THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL

MERLOT: Multimodal Neural Script Knowledge Models

Dan & Hao

2024/10/14

Introduction

Motivation

The human capacity for commonsense reasoning is shaped by how we experience causes and effects over time, which is a challenge to machines.

What is MERLOT?

—— Multimodal Event Representation Learning Over Time, which learns commonsense representations of multimodal events by self-supervised pretraining over 6M unlabelled YouTube videos.

How do we train MERLOT?

(a) Match individual video frames with contextualized representations of the associated transcripts.
(b) contextualize those frame-level representations over time by “unmasking" distant word-level corruptions and reordering scrambled video frames.

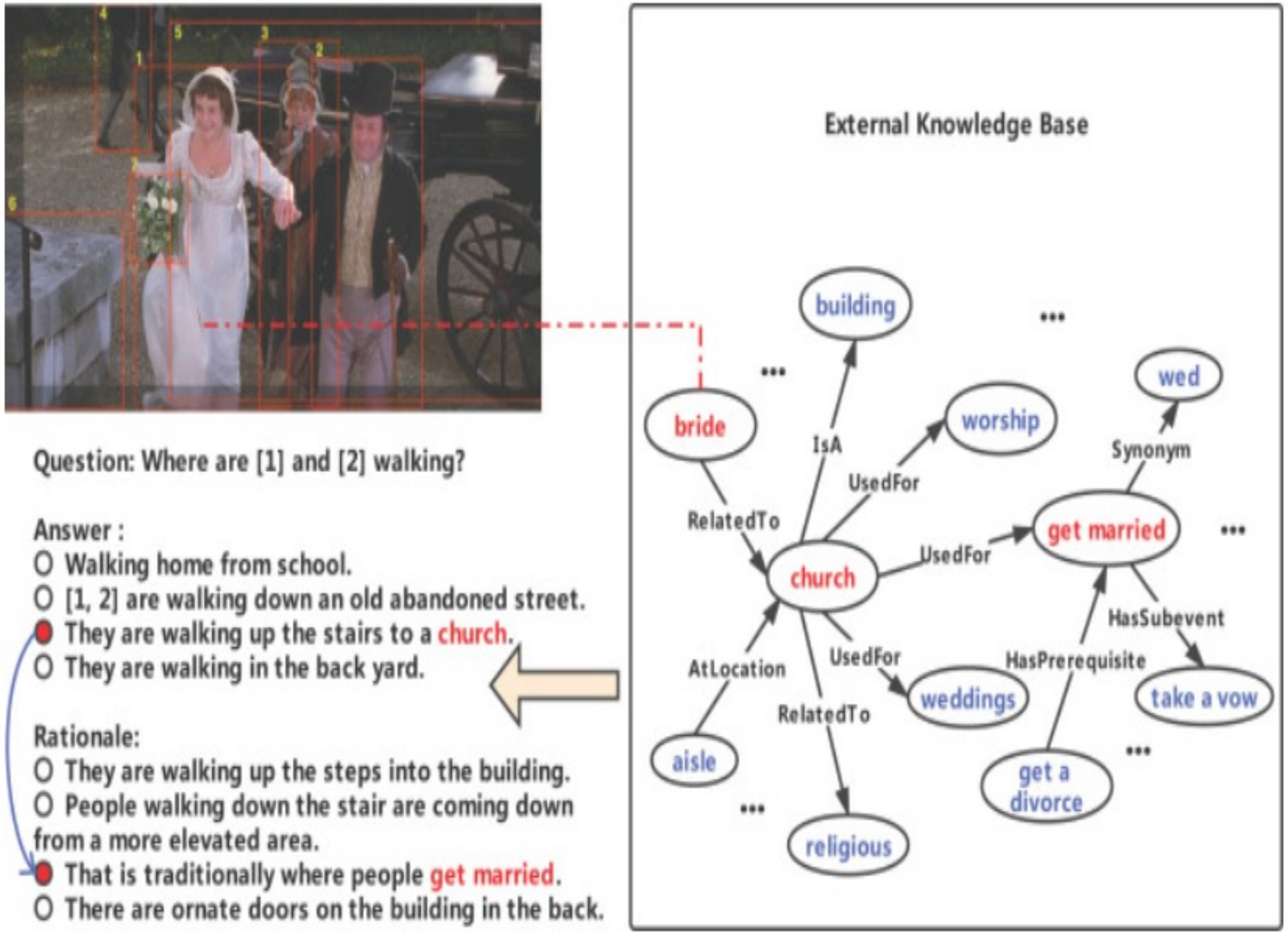
Related Work

RELATED WORK

Joint representations of text & images

The approaches of learning joint text-image representations of static images, and rely on significant human annotation in doing so.

