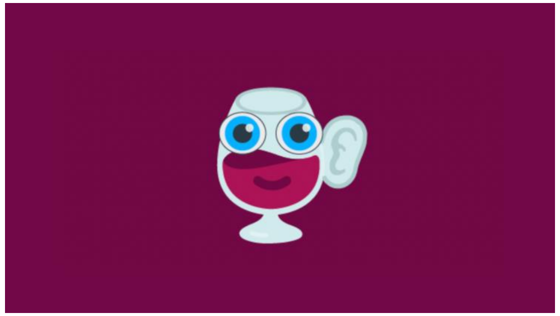
[CS6804] Paper Review



MERLOT RESERVE

Presenter: Chiawei Tang
Department of Computer Science
Virginia Tech
02/20/2023

https://reurl.cc/Dmv8dR

Outline

what we will lean in this presentation

- MERLOT
- VATT
- MERLOT RESERVE

MERLOT

Multimodal Event Representation Learning Over Time

Visual Commonsense Reasoning

- what might happen next
- what are people's intentions

Visual Commonsense Reasoning

- what might happen next
- what are people's intentions

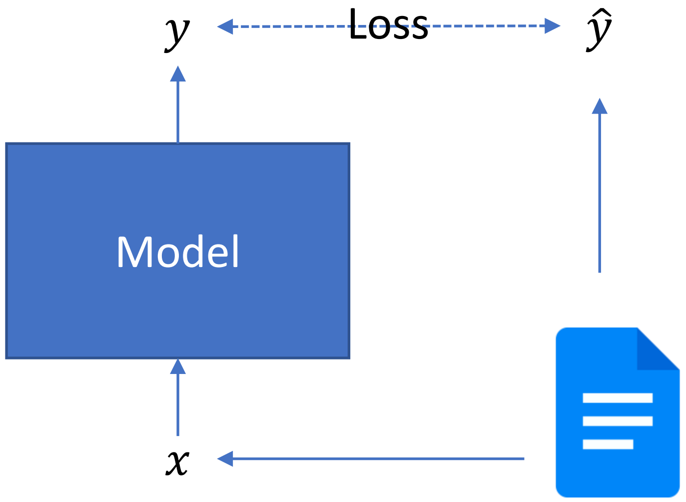
Vision

Language

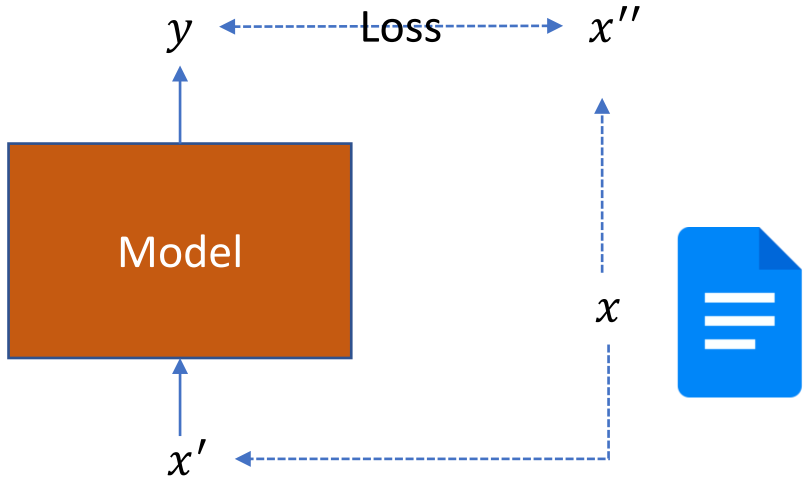
Audio

Self-Supervised Learning

Supervised

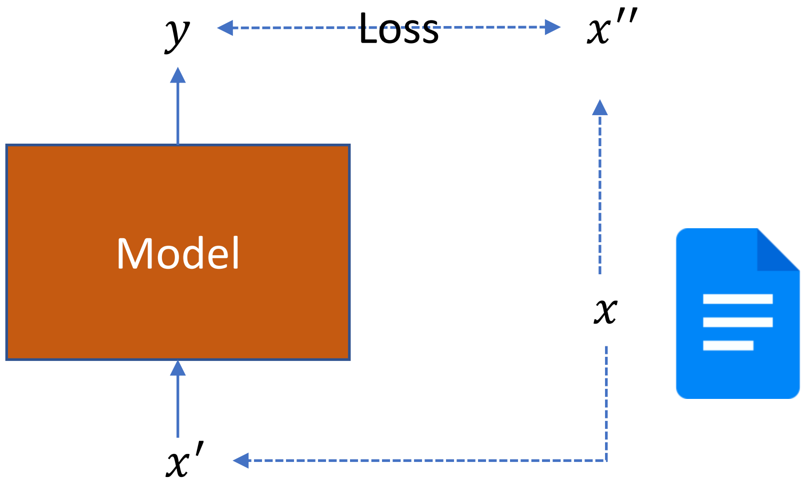


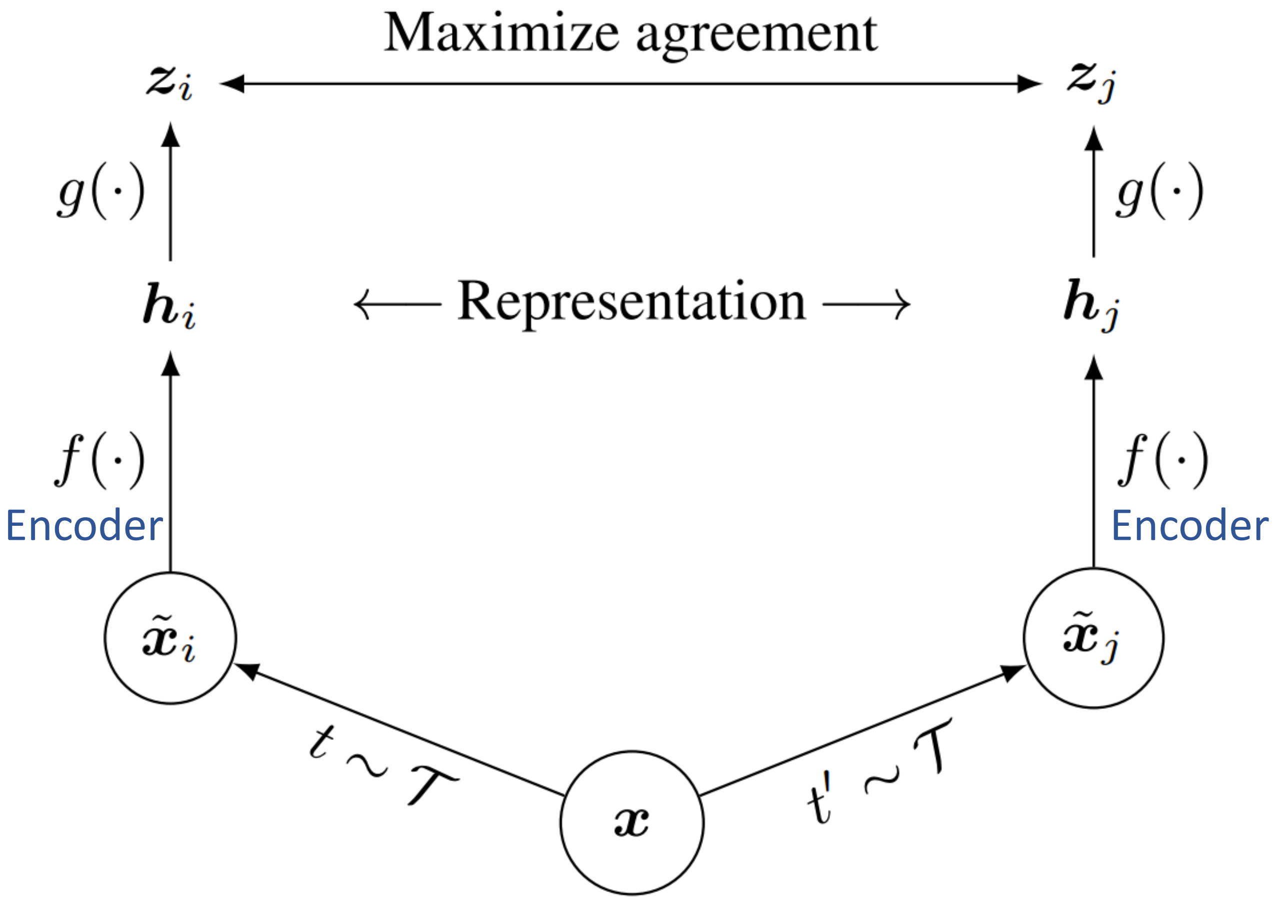
Self-supervised



Self-Supervised Learning

Self-supervised

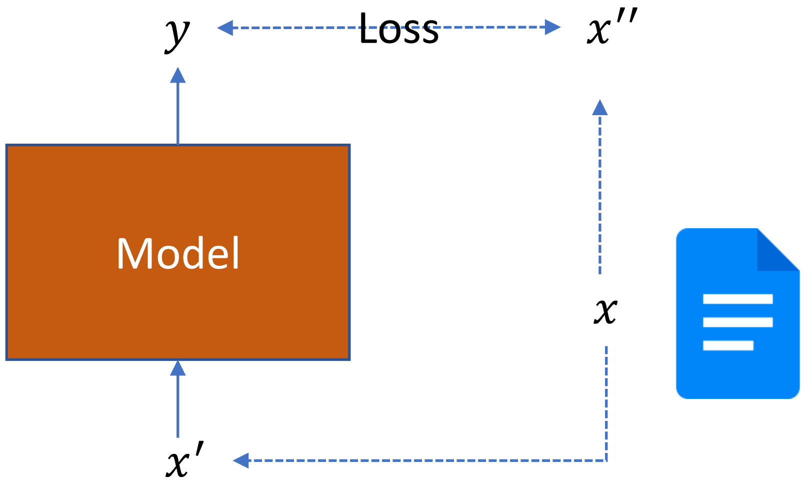


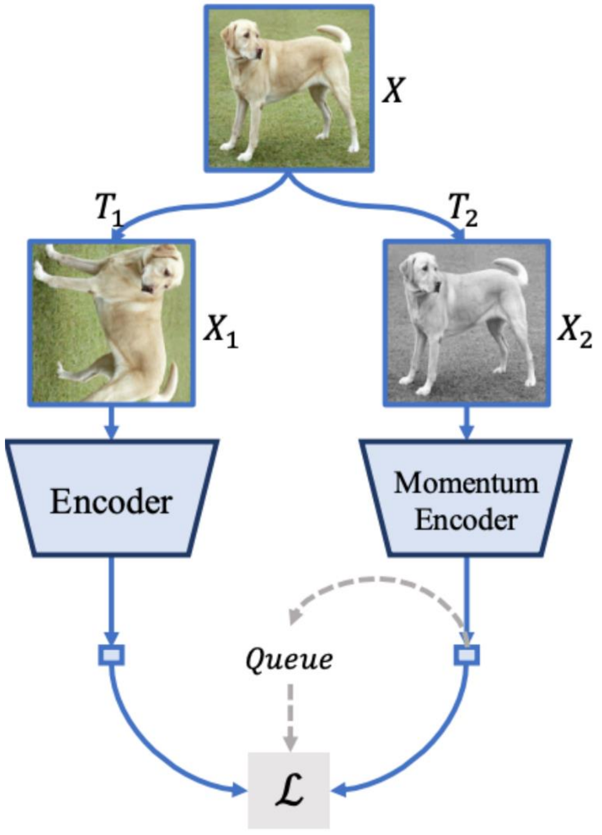


SimCLR: Chen et al, 2020

Self-Supervised Learning

Self-supervised



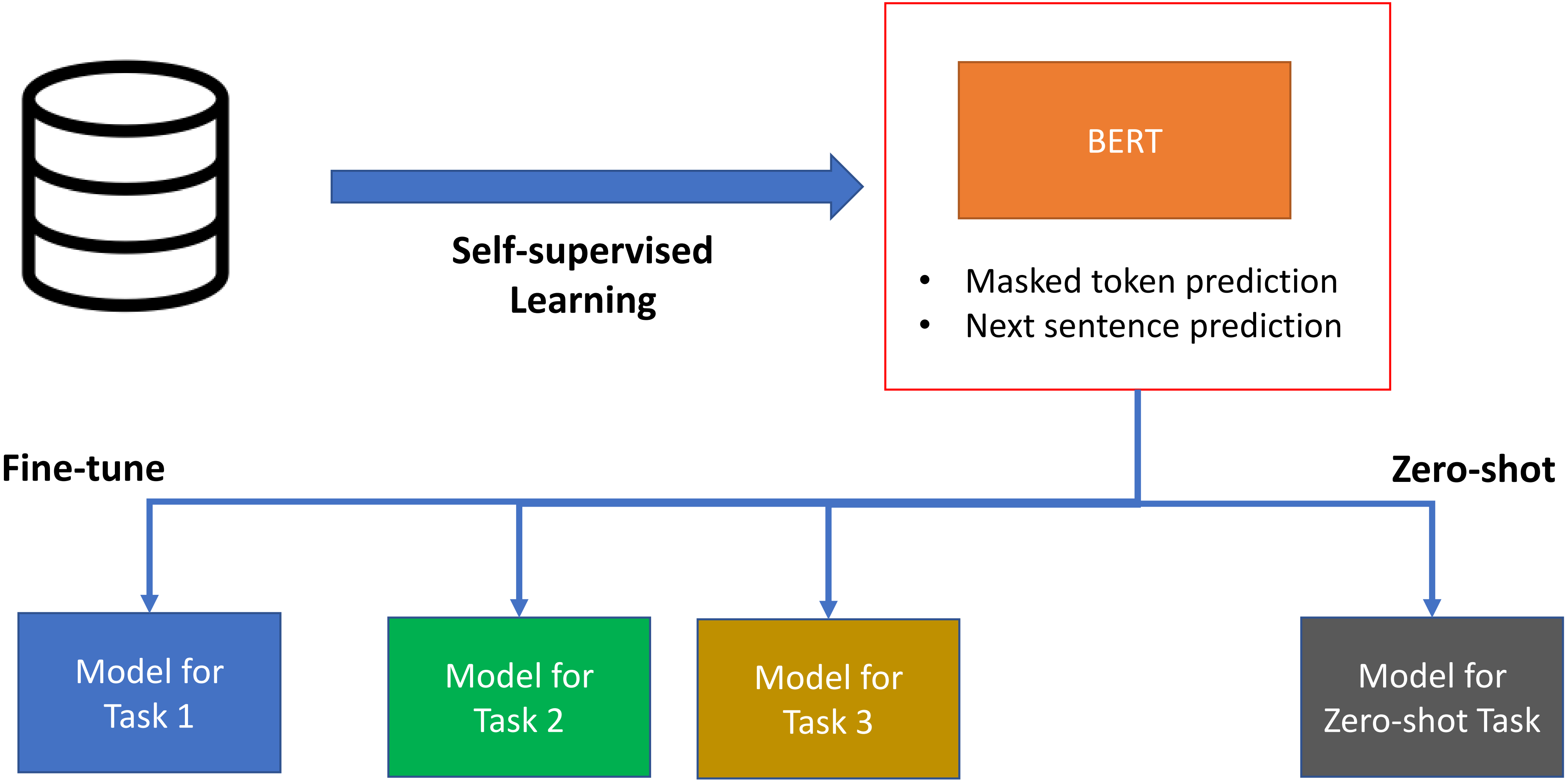


(a) MoCo

MoCo He et al. 2020

https://medium.com/geekculture/understanding-contrastive-learning-and-moco-efe491e4eed9

Self-Supervised Learning



Self-Supervised Learning

How to empower Vision Transformers with large-scale, unlabeled data?

https://www.youtube.com/watch?v=rgXxAFIBido

Self-Supervised Learning

How to empower Vision Transformers with large-scale, unlabeled data?

Self-supervision from the multimodal videos
(video frames, audio, and text)

https://www.youtube.com/watch?v=rgXxAFIBido

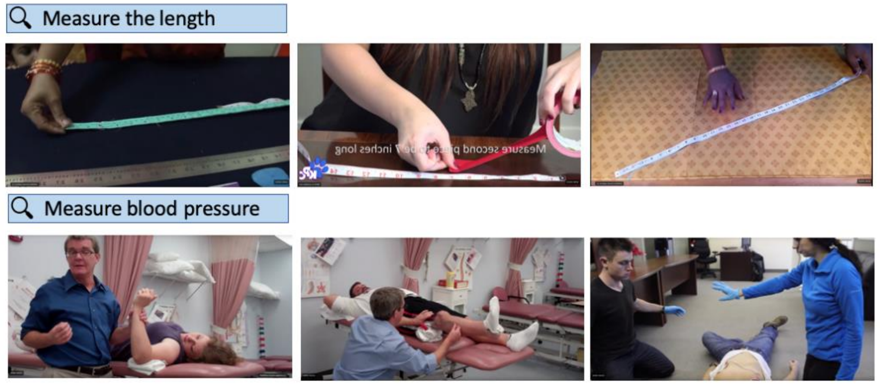
MERLOT

A model that learns commonsense representations of multimodal events by self-supervised pretraining over 6M unlabelled YouTube videos

Dataset

To learn about a broad range of objects, actions, and scenes

HowTo100M



VLOG

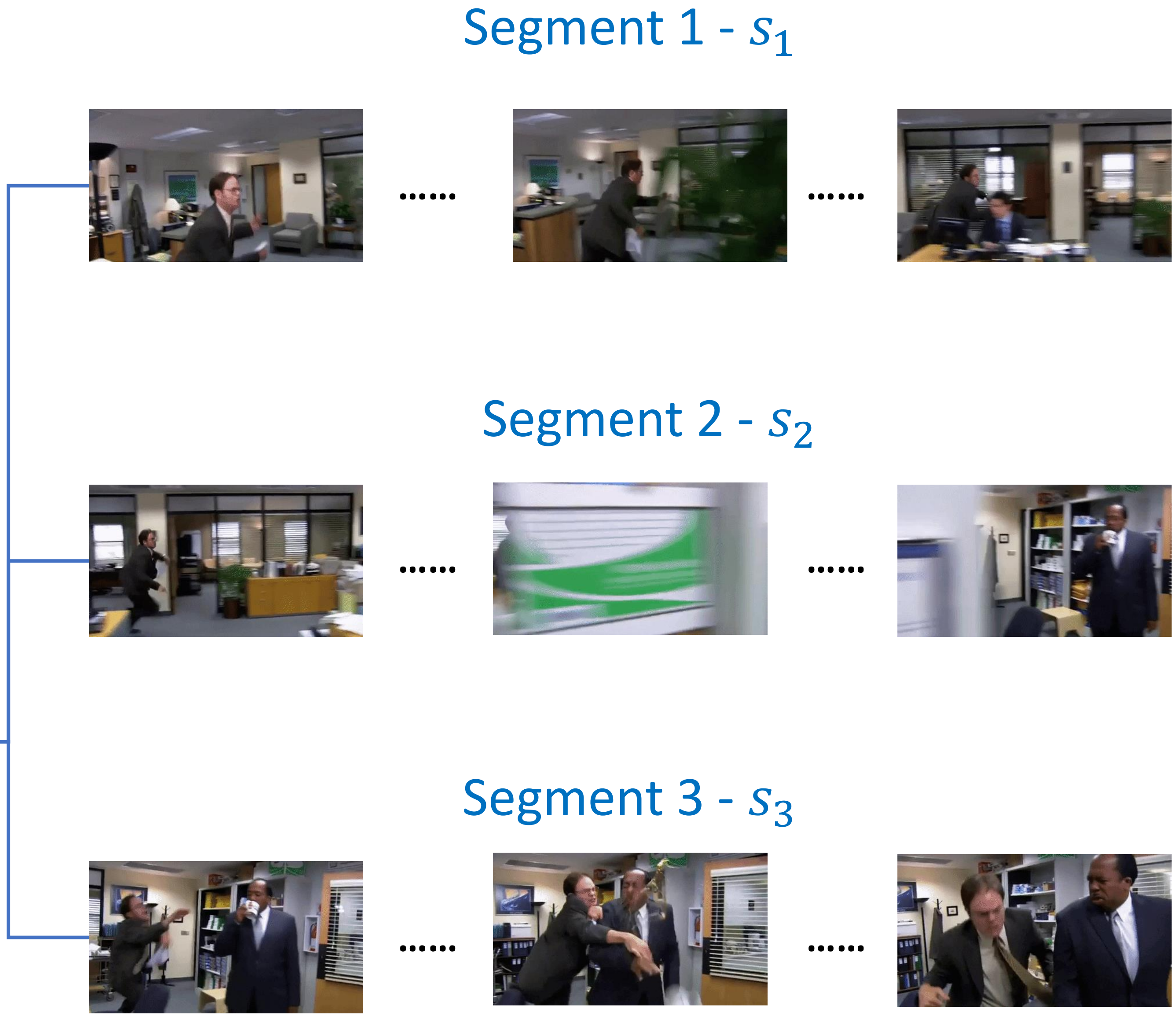
YouTube: Popular Topics

With ENGLISH ASR
Not too long (> 20 minutes)
Visually "ungrounded"
(video games commentaries)
Unlikely to contain objects

https://www.di.ens.fr/willow/research/howto100m/
https://web.eecs.umich.edu/~fouhey/2017/VLOG/index.html

MERLOT

32 Byte Pair Encode(BPE) tokens each



The Office

MERLOT

Segment 3 - s_3

....

