Language Technologies Institute

Carnegie Mellon University

Towards Real-World Multimodal AI

Louis-Philippe (LP) Morency

MultiComp Lab

PhD students:	**Chaitanya Ahuja, Volkan Cirik, Paul Liang, Victoria Lin, Hubert Tsai, Alexandria Vail, Torsten Wörtwein and Amir Zadeh**
Research assistants:	**Taylor Gautreaux, Martin Ma and Jed Yang**
Lab coordinator:	**Nicole Siverling**
Project assistant:	**John Friday**

Multimodal AI Technologies

Robots

Virtual Humans

Ubiquitous

Mobile

Wearable

Online

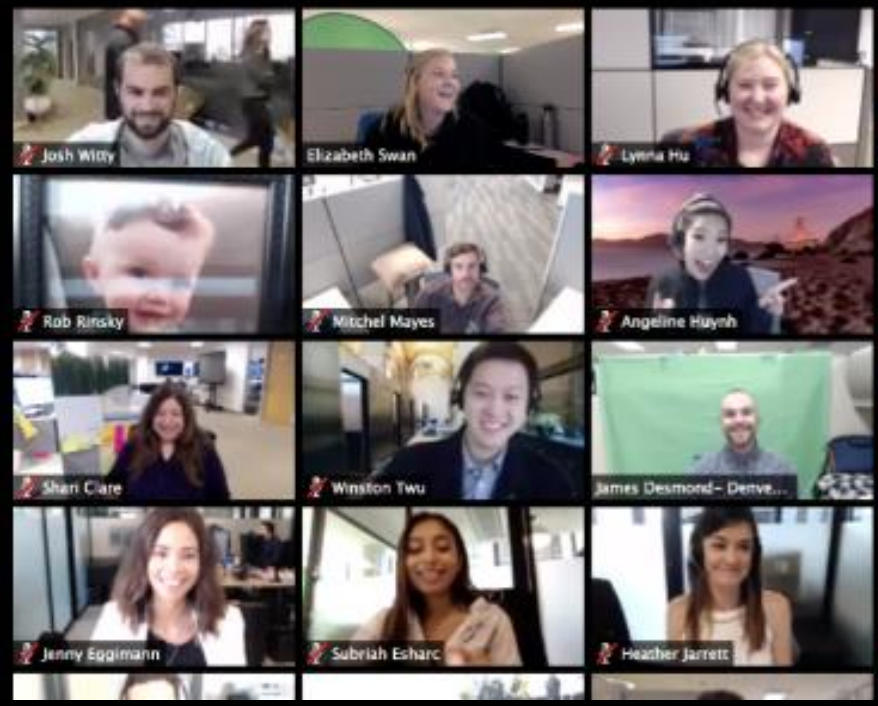
Multimodal AI Technologies

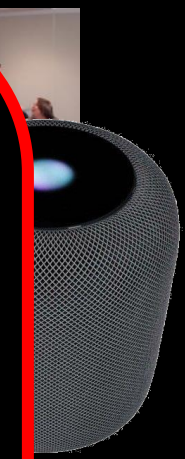
Robots

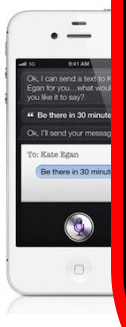
Virtual Humans

Ubiquitous

Video Conferencing







Multimodal Communicative Behaviors

Verbal

- **Lexicon**
  - Words
- **Syntax**
  - Part-of-speech
  - Dependencies
- **Pragmatics**
  - Discourse acts

Vocal

- **Prosody**
  - Intonation
  - Voice quality
- **Vocal expressions**
  - Laughter, moans

Visual

- **Gestures**
  - Head gestures
  - Eye gestures
  - Arm gestures
- **Body language**
  - Body posture
  - Proxemics
- **Eye contact**
  - Head gaze
  - Eye gaze
- **Facial expressions**
  - FACS action units
  - Smile, frowning

Multimodal AI

Verbal

We saw the yellow dog

Visual

Vocal

Multimodal AI

Disorders
- Depression
- Distress
- Autism

Social
- Leadership
- Empathy
- Engagement

Emotion
- Sentiment
- Persuasion
- Frustration

Core Challenges in Multimodal AI

- Representation
- Alignment
- Translation
- Fusion
- Co-learning

Multimodal Machine Learning: A Survey and Taxonomy

By Tadas Baltrusaitis, Chaitanya Ahuja, and Louis-Philippe Morency

(IEEE TPAMI journal, February 2019)

https://arxiv.org/abs/1705.09406

Graduate-level course on Multimodal Machine learning

(10th edition)

https://cmu-multicomp-lab.github.io/mmml-course/fall2020/

Real-World Multimodal AI Applications

Healthcare
Decision Support

Leadership and
Team Collaborations

Online Learning
and Education

Challenges for Real-World Multimodal AI

Core Challenges

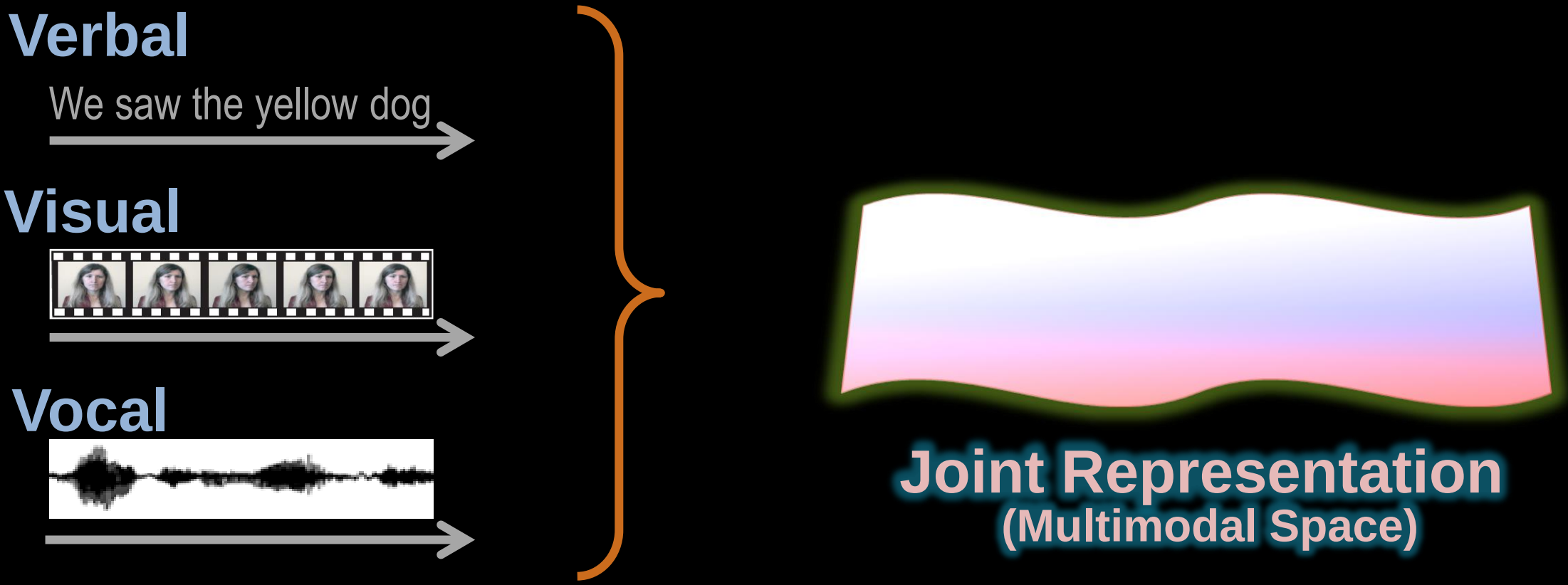
- Representation
- Alignment
- Translation
- Fusion
- Co-learning

Real-World Challenges

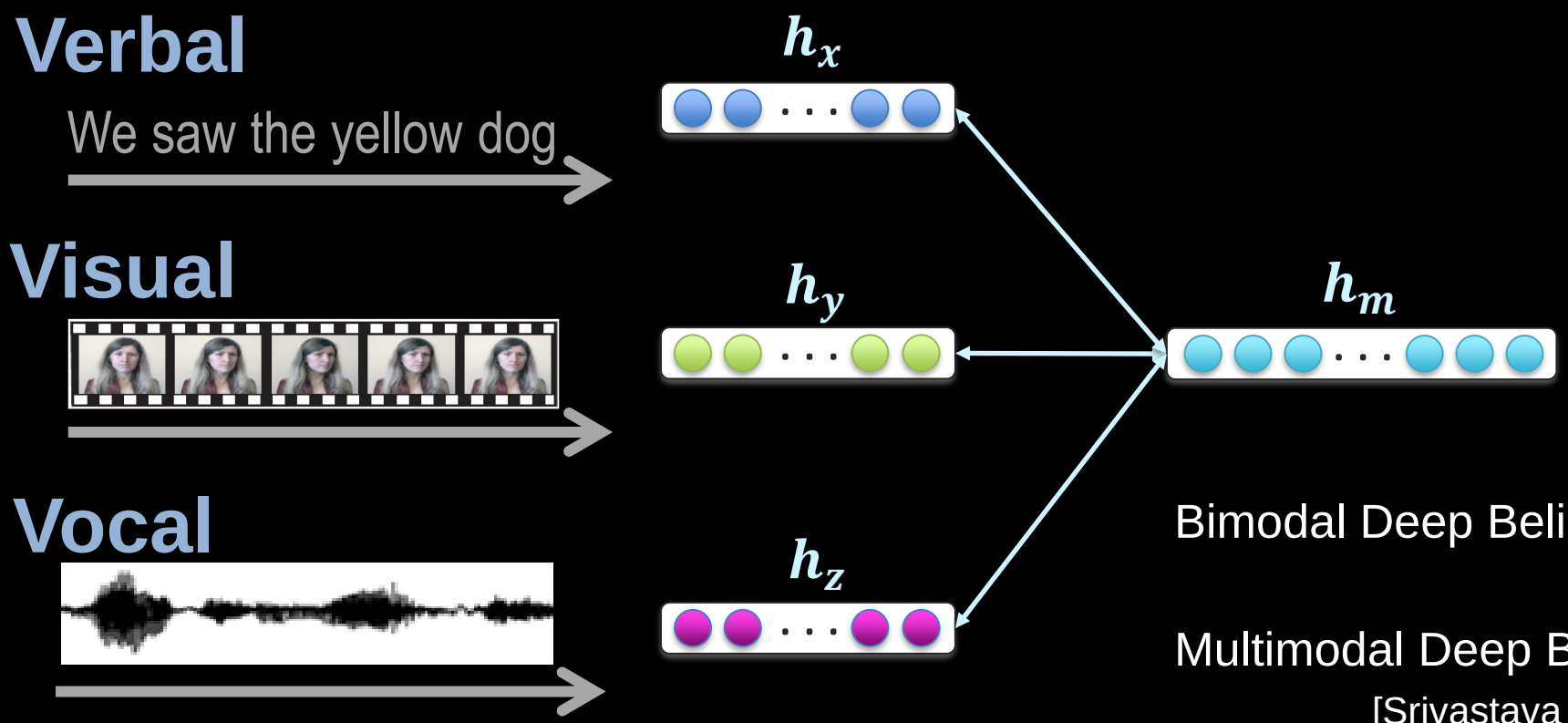
- Robustness
- Thrustworthy
- Variability
- Fairness
- Privacy

Core Challenge 1: Multimodal Representation

Definition: Learning how to represent and summarize multimodal data in away that exploits the **complementarity** and **redundancy**.



Multimodal Joint Representation: Previous Approaches



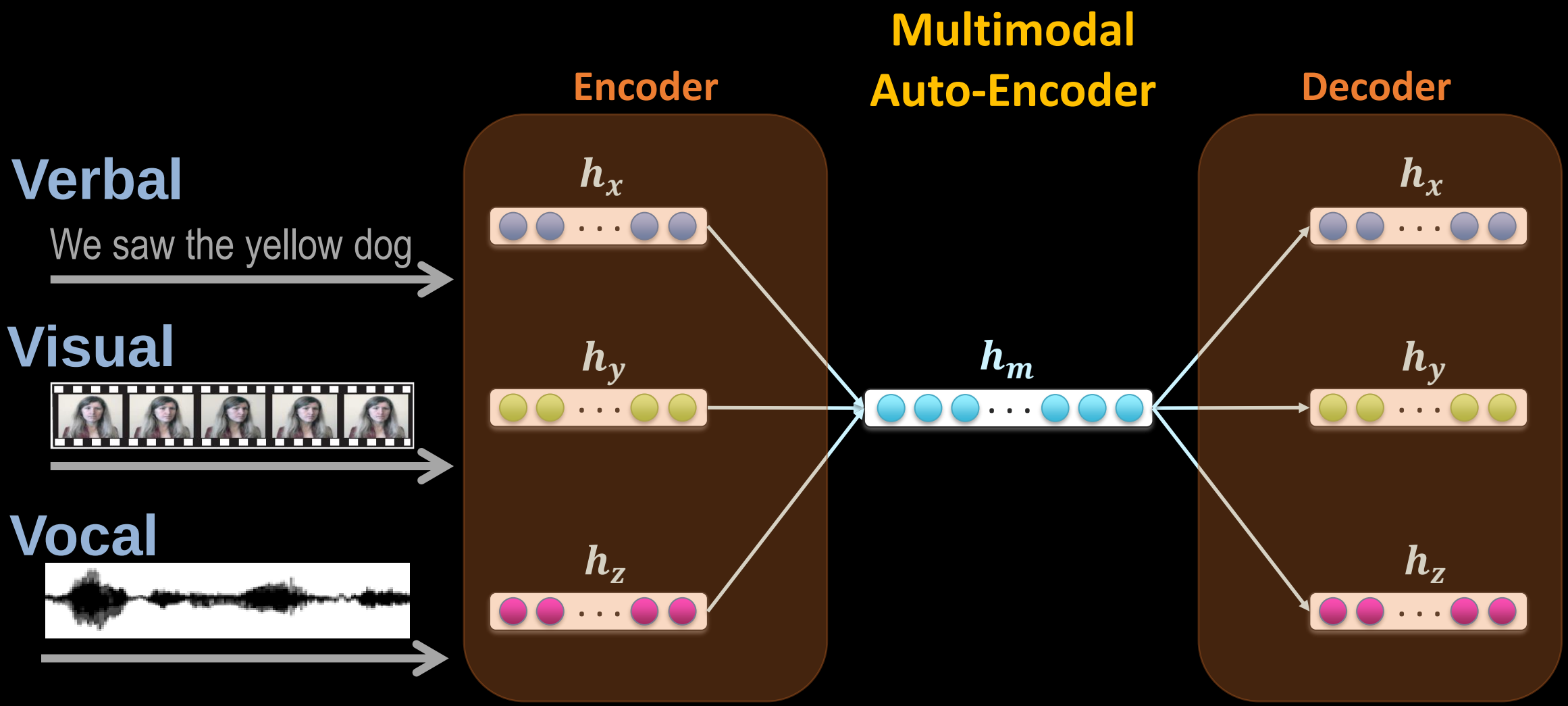
Bimodal Deep Belief Network

[Ngiam et al., ICML 2011]

