot	Random	25.0	25.0	25.0	25.0	25.0	6.2	2.3	0.1	0.1	0.1	0.5
	CLIP (VIT-B/16) [92]	39.8	40.5	35.5	36.0	38.0	16.5	12.8	2.3	2.0	3.0	11.9
	CLIP (RN50x16) [92]	39.9	41.7	36.5	**37.0**	38.7	13.4	14.5	2.1	2.3	2.3	9.7
	Just Ask (ZS)[123]										2.9	8.8
	RESERVE-B	44.4	40.1	38.1	35.0	39.4	17.9	15.6	2.7	26.1	3.7	10.8
	RESERVE-L	42.6	41.1	37.4	32.2	38.3	15.6	19.3	4.5	26.7	4.4	11.5
	RESERVE-B (+audio)	**44.8**	**42.4**	**38.8**	36.2	**40.5**	20.9	17.5	3.7	29.1	4.0	12.0
	RESERVE-L (+audio)	43.9	42.6	37.6	33.6	39.4	**23.2**	**23.7**	**4.8**	**31.0**	**5.8**	**13.6**

Table 5: Zero shot results. On STAR, RESERVE obtains state-of-the-art results, outperforming finetuned video models. It performs well on EPIC-Kitchens (verb and noun forecasting), along with LSMDC, despite their long-tail distributions. On MSR-VTT QA, it outperforms past work on weakly-supervised video QA. Further, it outperforms CLIP (that cannot handle dynamic situations), and benefits from audio when given.

[2] which learns to represent audio independently from vision (and so cannot early-fuse them), along with the larger MTV-Huge model [122] by 1.5%.

4.4. Zero-Shot Experiments

Next, we show that our model exhibits strong zero-shot performance for a variety of downstream tasks. Our zero-shot interface is enabled by our *contrastive span objective*. For QA tasks that require predicting an option from a label space of short phrases, we encode this label space as vectors, and predict the closest phrase to a MASKed input. We consider:

i. Situated Reasoning (STAR) [119]. This task requires the model to reason over short situations in videos, covering four axes: interaction, sequence, prediction, and feasibility. The model is given a video, a templated question, and 4 answer choices. We convert templated questions into literal statements (which are more similar to YouTube dialogue); the label space is the set of four options.

ii. Action Anticipation in Epic Kitchens [26]. Here, the goal is to predict *future actions* given a video clip, which requires reasoning temporally over an actor's motivations and intentions. The dataset has a long tail of rare action combinations, making zero-shot inference challenging (since we do not assume access to this prior). As such, prior work [46, 38] trains on the provided in-domain training set. To adapt RESERVE to this task, we provide it a single MASK token as text input, and use as our label space of all combinations of verbs and nouns in the vocabulary (e.g. 'cook apple, cook avocado', etc.).

iii. LSMDC [82, 96]. Models are given a video clip, along with a video description (with a MASK to be filled in). We compare it with the vocabulary used in prior work [128].

iv. MSR-VTT QA [120]. This is an open-ended video QA task about what is literally happening in a web video. We use GPT3 [16], prompted with a dozen (unlabelled) questions, to reword the questions into statements with MASKs. This introduces some errors, but minimizes domain shift. We use a label space of the top 1k options.

For these tasks, we use $N{=}8$ video segments (dilating time when appropriate), and provide audio input when possible. Details and prompts are in Appendix D. We compare against both finetuned and zeroshot models, including running CLIP [92] on all tasks. CLIP is a strong model for zero-shot classification, particularly when *encyclopedic knowledge about images* is helpful; our comparisons showcase where multimodal script knowledge helps.

