Natural Language Processing

AI51701/CSE71001

Lecture 21

11/30/2023

Instructor: Taehwan Kim

Announcements

- Final project report
  - 30% of your total grade, submit via BlackBoard-Assignment menu. Only one team member may submit it.
  - Due: on 12/11 at 11:59:00pm KST
  - The same lateness policy applies. No submission accepted after three late days – no exception!
  - **Detailed requirements** including section structure are stated.
  - Latex-created .pdf is required with the EMNLP 2023 style template.

Announcements

- Final project presentation schedule
  - 2-hour lecture on 12/12 instead of two lectures on 12/12,14
  - 10 min. presentation with 2 min. QnA for each team
    - Bring your laptop/tablet for presentation.
    - Do not change the slides after submission!
      - For fairness, I will deduct the score if it happens

Announcements

- Final project presentation schedule

```
>>> numpy.random.permutation(10)
array([3, 1, 0, 7, 6, 2, 9, 4, 8, 5])
>>>
```

	12/12
1	Dahee Kim, Song Kim
2	Jaewon Yang, Sungho Jeon
3	Jihyoung Jang, Eunchan Lee, Juhyeong Lee
4	Taegyeong Lee, Taesoo Kim, Sergey Pyatkovskiy
5	Jinwoo Lee, Seongouk Kim, Meraj Mammadov
6	Gaurav Saha, Youngbin Ki, Eldor Fozilov, Rogelio Ruzcko Tobias
7	Ryskul Tagmanova
8	Sangjune Park, Jongsung Lee
9	Jaemu Heo, Hyunmin Song
10	Namwoo Kwon, Gawon Choi

Multi Modal Learning

What is Multi-modal Learning?

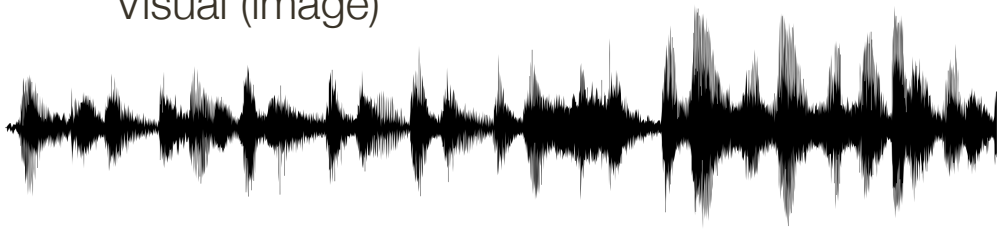
- **Modality**: refers to a certain type of information and/or representation format in which information is stored.
- **Sensory modality**: one or more primary channels of communication.

Visual (drawings)

Visual (image)

Owl Wisdom, Omens, Vision of the night No bird has as much myth and mystery surrounding it than the owl. Part of this mystical aura is due to the fact that the bird is nocturnal and the night time has always seemed mysterious to humans. The owl is a symbol of the feminine, the moon, and the night. Because of it association with the moon it has ties to fertility and seduction. The owl is bird of magic and darkness of prophecy and wisdom.

Natural Language (text)



Auditory (voice / sound)

Haptic / Touch

From UBC CPSC532L course

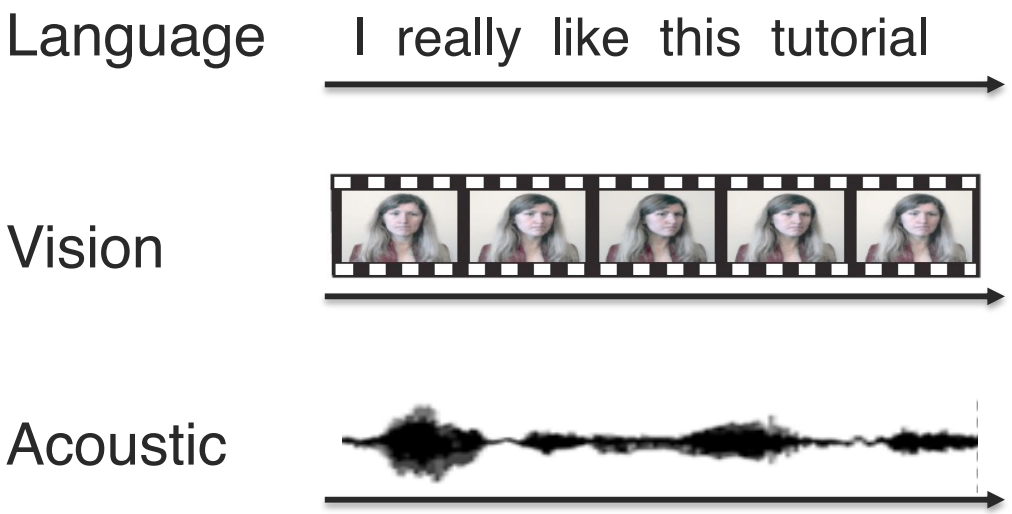
What is Multimodal Learning?

- **Multimodal Learning** (Multimodal Machine Learning) is the study of computer algorithms that learn and improve through the use and experience of data from multiple modalities
- **Multimodal Artificial Intelligence (AI)** studies computer agents able to demonstrate intelligence capabilities such as understanding, reasoning and planning, through multimodal experiences, and data

Multimodal AI is a superset of Multimodal ML

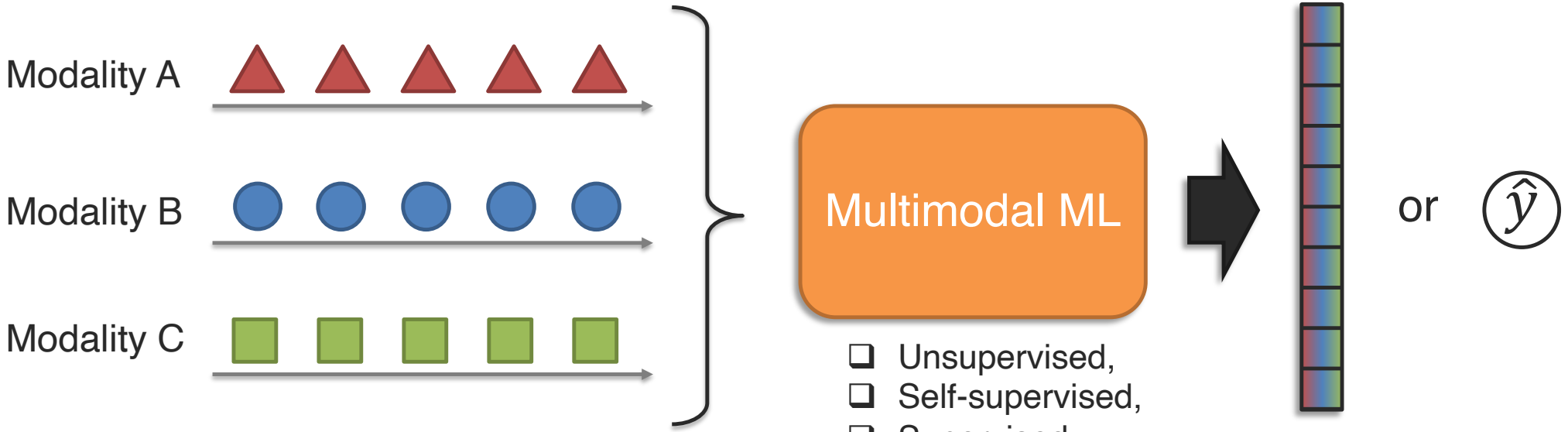
From slides by Louis-Philippe Morency

Multimodal Learning



From slides by Louis-Philippe Morency

Multimodal Learning



- ❑ Unsupervised,
- ❑ Self-supervised,
- ❑ Supervised,
- ❑ Reinforcement,

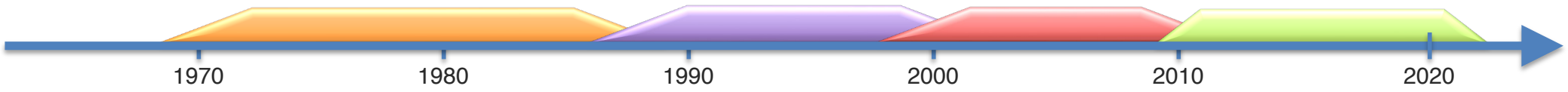
…

Structure: static, temporal, spatial, hierarchical

From slides by Louis-Philippe Morency

Prior Research in “Multimodal”

- Four eras of multimodal research
- The “behavioral” era (1970s until late 1980s)
- The “computational” era (late 1980s until 2000)
- The “interaction” era (2000 - 2010)
- The “deep learning” era (2010s until ...)



From slides by Louis-Philippe Morency

Behavioral Study of Multimodal

Language
and gestures

David McNeill

"For McNeill, gestures are in effect the speaker's thought in action, and integral components of speech, not merely accompaniments or additions."

McGurk effect

From slides by Louis-Philippe Morency

Behavioral Study of Multimodal

Language
and gestures

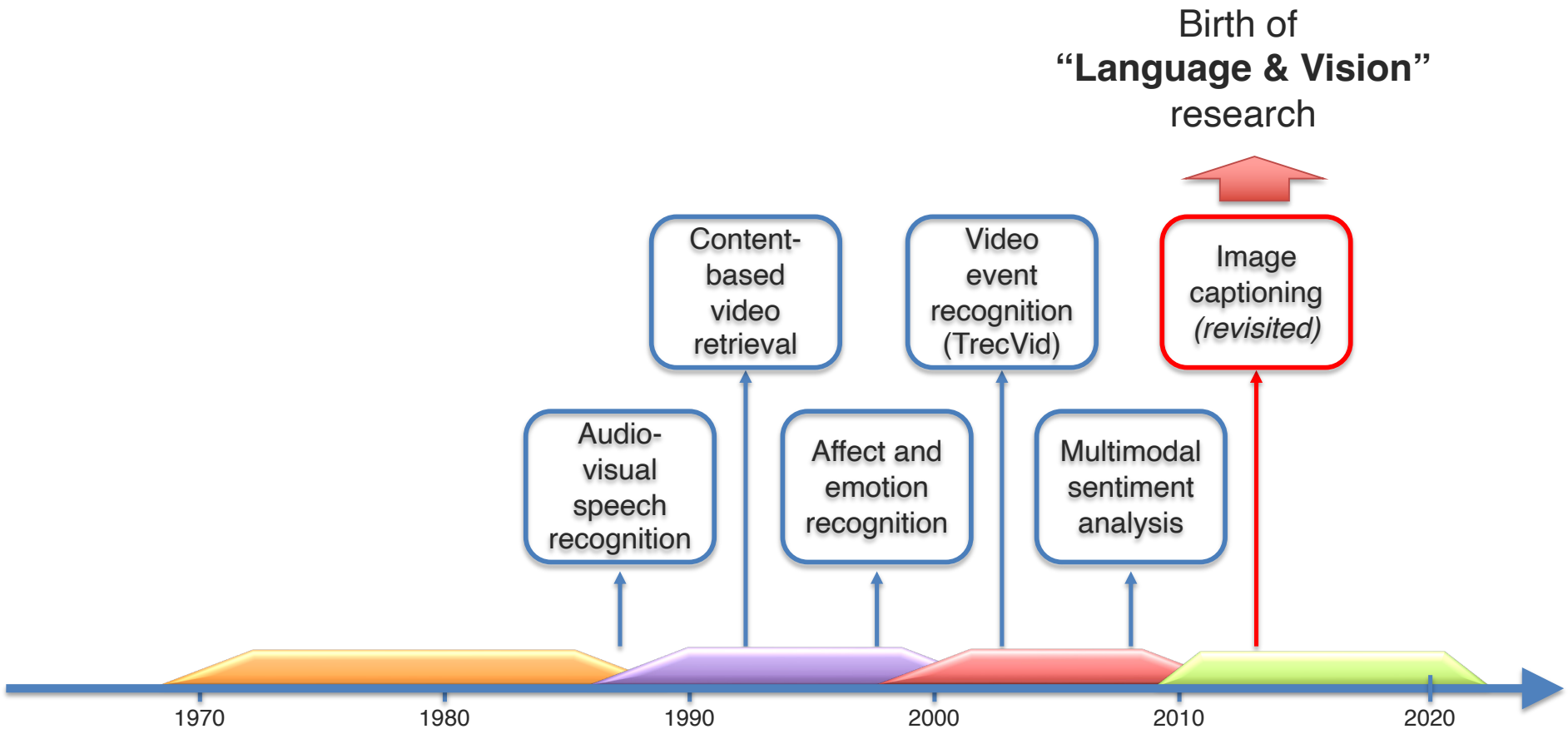
David McNeill

"For McNeill, gestures are in effect the speaker's thought in action, and integral components of speech, not merely accompaniments or additions."

McGurk effect

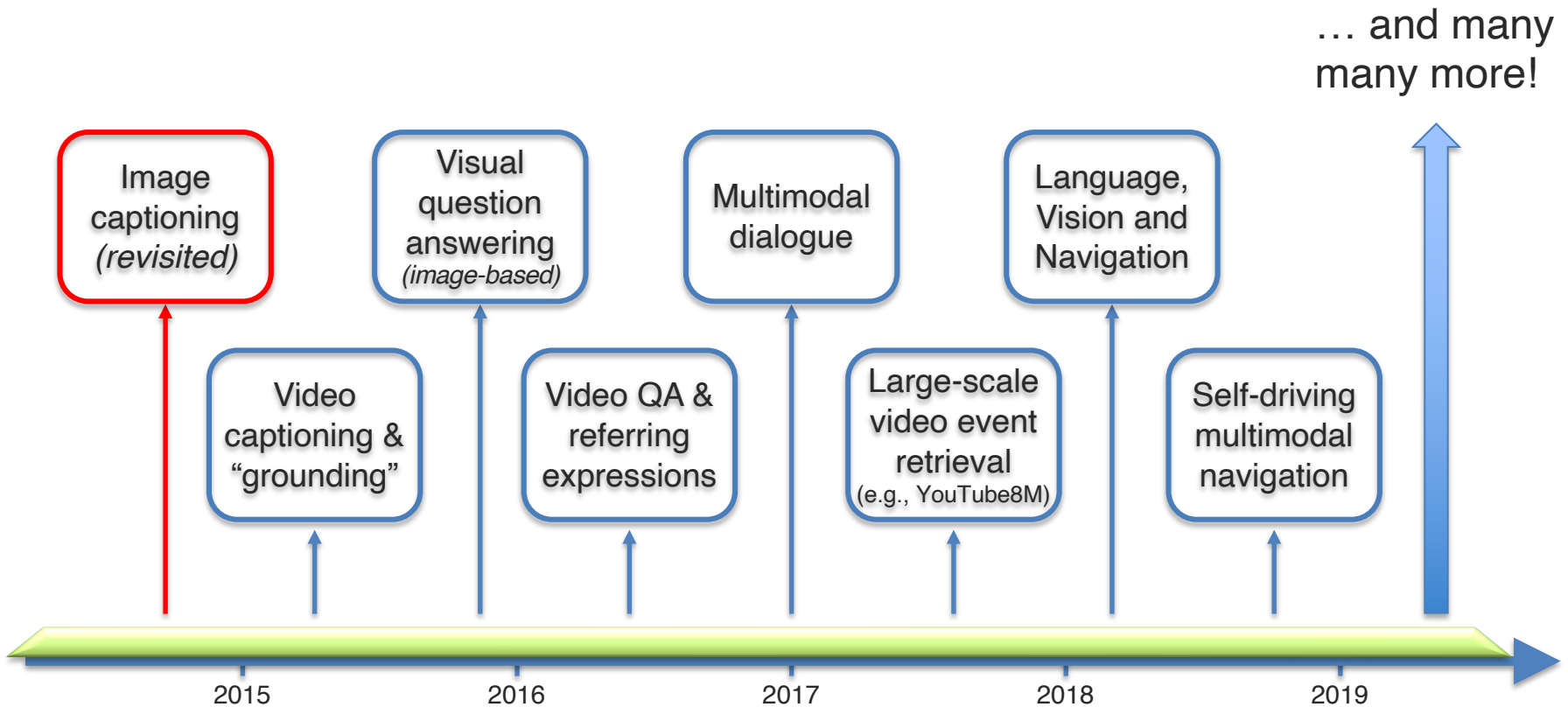
From slides by Louis-Philippe Morency

Multimodal Research Tasks



From slides by Louis-Philippe Morency

Multimodal Research Tasks



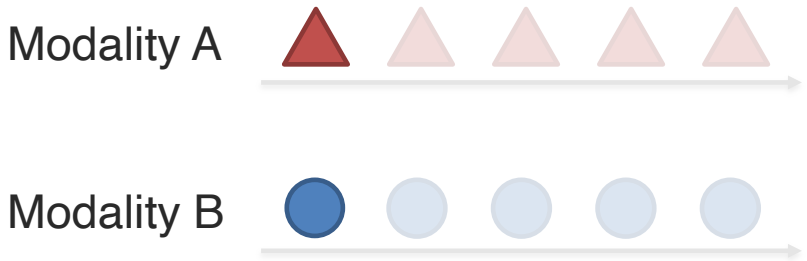
From slides by Louis-Philippe Morency

Challenge 1: Representation

Definition: Learning representations that reflect cross-modal interactions between individual elements, across different modalities

➡ This is a core building block for most multimodal modeling problems!

Individual elements:



It can be seen as a "local" representation

or

representation using holistic features

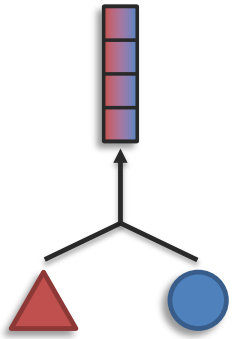
From slides by Louis-Philippe Morency

Challenge 1: Representation

Definition: Learning representations that reflect cross-modal interactions between individual elements, across different modalities

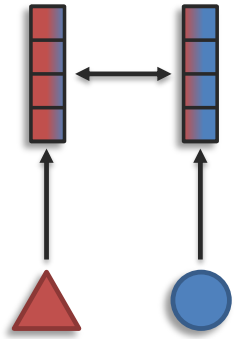
Sub-challenges:

Fusion



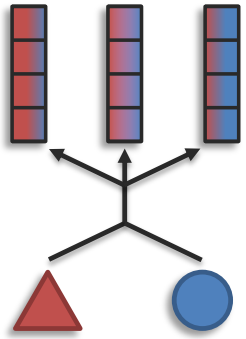
modalities **>** # representations

Coordination



modalities **=** # representations

Fission



modalities **<** # representations

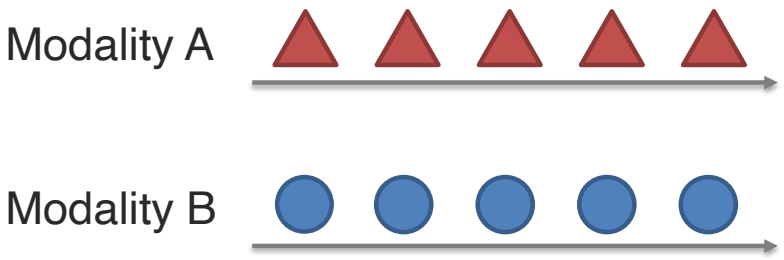
From slides by Louis-Philippe Morency

Challenge 2: Alignment

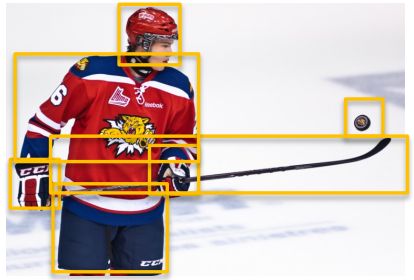
Definition: Identifying and modeling cross-modal connections between all elements of multiple modalities, building from the data structure

➡ Most modalities have internal structure with multiple elements

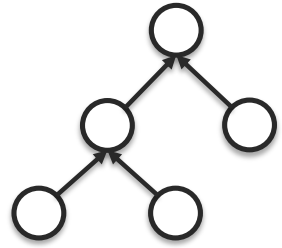
Elements with temporal structure:



Other structured examples:



Spatial



Hierarchical

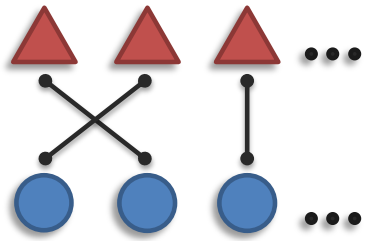
From slides by Louis-Philippe Morency

Challenge 2: Alignment

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

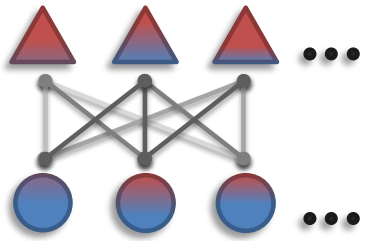
Sub-challenges:

Connections



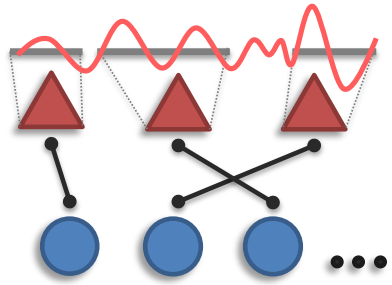
Explicit alignment
(e.g., grounding)

Aligned Representation



Implicit alignment
+ representation

Segmentation

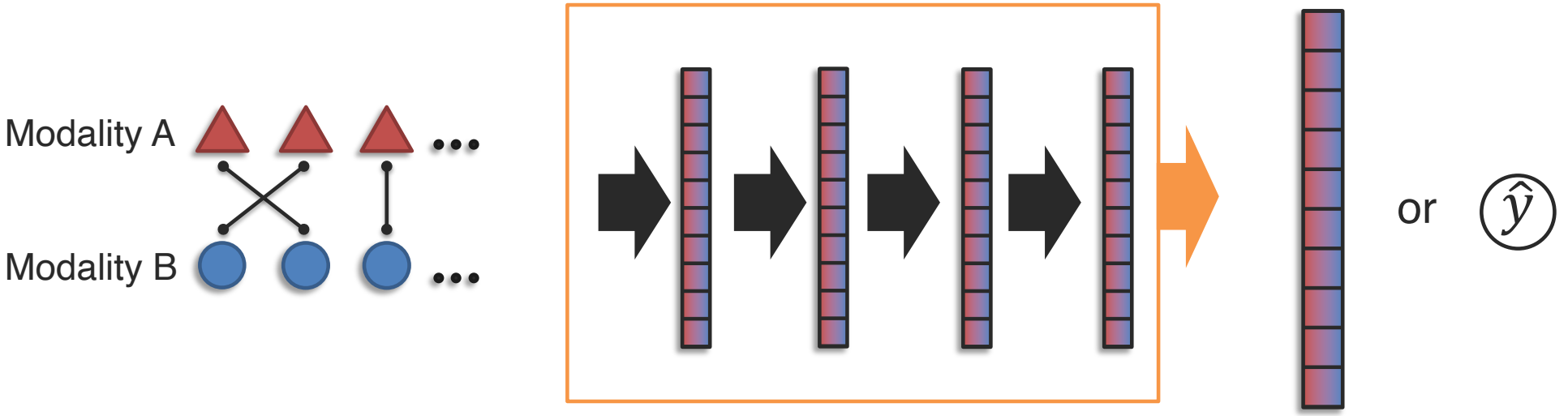


Granularity of
individual elements

From slides by Louis-Philippe Morency

Challenge 3: Reasoning

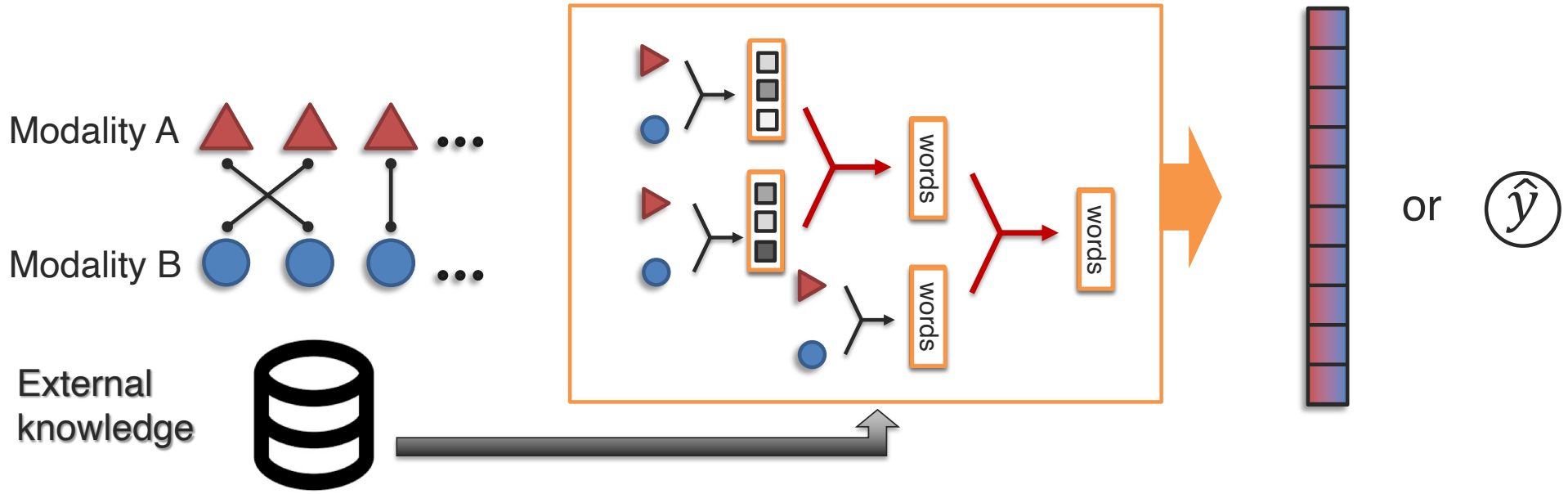
Definition: Combining knowledge, usually through multiple inferential steps, exploiting multimodal alignment and problem structure



From slides by Louis-Philippe Morency

Challenge 3: Reasoning

Definition: Combining knowledge, usually through multiple inferential steps, exploiting multimodal alignment and problem structure



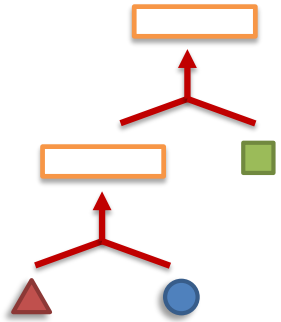
From slides by Louis-Philippe Morency

Challenge 3: Reasoning

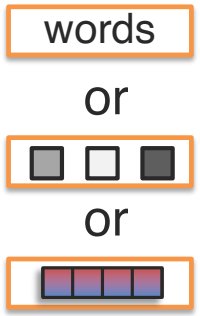
Definition: Combining knowledge, usually through multiple inferential steps, exploiting multimodal alignment and problem structure

Sub-challenges:

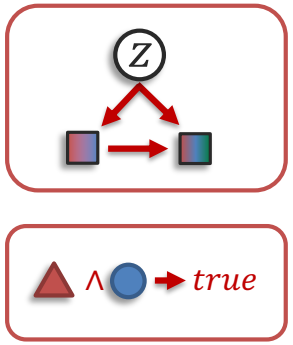
Structure Modeling



Intermediate concepts



Inference Paradigm



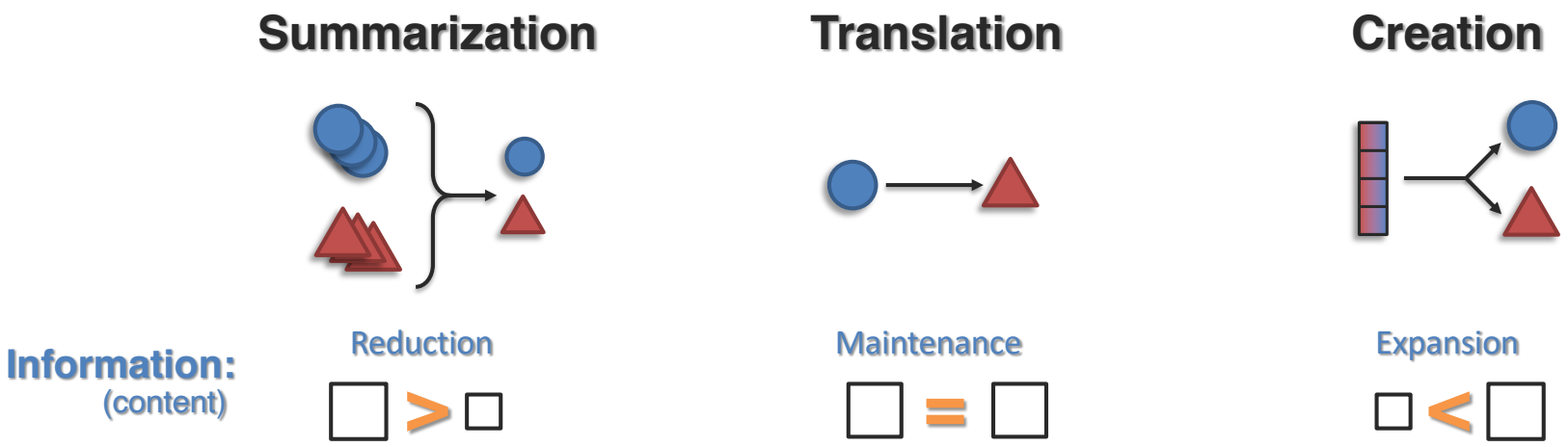
External Knowledge

From slides by Louis-Philippe Morency

Challenge 4: Generation

Definition: Learning a generative process to produce raw modalities that reflects cross-modal interactions, structure and coherence