Our approach learning dynamic visual representations purely from videos.



RELATED WORK

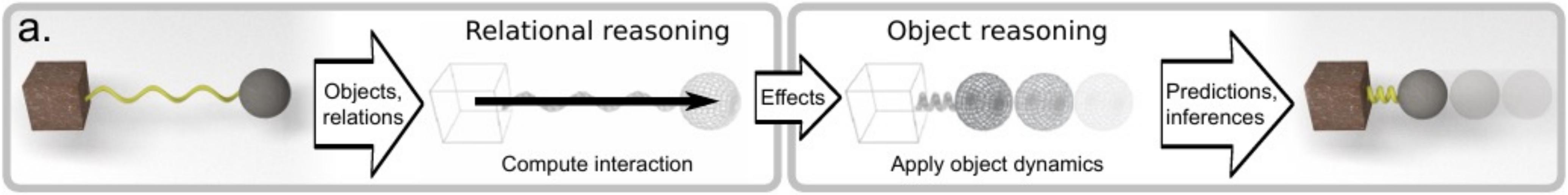
Learning from videos with ASR (Automatic Speech Recognition)

(1) Using web videos with ASR to build weakly-supervised object detectors
(2) Learning multimodal representations transferable to many tasks from uncurated sets of videos.
MERLOT is trained using a combination of objectives requiring no manual supervision, and performances better on downstream tasks.

Temporal ordering and forecasting

Past work uses Extrapolation (pixels, graphs, Euclidean distance, cycle consistency) and Deshuffling Objectives in videos.

Our method uses both language and vision as complementary views into the world to learn multimodal script knowledge representations, instead of just tracking what changes on-screen.



