

MERLOT RESERVE:



Neural *Script Knowledge* through Vision and Language and Sound

CVPR 2022, 170 citations

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

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

Rowan Zellers¹ Jiasen Lu² Ximing Lu² Youngjae Yu² Yanpeng Zhao³
Mohammadreza Salehi¹ Aditya Kusupati¹ Jack Hessel² Ali Farhadi¹ Yejin Choi²

¹Paul G. Allen School of Computer Science & Engineering, University of Washington
²Allen Institute for Artificial Intelligence ³University of Edinburgh

rowanzellers.com/merlotreserve

Group 6:

1. Reeshoon Sayera
2. Soumik Ghosh
3. Ifty Rezwan
4. Xiaohang Wang
5. Xitong Li

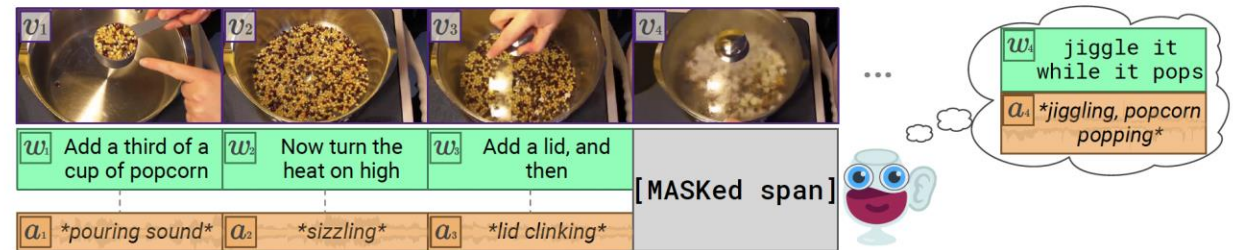


Figure 1: MERLOT RESERVE learns multimodal neural script knowledge representations of video – jointly reasoning over video frames, text, and audio. Our model is pretrained to predict which snippet of text (and audio) might be hidden by the MASK. This task enables it to perform well on a variety of vision-and-language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





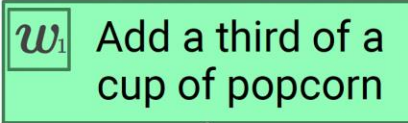
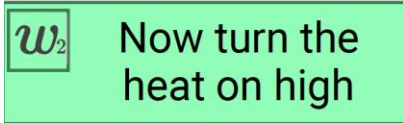
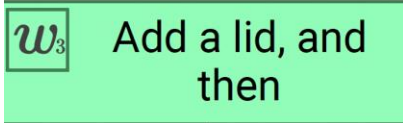
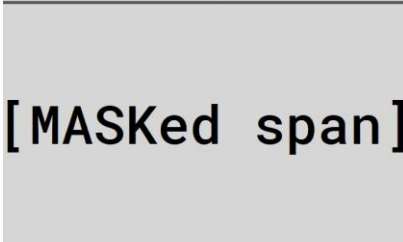

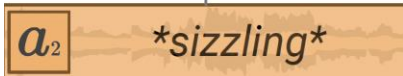
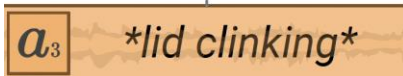
Outline

1. Introduction and Related Work
2. Model Architecture
3. Experiments and Ablation Study
4. Qualitative Analysis
5. Limitations and Potential Ideas

Introduction

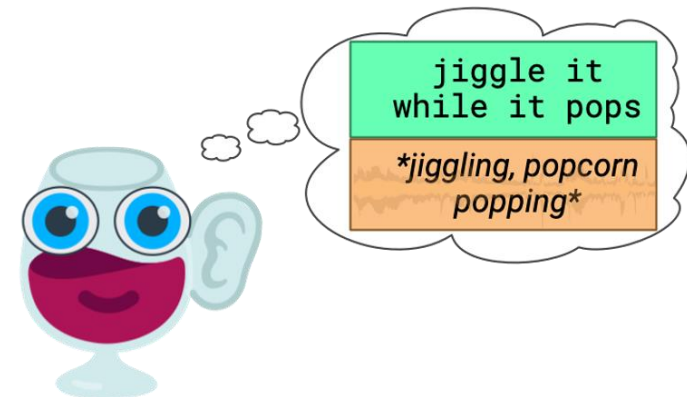
Script knowledge:

- Proposed by (Schank and Abelson, 1977)
- is a body of knowledge that describes **a typical sequence of actions** people do **in a particular situation**.

Learning from re-entry:

- **time-locked correlations** enable one modality to educate others.
- crucial for how we as humans learn visual and world knowledge.



Can we build machines that likewise learn vision, language, and sound together?

- Yes. MERLOT RESERVE.

Multimodal **E**vent **R**epresentation **L**earning **O**ver **T**ime, with **RE**-entrant **SupERV**sion of **E**vents.

MERLOT RESERVE learns from

- Video frames
- Subtitles
- Audio

Given a video:

- Replace **subtitles** and **audio** with MASK token.
- The model predicts by choosing the correct masked-out snippet.

Related Work

Joint representations of multiple modalities

Family of **VisualBert** models:

- Pretrain on images paired with literal captions

MERLOT:

- Learns a joint vision-text model but lacks audio

Co-supervision between modalities

Pitfall:

- Complex inter-modal (image-text) interactions ignored
- Learning simpler intra-modal (text-text) interactions

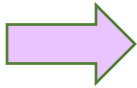
CLIP:

- Use objectives that cannot be shortcut with simple intra-modal(text-text) patterns.

MERLOT RESERVE combines these two lines of research.



Frame
Encoder



Going to pour these over top



Vision-
Language-
Audio
Temporal
Encoder

sound of candy melts getting poured

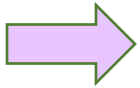


Audio
Encoder





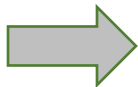
Frame
Encoder



Vision-
Language-
Audio
Temporal
Encoder

Going to pour these

MASK



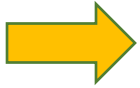
over top



sound of candy melts getting poured



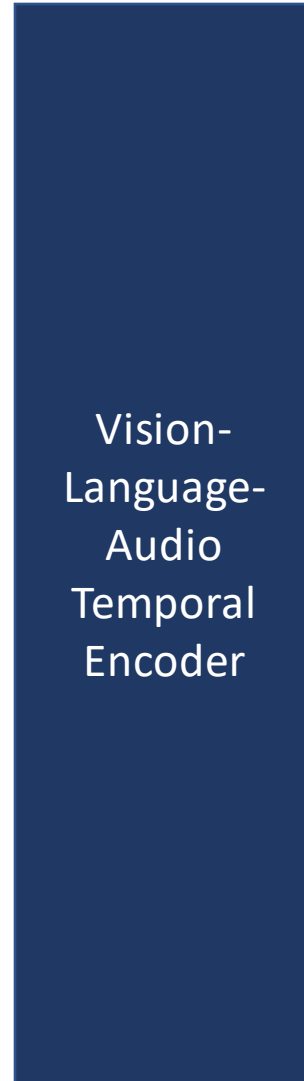
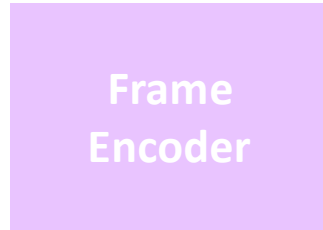
Audio
Encoder

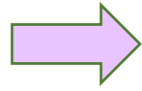
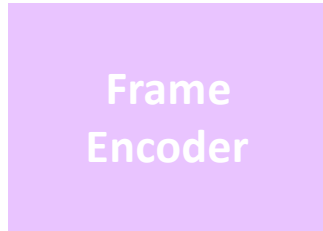