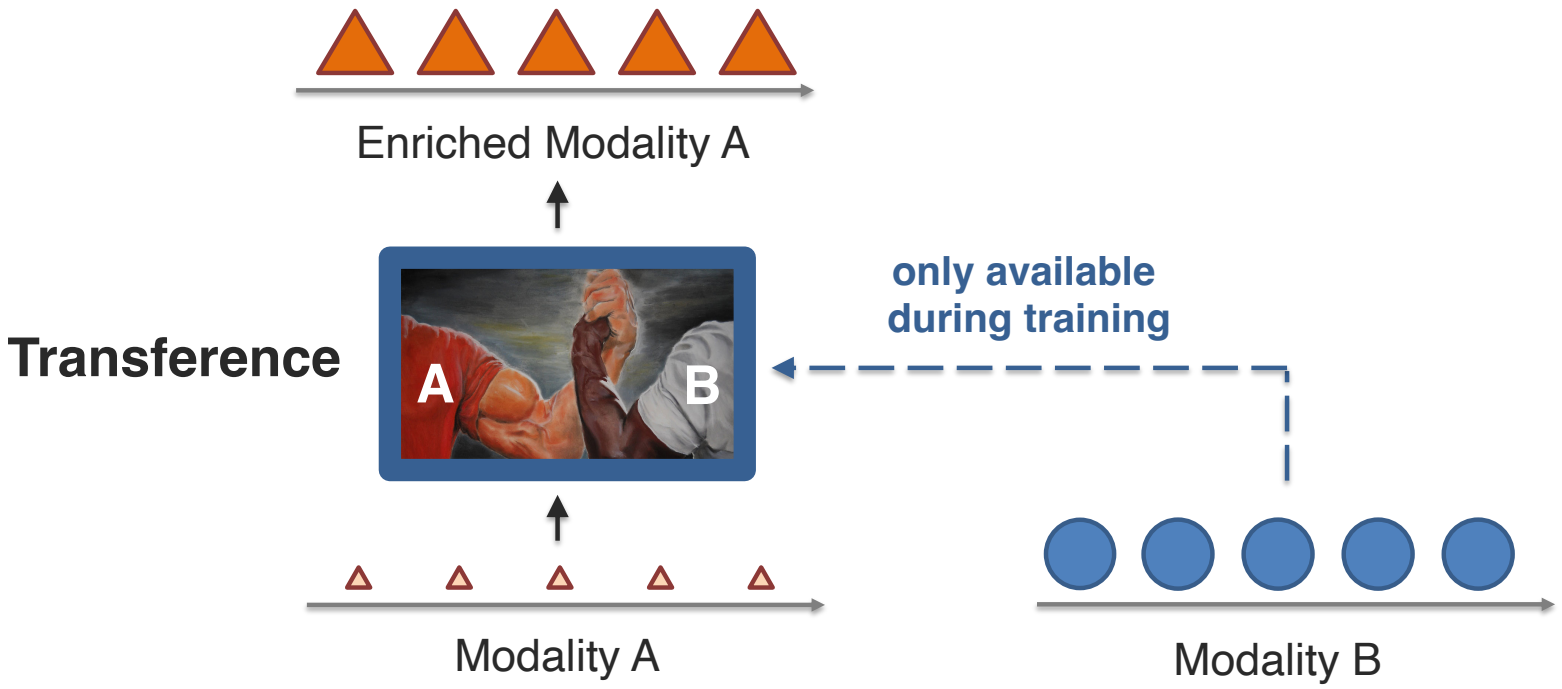
Sub-challenges:



From slides by Louis-Philippe Morency

Challenge 5: Transference

Definition: Transfer knowledge between modalities, usually to help the target modality which may be noisy or with limited resources



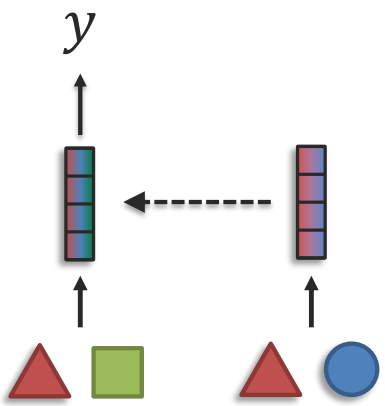
From slides by Louis-Philippe Morency

Challenge 5: Transference

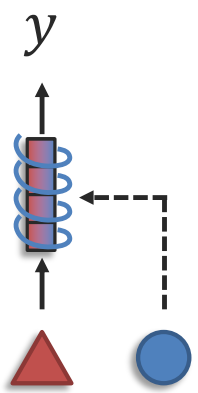
Definition: Transfer knowledge between modalities, usually to help the target modality which may be noisy or with limited resources

Sub-challenges:

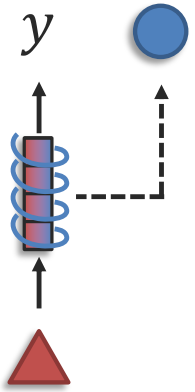
Transfer



Co-learning via representation



Co-learning via generation



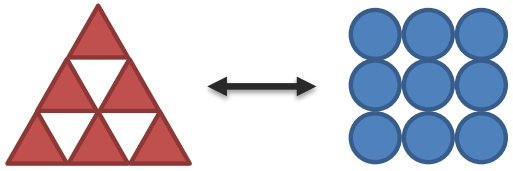
From slides by Louis-Philippe Morency

Challenge 6: Quantification

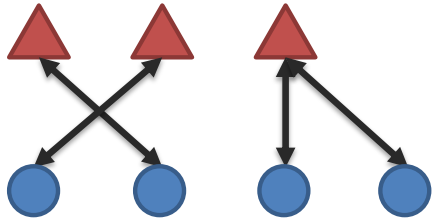
Definition: Empirical and theoretical study to better understand heterogeneity, cross-modal interactions and the multimodal learning process

Sub-challenges:

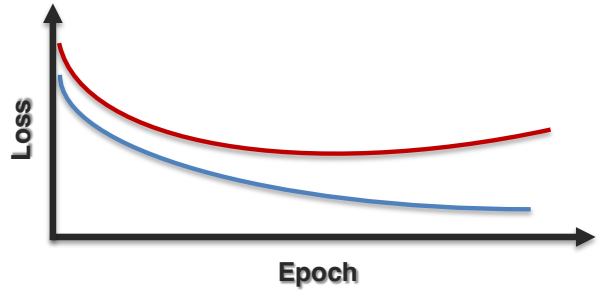
Heterogeneity



Interactions

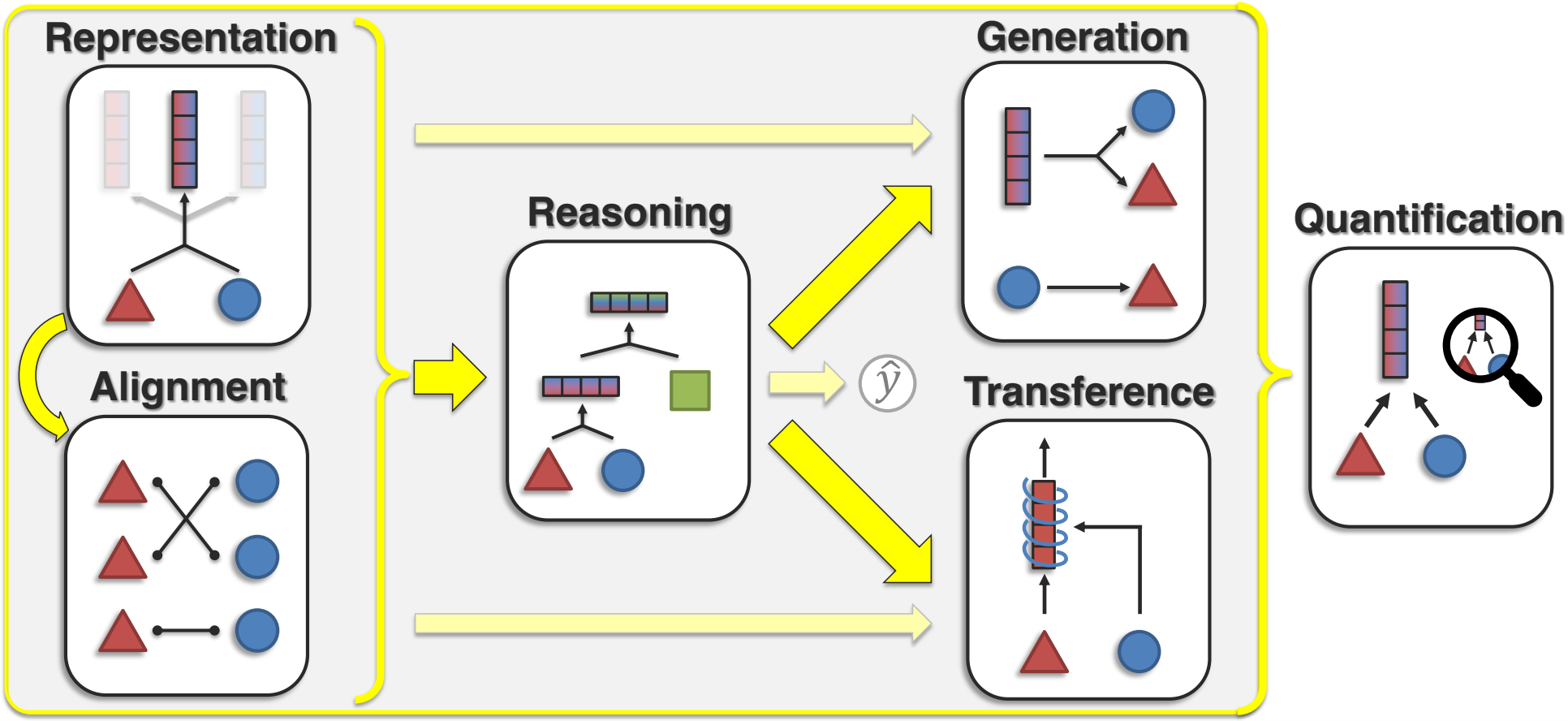


Learning



From slides by Louis-Philippe Morency

Core Multimodal Challenges



From slides by Louis-Philippe Morency

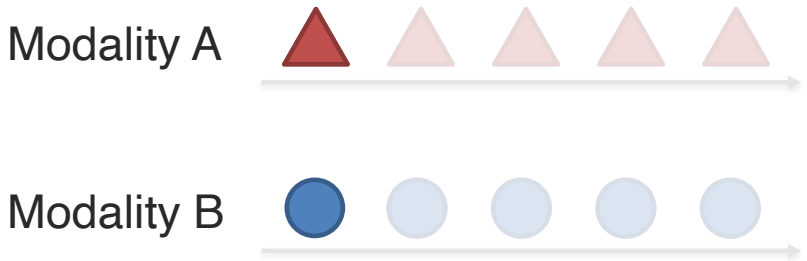
Representation

Challenge 1: Representation

Definition: Learning representations that reflect cross-modal interactions between individual elements, across different modalities

➡ This is a core building block for most multimodal modeling problems!

Individual elements:



It can be seen as a "local" representation

or

representation using holistic features

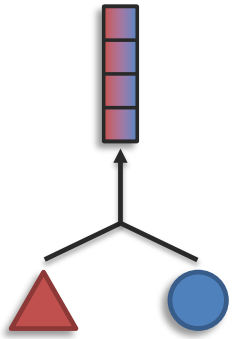
From slides by Louis-Philippe Morency

Challenge 1: Representation

Definition: Learning representations that reflect cross-modal interactions between individual elements, across different modalities

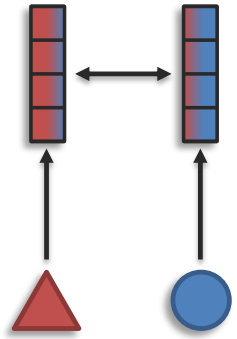
Sub-challenges:

Fusion



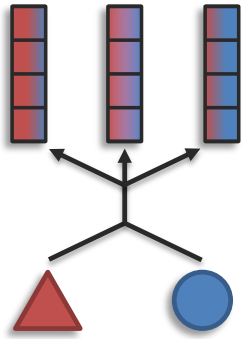
modalities **>** # representations

Coordination



modalities **=** # representations

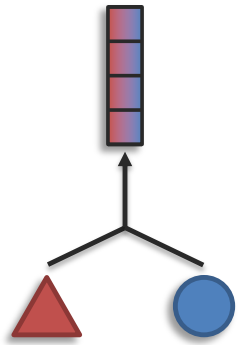
Fission



modalities **<** # representations

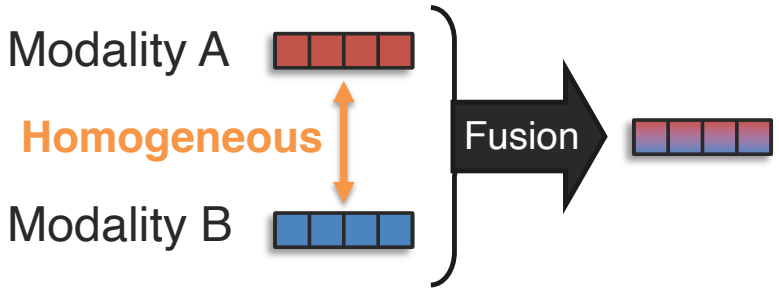
From slides by Louis-Philippe Morency

Sub-Challenge 1a: Representation Fusion

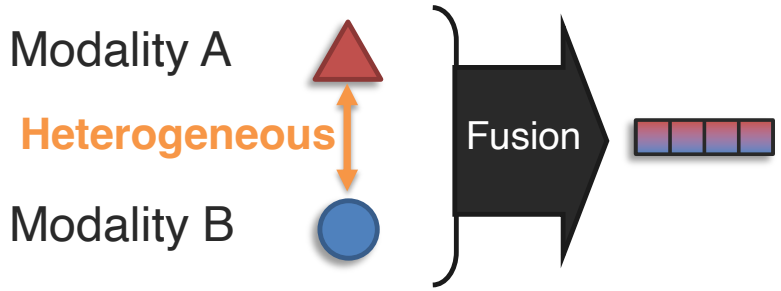


Definition: Learn a joint representation that models cross-modal interactions between individual elements of different modalities

Basic fusion:



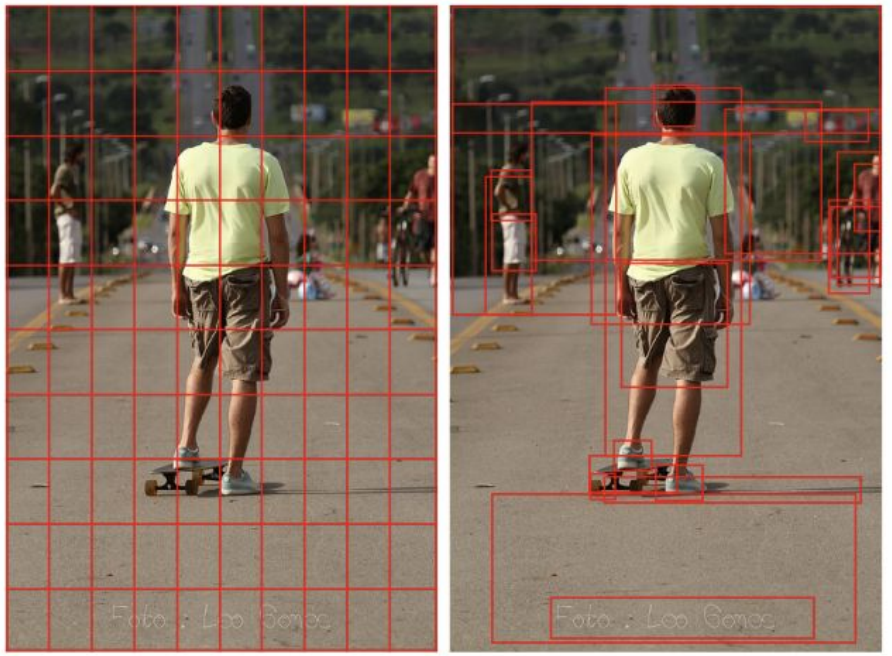
Complex fusion:



From slides by Louis-Philippe Morency

Features

- Featurizing text: Batch_size x Sequence_length x Hidden_size.
- Featurizing images:
  - Sparse “region” features:
    - Object detectors
  - Dense features:
    - ConvNet layer(s) or feature maps
    - Vision Transformer layers

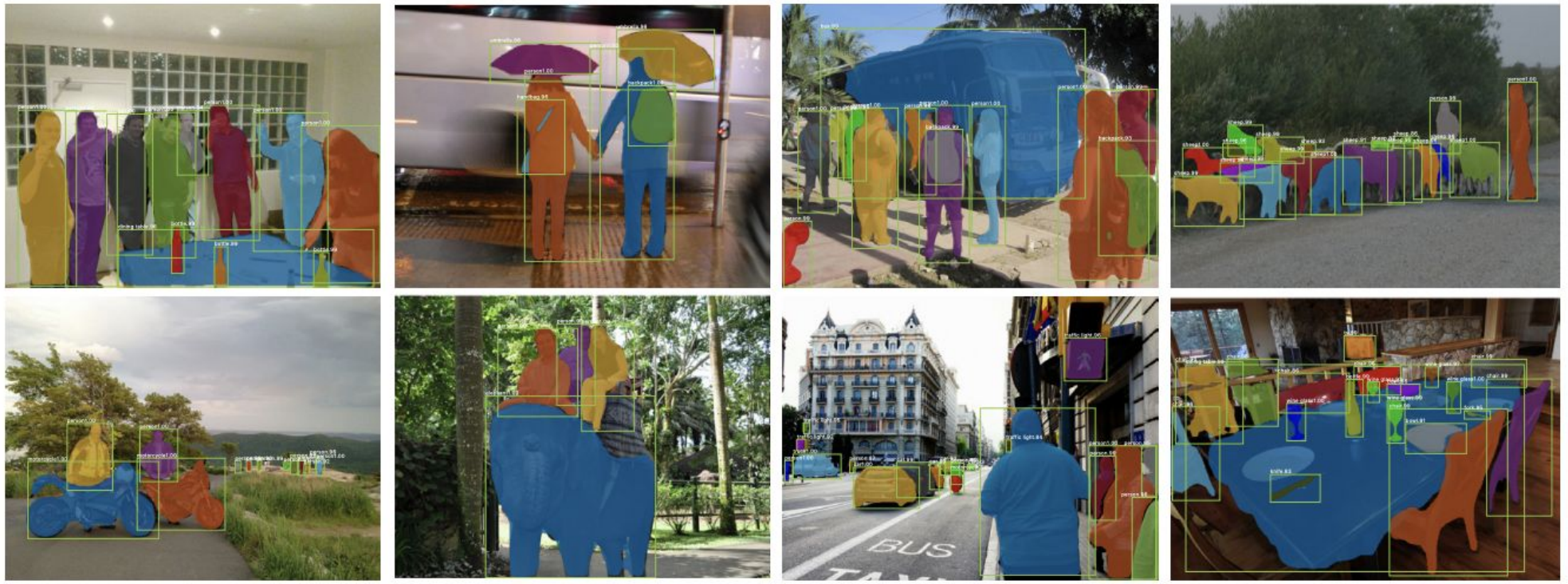


Anderson et al., 2018

From Stanford 224n course

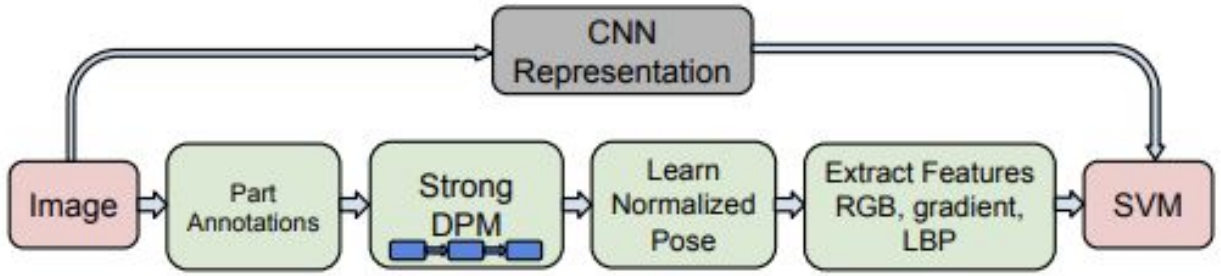
Region features

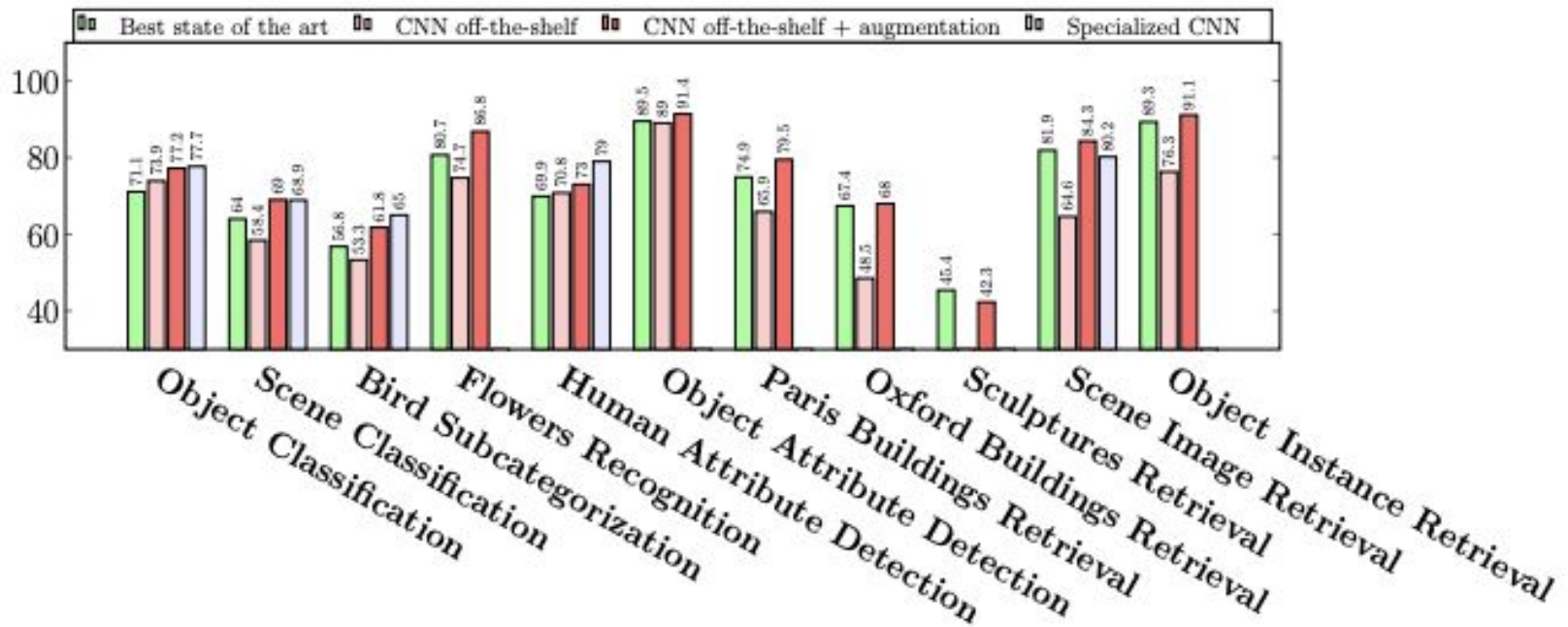
- R-CNN (Girshick et al., 2014); Fast R-CNN (Girshick, 2015); Faster R-CNN (Ren et al., 2015); YOLO (you only look once); …



From Stanford 224n course

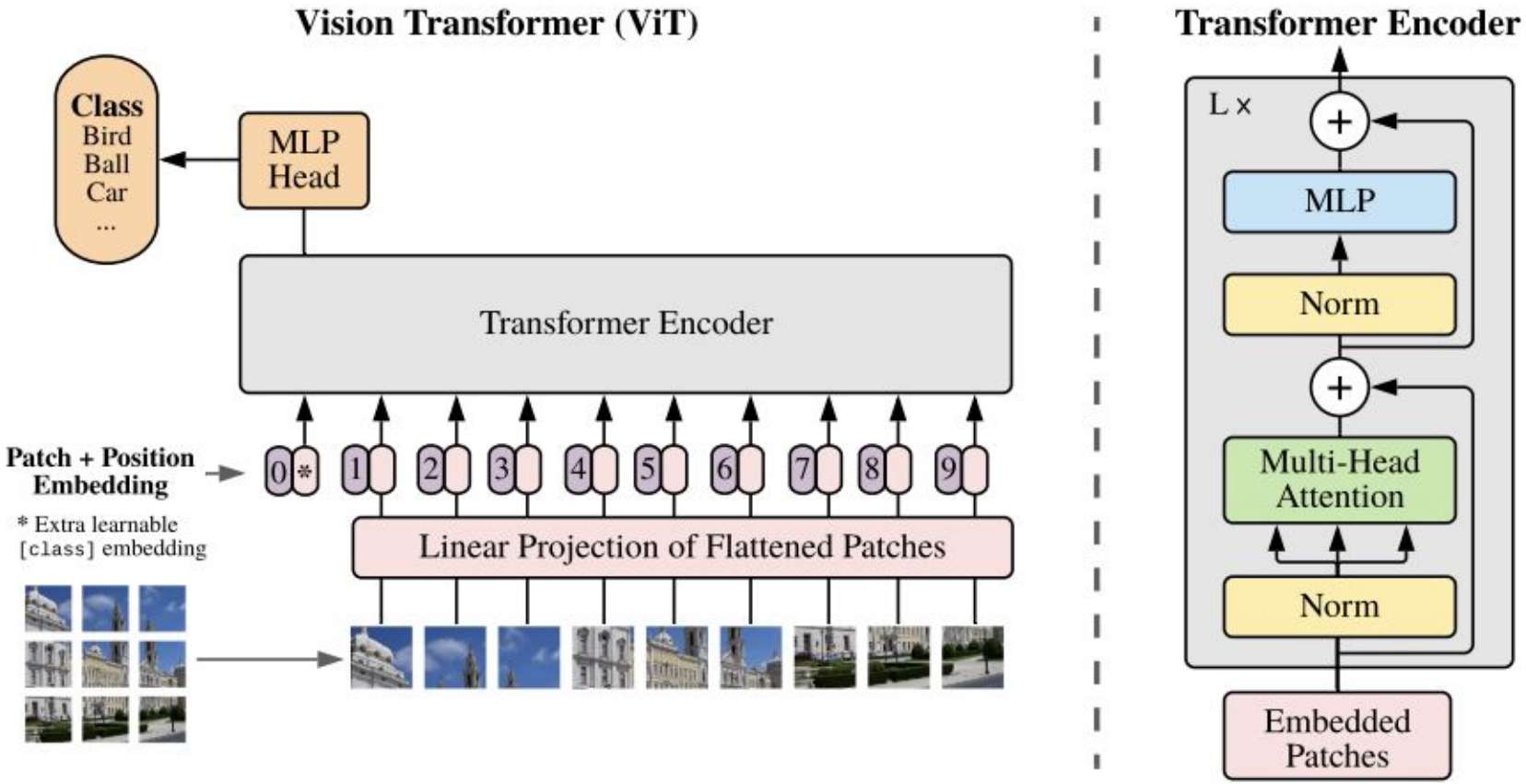
“Off the shelf” ConvNet features (Razavian et al., 2014)





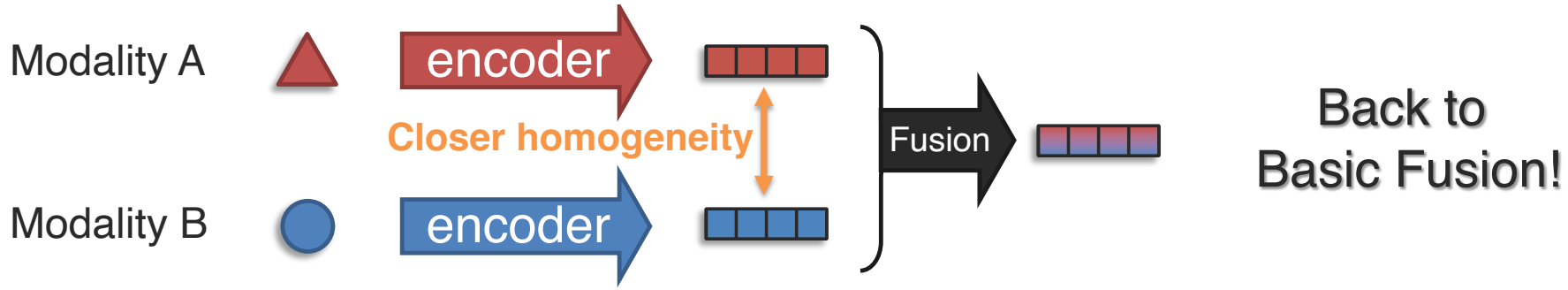
From Stanford 224n course

Vision Transformers (Dosovitskiy et al., 2020)



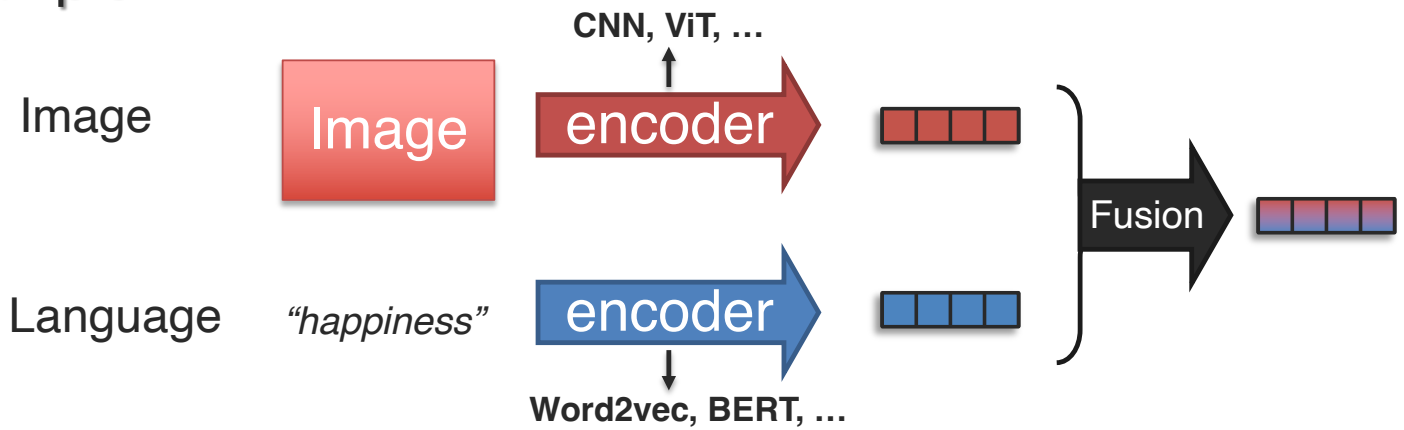
From Stanford 224n course

Fusion with Unimodal Encoders



Back to Basic Fusion!