....

I_1 I_{t-n} I_t

w_1 At 8 a.m. today, someone poisons the coffee.
w_2 Do not drink the coffee.
w_3 No~~~

...

- an image frame I_t, extracted from the middle timestep of the segment
- the words w_t spoken during the segment, with a total length of L tokens.

The Office

MERLOT

I_t

Image Encoder

Visual Transformer

Joint Vision & Language Transformer Encoder (RoBERTa)

At 8 a.m. today, someone poisons the coffee.

w_t

Word Embed

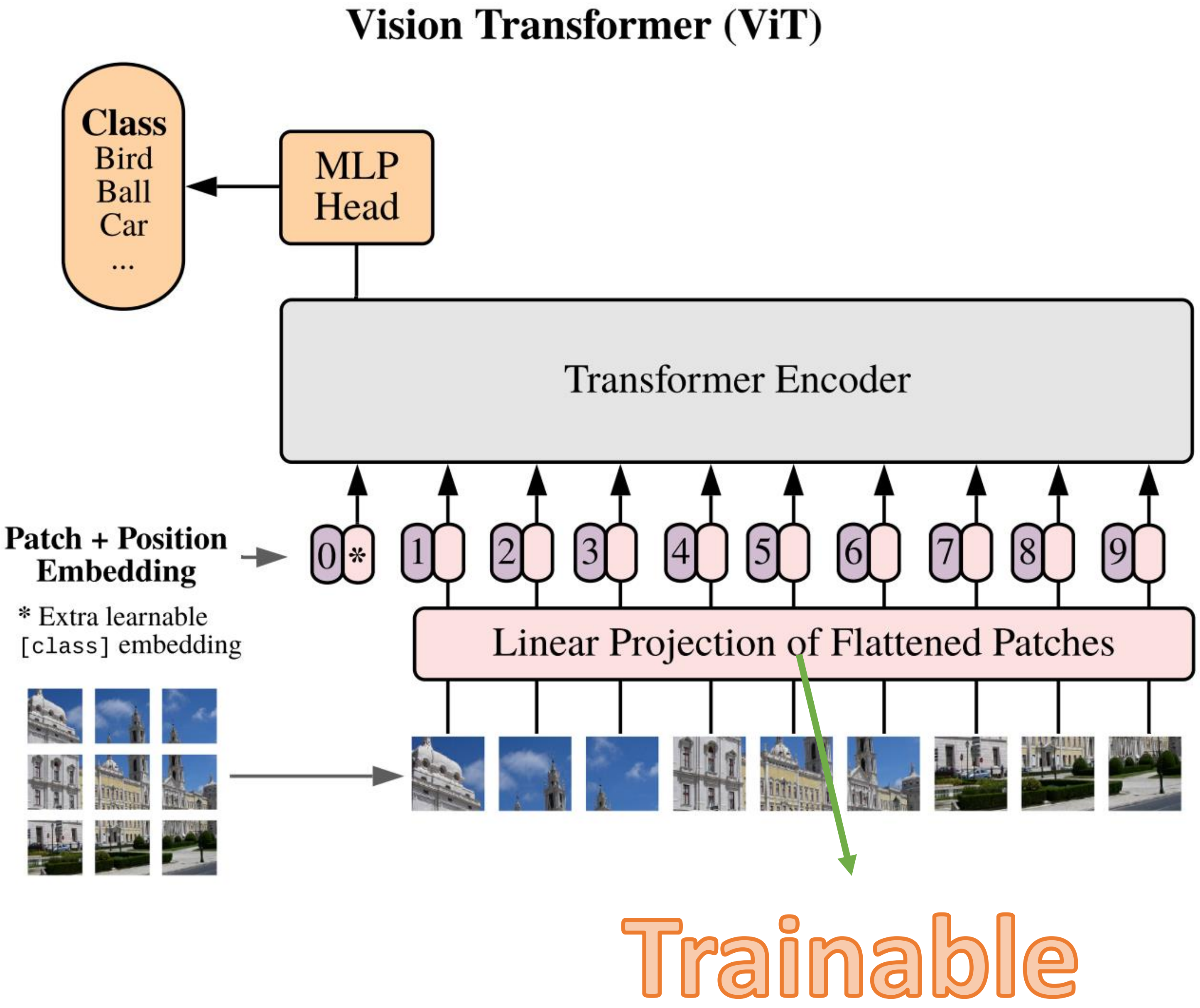
BPE

MERLOT

Image Encoder

Different image size for videos

https://arxiv.org/abs/2010.11929



Vision Transformer

1. Split an image into patches
2. Flatten the patches
3. Produce lower-dimensional linear embeddings from the flattened patches
4. Add positional embeddings
5. Feed the sequence as an input to a standard transformer encoder
6. Pretrain the model with image labels
7. Finetune on the downstream dataset for image classification

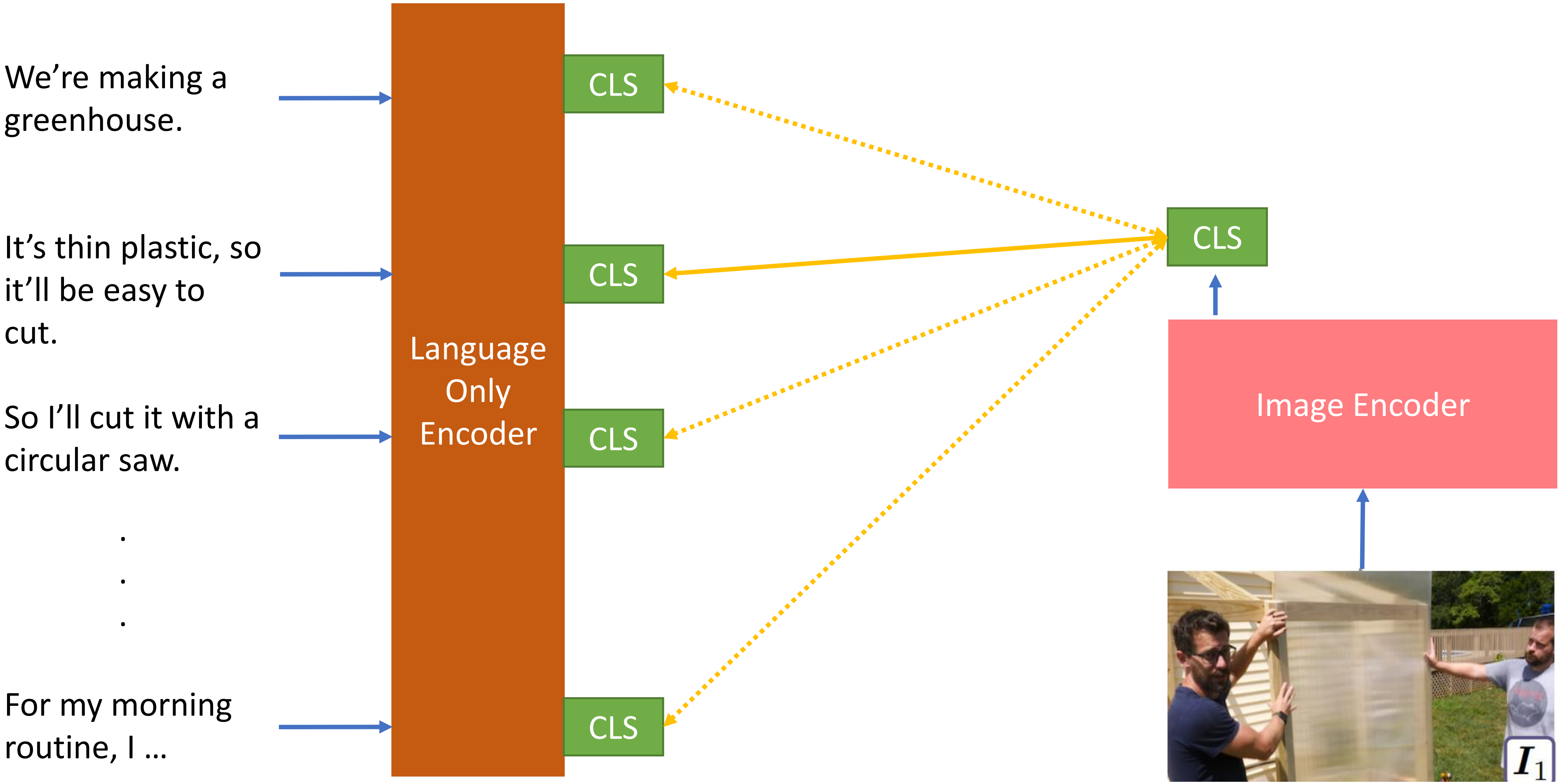
https://arxiv.org/pdf/2010.11929.pdf
https://theaisummer.com/vision-transformer/

MERLOT

- Contrastive frame-transcript matching
- (Attention) Masked Language Modeling
- Temporal Reordering

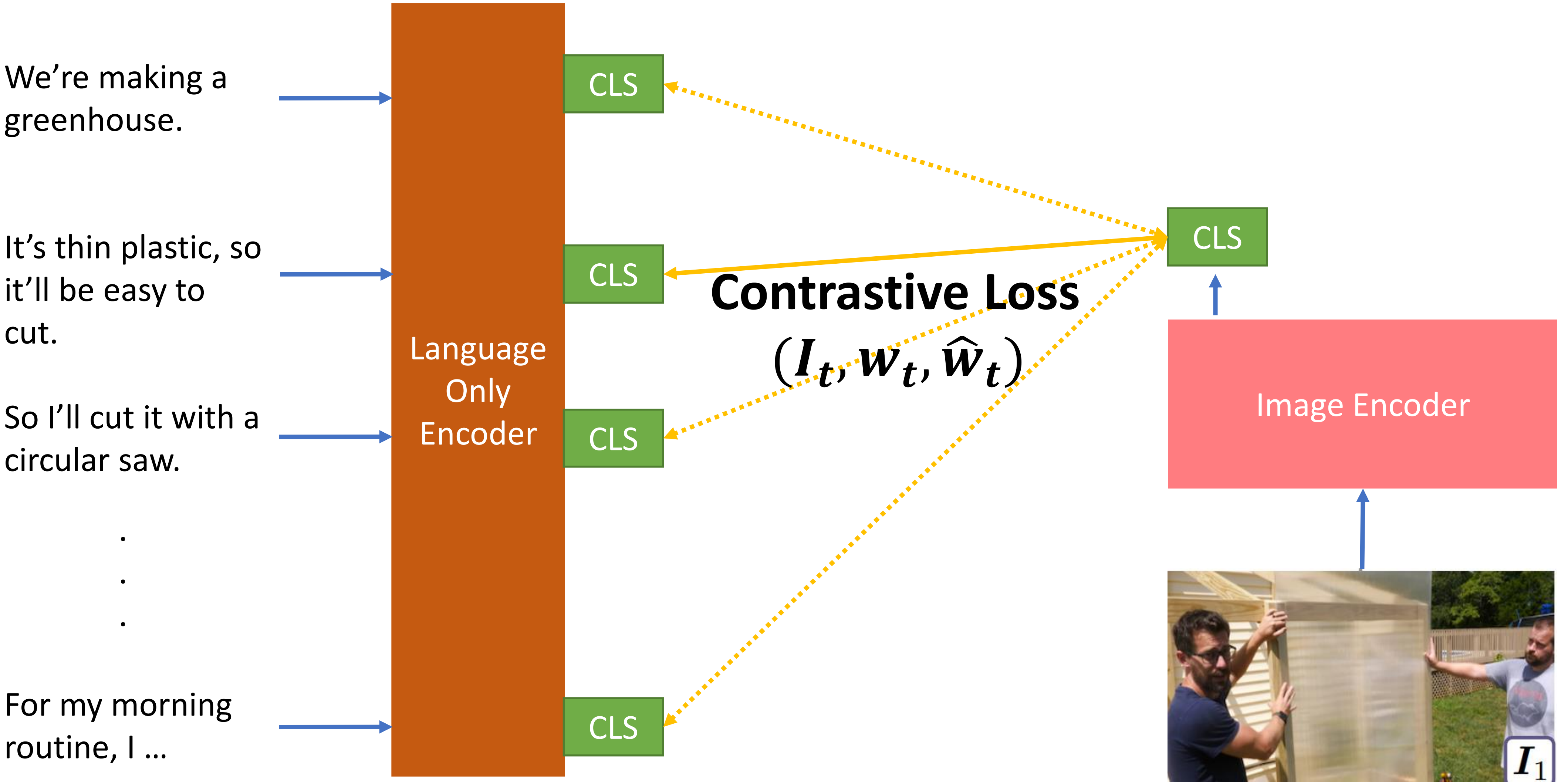
MERLOT

Contrastive frame-transcript matching



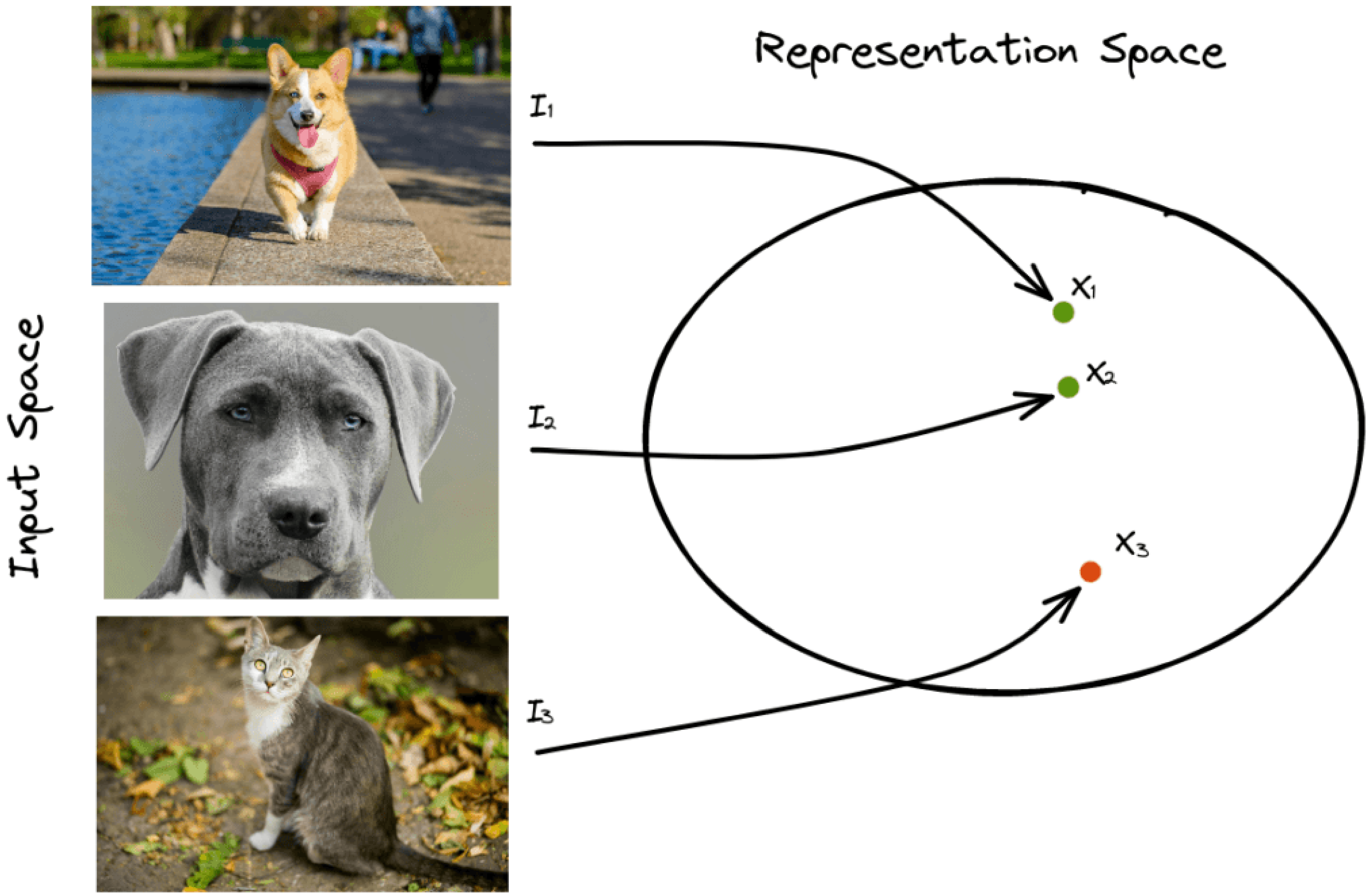
MERLOT

Contrastive frame-transcript matching



Contrastive Loss

Contrastive loss is one of the first training objectives that was used for contrastive learning. **It takes as input a pair of samples that are either similar or dissimilar, and it brings similar samples closer and dissimilar samples far apart.**



https://www.baeldung.com/cs/contrastive-learning#:~:text=Contrastive%20Loss,and%20dissimilar%20samples%20far%20apart.

Contrastive Loss

Contrastive loss is one of the first training objectives that was used for contrastive learning. **It takes as input a pair of samples that are either similar or dissimilar, and it brings similar samples closer and dissimilar samples far apart.**

$$\ell_{i,j} = -\log \frac{\exp(\mathrm{sim}(\boldsymbol{z}_i, \boldsymbol{z}_j)/\tau)}{\sum_{k=1}^{2N} \mathbb{1}_{[k \neq i]} \exp(\mathrm{sim}(\boldsymbol{z}_i, \boldsymbol{z}_k)/\tau)},$$

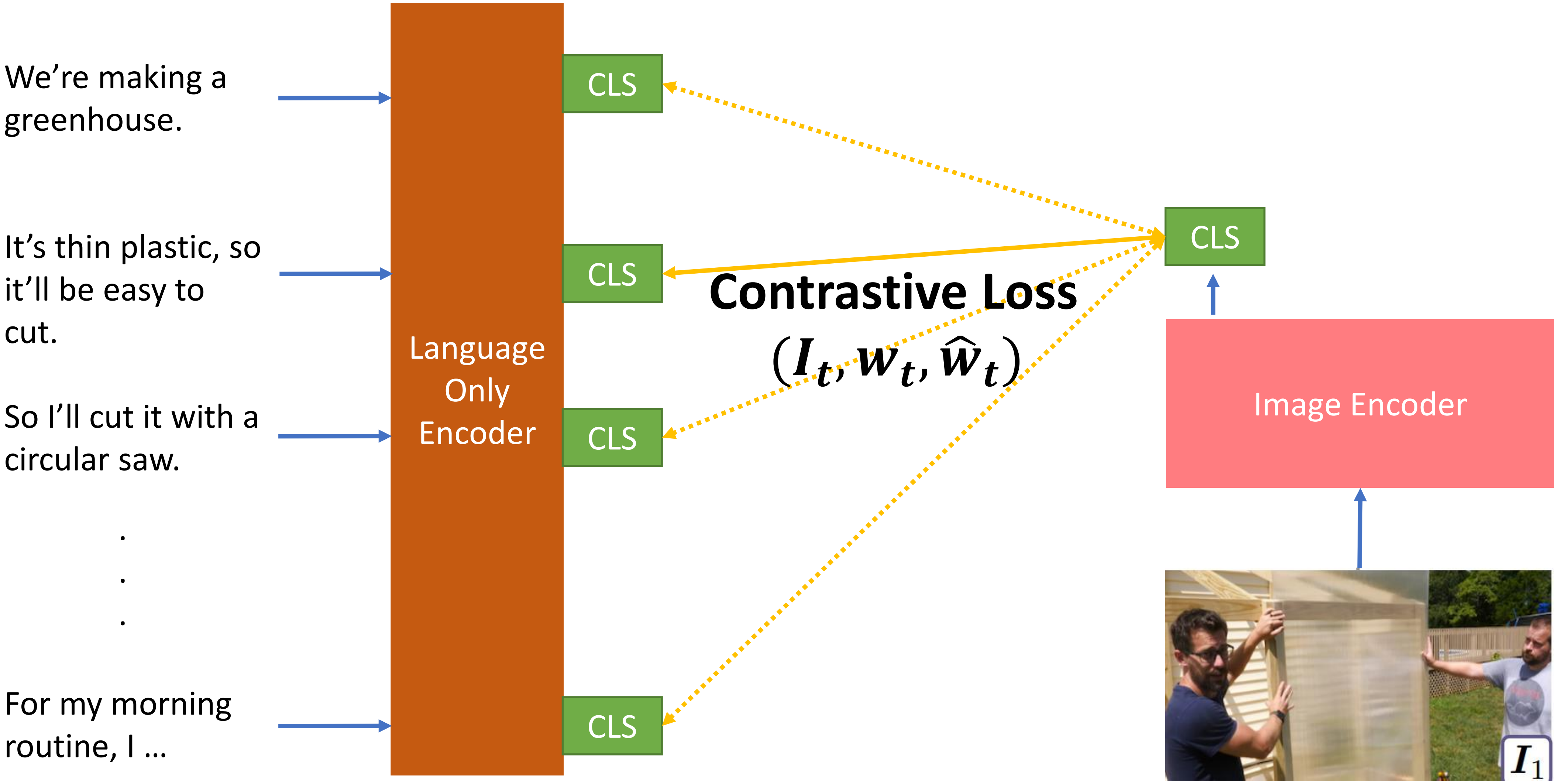
- similar vectors to be as close to 1 as possible, since -log(1) = 0
- negative examples to be close to 0 , since any non-zero values will reduce the value of similar vectors

Softmax $\sigma(\mathbf{z})_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}}$ for $i = 1, \dots, K$ and $\mathbf{z} = (z_1, \dots, z_K) \in \mathbb{R}^K$

https://www.baeldung.com/cs/contrastive-learning#:~:text=Contrastive%20Loss,and%20dissimilar%20samples%20far%20apart.