Going to pour these over top

sound of candy melts getting poured

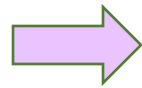
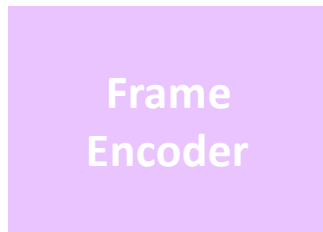
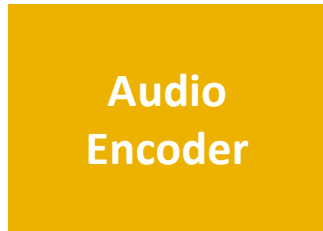




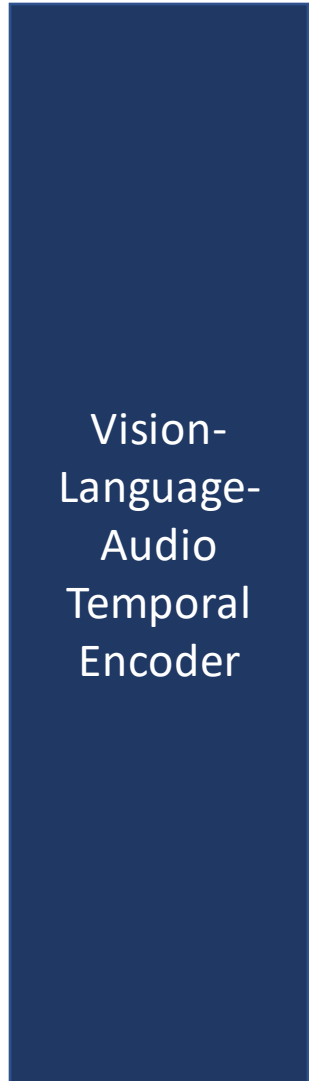
Going to pour these over top

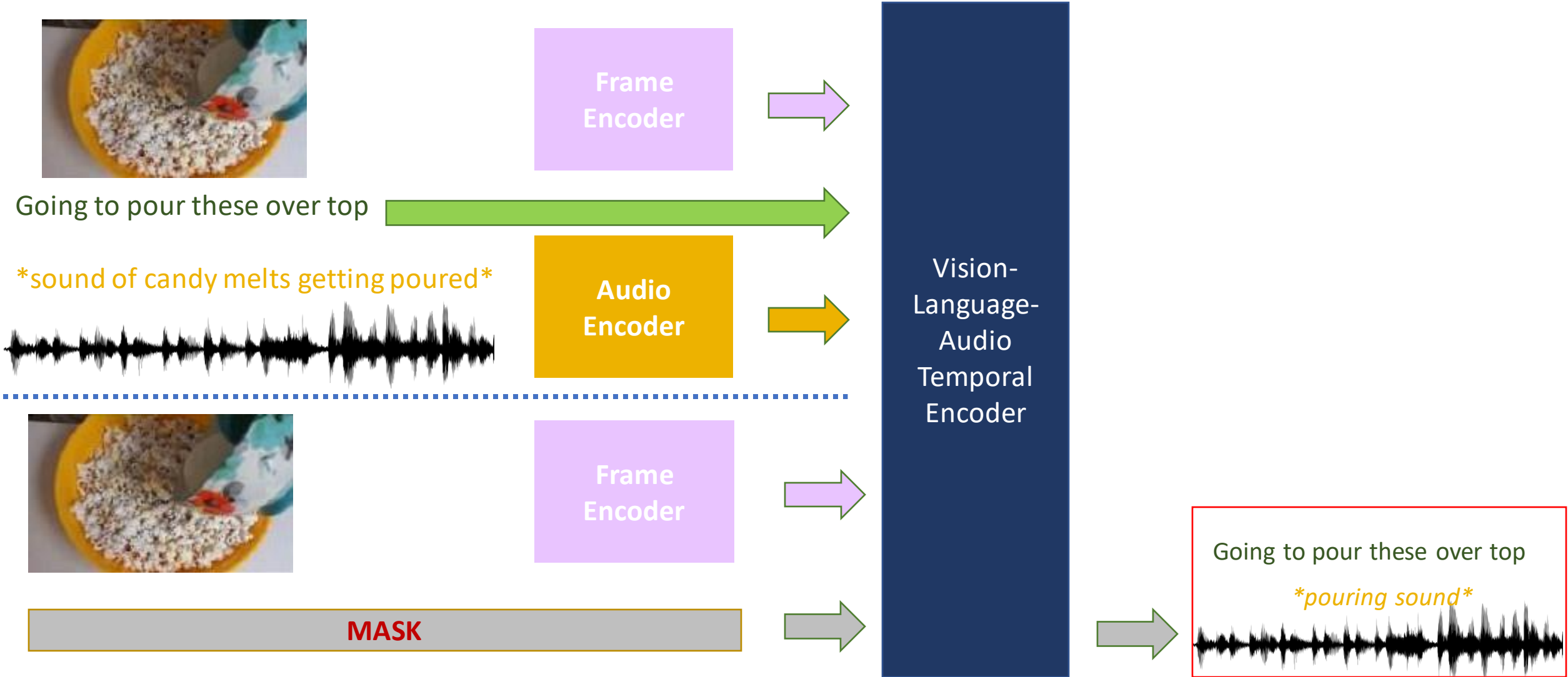


sound of candy melts getting poured

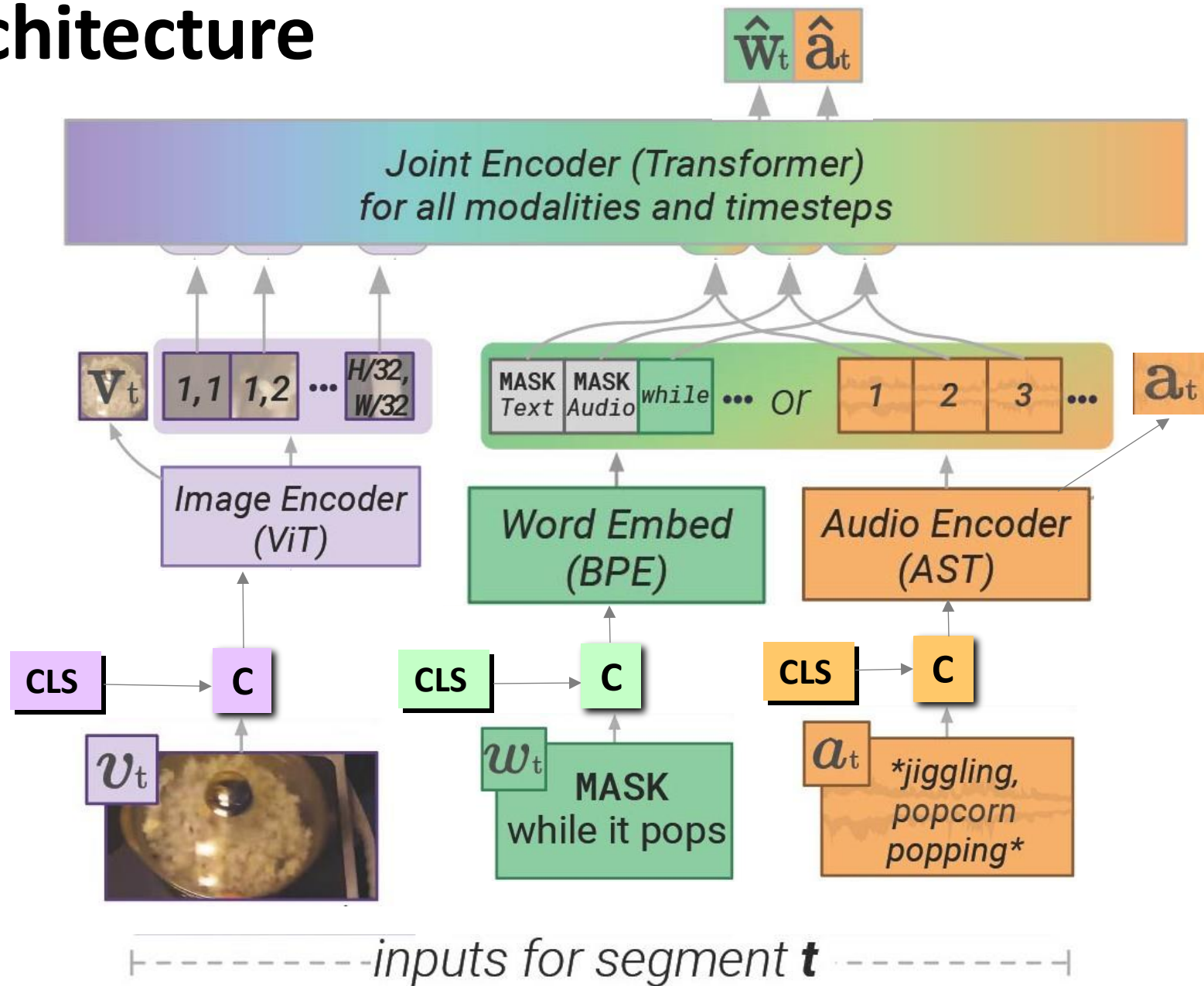


MASK





Model Architecture



Avoiding Shortcut Learning

Shortcut Learning: Low training loss, but poor representations

How to avoid shortcut learning?