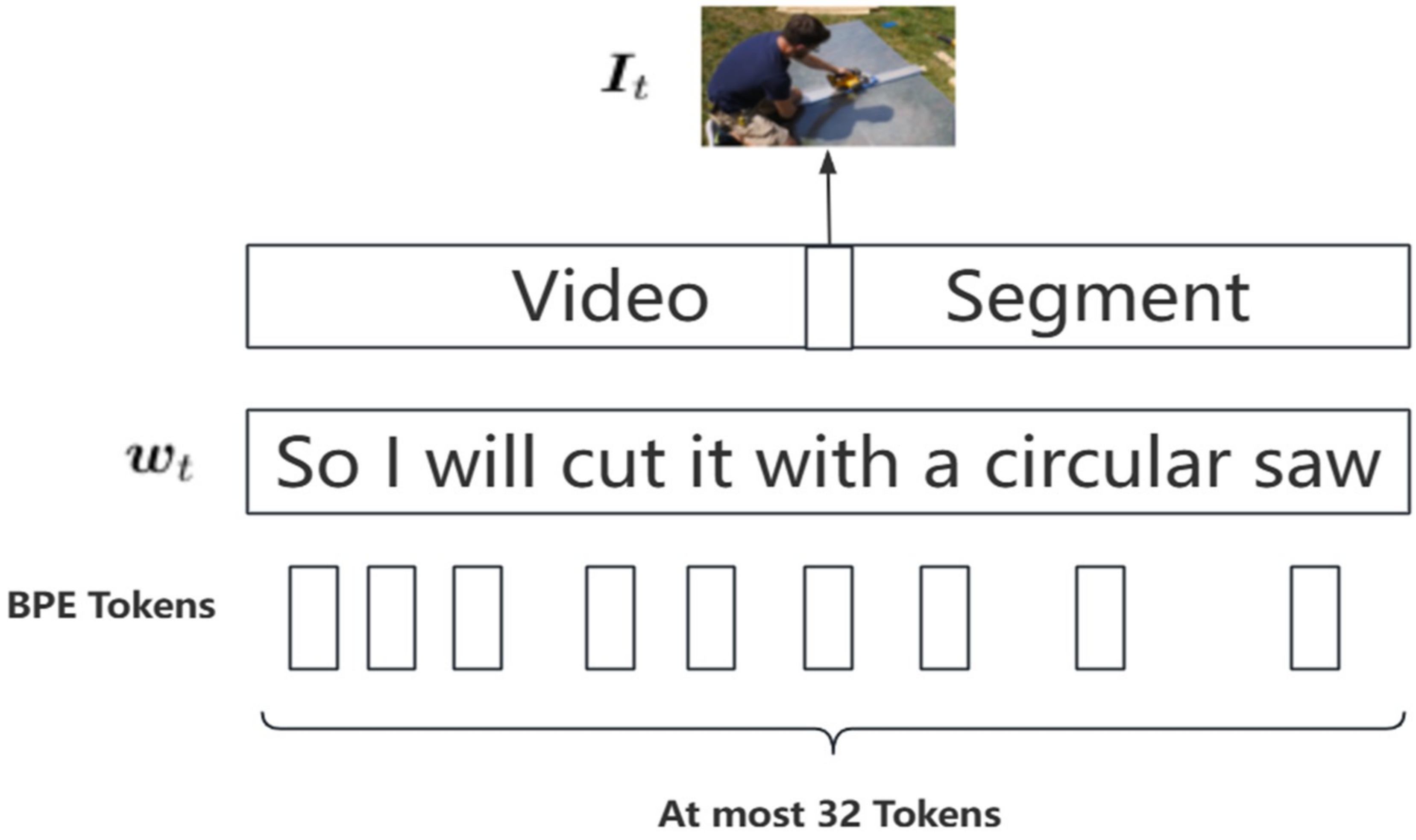
Method

YT-Temporal-180M

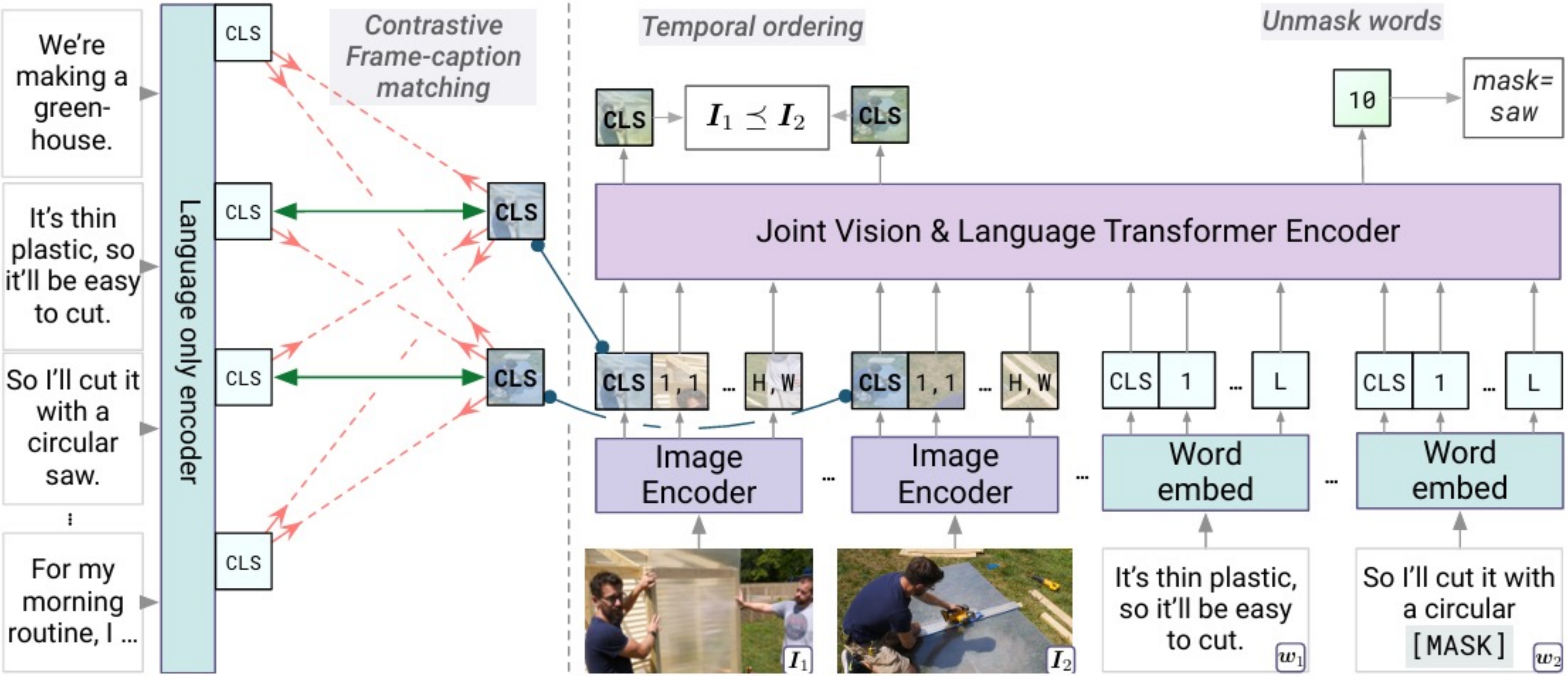
A dataset for learning multimodal script knowledge, derived from 6 million public YouTube videos.

Intentionally spans many domains, datasets, and topics to encourage the model to learn about a broad range of objects, actions, and scenes.