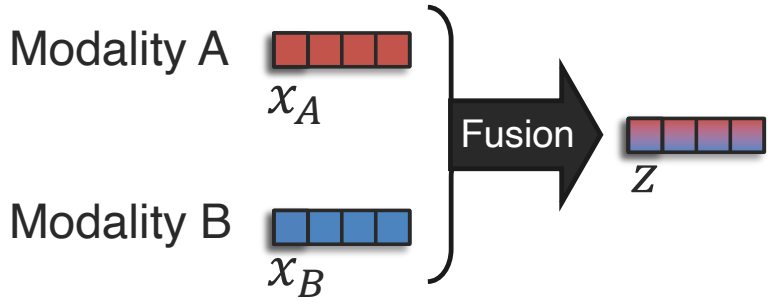
Example:



➡ Unimodal encoders can be jointly learned with fusion network, or pre-trained

From slides by Louis-Philippe Morency

Basic Concepts for Representation Fusion (aka, Basic Fusion)



Goal: Model *cross-modal interactions* between the multimodal elements

➡ Let's study the univariate case first
(only 1-dimensional features)

Linear regression:

$$z = w_0 + w_1 x_A + w_2 x_B + w_3 (x_A \times x_b) + \epsilon$$

w_0: constant
$w_1 x_A + w_2 x_B$: Additive terms
$w_3 (x_A \times x_b)$: Multiplicative term
ϵ: error

① Additive interaction:

$$z = w_1 x_A + w_2 x_B + \epsilon$$

② Multiplicative interaction:

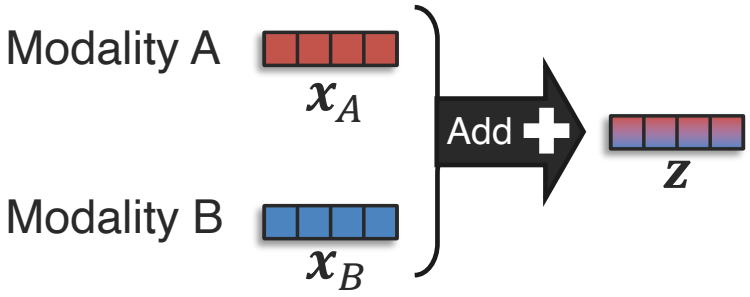
$$z = w_3 (x_A \times x_b) + \epsilon$$

③ Additive and multiplicative interactions:

$$z = w_1 x_A + w_2 x_B + w_3 (x_A \times x_b) + \epsilon$$

From slides by Louis-Philippe Morency

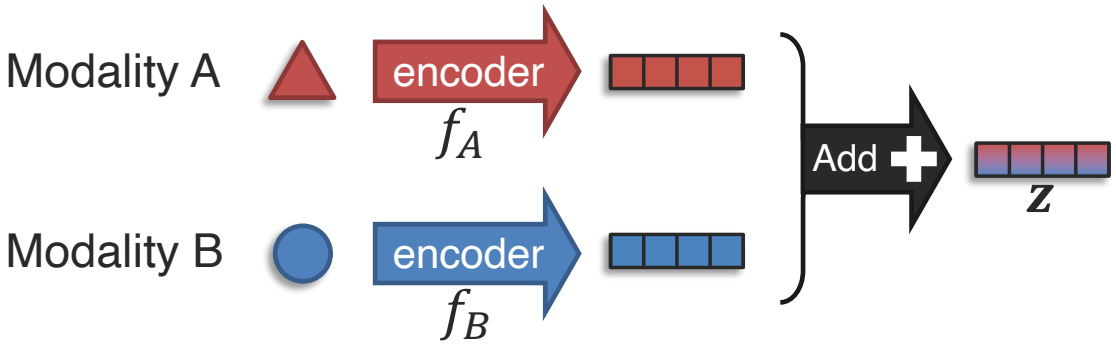
Additive Fusion



Additive fusion:

$$z = w_1 x_A + w_2 x_B = W \cdot \begin{bmatrix} x_A \\ x_B \end{bmatrix}$$

With unimodal encoders:



Additive fusion:

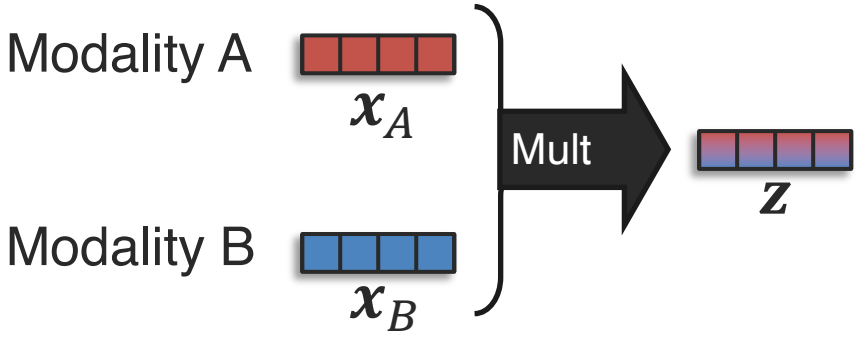
$$z = f_A(\blacktriangle) + f_B(\bullet)$$

⇒ It could be seen as an ensemble approach
(late fusion)

Early, middle, and late fusion

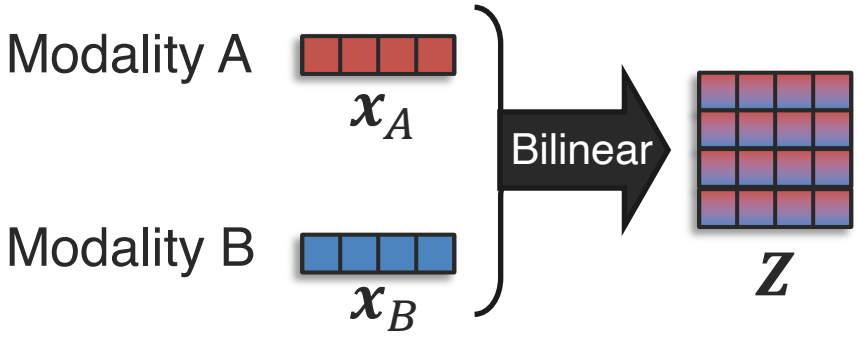
- Suppose we have a binary classifier MLP and two input vectors: **u**, **v**
- Early - mix inputs:
  - $\sigma(W2\sigma(W1[\mathbf{u},\mathbf{v}]+b1)+b2)$
- Middle - concatenate features:
  - $\sigma(W2[\sigma(W1[\mathbf{u}]+b1), \sigma(W'1[\mathbf{v}]+b'1)] +b2)$
- Late - combine final scores:
  - $1/2\ (\sigma(W2\sigma(W1[u]+b1)+b2) + \sigma(V2\sigma(V1[\mathbf{v}]+b'1)+b'2))$

Multiplicative Fusion



Multiplicative fusion:

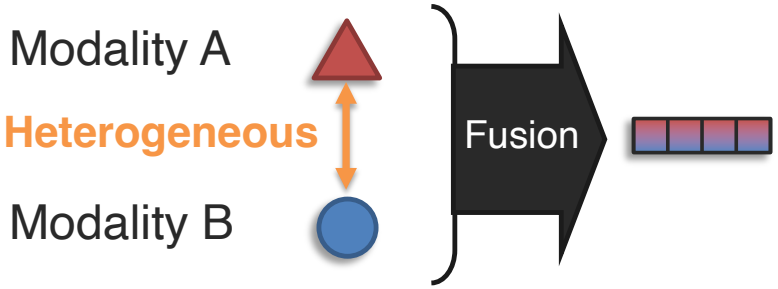
$$z = w(x_A \times x_B)$$



Bilinear Fusion:

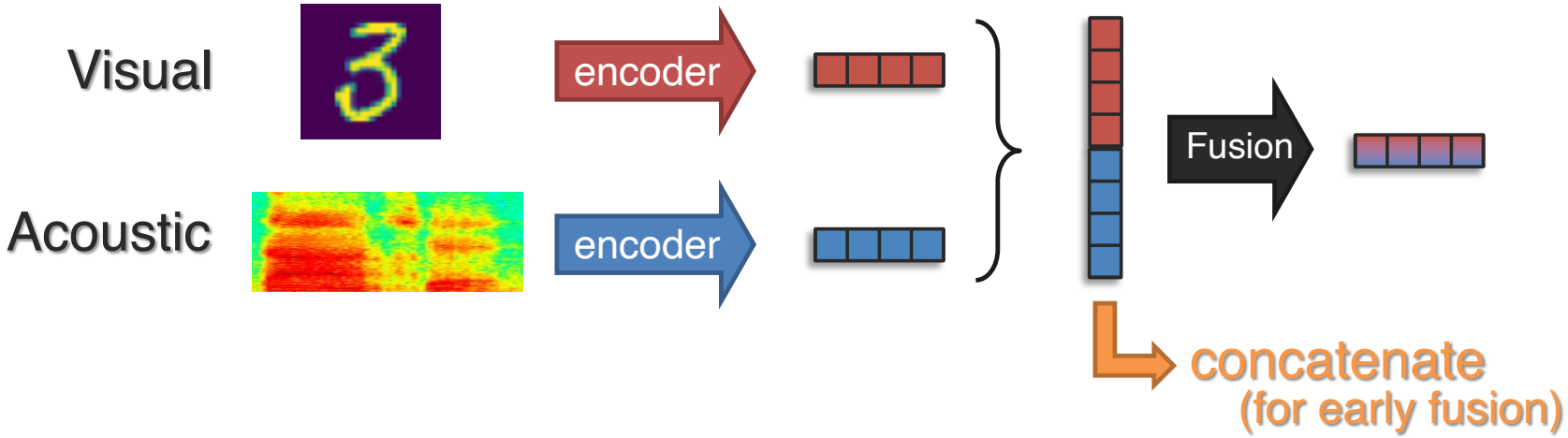
$$Z = w(x_A^T \cdot x_B)$$

Complex Fusion

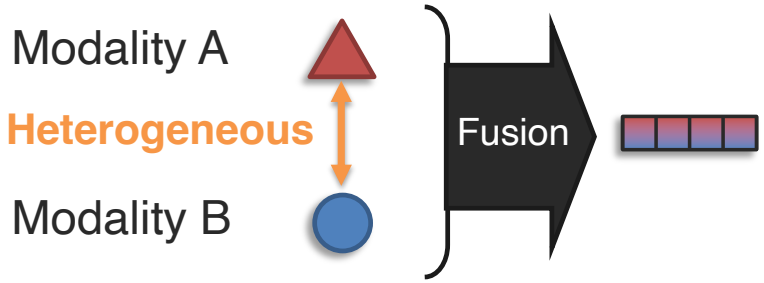


Open Challenge!

Example: From Early Fusion…

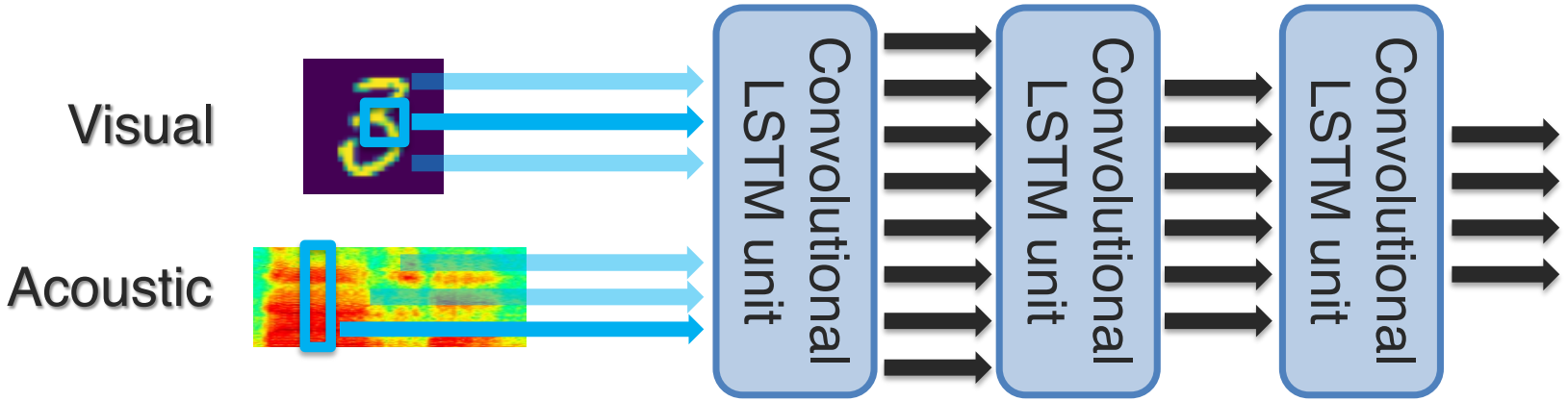


Complex Fusion



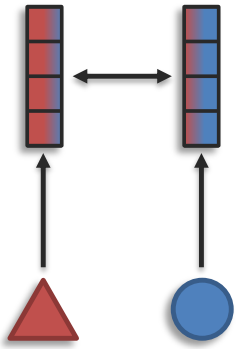
Open Challenge!

Example: From Early Fusion… to Very Early Fusion (inspired by human brain)



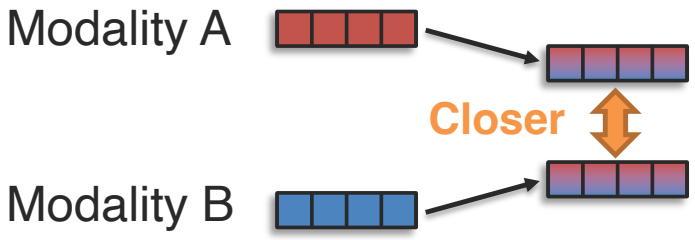
Barnum, et al. “On the Benefits of Early Fusion in Multimodal Representation Learning." *arxiv 2022*

Sub-Challenge 1b: Representation Coordination

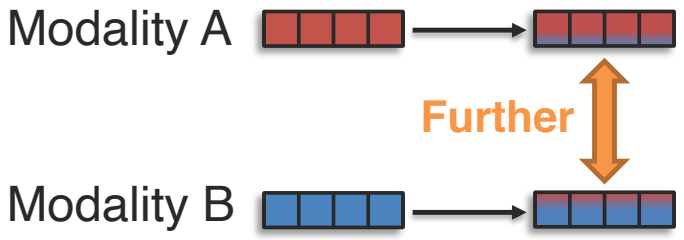


Definition: Learn multimodally-contextualized representations that are coordinated through their cross-modal interactions

Strong Coordination:

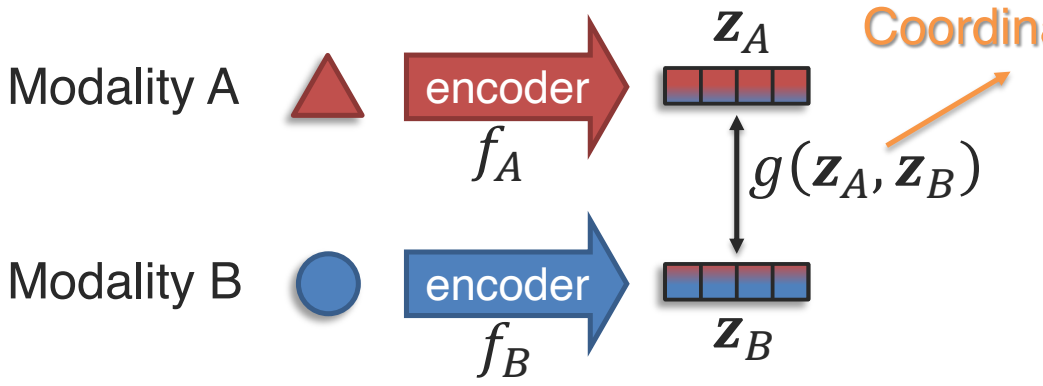


Partial Coordination:



From slides by Louis-Philippe Morency

Coordination Function



Learning with coordination function:

$$\mathcal{L} = g\big(f_A(\blacktriangle), f_B(\bullet)\big)$$

with model parameters θ_g, θ_{f_A} and θ_{f_B}

Examples of coordination function:

① Cosine similarity:

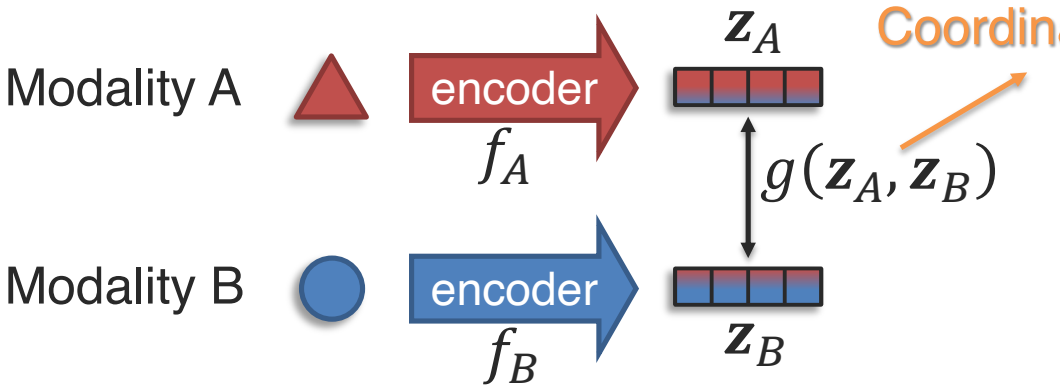
$$g(\mathbf{z}_A, \mathbf{z}_B) = \frac{\mathbf{z}_A \cdot \mathbf{z}_B}{\|\mathbf{z}_A\| \|\mathbf{z}_B\|}$$

Strong coordination!

➡ For normalized inputs (e.g., $\mathbf{z}_A - \overline{\mathbf{z}_A}$) , equivalent to *Pearson correlation coefficient*

From slides by Louis-Philippe Morency

Coordination Function



Learning with coordination function:

$$\mathcal{L} = g\big(f_A(\blacktriangle), f_B(\bullet)\big)$$

with model parameters θ_g, θ_{f_A} and θ_{f_B}

Examples of coordination function:

② Kernel similarity functions:

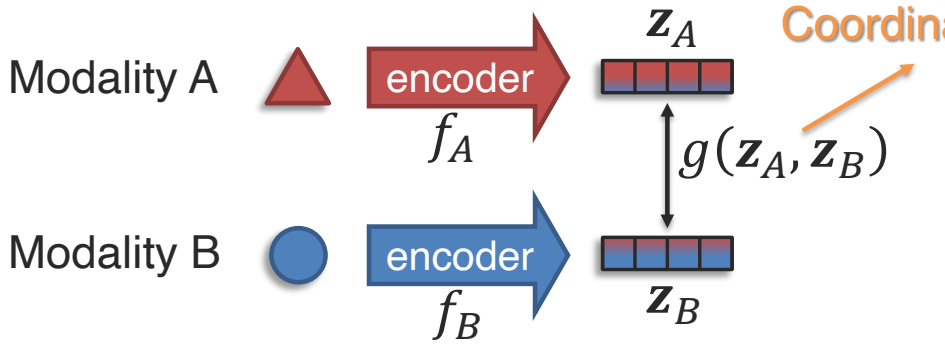
$$g(z_A, z_B) = k(z_A, z_B)$$

- Linear
- Polynomial
- Exponential
- RBF

➡ All these examples bring relatively strong coordination between modalities

From slides by Louis-Philippe Morency

Coordination Function



Learning with coordination function:

$$\mathcal{L} = g\big(f_A(\blacktriangle), f_B(\bullet)\big)$$

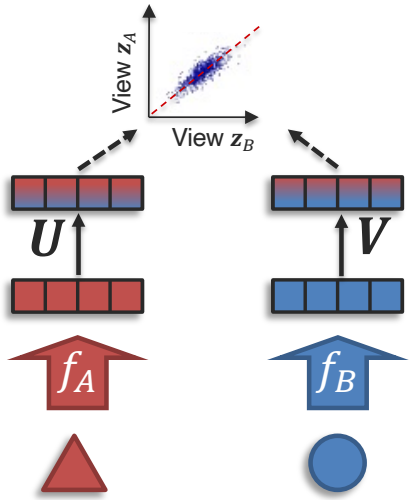
with model parameters θ_g, θ_{f_A} and θ_{f_B}

Examples of coordination function:

③ Canonical Correlation Analysis (CCA):

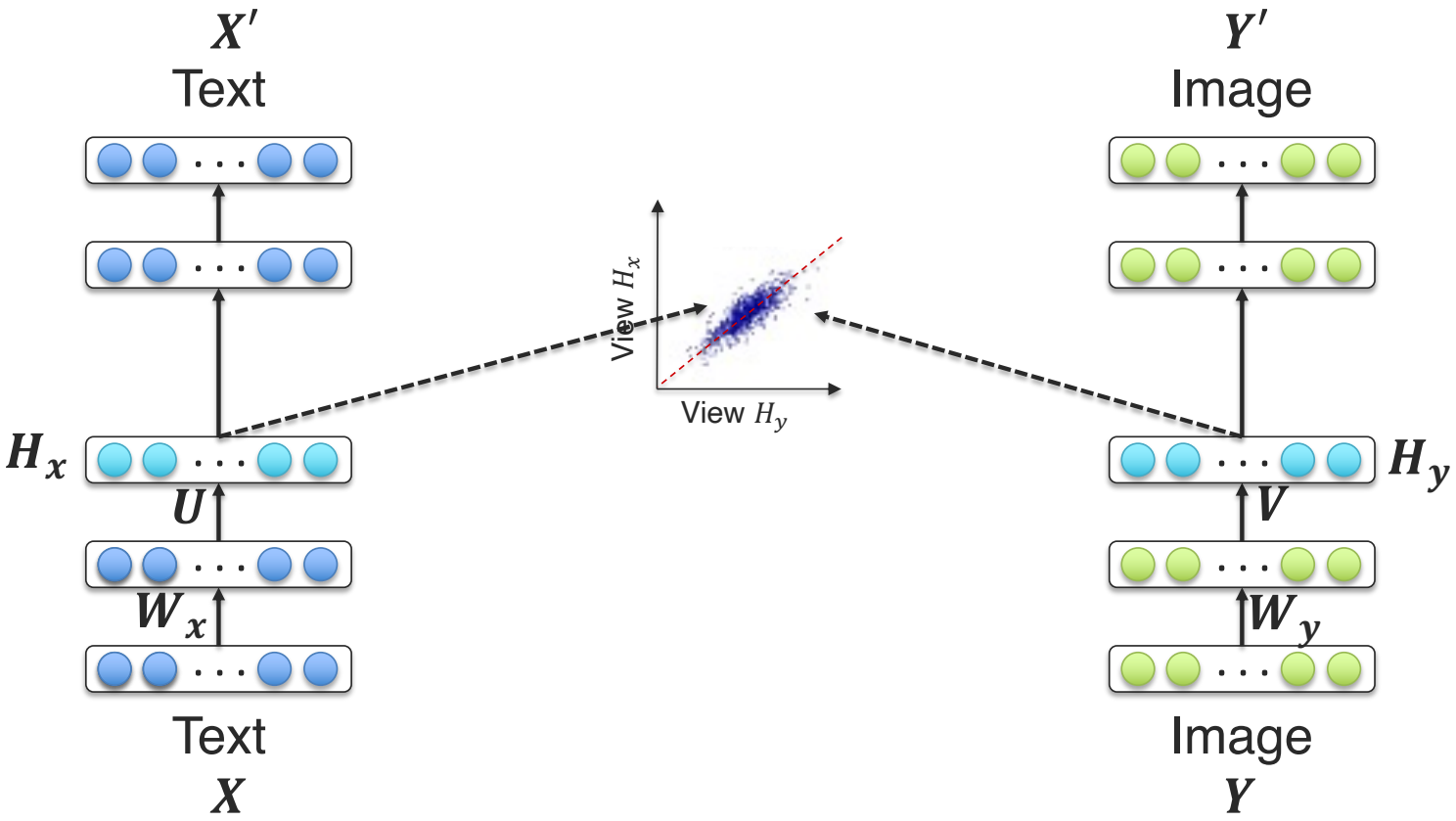
$$\underset{\boldsymbol{V},\boldsymbol{U},f_A,f_B}{\operatorname{argmax}}\ corr(\boldsymbol{z}_A, \boldsymbol{z}_B)$$

➡ CCA includes multiple projections, all orthogonal with each others



From slides by Louis-Philippe Morency

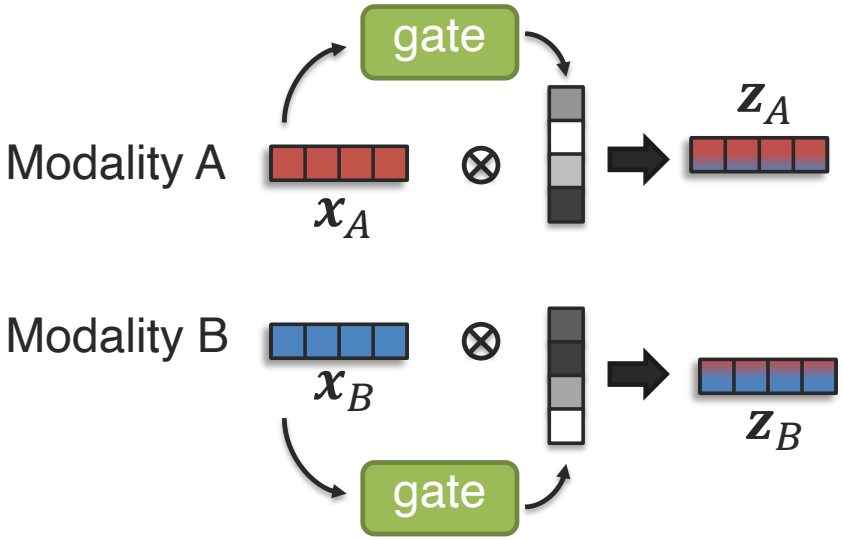
Deep Canonically Correlated Autoencoders (DCCAE)



Wang et al., On deep multi-view representation learning, PMLR 2015

From slides by Louis-Philippe Morency

Gated Coordination



Gated coordination:

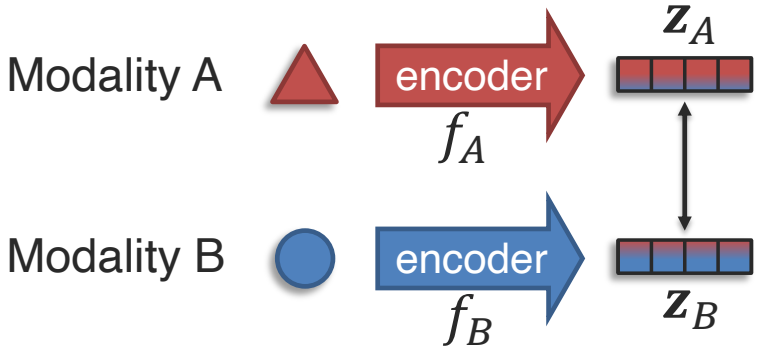
$$z_A = g_A(x_A, x_B) \cdot x_A$$

$$z_B = g_B(x_A, x_B) \cdot x_B$$

Related to attention modules in transformers

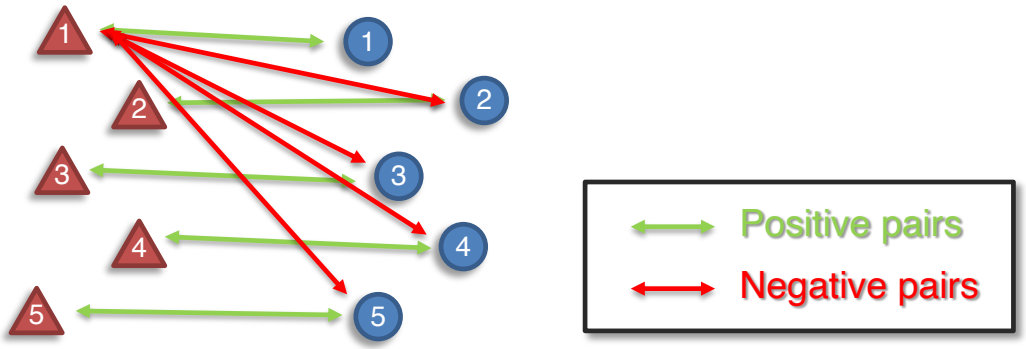
From slides by Louis-Philippe Morency

Coordination with Contrastive Learning



Paired data: $\{ \blacktriangle, \bullet \}$

(e.g., images and text descriptions)



