Language Technologies Institute

Carnegie Mellon University

# Multimodal Machine Learning

## Lecture 4.2: Multimodal Alignment

Louis-Philippe Morency

** Original course co-developed with Tadas Baltrusaitis. Major revision co-developed with Paul Liang. Co-lecturers included Yonatan Bisk, Daniel Fried and Carlos Busso.*

# Administrative Stuff

# Upcoming Course Assignments

**Proposal, aka First Assignment** (due Sunday 9/20 at 8pm ET)

- Literature review + initial research ideas

**Week 5 Reading Assignment** (due Friday 9/11 at 8pm ET)

- Two papers (with a focus on multimodal alignment)

**Proposal presentations** (next week)

- Be sure to signup for a 15-mins timeslot!
- Designed to give feedback about your research ideas

**Second project assignment** (due Sunday 10/4 at 8pm ET)

- Look at your data, before modeling it! ☺

**No lectures next week!**

No lectures on Tuesday 9/22 and Thursday 9/24

# Attendance using Poll Everywhere

1. Scan QR Code
   *or*
   https://pe.app/mmml

2. Enter your AndrewID
   (as your “name tag”)

❌ *Do not try to sign-in*

3. Click ”Share location”

Language Technologies Institute

Carnegie Mellon University

# Multimodal Machine Learning

## Lecture 4.2: Multimodal Alignment

Louis-Philippe Morency

** Original course co-developed with Tadas Baltrusaitis. Major revision co-developed with Paul Liang. Co-lecturers included Yonatan Bisk, Daniel Fried and Carlos Busso.*

# Lecture objectives

- Discrete alignment
  - Global alignment
    - Assignment problem and optimal transport
- Continuous alignment
  - Continuous warping
    - Dynamic time warping
  - Discretization and segmentation
- Multimodal Alignment + Representation
  - Cross-modal and concatenated architectures

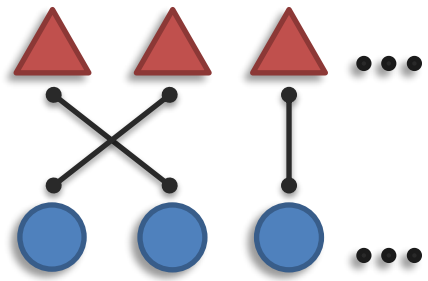
# Discrete Alignment

# Challenge 2: Alignment

**Definition:** Identifying and modeling cross-modal connections between all elements of multiple modalities, building from the data structure

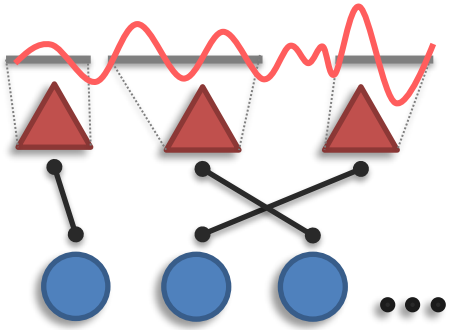
## Sub-challenges:

### Discrete Alignment



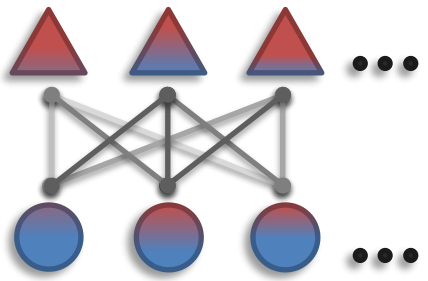
Discrete elements and connections

### Continuous Alignment



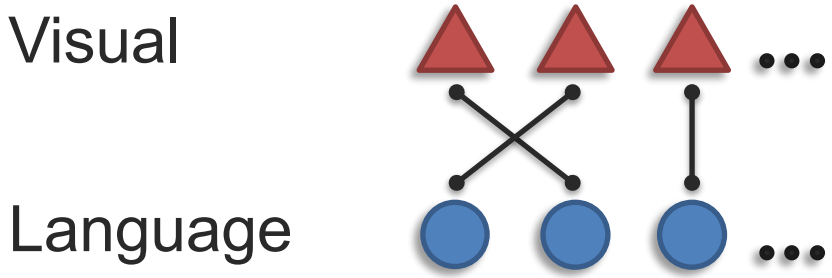
Segmentation and continuous warping

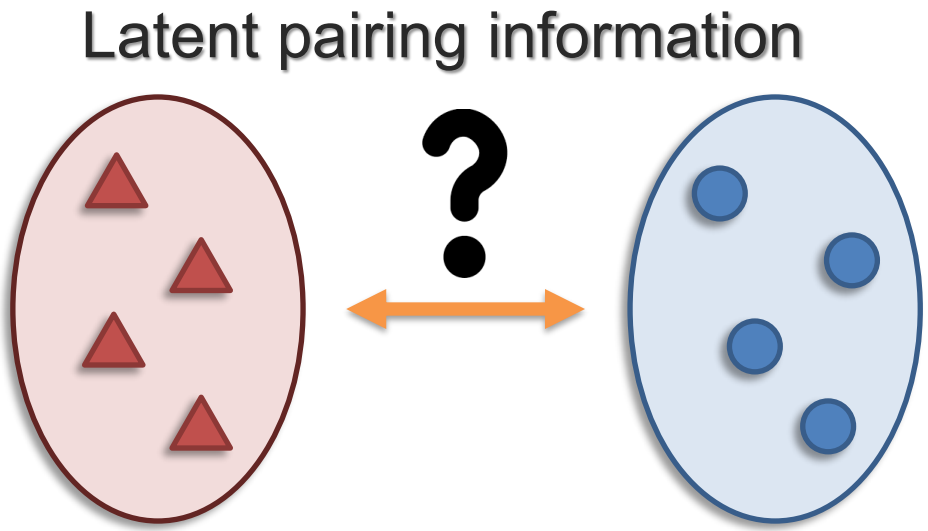
### Contextualized Representation



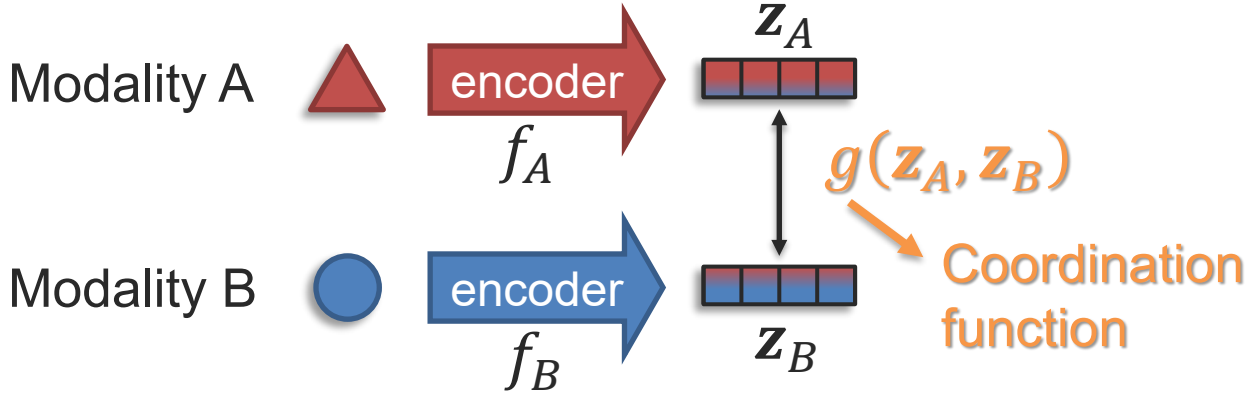
Alignment + representation

# Global Alignment

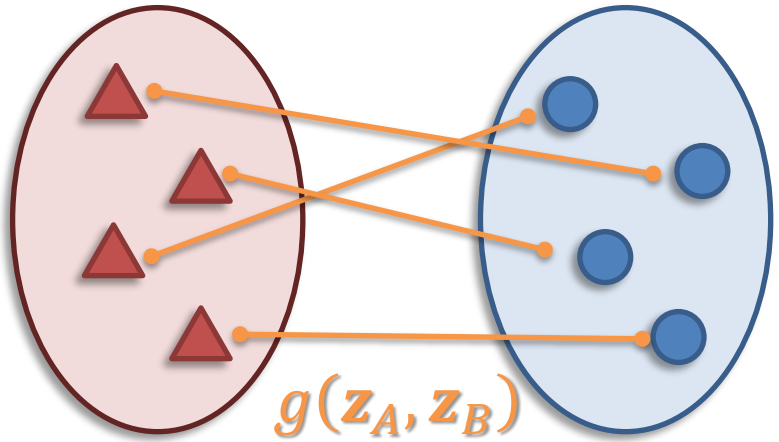



Jointly optimize representation + global alignment:

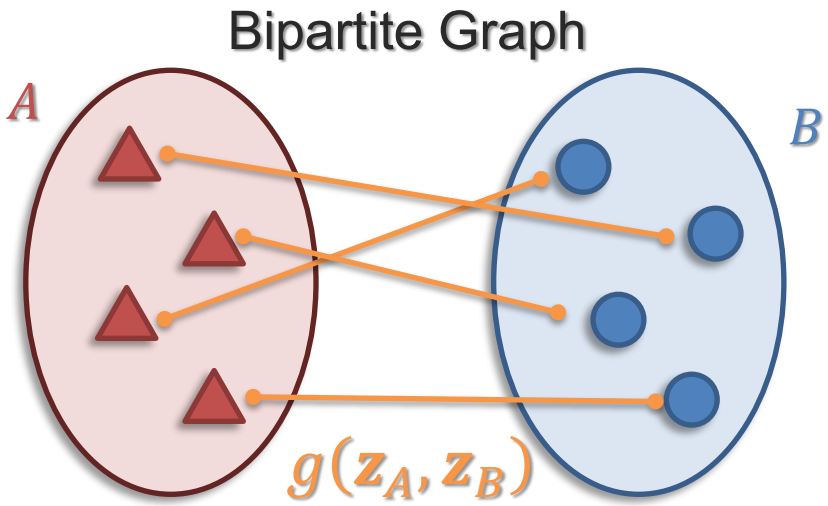



(representation)




(global alignment)

# Assignment Problem




**Initial assumptions:**

- Same number of elements in A and B modalities
- 1-to-1 "hard" alignment between elements
- All elements assigned (aka "perfect matching")

➡ How to solve?

Naive solution: check all assignments

Better solution: Linear Programming

Assignment: (vector of indices)

~~$f: A \rightarrow B$~~

$x_{ij} = 1$ when matching connection, otherwise 0

Similarity weights:

~~$w_{(i,f(i))} = g(\boldsymbol{z}_A^i, \boldsymbol{z}_B^{f(i)})$~~

$w_{(i,j)} = g(\boldsymbol{z}_A^i, \boldsymbol{z}_B^j)$

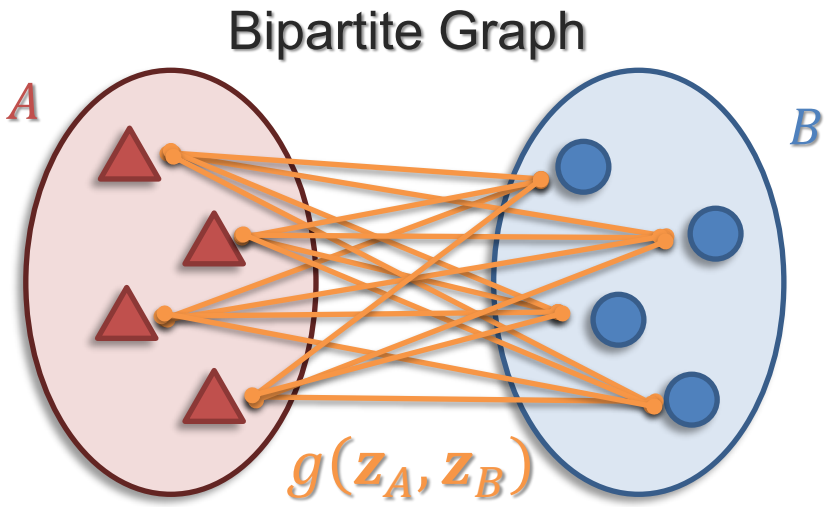
Maximize:

~~$\max_{f \in \mathrm{Perm}(N)} \sum_{i=1}^{N} w_{i,f(i)}$~~

$\max_{\{x_{ij}\}} \sum_{(i,j) \in A \times B} w_{i,j} x_{ij}$

➡ Can be solved with simplex algorithm

# Optimal transport




**New assumptions:**

- Different number of elements in A and B modalities
- Many-to-many "soft" alignment between elements

➡ It can be seen as "transporting" elements from modality A to modality B (and vice-versa)

Assignments: $x_{(i,j)}$: soft alignment between $\boldsymbol{z}_A^i$ and $\boldsymbol{z}_B^j$