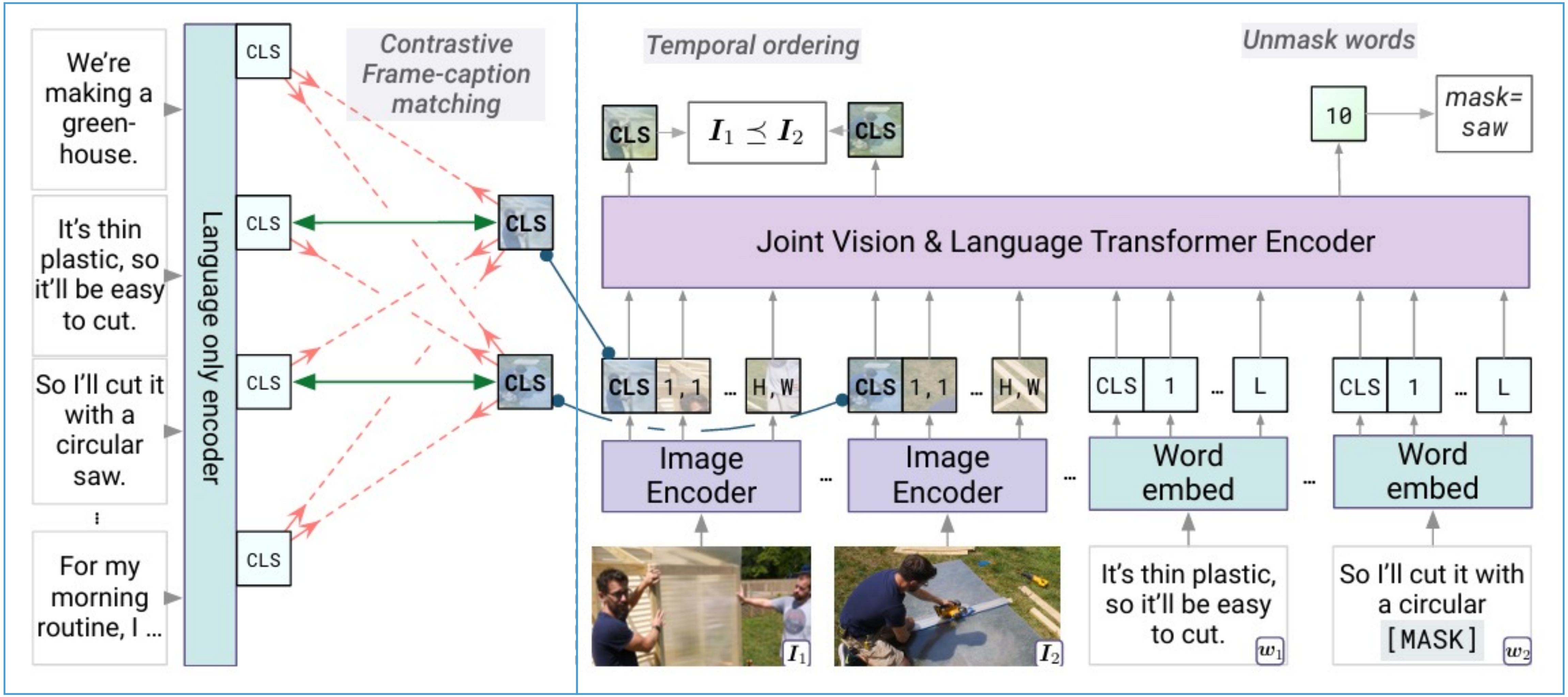
METHOD - ARCHITECTURE



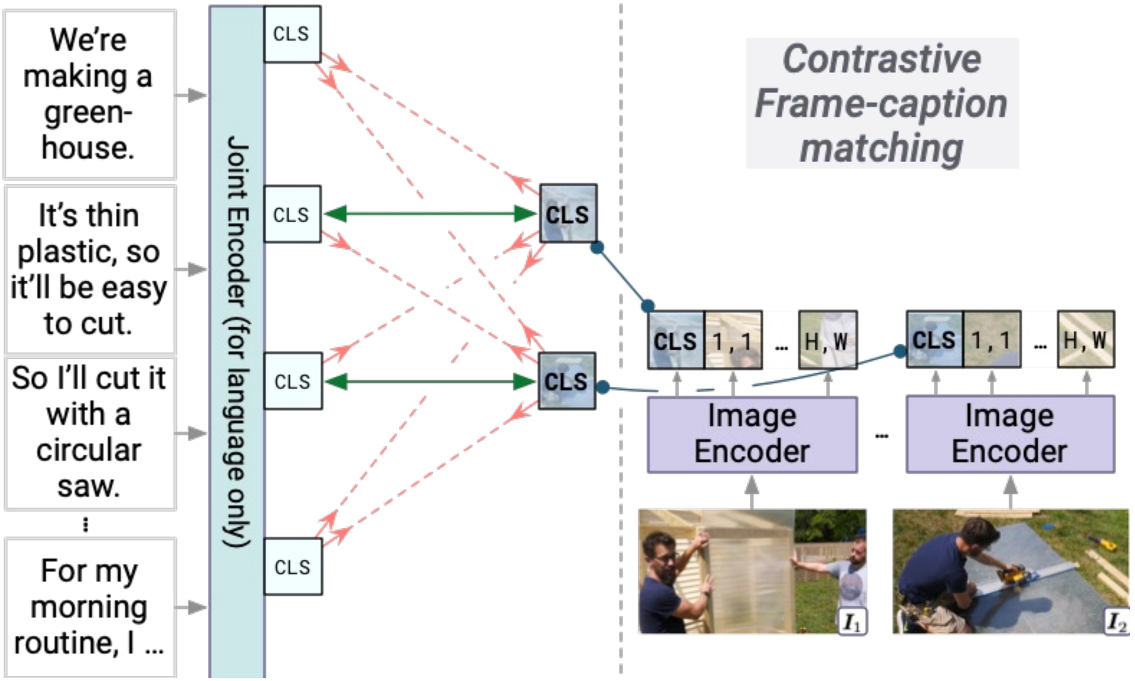
METHOD - ARCHITECTURE



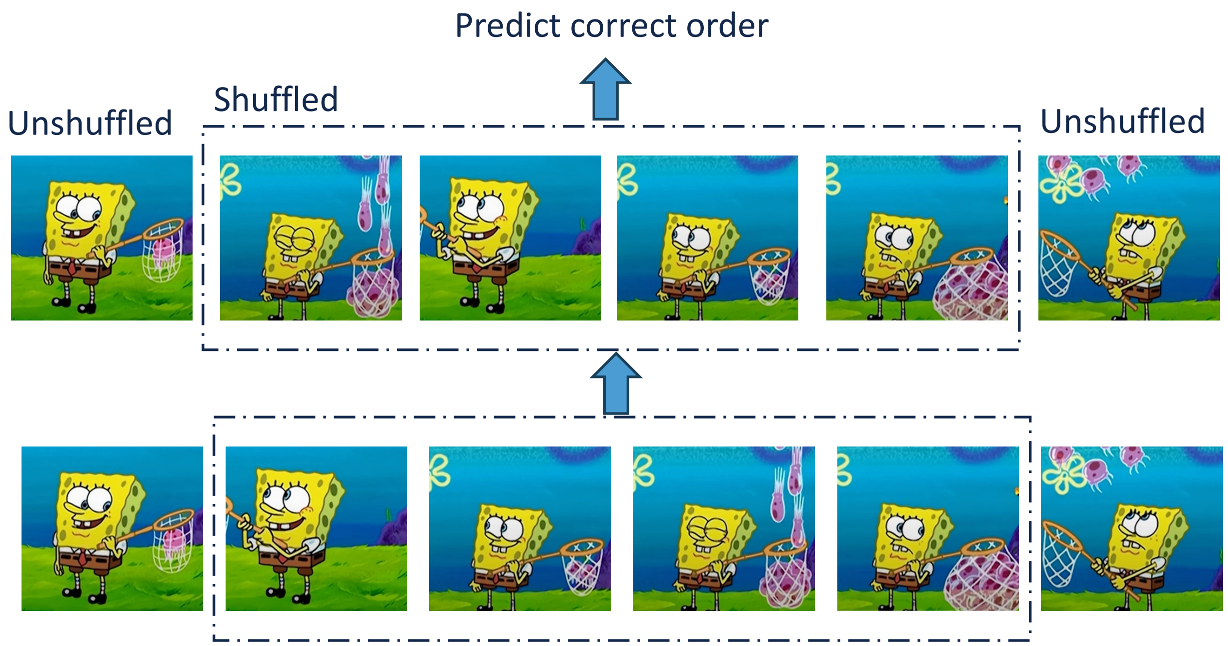
METHOD - ARCHITECTURE