Contrastive loss:

➡ brings positive pairs closer and pushes negative pairs apart

Simple contrastive loss:

$$\max\{0, \alpha + sim(\boldsymbol{z}_A, \boldsymbol{z}_B^+) - sim(\boldsymbol{z}_A, \boldsymbol{z}_B^-)\}$$

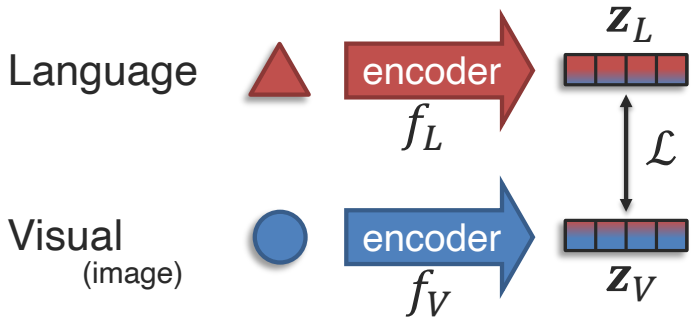
positive pairs

negative pair

Similarity functions are often cosine similarity

From slides by Louis-Philippe Morency

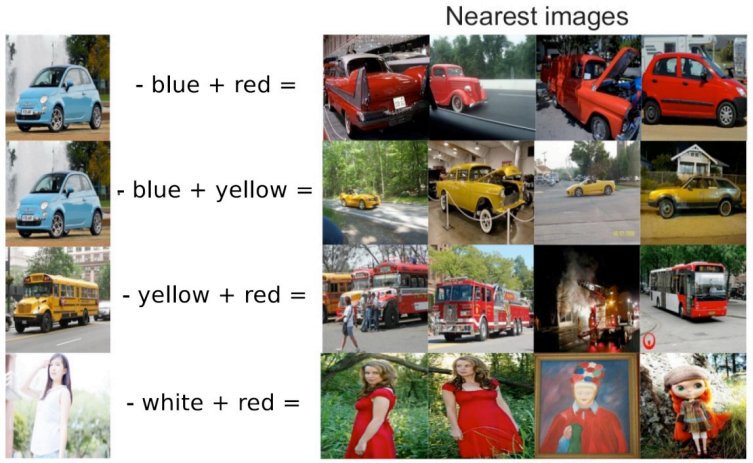
Example – Visual-Semantic Embeddings

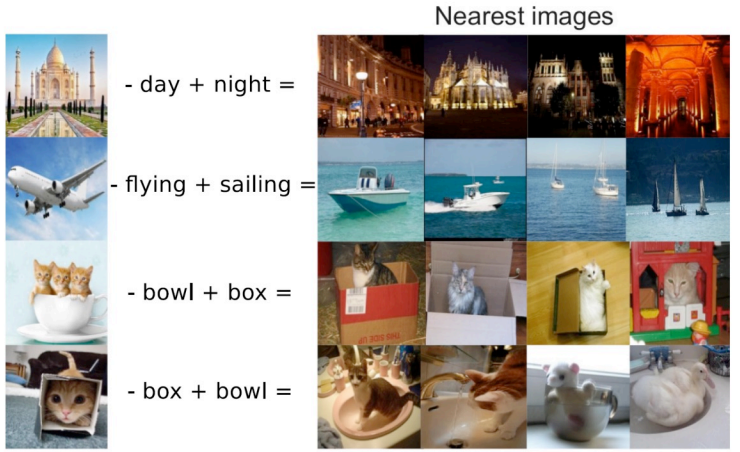


Two contrastive loss terms:

$$\max\{0, \alpha + sim(\boldsymbol{z}_L, \boldsymbol{z}_V^+) - sim(\boldsymbol{z}_L, \boldsymbol{z}_V^-)\}$$

$$+ \max\{0, \alpha + sim(\boldsymbol{z}_V, \boldsymbol{z}_L^+) - sim(\boldsymbol{z}_V, \boldsymbol{z}_L^-)\}$$

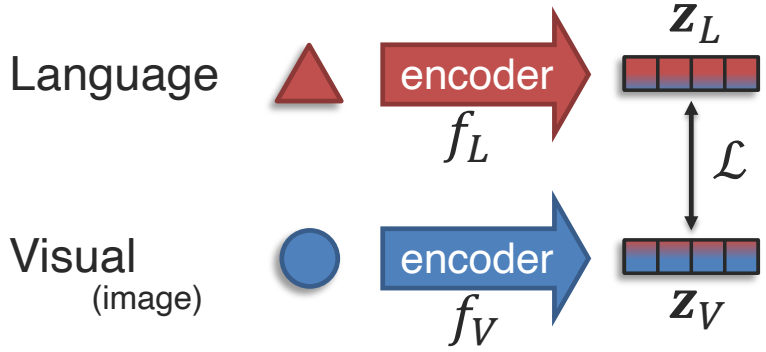




Kiros et al., Unifying Visual-Semantic Embeddings with Multimodal Neural Language Models, NIPS 2014

From slides by Louis-Philippe Morency

Example – CLIP (Contrastive Language–Image Pre-training)



Popular contrastive loss: InfoNCE

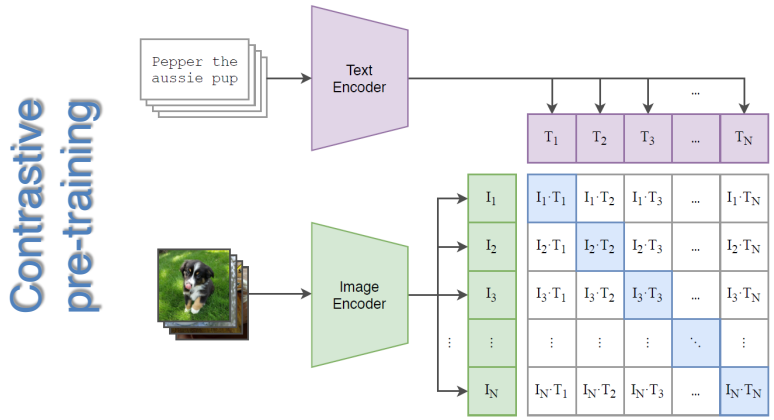
$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N} log \frac{sim(z_A^i, z_B^i)}{\sum_{j=1}^{N} sim(z_A^i, z_B^j)}$$

positive pairs

negative pairs and positive pairs

Similarity function can be cosine similarity

Positive and negative pairs:



➡ f_L and f_V are great encoders for language-vision tasks

➡ z_L and z_V are coordinated but not identical representation spaces

Radford et al., Learning Transferable Visual Models From Natural Language Supervision, arxiv 2021

From slides by Louis-Philippe Morency

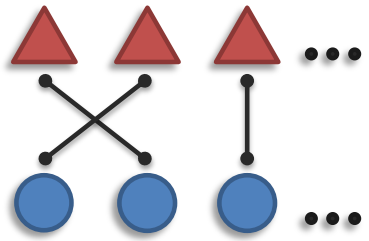
Alignment

Challenge 2: Alignment

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

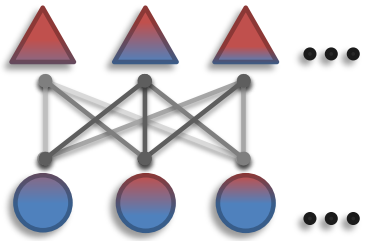
Sub-challenges:

Connections



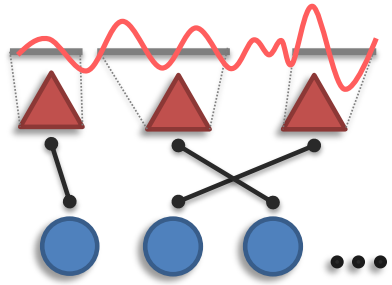
Explicit alignment
(e.g., grounding)

Aligned Representation



Implicit alignment
+ representation

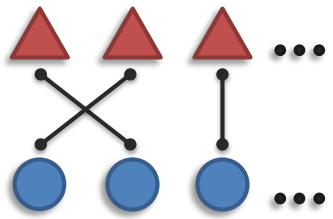
Segmentation



Granularity of
individual elements

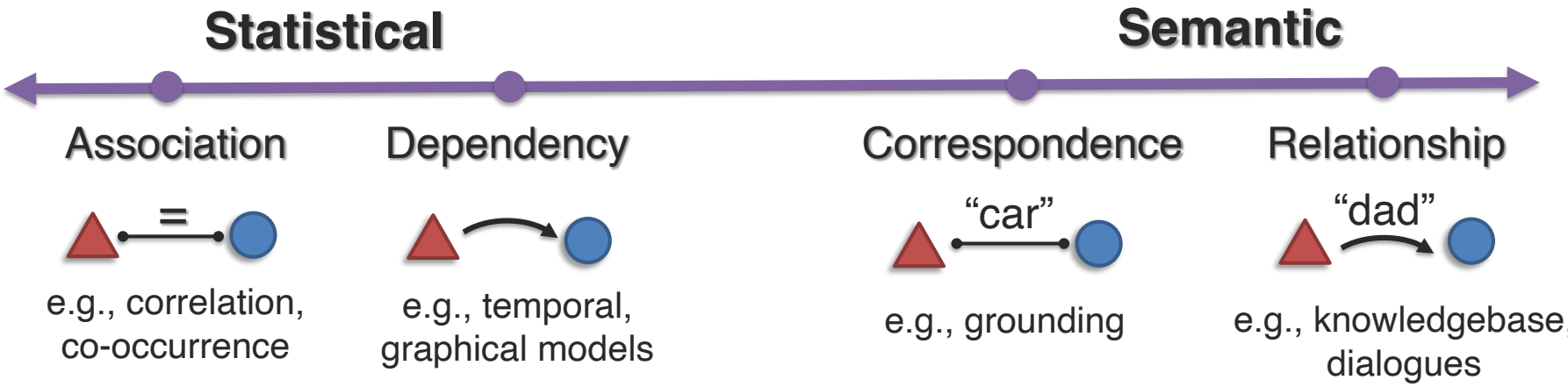
From slides by Louis-Philippe Morency

Sub-Challenge 2a: Connections



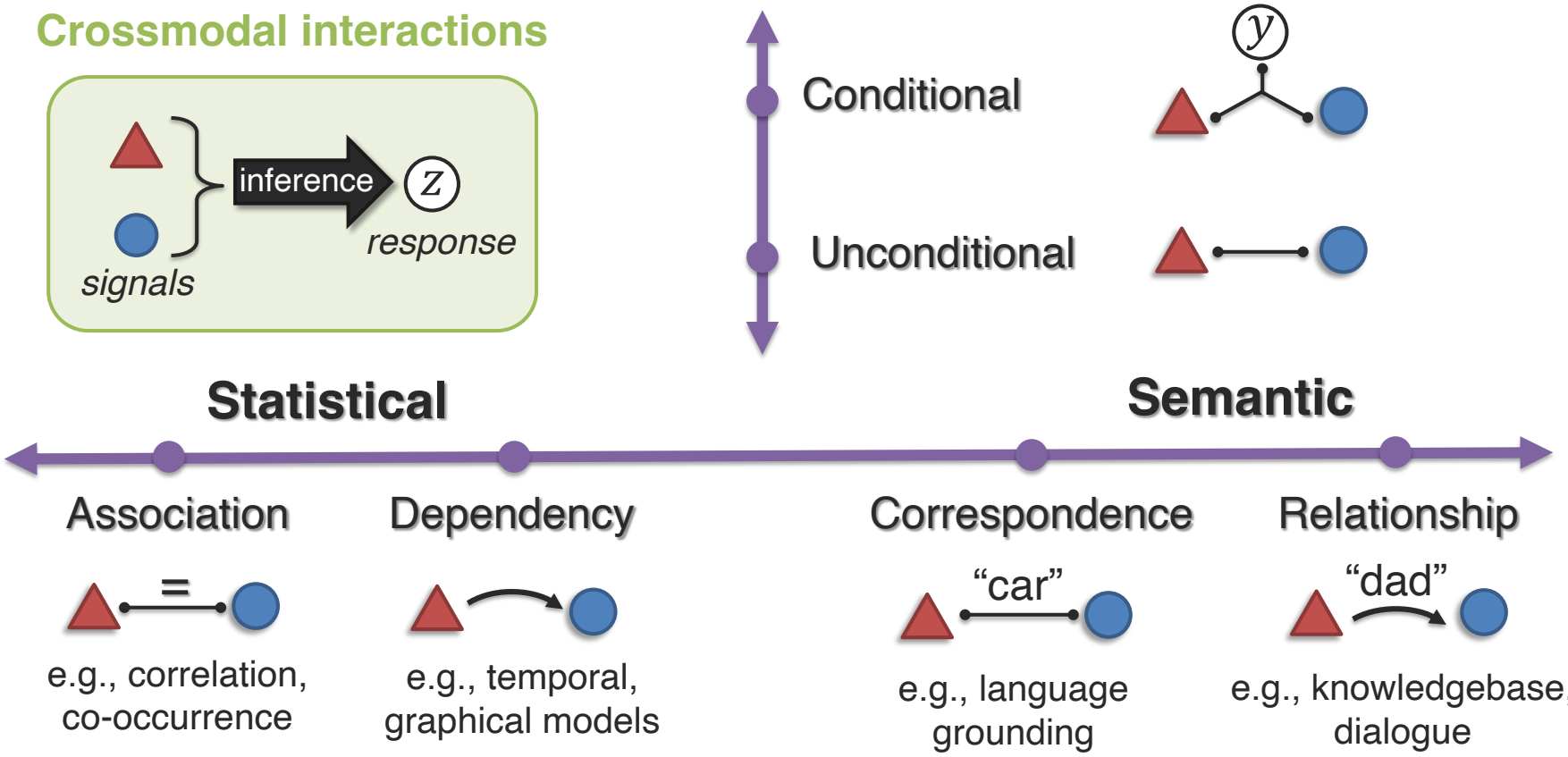
Definition: Identifying connections between elements of multiple modalities

Why should 2 elements be connected?



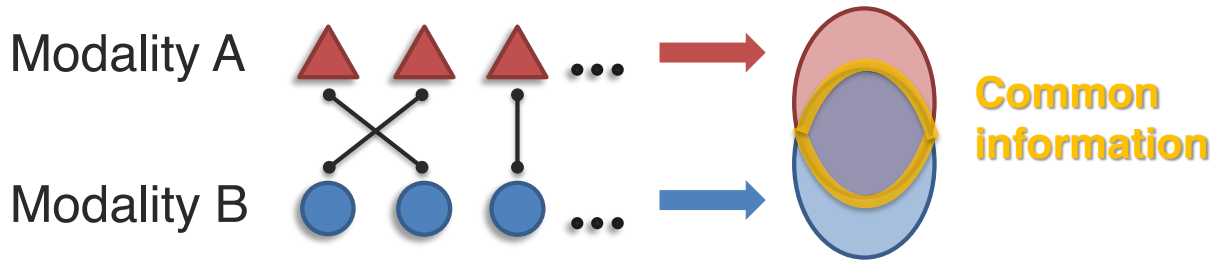
From slides by Louis-Philippe Morency

Sub-Challenge 2a: Connections



From slides by Louis-Philippe Morency

Sub-Challenge 2a: Connections



Common information

Example: language grounding

A woman reading newspaper

Statistical — **Semantic**

Association

e.g., correlation, co-occurrence

Dependency

e.g., temporal, graphical models

Correspondence

e.g., grounding

Relationship

e.g., knowledgebase, dialogue

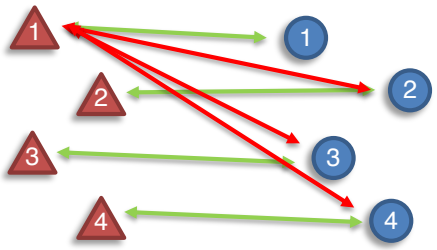
From slides by Louis-Philippe Morency

Language Grounding – Supervised Approach

Visual

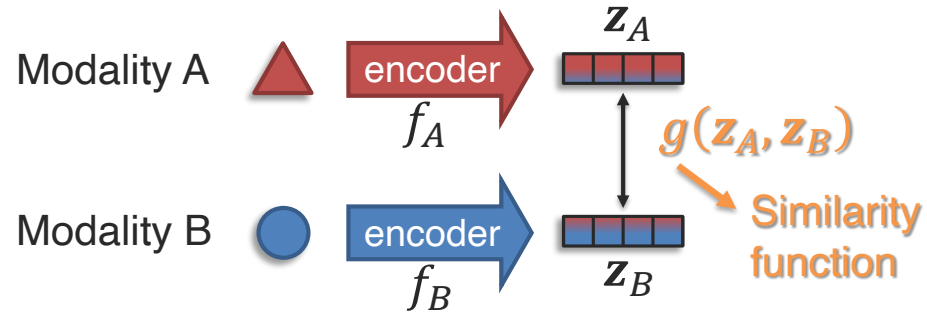
Language

Supervision: Paired data



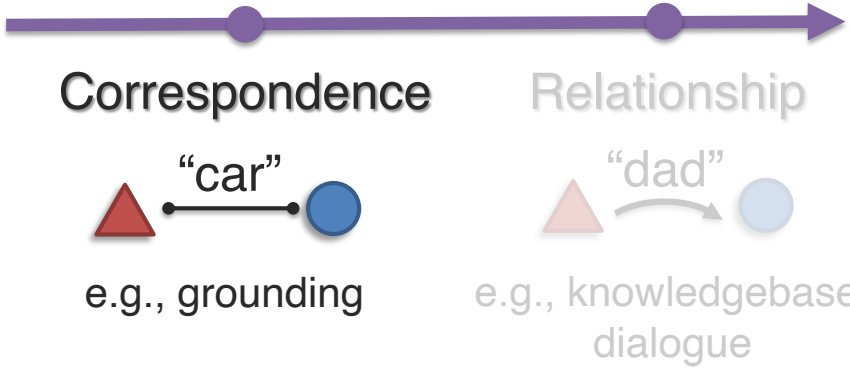
A woman reading newspaper

Learning coordinated representations:



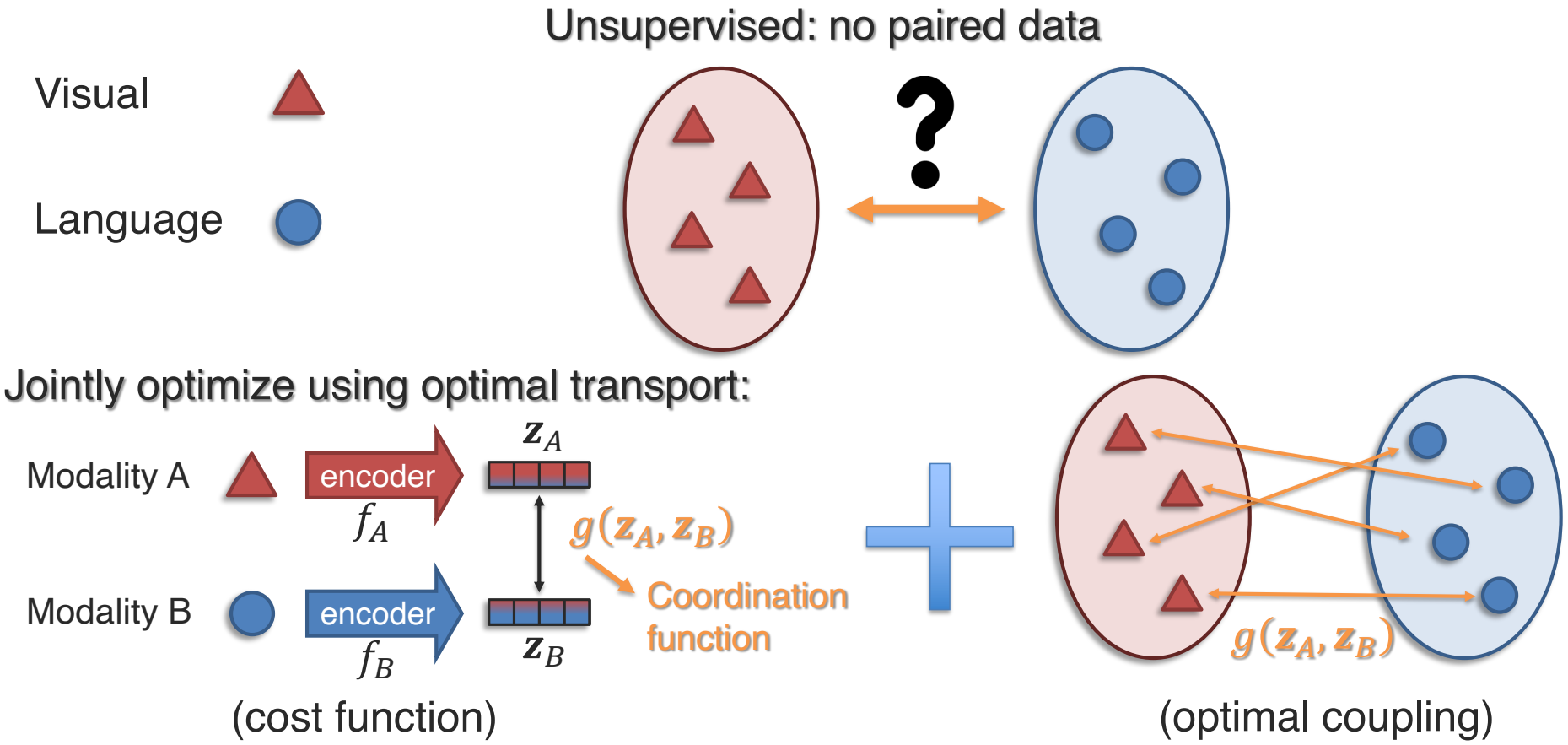
or contrastive learning

Semantic



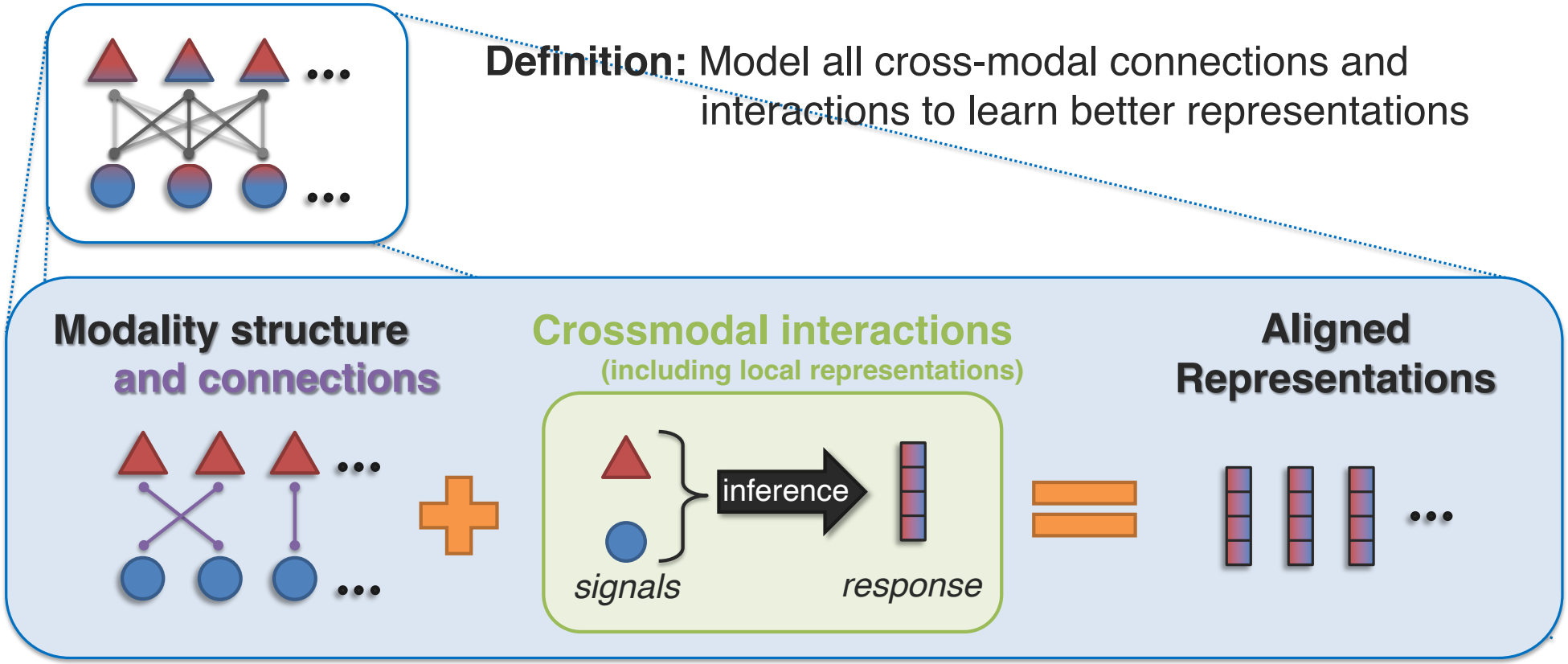
From slides by Louis-Philippe Morency

Language Grounding – Unsupervised Approach



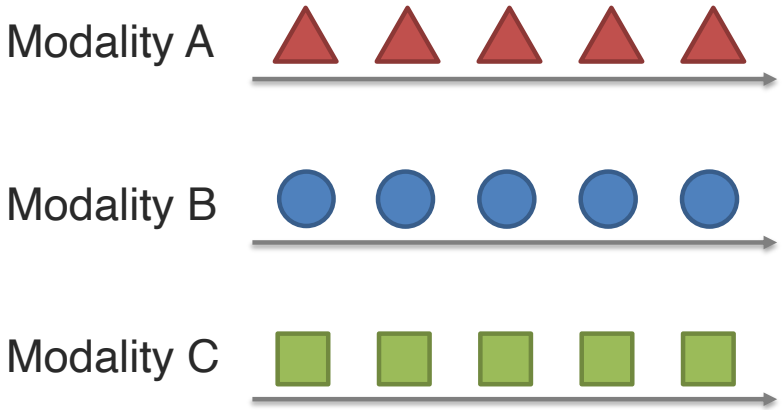
From slides by Louis-Philippe Morency

Sub-Challenge 2b: Aligned Representations



From slides by Louis-Philippe Morency

Aligned Representations – A Popular Approach



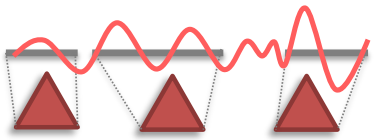
Assumptions:

① Segmented elements

objects

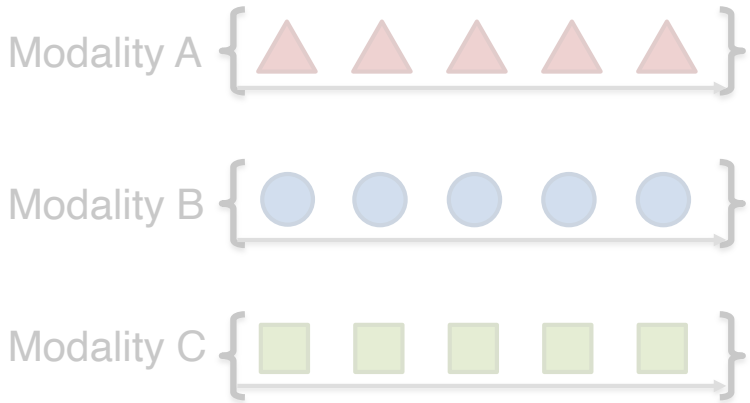
Image

➡ More about this assumption in the next sub-challenge



From slides by Louis-Philippe Morency

Aligned Representations – A Popular Approach



Assumptions:

① Segmented elements

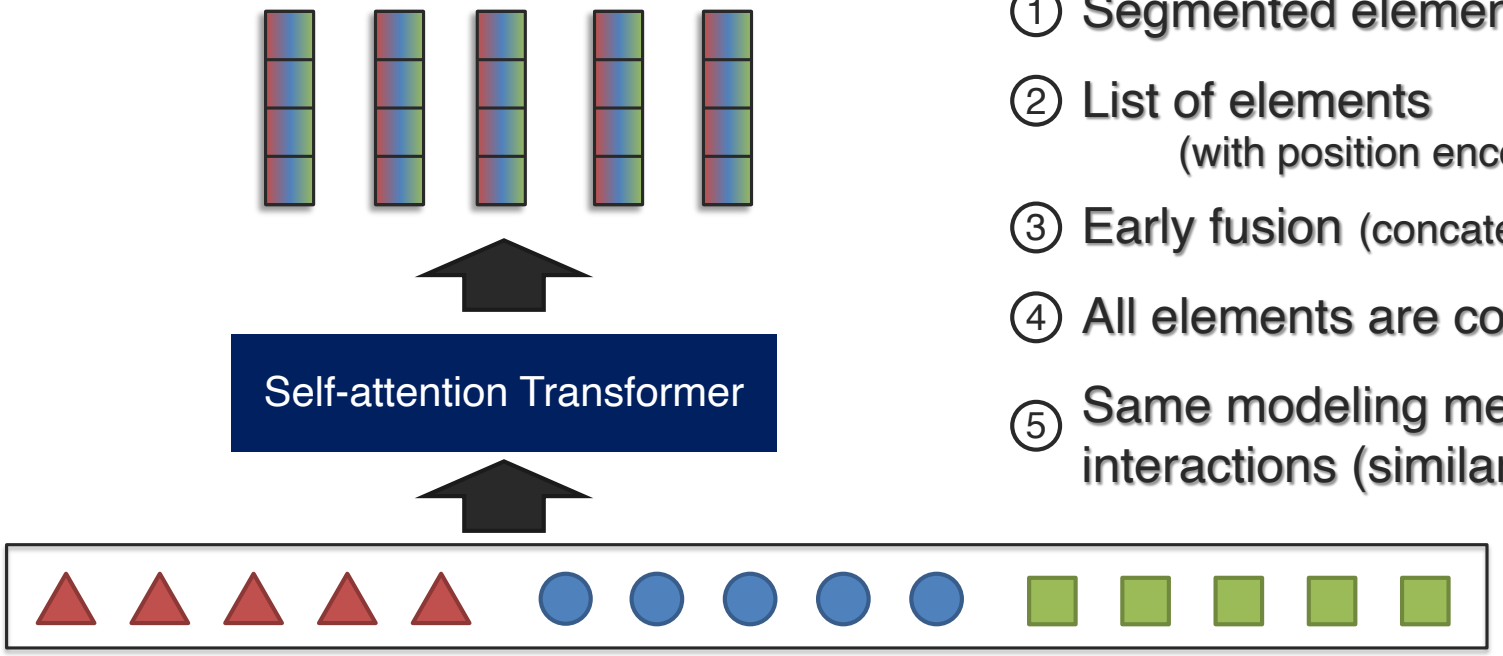
② List of elements
(with position encodings)

③ Early fusion (concatenated modalities)

Concatenate all modalities

From slides by Louis-Philippe Morency

Aligned Representations – A Popular Approach



Assumptions:

① Segmented elements

② List of elements
(with position encodings)

③ Early fusion (concatenated modalities)

④ All elements are connected

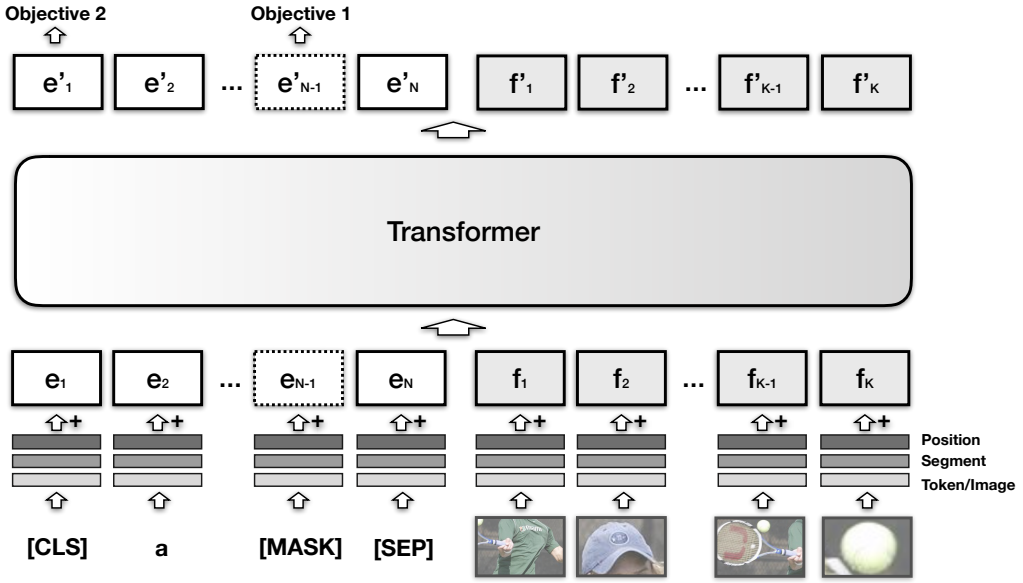
⑤ Same modeling method for all interactions (similarity kernels)

From slides by Louis-Philippe Morency

Aligned Representation – Early Fusion

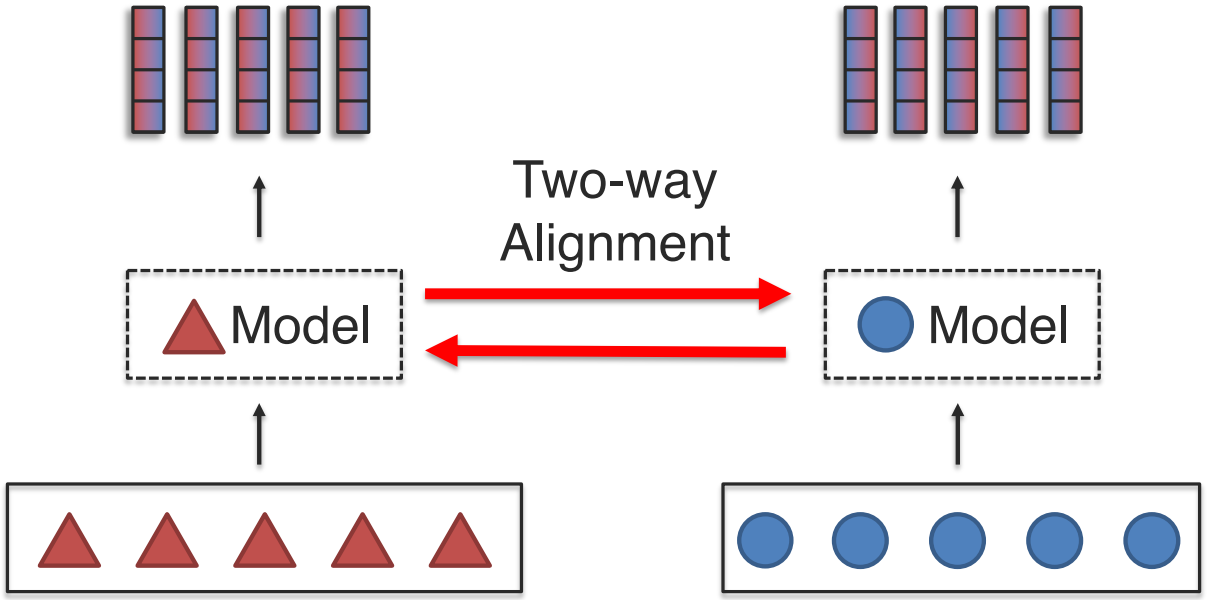
VisualBERT

A person hits a ball with a tennis racket



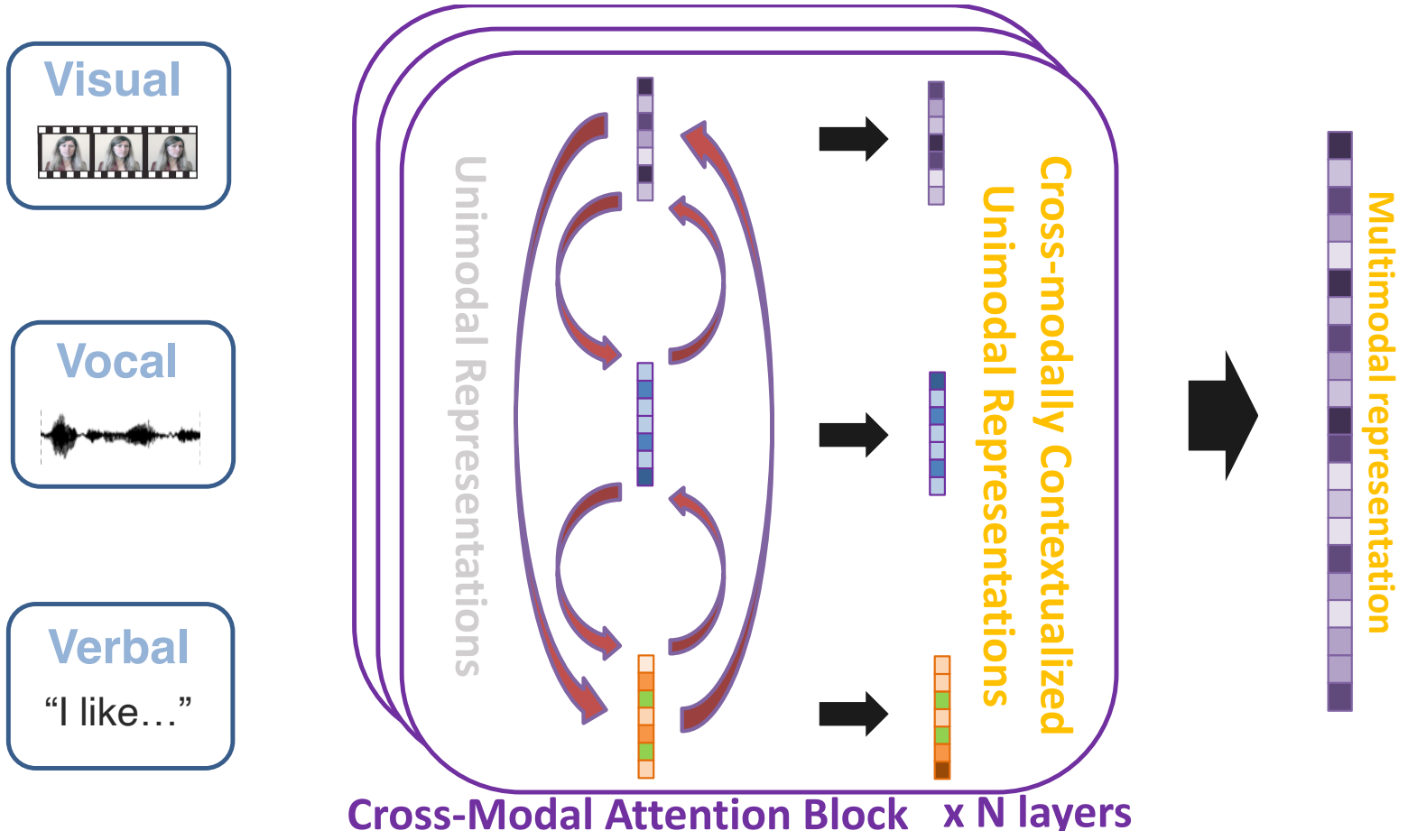
Li et al., VisualBERT: A Simple and Performant Baseline for Vision and Language, arxiv 2019

Aligned Representations – Two-Way Directional Alignment



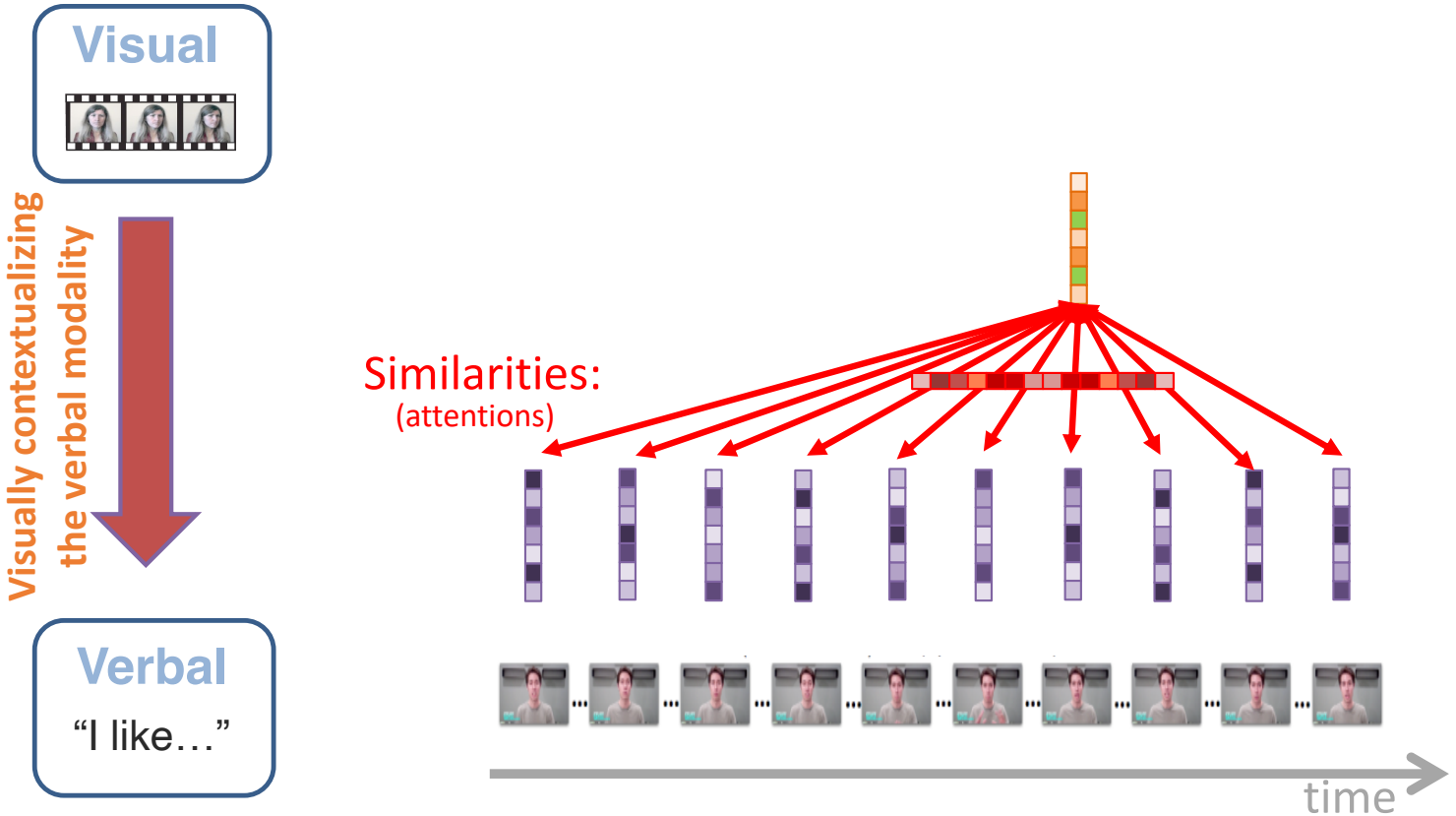
From slides by Louis-Philippe Morency

Multimodal Transformer – Pairwise Cross-Modal



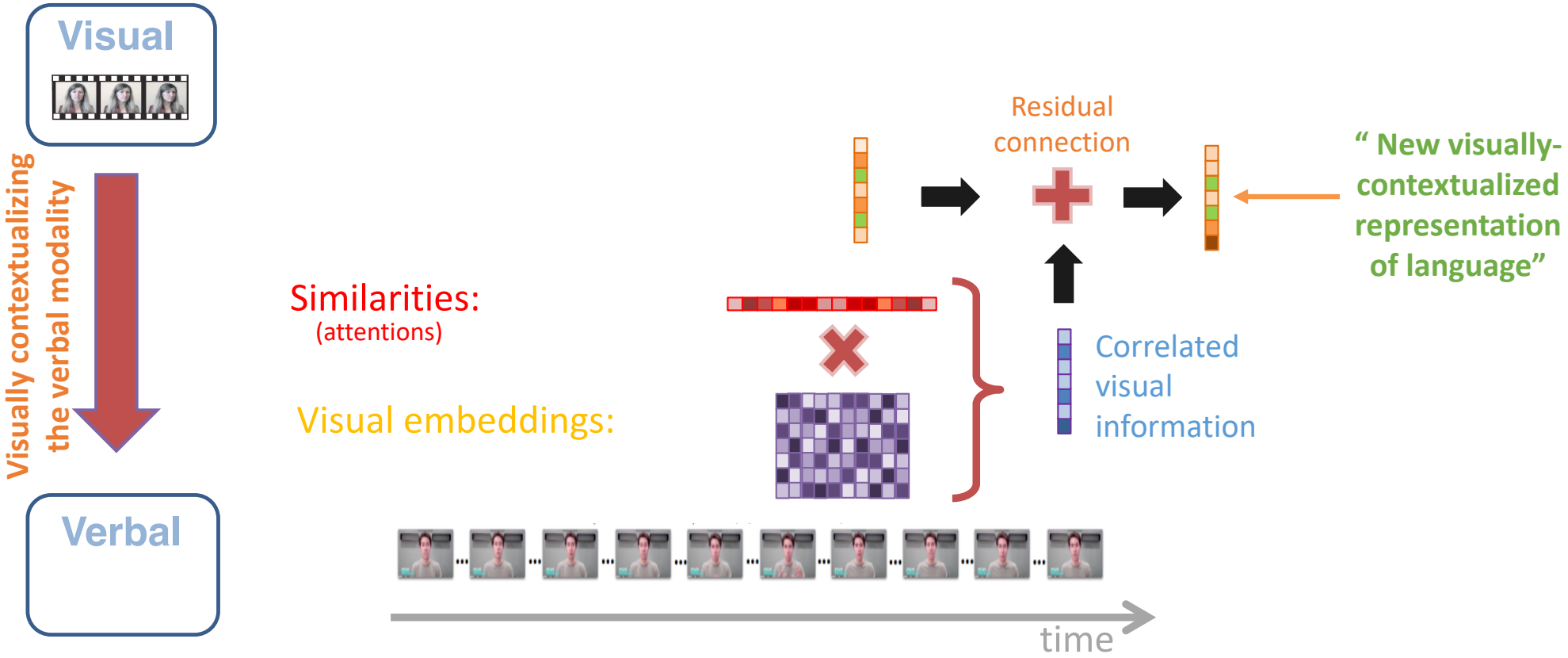
From slides by Louis-Philippe Morency

Cross-Modal Transformer Module ($\boldsymbol{V} \rightarrow \boldsymbol{L}$)



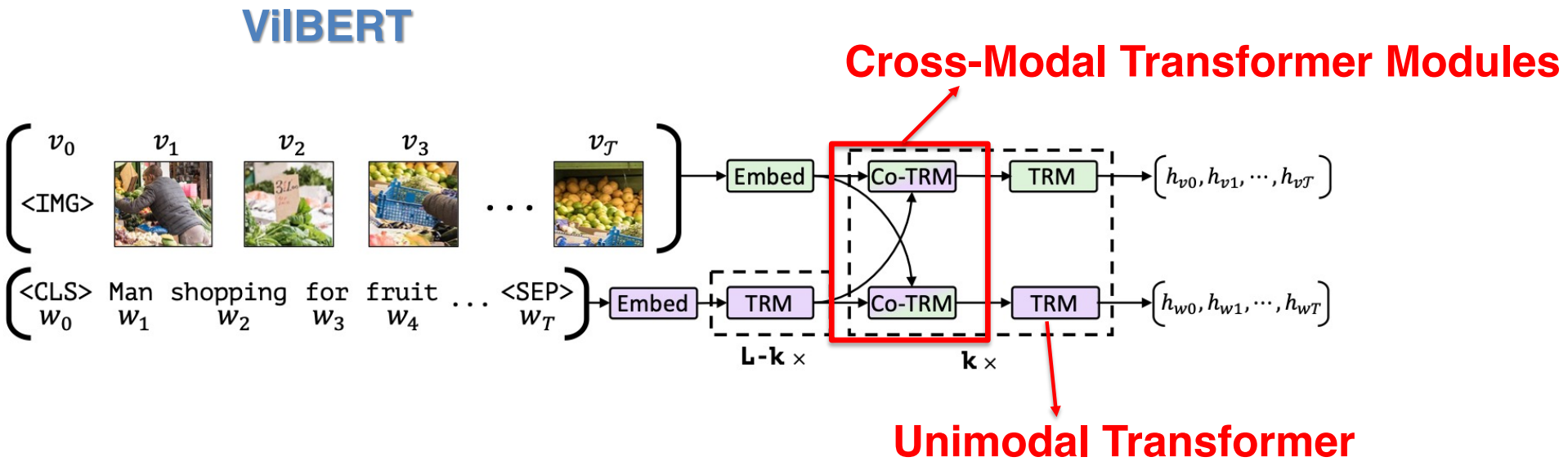
From slides by Louis-Philippe Morency

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



From slides by Louis-Philippe Morency

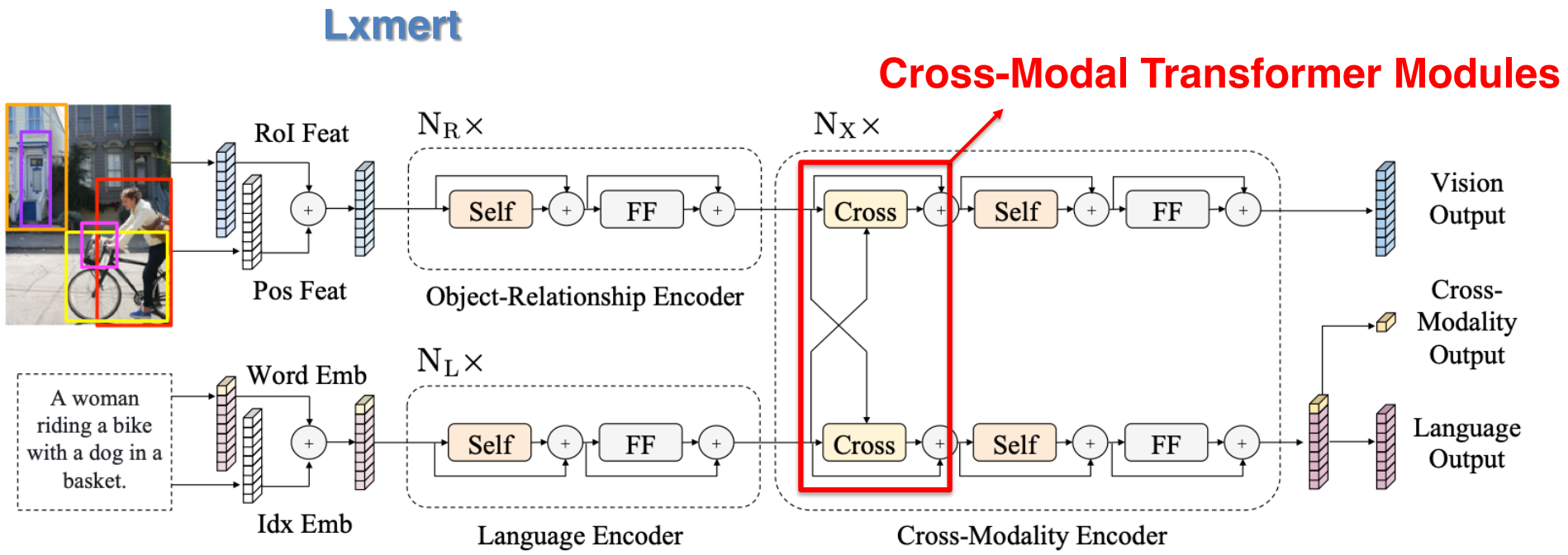
Example of Two-Way Directional Alignment



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

From slides by Louis-Philippe Morency

Example of Two-Way Directional Alignment



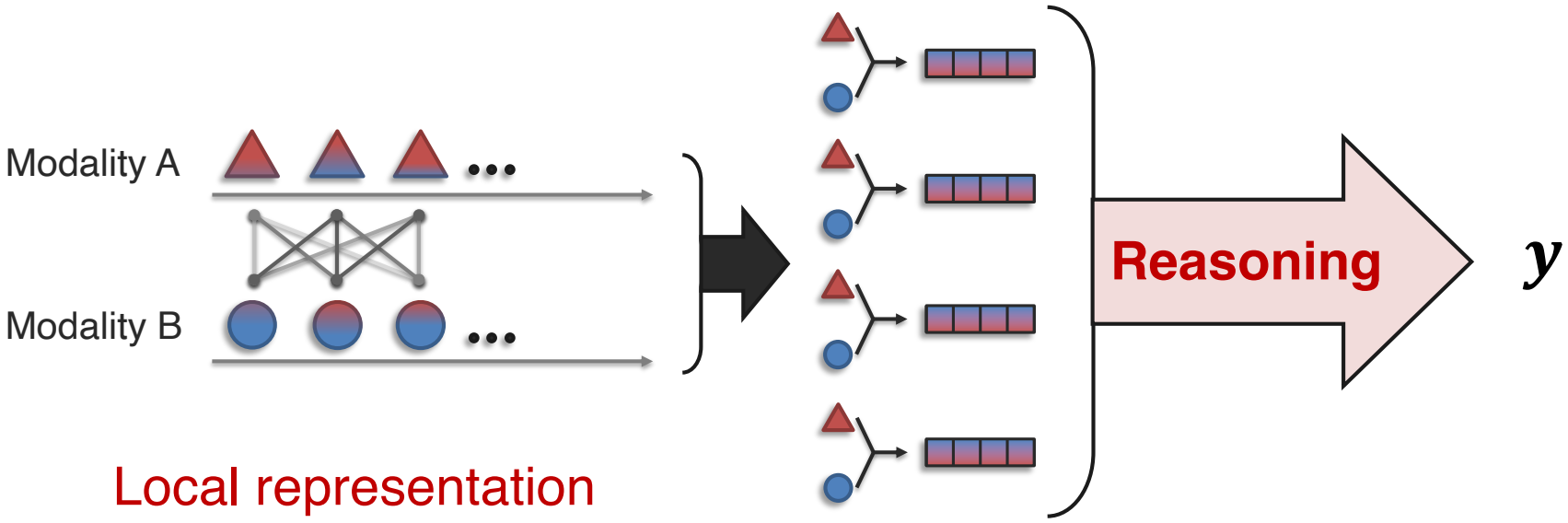
Tan, Hao, and Mohit Bansal. "Lxmert: Learning cross-modality encoder representations from transformers." *arXiv* (August 20, 2019).

From slides by Louis-Philippe Morency

Reasoning

Reasoning

Definition: Combining knowledge, usually through multiple inferential steps, exploiting multimodal alignment and problem structure.



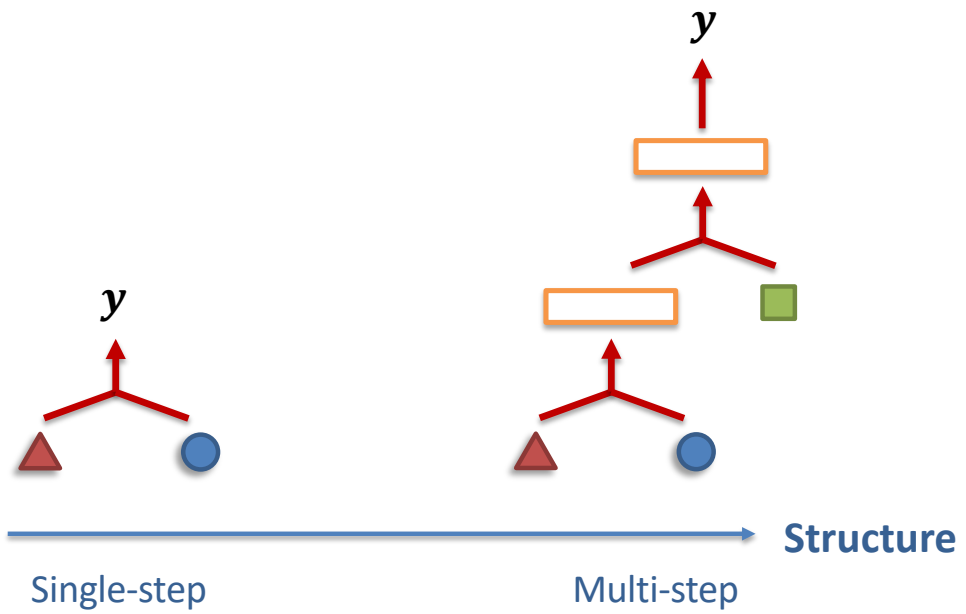
Local representation

+ Aligned representation

From slides by Louis-Philippe Morency

Sub-Challenge 3a: Structure Modeling

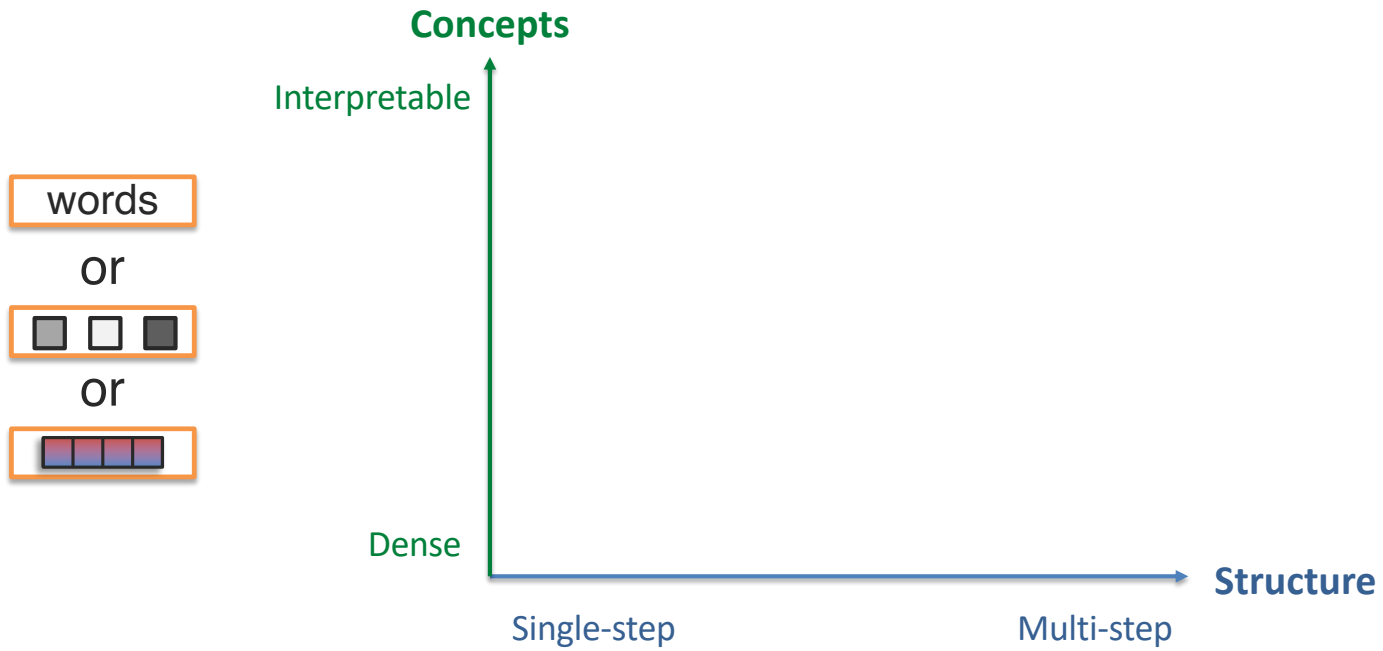
Definition: Defining or learning the relationships over which reasoning occurs.