Multimodal Deep Boltzmann Machine

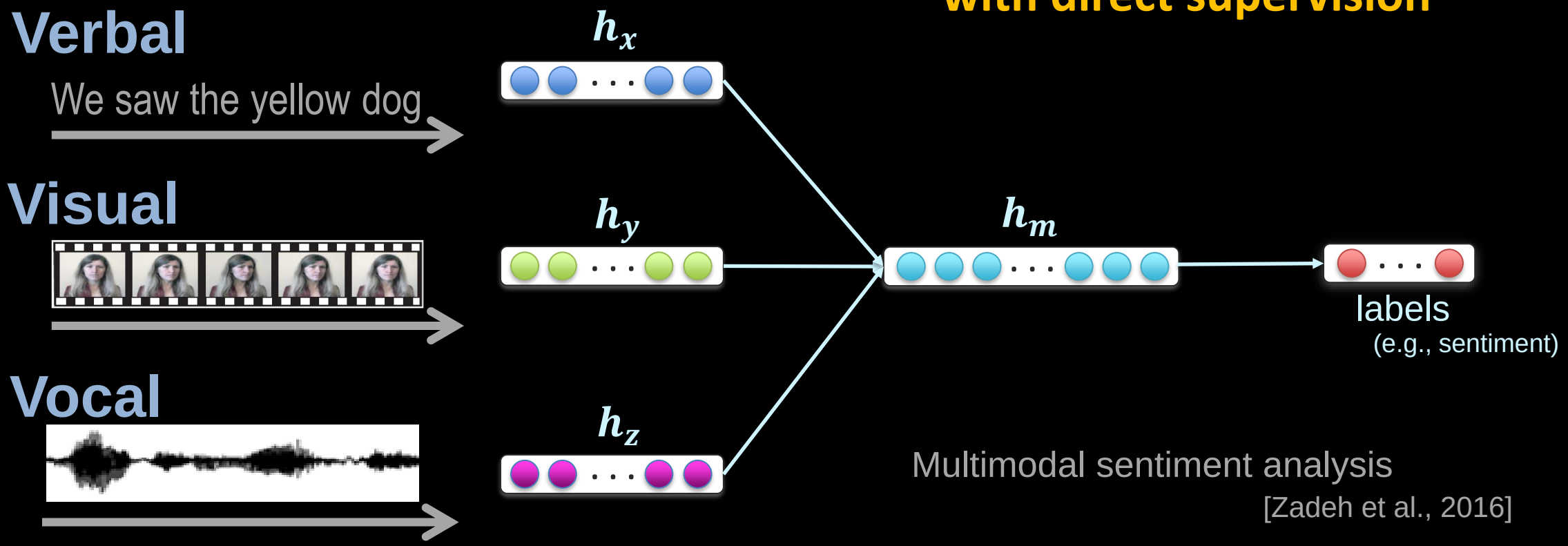
[Srivastava and Salahutdinov, NIPS 2012]

Multimodal Joint Representation : Previous Approaches



Multimodal Joint Representation: Previous Approaches

Representation learned with direct supervision



Multimodal sentiment analysis

[Zadeh et al., 2016]

Cross-Modal Interactions

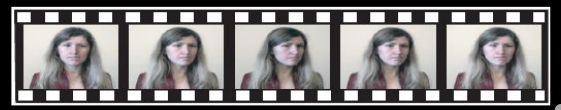
Unimodal

Bimodal

Trimodal

Verbal

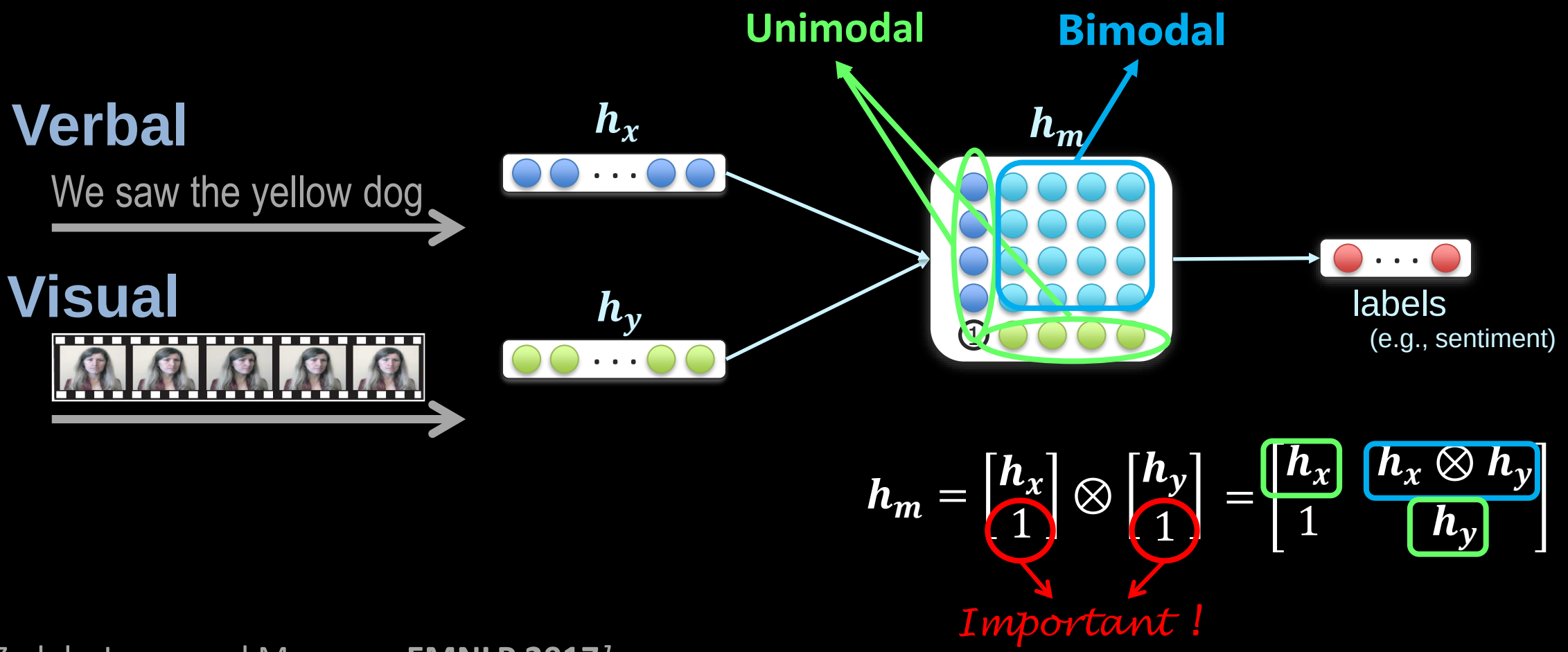
We saw the yellow dog



Visual

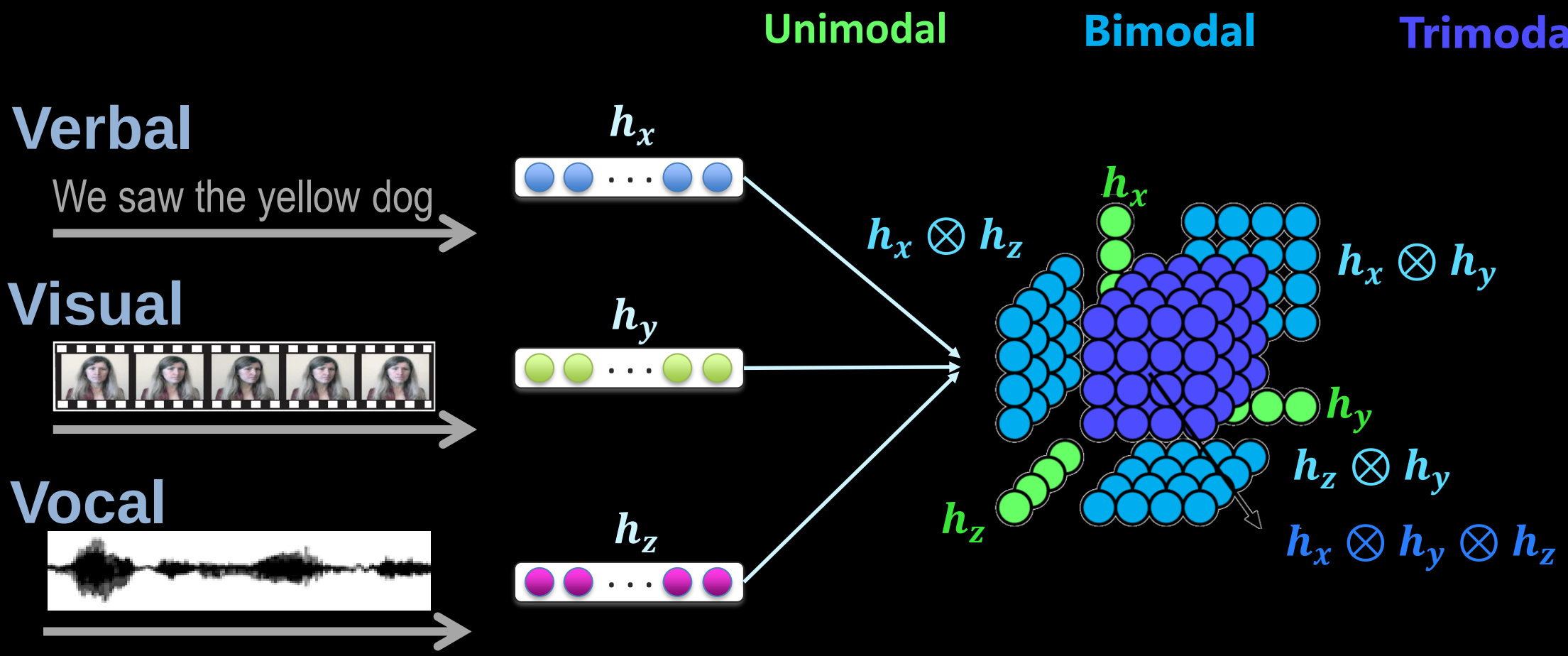
Vocal

Representation using Tensor Fusion Network



[Zadeh, Jones and Morency, **EMNLP 2017**]

Cross-Modal Interactions

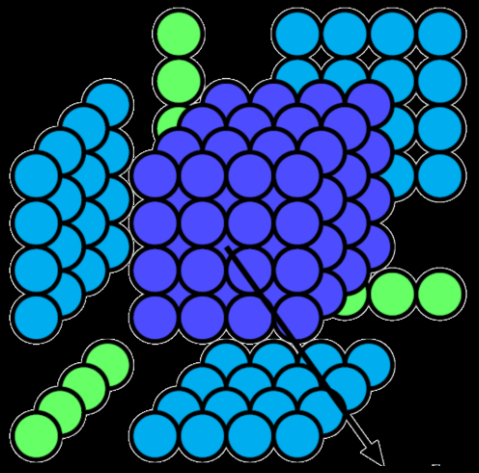


[Zadeh, Jones and Morency, **EMNLP 2017**]

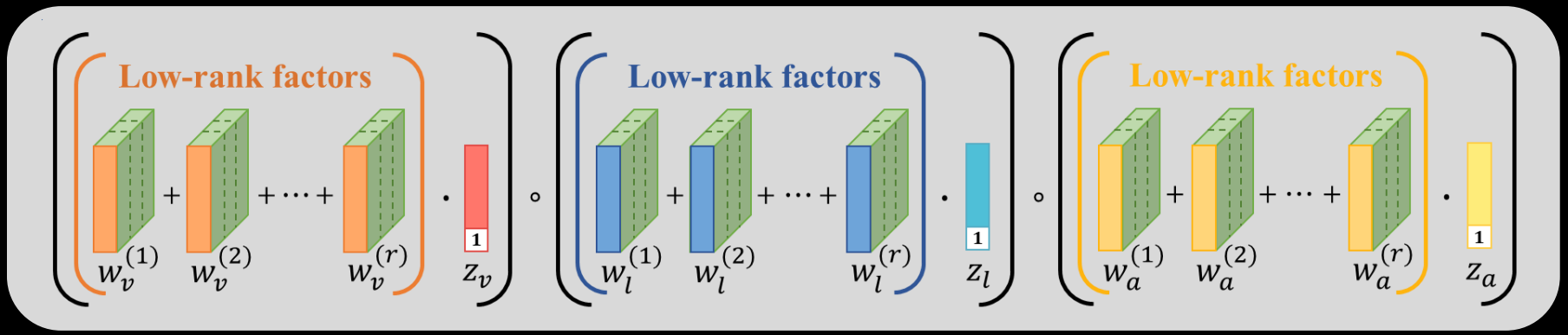
Improving Efficiency of Multimodal Representations

Tensor Fusion Network: Explicitly models unimodal, bimodal and trimodal interactions

[Zadeh, Jones and Morency, **EMNLP 2017**]



Efficient Low-rank Multimodal Fusion



Canonical Polyadic Decomposition

[Liu, Shen, Bharadwaj, Liang, Zadeh and Morency, **ACL 2018**]

Representation + Robustness

Missing data!

Verbal

Visual

Vocal

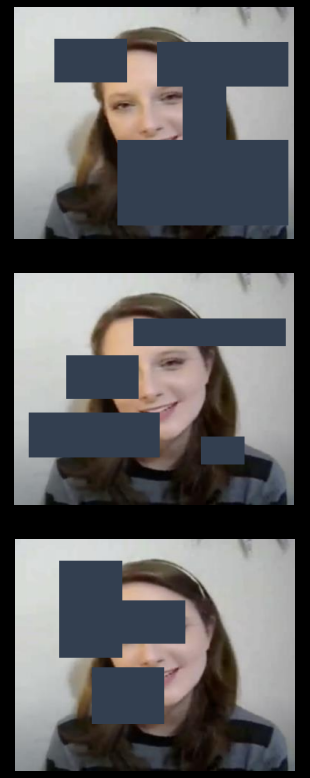
One option:
Variational Auto-Encoder

Encoder

Decoder

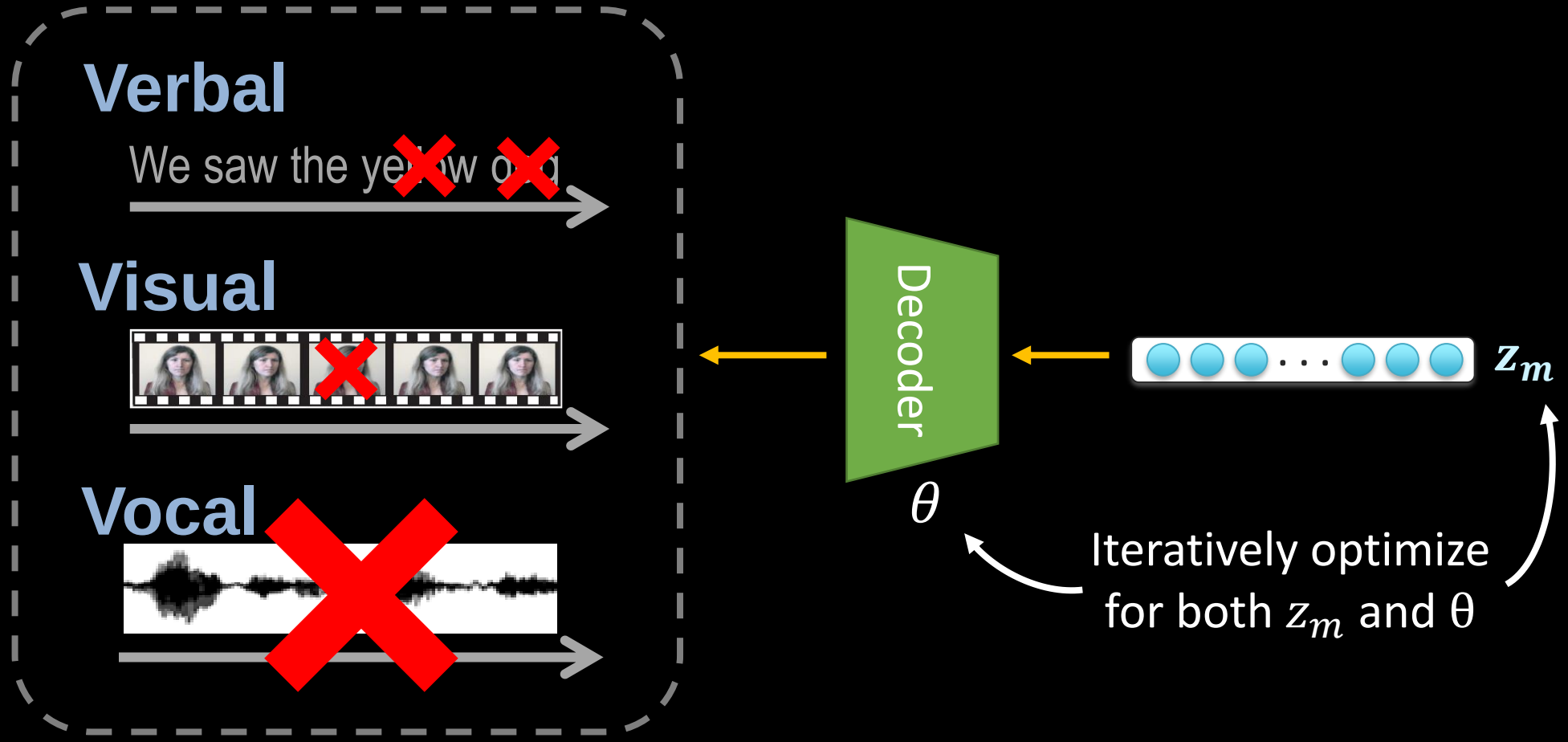
z_m

Robust to different missing patterns?