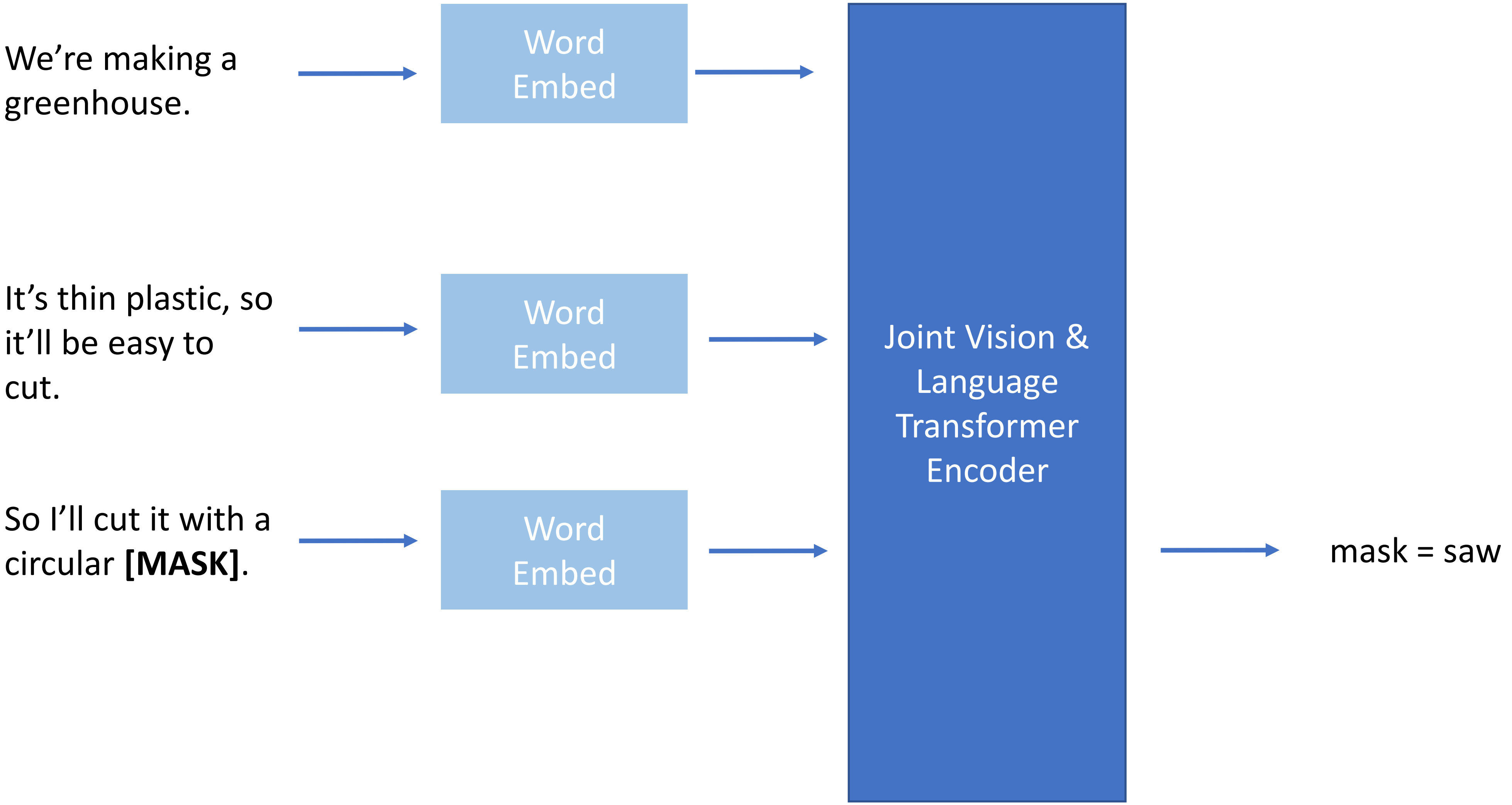
MERLOT

Contrastive frame-transcript matching



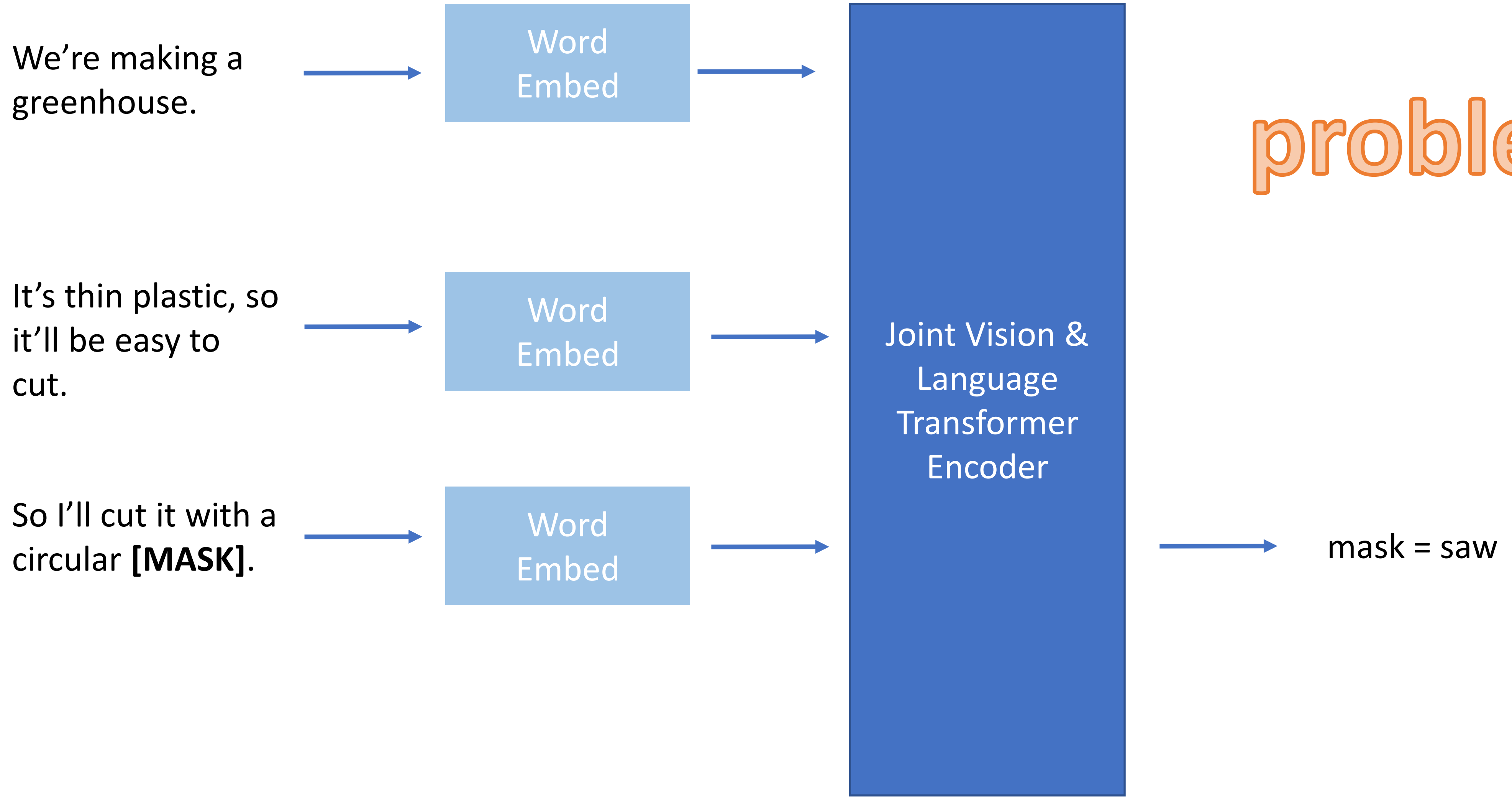
MERLOT

(*Attention*) Masked Language Modeling



MERLOT

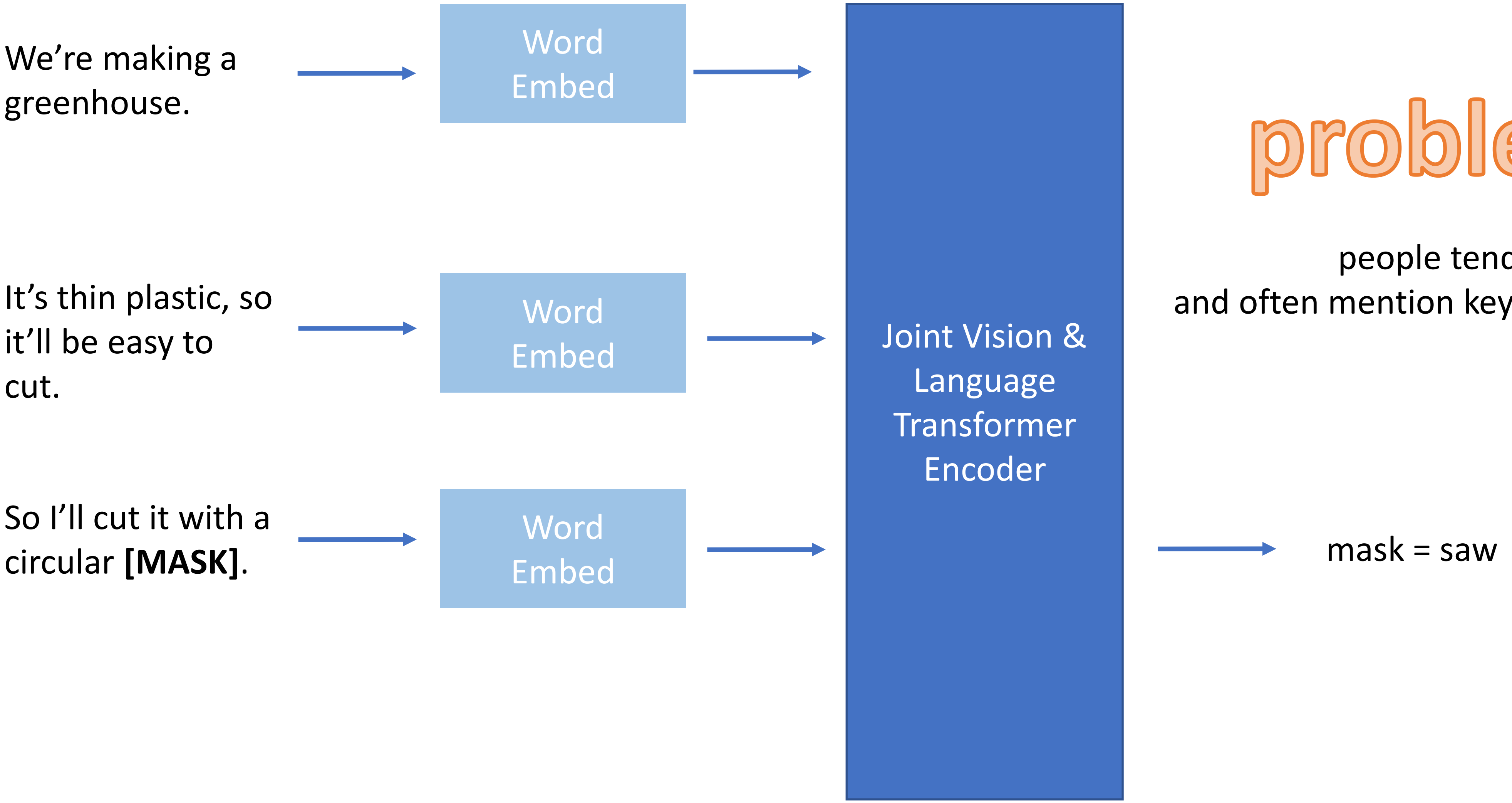
(*Attention*) Masked Language Modeling



problems?

MERLOT

(*Attention*) Masked Language Modeling



problems?

people tend to ramble,
and often mention key objects multiple times

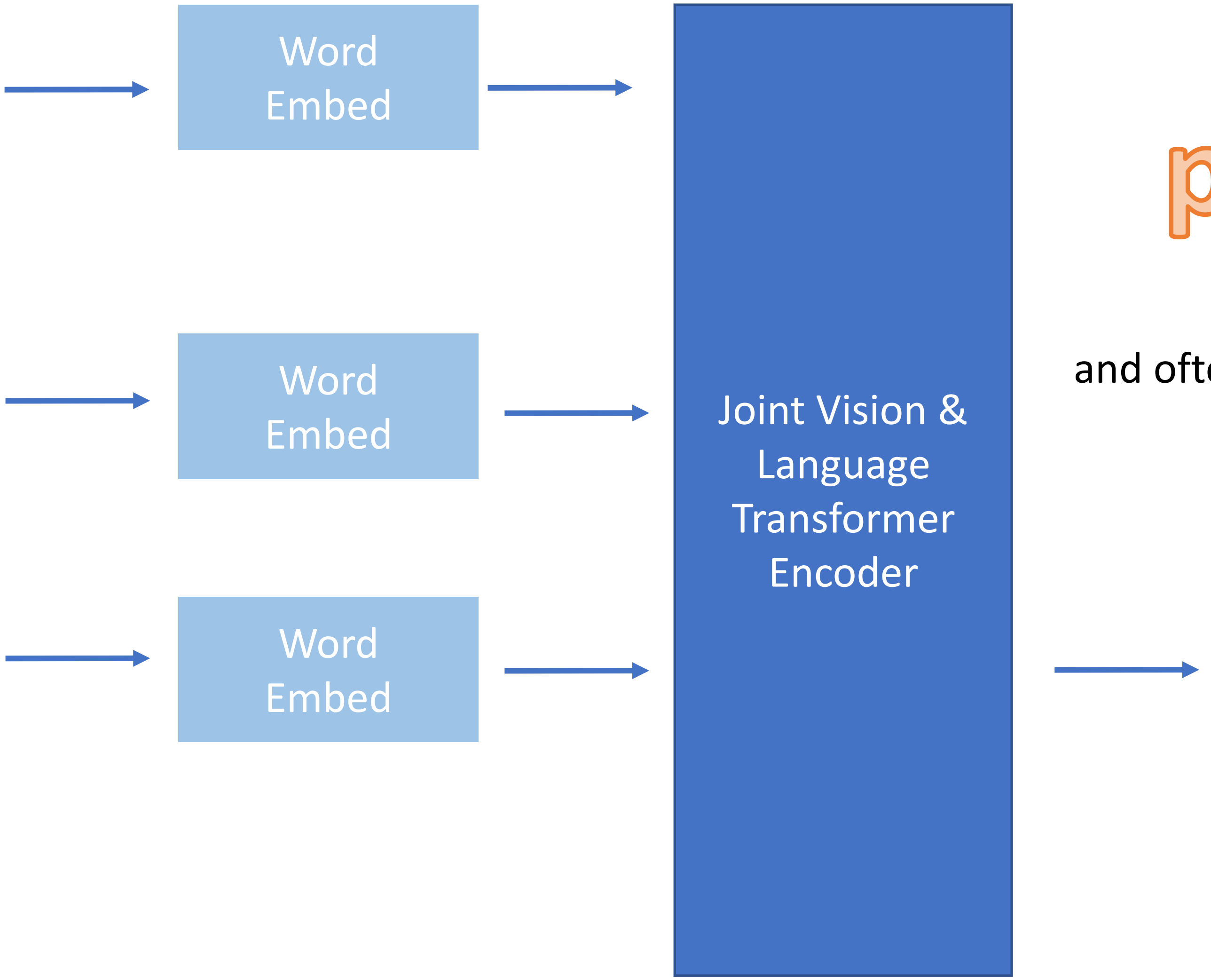
MERLOT

(*Attention*) Masked Language Modeling

We're making a greenhouse.

It's thin plastic, so it'll be easy to cut.

So I'll cut it with a circular **[MASK]**.



mask = saw

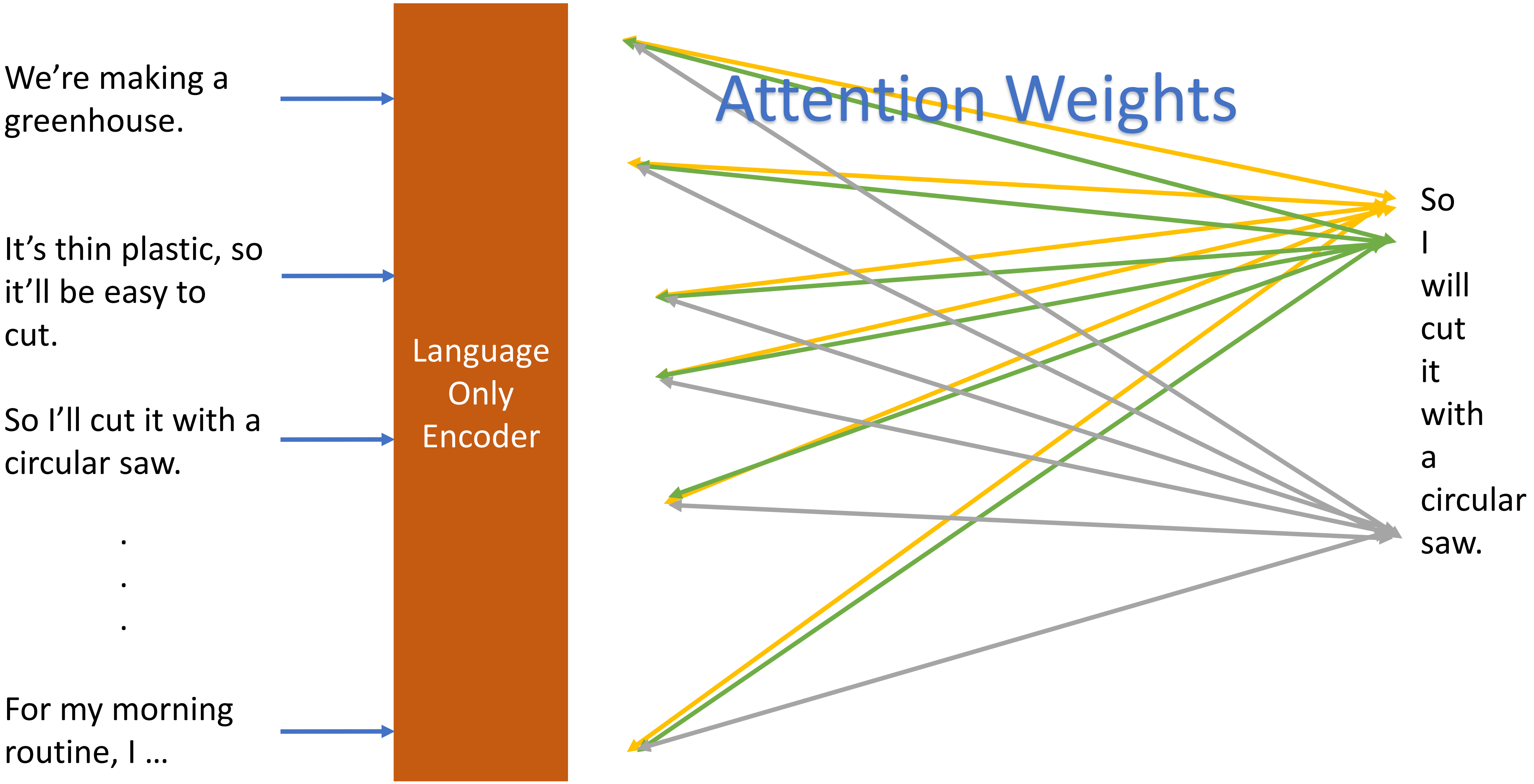
problems?