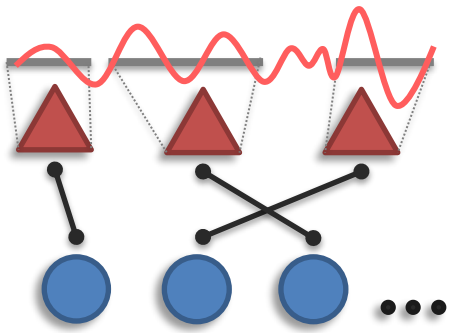
Similarity weights: $w_{(i,j)} = g(\boldsymbol{z}_A^i, \boldsymbol{z}_B^j)$

Maximize: $$\max_{\{x_{ij}\}} \sum_{(i,j)\in A\times B} w_{i,j} x_{ij}$$

➡ Wassertein distance gives optimal transport

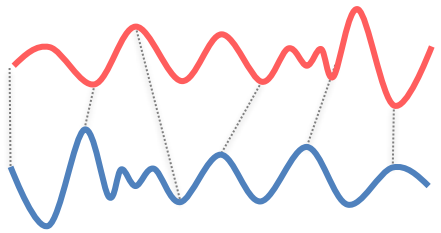
# Continuous Alignment

# Challenge 2b: Continuous Alignment

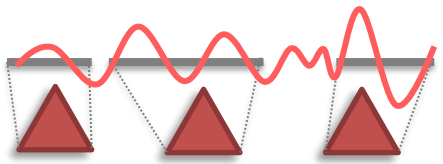


**Definition:** Model alignment between modalities with continuous signals and no explicit elements

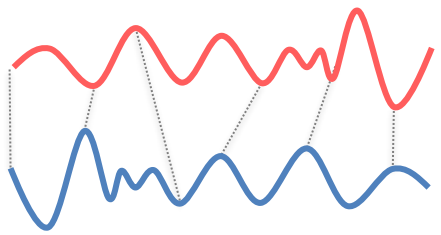
Continuous warping



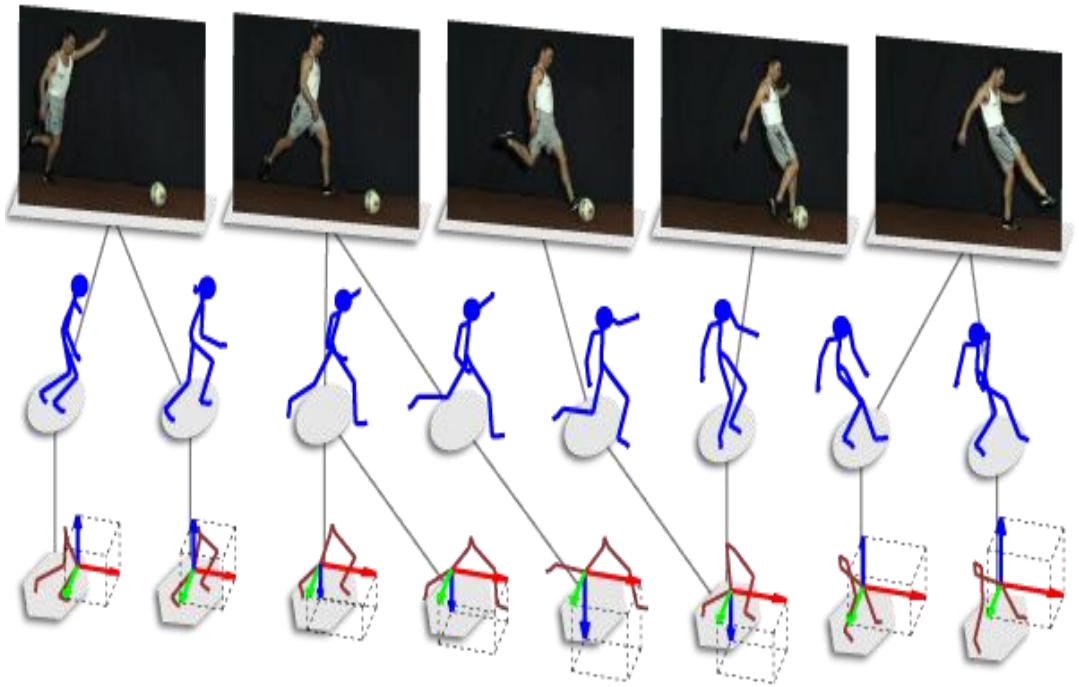
Discretization
(segmentation)



# Continuous Warping – Example



Aligning video sequences



# Dynamic Time Warping (DTW)

We have two unaligned temporal unimodal signals

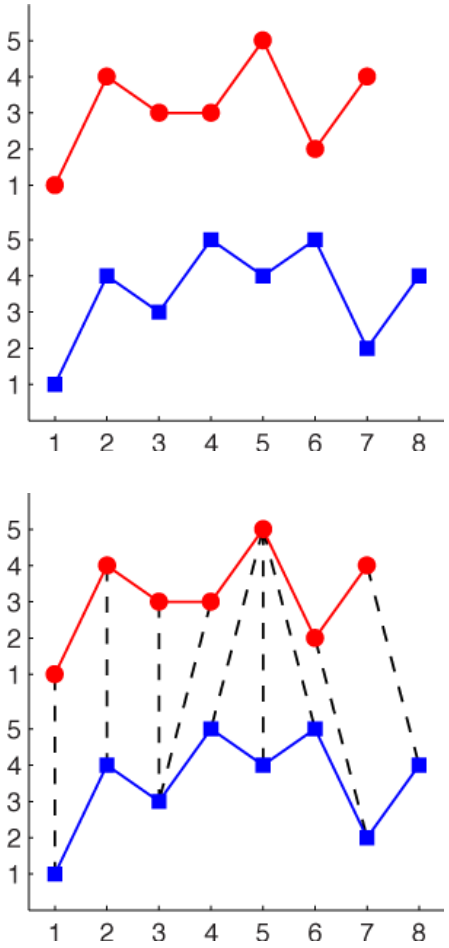
- $\mathbf{X} = [\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_{n_x}] \in \mathbb{R}^{d \times n_x}$
- $\mathbf{Y} = [\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_{n_y}] \in \mathbb{R}^{d \times n_y}$

Find set of indices to minimize the alignment difference:

$$L(\boldsymbol{p}^x, \boldsymbol{p}^y) = \sum_{t=1}^{l} \left\| \boldsymbol{x}_{\boldsymbol{p}_t^x} - \boldsymbol{y}_{\boldsymbol{p}_t^y} \right\|_2^2$$

where $\boldsymbol{p}^x$ and $\boldsymbol{p}^y$ are index vectors of same length

Dynamic Time Warping is designed to find these index vectors!

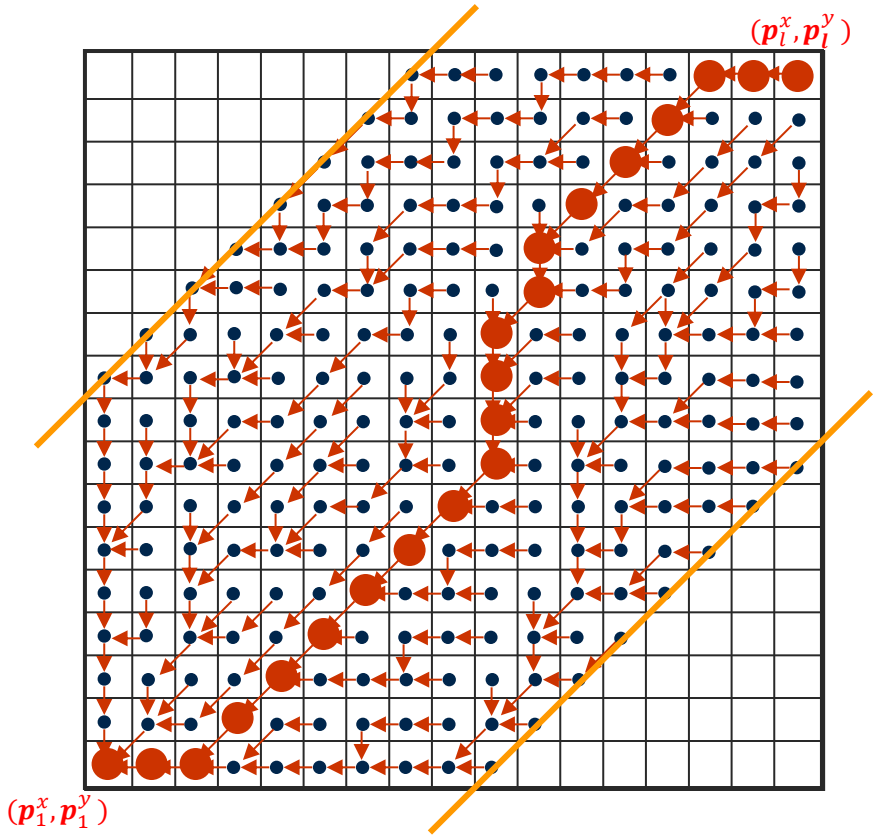


# Dynamic Time Warping (DTW)

Lowest cost path in a cost matrix

- Restrictions?
  - Monotonicity – no going back in time
  - Continuity - no gaps
  - Boundary conditions - start and end at the same points
  - Warping window - don't get too far from diagonal
  - Slope constraint – do not insert or skip too much

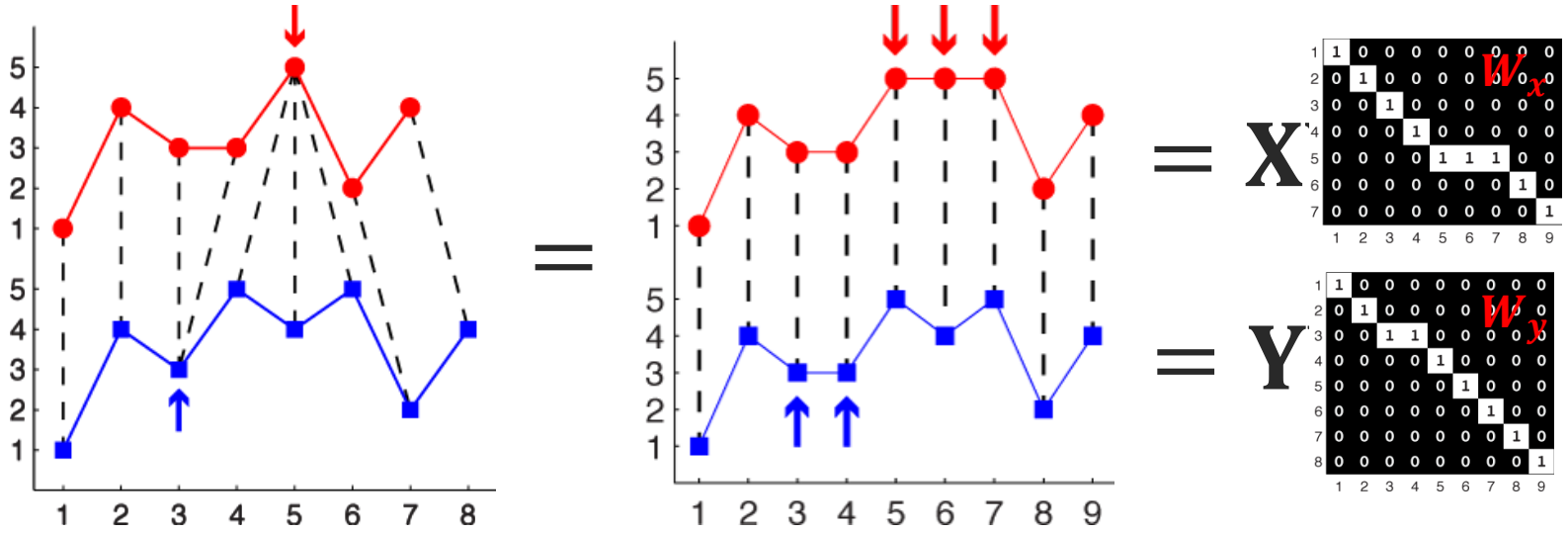
Solved using dynamic programming while respecting the restrictions



# DTW alternative formulation

$$L(\boldsymbol{p}^x, \boldsymbol{p}^y) = \sum_{t=1}^{l} \left\| \boldsymbol{x}_{p_t^x} - \boldsymbol{y}_{p_t^y} \right\|_2^2$$

Replication doesn't change the objective!