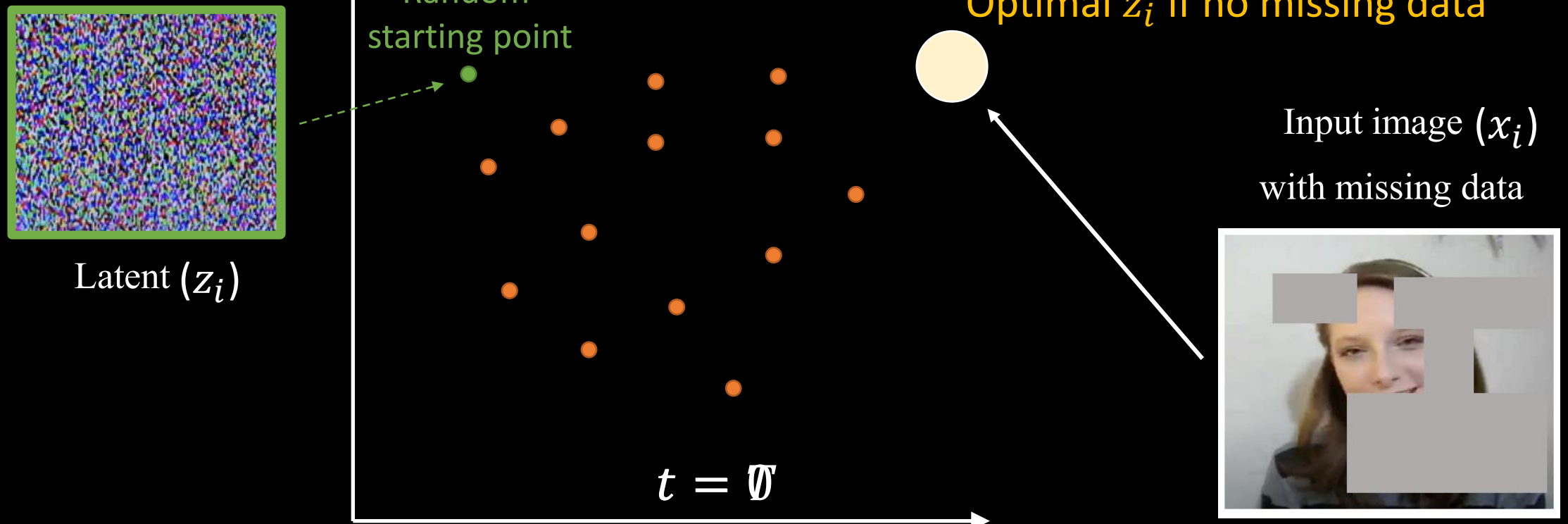


Variational Auto-Decoder [Zadeh et al., 2019]

Missing data!

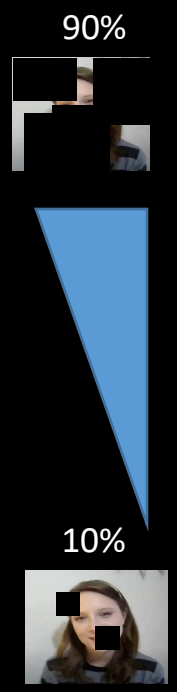


Variational Auto-Decoder [Zadeh et al., 2019]

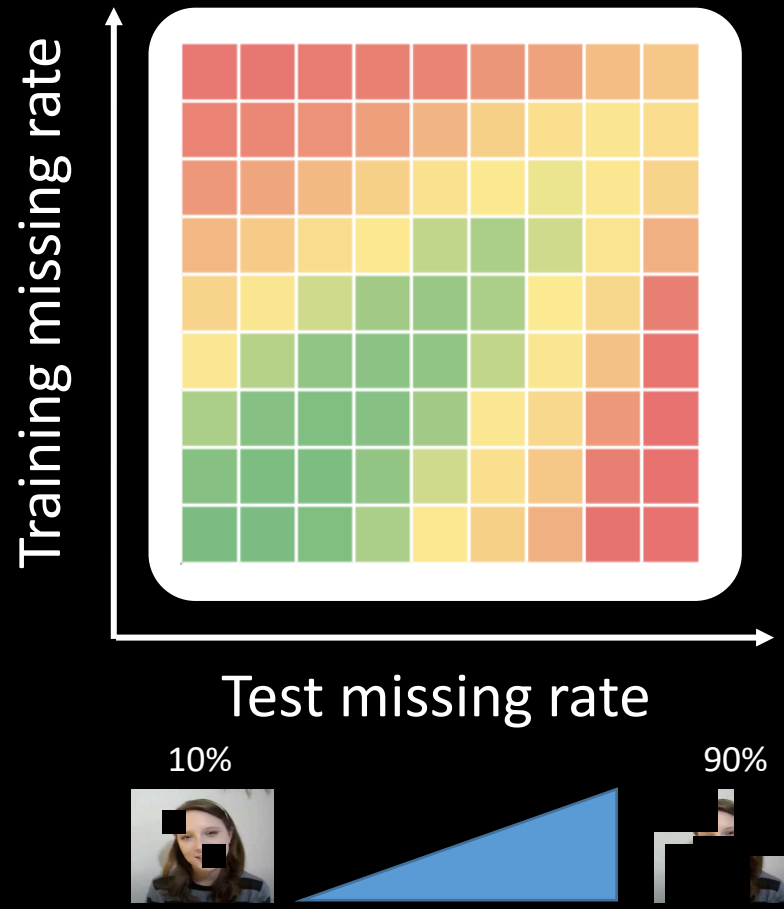


Variational Auto-Decoder [Zadeh et al., 2019]

Robust to different missing patterns?

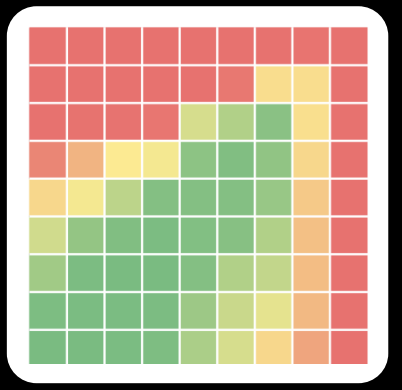


Variational Auto-Encoder

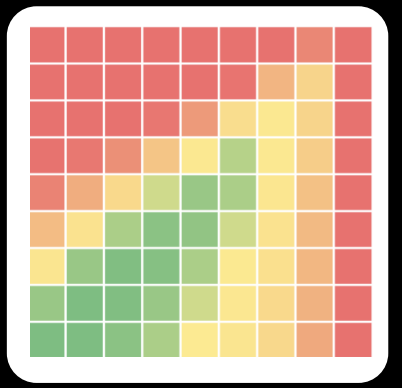


Variational Auto-Decoder

MisGAN

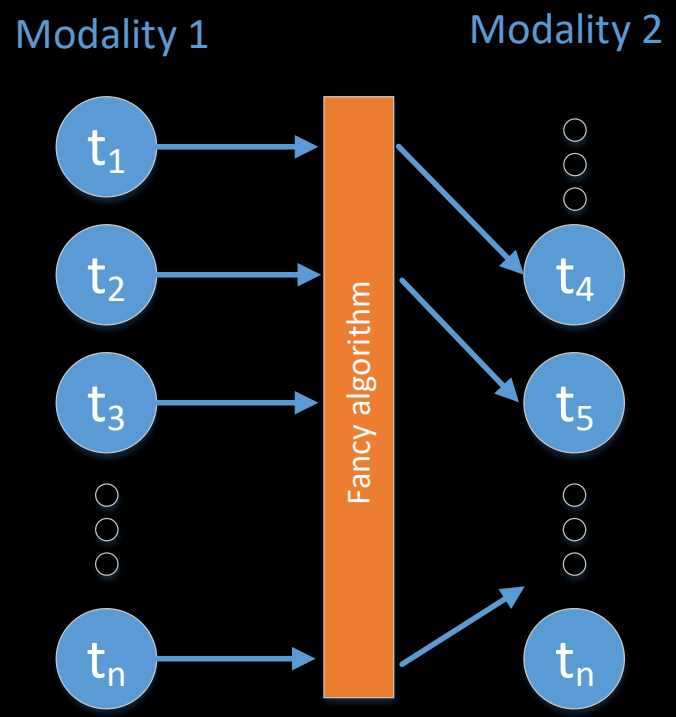


GAIN



Core Challenge 2: Alignment

Definition: Identify the direct relations between (sub)elements from two or more different modalities.



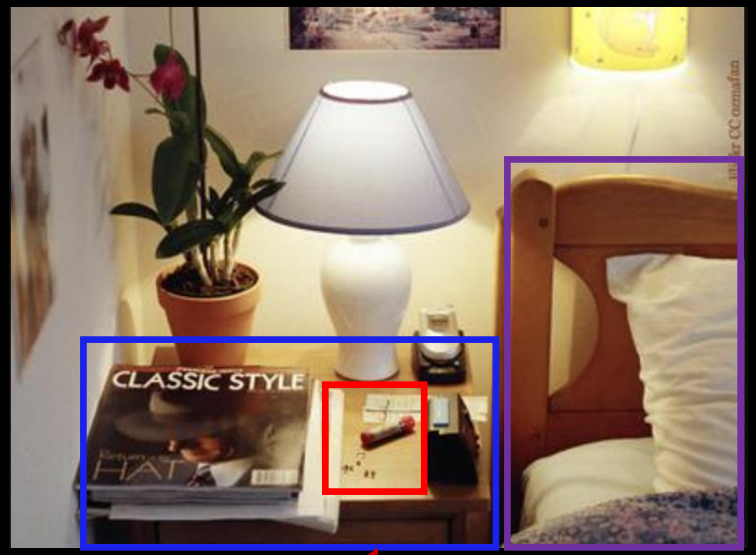
Grounding: Linking Language and the Perceived World

"Could you bring my pills? They should be on top of the nightstand on the left of the bed."

Spatial Grounding

Grounding: Linking Language and the Perceived World

"Could you bring my pills? They should be on top of the nightstand on the left of the bed."



Spatial Grounding: Object entities

Grounding: Linking Language and the Perceived World

"Could you bring my pills? They should be on top of the nightstand on the left of the bed."

How can we thrust that all grounding elements are properly modeled?

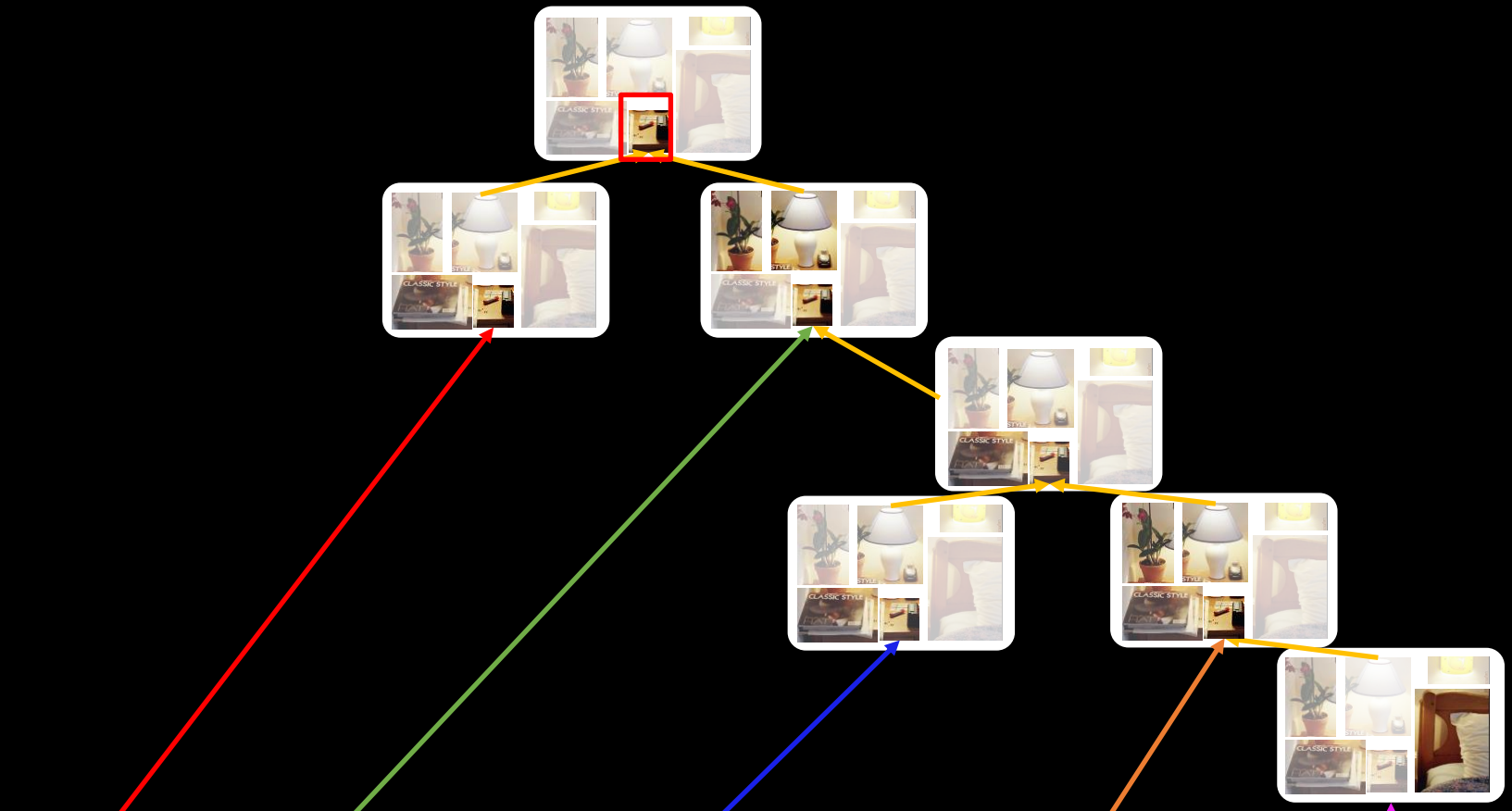
Solution: Interpretability

Spatial Grounding: Object entities + Relationships

GroundNet: Using Linguistic Syntax to Guide Grounding

[Cirik et al., AAAI 2018]