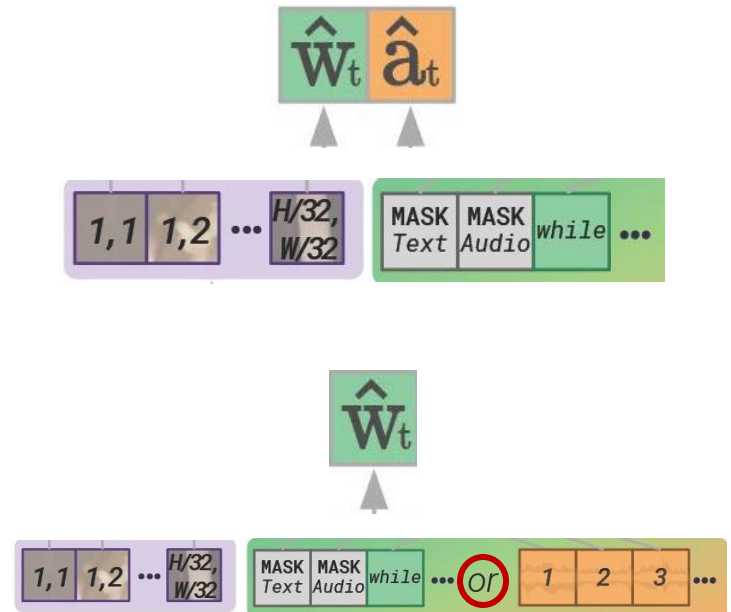
Train on two types of masked videos:

➤ Audio only as *target*:

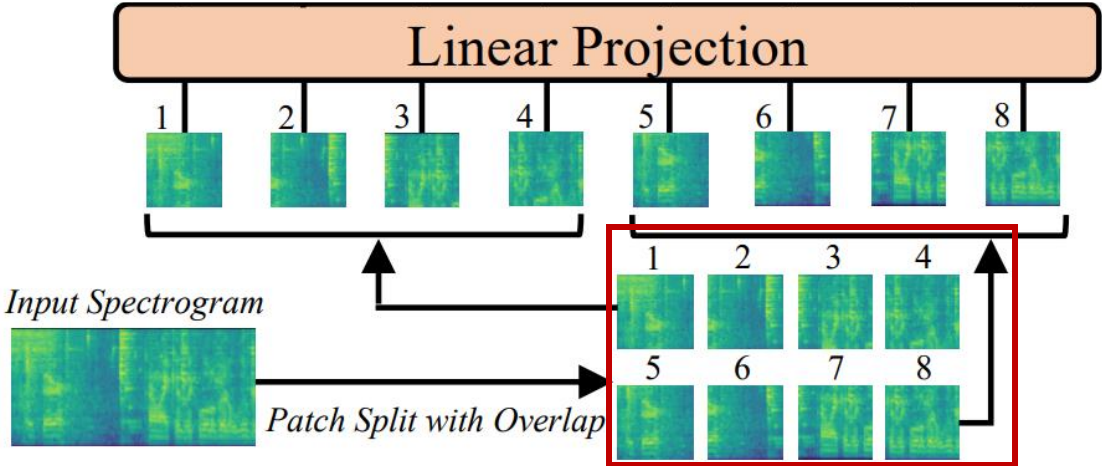
- Provide *video frames* and *text*
- Infer *text* and *audio* representations in *MASKed tokens*

➤ Audio as *input*:

- Provide *video frames* and (*text* or *audio*)
- Infer only *text* representations in *MASKed tokens*



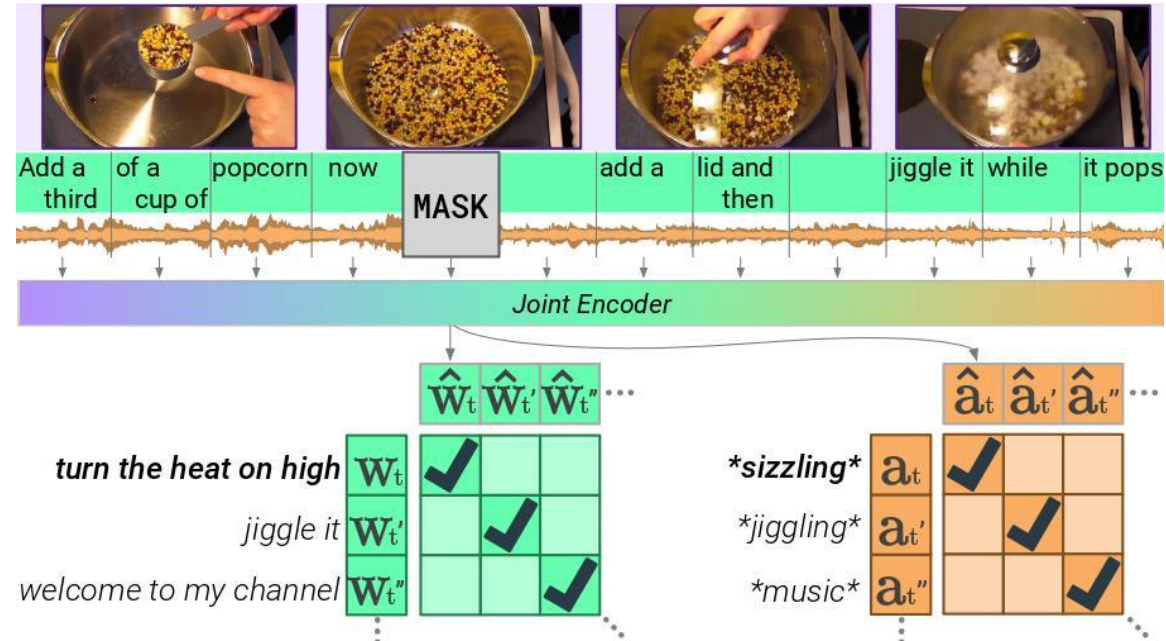
AST: Audio Spectrogram Transformer



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

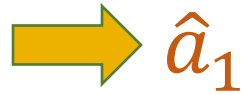


Contrastive Masked Span



$$\mathcal{L}_{\text{contrastive masked span}} = \mathcal{L}_{\text{text}} + \mathcal{L}_{\text{audio}} \quad (\text{Proposed Novel Loss})$$