Results. Table 5 shows our model performs competitively:

i. On STAR, it obtains state-of-the-art results, with performance gain when audio is included. Interestingly, RESERVE-B outperforms its larger variant; we hypothesize that this is due to limited prompt searching around question templates. We qualitatively observed that RESERVE-L sometimes excludes topically correct options if they sound grammatically strange (to it).

ii. On EPIC-Kitchens, our model obtains strong results at correctly anticipating the verb and noun - despite the heavy-tailed nature of both distributions. It is worse on getting both right ('action'), we suspect that this might be due to priors (motifs) between noun and verb [129]. These are easy to learn given access to training data, but we exclude these as we consider the zero-shot task.

iii. On LSMDC, our model obtains strong results at filling-in-the-blank, likewise despite a heavy (unseen) frequency bias. Notably, it outperforms CLIP significantly, with CLIP often preferring templates that use visually-relevant words, even if they don't make sense as a whole. For instance, given a clip of a mailman, CLIP chooses 'the mailman smiles off,' versus 'the mailman takes off.'

iv. Finally, our model performs well on MSR-VTT QA, outperforming past work that directly rewords subtitled instructional videos into video QA instances [123].

5. Qualitative Analysis: Why does audio help?

What can RESERVE learn from both text *and* audio? Three validation set examples are shown in Figure 5. The model is given the displayed text and video frames, and must

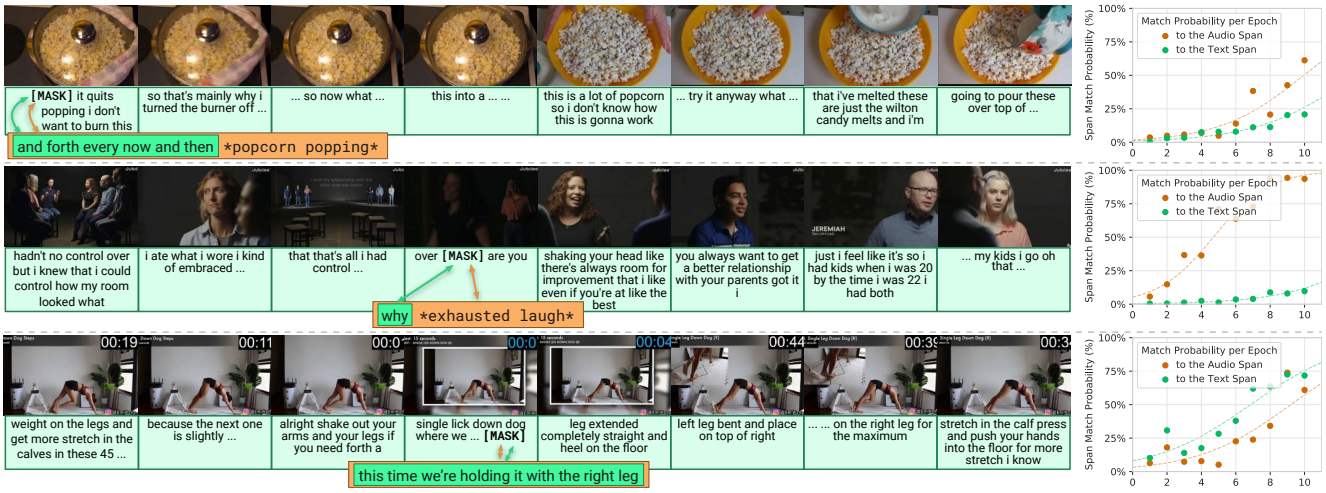

Figure 5: **Exploring MASKed audio self-supervision**. Shown are example videos from our validation set, with predictions from RESERVE-B. During pretraining, our model progressively learns to pick up on audio-specific clues. It seems to recognize physical dynamics of *cooking popcorn*, matching the first row to its MASKed audio. Likewise, it seems to use social reasoning to match the second row to its audio. Both of these clues are orthogonal to what the subtitles provide.

match the MASK to the correct missing text and audio span (out of 48k total in the batch). The plots show RESERVE-B's probability of correctly identifying the correct audio or text span, as it progresses through 10 epochs of pretraining.

Audio's supervisory signal. In the first two rows of Figure 5, audio provides orthogonal supervision to text:

1. In the first row, the MASKed audio contains the sound of popcorn pops slowing. By the final epoch, RESERVE-B selects this specific auditory cue with 60% probability, over others (including from adjacent segments, at different stages of popping). Here, sound provides signal for joint vision-text understanding of the situation, as evidenced by its greater match probability.
2. The second row contains only the text 'why,' with the audio providing greatly more information — a female-presenting speaker (shown in the next frame) laughs, astonished that the child (in the frame afterwards) might want a better relationship with their parents.
3. In the third row, matching performance is similar between modalities, possibly as the yogi is narrating over a (muted) video recording, and not adding much information.