Alternative objective:

$$L(\boldsymbol{W}_x, \boldsymbol{W}_y) = \left\| \boldsymbol{X}\boldsymbol{W}_x - \boldsymbol{Y}\boldsymbol{W}_y \right\|_F^2$$

Frobenius norm $\|\boldsymbol{A}\|_F^2 = \sum_i \sum_j |a_{i,j}|^2$

$\boldsymbol{X}, \boldsymbol{Y}$ – original signals (same #rows, possibly different #columns)

$\boldsymbol{W}_x, \boldsymbol{W}_y$ - alignment matrices

A differentiable version of DTW also exists…
https://arxiv.org/pdf/1703.01541.pdf

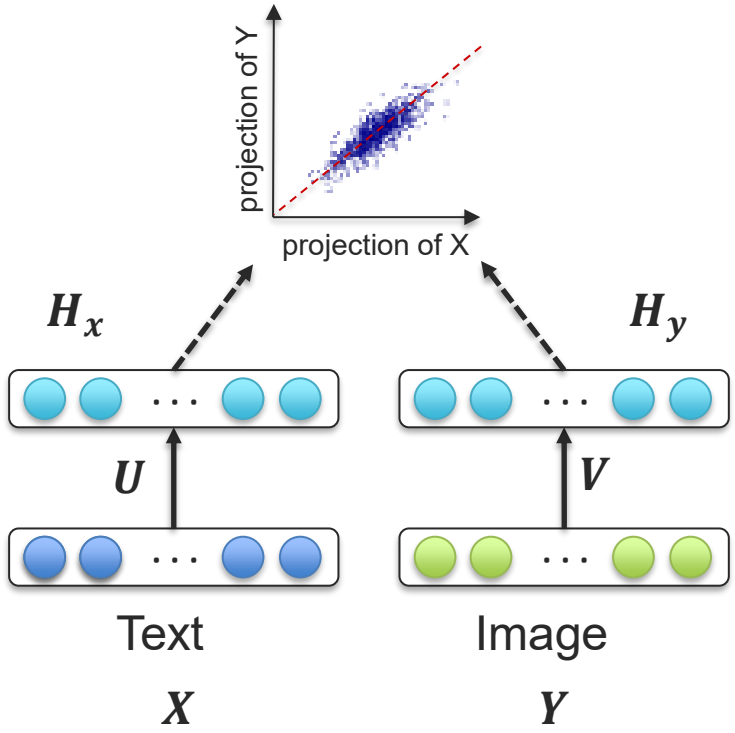
# Canonical Correlation Analysis – Reminder

CCA loss can also be re-written as:

$$L(\boldsymbol{U}, \boldsymbol{V}) = \|\mathbf{U}^T\mathbf{X} - \mathbf{V}^T\mathbf{Y}\|_F^2$$

subject to:

$$\boldsymbol{U^T\Sigma_{YY}U} = \boldsymbol{V^T\Sigma_{YY}V} = \boldsymbol{I},\ \boldsymbol{u_{(j)}^T\Sigma_{XY}v_{(i)}} = \mathbf{0}$$



# Canonical Time Warping

Dynamic Time Warping + Canonical Correlation Analysis = Canonical Time Warping

$$L(\boldsymbol{U}, \boldsymbol{V}, \boldsymbol{W_x}, \boldsymbol{W_y}) = \left\| \mathbf{U}^T \mathbf{X} \mathbf{W_x} - \mathbf{V}^T \mathbf{Y} \mathbf{W_y} \right\|_F^2$$

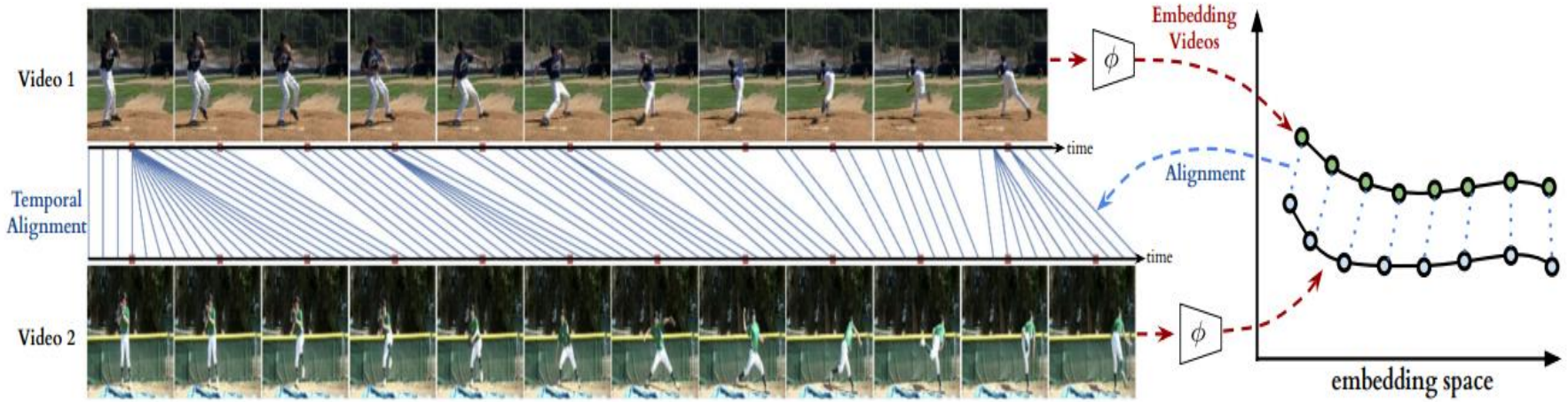
Allows to align multi-modal or multi-view (same modality but from a different point of view)

- $\boldsymbol{W_x}, \boldsymbol{W_y}$ – temporal alignment
- $\boldsymbol{U}, \boldsymbol{V}$ – cross-modal (spatial) alignment

[Canonical Time Warping for Alignment of Human Behavior, Zhou and De la Tore, 2009]

# Temporal Alignment and Neural Representation Learning

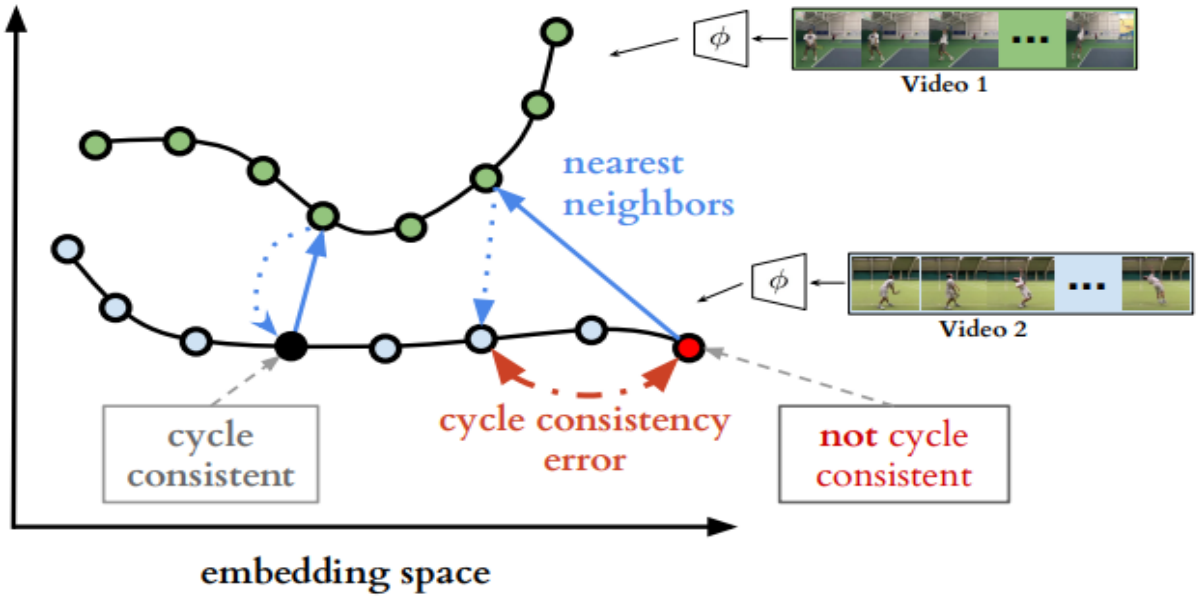
**Premise:** we have paired video sequences that can be be temporally aligned




How can we define a loss function to enforce the alignment between sequences while at the same time learning good representations?

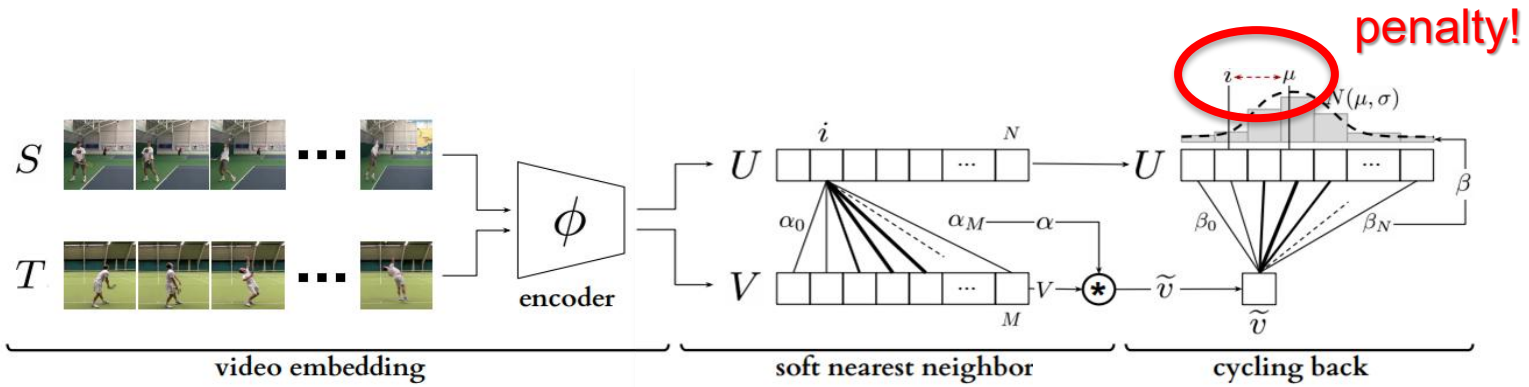
# Temporal Cycle-Consistency Learning

Solution: Representation learning by enforcing **Cycle consistency**




**Main idea:** My closest neighbor also views me as their closest neighbor

# Temporal Cycle-Consistency Learning




Compute “soft” / “weighted” nearest neighbour:

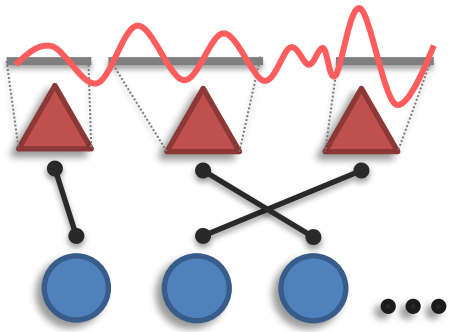
distances: $\alpha_j = \frac{e^{-||u_i - v_j||^2}}{\sum_k^M e^{-||u_i - v_k||^2}}$ Soft nearest neighbor: $\widetilde{v} = \sum_j^M \alpha_j v_j,$

Find the nearest neighbor the other way and then penalize the distance:

$$\beta_k = \frac{e^{-||\widetilde{v} - u_k||^2}}{\sum_j^N e^{-||\widetilde{v} - u_j||^2}}$$

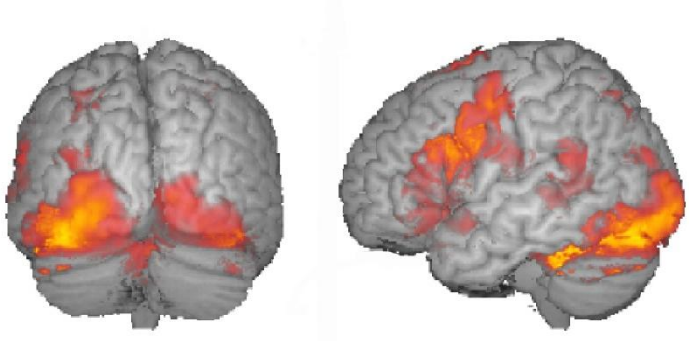
$$L_{cbr} = \frac{|i - \mu|^2}{\sigma^2} + \lambda \log(\sigma)$$

# Discretization (aka Segmentation)

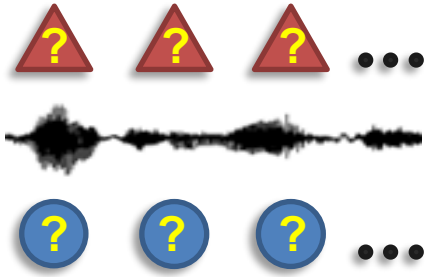


**Common assumptions:** ① Segmented elements

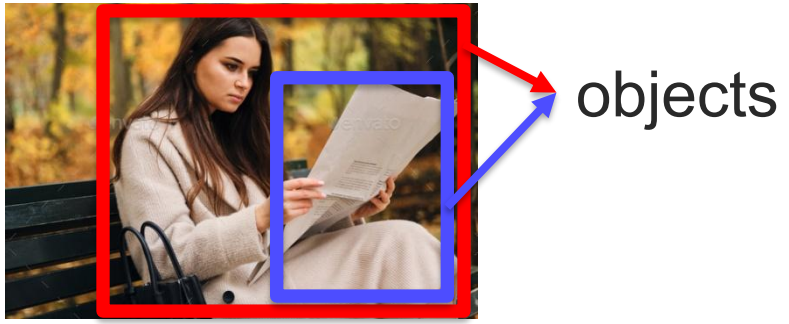
Examples:



Medical imaging



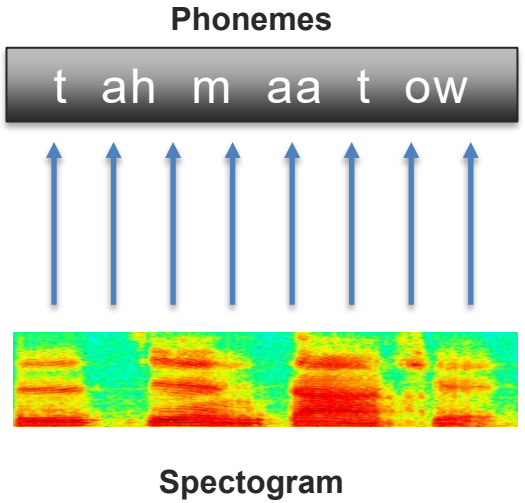
Signals




Images

# Discretization – Example

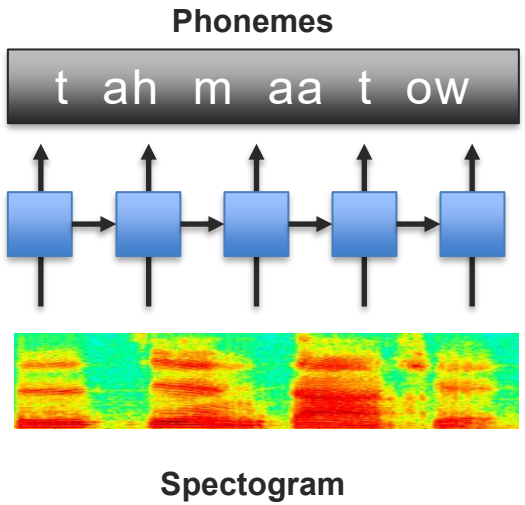
## Sequence Labeling and Alignment



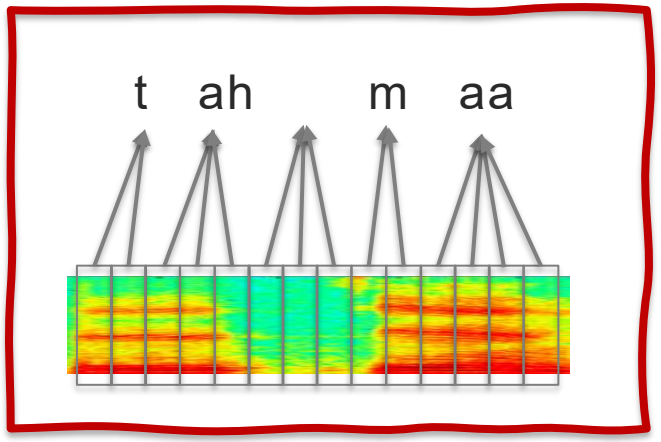

**How can we predict the sequence of phoneme labels?**

# Discretization – Example

## Sequence Labeling and Alignment




**Challenge: many-to-1 alignment**




**How can we predict the sequence of phoneme labels?**

# Discretization – A Classification Approach

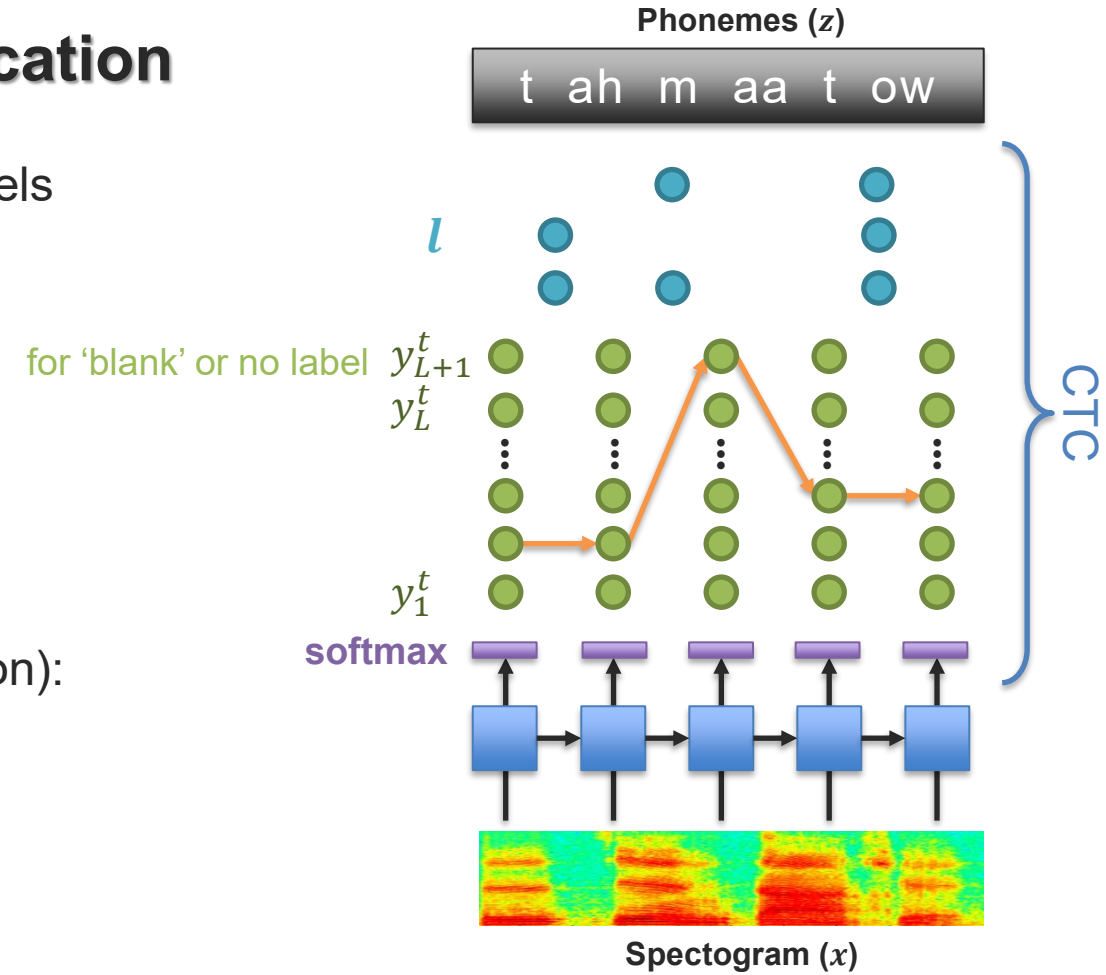
## Connectionist Temporal Classification

4. Most probable sequence labels

3. Predicted labels $l$

2. Path $\pi$ over the activations:

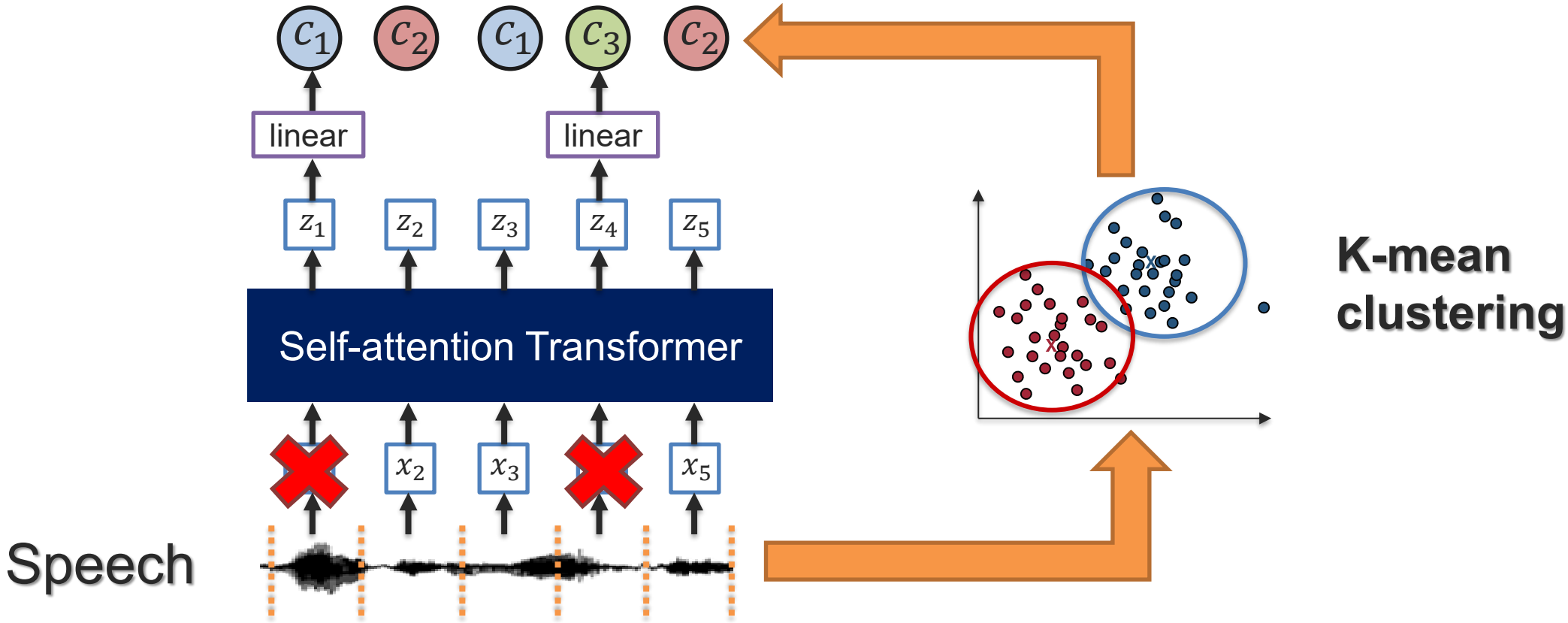
1. Output activations (distribution):




Grave et al., Connectionist Temporal Classification: Labelling Unsegmented Sequence Data with Recurrent Neural Networks, ICML 2006

# Discretization and Representation – Cluster-based Approaches

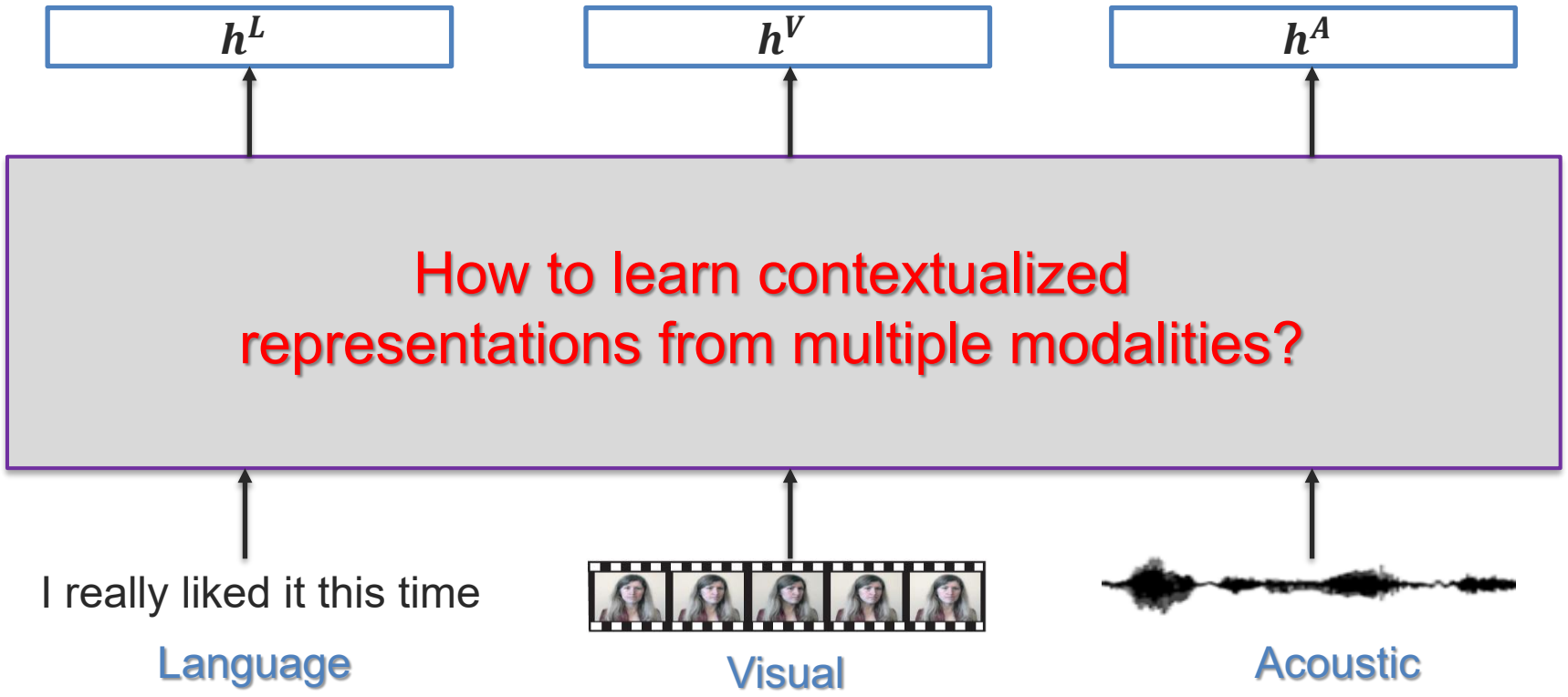
## HUBERT: Hidden-Unit BERT




Hsu et al., HuBERT: Self-Supervised Speech Representation Learning by Masked Prediction of Hidden Units, arxiv 2021