people tend to ramble,
and often mention key objects multiple times

Attention Masking

MERLOT

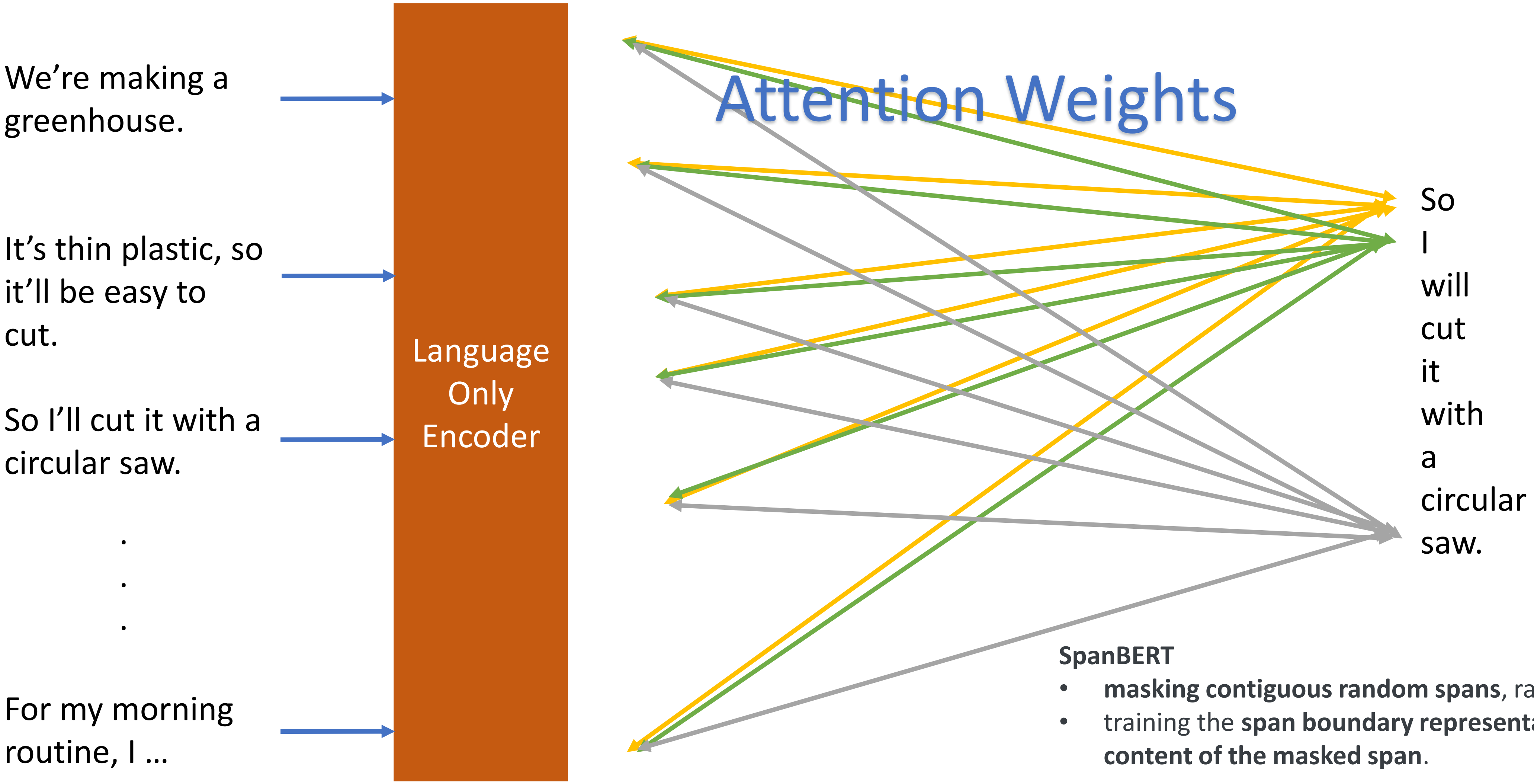
(*Attention*) Masked Language Modeling



https://arxiv.org/pdf/1907.10529.pdf

MERLOT

(*Attention*) Masked Language Modeling

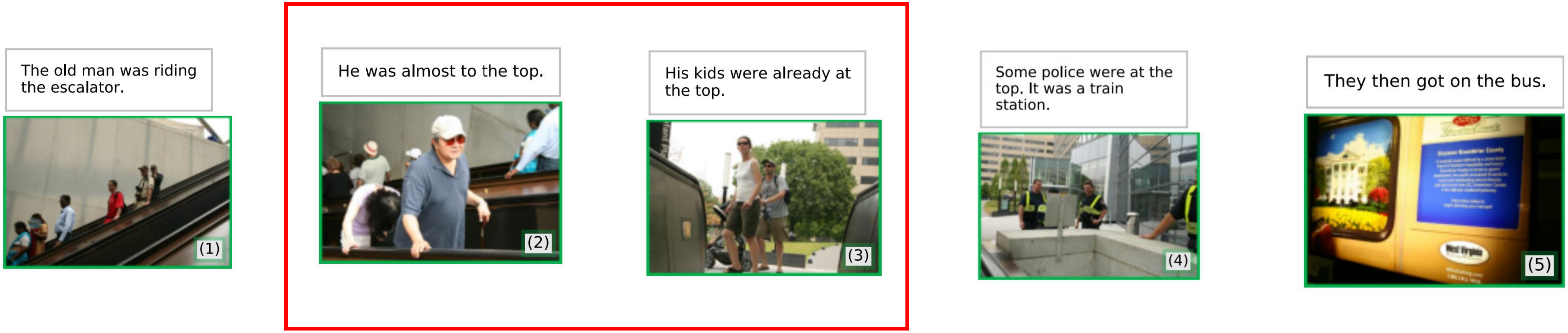


SpanBERT

- **masking contiguous random spans**, rather than random tokens
- training the **span boundary representations to predict the entire content of the masked span**.

MERLOT

Temporal Reordering

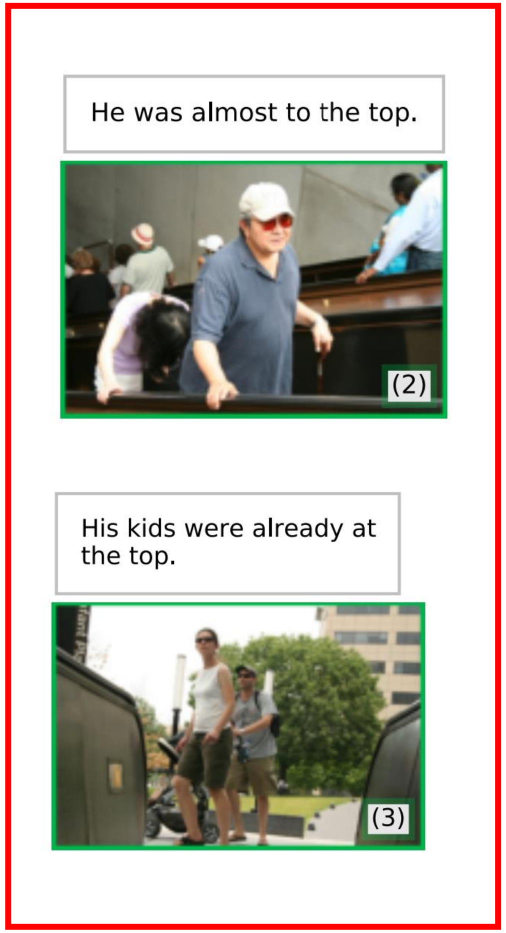


Replace segment-level position embeddings

[image_t] -> [image_unk_0]

MERLOT

Temporal Reordering



Reordering Loss

$concat(h_{ti}, h_{ti})$

MLP

Cross Entropy

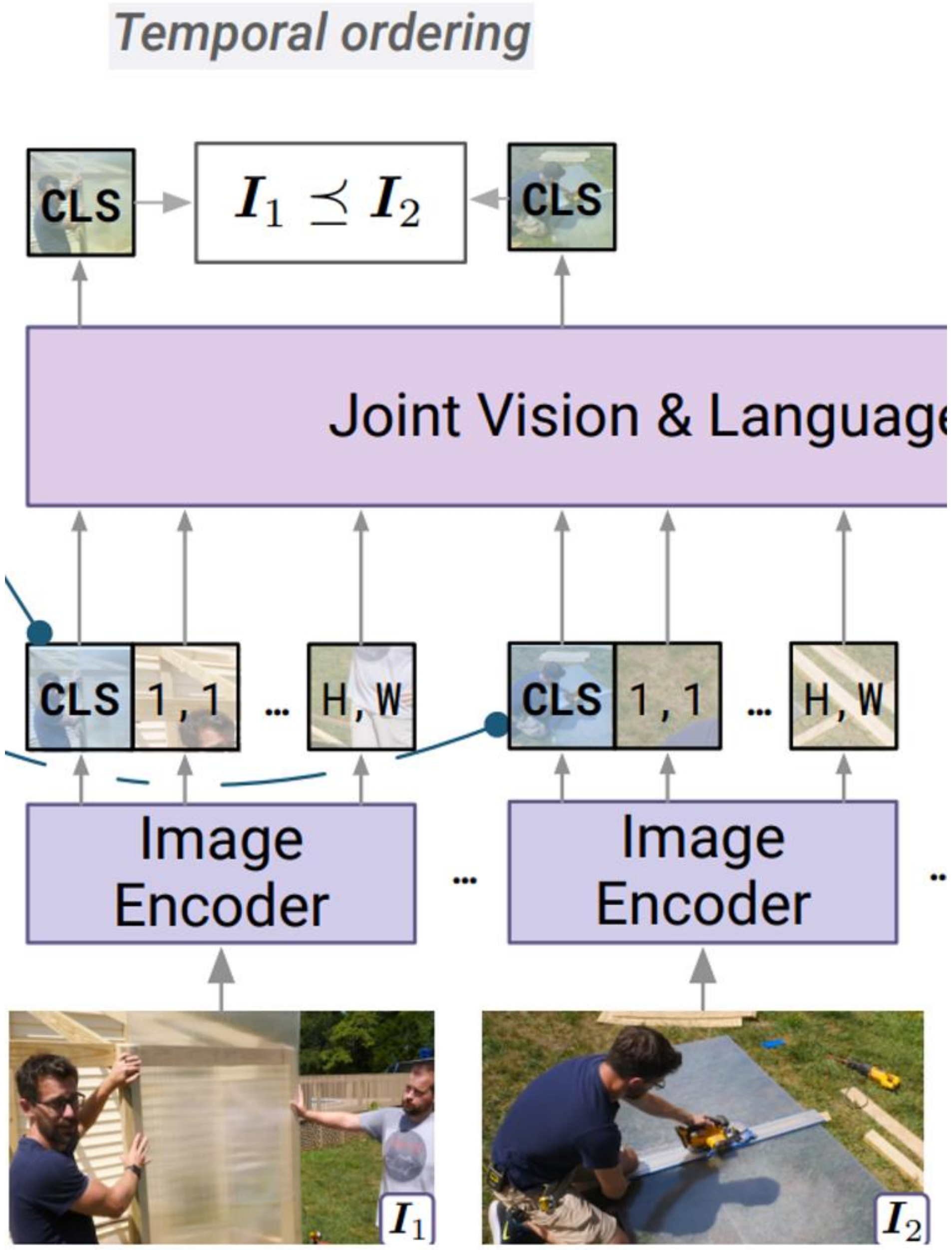
$t_i < t_j$ or $t_i > t_j$

MERLOT

Temporal Reordering

Reordering Loss

$concat(h_{ti}, h_{ti})$ → MLP → Cross Entropy $t_i < t_j$ or $t_i > t_j$



MERLOT

Architecture

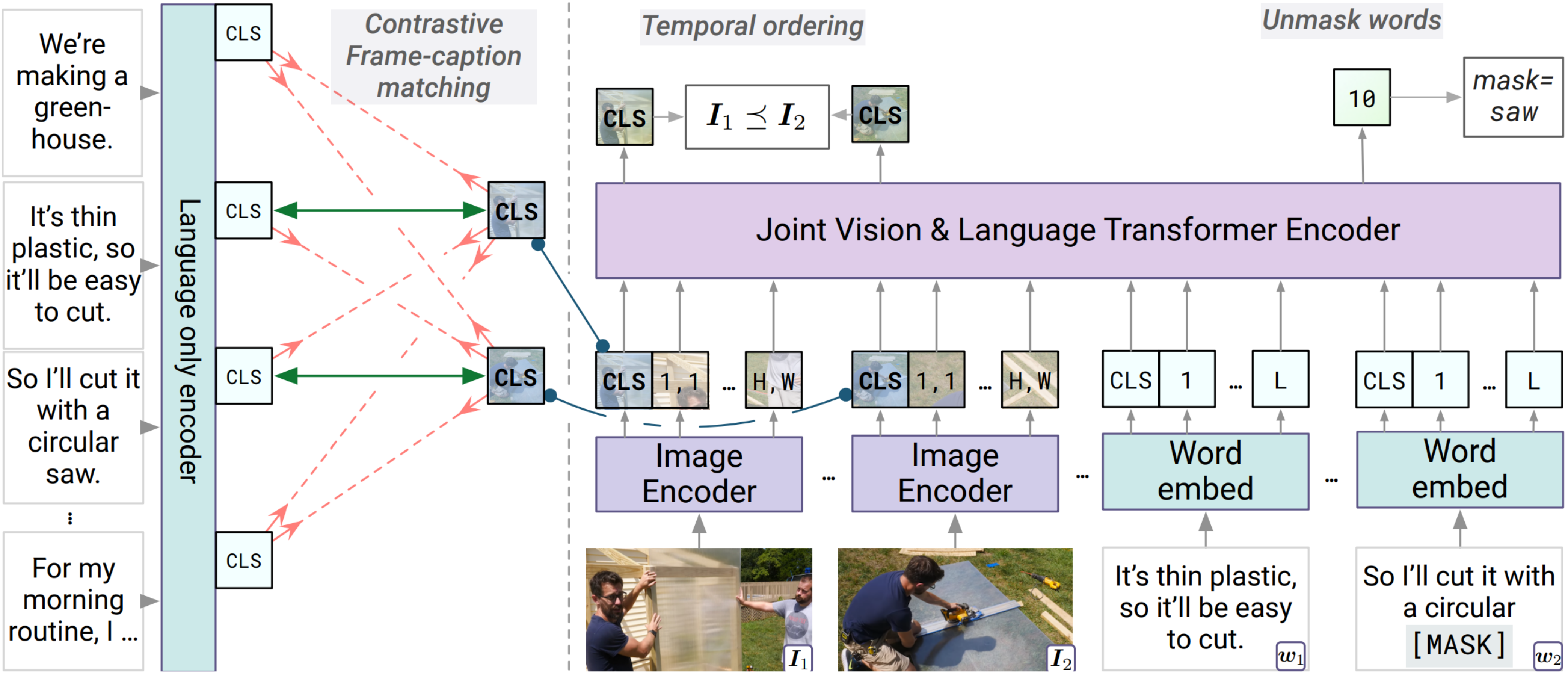
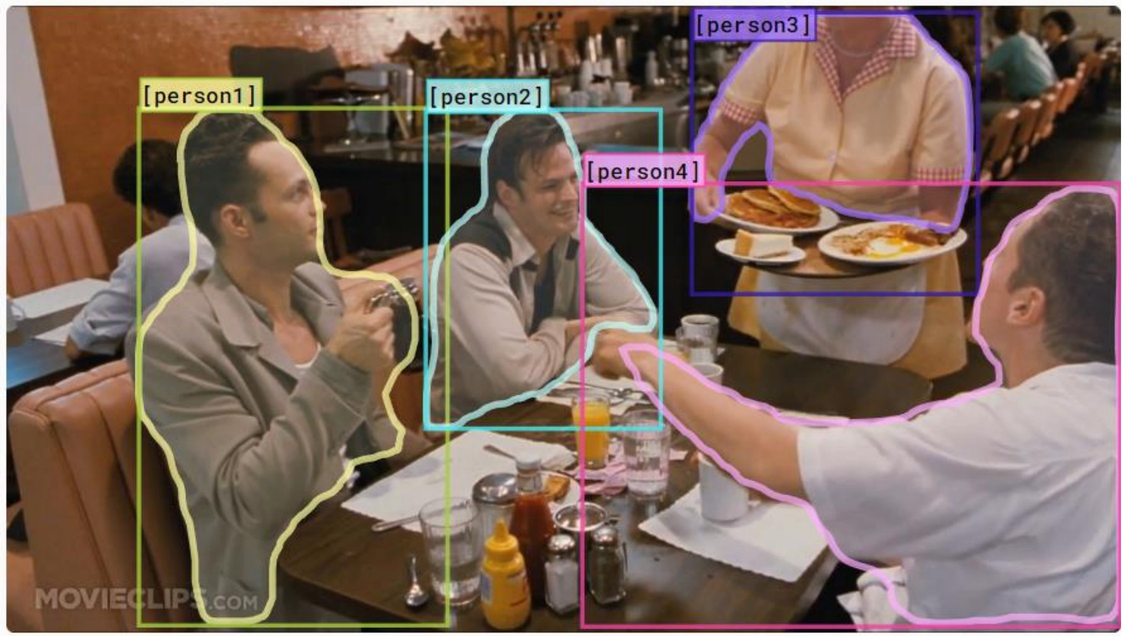


Image Tasks

Visual Commonsense Reasoning



Why is [person4] pointing at [person1]?

a) He is telling [person3] that [person1] ordered the pancakes.
b) He just told a joke.
c) He is feeling accusatory towards [person1].
d) He is giving [person1] directions.

Rationale: I think so because...

a) [person1] has the pancakes in front of him.
b) [person4] is taking everyone's order and asked for clarification.
c) [person3] is looking at the pancakes both she and [person2] are smiling slightly.
d) [person3] is delivering food to the table, and she might not know whose order is whose.

- what might happen next
- what are people’s intentions

https://visualcommonsense.com/

Image Tasks

Visual Commonsense Reasoning

	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**

Unsupervised ordering of Visual Stories

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

https://visualcommonsense.com/

Video Tasks

TVQA

TVQA links depicted objects to visual concepts in questions and answers.

Question: What is Sheldon holding when he is talking to Howard about the sword?
Correct Answer: A computer.

Question: Who is talking to Howard when he is in the kitchen upset?
Correct Answer: Raj is talking to Howard.

- what might happen next
- what are people's intentions

https://tvqa.cs.unc.edu/

Video Tasks

Tasks	Split	Vid. Length	ActBERT [127]	ClipBERT$_{8x2}$ [67]	SOTA	MERLOT
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**
ActivityNetQA	Test	Long	-	-	38.9 [118]	**41.4**
Drama-QA	Val	Long	-	-	81.0 [56]	**81.4**
TVQA	Test	Long	-	-	76.2 [56]	**78.7**
TVQA+	Test	Long	-	-	76.2 [56]	**80.9**
VLEP	Test	Long	-	-	67.5 [66]	**68.4**

Zero-shot Ordering

The old man was riding the escalator.

He was almost to the top.

His kids were already at the top.

Some police were at the top. It was a train station.

They then got on the bus.

I went to the fair with my kids last weekend.

There were a lot of people there.

They also had a barn.

We got to see a lot of animals.