"Pills on top of the nightstand on the left of the bed"



pills | on top of | the nightstand | on the left of | the bed

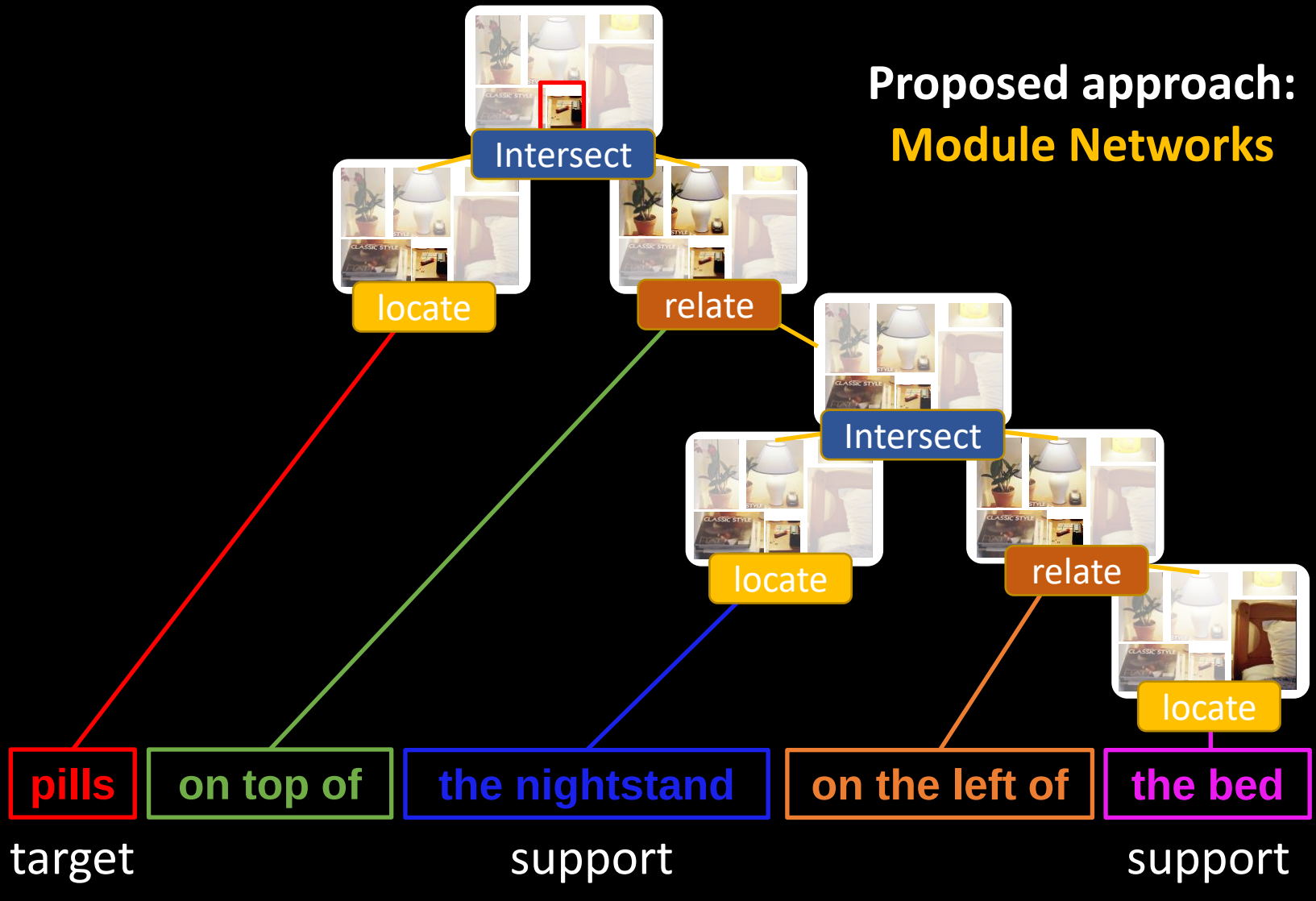
target | support | support

GroundNet: Using Linguistic Syntax to Guide Grounding

[Cirik et al., AAAI 2018]

"Pills on top of the nightstand on the left of the bed"

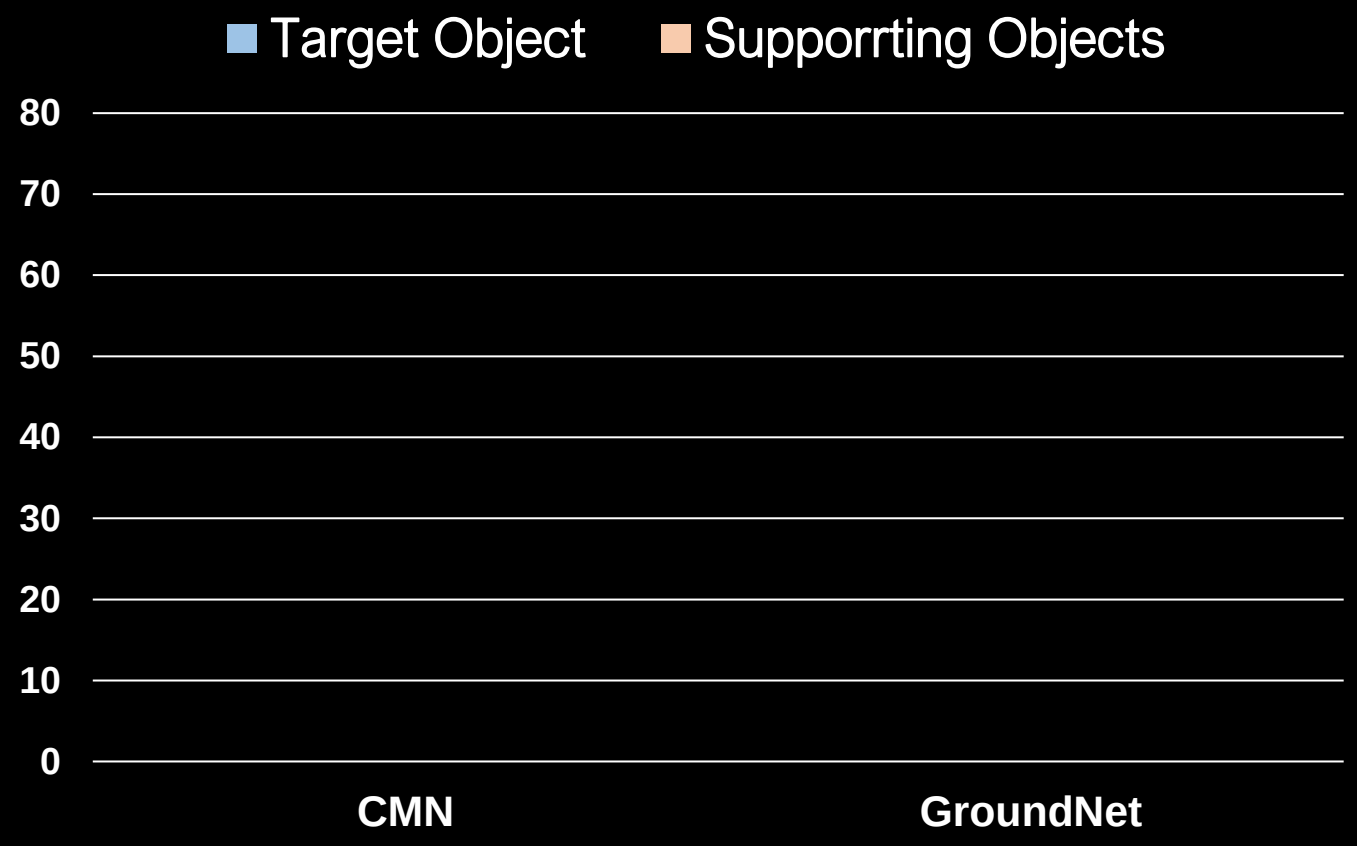
Proposed approach:
Module Networks



GroundNet: Using Linguistic Syntax to Guide Grounding

[Cirik et al., AAAI 2018]

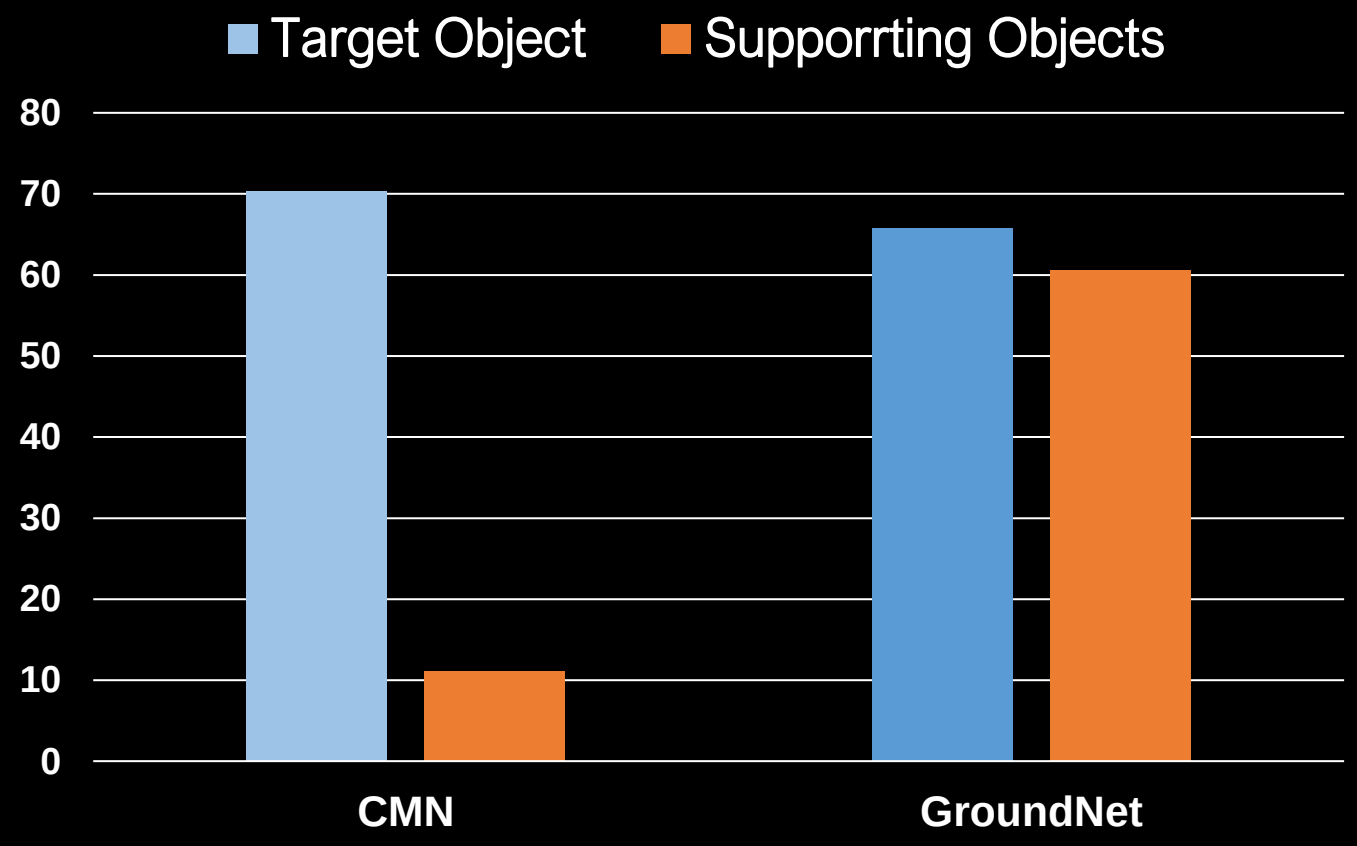
"Pills on top of the nightstand on the left of the bed"



GroundNet: Using Linguistic Syntax to Guide Grounding

[Cirik et al., AAAI 2018]

"Pills on top of the nightstand on the left of the bed"



Refer360: Language-to-Action Dataset [Cirik et al., ACL 2020]

Multi-step instruction:

1. **Go** to the **entrance** of the **lounge area**.
2. **On your right** there will be **a bar**.
3. **On top** of the **counter**, you will see **a box**. **Bring** me **that**.

Dataset:

- 17,135 annotated instances
- 2,000 panaromic 360 degrees scenes
- 43.8 average number of words per instructions

https://github.com/volkancirik/refer360

Alignment and Representation

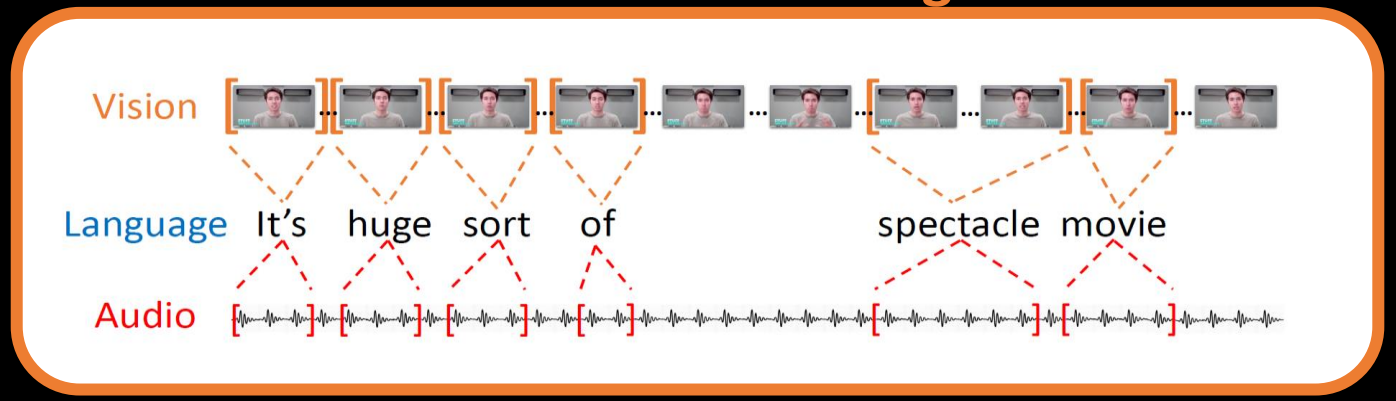
Visual

Vocal

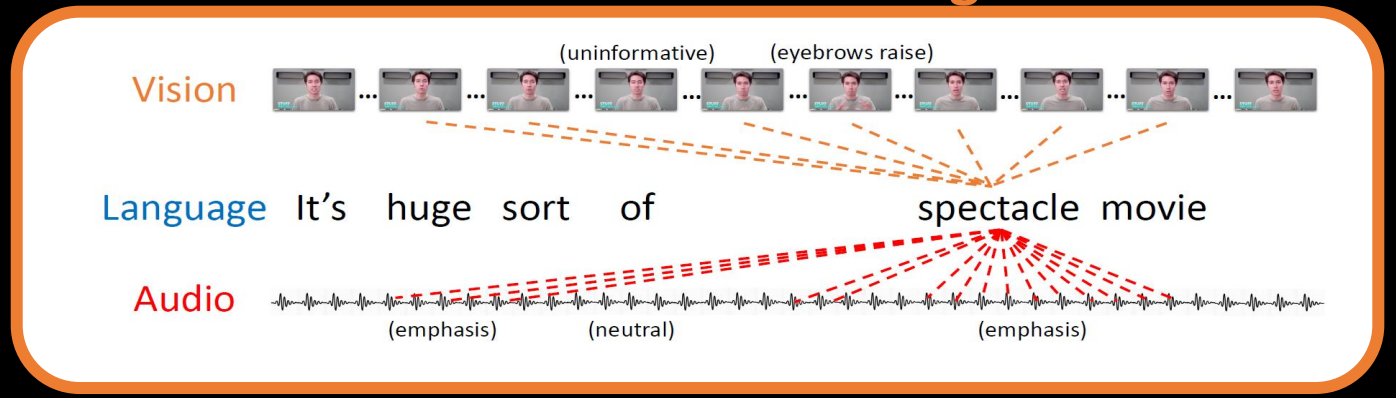
Verbal

"I like…"