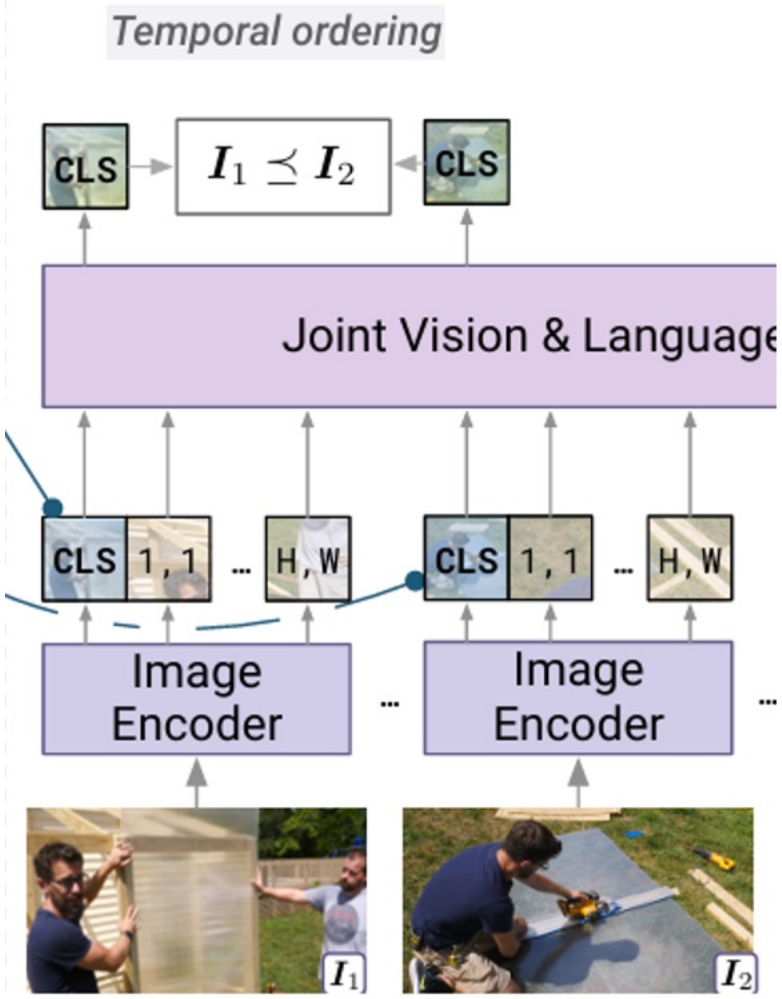


- **Contrastive frame-transcript matching**
  - Use language-only encoder to extract hidden states of video transcripts to see whether the frames and the subtitles are matched