# Alignment + Representation

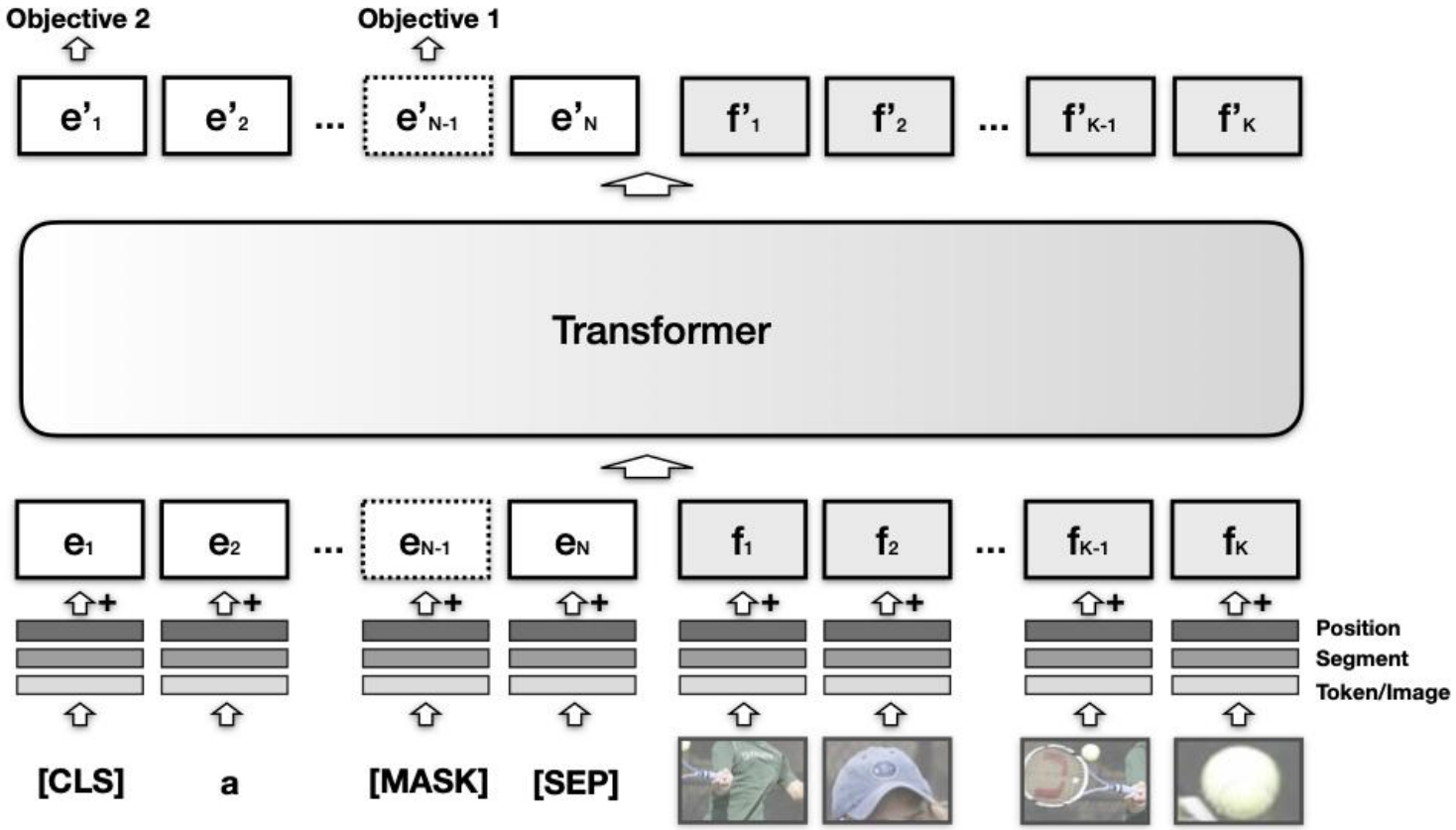
# Multimodal Alignment and Representation




Option 1: Concatenate modalities and learn BERT transformer

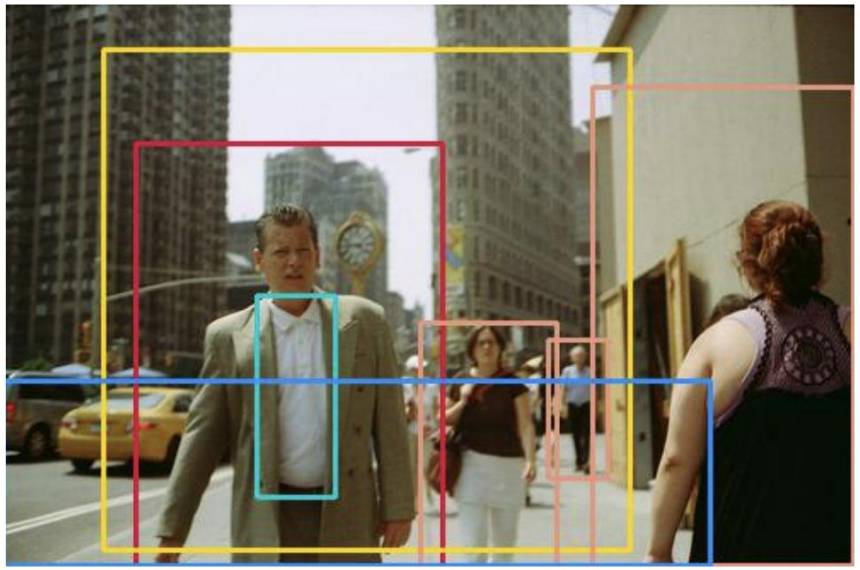
# VisualBERT

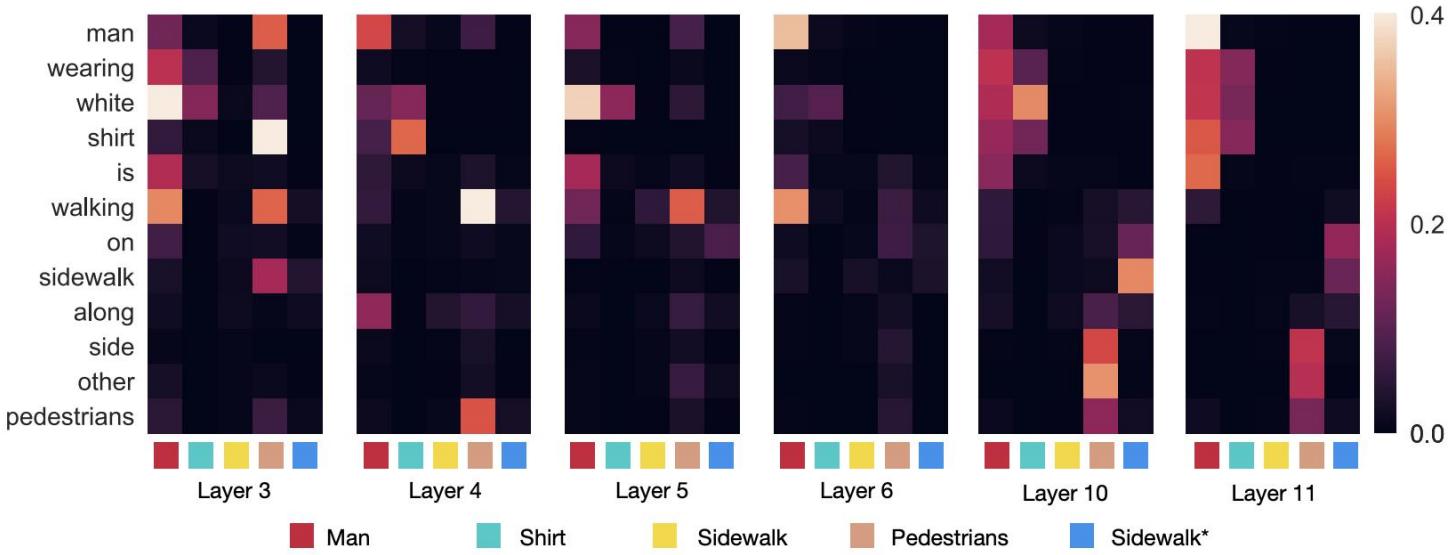
A person hits a ball with a tennis racket




Li, Liunian Harold, et al. "Visualbert: A simple and performant baseline for vision and language." *ACL* (2020).