Vision-
Language-
Audio
Temporal
Encoder



Vision-
Language-
Audio
Temporal
Encoder

\hat{a}_1

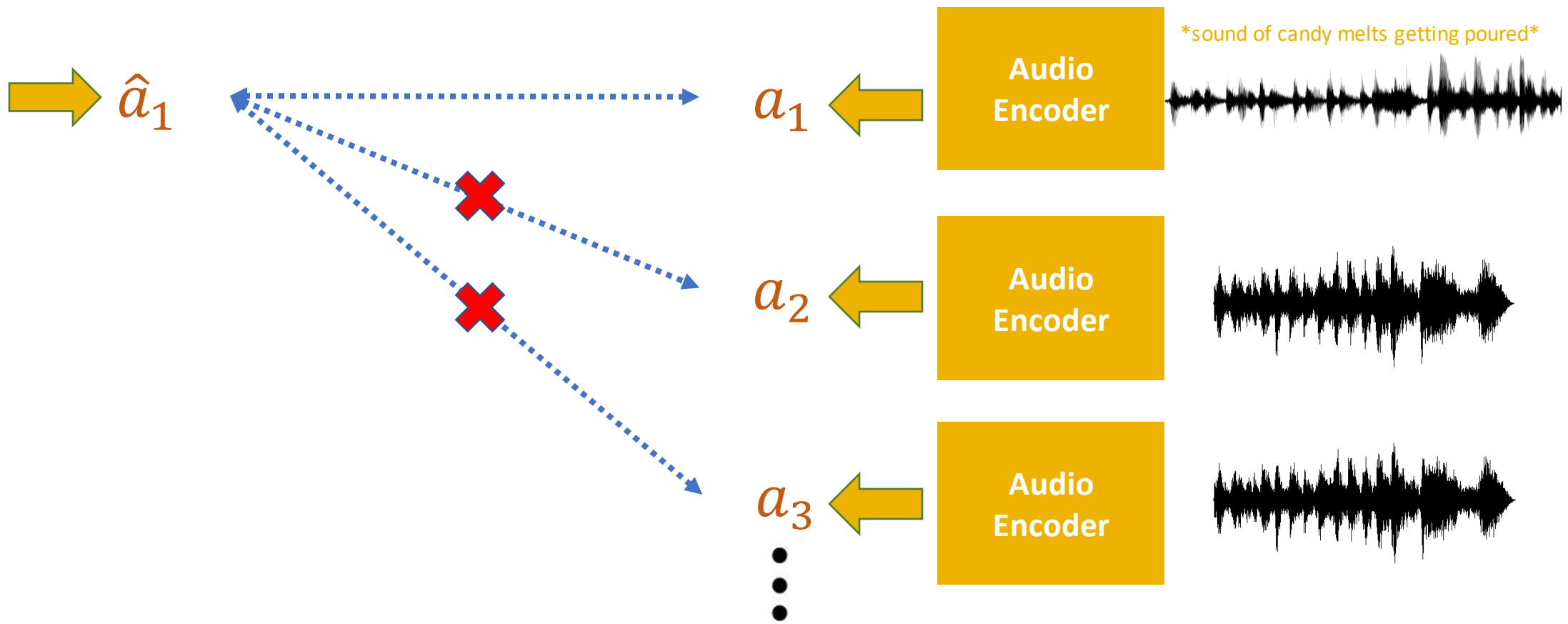


a_1

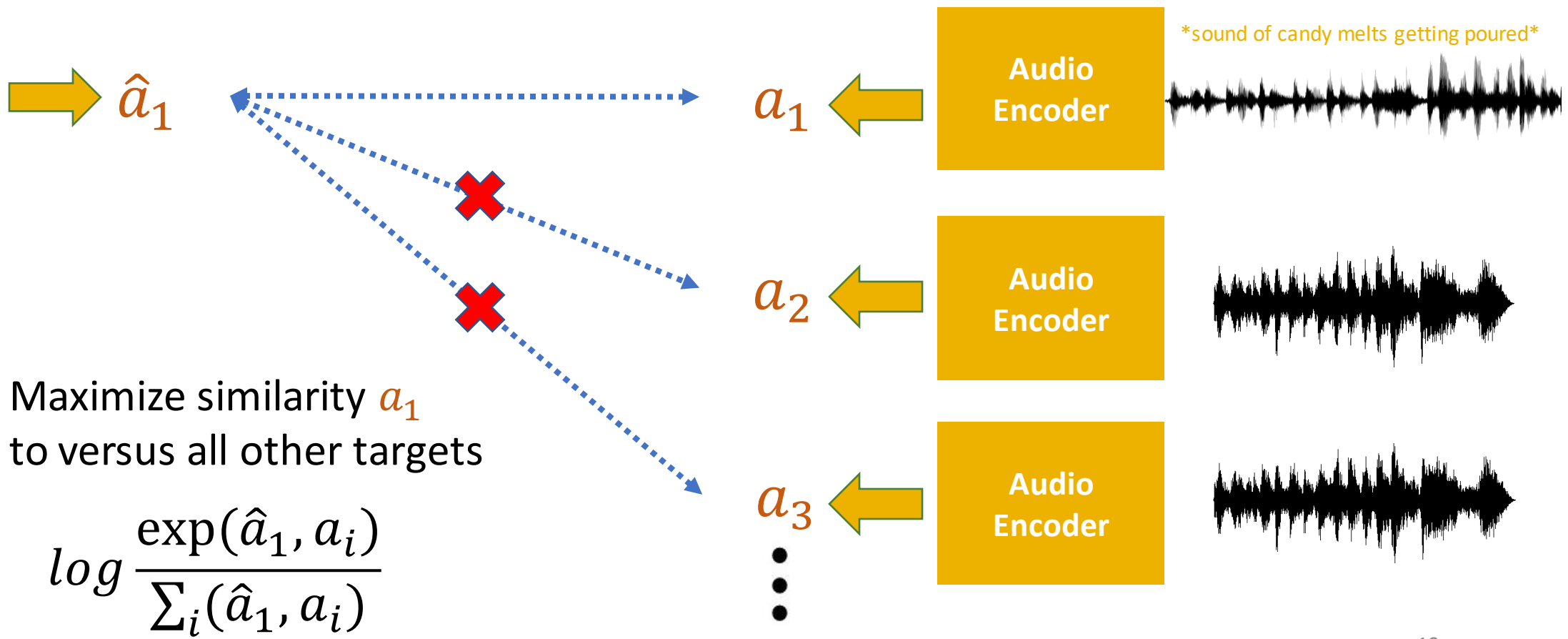
Audio
Encoder

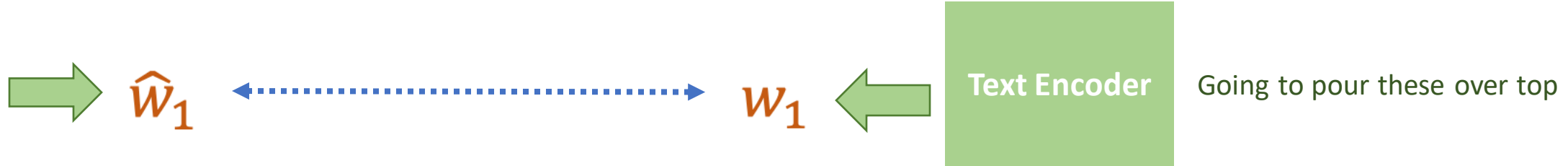


Vision-
Language-
Audio
Temporal
Encoder

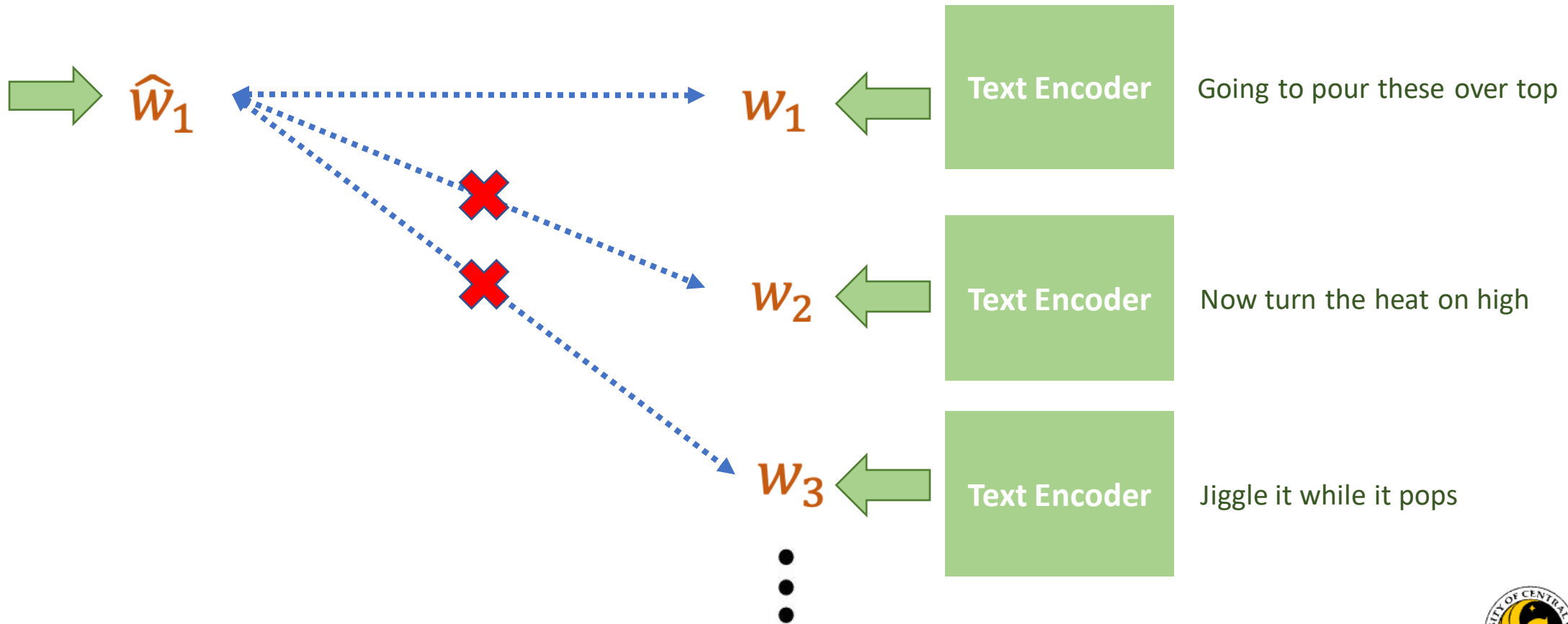


Vision-
Language-
Audio
Temporal
Encoder

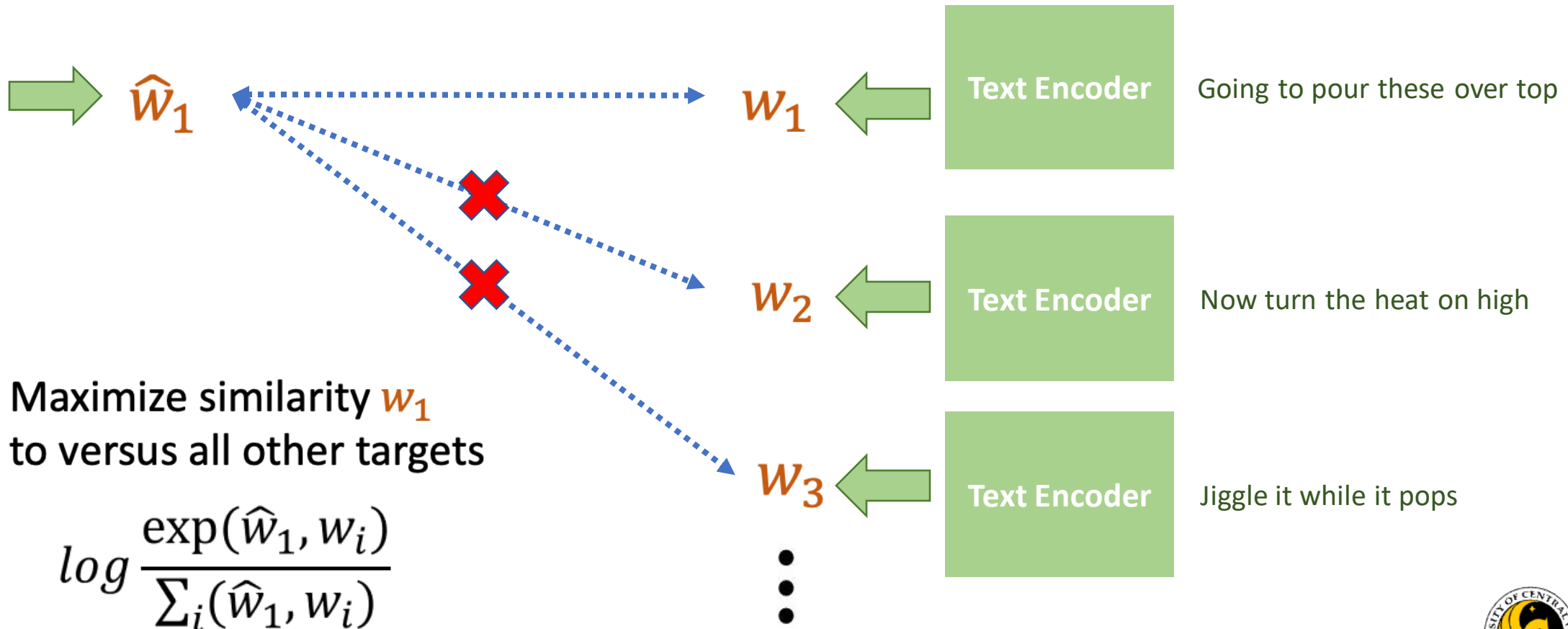




Vision-
Language-
Audio
Temporal
Encoder



Vision-
Language-
Audio
Temporal
Encoder



Contrastive Masked Span (Text & Audio)

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

$\mathcal{L}_{\text{text} \rightarrow \text{mask}}$ is the transpose of $\mathcal{L}_{\text{mask} \rightarrow \text{text}}$

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

$\mathcal{L}_{\text{audio} \rightarrow \text{mask}}$ is the transpose of $\mathcal{L}_{\text{mask} \rightarrow \text{audio}}$

Total Text Loss:

$$\mathcal{L}_{\text{text}} = \mathcal{L}_{\text{mask} \rightarrow \text{text}} + \mathcal{L}_{\text{text} \rightarrow \text{mask}}$$

Total Audio Loss:

$$\mathcal{L}_{\text{audio}} = \mathcal{L}_{\text{mask} \rightarrow \text{audio}} + \mathcal{L}_{\text{audio} \rightarrow \text{mask}}$$

Contrastive Span Training

- **Consists of two parts:**

- Contrastive Masked Span (Proposed Novel Loss)

$$\mathcal{L}_{\text{contrastive masked span}} = \mathcal{L}_{\text{text}} + \mathcal{L}_{\text{audio}}$$