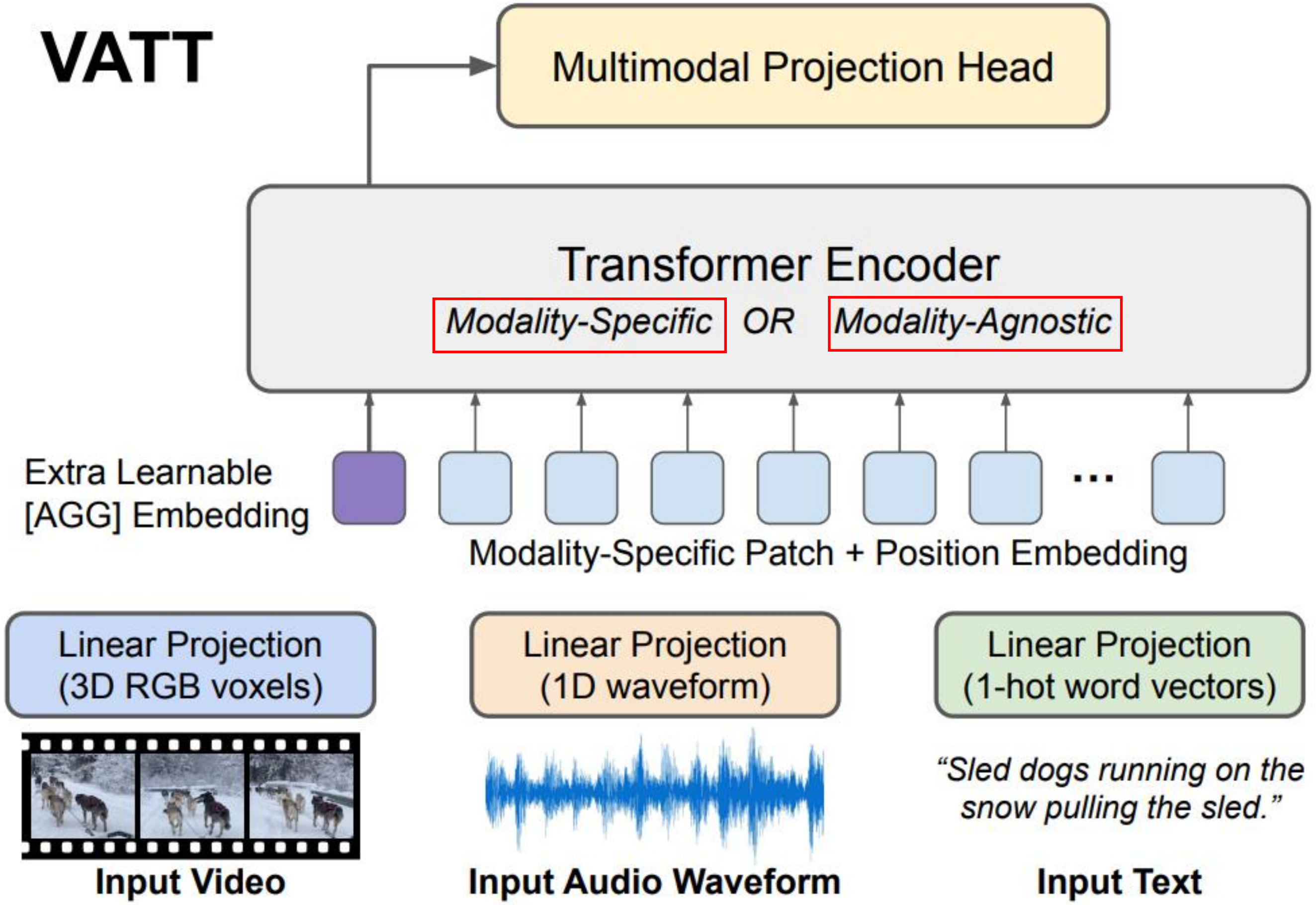
We can't wait to go back later.

VATT

Transformers for Multimodal Self-Supervised Learning from Raw Video, Audio and Text

VATT

One backbone share different modalities

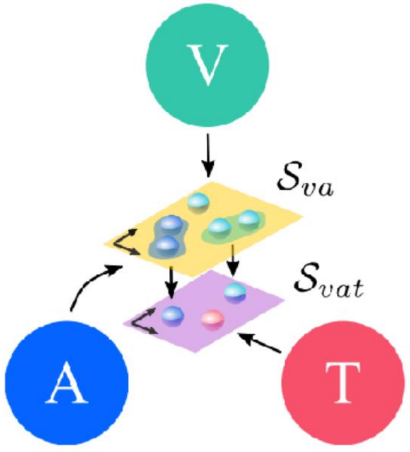


https://openreview.net/pdf?id=RzYrn625bu8

VATT

Modality-Specific

- Video, audio, and text inputs have respective feature extractors
- Each feature extractor has different architecture according to the modality.



(c) FAC

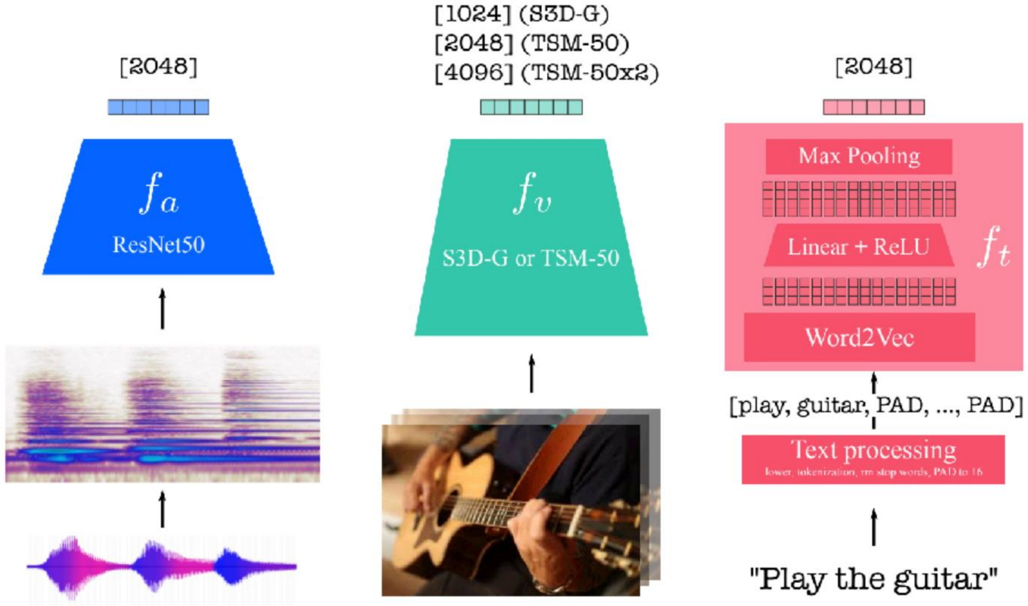
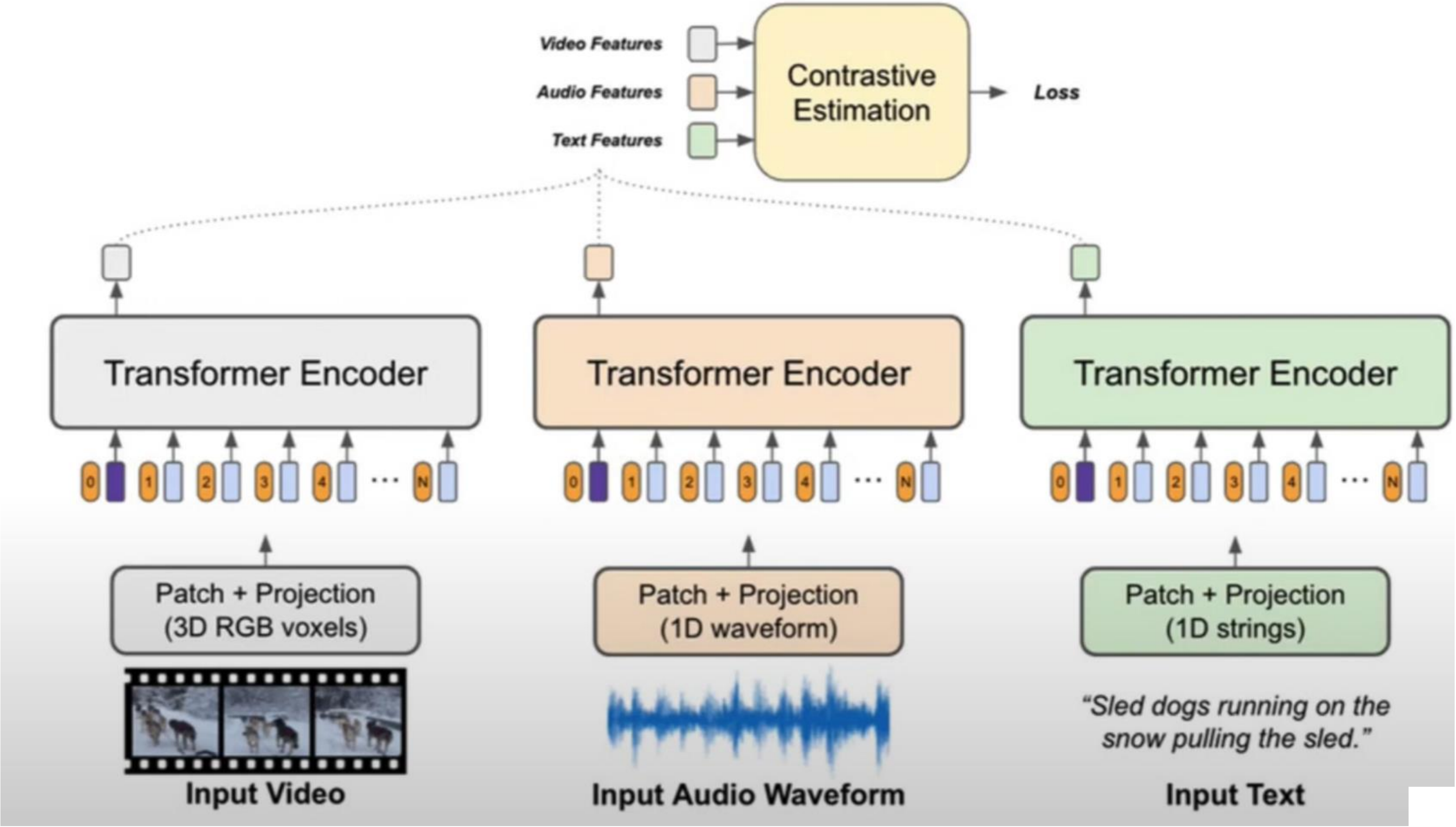


Figure 2: Backbone architecture for audio, vision and text.

FAC: Fine and Coarse

VATT

Modality-Specific

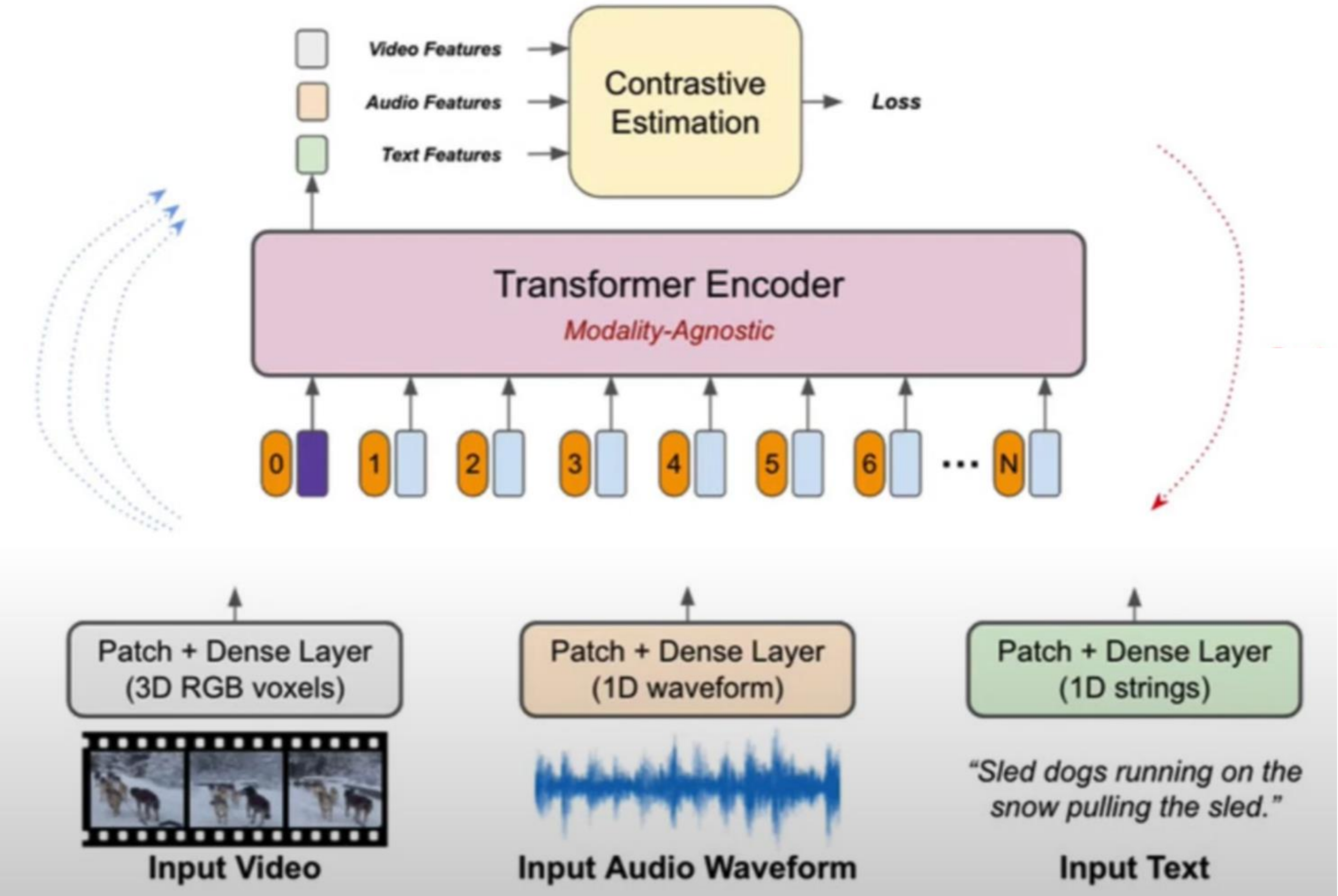


- Video, audio, and text inputs have respective feature extractors
- Each feature extractor has different architecture according to the modality.

https://www.youtube.com/watch?v=V3gY_hyATU8

VATT

Modality-Agnostic

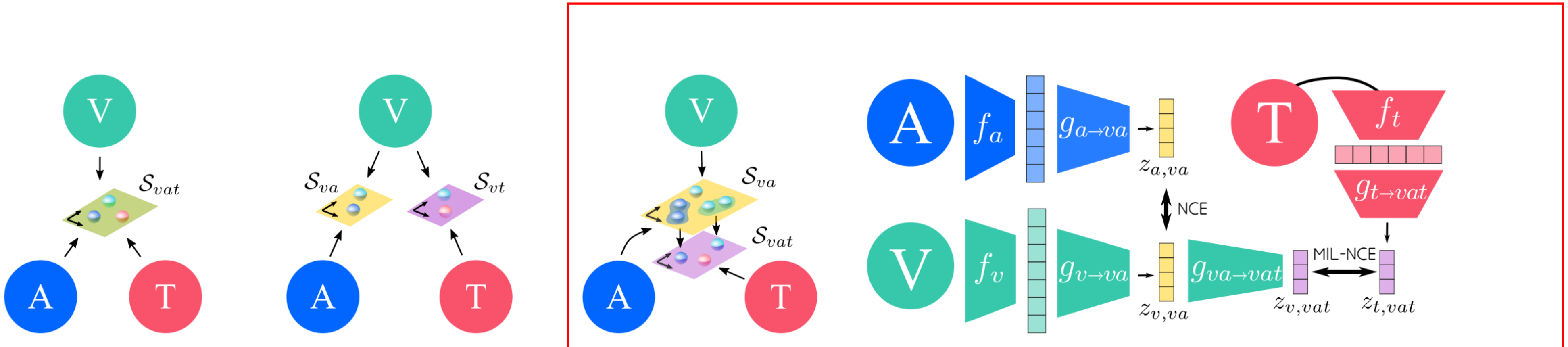


https://www.youtube.com/watch?v=V3gY_hyATU8

VATT

Common Space Projection

- Multi-modal features need to be projected to common space for feature fusion, but different modalities have different levels of semantic granularity
- Vision & audio: fine-grained space
- Vision + audio & text: the lower dimensional coarse-grained space.



(a) **Shared** (b) **Disjoint** (c) **FAC** (d) **FAC details**

https://arxiv.org/pdf/2006.16228.pdf

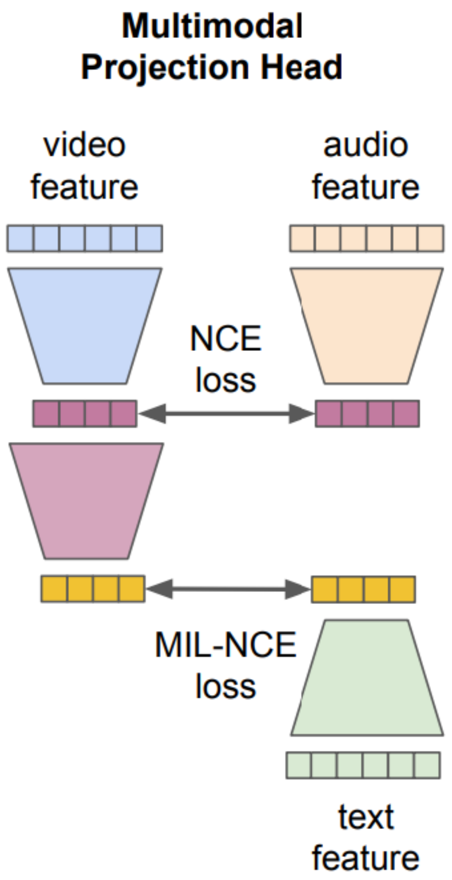
VATT

Multimodal Contrastive Learning

$$\mathcal{L} = \text{NCE}(\boldsymbol{z}_{v,va}, \boldsymbol{z}_{a,va}) + \lambda \text{MIL-NCE}(\boldsymbol{z}_{v,vt}, \{\boldsymbol{z}_{t,vt}\})$$

$$\text{NCE}(\boldsymbol{z}_{v,va}, \boldsymbol{z}_{a,va}) = -\log\left(\frac{\exp(\boldsymbol{z}_{v,va}^\top \boldsymbol{z}_{a,va}/\tau)}{\exp(\boldsymbol{z}_{v,va}^\top \boldsymbol{z}_{a,va}/\tau) + \sum_{z'\in\mathcal{N}} \exp(\boldsymbol{z'}_{v,va}^\top \boldsymbol{z'}_{a,va}/\tau)}\right)$$

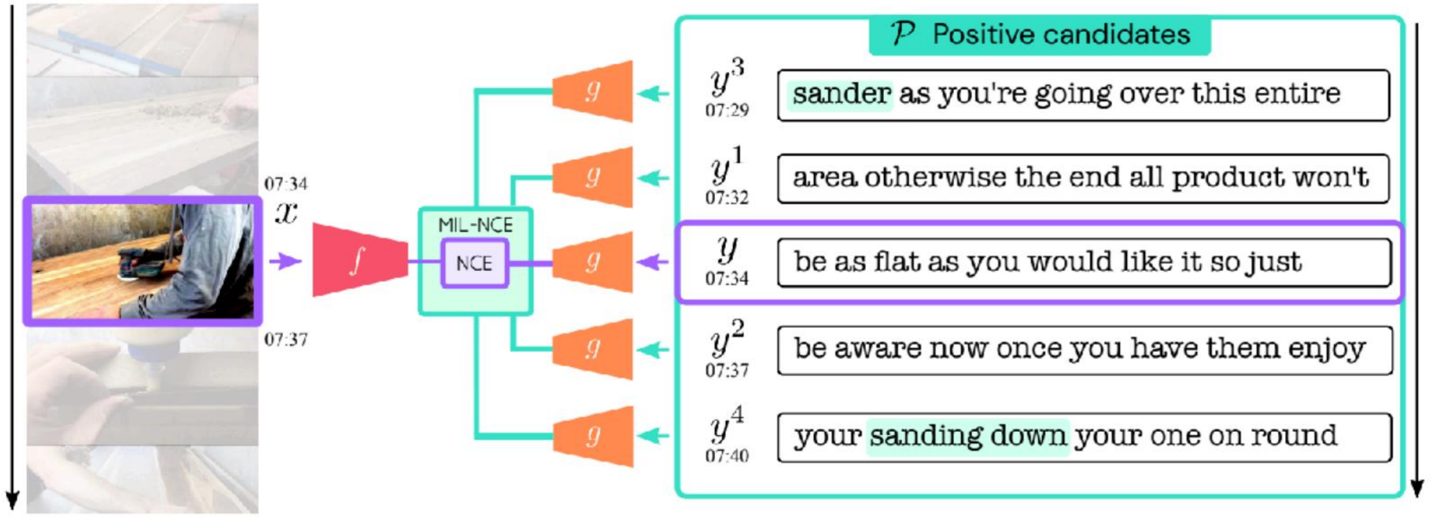
$$\text{MIL-NCE}(\boldsymbol{z}_{v,vt}, \{\boldsymbol{z}_{t,vt}\}) = -\log\left(\frac{\sum_{\boldsymbol{z}_{t,vt}\in\mathcal{P}} \exp(\boldsymbol{z}_{v,vt}^\top \boldsymbol{z}_{t,vt}/\tau)}{\sum_{\boldsymbol{z}_{t,vt}\in\mathcal{P}} \exp(\boldsymbol{z}_{v,vt}^\top \boldsymbol{z}_{t,vt}/\tau) + \sum_{z'\in\mathcal{N}} \exp(\boldsymbol{z'}_{v,vt}^\top \boldsymbol{z'}_{t,vt}/\tau)}\right)$$



VATT

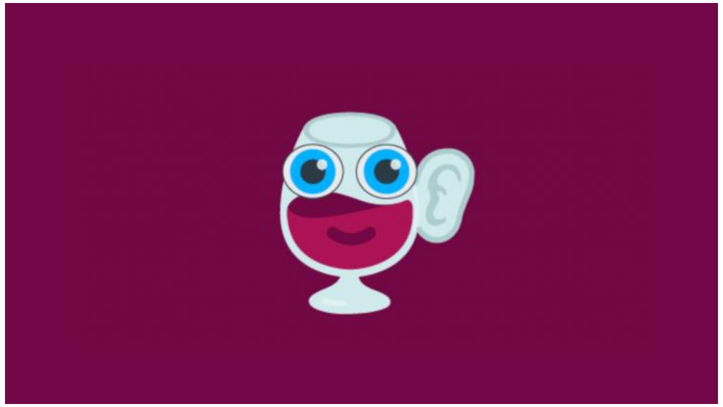
Multimodal Contrastive Learning

- Vision & Text: Multiple-Instance-Learning-NCE (MIL-NCE) loss
- For multiple positive pairs of video & text, a video is matched to multiple text inputs that are temporally close to the video input.