Role of text. Text is still a crucial complement to audio, in terms of the supervision it provides. Consider the second row: RESERVE-B learns to match the audio almost perfectly (perhaps reasoning that the speaker is shown in the next frame, and is laughing). In later epochs, its text-match probability increases: knowing that a 'why' question is likely to be asked is a valid *social* inference to make about this (tense) situation.

Learning through multimodal reentry. Developmental psychologists have hypothesized that human children learn by *reentry*: learning connections between all senses as they interact with the world [35, 100]. Using a held-out modality (like audio) might support learning a better world representation (from e.g. vision and text), by forcing models to abstract away from raw perceptual input. Our work suggests that reentry has potential for machines as well.

6. Conclusion, Limitations, Broader Impact

We introduced RESERVE, which learns jointly through sound, language, and vision, guided through a new pretraining objective. Our model performs well in both finetuned and zero-shot settings, yet it has limitations. Our model only learns from 40-second long videos; relies on ASR models for subtitles, and can only match (not generate) text and audio.

Still, we foresee broad possible societal impact of this line of work. Video-pretrained models might someday assist low vision or d/Deaf users [76, 48]. Yet, the same technology can have impacts that we authors consider to be negative, including surveillance, or applications that hegemonize social biases. We discuss these further in Appendix A: key dimensions include respecting user privacy during dataset collection, exploring biases in YouTube data, dual use, and energy consumption. We discuss our plan to *release our model and data* for research use so others can critically study this approach to learning script knowledge.

Acknowledgements

We thank the anonymous reviewers, as well as Jae Sung Park, Oren Etzioni, Gabriel Ilharco, and Mitchell Wortsman for feedback on this work. Thanks also to Zak Stone and the Google Cloud TPU team for providing access to the TPU machines used for conducting experiments. Thanks to James Bradbury and Skye Wanderman-Milne for help with JAX on TPUs. Thanks to the AI2 ReVIZ team, including Jon Borchardt and M Kusold, for help with the demo. This work was funded by DARPA MCS program through NIWC Pacific (N66001-19-2-4031), and the Allen Institute for AI. Last, but not least, thanks to the YouTubers whose work and creativity helps machines to learn about the multimodal world.

References

[1] Sami Abu-El-Haija, Nisarg Kothari, Joonseok Lee, Paul Natsev, George Toderici, Balakrishnan Varadarajan, and Sudheendra Vijayanarasimhan. Youtube-8m: A large-scale video classification benchmark. *arXiv preprint arXiv:1609.08675*, 2016. 14

[2] 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. 2, 7, 20, 25

[3] 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. 2

[4] 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. 6

[5] 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. 18, 19

[6] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018. 2

[7] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623, 2021. 14, 16

[8] Emily M. Bender and Alexander Koller. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online, July 2020. Association for Computational Linguistics. 16

[9] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? *arXiv preprint arXiv:2102.05095*, 2021. 7

[10] Stella Biderman, Sid Black, Charles Foster, Leo Gao, Eric Hallahan, Horace He, Ben Wang, and Phil Wang. Rotary embeddings: A relative revolution, 2021. [Online; accessed]. 17

[11] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963*, 2021. 16

[12] Sophie Bishop. Anxiety, panic and self-optimization: Inequalities and the youtube algorithm. *Convergence*, 24(1):69–84, 2018. 16

[13] Yonatan Bitton, Gabriel Stanovsky, Michael Elhadad, and Roy Schwartz. Data efficient masked language modeling for vision and language. *arXiv preprint arXiv:2109.02040*, 2021. 2

[14] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021. 14

[15] Luke Breitfeller, Emily Ahn, David Jurgens, and Yulia Tsvetkov. Finding microaggressions in the wild: A case for locating elusive phenomena in social media posts. 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 1664–1674, 2019. 16