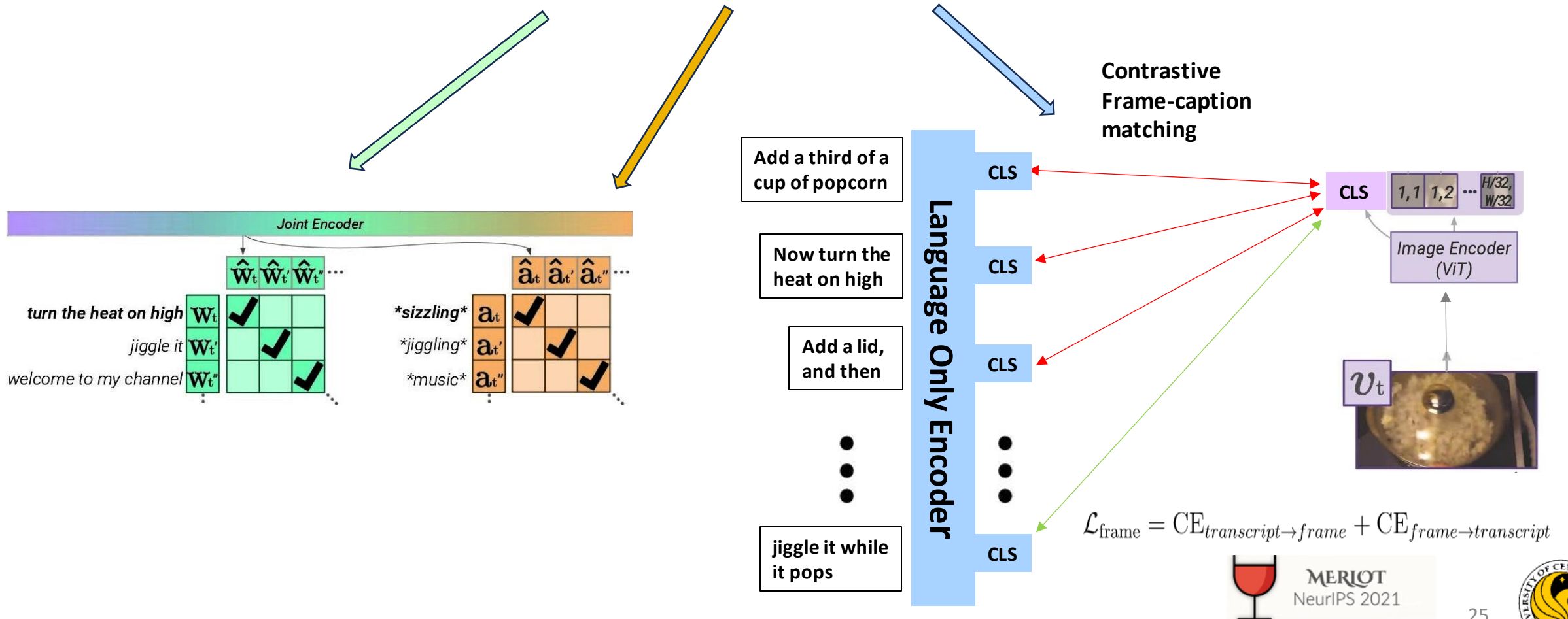
- Contrastive Frame-Transcript Matching (From **MERLOT**)

$$\mathcal{L}_{\text{contrastive frame-transcript matching}} = \mathcal{L}_{\text{frame}}$$

$$\mathcal{L}_{\text{contrastive span training}} = \mathcal{L}_{\text{text}} + \mathcal{L}_{\text{audio}} + \mathcal{L}_{\text{frame}} \quad (\text{Total Loss})$$

Contrastive Span Training

$$\mathcal{L}_{\text{contrastive span training}} = \mathcal{L}_{\text{text}} + \mathcal{L}_{\text{audio}} + \mathcal{L}_{\text{frame}}$$



Pretraining Setup

- **Two models**

-  **Reserve-B**

- Hidden layer size of 768
 - 12-layer ViT-B/16 image encoder
 - 12-layer joint encoder

-  **Reserve-L**

- Hidden size of 1024
 - 24-layer ViT-L/16 image encoder
 - 24-layer joint encoder

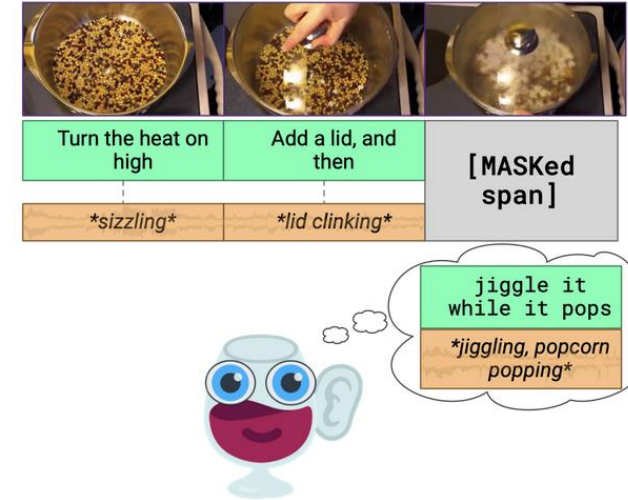
- **Training setup**

- Trained on 512 TPU v3 accelerators
 - Reserve-B took 5 days
 - Reserve-L took 16 days

Pretraining Dataset

- New dataset

- 20 million English subtitled Youtube videos
- 1 billion frames
- Steps taken to protect user privacy
- Directly scraped from public, large and monetized channels



Details:

Year Created: 2022

Size: 20 million videos

Number of Categories: n/a

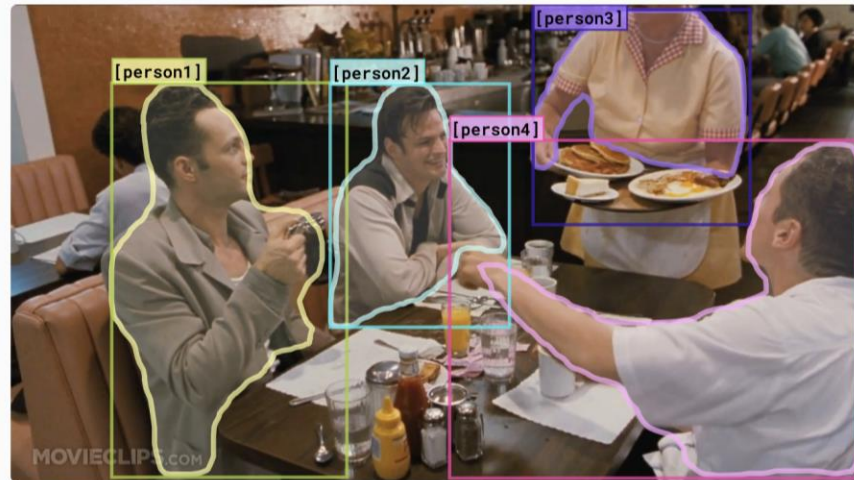
COVE Profile Last Updated: Mar. 30th, 2022

Tasks: multimodal

Topics: visual reasoning

Data Types: video

Ablations



hide all show all [person1] [person2] [person3] [person4]
more objects »

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...

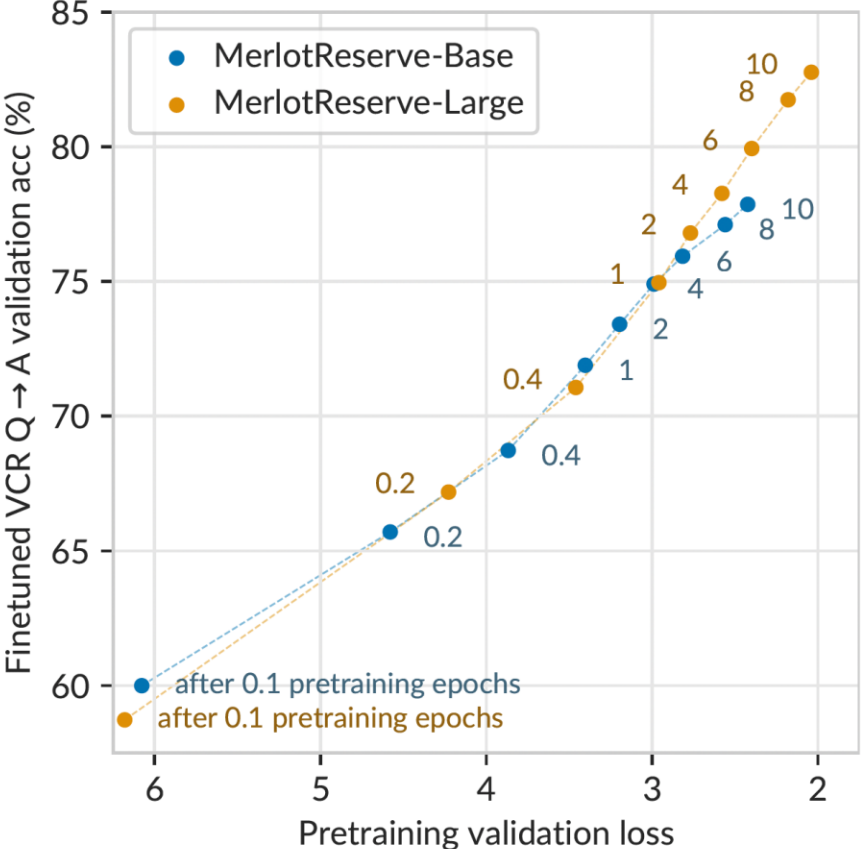
- a) [person1] has the pancakes in front of him.
- b) [person4] is taking everyone's order and asked for clarification.
- c) [person3] is looking at the pancakes both she and [person2] are smiling slightly.
- d) [person3] is delivering food to the table, and she might not know whose order is whose.