From slides by Louis-Philippe Morency

Sub-Challenge 3b: Intermediate Concepts

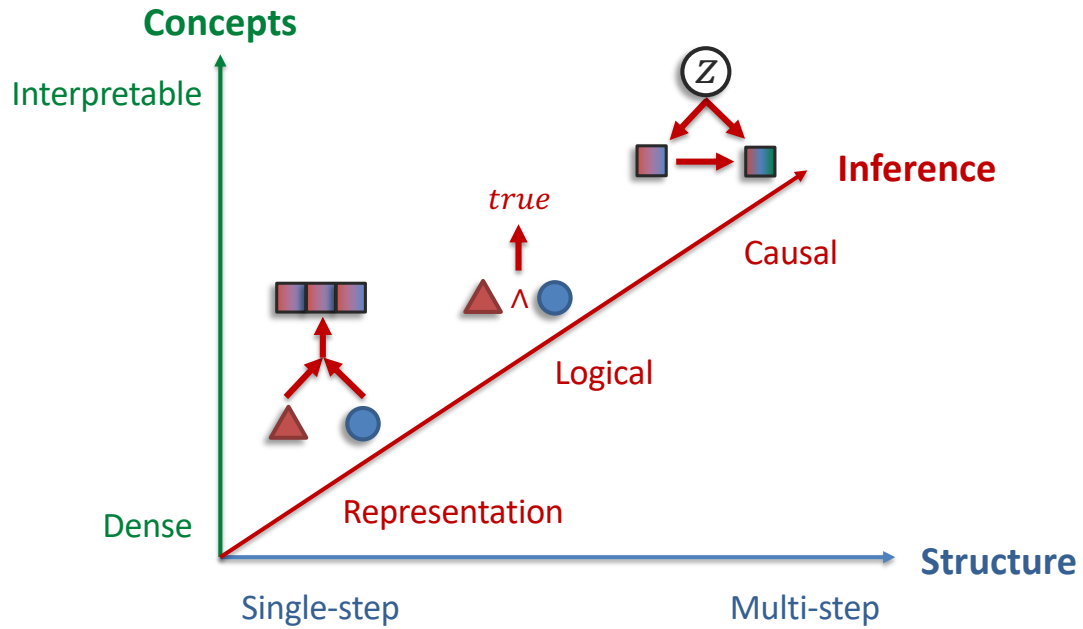
Definition: The parameterization of individual multimodal concepts in the reasoning process.



From slides by Louis-Philippe Morency

Sub-Challenge 3c: Inference Paradigm

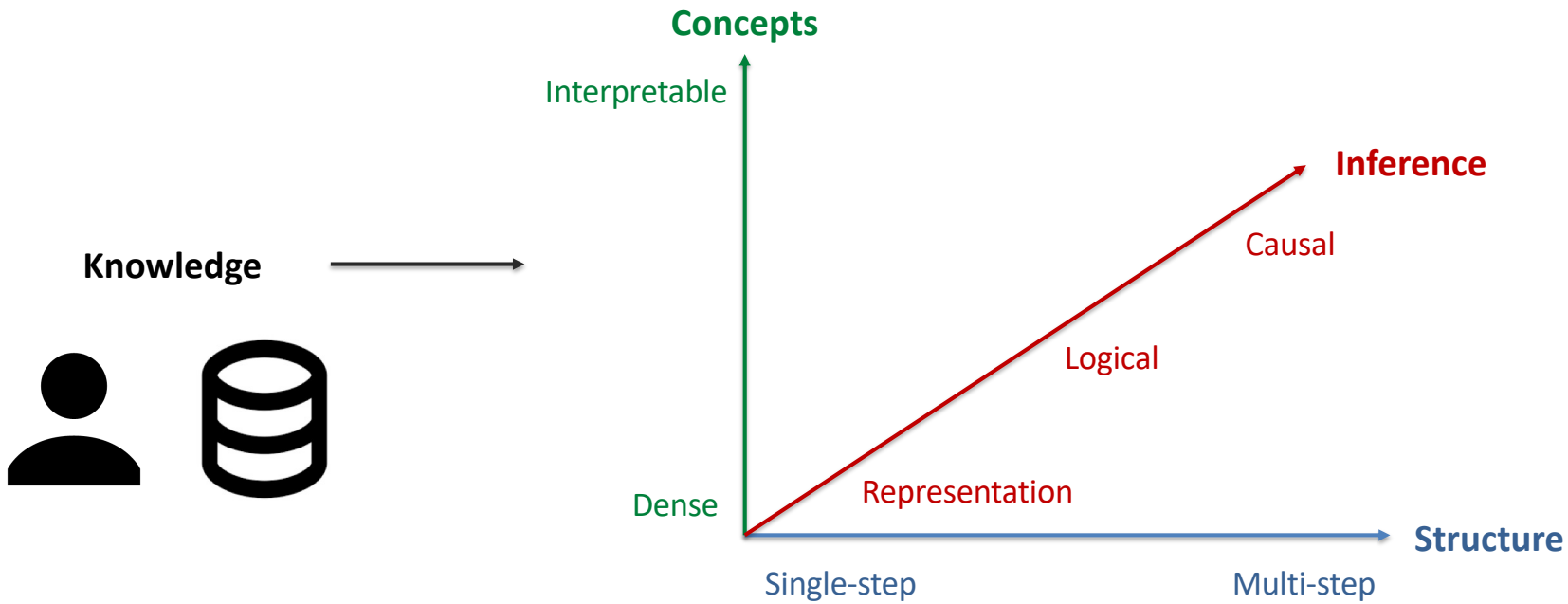
Definition: How increasingly abstract concepts are inferred from individual multimodal evidence



From slides by Louis-Philippe Morency

Sub-Challenge 3d: External Knowledge

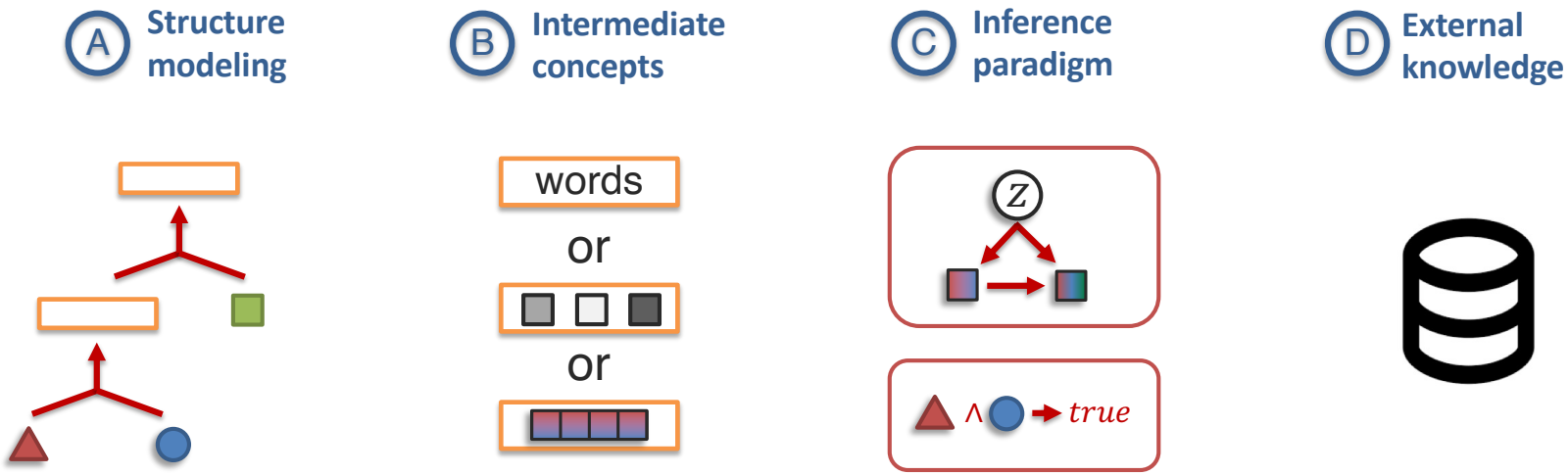
Definition: Leveraging external knowledge in the study of structure, concepts, and inference.



From slides by Louis-Philippe Morency

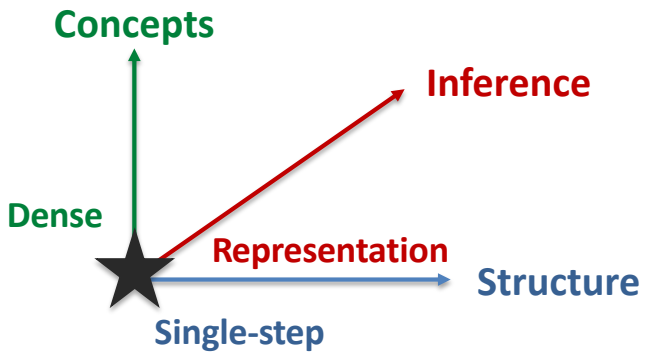
Reasoning

Definition: Combining knowledge, usually through multiple inferential steps, exploiting multimodal alignment and problem structure.



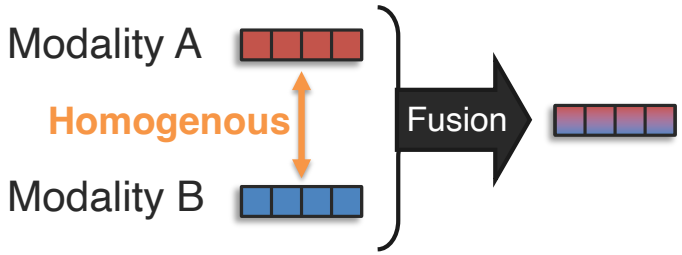
From slides by Louis-Philippe Morency

Reasoning

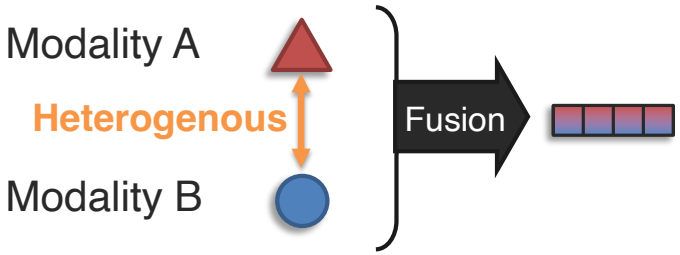


Recall representation fusion!

Basic fusion:



Complex fusion:

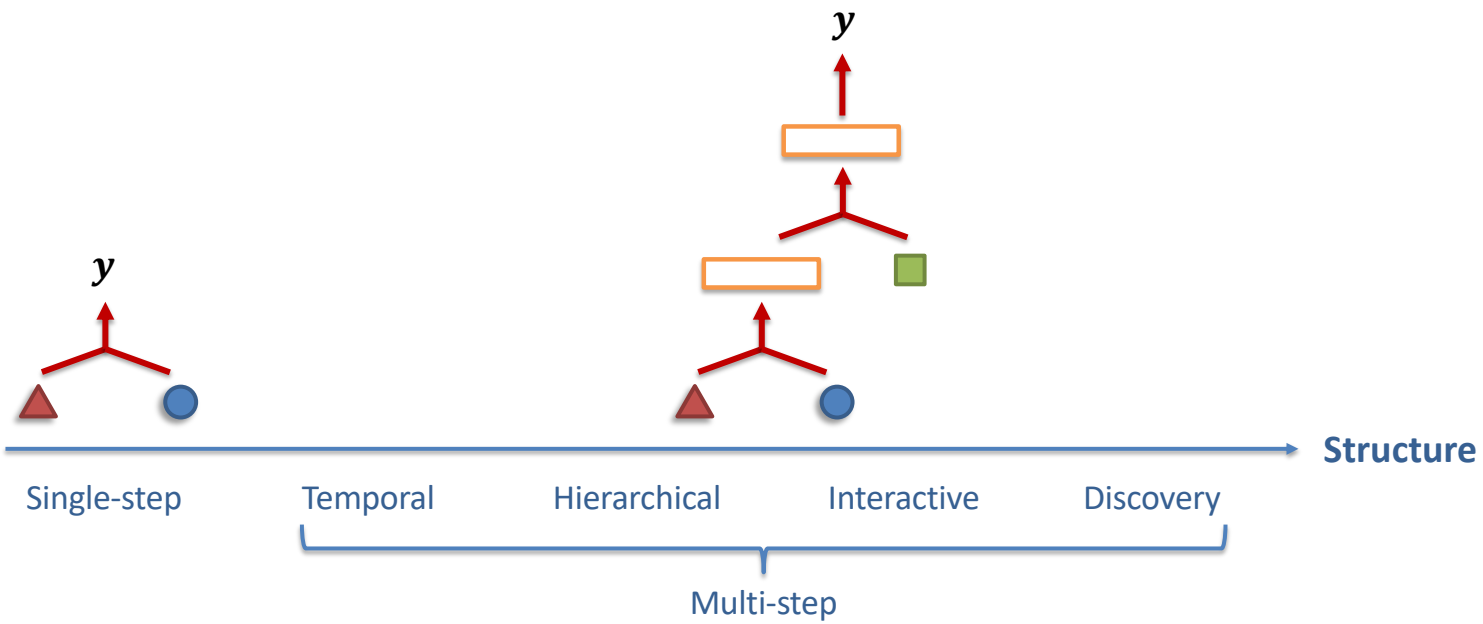


Ideas also apply here, but can we be explicitly interpretable and robust?

From slides by Louis-Philippe Morency

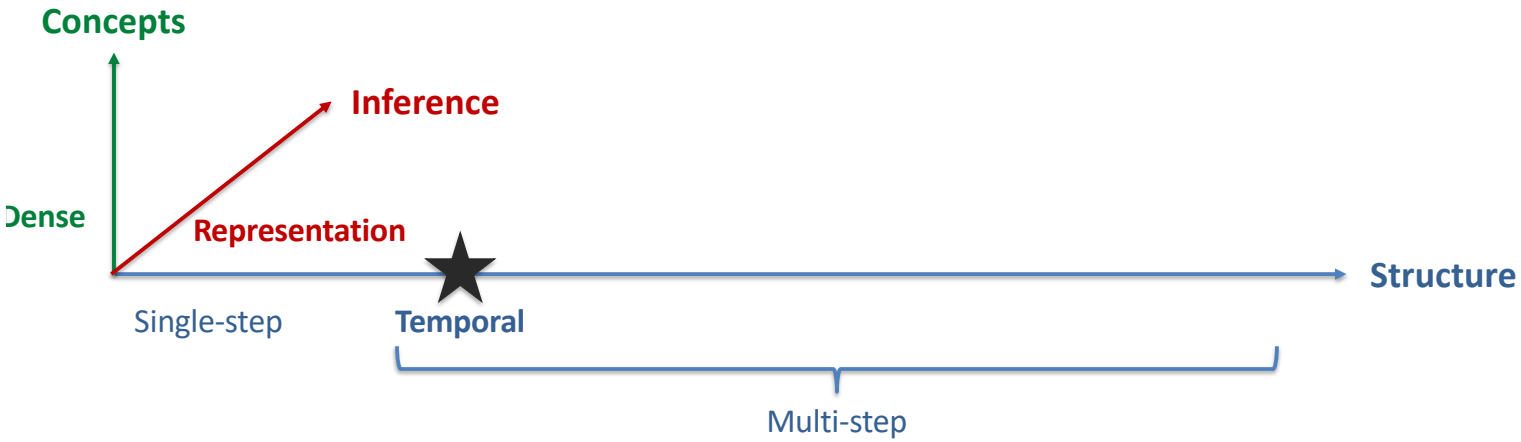
Sub-Challenge 3a: Structure Modeling

Definition: Defining or learning the relationships over which composition occurs.



From slides by Louis-Philippe Morency

Sub-Challenge 3a: Structure Modeling

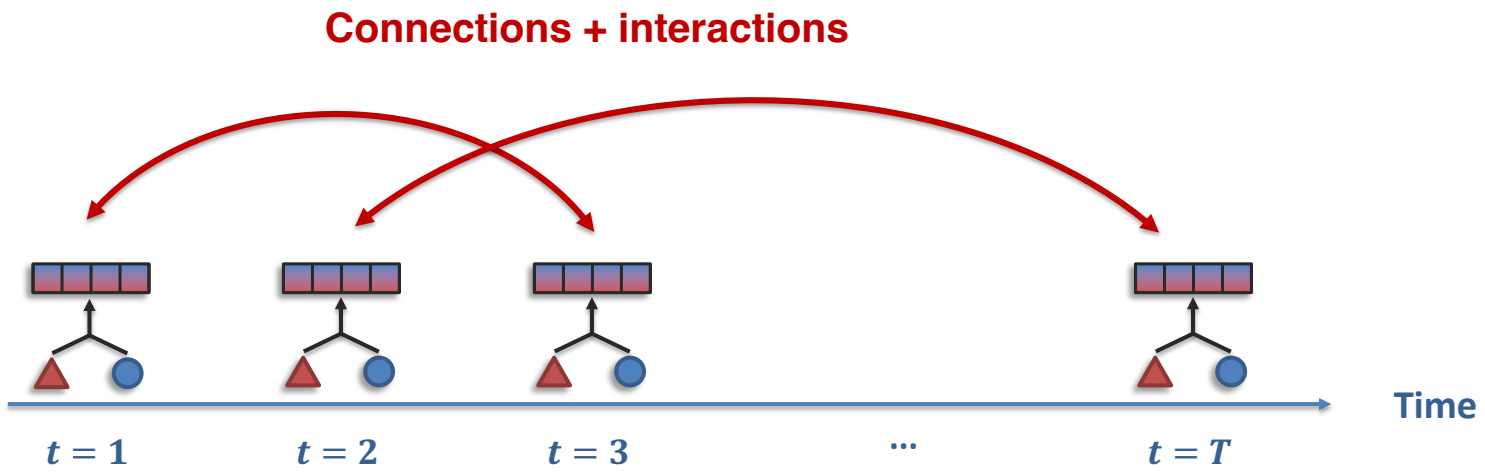


From slides by Louis-Philippe Morency

Temporal Structure

Temporal structure in multi-view sequences

How can we capture cross-modal interactions across time?

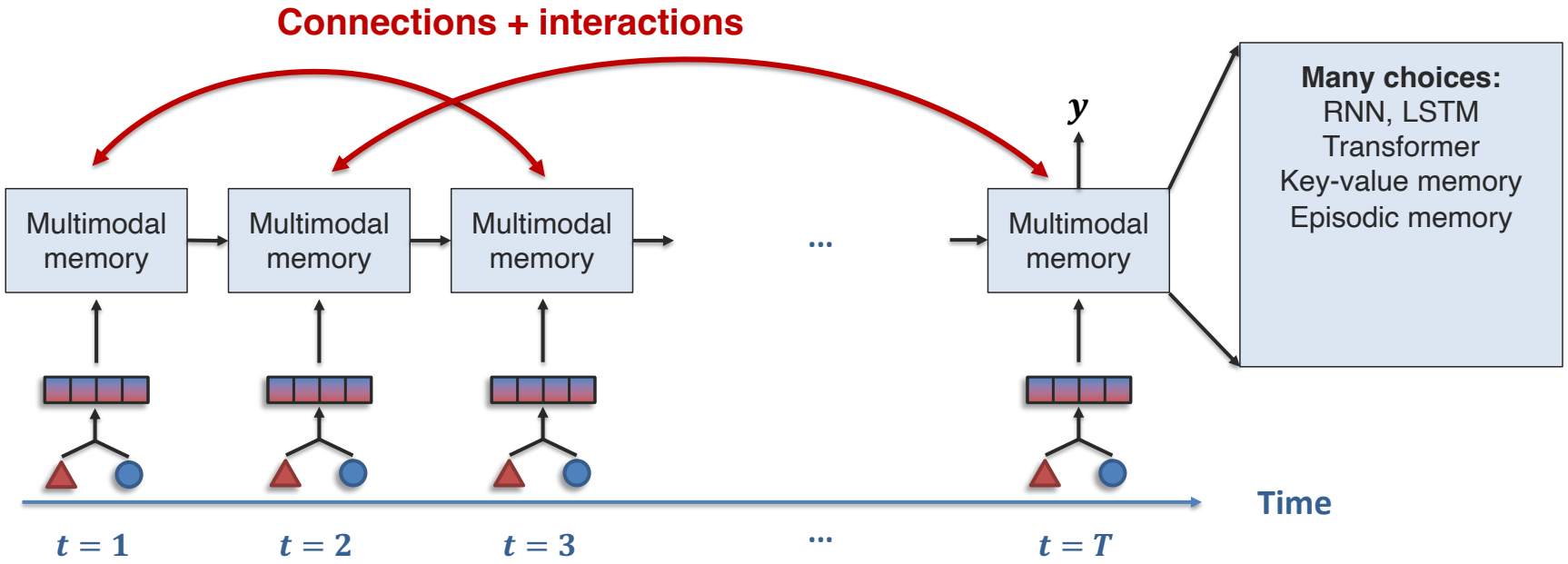


From slides by Louis-Philippe Morency

Temporal Structure

Temporal structure in multi-view sequences

Key ideas: memory to capture cross-modal interactions across time

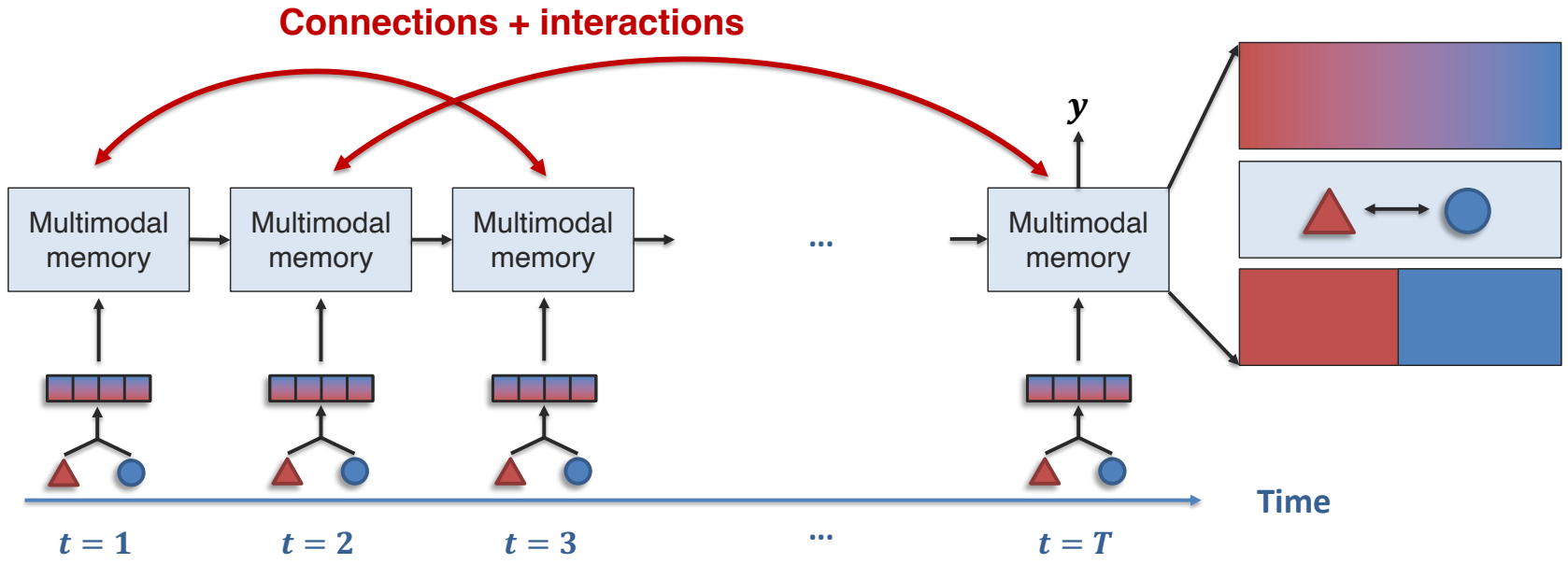


From slides by Louis-Philippe Morency

Temporal Structure

Temporal structure in multi-view sequences

Structuring multimodal memory: ideas from representation fusion, coordination, and fission



[Rajagopalan et al., Extending Long Short-Term Memory for Multi-View Structured Learning. ECCV 2016]

From slides by Louis-Philippe Morency

Temporal Structure

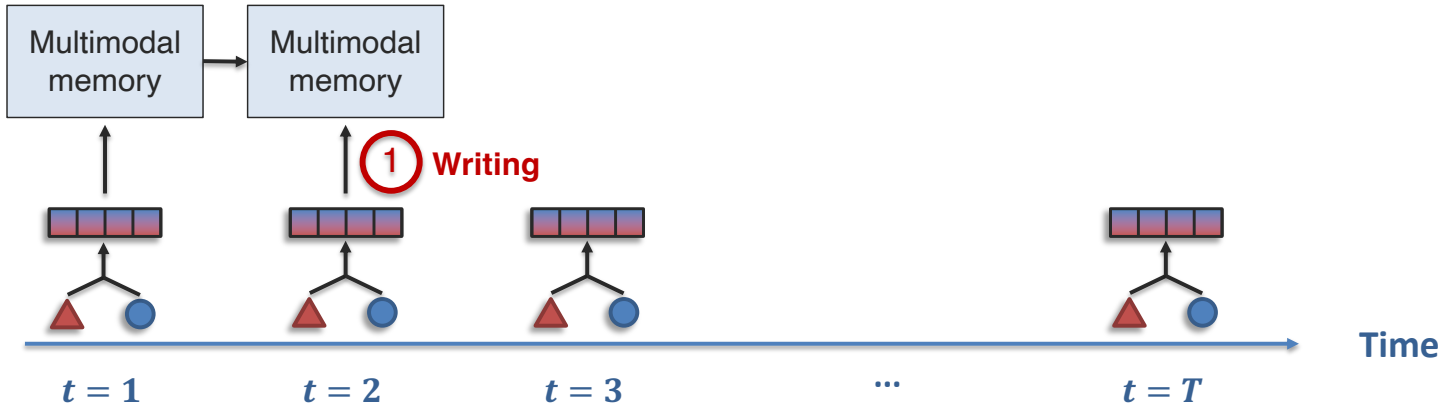
Temporal structure in multi-view sequences

1 **Writing** Coordination function measuring **similarity** between feature and memory to weight feature:

$$w_t = g(\boldsymbol{z}_t, \boldsymbol{M}_t) = \boldsymbol{z}_t \cdot \boldsymbol{M}_t$$

Write = $w_t \boldsymbol{z}_t$

Recall representation coordination, attention models, Transformers, LSTMs etc.