[16] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020. 7, 14, 16

[17] Qiong Cao, Li Shen, Weidi Xie, Omkar M Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, pages 67–74. IEEE, 2018. 14, 15

[18] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, et al. Extracting training data from large language models. *arXiv preprint arXiv:2012.07805*, 2020. 14

[19] Joao Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A short note about kinetics-600. *arXiv preprint arXiv:1808.01340*, 2018. 2, 6, 25

[20] Robin S Chapman. Children's language learning: An interactionist perspective. *The Journal of Child Psychology and Psychiatry and Allied Disciplines*, 41(1):33–54, 2000. 1

[21] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020. 2, 4, 6, 17, 18

[22] Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying vision-and-language tasks via text generation. In *ICML*, 2021. 2

[23] Kyunghyun Cho. Tweet. "important dependences between the image features and words/phrases in the description could be explained away by the dependencies among words/phrases". 5

[24] Christopher Clark, Mark Yatskar, and Luke Zettlemoyer. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases. In *EMNLP*, pages 4069–4082, Hong Kong, China, Nov. 2019. Association for Computational Linguistics. 2

[25] Matthew Crain. The limits of transparency: Data brokers and commodification. *new media & society*, 20(1):88–104, 2018. 14

[26] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, , Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision (IJCV)*, 2021. 2, 7, 25

[27] Karan Desai and Justin Johnson. Virtex: Learning visual representations from textual annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11162–11173, 2021. 5

[28] Sunipa Dev, Masoud Monajatipoor, Anaelia Ovalle, Arjun Subramonian, Jeff M Phillips, and Kai-Wei Chang. Harms of gender exclusivity and challenges in non-binary representation in language technologies. *arXiv preprint arXiv:2108.12084*, 2021. 16

[29] 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. 2, 5, 14

[30] Travis L Dixon. Crime news and racialized beliefs: Understanding the relationship between local news viewing and perceptions of african americans and crime. *Journal of Communication*, 58(1):106–125, 2008. 27

[31] Travis L Dixon and Daniel Linz. Overrepresentation and underrepresentation of african americans and latinos as lawbreakers on television news. *Journal of communication*, 50(2):131–154, 2000. 27

[32] Jesse Dodge, Maarten Sap, Ana Marasovic, William Agnew, Gabriel Ilharco, Dirk Groeneveld, and Matt Gardner. Documenting the english colossal clean crawled corpus. *CoRR*, abs/2104.08758, 2021. 16

[33] Carl Doersch, Abhinav Gupta, and Alexei A Efros. Unsupervised visual representation learning by context prediction. In *Proceedings of the IEEE international conference on computer vision*, pages 1422–1430, 2015. 4

[34] 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. 3, 4, 6, 19

[35] Gerald M Edelman. Neural darwinism: selection and reentrant signaling in higher brain function. *Neuron*, 10(2):115–125, 1993. 1, 8

[36] David F Fouhey, Wei-cheng Kuo, Alexei A Efros, and Jitendra Malik. From lifestyle vlogs to everyday interactions. In *CVPR*, 2018. 27

[37] Christian Fuchs. An alternative view of privacy on facebook. *Information*, 2(1):140–165, 2011. 14

[38] Antonino Furnari and Giovanni Maria Farinella. Rolling-unrolling lstms for action anticipation from first-person video. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 2020. 7, 25

[39] 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. 6

[40] Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020. 23

[41] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumeé III, and Kate Crawford. Datasheets for datasets. *arXiv preprint arXiv:1803.09010*, 2018. 27

[42] Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, 2020. 16

[43] Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *Proc. IEEE ICASSP 2017*, New Orleans, LA, 2017. 29

[44] Tarleton Gillespie. Content moderation, ai, and the question of scale. *Big Data & Society*, 7(2):2053951720943234, 2020. 16

[45] Franklin D Gilliam Jr, Shanto Iyengar, Adam Simon, and Oliver Wright. Crime in black and white: The violent, scary world of local news. *Harvard International Journal of press/politics*, 1(3):6–23, 1996. 27