(a) **Examples of positive candidates**

End-to-End Learning of Visual Representation from Uncurated Instructional Videos (CVPR'20)



MERLOT RESERVE

Neural Script Knowledge through Vision and Language and Sound

MERLOT RESERVE

Segment 3 - s_3 (5 seconds)

....

....

v_1 v_{t-n} v_t

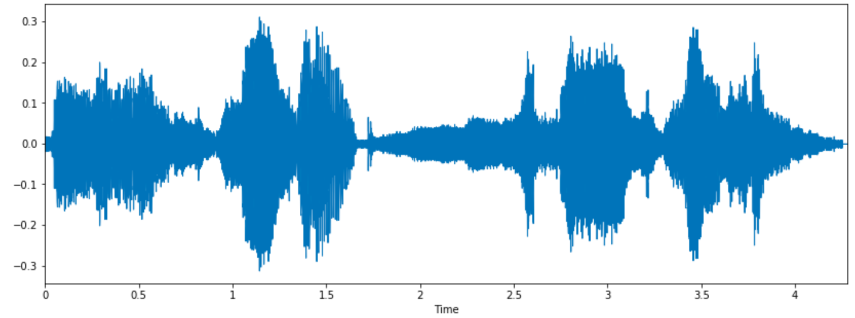
w_1 At 8 a.m. today, someone poisons the coffee.

w_2 Do not drink the coffee.

w_3 No~~~

...

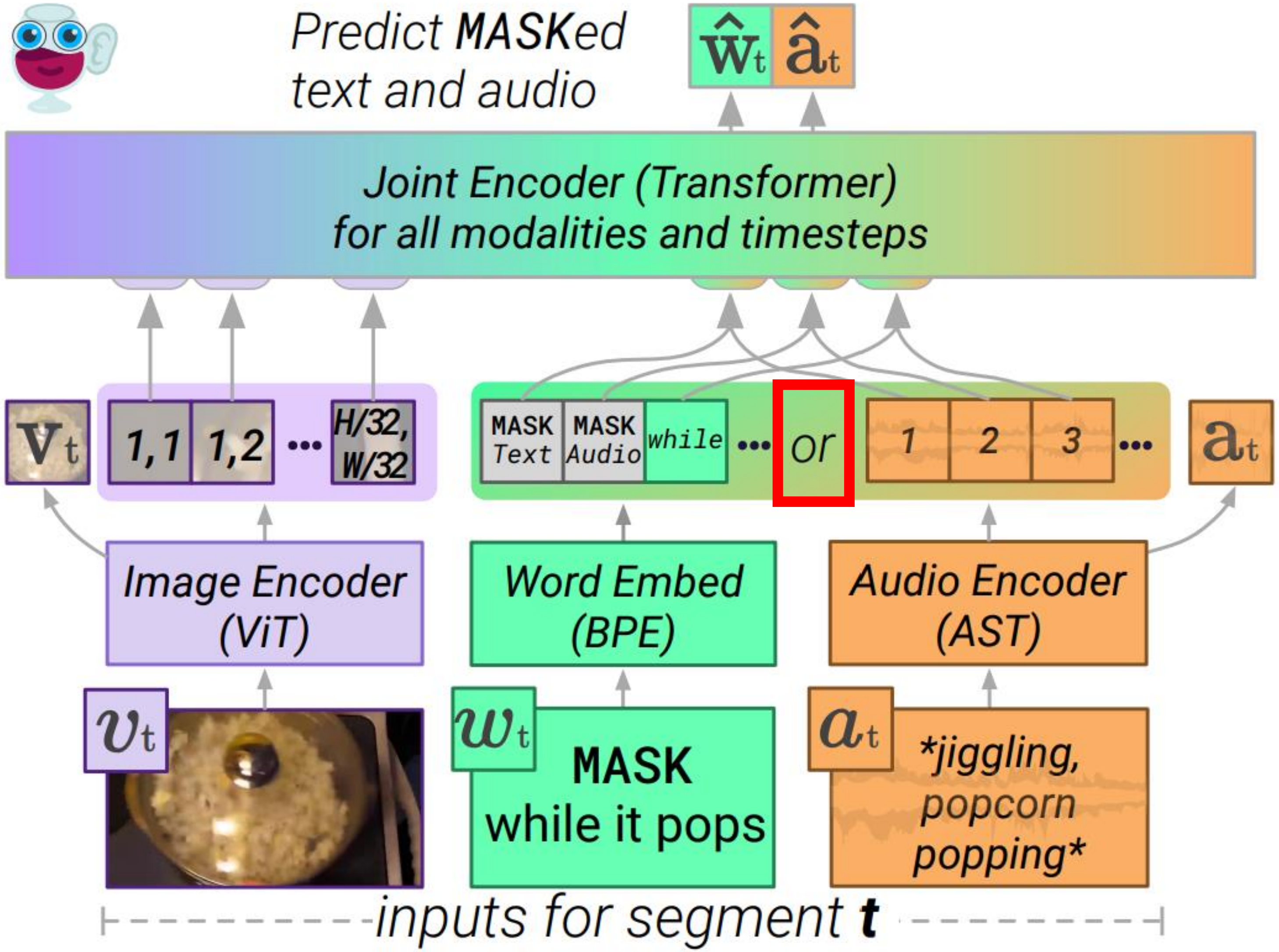
a_t



- A frame v_t, from the middle of the segment
- The ASR tokens w_t spoken during the segment
- The audio a_t of the segment.

MERLOT RESERVE

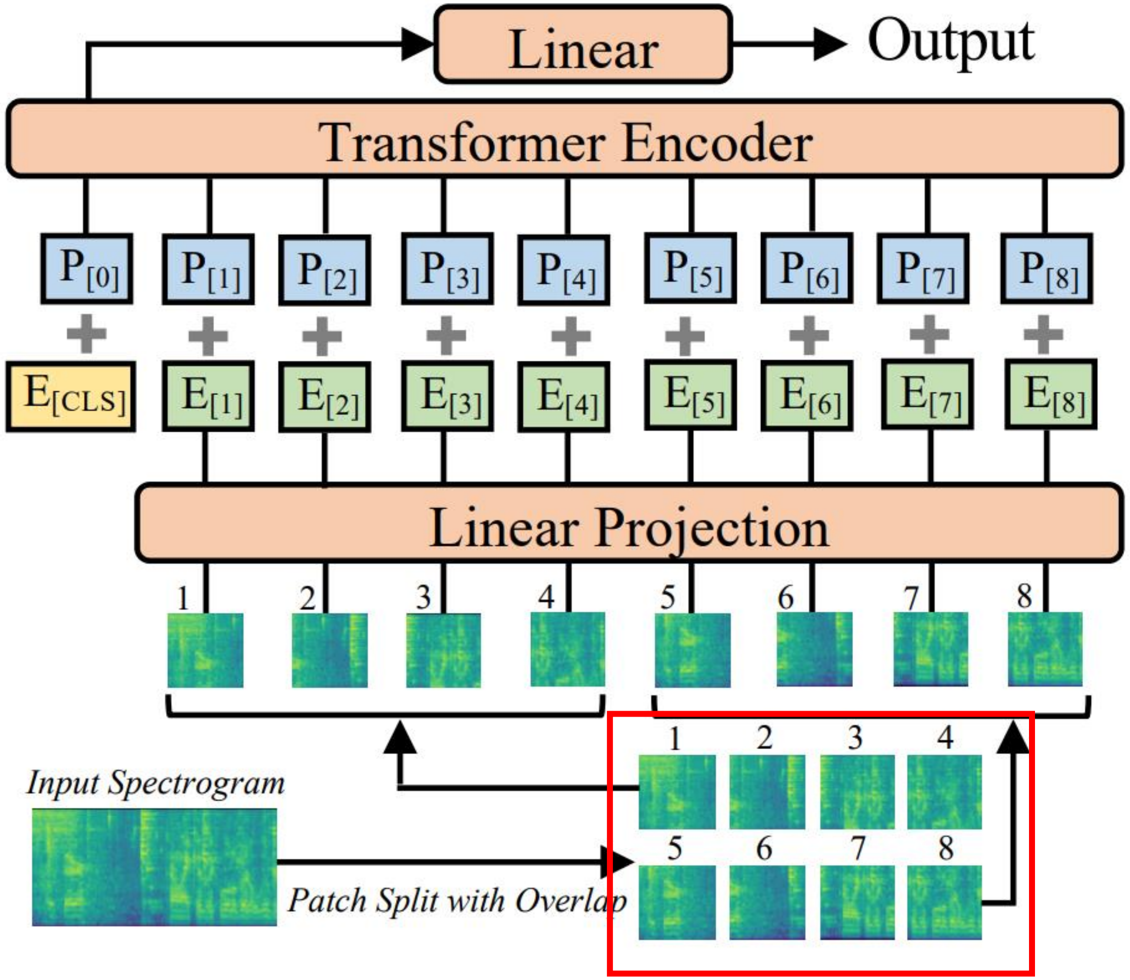
As the text w_t was automatically transcribed by a model given audio a_t, it is reasonable to assume that it contains strictly less information content. Thus, for each segment s_t, the paper provides models with exactly one of text **or** audio.



MERLOT RESERVE

Audio Encoder

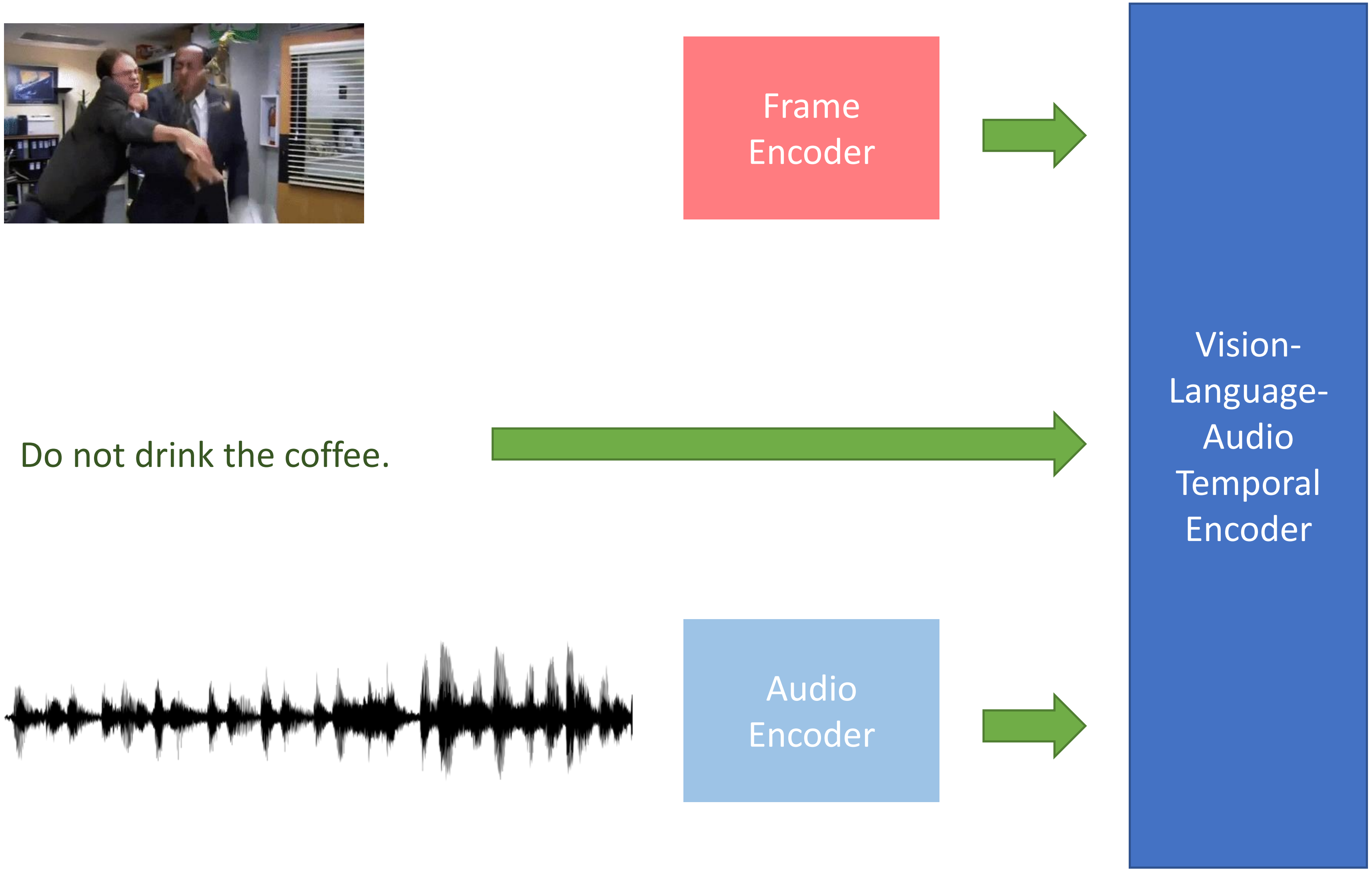
AST:Audio Spectrogram Transformer



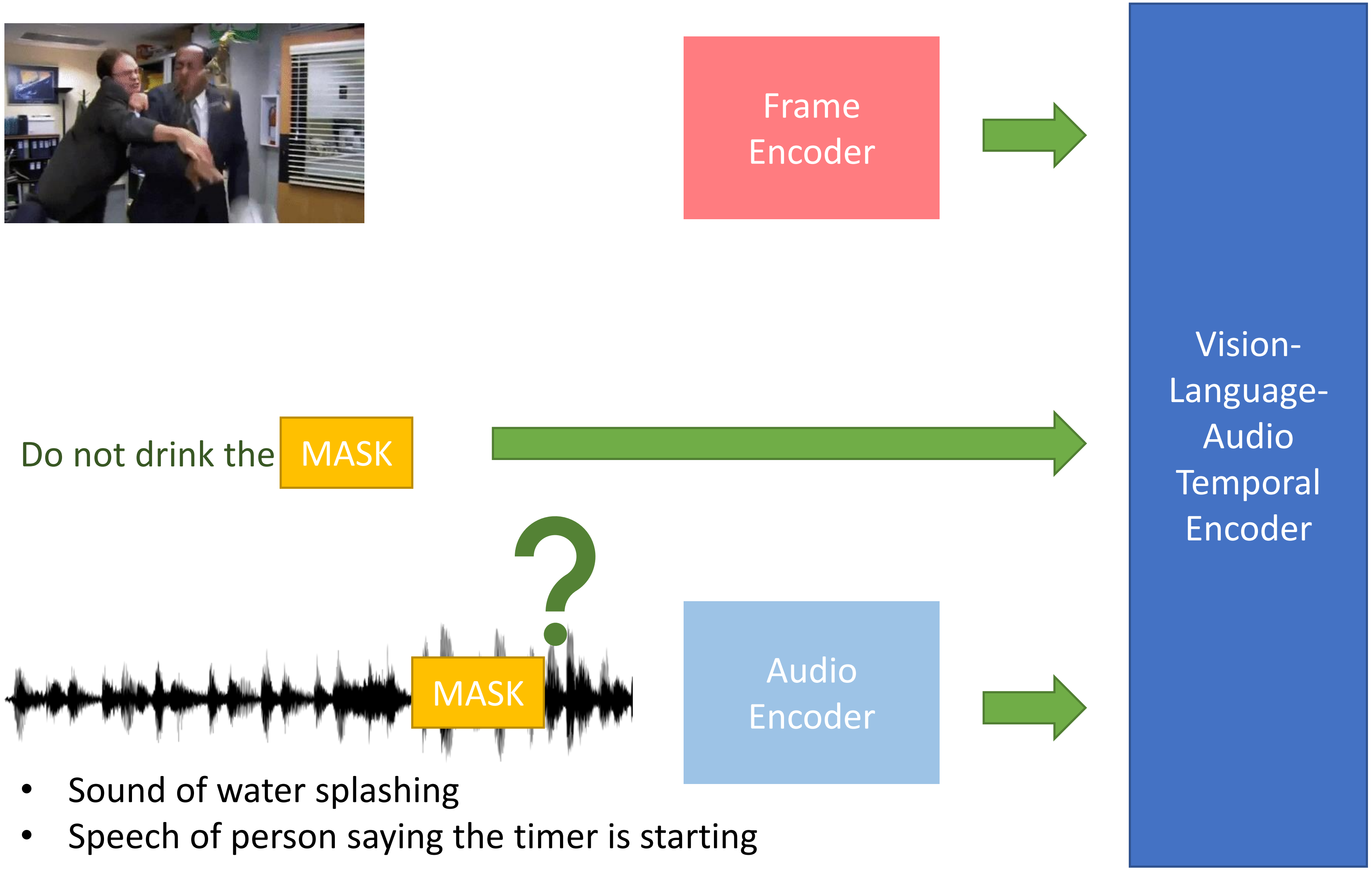
Split the audio a_t in each segment into three equal-sized subsegments