Predefined Word-level alignment



Automatic Cross-Modal alignment



time

Multimodal representation

Representation

Alignment

Multimodal Transformer [Tsai et al., ACL 2019]

Alignment

Representation

Visual

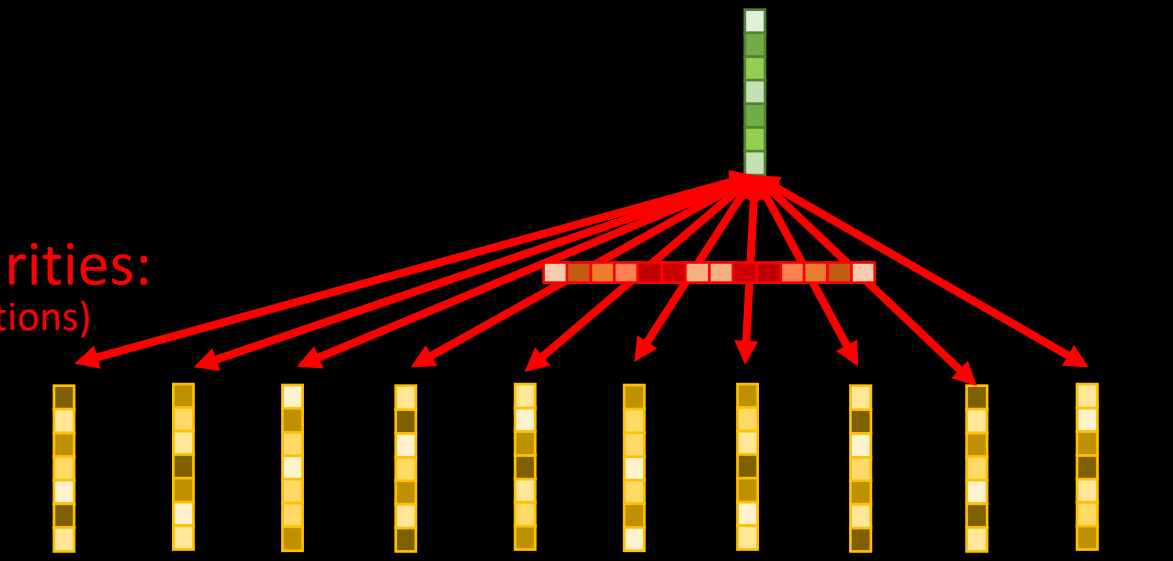
Visually contextualizing the verbal modality

Verbal

"I like…"

"spectacle"

Similarities:
(attentions)



time

Multimodal Transformer [Tsai et al., ACL 2019]

Alignment

Representation

Visual

Visually contextualizing the verbal modality

Verbal

"I like…"

"spectacle"

Residual connection

" New visually-contextualized representation of language"

Similarities:
(attentions)

Visual embeddings:

Correlated visual information

time

Multimodal Transformer [Tsai et al., ACL 2019]

Alignment

Representation

Visual

Vocal

Verbal

"I like…"

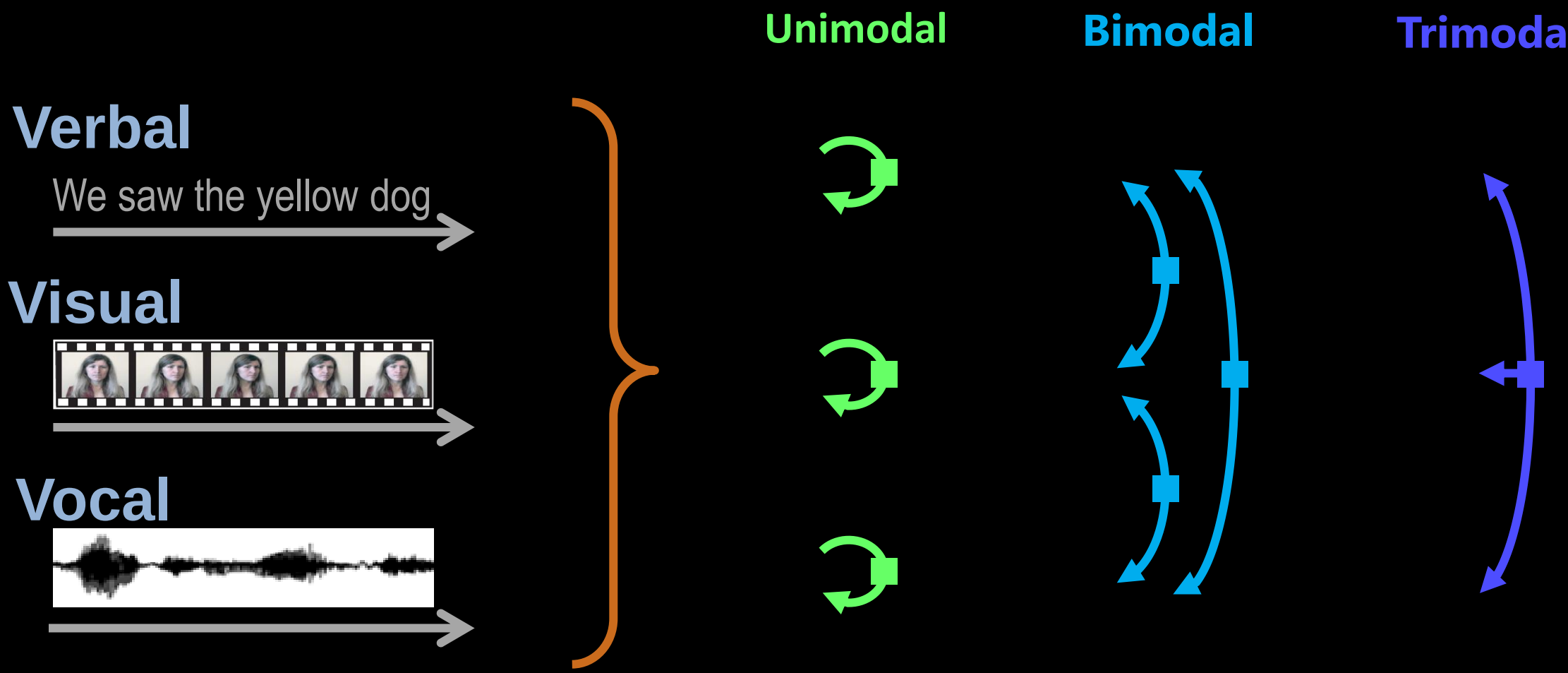
Unimodal Representations

Cross-modally Contextualized Unimodal Representations

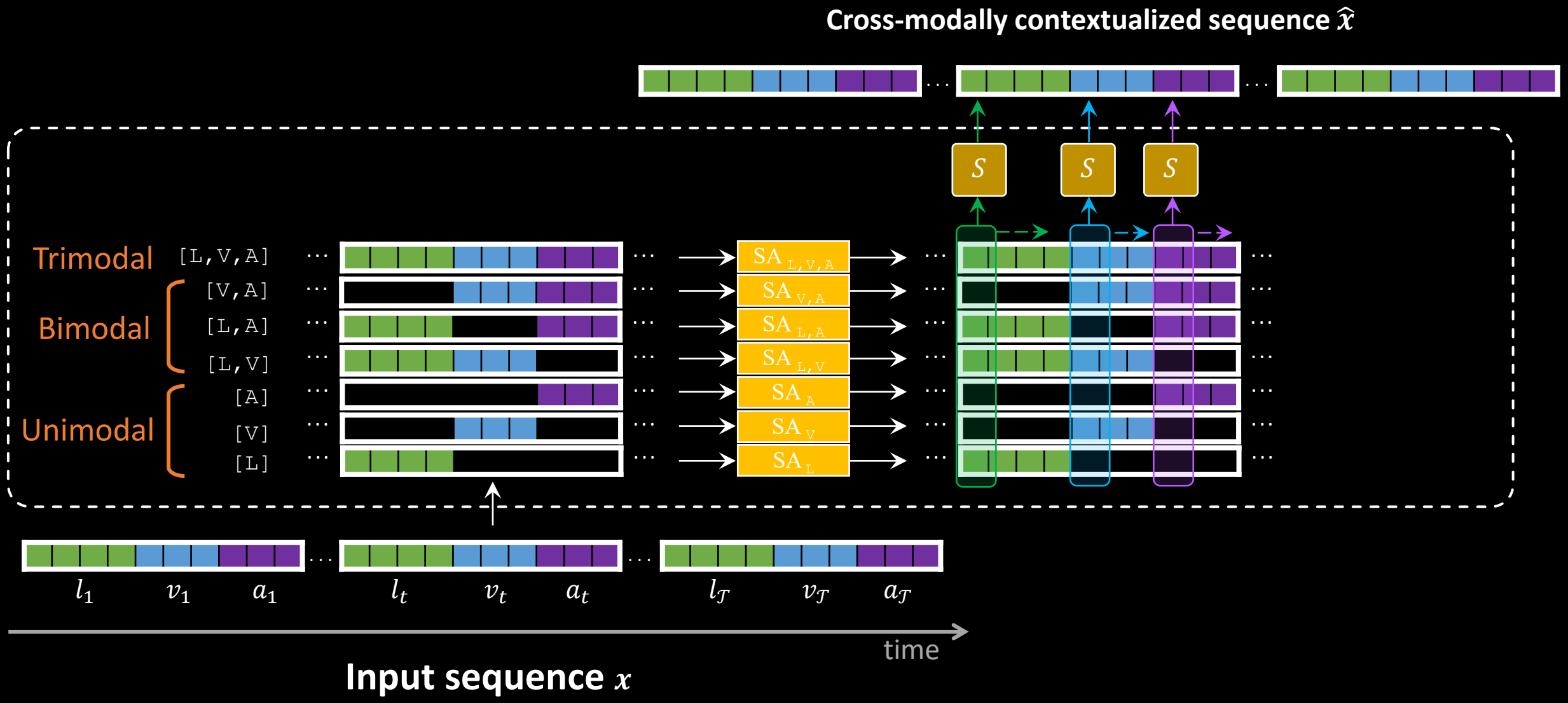
Multimodal representation

Cross-Modal Attention Block x N layers

Model Cross-Modal Interactions with Transformers?



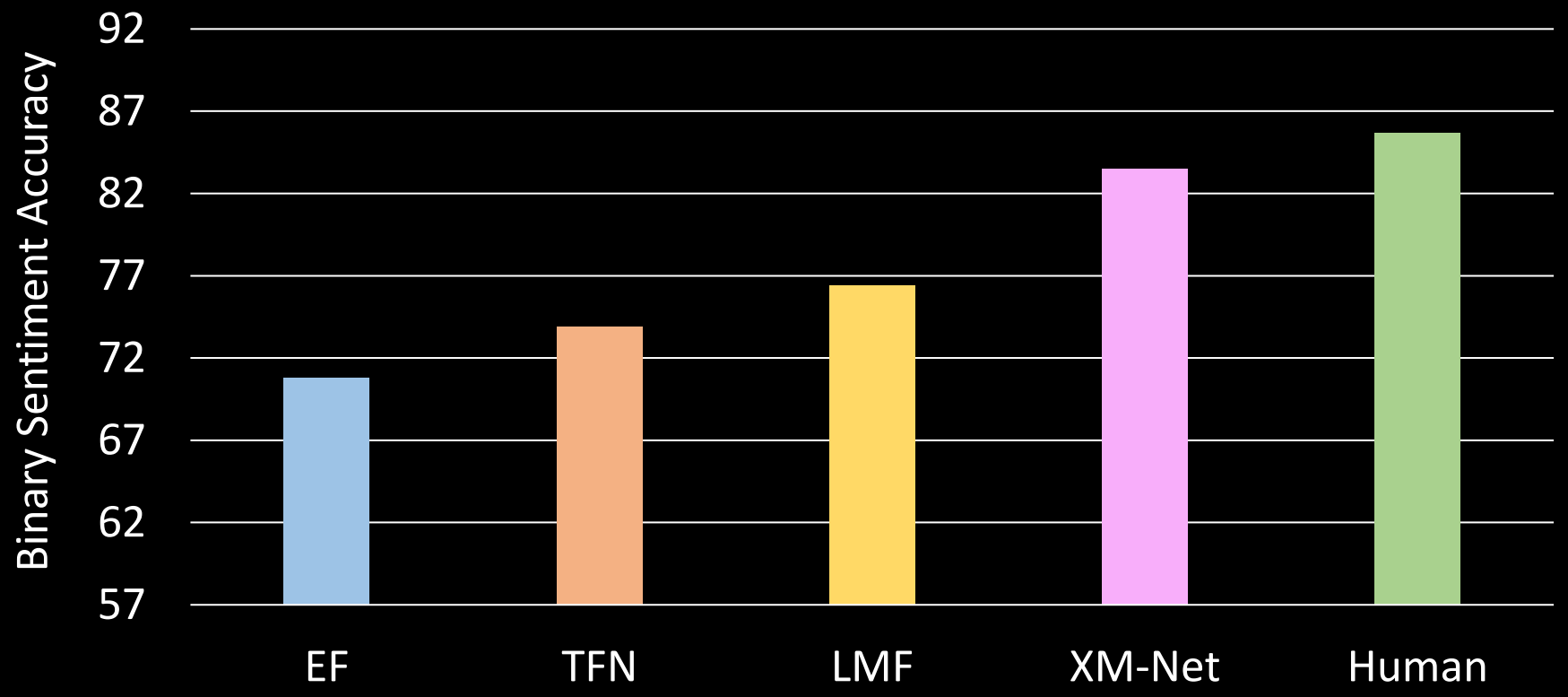
XM-Net: Cross-Modal Transformer Network [Zadeh et al., 2020]