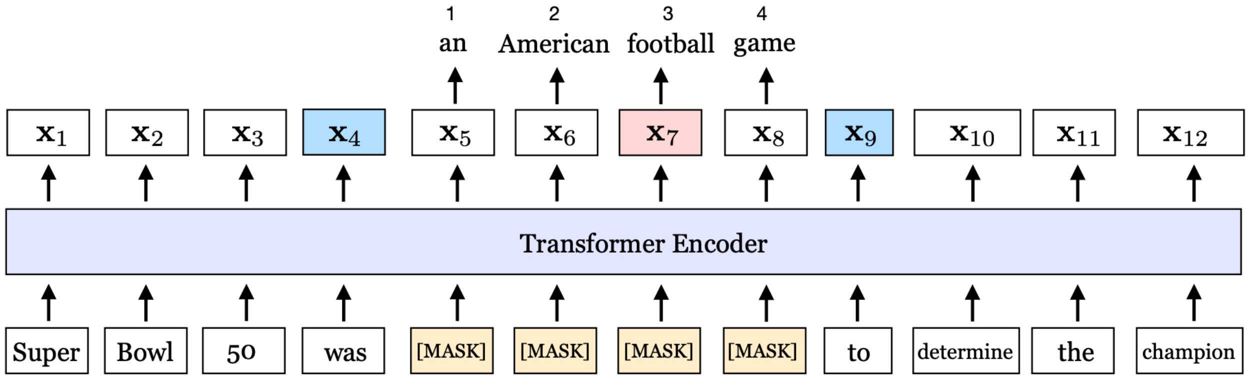
- **Temporal Reordering**
  - Have the model order the image frames in a video to force it to explicitly learn temporal reasoning