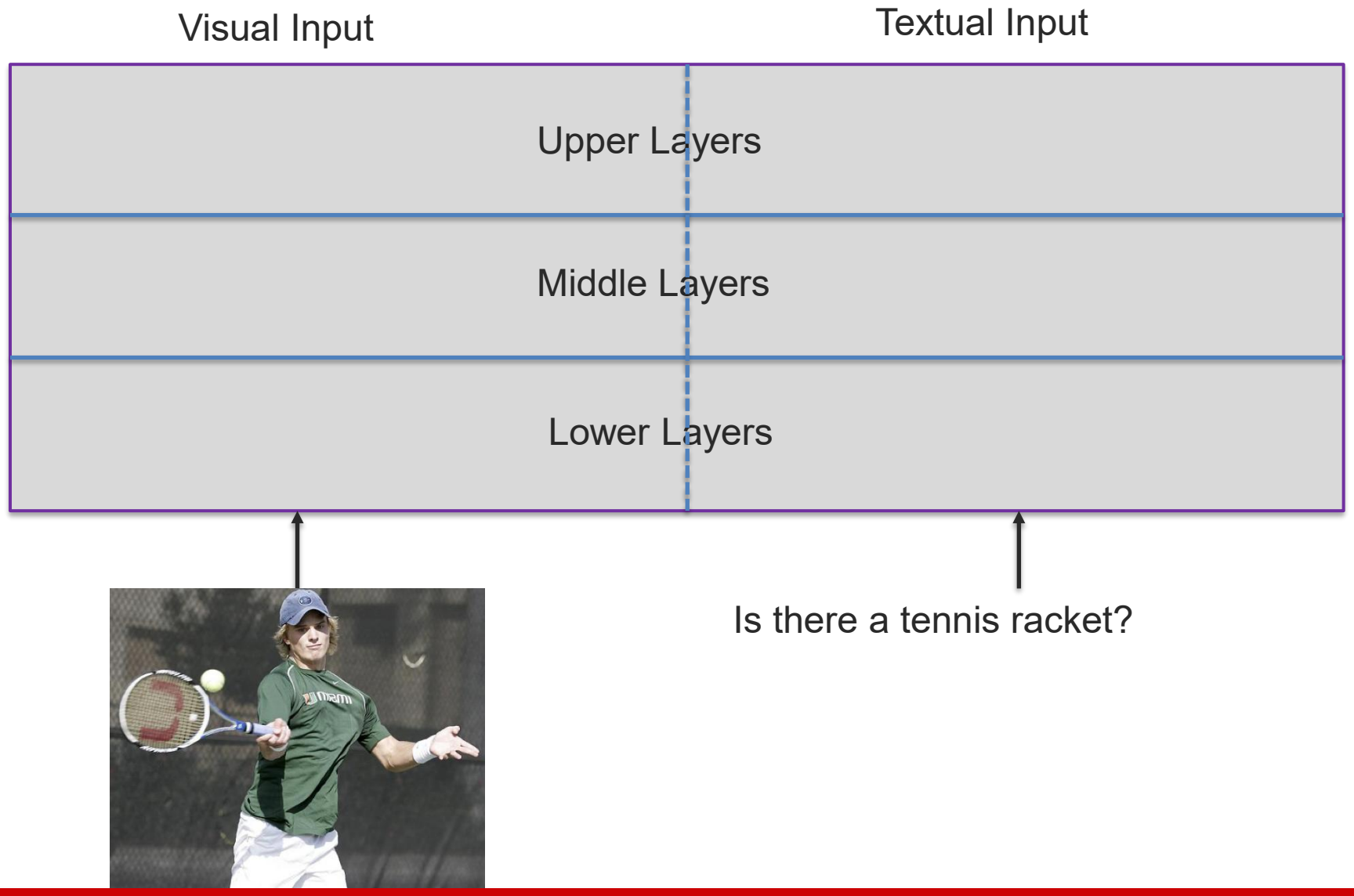
# VisualBERT



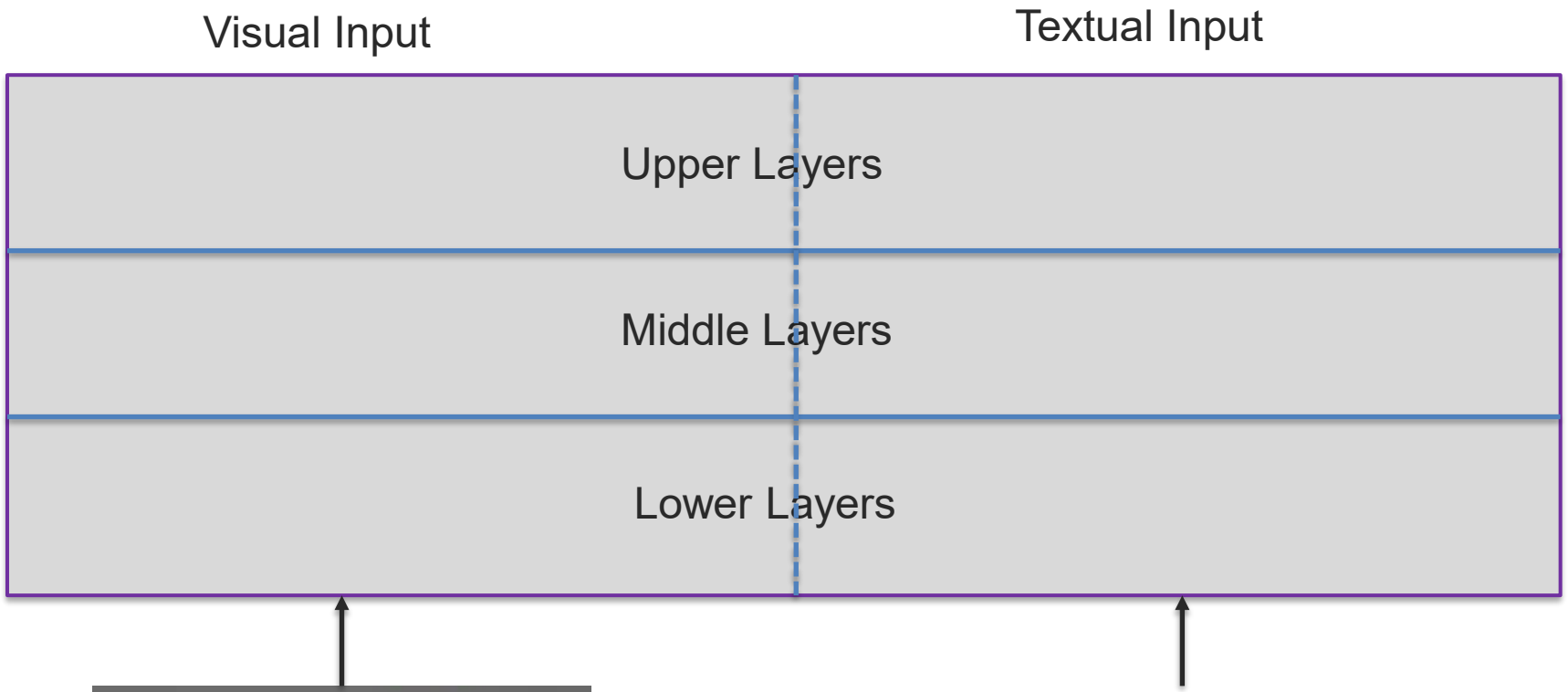



Li, Liunian Harold, et al. "Visualbert: A simple and performant baseline for vision and language." *ACL* (2020).

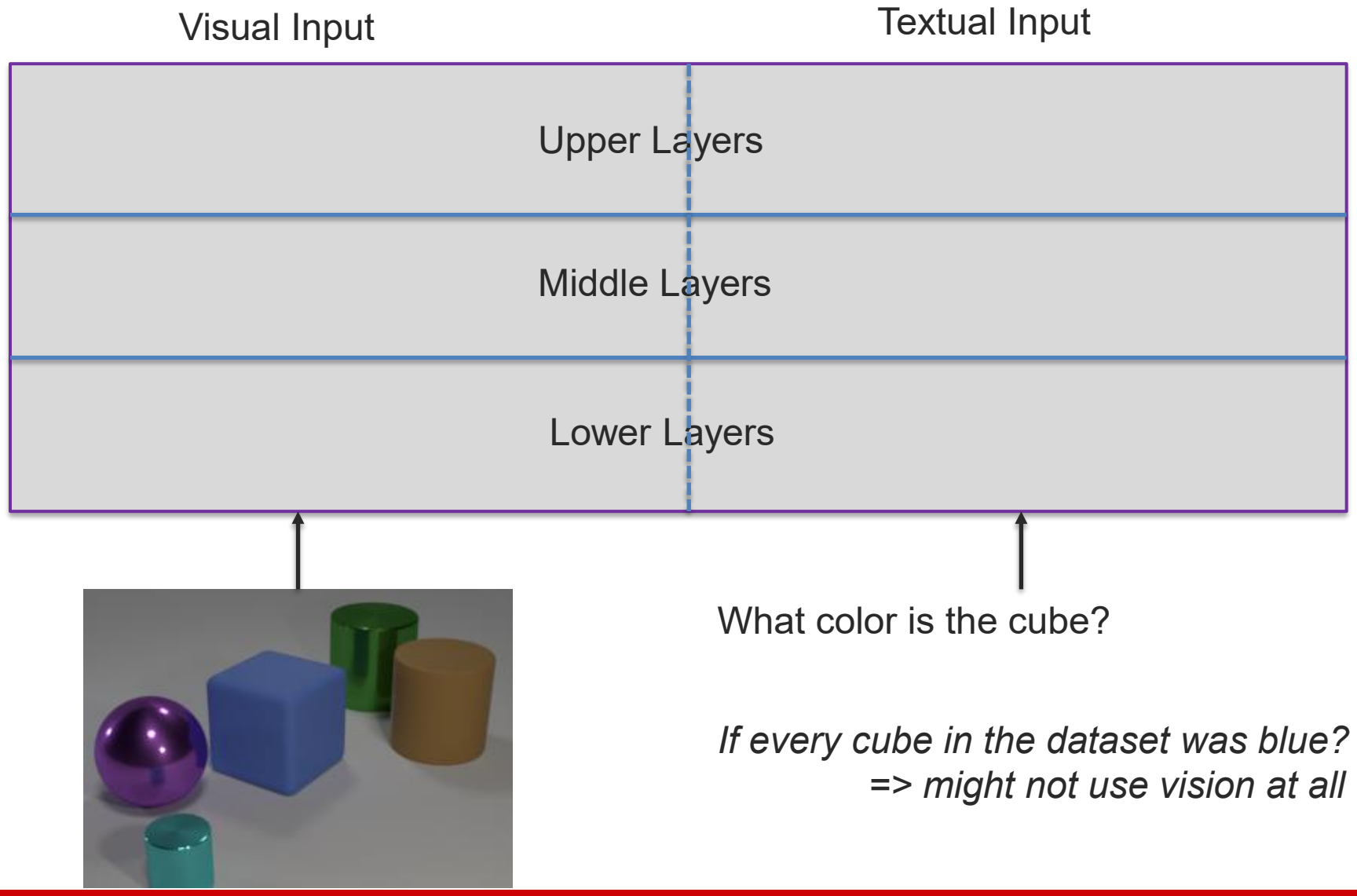
# Where Do Cross-Modal Interactions Happen?