[46] Rohit Girdhar and Kristen Grauman. Anticipative video transformer. *arXiv preprint arXiv:2106.02036*, 2021. 7, 25

[47] Yuan Gong, Yu-An Chung, and James Glass. Ast: Audio spectrogram transformer. *arXiv preprint arXiv:2104.01778*, 2021. 3, 18

[48] Steven M Goodman, Ping Liu, Dhruv Jain, Emma J McDonnell, Jon E Froehlich, and Leah Findlater. Toward user-driven sound recognizer personalization with people who are d/deaf or hard of hearing. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 5(2):1–23, 2021. 8

[49] Daniel Gordon, Kiana Ehsani, Dieter Fox, and Ali Farhadi. Watching the world go by: Representation learning from unlabeled videos. *arXiv preprint arXiv:2003.07990*, 2020. 5

[50] Jonathan Gordon and Benjamin Van Durme. Reporting bias and knowledge acquisition. In *Proceedings of the 2013 workshop on Automated knowledge base construction*, pages 25–30. ACM, 2013. 16

[51] 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. 2

[52] Ben Green. Good" isn't good enough. In *Proceedings of the AI for Social Good workshop at NeurIPS*, 2019. 16

[53] Daniel Griffin and Jae Lim. Signal estimation from modified short-time fourier transform. *IEEE Transactions on acoustics, speech, and signal processing*, 32(2):236–243, 1984. 18

[54] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. Audioclip: Extending clip to image, text and audio. *arXiv preprint arXiv:2106.13043*, 2021. 2, 29

[55] Foad Hamidi, Morgan Klaus Scheuerman, and Stacy M Branham. Gender recognition or gender reductionism? the social implications of embedded gender recognition systems. In *Proceedings of the 2018 chi conference on human factors in computing systems*, pages 1–13, 2018. 16

[56] Donna Haraway. Situated knowledges: The science question in feminism and the privilege of partial perspective. *Feminist studies*, 14(3):575–599, 1988. 16

[57] 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. 18

[58] Don Heider. *White news: Why local news programs don't cover people of color*. Routledge, 2014. 27

[59] Jack Hessel and Lillian Lee. Does my multimodal model learn cross-modal interactions? it's harder to tell than you might think! In *EMNLP*, 2020. 2

[60] 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*, Nov. 2019. 4

[61] Dirk Hovy and Shrimai Prabhumoye. Five sources of bias in natural language processing. *Language and Linguistics Compass*, 15(8):e12432, 2021. 14

[62] Gabriel Ilharco, Yuan Zhang, and Jason Baldridge. Large-scale representation learning from visually grounded untranscribed speech. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 55–65, 2019. 2

[63] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. *arXiv preprint arXiv:2102.05918*, 2021. 2, 4

[64] 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. 5, 20

[65] Ruogu Kang, Laura Dabbish, Nathaniel Fruchter, and Sara Kiesler. "my data just goes everywhere:" user mental models of the internet and implications for privacy and security. In *Eleventh Symposium On Usable Privacy and Security ({SOUPS} 2015)*, pages 39–52, 2015. 14, 27

[66] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020. 6, 19

[67] Os Keyes, Zoë Hitzig, and Mwenza Blell. Truth from the machine: artificial intelligence and the materialization of identity. *Interdisciplinary Science Reviews*, 46(1-2):158–175, 2021. 16

[68] Seonhoon Kim, Seohyeong Jeong, Eunbyul Kim, Inho Kang, and Nojun Kwak. Self-supervised pre-training and contrastive representation learning for multiple-choice video qa. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13171–13179, 2021. 6

[69] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. 4

[70] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16*, pages 491–507. Springer, 2020. 4

[71] Jatin Lamba, Jayaprakash Akula, Rishabh Dabral, Preethi Jyothi, Ganesh Ramakrishnan, et al. Cross-modal learning for audio-visual video parsing. *arXiv preprint arXiv:2104.04598*, 2021. 2

[72] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. In *International Conference on Learning Representations*, 2019. 6, 19