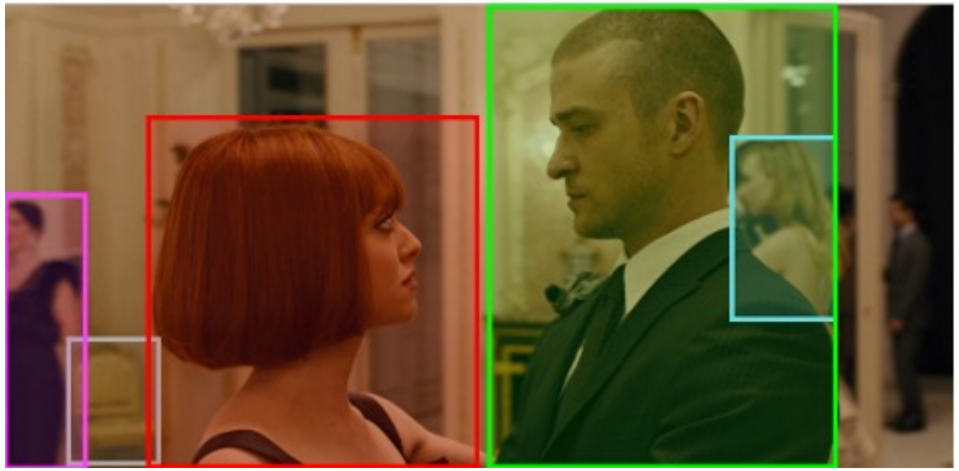
- **(Attention) Masked Language Modeling**
  - BERT-style masking: randomly replace 20% words with a MASK token & reconstruct
  - OUR method: attention masking
    - 50% time: randomly replace with a MASK token
    - Another 50% time: mask out of the top 20% most-attended-to-tokens
    - apply SpanBERT masking

$$\mathcal{L}(\text{football}) = \mathcal{L}_{\text{MLM}}(\text{football}) + \mathcal{L}_{\text{SBO}}(\text{football})$$
$$= -\log P(\text{football} \mid \mathbf{x}_7) - \log P(\text{football} \mid \mathbf{x}_4, \mathbf{x}_9, \mathbf{p}_3)$$



Experiments

- **VCR**
  - Models must answer commonsense visual questions about images.



	Q→A	QA→R	Q→AR
ViLBERT [75]	73.3	74.6	54.8
Unicoder-VL [68]	73.4	74.4	54.9
VLBERT [69]	73.8	74.4	55.2
UNITER [22]	75.0	77.2	58.2
VILLA [36]	76.4	79.1	60.6
ERNIE-ViL [119]	77.0	80.3	62.1
MERLOT (base-sized)	**80.6**	**80.4**	**65.1**

Table 1: Results on VCR [123]. We compare against SOTA models of the same 'base' size as ours (12-layer vision-and-language Transformers). MERLOT performs best on all metrics.

- **Unsupervised Ordering of Visual Stories**
  - Visual Stories dataset: 5 images and captions in a certain order
  - Task: must match frames to the captions
  - With no fine-tuning, MERLOT has strong capability to reason about past and future events from temporal visual stories.

	Spearman (↑)	Pairwise acc (↑)	Distance (↓)
CLIP [89]	.609	78.7	.638
UNITER [22]	.545	75.2	.745
MERLOT	**.733**	**84.5**	**.498**

Table 2: Results unscrambling SIND visual stories[50, 2]. Captions are provided in the correct order; models must arrange the images temporally. MERLOT performs best on all metrics by reasoning over the entire story, instead of independently matching images with captions.

- **Video Reasoning:**
  - Achieved SOTA on 12 video reasoning tasks

	Tasks	Split	Vid. Length	ActBERT [127]	ClipBERT$_{8x2}$ [67]	SOTA	MERLOT
244K QA, 10K 10s clips	MSRVTT-QA	Test	Short	-	37.4	41.5 [118]	**43.1**
	MSR-VTT-MC	Test	Short	88.2	-	88.2 [127]	**90.9**
	TGIF-Action	Test	Short	-	82.8	82.8 [67]	**94.0**
	TGIF-Transition	Test	Short	-	87.8	87.8 [67]	**96.2**
	TGIF-Frame QA	Test	Short	-	60.3	60.3 [67]	**69.5**
	LSMDC-FiB QA	Test	Short	48.6	-	48.6 [127]	**52.9**
	LSMDC-MC	Test	Short	-	-	73.5 [121]	**81.7**
58K QA, 5.8K videos	ActivityNetQA	Test	Long	-	-	38.9 [118]	**41.4**
18K MCQ, 24K 1min clips	Drama-QA	Val	Long	-	-	81.0 [56]	**81.4**
152K QA, 21.8K 1min clips, 460hrs	TVQA	Test	Long	-	-	76.2 [56]	**78.7**
30K QA, 4.2K 1min clips, 310K BB	TVQA+	Test	Long	-	-	76.2 [56]	**80.9**
28.7K Binary QA, 10K clips	VLEP	Test	Long	-	-	67.5 [66]	**68.4**

Table 3: 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.