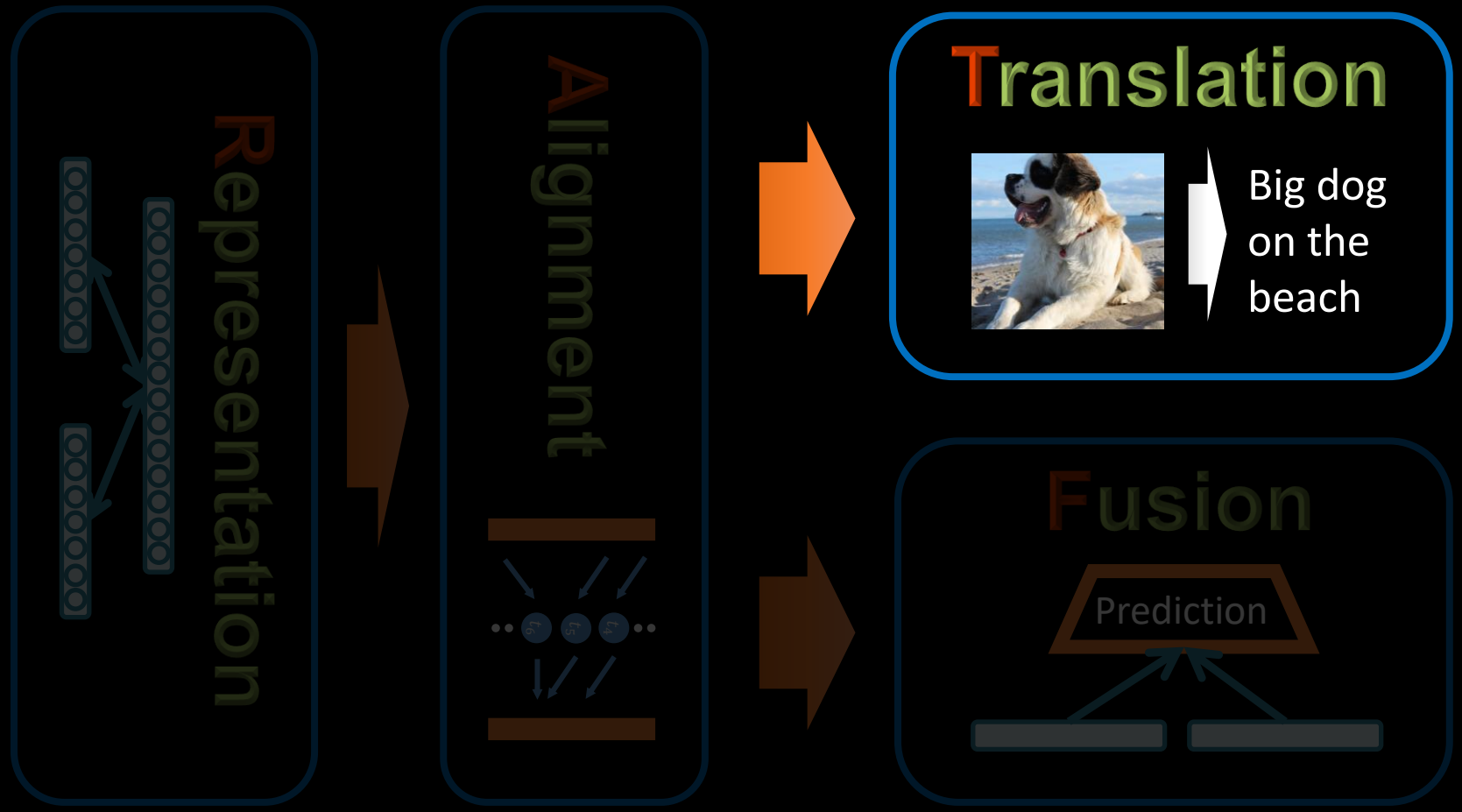
XM-Net: Cross-Modal Transformer Network [Zadeh et al., 2020]

Results on CMU-MOSI Dataset

(multimodal sentiment analysis)



Multimodal AI – Core Challenges [Survey: TPAMI 2019]



Multimodal Translation: Speech-to-Gestures

[IVA 2015]

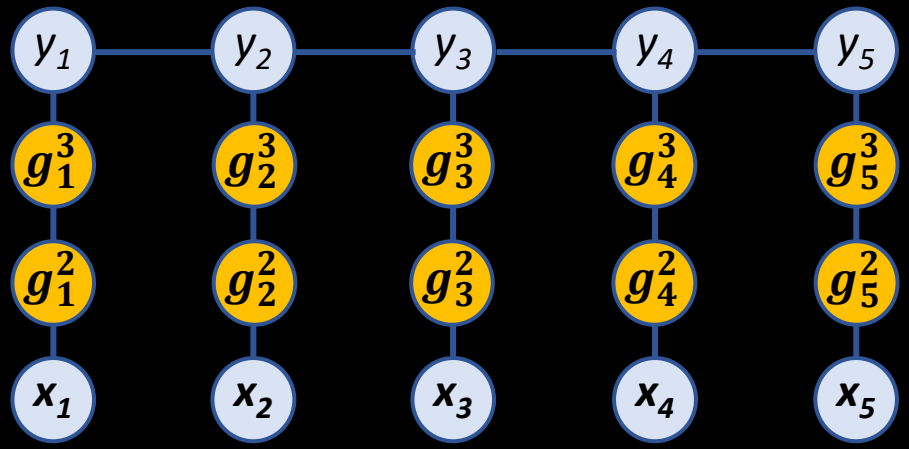
Visual

Vocal

Verbal

"I like…"

Deep Conditional Neural Field

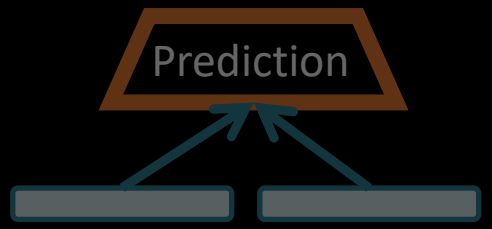


"I like this presentation very much"

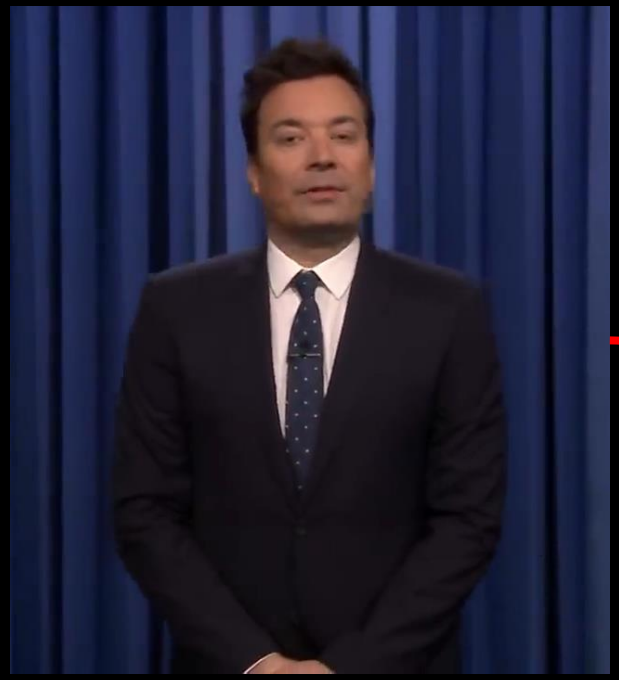
Translation

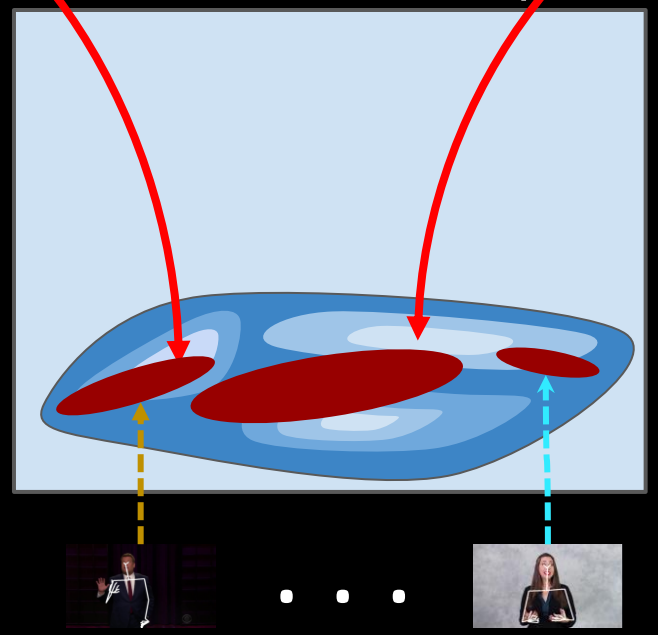
Big dog on the beach

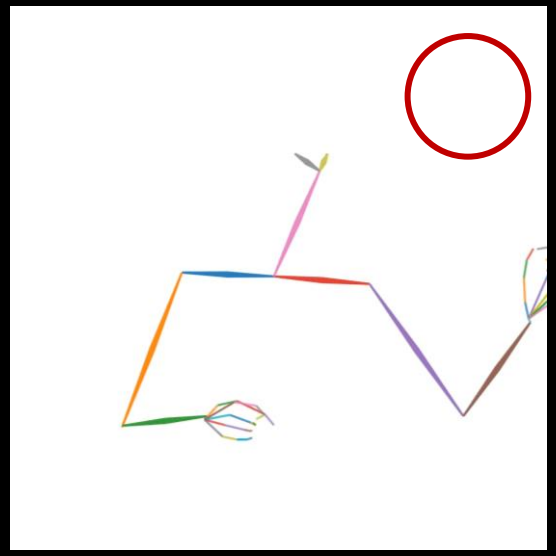
Fusion



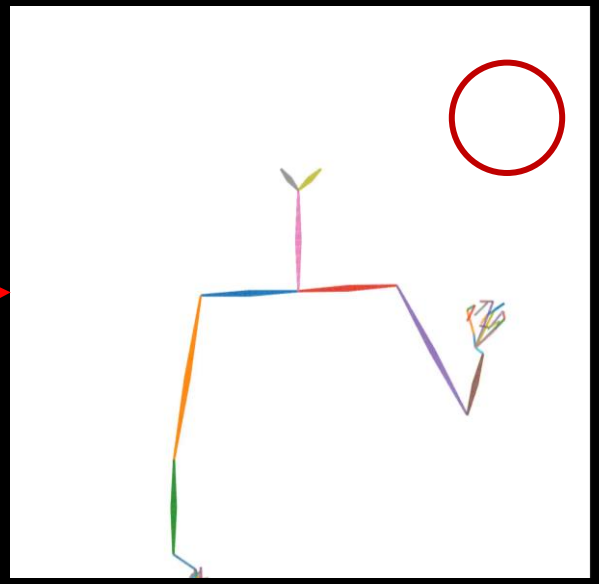
Nonverbal Signatures: Idiosyncrasy and Variability