https://arxiv.org/pdf/2104.01778.pdf

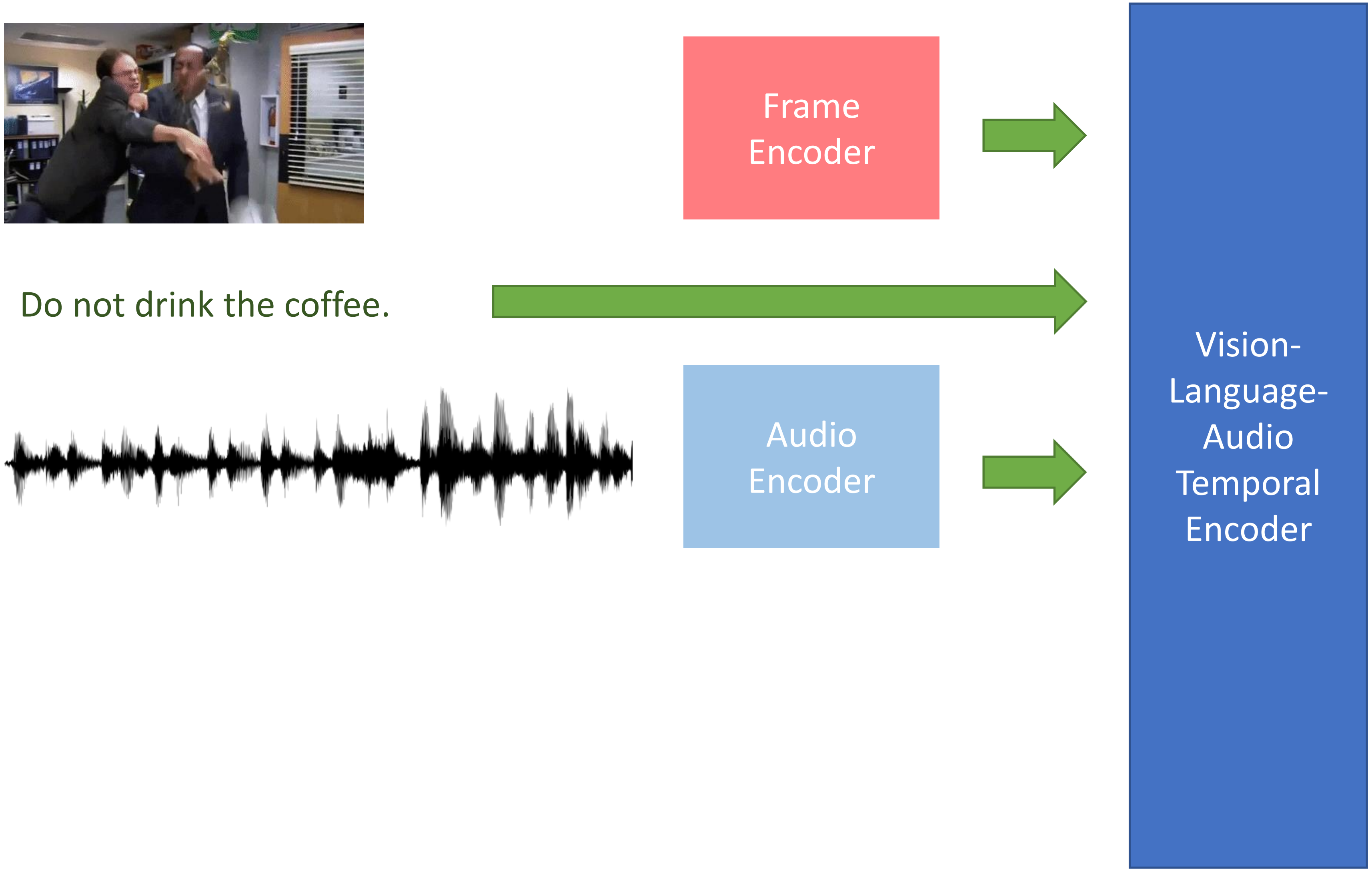
MERLOT RESERVE



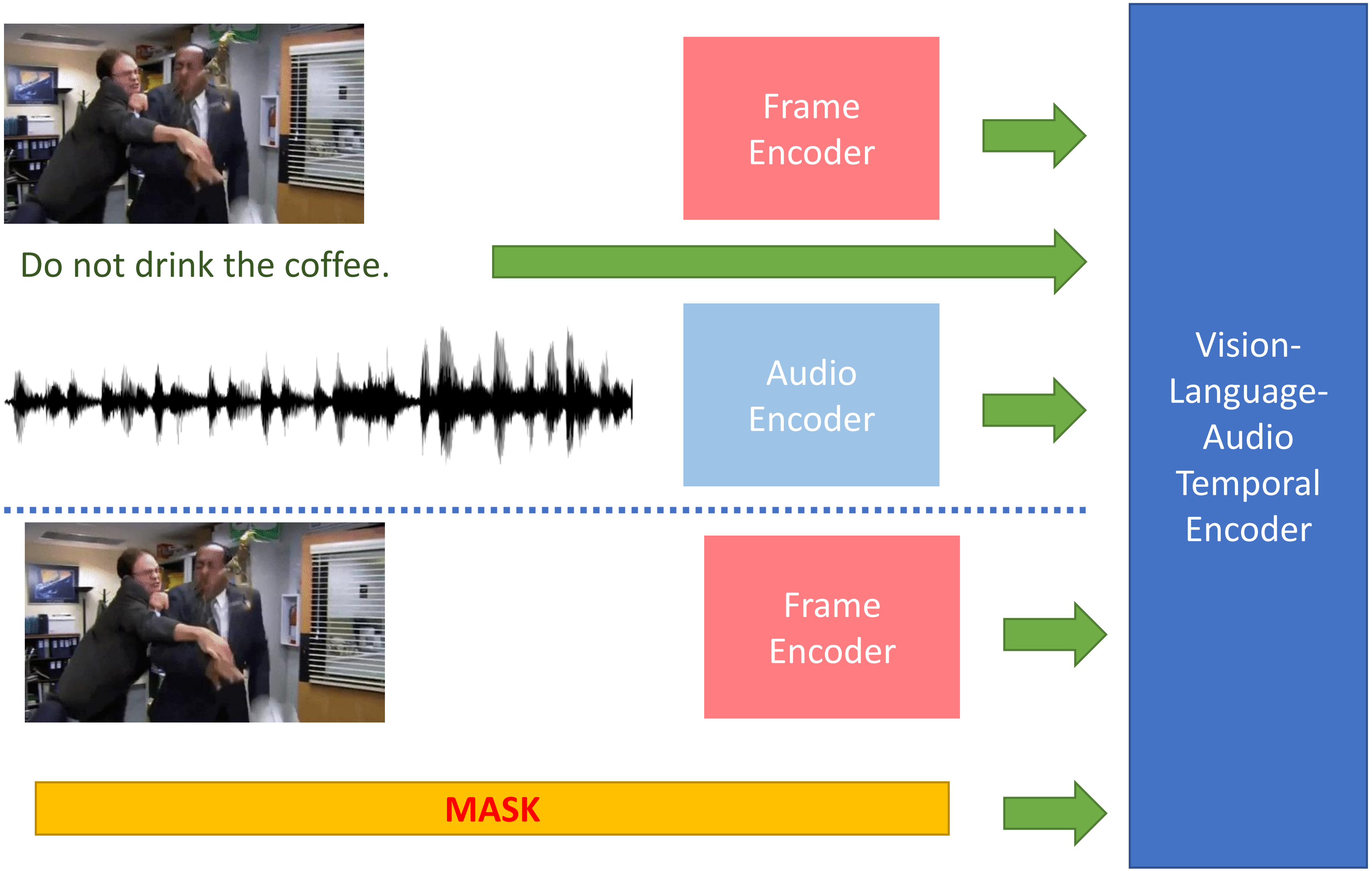
MERLOT RESERVE



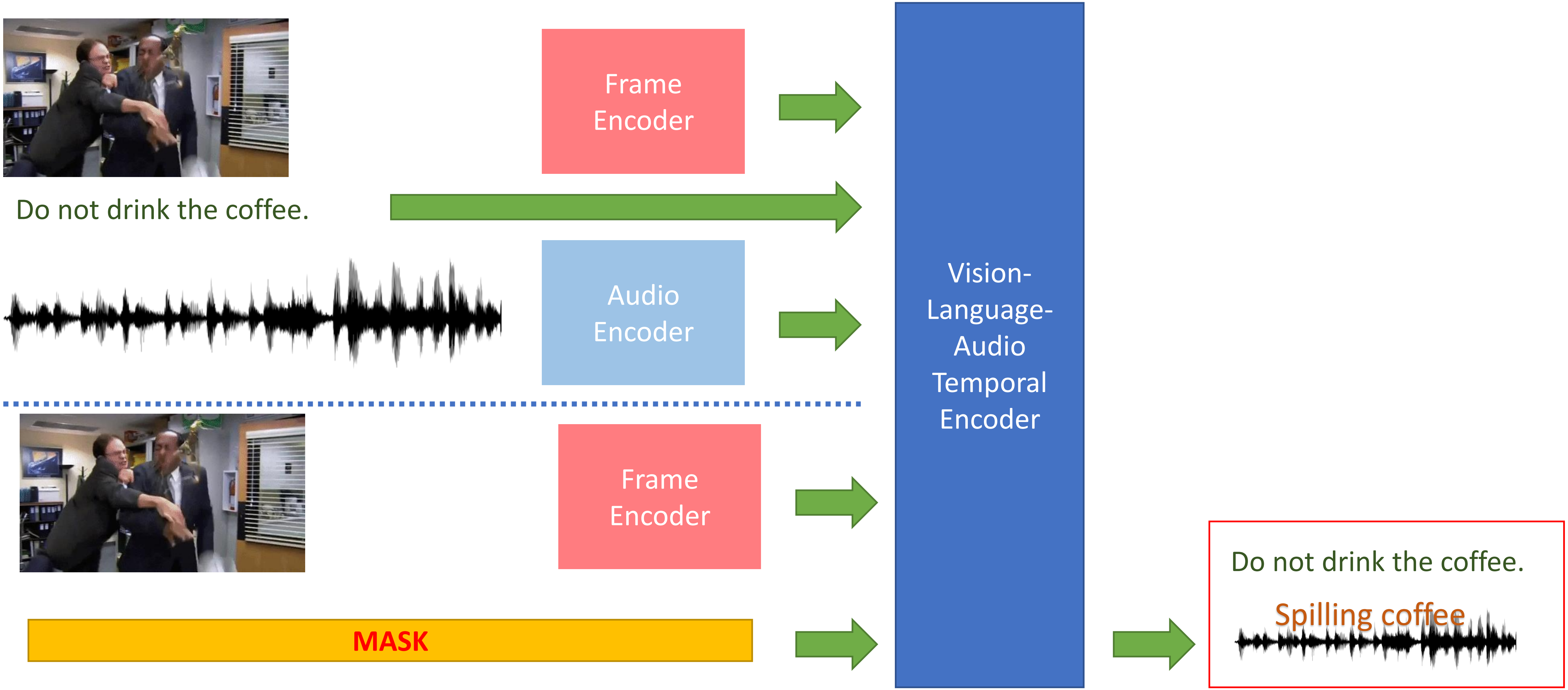
MERLOT RESERVE



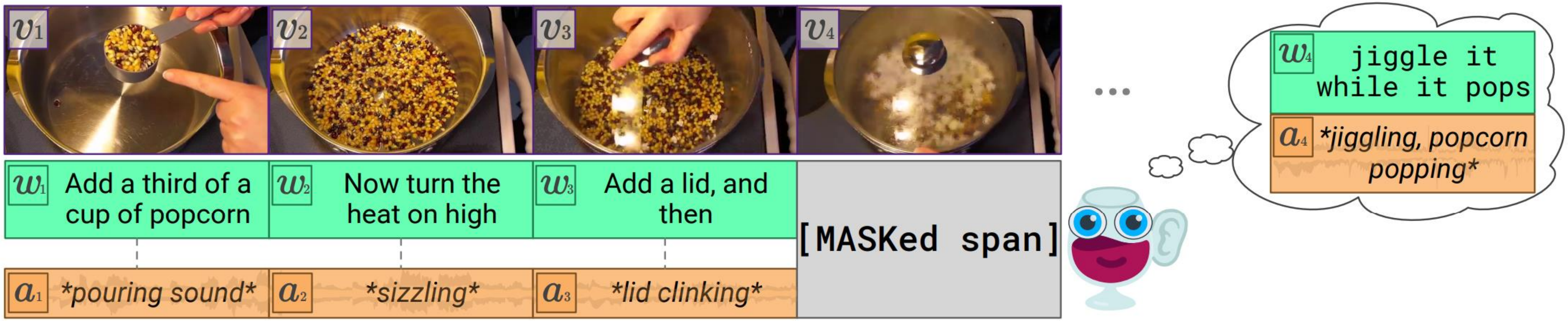
MERLOT RESERVE



MERLOT RESERVE

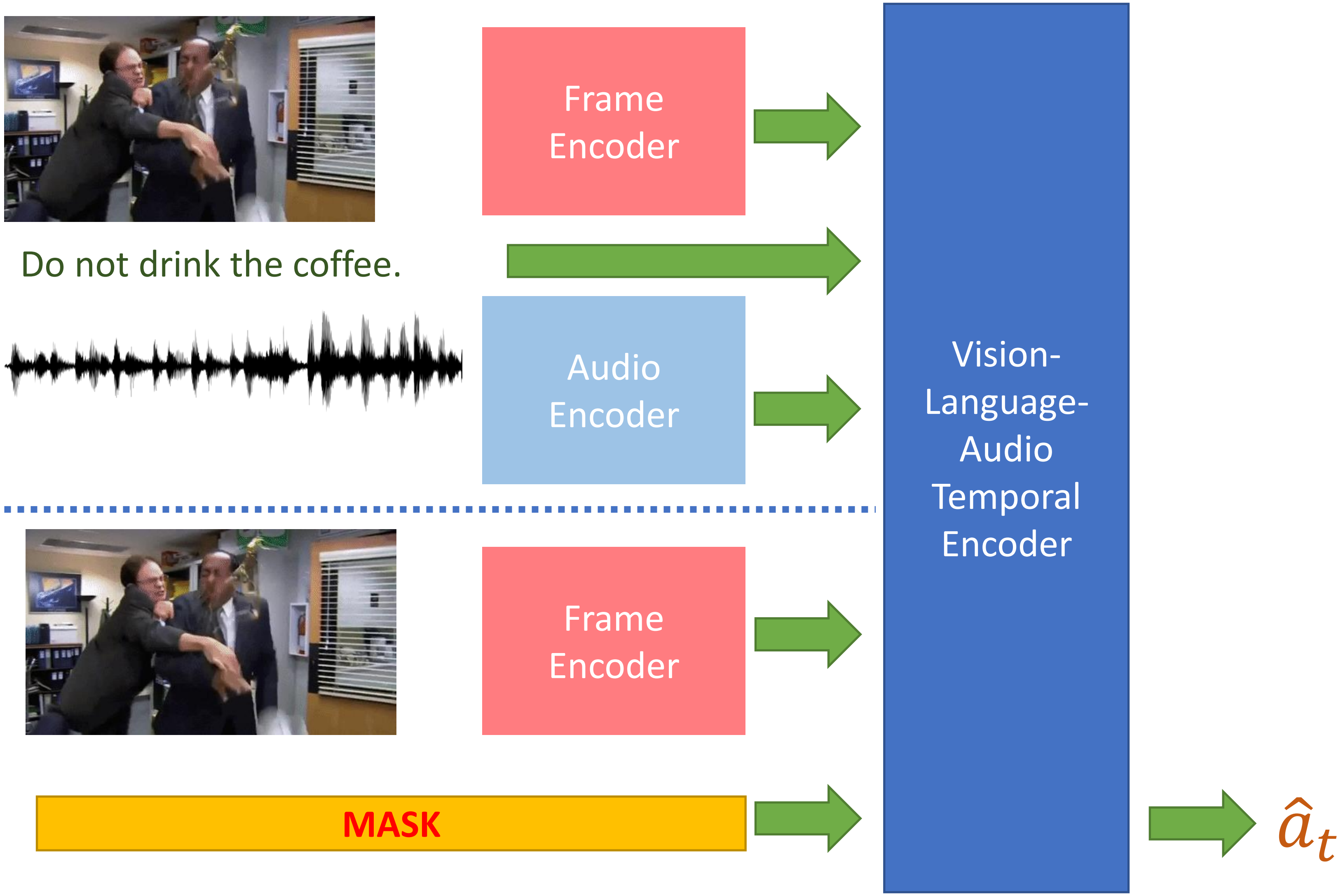


MERLOT RESERVE

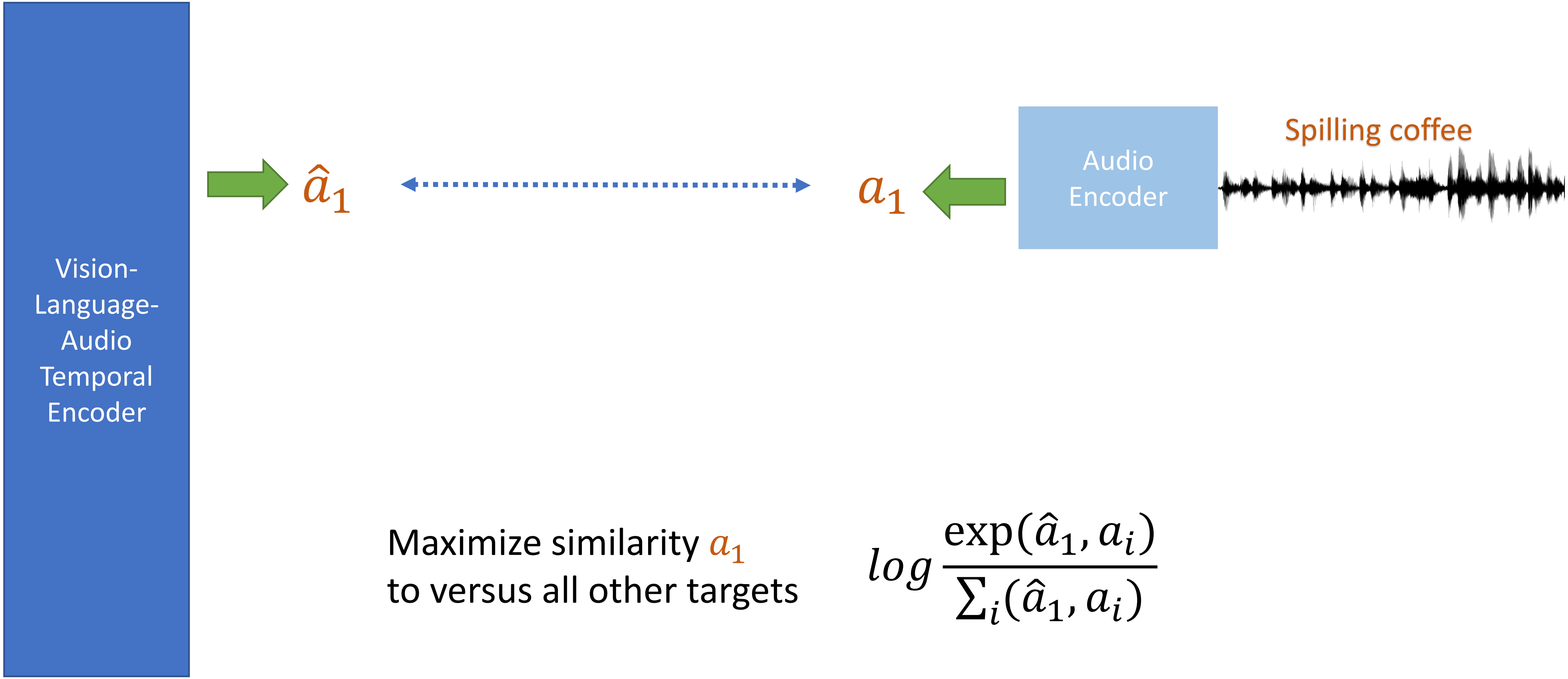


https://arxiv.org/pdf/2201.02639.pdf

MERLOT RESERVE



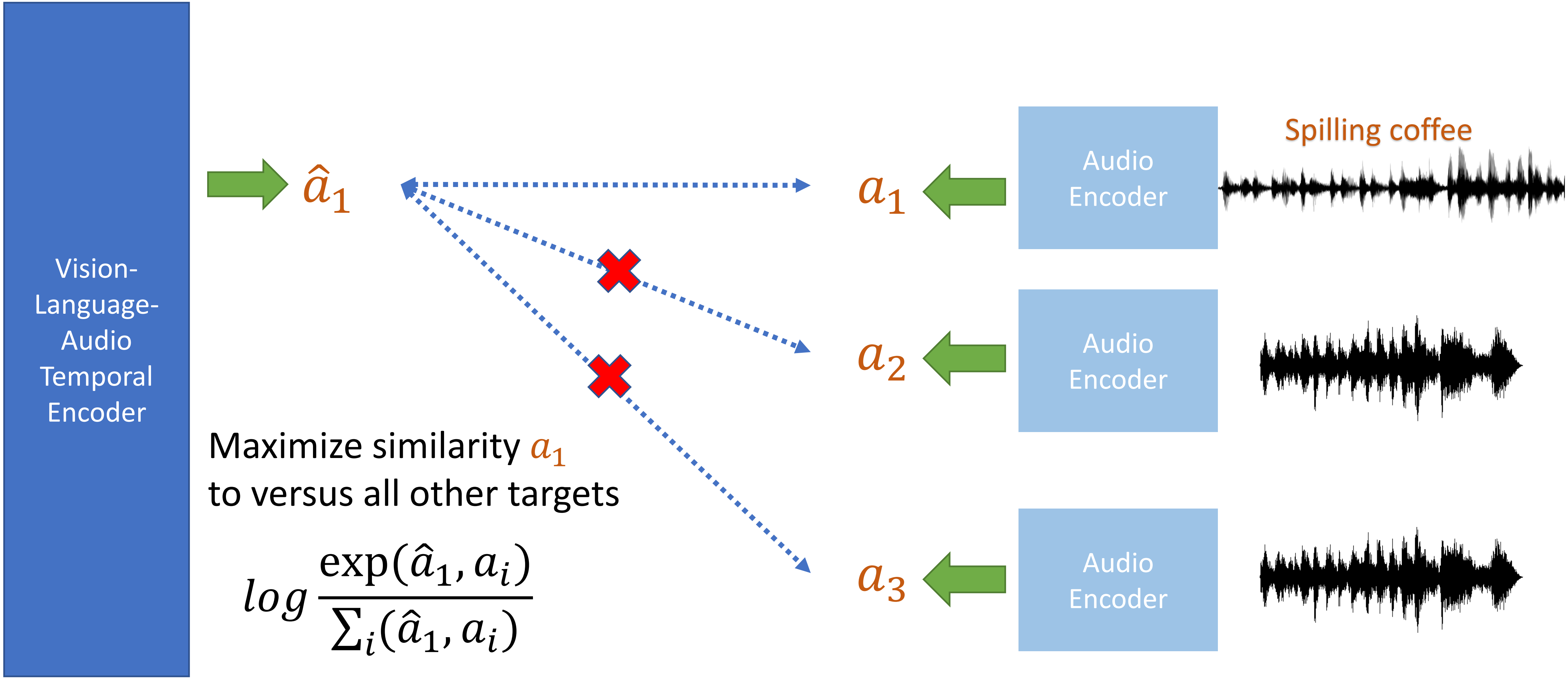
MERLOT RESERVE



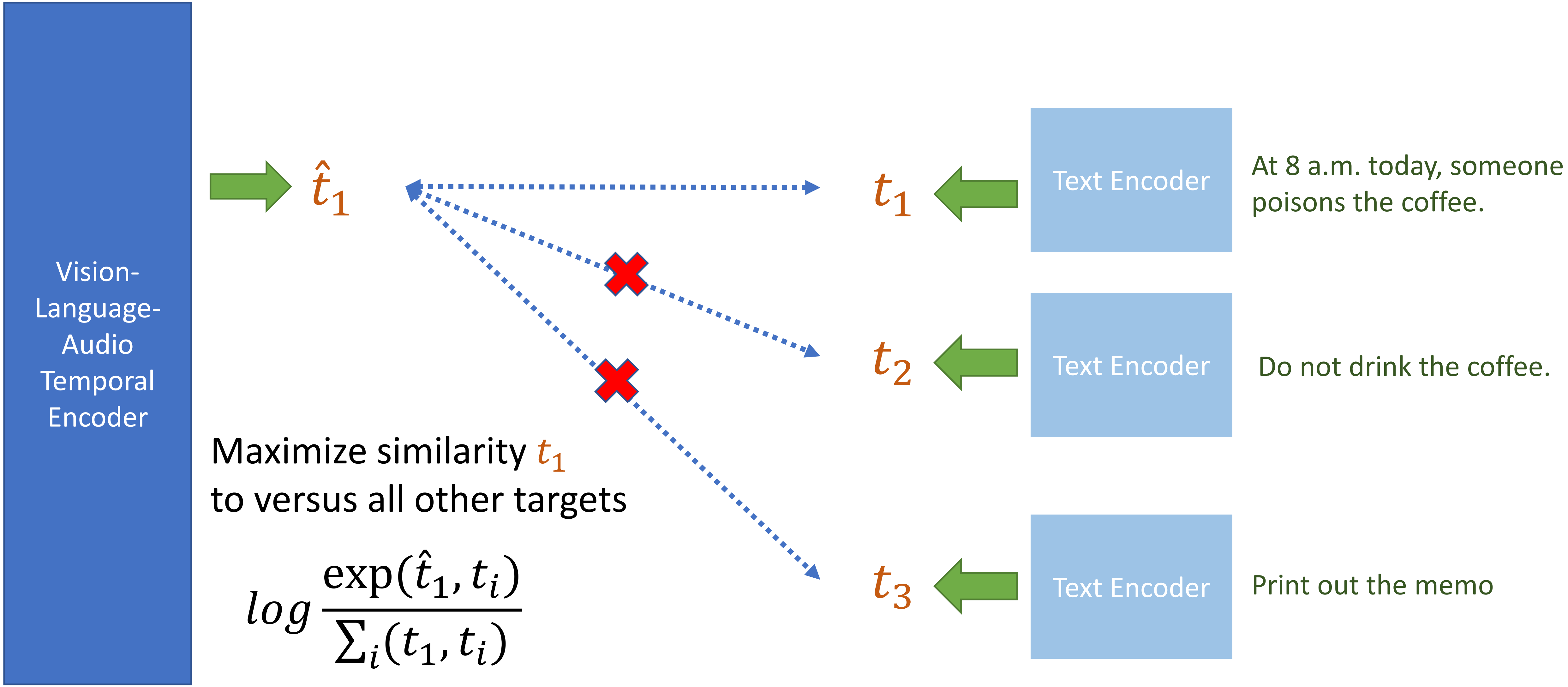
Maximize similarity a_1
to versus all other targets

$$log \frac{\exp(\hat{a}_1, a_i)}{\sum_i(\hat{a}_1, a_i)}$$

MERLOT RESERVE



MERLOT RESERVE



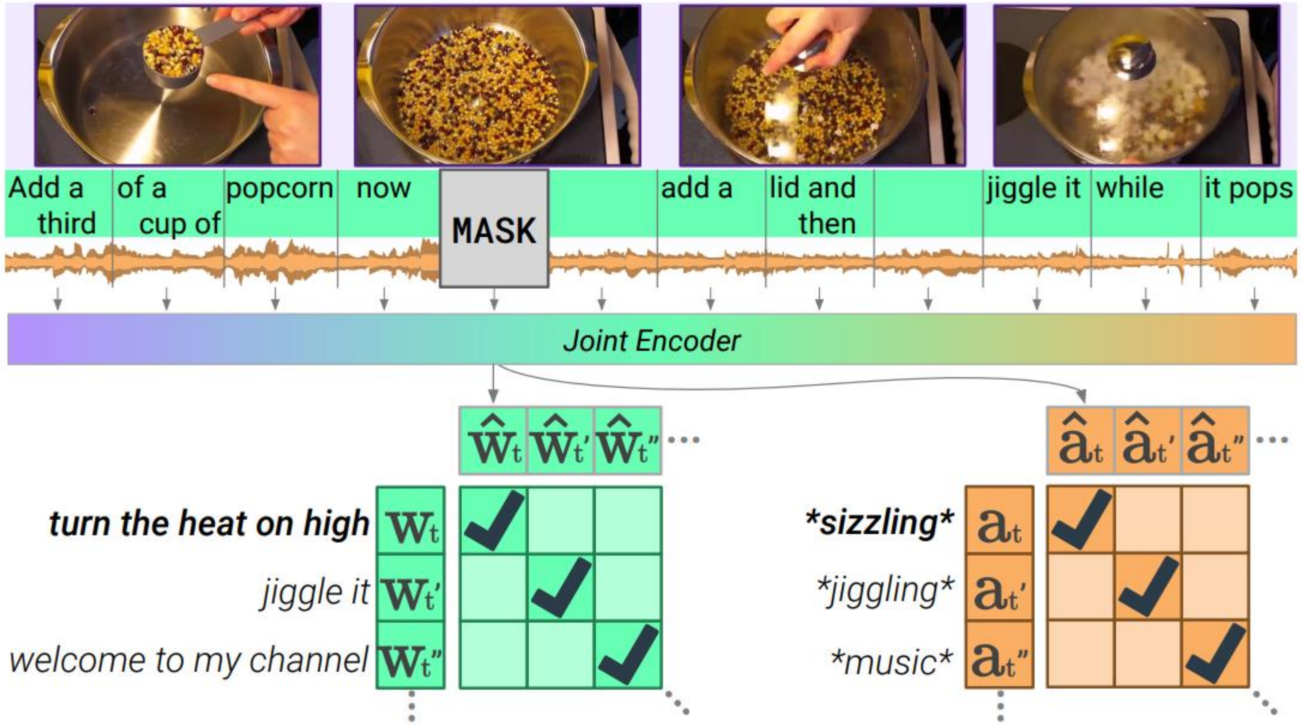
Maximize similarity t_1
to versus all other targets

$$log \frac{\exp(\hat{t}_1, t_i)}{\sum_i (t_1, t_i)}$$

MERLOT RESERVE

Contrastive Span Training

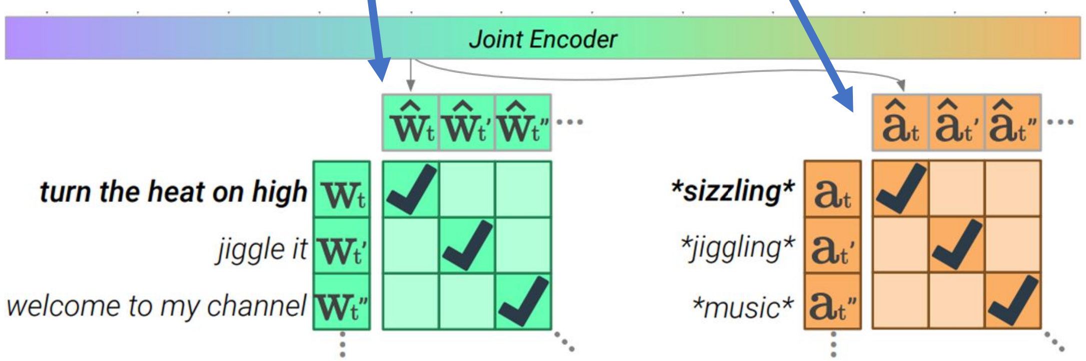
maximize the similarity to the encodings of the text w_t and audio a_t

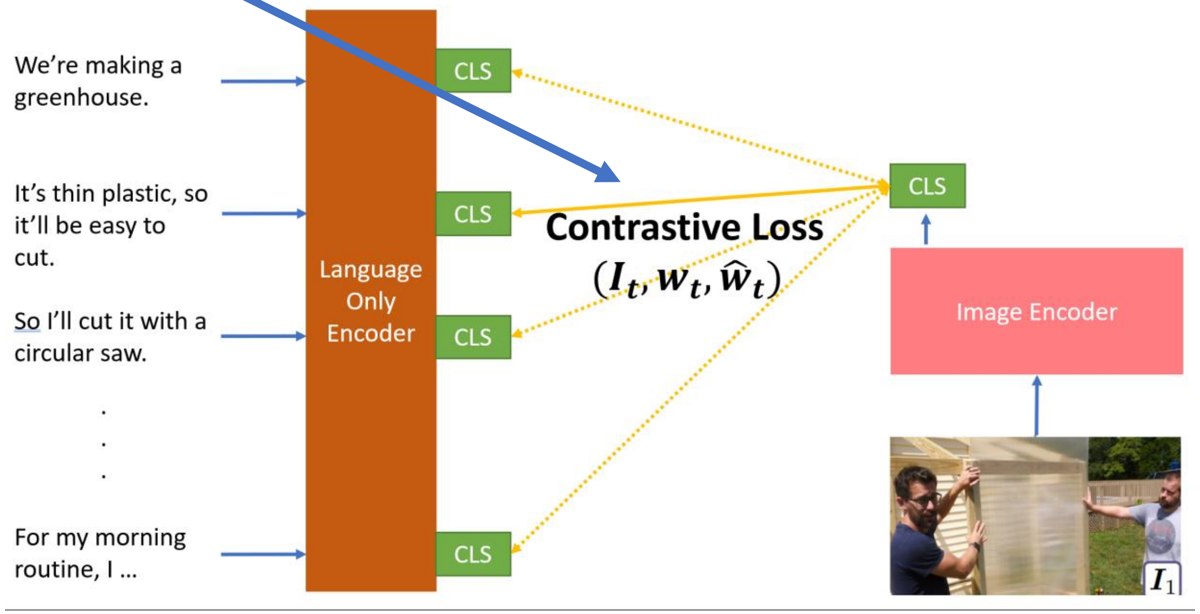


$$\mathcal{L}_{\text{mask}\rightarrow\text{text}} = \frac{1}{|\mathcal{W}|} \sum_{\mathbf{w}_t \in \mathcal{W}} \left(\log \frac{\exp(\sigma \hat{\mathbf{w}}_t \cdot \mathbf{w}_t)}{\sum_{\mathbf{w} \in \mathcal{W}} \exp(\sigma \hat{\mathbf{w}}_t \cdot \mathbf{w})} \right)$$

MERLOT RESERVE

$$L = L_{text} + L_{audio} + L_{frame}$$





Ablations

Visual Commonsense Reasoning

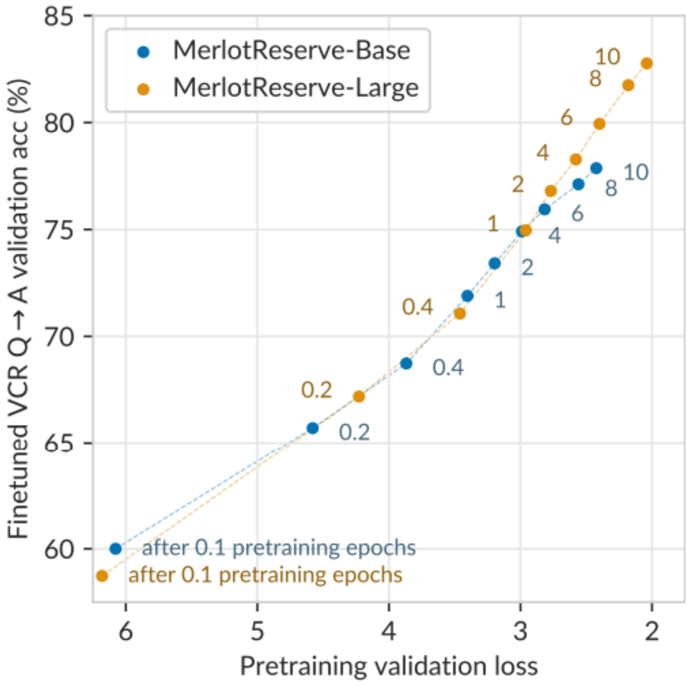
- contrastive span pretraining outperforms mask LM
- improved performance when audio is used both as input and target

	Configuration *for one epoch of pretraining*	VCR val Q→A (%)
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**

Image Tasks

Visual Commonsense Reasoning

Dataset - YT-Temporal-1B.



Pretraining progress

	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**

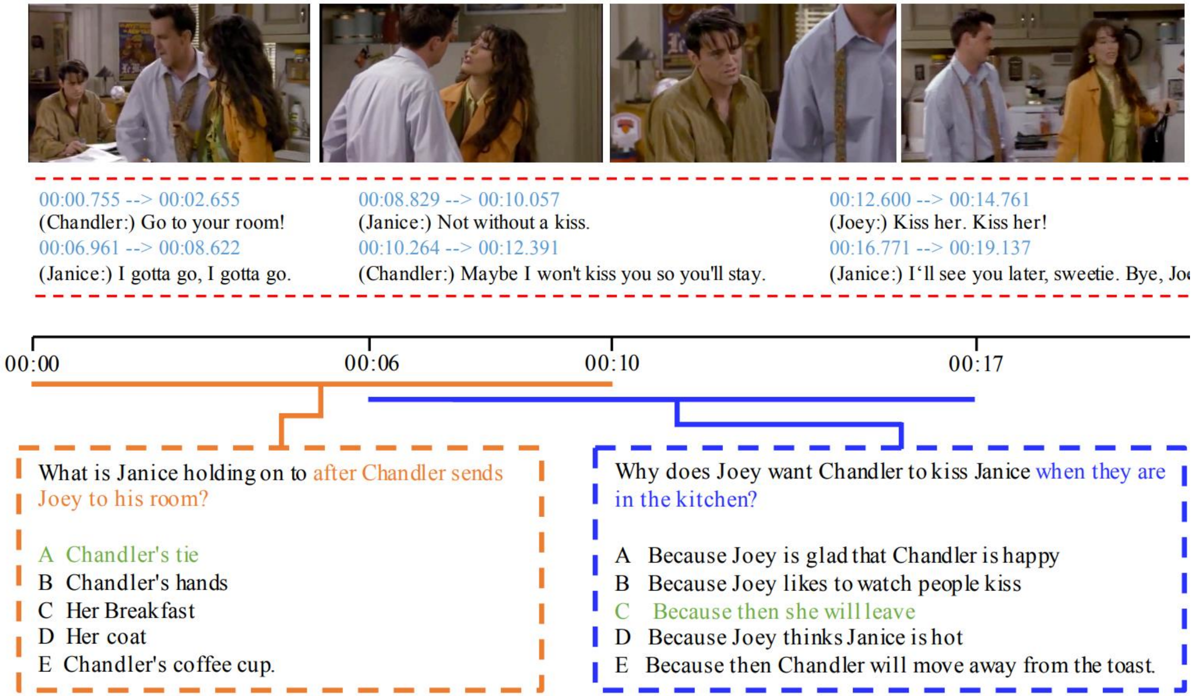
Results on VCR

The joint encoder is a 12-layer, 768-dimensional Transformer

Video Tasks

TVQA

TVQA links depicted objects to visual concepts in questions and answers.



	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**

Results on TQVA

https://tvqa.cs.unc.edu/

https://www.deepmind.com/open-source/kinetics

Activity Recognition

Kinetics-600

No Transcripts