# Where Do Cross-Modal Interactions Happen?

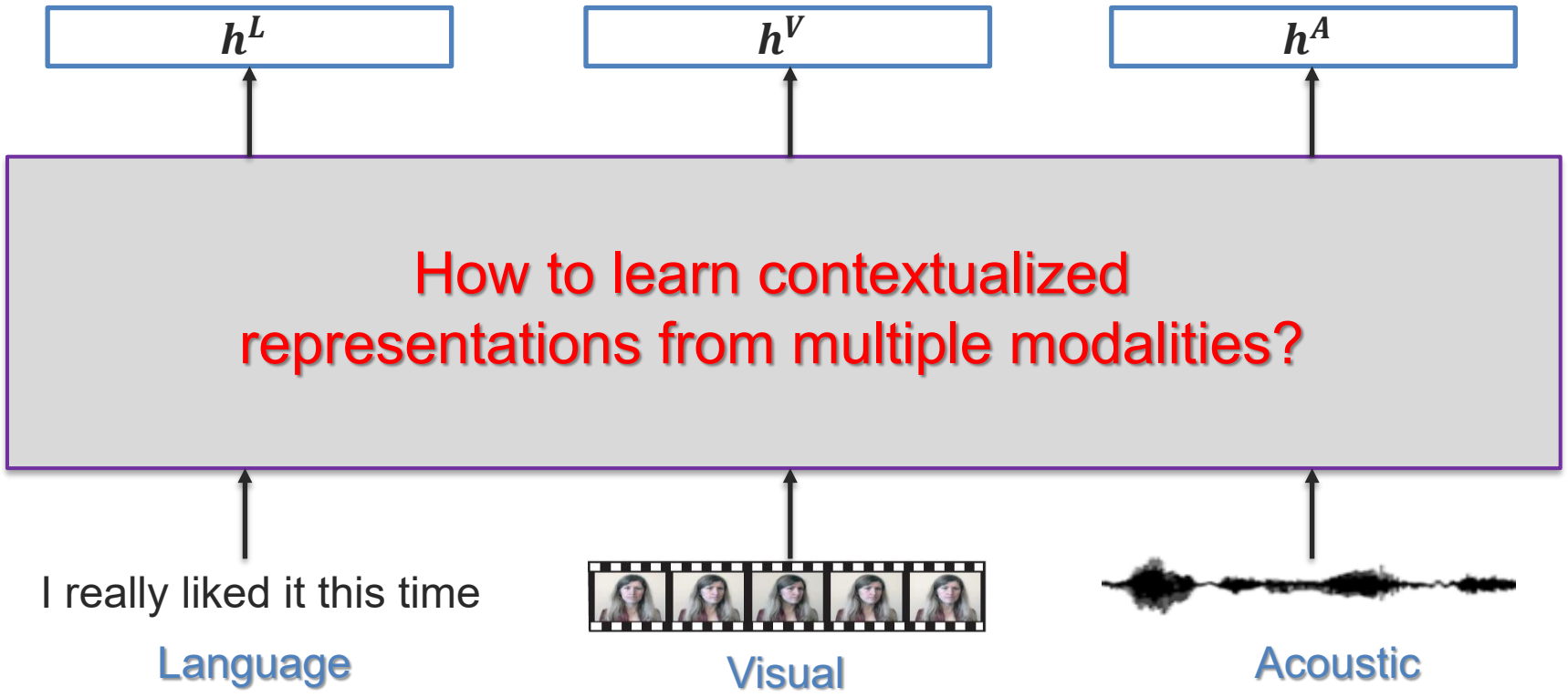


# Do Cross-Modal Interactions Happen?



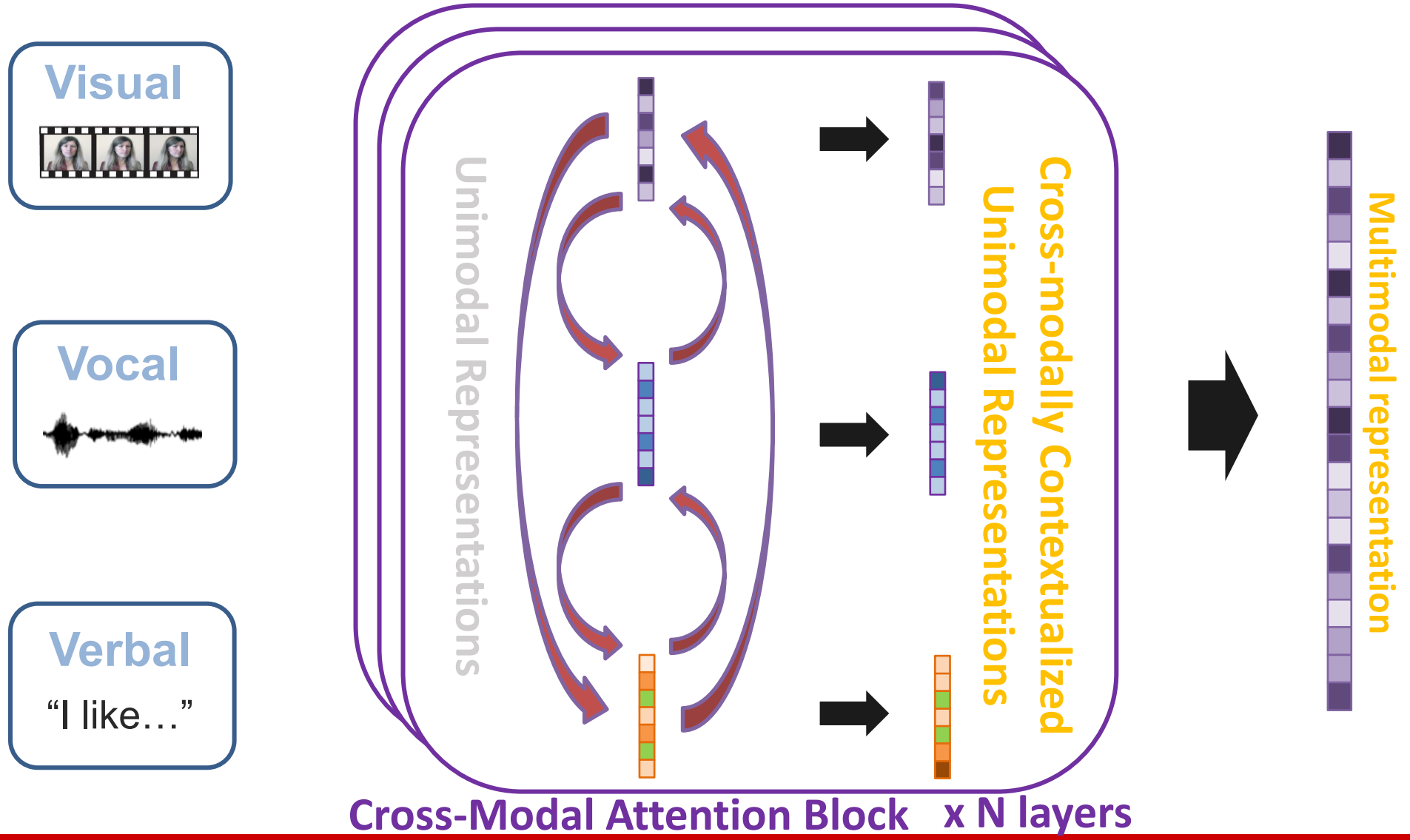

*If every cube in the dataset was blue?*
*=> might not use vision at all*

# Multimodal Embeddings

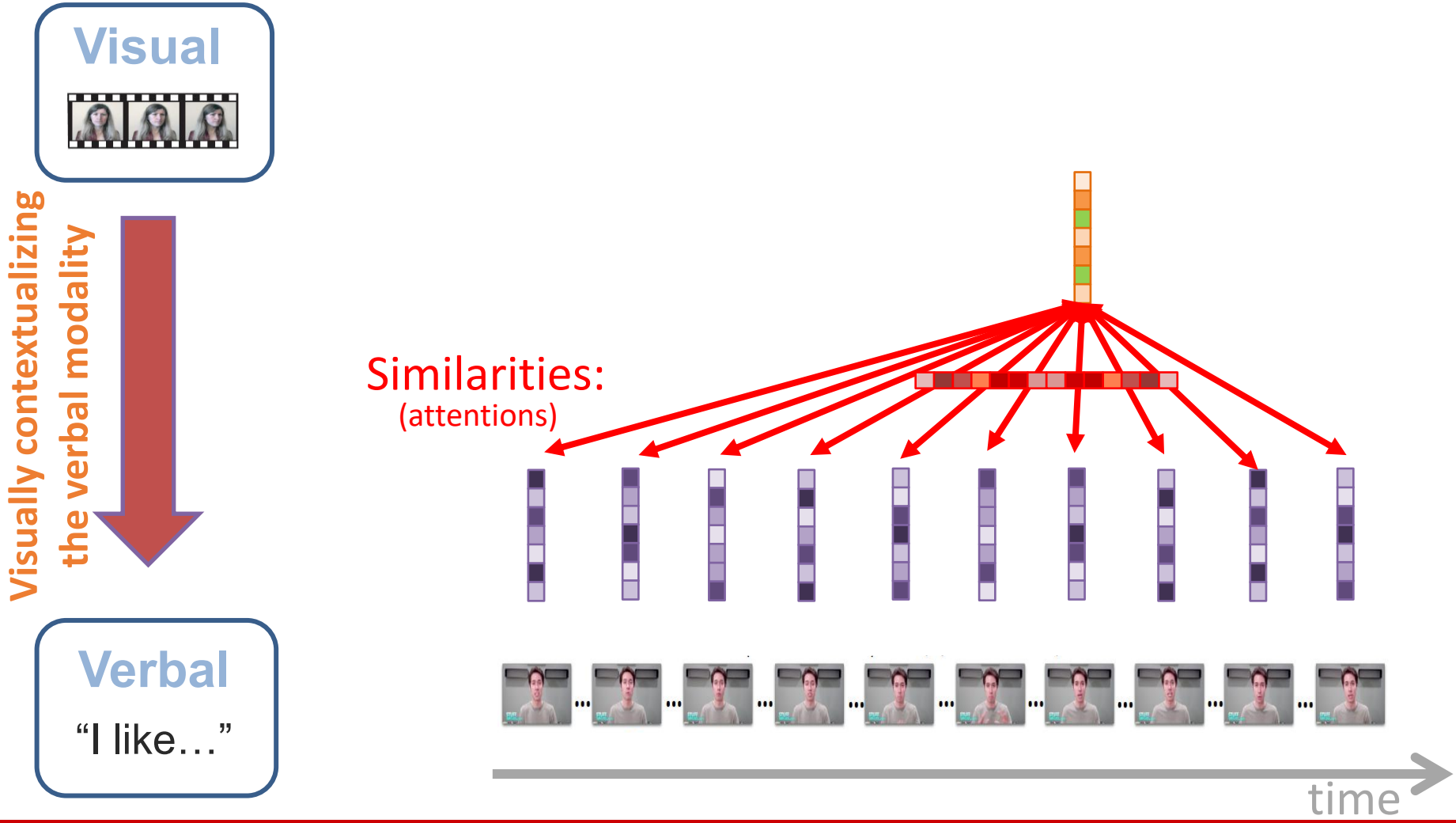



Option 2: Look at pairwise interactions between modalities

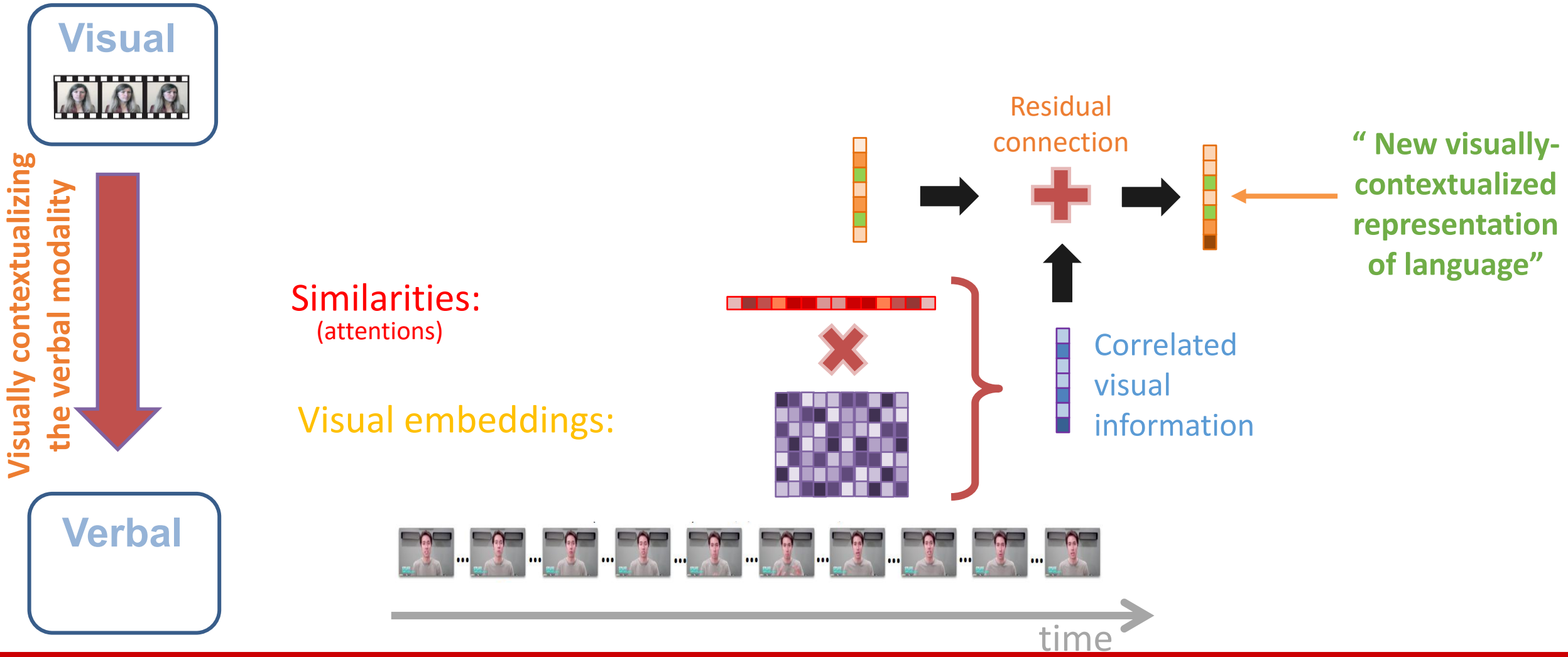
# Multimodal Transformer – Pairwise Cross-Modal



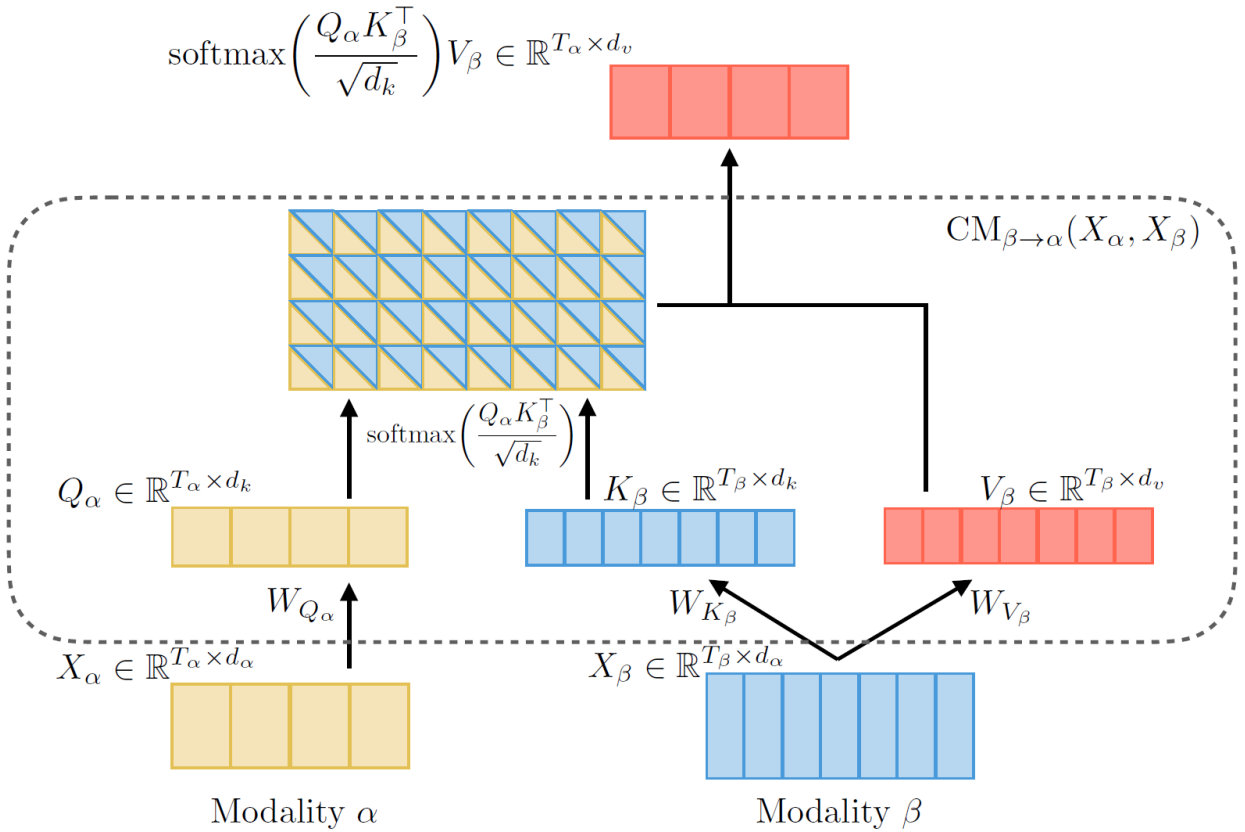
# Cross-Modal Transformer Module ($V \rightarrow L$)