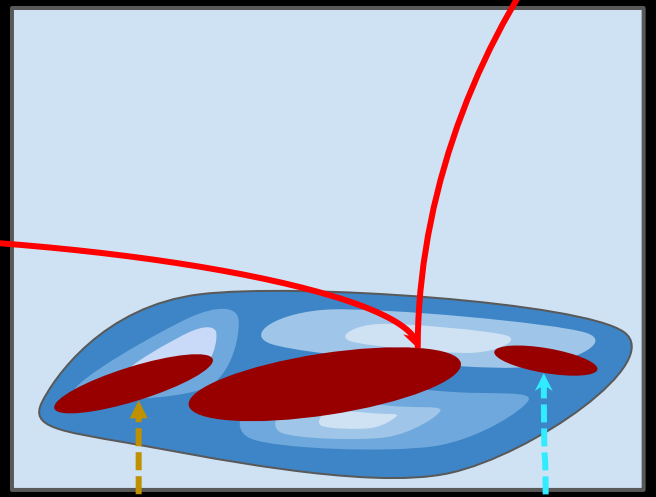


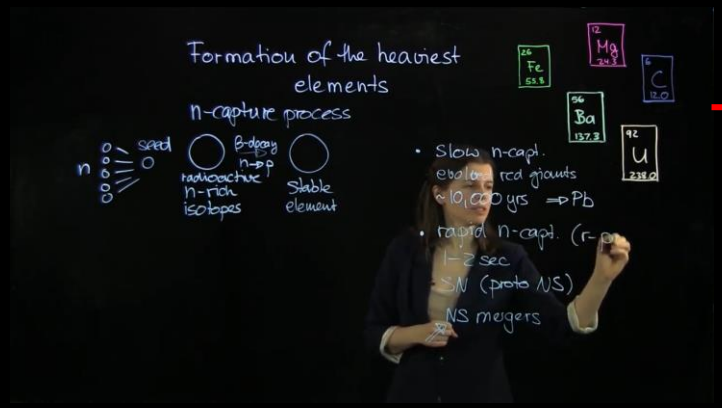


Style Transfer



Multimodal Gesture Space





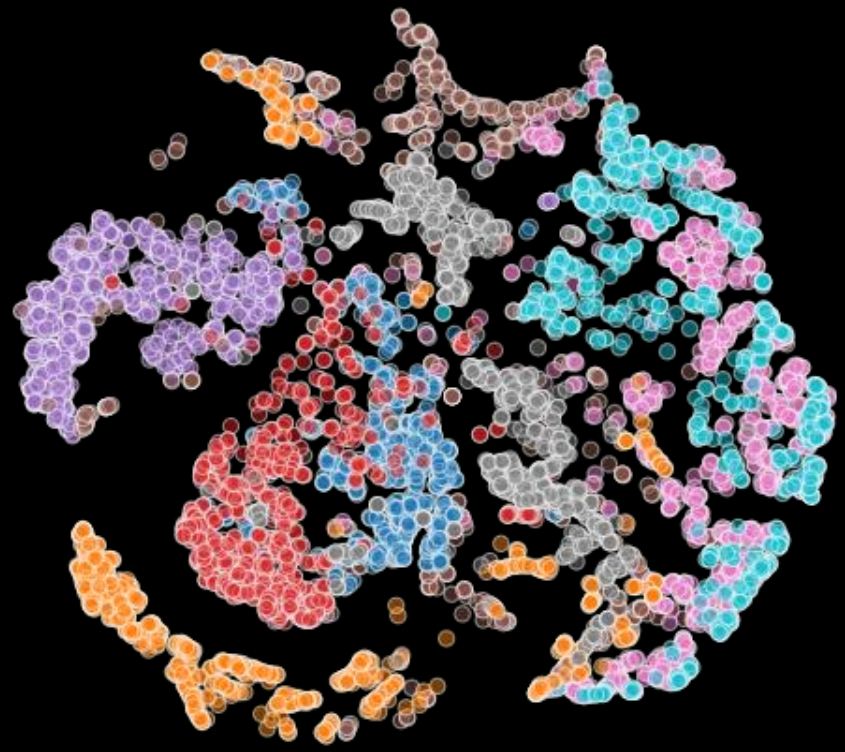
(Source) Speaker A

Driven by audio from speaker B

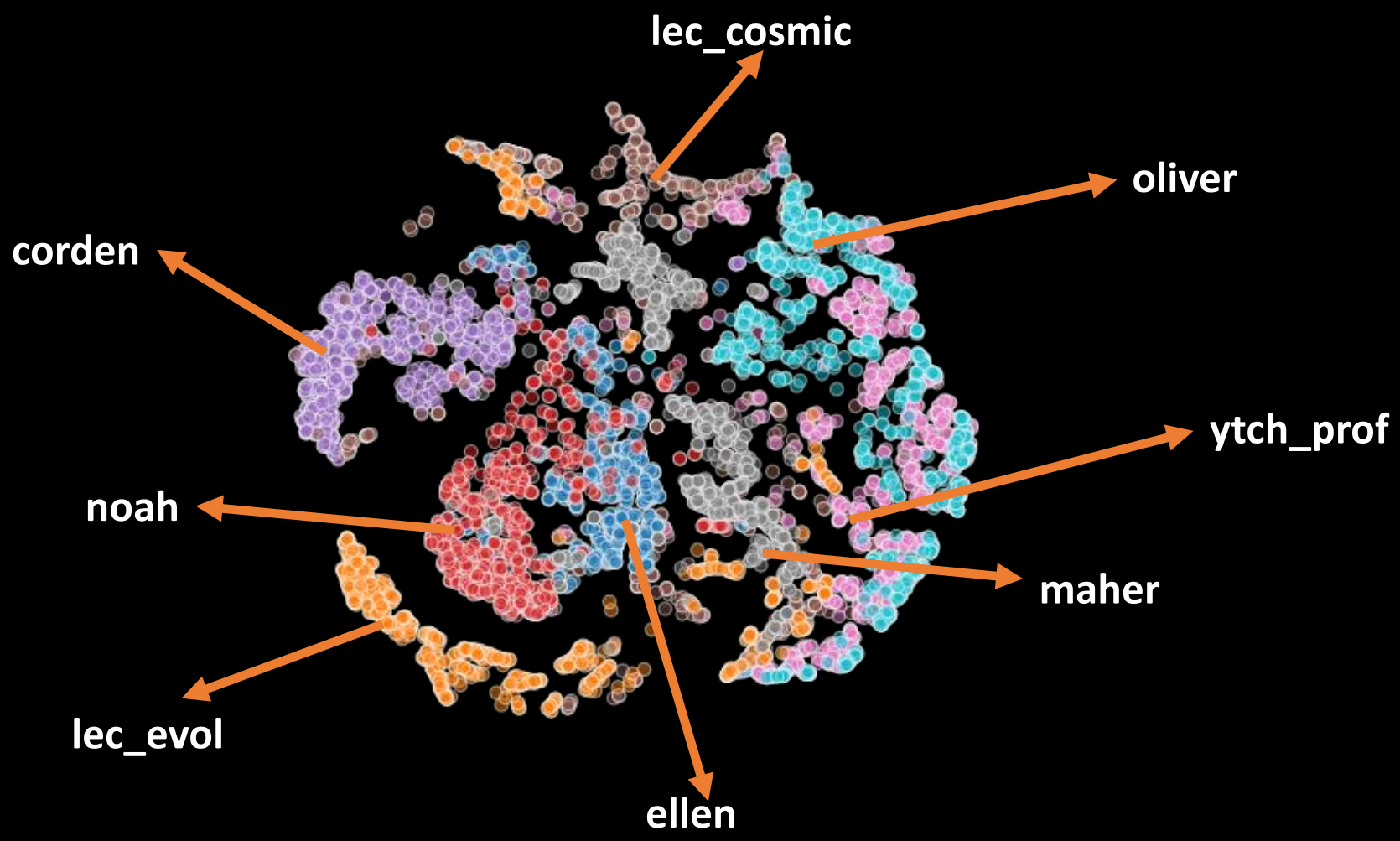
(Target) Speaker B

Multimodal Gesture Space

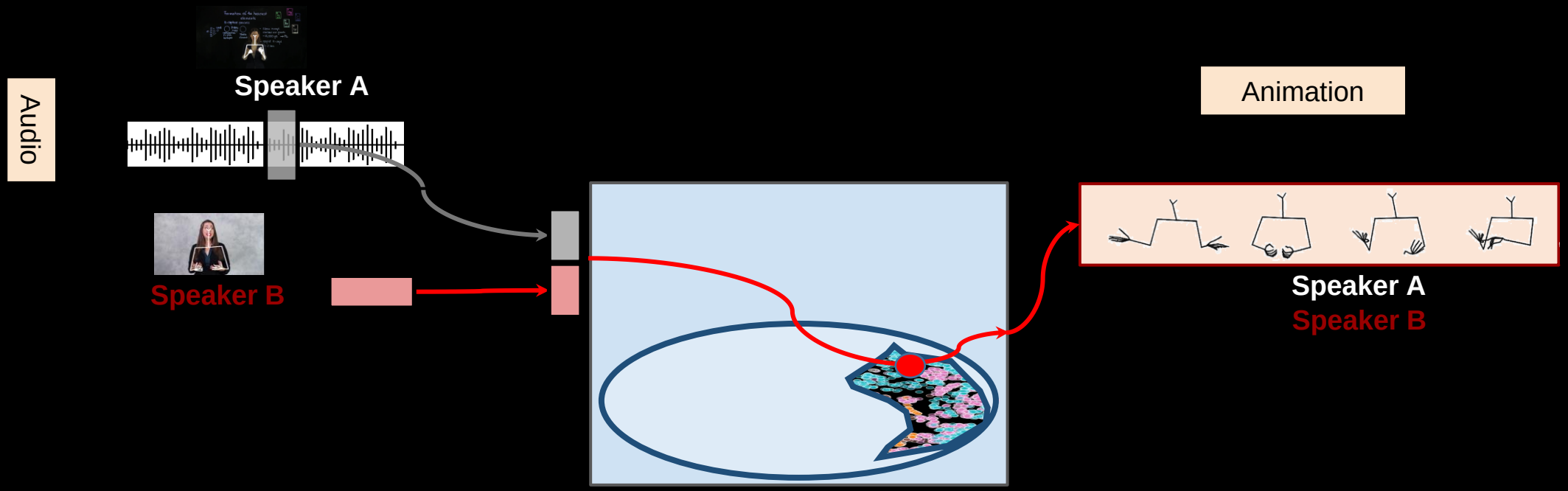
- What does this space represent?
- How do we use this gesture space to generate stylized gestures?
- How do we learn this gesture space?



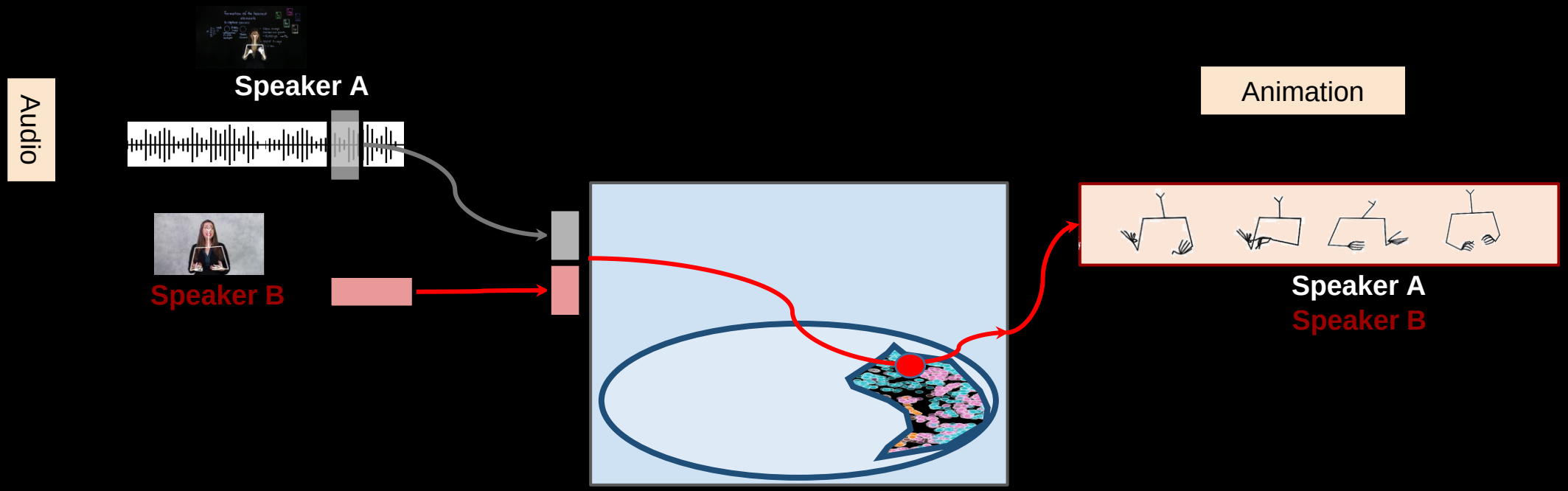
Multimodal Gesture Space



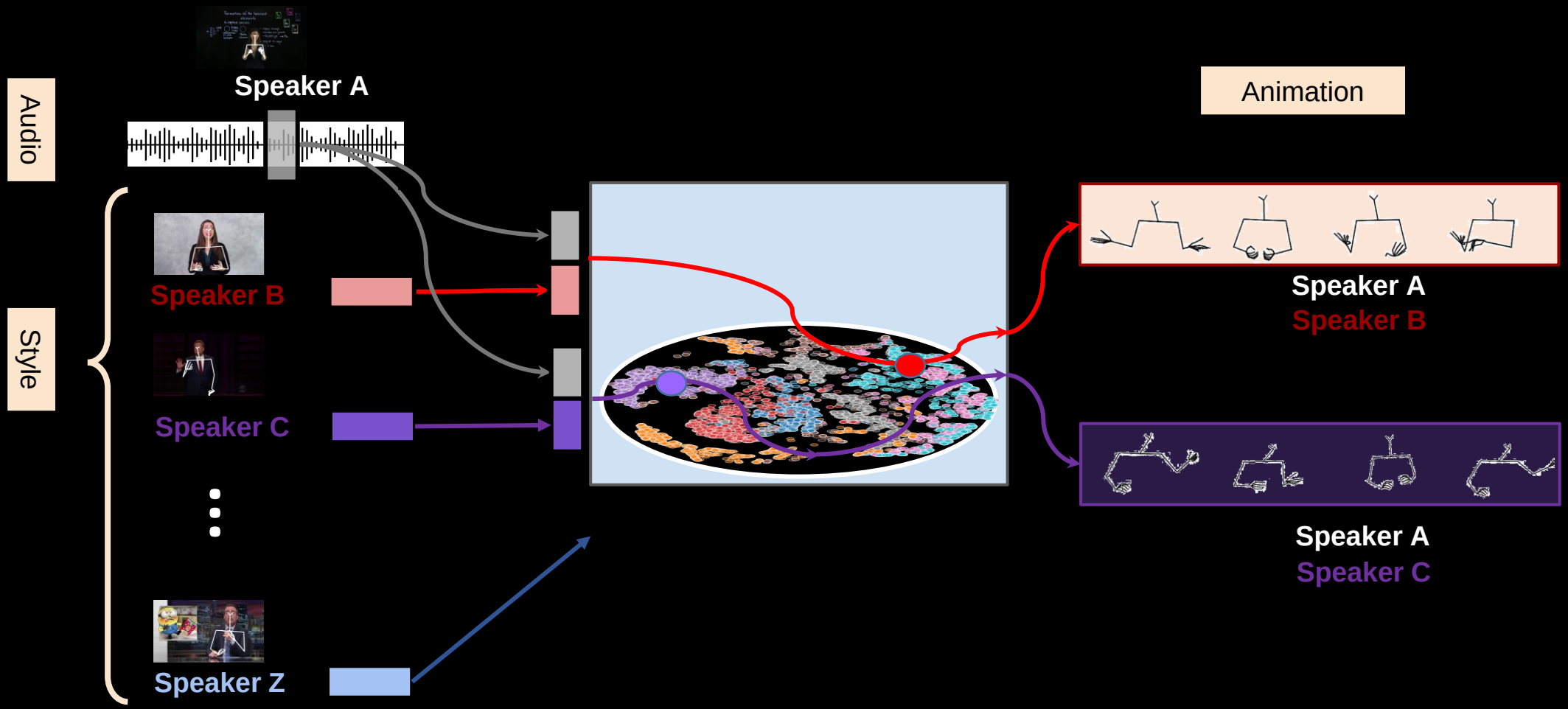
Stylized Co-speech gesture generation [Ahuja et al., 2020]



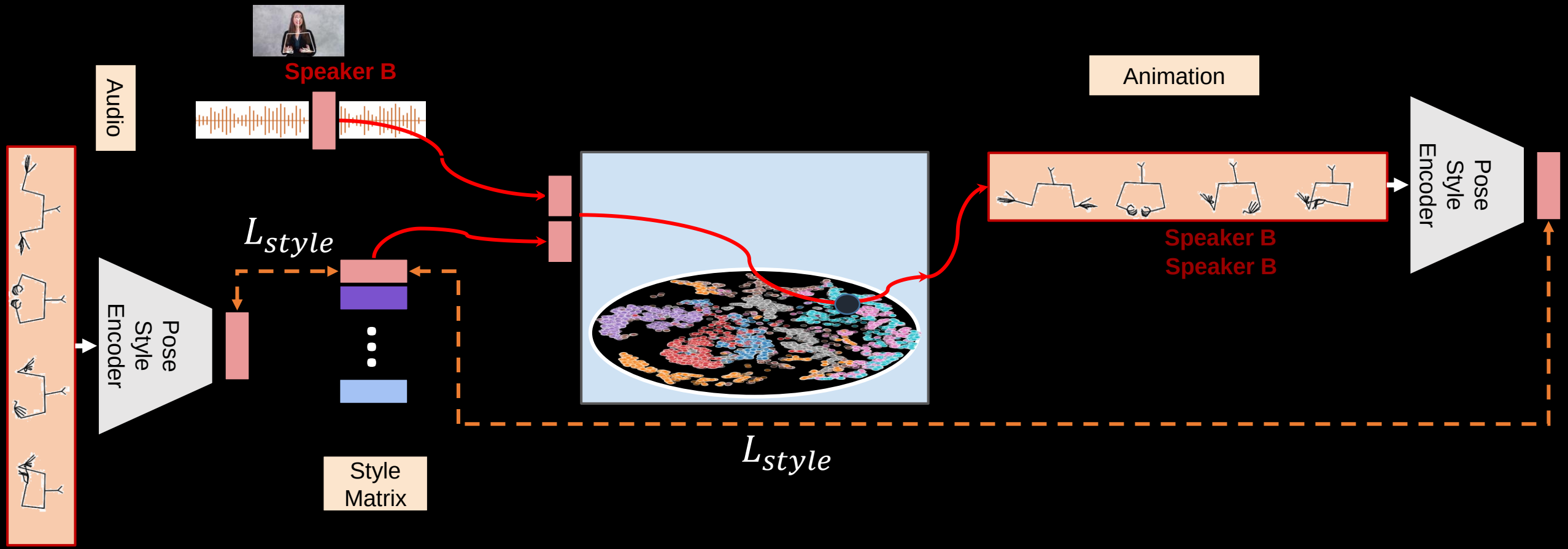
Stylized Co-speech gesture generation [Ahuja et al., 2020]



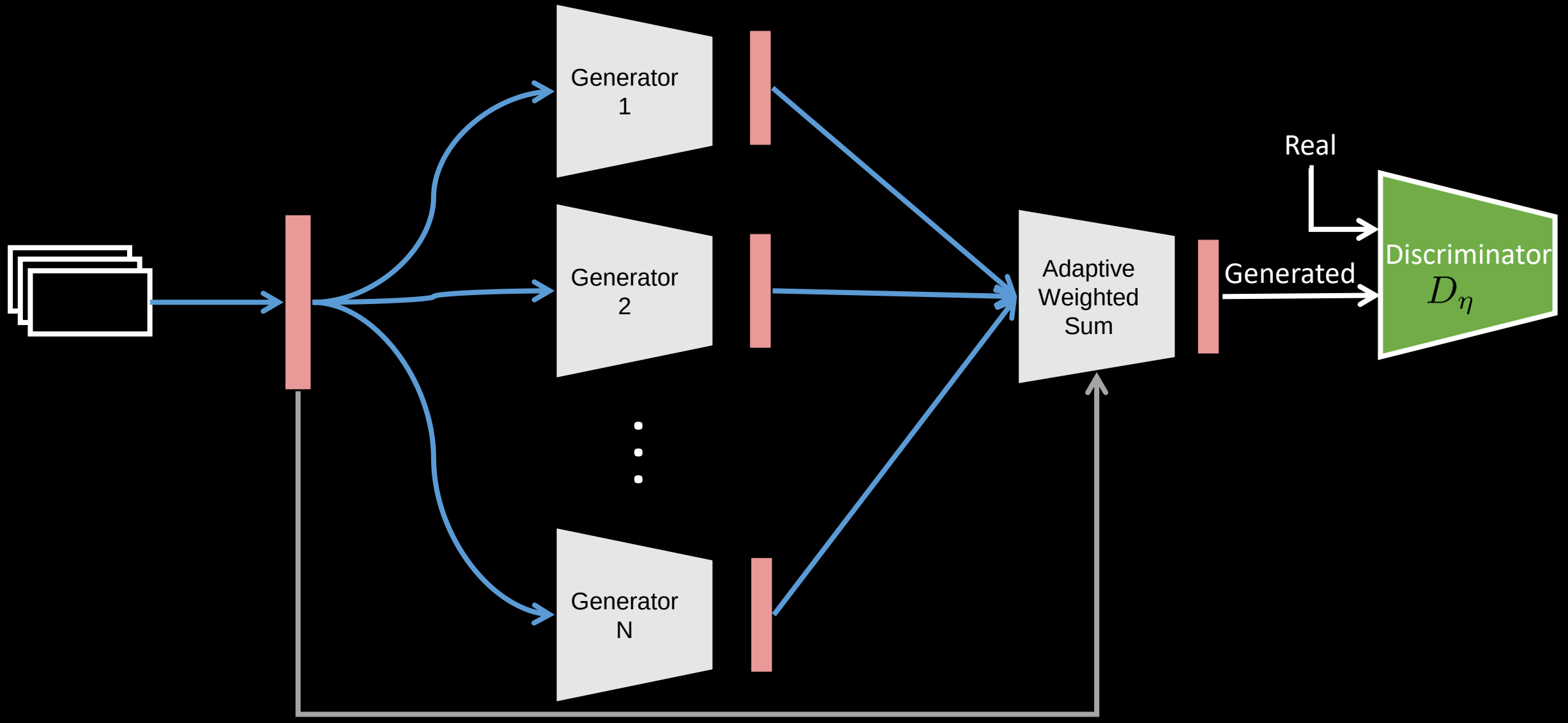
Stylized Co-speech gesture generation [Ahuja et al., 2020]