[73] Cheolhyoung Lee, Kyunghyun Cho, and Wanmo Kang. Mixout: Effective regularization to finetune large-scale pretrained language models. In *International Conference on Learning Representations*, 2019. 24

[74] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7331–7341, 2021. 2, 7

[75] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. Tvqa: Localized, compositional video question answering. In *EMNLP*, 2018. 2, 6

[76] Marco Leo, G Medioni, M Trivedi, Takeo Kanade, and Giovanni Maria Farinella. Computer vision for assistive technologies. *Computer Vision and Image Understanding*, 154:1–15, 2017. 8

[77] 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. 2, 6

[78] Xudong Lin, Gedas Bertasius, Jue Wang, Shih-Fu Chang, Devi Parikh, and Lorenzo Torresani. Vx2text: End-to-end learning of video-based text generation from multimodal inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7005–7015, 2021. 2

[79] Jing Liu, Xinxin Zhu, Fei Liu, Longteng Guo, Zijia Zhao, Mingzhen Sun, Weining Wang, Jinqiao Wang, and Hanqing Lu. Opt: Omni-perception pre-trainer for cross-modal understanding and generation. *arXiv preprint arXiv:2107.00249*, 2021. 2

[80] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 4

[81] 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. 2, 6

[82] Tegan Maharaj, Nicolas Ballas, Anna Rohrbach, Aaron C Courville, and Christopher Joseph Pal. A dataset and exploration of models for understanding video data through fill-in-the-blank question-answering. In *Computer Vision and Pattern Recognition (CVPR)*, 2017. 7, 26

[83] Alice E Marwick and danah boyd. Networked privacy: How teenagers negotiate context in social media. *New media & society*, 16(7):1051–1067, 2014. 14

[84] 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. 14

[85] 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. 27

[86] Heather Molyneaux, Susan O'Donnell, Kerri Gibson, Janice Singer, et al. Exploring the gender divide on youtube: An analysis of the creation and reception of vlogs. *American Communication Journal*, 10(2):1–14, 2008. 16

[87] Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. Voxceleb: a large-scale speaker identification dataset. *arXiv preprint arXiv:1706.08612*, 2017. 14, 15, 28, 29

[88] David Patterson, Joseph Gonzalez, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean. Carbon emissions and large neural network training. *arXiv preprint arXiv:2104.10350*, 2021. 17

[89] Jean Piaget and Margaret Trans Cook. The origins of intelligence in children. 1952. 1

[90] Karol J. Piczak. ESC: Dataset for Environmental Sound Classification. In *Proceedings of the 23rd Annual ACM Conference on Multimedia*, pages 1015–1018. ACM Press. 28, 29

[91] Yael Pritch, Sarit Ratovitch, Avishai Hendel, and Shmuel Peleg. Clustered synopsis of surveillance video. In *2009 Sixth IEEE international conference on advanced video and signal based surveillance*, pages 195–200. IEEE, 2009. 16

[92] 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. 2, 4, 7, 14, 15, 16, 23, 25, 29

[93] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67, 2020. 14, 16, 20, 27

[94] Micah Rajunov and A Scott Duane. *Nonbinary: Memoirs of Gender and Identity*. Columbia University Press, 2019. 16

[95] Manoel Horta Ribeiro, Raphael Ottoni, Robert West, Virgílio AF Almeida, and Wagner Meira Jr. Auditing radicalization pathways on youtube. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 131–141, 2020. 16

[96] 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. 2, 7, 26

[97] Andrew Rouditchenko, Angie Boggust, David Harwath, Brian Chen, Dhiraj Joshi, Samuel Thomas, Kartik Audhkhasi, Hilde Kuehne, Rameswar Panda, Rogerio Feris, et al. Avlnet: Learning audio-visual language representations from instructional videos. *arXiv preprint arXiv:2006.09199*, 2020. 2, 3

[98] Justin Salamon, Christopher Jacoby, and Juan Pablo Bello. A dataset and taxonomy for urban sound research. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 1041–1044, 2014. 28, 29

[99] 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. 1

[100] Linda Smith and Michael Gasser. The development of embodied cognition: Six lessons from babies. *Artificial life*, 11(1-2):13–29, 2005. 1, 8