# Cross-Modal Transformer Module ($V \rightarrow L$)

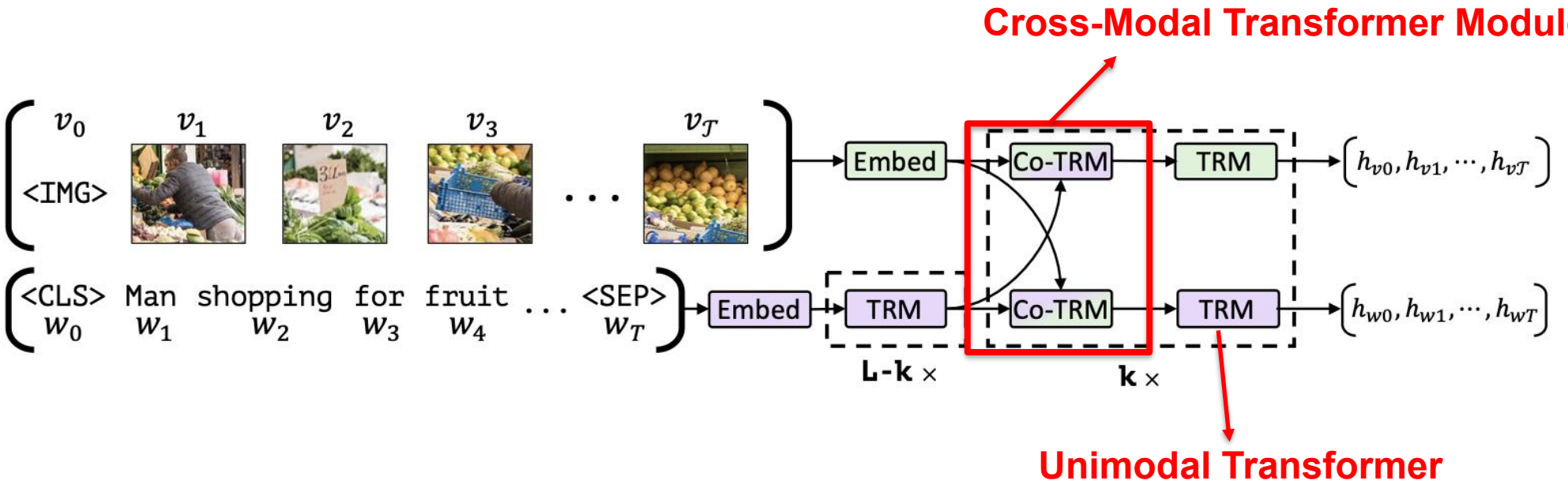


# Cross-Modal Transformer Module ($\beta \rightarrow \alpha$)



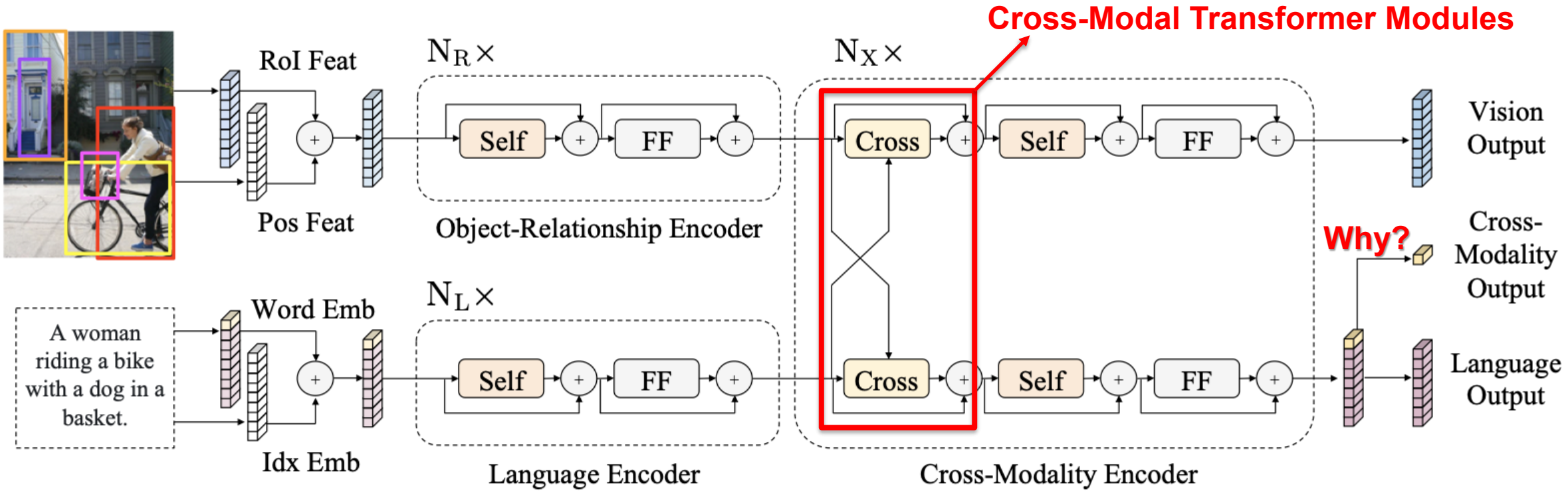

Tsai et al., Multimodal Transformer for Unaligned Multimodal Language Sequences, ACL 2019

# ViLBERT



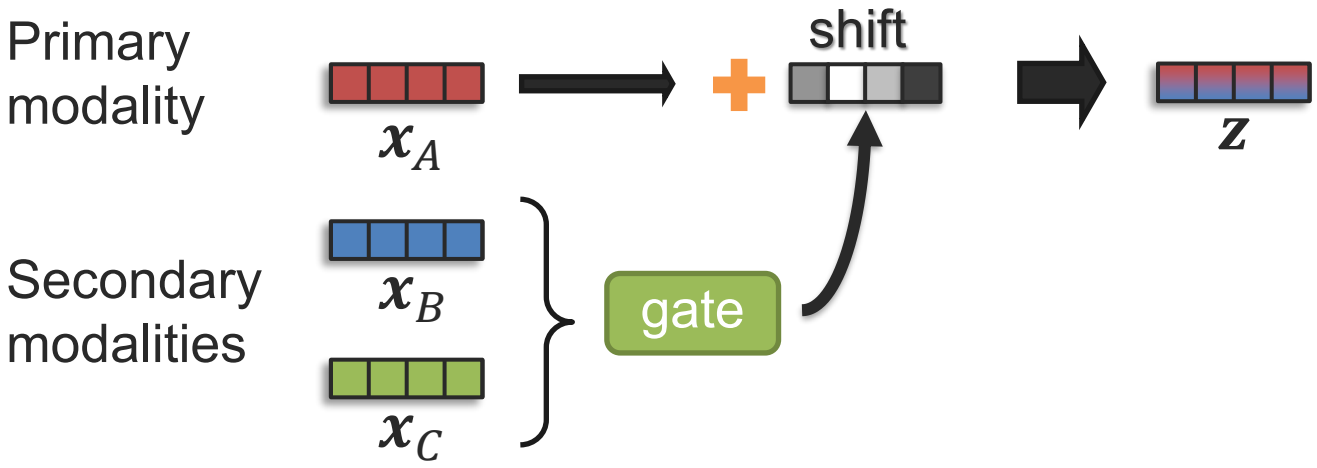

Lu, Jiasen, et al. "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks." *arXiv* (August 6, 2019).

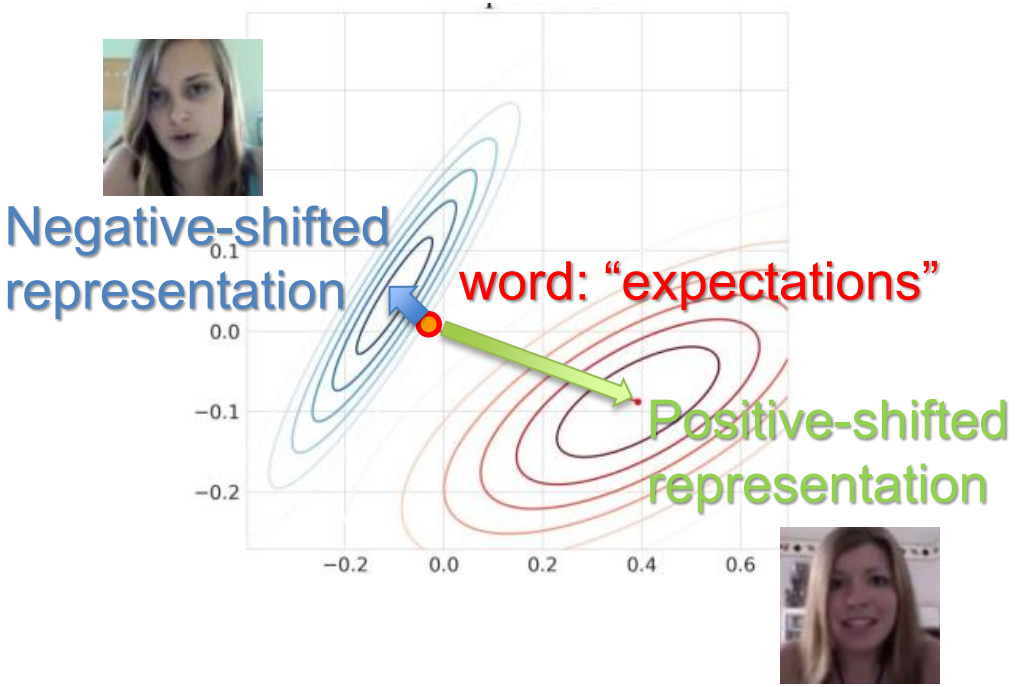
# LXMERT




Tan, Hao, and Mohit Bansal. "Lxmert: Learning cross-modality encoder representations from transformers." *arXiv* (August 20, 2019).

# Reminder: Modality-Shifting Fusion




## Example with language modality:

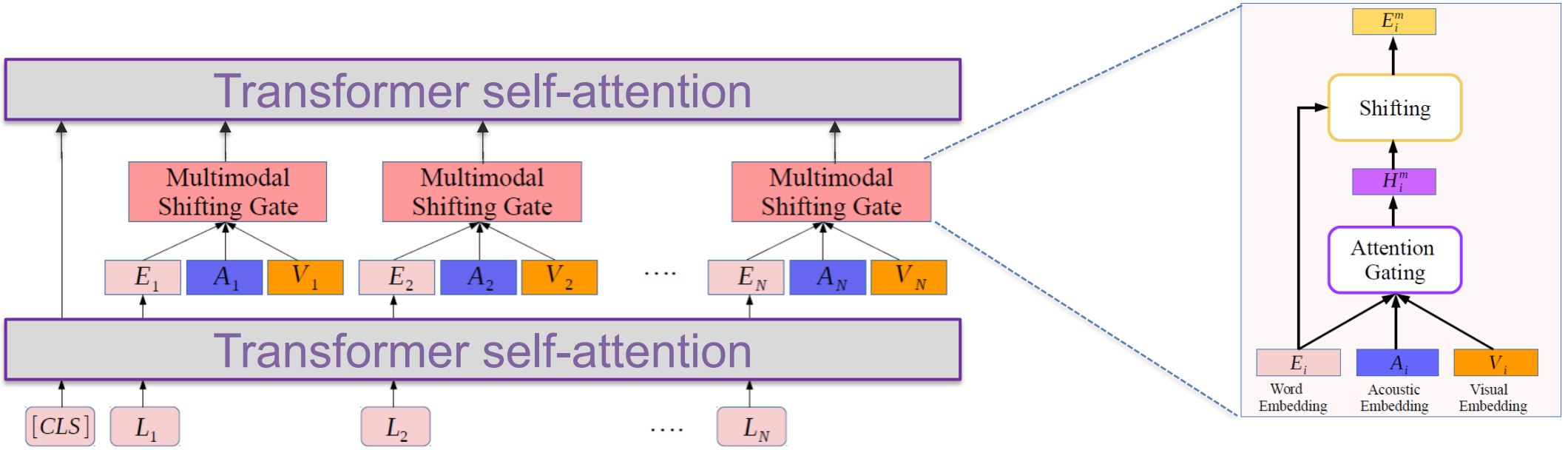
Primary modality: language

Secondary modalities: acoustic and visual

Wang et al., Words Can Shift: Dynamically Adjusting Word Representations Using Nonverbal Behaviors, AAAI 2019

# Modality-Shifting with Transformers

## Multimodal Adaptation Gate (MAG) + BERT




Rahman et al., Integrating Multimodal Information in Large Pretrained Transformers, ACL 2020