[Wang et al., Multimodal Memory Modelling for Video Captioning, CVPR 2018]

From slides by Louis-Philippe Morency

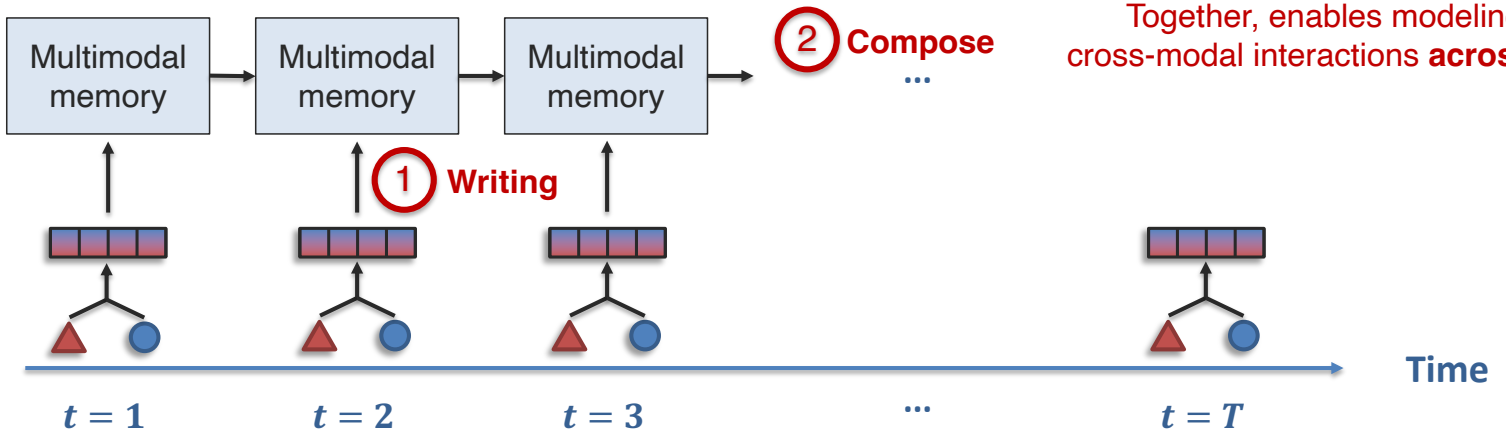
Temporal Structure

Temporal structure in multi-view sequences

2 **Compose** Weighted function to **compose** previous memory and new addition

$$\boldsymbol{M}_{t+1} = (1 - \boldsymbol{\alpha}_t)\boldsymbol{M}_t + \boldsymbol{\alpha}_t \text{Write}$$

$\boldsymbol{\alpha}_t$ learnable in LSTM/RNN/Transformers, or similarity function in parameterized memory.



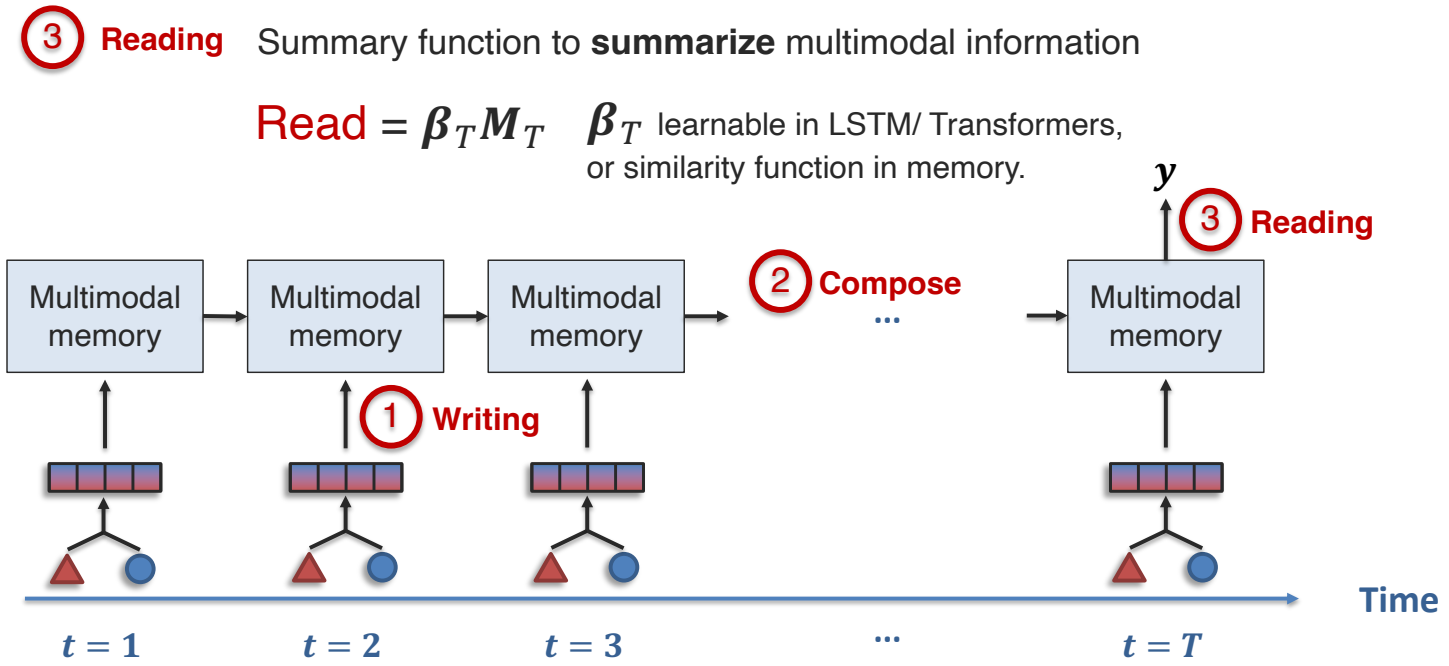
Together, enables modeling of cross-modal interactions **across time.**

[Xiong et al., Dynamic Memory Networks for Visual and Textual Question Answering. arXiv 2016]

From slides by Louis-Philippe Morency

Temporal Structure

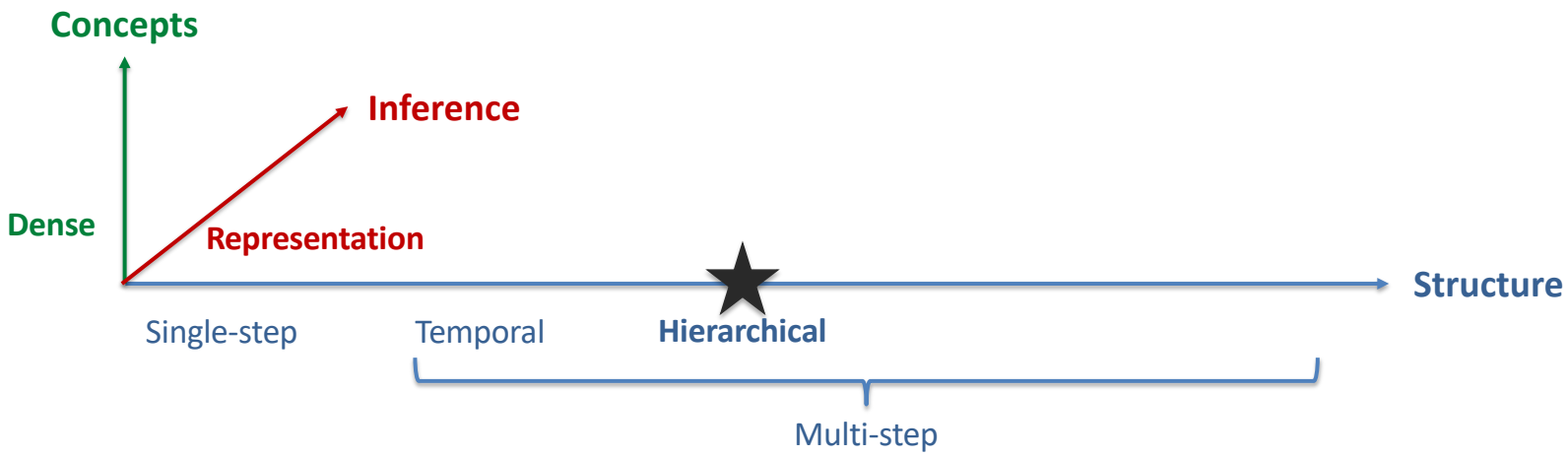
Temporal structure in multi-view sequences



[Hazarika et al., ICON: Interactive Conversational Memory Network for Multimodal Emotion Detection. EMNLP 2018]

From slides by Louis-Philippe Morency

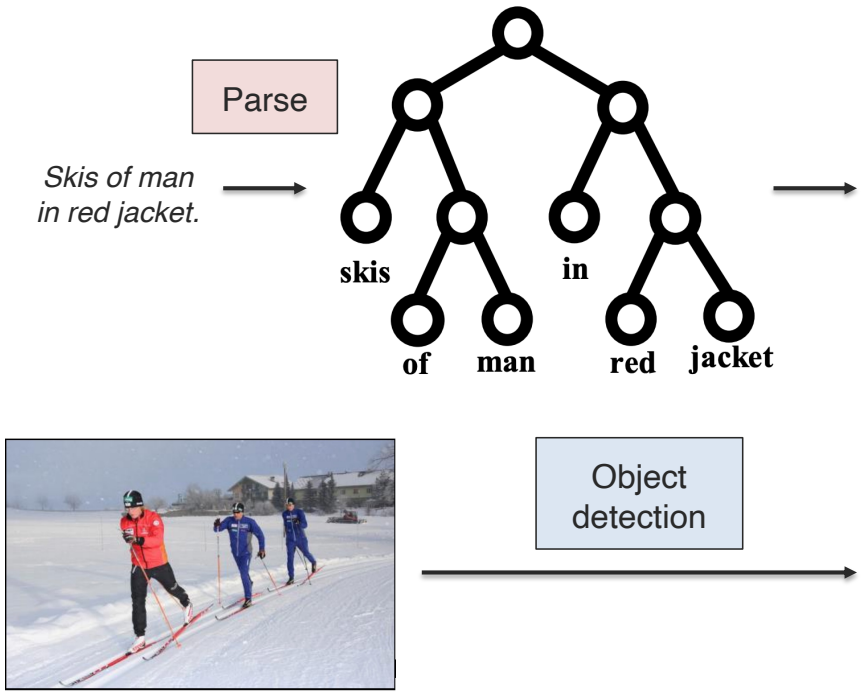
Sub-Challenge 3a: Structure Modeling



From slides by Louis-Philippe Morency

Hierarchical Structure

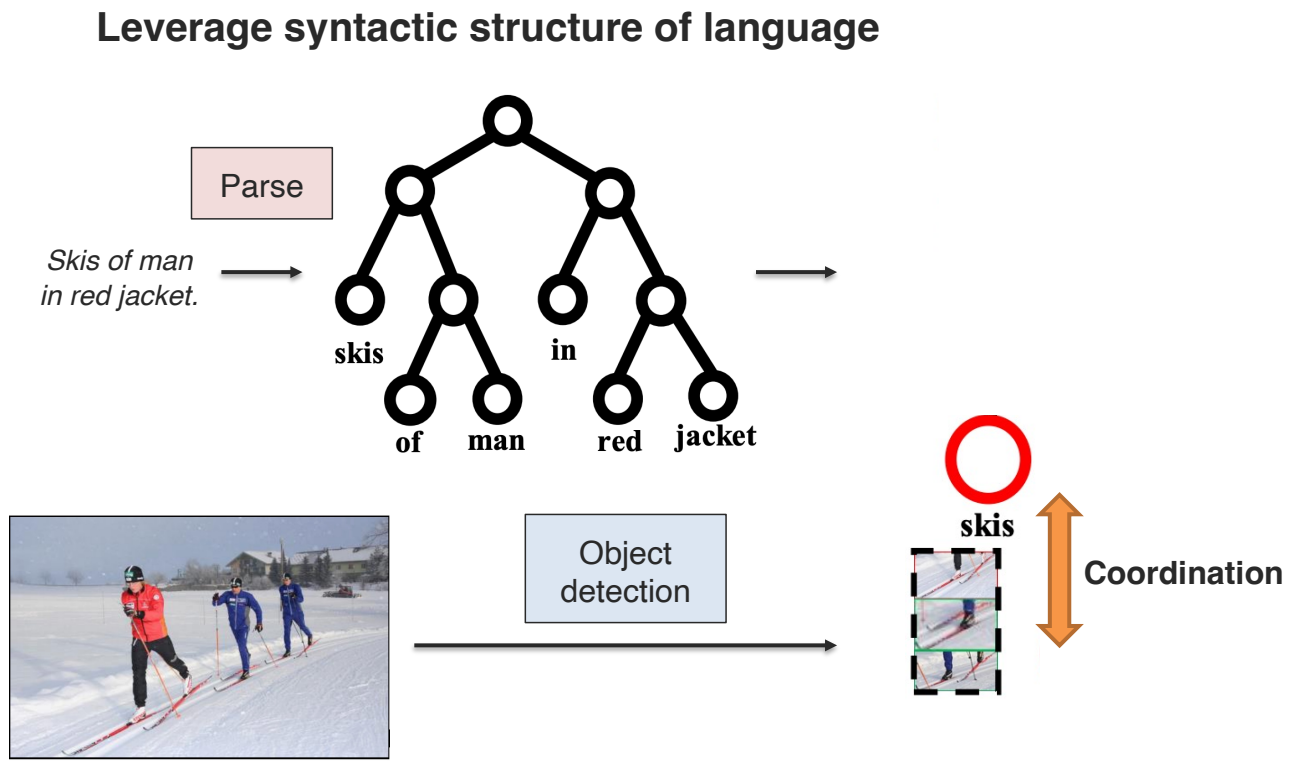
Leverage syntactic structure of language



[Hong et al., Learning to Compose and Reason with Language Tree Structures for Visual Grounding. IEEE TPAMI 2019]

From slides by Louis-Philippe Morency

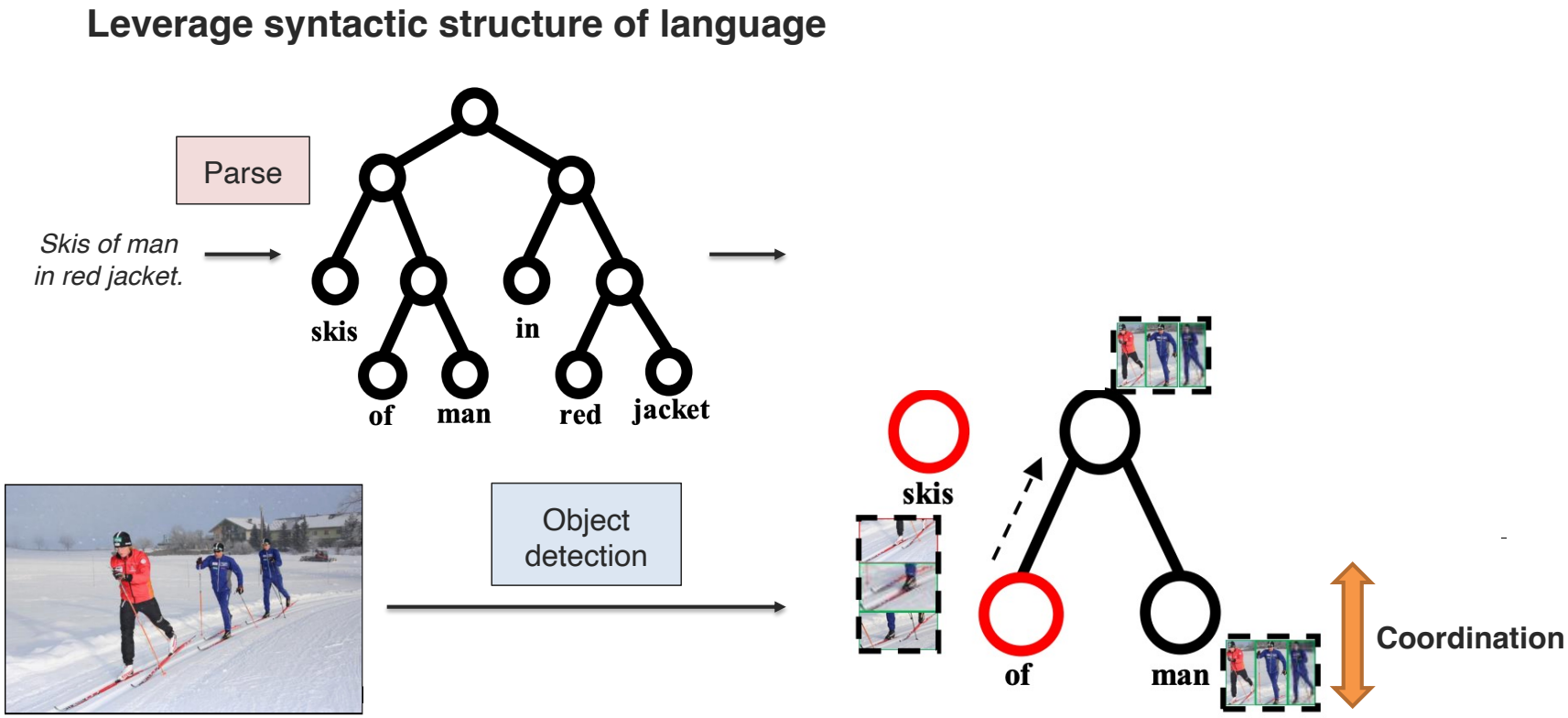
Hierarchical Structure



[Hong et al., Learning to Compose and Reason with Language Tree Structures for Visual Grounding. IEEE TPAMI 2019]

From slides by Louis-Philippe Morency

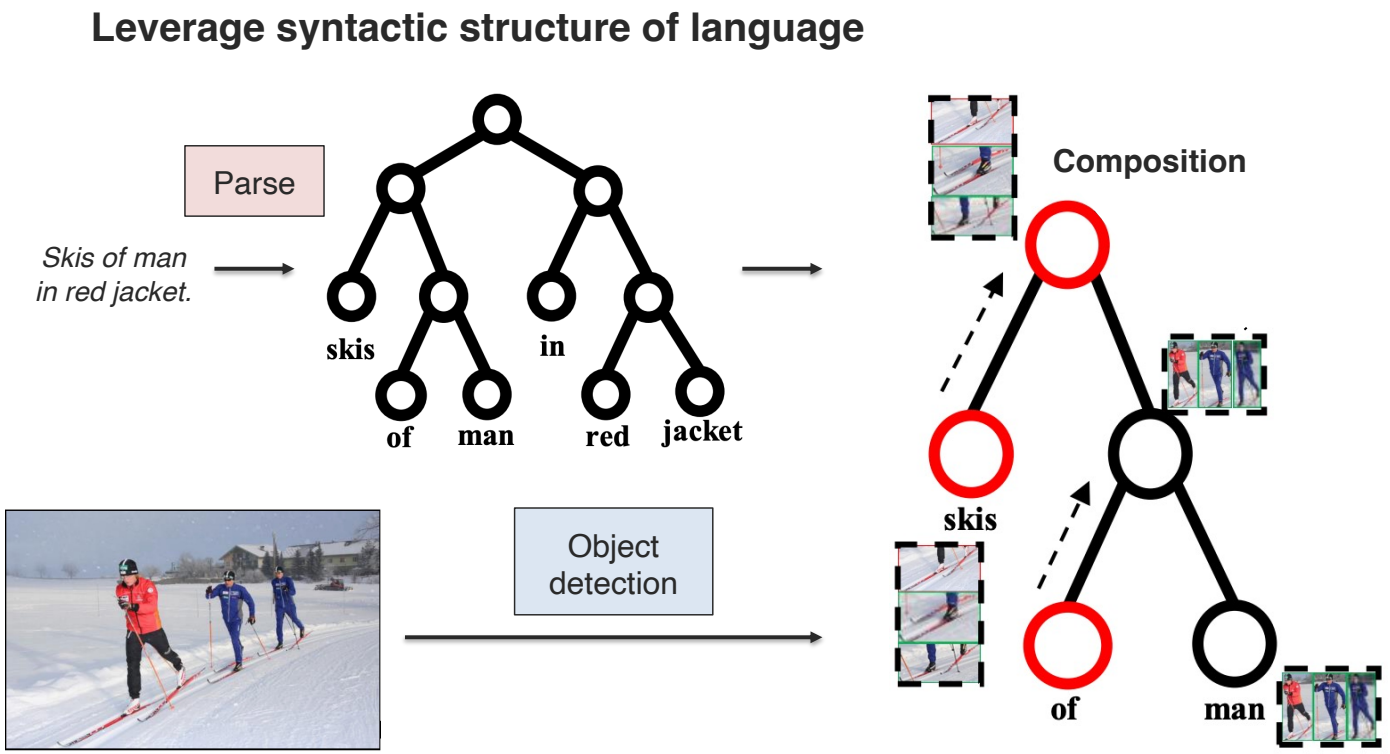
Hierarchical Structure



[Hong et al., Learning to Compose and Reason with Language Tree Structures for Visual Grounding. IEEE TPAMI 2019]

From slides by Louis-Philippe Morency

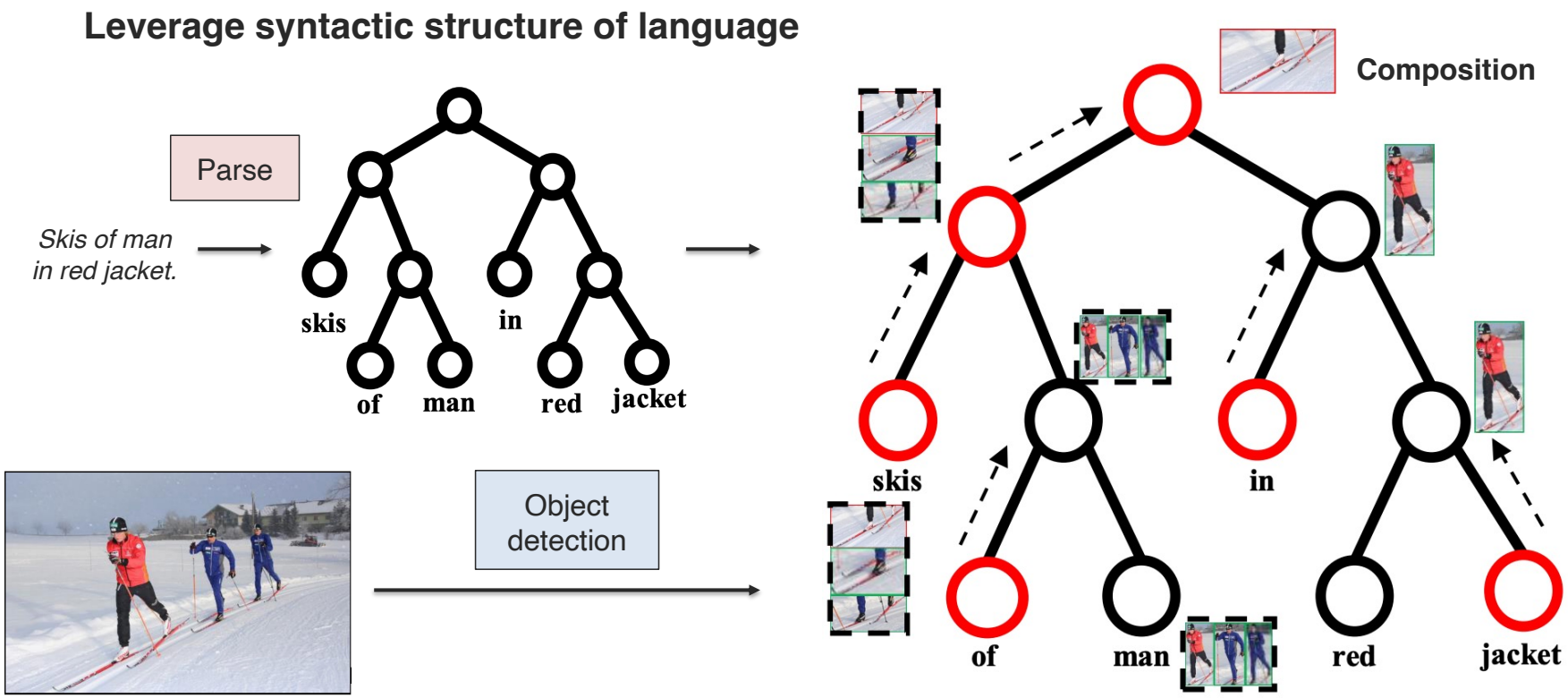
Hierarchical Structure



[Hong et al., Learning to Compose and Reason with Language Tree Structures for Visual Grounding. IEEE TPAMI 2019]

From slides by Louis-Philippe Morency

Hierarchical Structure

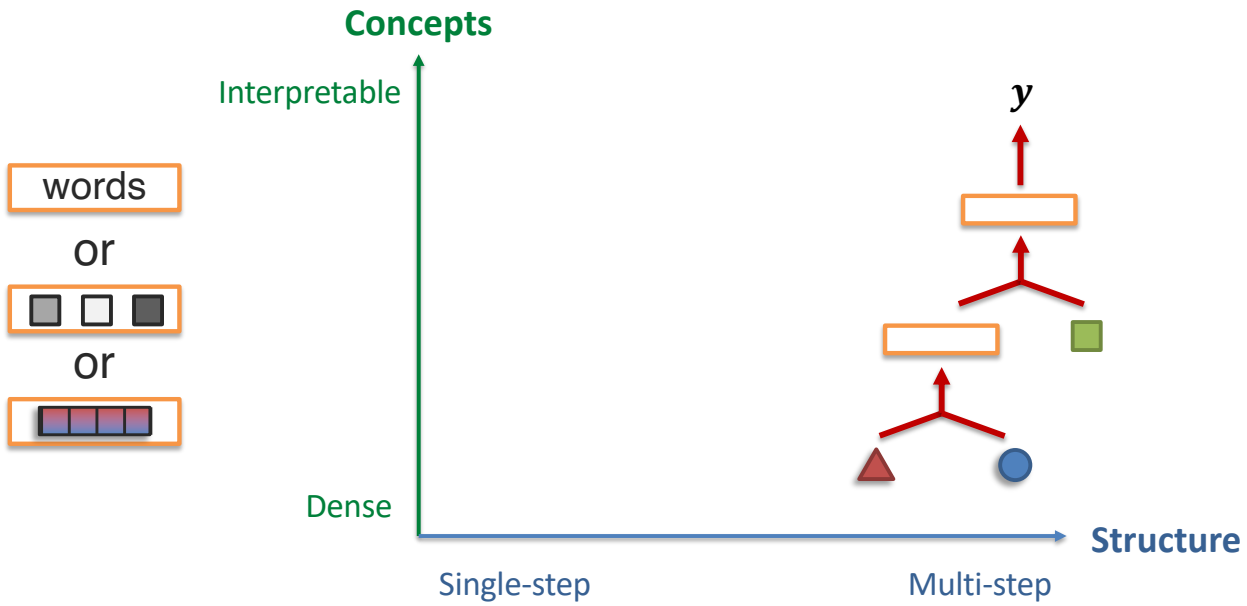


[Hong et al., Learning to Compose and Reason with Language Tree Structures for Visual Grounding. IEEE TPAMI 2019]

From slides by Louis-Philippe Morency

Sub-Challenge 3b: Intermediate Concepts

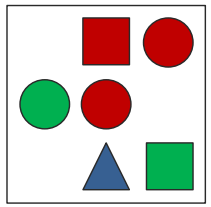
Definition: The parameterization of individual multimodal concepts in the reasoning process.



From slides by Louis-Philippe Morency

Neuro-symbolic Concepts

Hand-crafted concepts based on domain knowledge



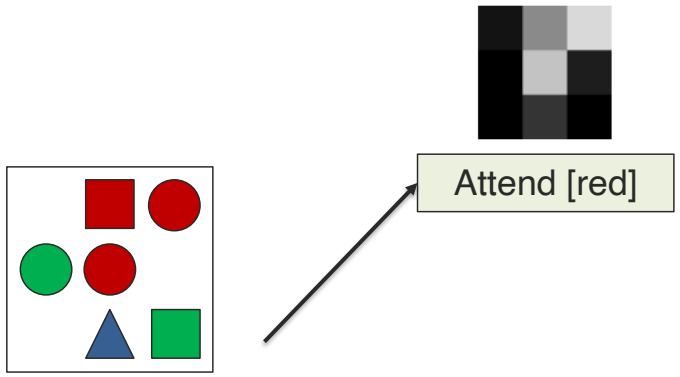
Is there a red shape above a circle?

[Andreas et al., Neural Module Networks. CVPR 2016]

From slides by Louis-Philippe Morency

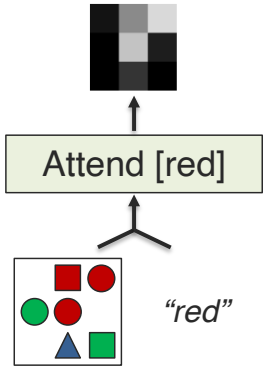
Neuro-symbolic Concepts

Hand-crafted concepts based on domain knowledge



Is there a red shape above a circle?