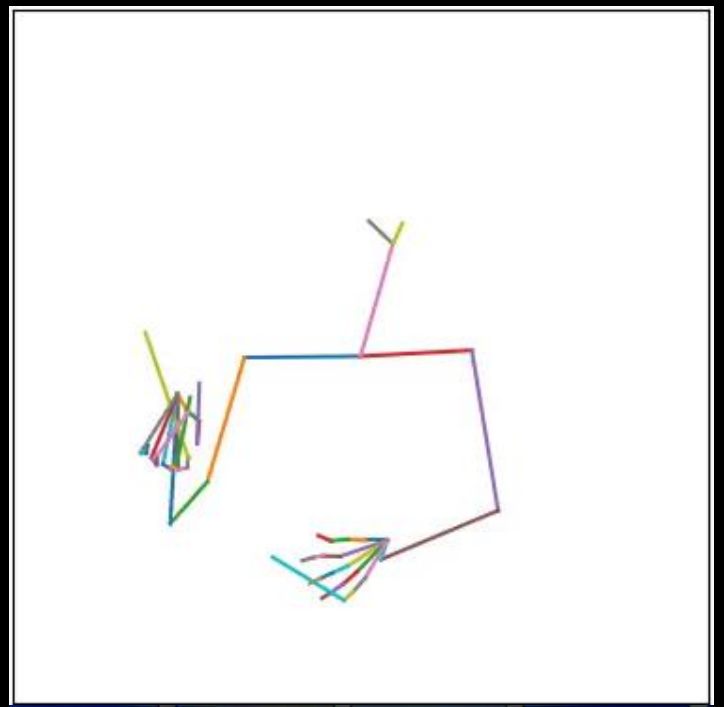
How do we learn the gesture space?



Pose Decoder: Conditional MixGAN

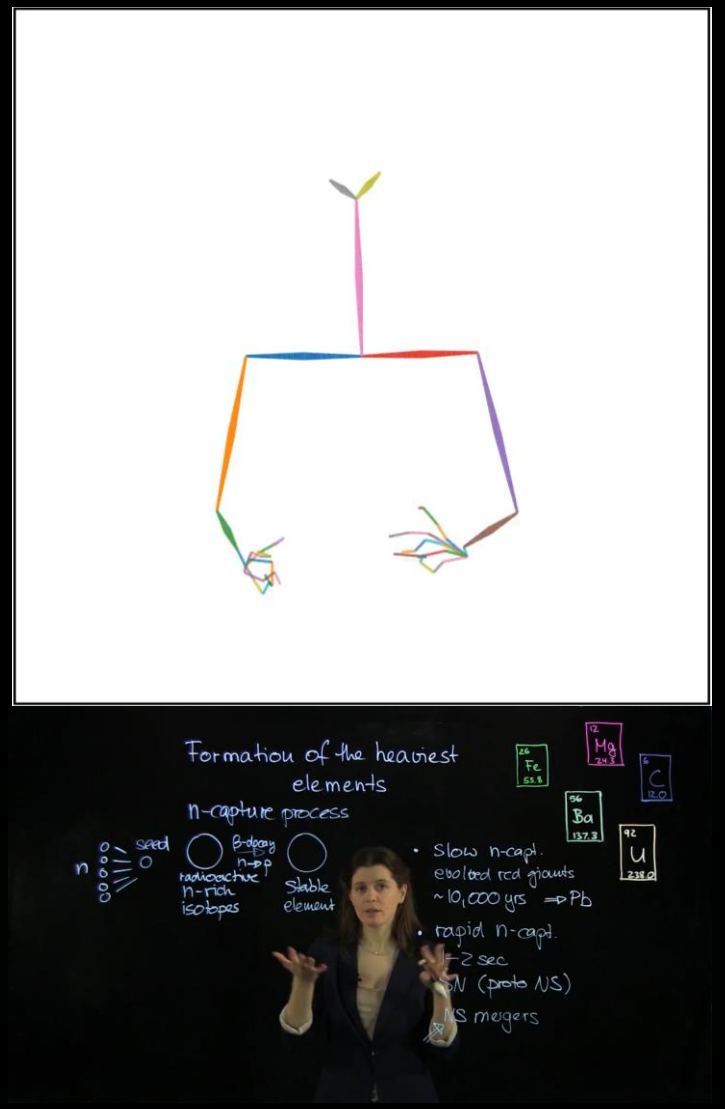


Original Animation

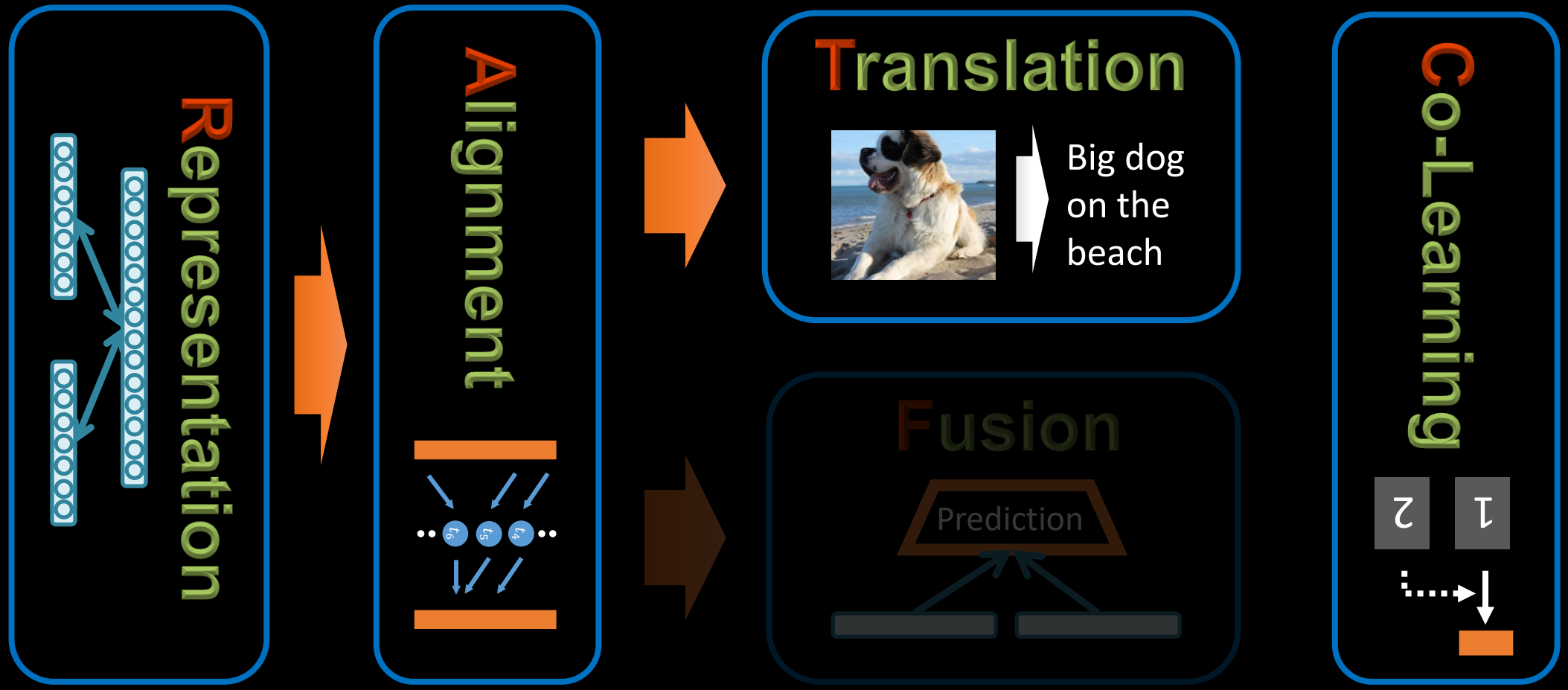


Style Transfer

Generated Animation



Multimodal AI – Core Challenges [Survey: TPAMI 2019]



Challenges for Real-World Multimodal AI

Core Challenges

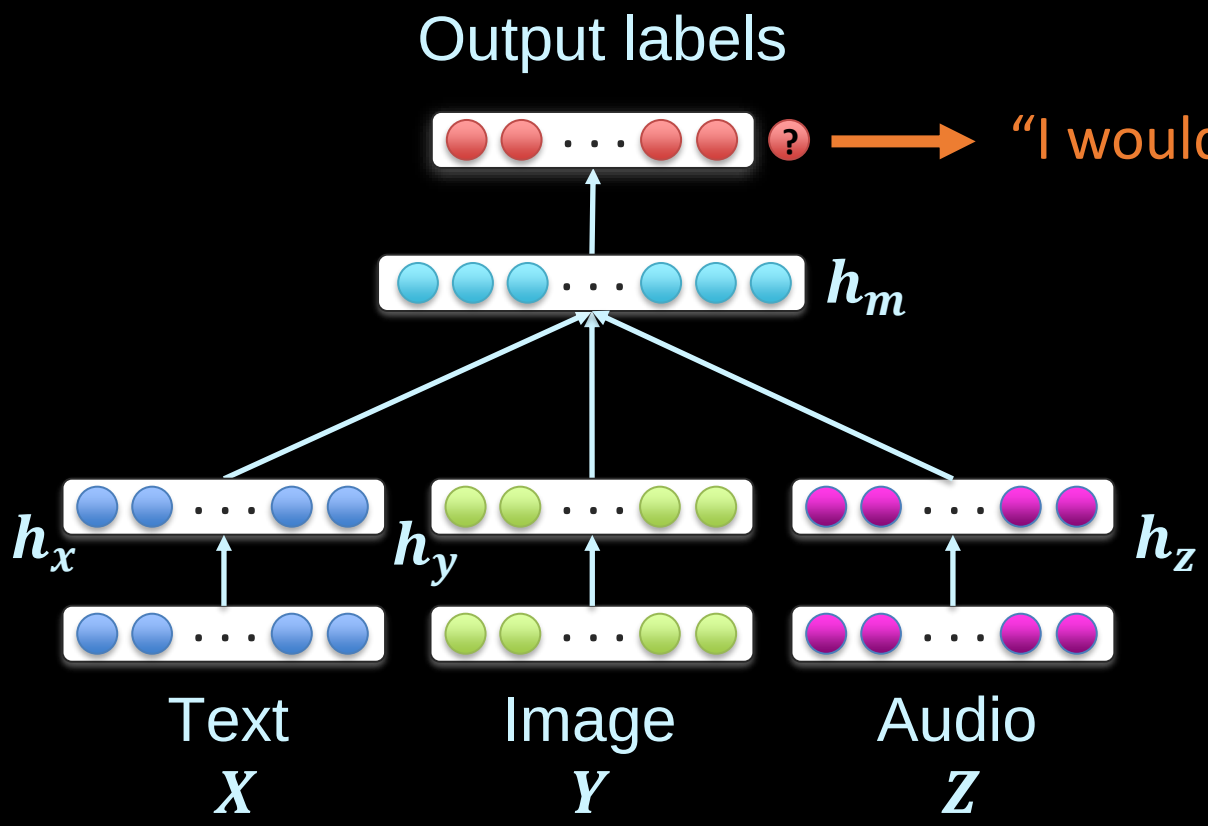
- Representation
- Alignment
- Fusion
- Translation
- Co-learning

Real-World Challenges

- Robustness
- Variability
- Thrustworthy
- Fairness
- Privacy

Thrustworthy

Deep Gambler: Learning to Abstain [Neurips, 2019]



Analogy: Horse race gambling
(portfolio theory)

Balance between:

1. betting for one of the labels when confident
2. Reserving one's winnings (abstaining) when not confident

Challenges for Real-World Multimodal AI

Core Challenges

- Representation
- Alignment
- Fusion
- Translation
- Co-learning

Real-World Challenges

- Robustness
- Variability
- Thrustworthy
- Fairness
- Privacy

Fairness

Toward Debiasing Sentence Representations [ACL 2020]

“The boy is coding.” OR “The girl is coding.”

“The boys at the playground.”
OR
“The girls at the playground.”

RESEARCH QUESTION: How to debias multimodal representations?

Challenges for Real-World Multimodal AI

Core Challenges

- Representation
- Alignment
- Fusion
- Translation
- Co-learning

+

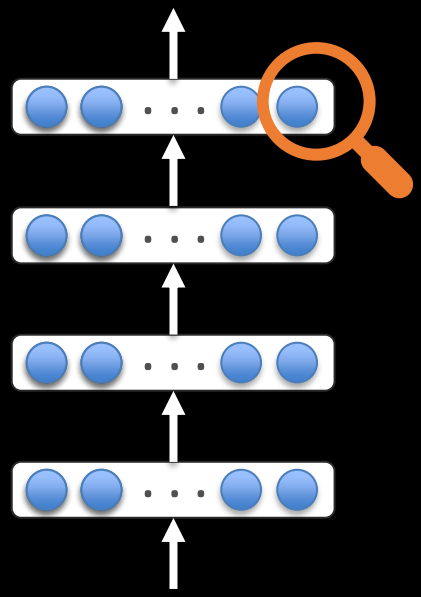
Real-World Challenges

- Robustness
- Variability
- Thrustworthy
- Fairness
- Privacy

Privacy

Privacy-Preserving ML [Neurips-W, 2020]

Mental health markers
(e.g., mood)

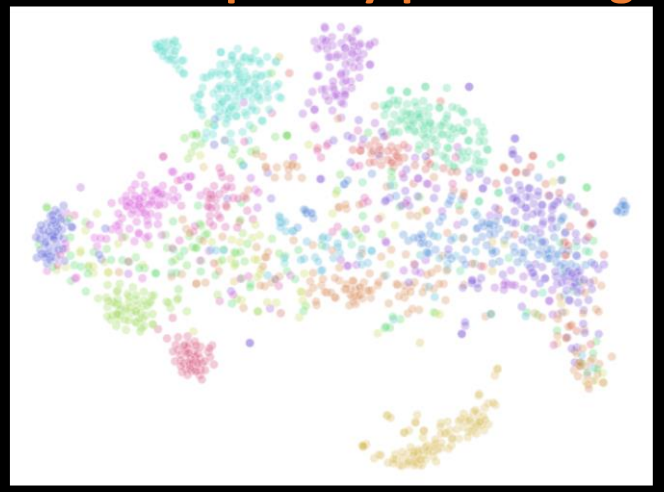


Mobile data

- Keystrokes
- Active apps
- GPS information

…

without privacy-preserving



with privacy-preserving

Each color represents a different user

→ User data is clustered

Using **Selective-Additive Learning** to hide identifiable information

Towards Real-World Multimodal AI

Core Challenges

- Representation
- Alignment
- Fusion
- Translation
- Co-learning

Real-World Challenges

- Robustness
- Variability
- Thrustworthy
- Fairness
- Privacy

MERCI !

http://multicomp.cs.cmu.edu/