[101] Nick Srnicek. *Platform capitalism*. John Wiley & Sons, 2017. 16

[102] Michael Strangelove. *Watching YouTube*. University of Toronto press, 2020. 14, 16

[103] 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. 17

[104] Jianlin Su, Yu Lu, Shengfeng Pan, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *arXiv preprint arXiv:2104.09864*, 2021. 17

[105] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. VideoBERT: A joint model for video and language representation learning. In *ICCV*, 2019. 2

[106] Hao Tan and Mohit Bansal. LXMERT: Learning cross-modality encoder representations from transformers. In *EMNLP*, 2019. 2, 4, 5, 17

[107] Rachael Tatman. Gender and dialect bias in youtube's automatic captions. *EACL 2017*, page 53, 2017. 16

[108] Hugo Touvron, Andrea Vedaldi, Matthijs Douze, and Herve Jegou. Fixing the train-test resolution discrepancy. *Advances in Neural Information Processing Systems*, 32:8252–8262, 2019. 18

[109] Aisha Urooj, Amir Mazaheri, Mubarak Shah, et al. Mmft-bert: Multimodal fusion transformer with bert encodings for visual question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 4648–4660, 2020. 6

[110] 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. 3, 18

[111] Aurélie Villard, Alan Lelah, and Daniel Brissaud. Drawing a chip environmental profile: environmental indicators for the semiconductor industry. *Journal of Cleaner Production*, 86:98–109, 2015. 17

[112] 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. 4

[113] Feng Wang and Huaping Liu. Understanding the behaviour of contrastive loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2495–2504, 2021. 4

[114] Luyu Wang, Pauline Luc, Adria Recasens, Jean-Baptiste Alayrac, and Aaron van den Oord. Multimodal self-supervised learning of general audio representations. *arXiv preprint arXiv:2104.12807*, 2021. 2

[115] Yuxuan Wang, RJ Skerry-Ryan, Daisy Stanton, Yonghui Wu, Ron J Weiss, Navdeep Jaitly, Zongheng Yang, Ying Xiao, Zhifeng Chen, Samy Bengio, et al. Tacotron: Towards end-to-end speech synthesis. *arXiv preprint arXiv:1703.10135*, 2017. 18

[116] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*, 2021. 2

[117] Zeerak Waseem, Smarika Lulz, Joachim Bingel, and Isabelle Augenstein. Disembodied machine learning: On the illusion of objectivity in nlp. *arXiv preprint arXiv:2101.11974*, 2021. 16

[118] Georg Wiese, Dirk Weissenborn, and Mariana Neves. Neural domain adaptation for biomedical question answering. In *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*, pages 281–289, 2017. 24

[119] Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B. Tenenbaum, and Chuang Gan. Star: A benchmark for situated reasoning in real-world videos. *2021 Conference on Neural Information Processing Systems*, 2021. 2, 7, 25

[120] 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. 2, 7, 26

[121] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, and Florian Metze Luke Zettlemoyer Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084*, 2021. 2

[122] Shen Yan, Xuehan Xiong, Anurag Arnab, Zhichao Lu, Mi Zhang, Chen Sun, and Cordelia Schmid. Multi-view transformers for video recognition. *arXiv preprint arXiv:2201.04288*, 2022. 7

[123] 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. 2, 7, 27

[124] 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. 2, 6

[125] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021. 7

[126] 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. 2, 5, 18

[127] 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. 14, 16

[128] Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. Merlot: Multimodal neural script knowledge models. *arXiv preprint arXiv:2106.02636*, 2021. 1, 2, 4, 5, 6, 7, 14, 16, 17, 18, 20, 24, 26, 27, 28

[129] Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *Conference on Computer Vision and Pattern Recognition*, 2018. 7

[130] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers, 2021. 4, 6, 17

[131] 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. 17

[132] 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. 2

[133] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, 2017. 16

[134] Shoshana Zuboff. Big other: surveillance capitalism and the prospects of an information civilization. *Journal of Information Technology*, 30(1):75–89, 2015. 16