- Model analyzes a movie image and question.
- Chooses correct answer from four choices (Q→A).
- Selects justification for answer from four rationales (QA→R).
- Success judged on accurate answer and rationale selection (Q→AR).

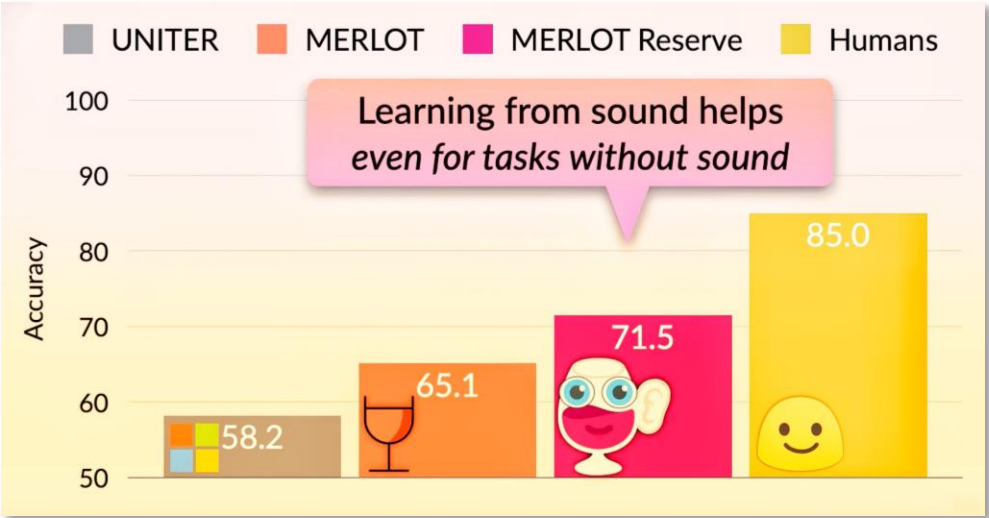
Configuration for one epoch of pretraining	VCR Q→A	val (%)
V+T Mask LM [29, 106, 128]	67.2	
V+T VirTex-style [27]	67.8	
V+T Contrastive Span	69.7	
V+T+A Audio as target	70.4	
V+T+A Audio as input and target	70.7	
V+T+A Audio as input and target, w/o strict localization	70.6	
V+T+A RESERVE-B		71.9

Visual Commonsense Reasoning – Image Task

Trained for 10 epochs on YT-Temporal-1B



Model	VCR test (acc; %)			GFLOPS
	Q→A	QA→R	Q→AR	
Caption/Obj/Det-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	303
RESERVE-B	79.3	78.7	62.6	146
RESERVE-L	84.0	84.9	72.0	341



TVQA: Television Question Answering – Video Task



00:00.755 --> 00:02.655

(Chandler:) Go to your room!

00:06.961 --> 00:08.622

(Janice:) I gotta go, I gotta go.

00:08.829 --> 00:10.057

(Janice:) Not without a kiss.

00:10.264 --> 00:12.391

(Chandler:) Maybe I won't kiss you so you'll stay.

00:12.600 --> 00:14.761

(Joey:) Kiss her. Kiss her!

00:16.771 --> 00:19.137

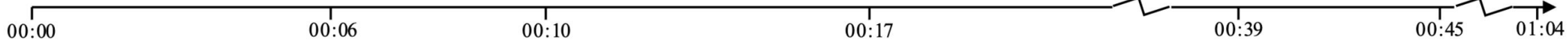
(Janice:) I'll see you later, sweetie. Bye, Joey.

00:39.327 --> 00:40.760

(Chandler:) She makes me happy. ...

00:41.596 --> 00:44.087

(Joey:) Okay. All right.



What is Janice holding on to **after Chandler sends Joey to his room?**

- A Chandler's tie
- B Chandler's hands
- C Her Breakfast
- D Her coat
- E Chandler's coffee cup.

Why does Joey want Chandler to kiss Janice **when they are in the kitchen?**

- A Because Joey is glad that Chandler is happy
- B Because Joey likes to watch people kiss
- C **Because then she will leave**
- D Because Joey thinks Janice is hot
- E Because then Chandler will move away from the toast.

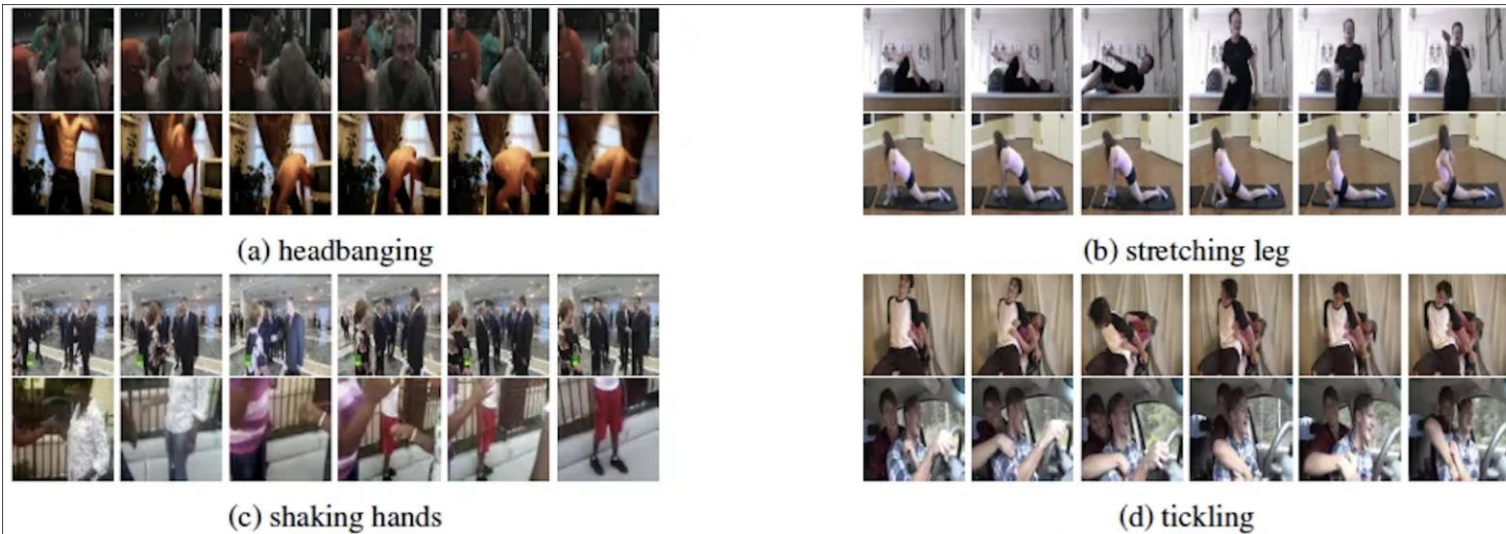
What is on the couch behind Joey **when he is at the counter?**

- A A chick
- B **A soccer ball**
- C A duck
- D A pillow
- E Janice's coat



Kinetics-600 Activity Recognition

- Demonstrated a **1.7%** improvement by integrating **audio** and **visual** data.
- Showcased the advantage of **joint representation of video frames and sound**.