- **Context Size**
  - Pretraining on more segments at once improves performance
    - more context -> language-only representation learning
  - Attention Masking can counteract this issue

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, attention masking	**75.2**	**75.8**

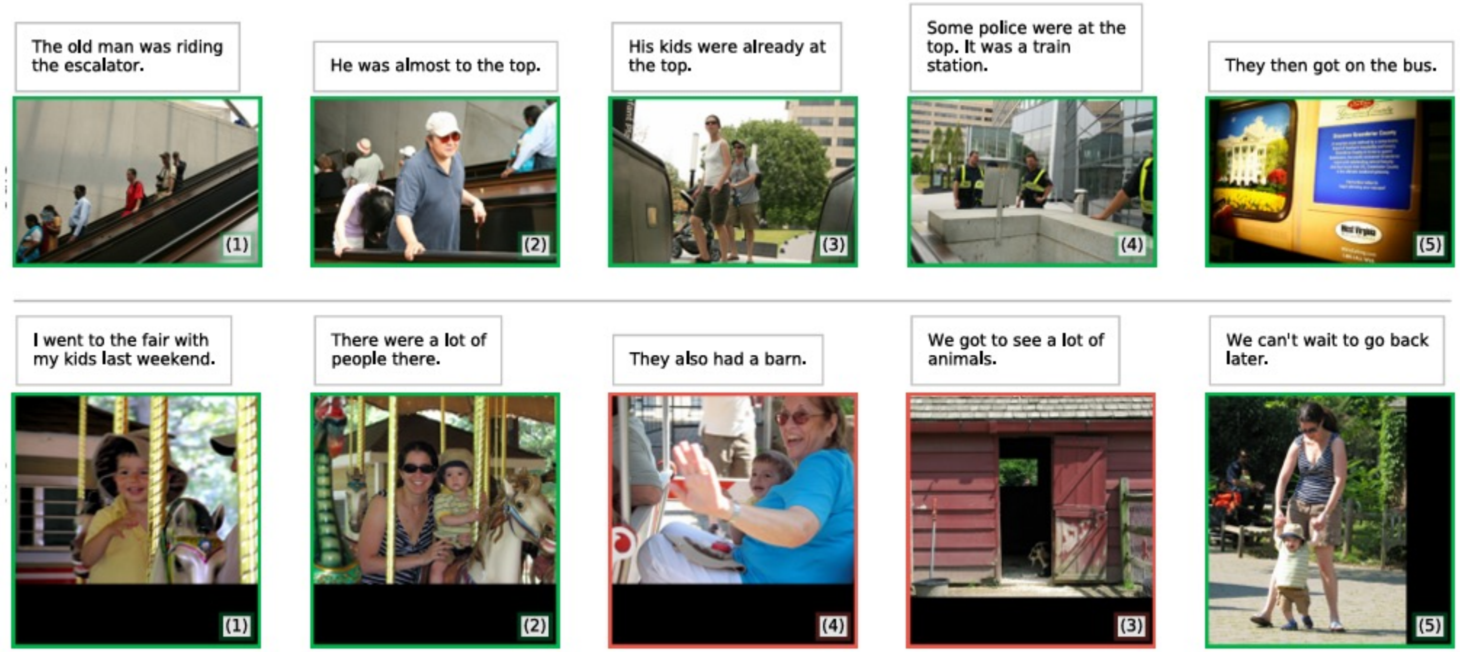
- **Dataset**
  - Perform better on YT-Temporal-180M, even when controlled for size
  - Using raw ASR reduces performance

Dataset	VCR
Conceptual ∪ COCO	58.9
HowTo100M	66.3
🍷 YT-Temporal-180M	**75.2**
HowTo100M-sized YT-Temporal-180M	72.8
YTT180M, raw ASR	72.8

- **Losses**
  - Removing contrastive V-L loss makes performance drop significantly
  - The temporal ordering loss is not as important for downstream finetuning

Training setup	VCR	TVQA+
No contrastive V-L loss	57.5	67.6
No temporal ordering loss	**75.5**	75.6
🍷 All losses	75.2	**75.8**

- **Zero-shot Story Ordering**
  - To match correct frames with the sorted captions
  - Interesting reason for the wrong one



Conclusion

CONCLUSION

- MERLOT demonstrates a novel way of **multimodal learning and temporal reasoning** by both visual frames and transcripts
- MERLOT is scalable to massive datasets without human annotations via **self-supervised learning objectives**
- Outperforms previous SOTA methods on various video QA tasks, benefiting from pretraining on a large, diverse dataset (YT-Temporal-180M)

Limitations

1. Finer-grained temporal reasoning pretraining objectives vs. frame ordering needs to be explored

e.g. a temporal frame localization within transcripts

2. Multilingual videos and communities on YouTube are not included
3. Social Biases

Thank You!