The **Kinetics-600** is a large-scale action recognition dataset which consists of around 480K videos from 600 action categories. The 480K videos are divided into 390K, 30K, 60K for training, validation and test sets, respectively.

		Kinetics-600 (%)	
	Model	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**

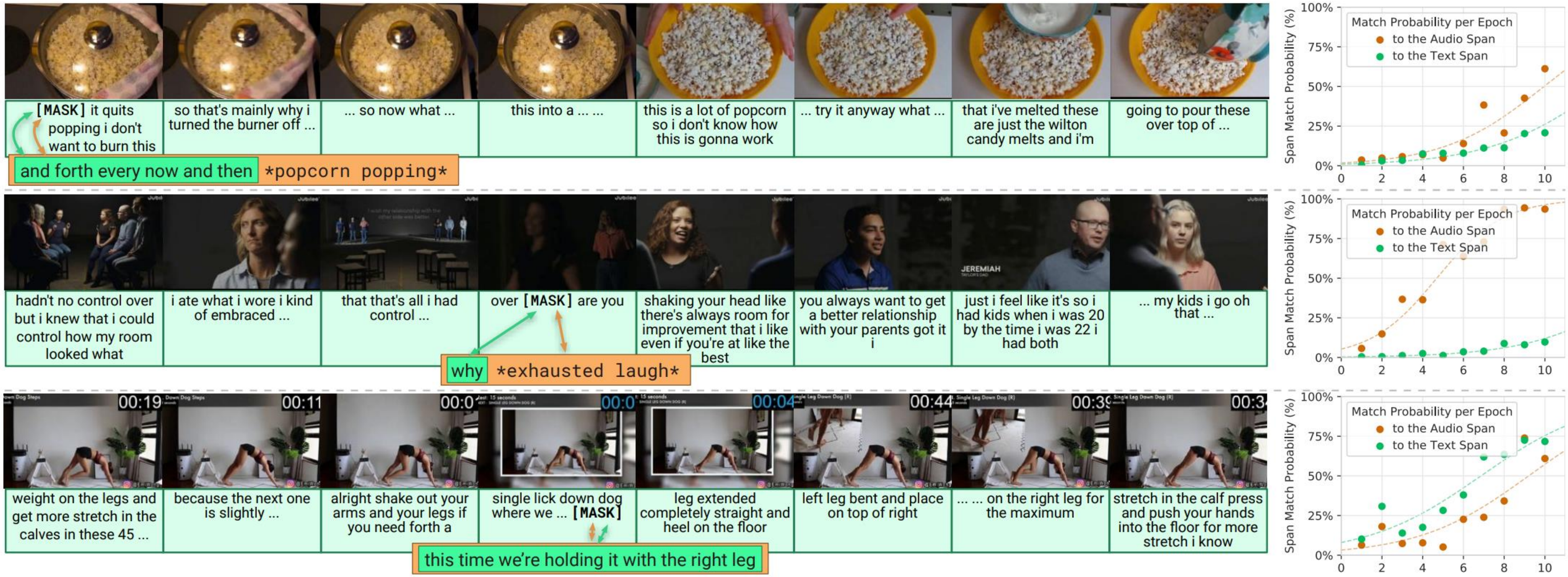
Results on Kinetics-600

Zero-Shot

	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]					AVT+ [46]			MERLOT [128]		
		39.8	43.6	32.3	31.4	36.7	28.2	32.0	15.9	52.9	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**

- Situated Reasoning (STAR): The model is given a video, a templated question, and 4 answer choices.
- Action Anticipation in Epic Kitchens: Here, the goal is to predict future actions given a video clip
- LSMDC: Videos with captions (with a MASK to be filled in)
- MSR-VTT QA: Open-ended video QA

Why does audio help?



Strength

- A new, extensive dataset named YT-Temporal-1B has been introduced as a competitor to HowTo100M and YT-Temporal-180M, which offers broader content coverage beyond just instructional videos.
- A new multimodal model to acquire knowledge from vast amounts of videos featuring accompanying vision, language and audio, achieving new SOTA.
- A new objective learning function that maximize the similarity of positive text and audio pair

https://openreview.net/forum?id=CRFSrgYtV7m

Weakness

- The paper discusses the model training process using millions of YouTube videos, which has sparked privacy apprehensions. Moreover, accessing the dataset may be challenging, as the only provided links to the associated YouTube videos are unlikely to ensure long-term availability of the data.

- The visual representations are not evaluated

- Since the model only utilizes a single frame per segment as input, it remains unclear how it can effectively represent dynamic scenes.
  - we would argue that in the context of learning multimodal script knowledge, giving the model a single frame might actually be a strength. The reason being is that our goal is for MERLOT to infer what’s going on in the world, temporally, through (partial) observations of both vision and language.

https://openreview.net/forum?id=CRFSrgYtV7m

Future Works

- An end-to-end model with feature extractions
- Multiple frames per segment
- Contrastive loss to align image and text/audio pair

https://openreview.net/forum?id=CRFSrgYtV7m

Q&A