Training – 390,000
 Validation – 30,000
 Testing – 60,000

	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

No Transcripts Used

Zero Shot Experiments

- **Situated Reasoning(STAR):**
 - Short Situations
 - Four Classes
- **LSMDC:**
 - Video & Description
 - Fill in the Blanks
- **Epic Kitchens:**
 - Predict Future Actions
 - Long tail distribution
- **MSR-VTT QA:**
 - Open Ended Video QA

Model	Situated Reasoning (STAR)					EPIC-Kitchens			LSMDC	MSR-VTT QA	
	Interaction	Sequence	Prediction	Feasibility	Overall	Verb	Noun	Action	(FiB test %) Acc	(test acc %) top1	(test acc %) top5
Supervised SoTA	ClipBERT [74]					AVT+ [46]			MERLOT [128]		
Random	39.8	43.6	32.3	31.4	36.7	28.2	32.0	15.9	52.9	43.1	
CLIP (ViT-B/16) [92]	25.0	25.0	25.0	25.0	25.0	6.2	2.3	0.1	0.1	0.1	0.5
CLIP (RN50x16) [92]	39.8	40.5	35.5	36.0	38.0	16.5	12.8	2.3	2.0	3.0	11.9
Just Ask (ZS)[123]	39.9	41.7	36.5	37.0	38.7	13.4	14.5	2.1	2.3	2.3	9.7
RESERVE-B										2.9	8.8
RESERVE-B (+audio)	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

zero-shot

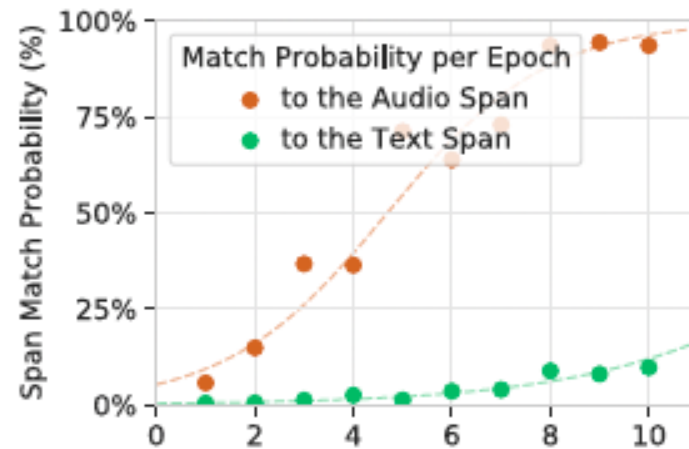


Qualitative Analysis - Why does audio help?

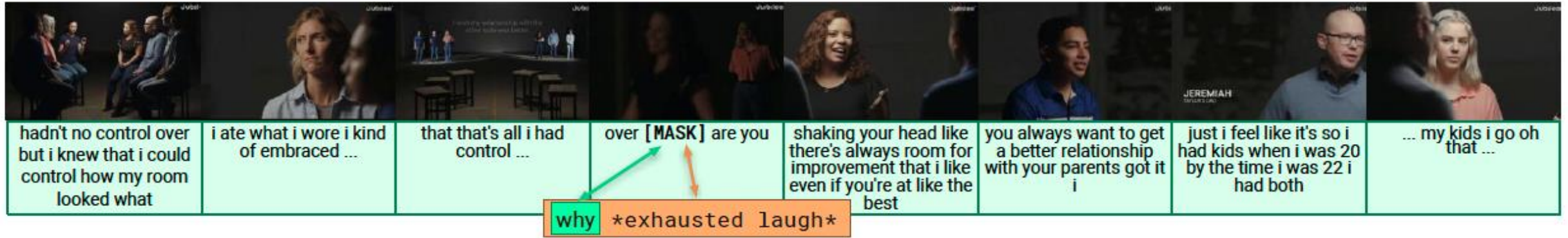


 [MASK] it quits popping i don't want to burn this	so that's mainly why i turned the burner off so now what ...	this into a	this is a lot of popcorn so i don't know how this is gonna work	... try it anyway what ...	that i've melted these are just the wilton candy melts and i'm	going to pour these over top of ...
--	--	---------------------	--------------------	---	----------------------------	--	-------------------------------------

and forth every now and then *popcorn popping*



Qualitative Analysis - Why does audio help?



hadn't no control over but i knew that i could control how my room looked what

i ate what i wore i kind of embraced ...

that that's all i had control ...

over [MASK] are you

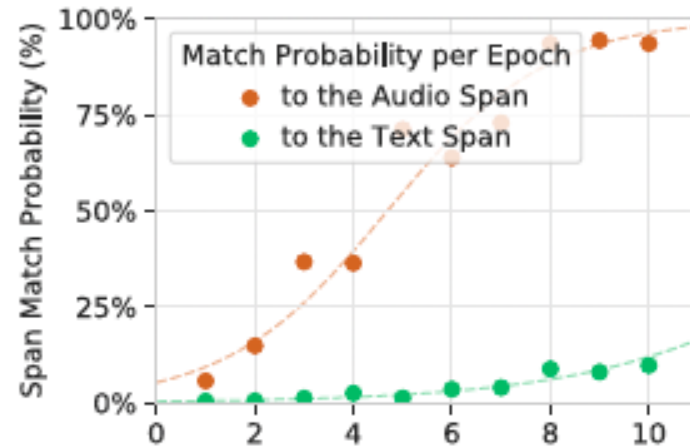
shaking your head like there's always room for improvement that i like even if you're at like the best

you always want to get a better relationship with your parents got it i

just i feel like it's so i had kids when i was 20 by the time i was 22 i had both

... my kids i go oh that ...

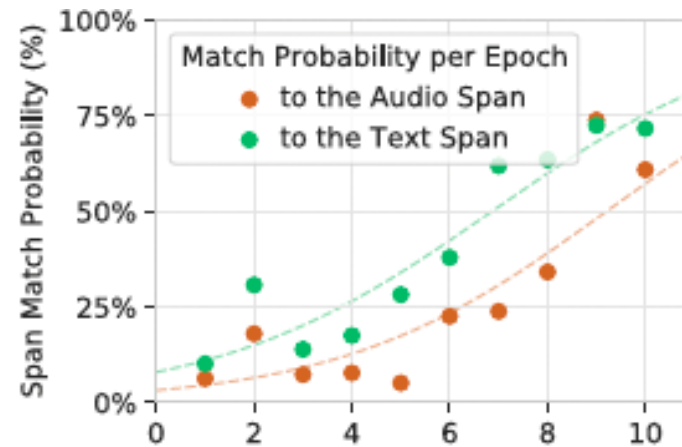
why *exhausted laugh*



Qualitative Analysis - Why does audio help?

00:19	00:11	00:00	00:00	00:04	00:44	00:39	00:39
weight on the legs and get more stretch in the calves in these 45 ...	because the next one is slightly ...	alright shake out your arms and your legs if you need forth a	single lick down dog where we ... [MASK]	leg extended completely straight and heel on the floor	left leg bent and place on top of right on the right leg for the maximum	stretch in the calf press and push your hands into the floor for more stretch i know

this time we're holding it with the right leg



Inference Demo

Video images used



Masked Dialogue

In this video I'll be MASK

Possible Answers

Making a lemonade	Going backpacking	
Archery	Making coffee	Arm wrestling
Horseback riding	Baking cookies	
Brushing hair	Cricket	Painting

Top Five Answers with audio

Making coffee	99%
Drinking coffee	0.6%
Making a lemonade	0.1%
Making an omelette	0.1%
Starting a campfire	0.1%

Inference Demo

Video images used



Masked Dialogue

In this video I'll be MASK

Possible Answers

Making a lemonade Going backpacking
Archery Making coffee Arm wrestling
Horseback riding Baking cookies
Brushing hair Cricket Painting

Top Five Answers without audio

Making coffee	96.7%
Drinking coffee	2.1%
Making a lemonade	0.4%
Starting a campfire	0.3%
Making a cake	0.1%

Limitations & Potential Ideas

- **Privacy concerns** and **dataset access issues** stem from YouTube videos.
- **Noisy training** text generated from YouTube's **ASR**.
- Use of a **single frame per segment**.
 - **Potential idea:** Use a cheap learned pooling mechanism.
- **Only contrastive losses** were used in the model training. **No generative loss**.
 - **Potential ideas:** Perform ablation studies based on above statements and also test the model with addition of generative losses.
- Model can **only assess 40 seconds** of a video.

Thank you.