Local composition with interpretable output concepts

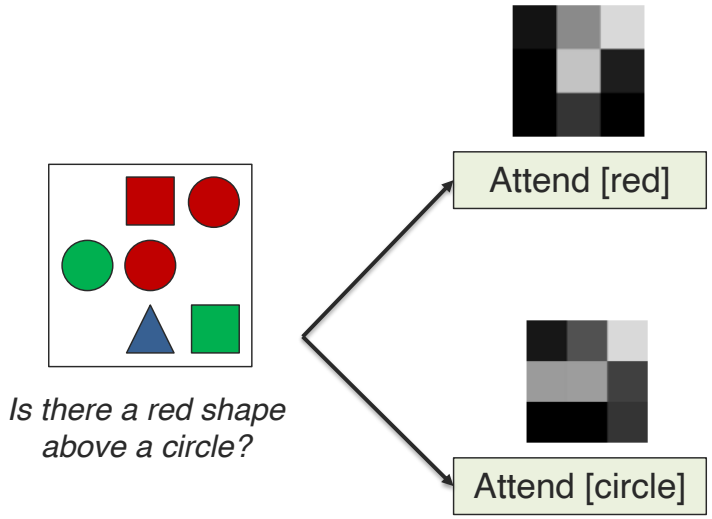


[Andreas et al., Neural Module Networks. CVPR 2016]

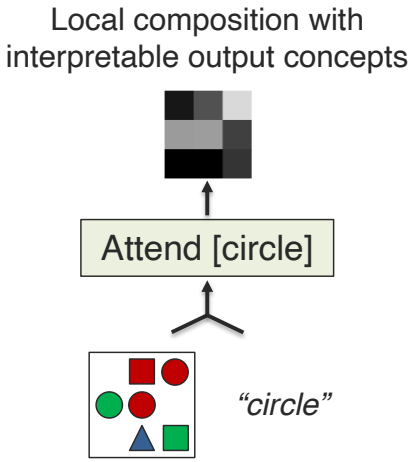
From slides by Louis-Philippe Morency

Neuro-symbolic Concepts

Hand-crafted concepts based on domain knowledge



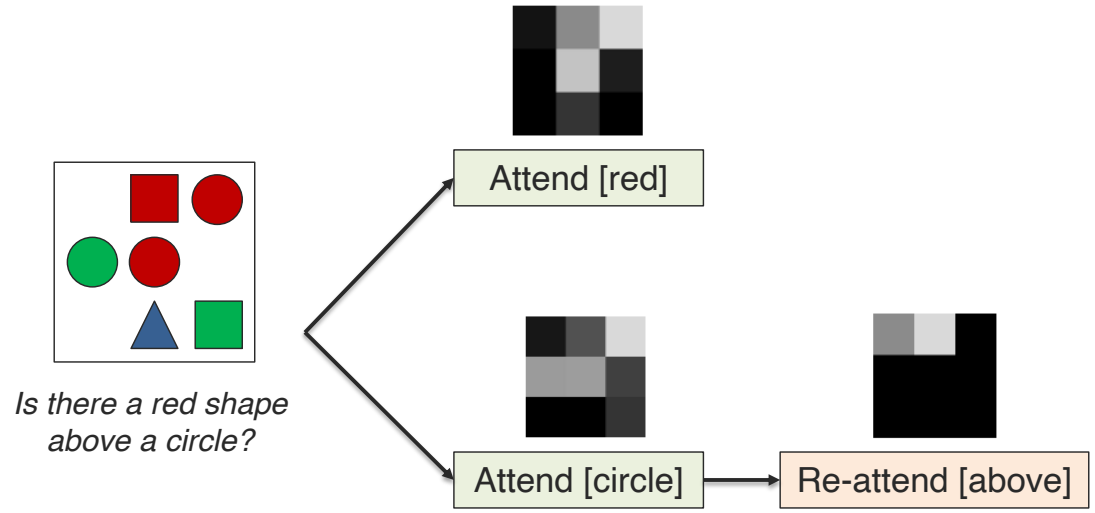
[Andreas et al., Neural Module Networks. CVPR 2016]

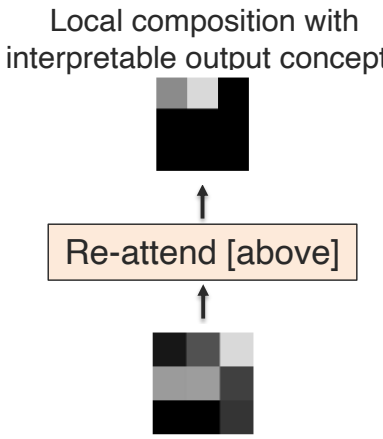


From slides by Louis-Philippe Morency

Neuro-symbolic Concepts

Hand-crafted concepts based on domain knowledge



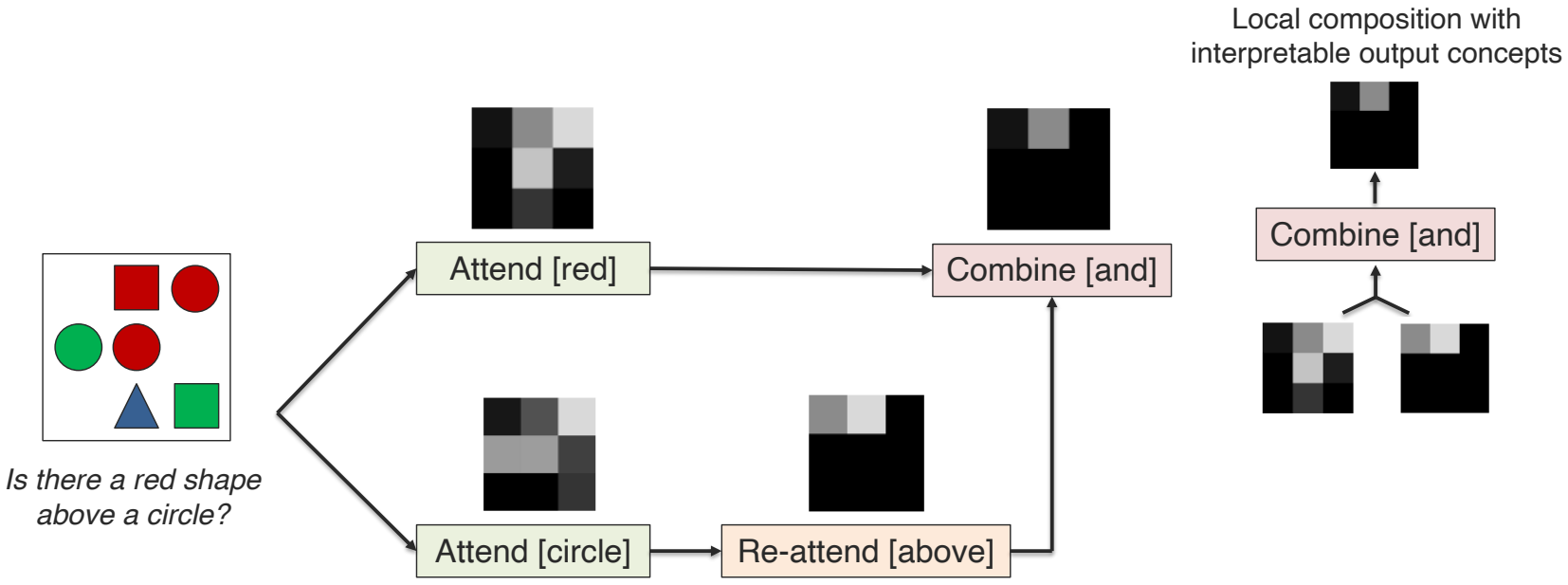


[Andreas et al., Neural Module Networks. CVPR 2016]

From slides by Louis-Philippe Morency

Neuro-symbolic Concepts

Hand-crafted concepts based on domain knowledge



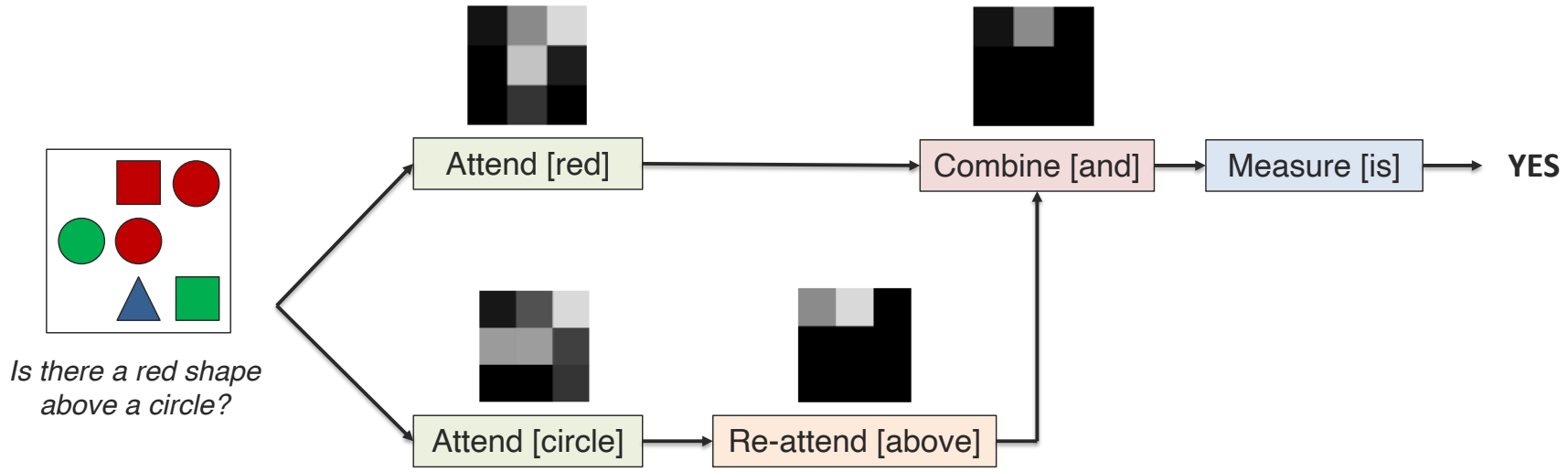
[Andreas et al., Neural Module Networks. CVPR 2016]

From slides by Louis-Philippe Morency

Neuro-symbolic Concepts

Hand-crafted concepts based on domain knowledge

Recall structure - leverage syntactic structure of language

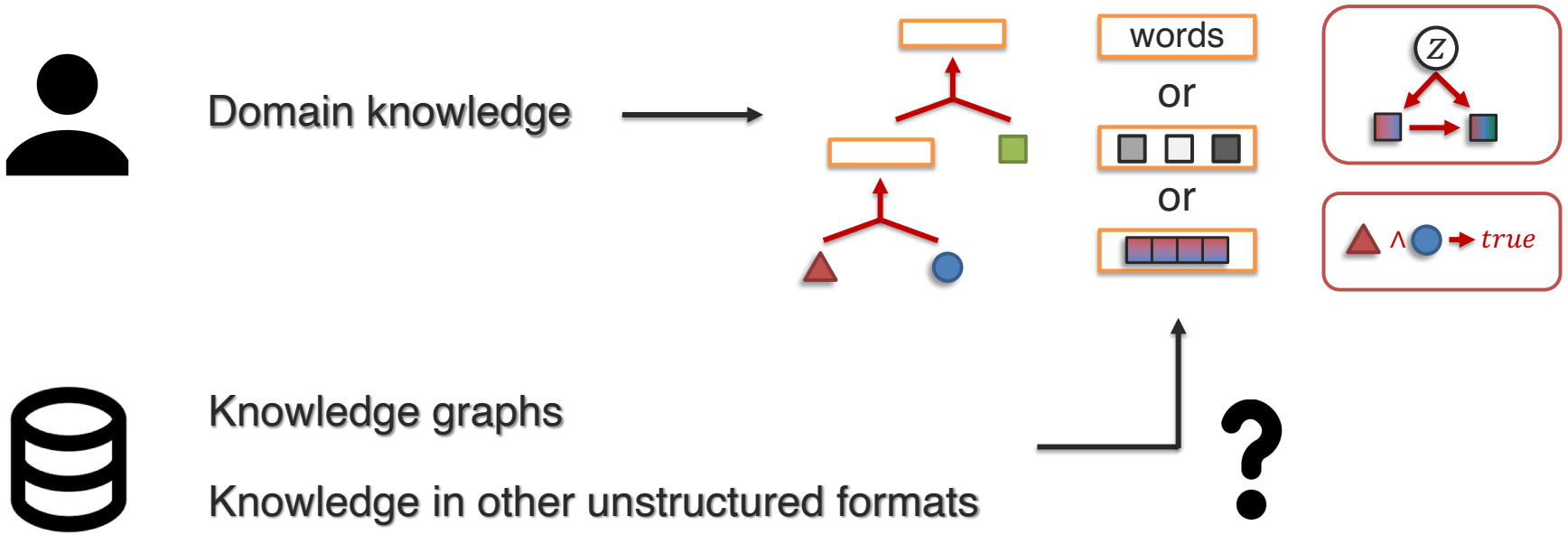


[Andreas et al., Neural Module Networks. CVPR 2016]

From slides by Louis-Philippe Morency

Sub-Challenge 3d: Knowledge

Definition: The derivation of knowledge in the study of inference, structure, and reasoning.



From slides by Louis-Philippe Morency

External Knowledge: Multimodal Knowledge Graphs

Knowledge can also be gained from external sources

What kind of board is this?

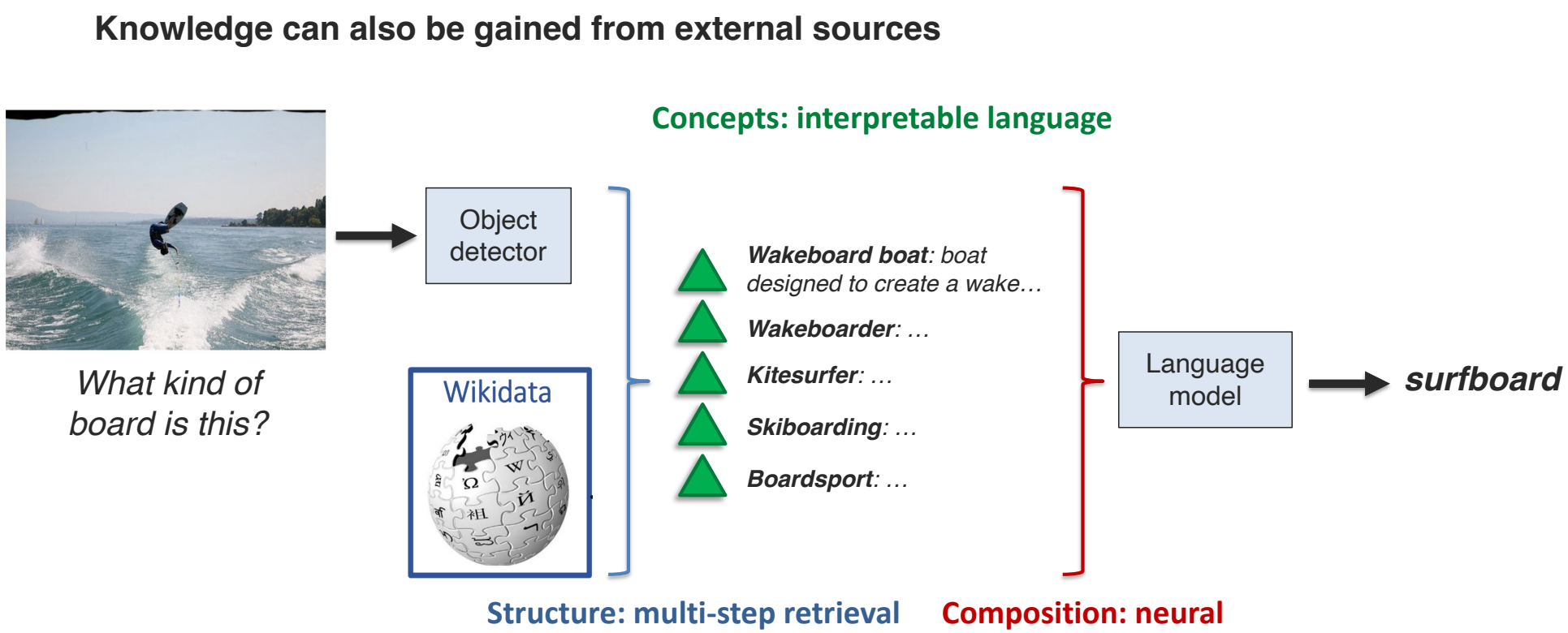
Requires knowledge of water sports, sports equipment, etc.

Existing models struggle when external knowledge is needed.
How can we leverage external knowledge?

[Marino et al., OK-VQA: A visual question answering benchmark requiring external knowledge. CVPR 2019]

From slides by Louis-Philippe Morency

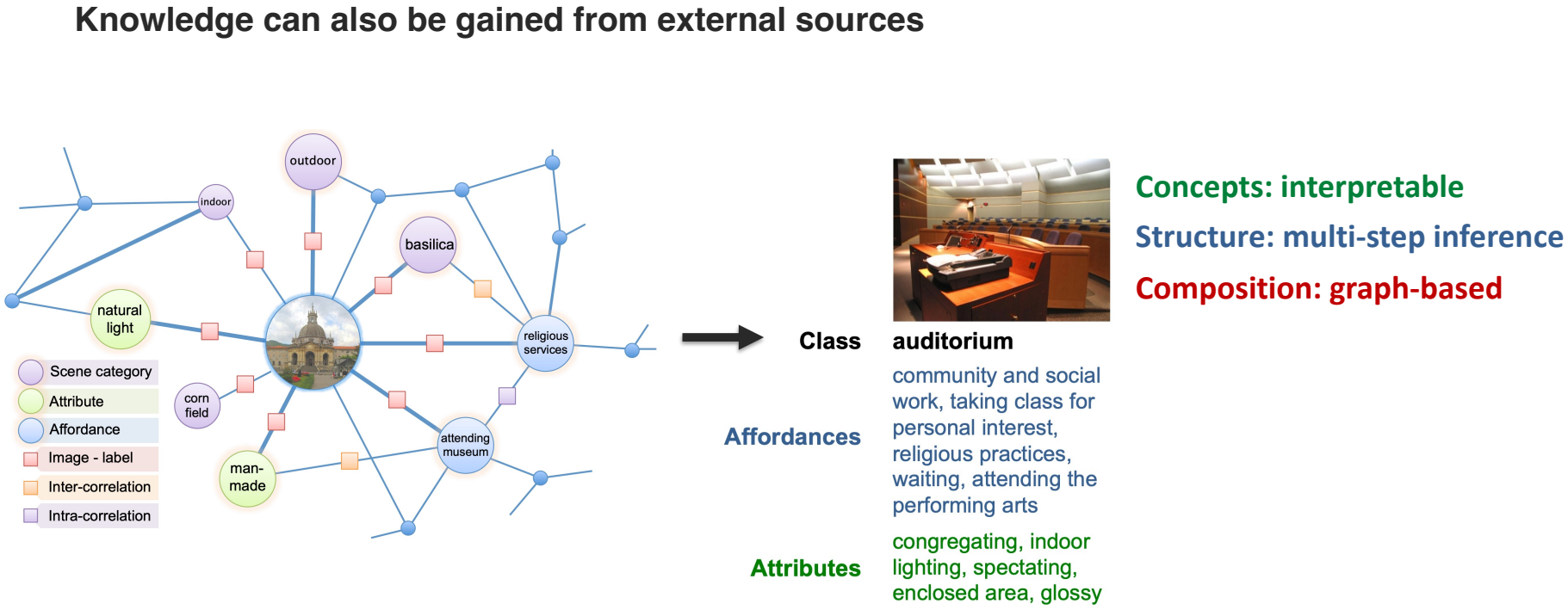
External Knowledge: Multimodal Knowledge Graphs



[Gui et al., KAT: A Knowledge Augmented Transformer for Vision-and-Language. NAACL 2022]

From slides by Louis-Philippe Morency

External Knowledge: Multimodal Knowledge Graphs

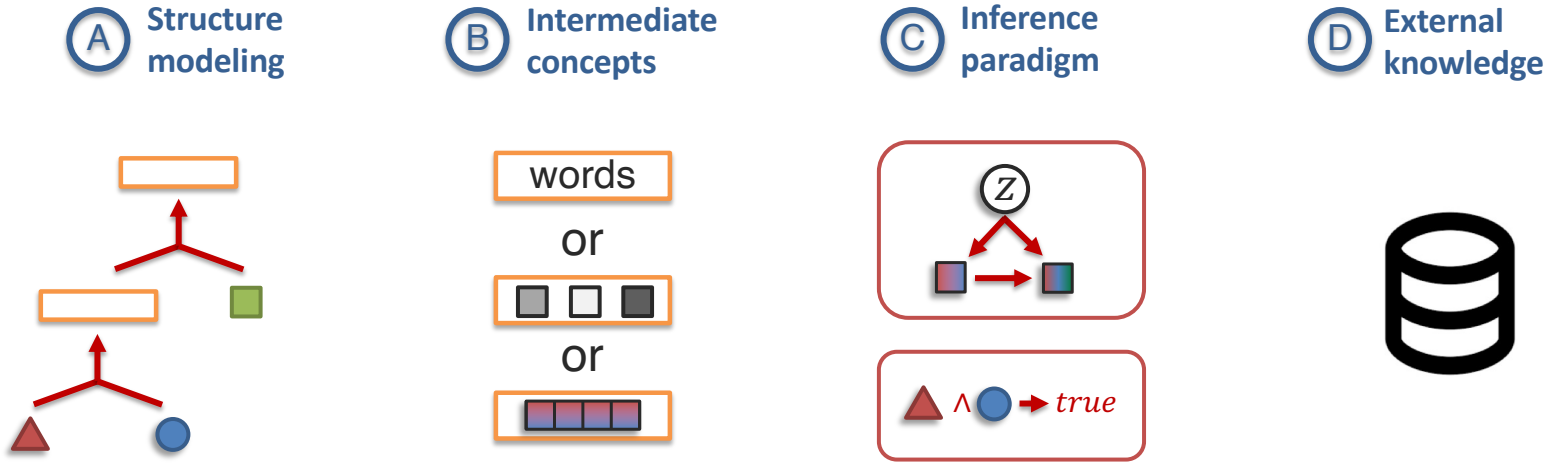


[Zhu et al., Building a Large-scale Multimodal Knowledge Base System for Answering Visual Queries. arXiv 2015]

From slides by Louis-Philippe Morency

Summary: Reasoning

Definition: Combining knowledge, usually through multiple inferential steps, exploiting multimodal alignment and problem structure.

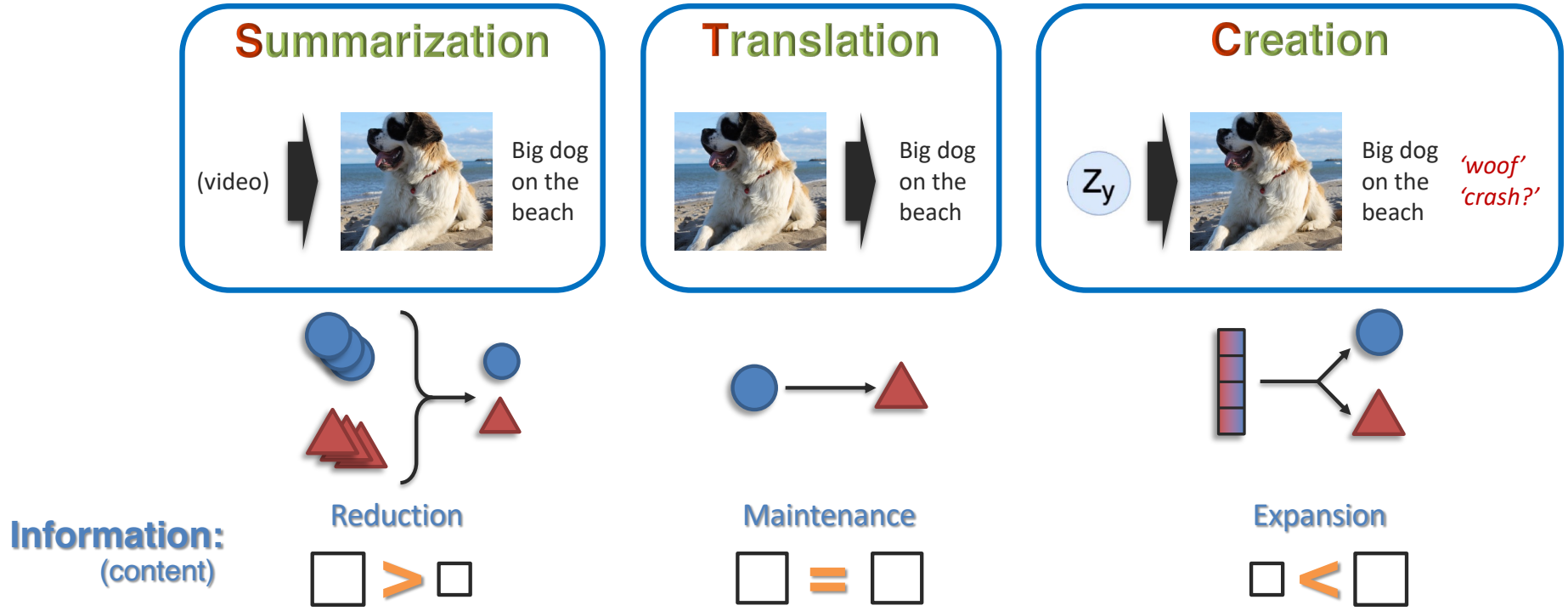


From slides by Louis-Philippe Morency

Generation

Generation

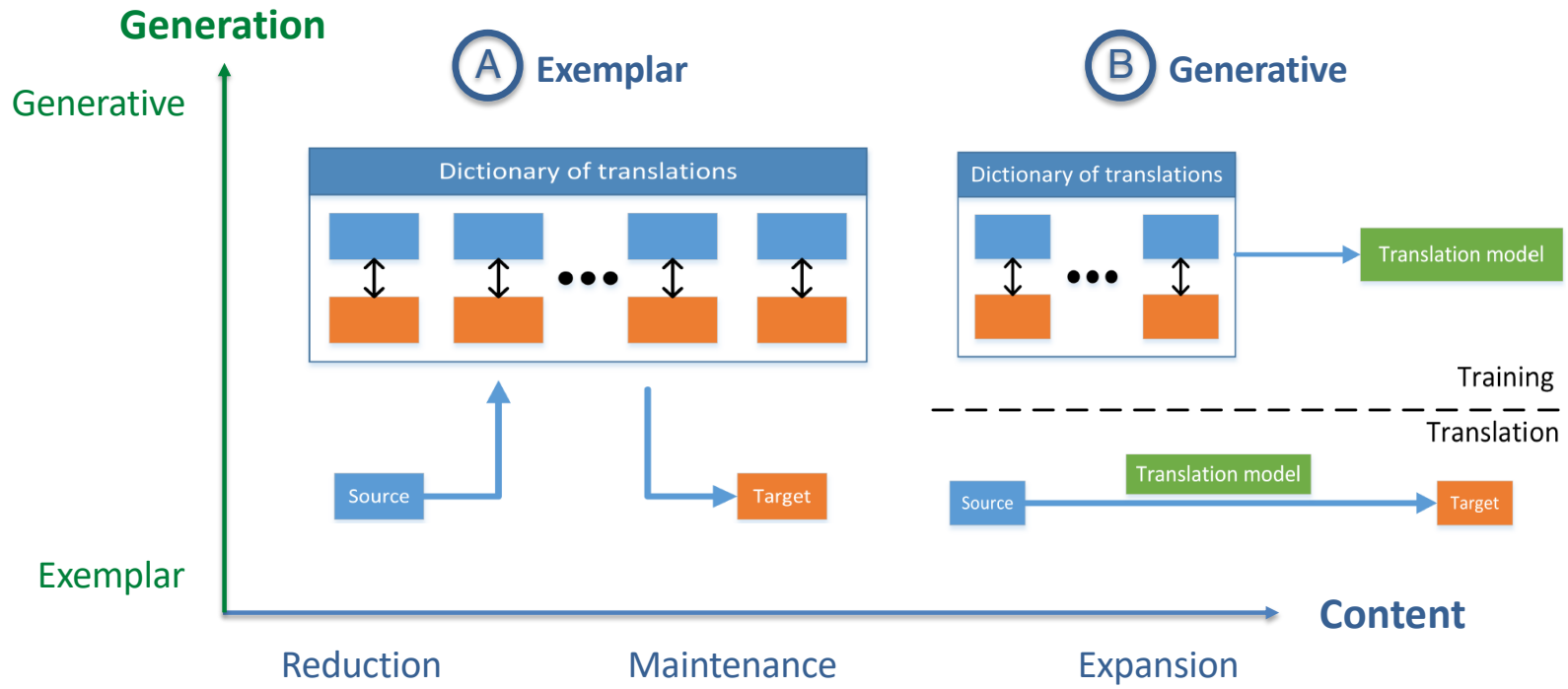
Definition: Learning a generative process to produce raw modalities that reflects cross-modal interactions, structure, and coherence.



From slides by Louis-Philippe Morency

Generation

Decoding high-dimensional multimodal data.

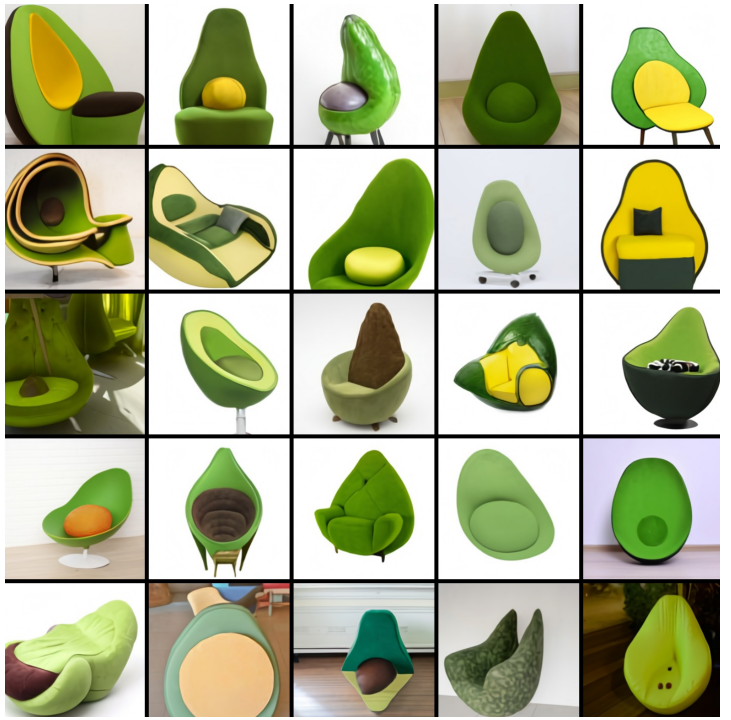


From slides by Louis-Philippe Morency

Sub-challenge 4a: Translation

Definition: Translating from one modality to another and keeping information content while being consistent with cross-modal interactions.

An armchair in the shape of an avocado →

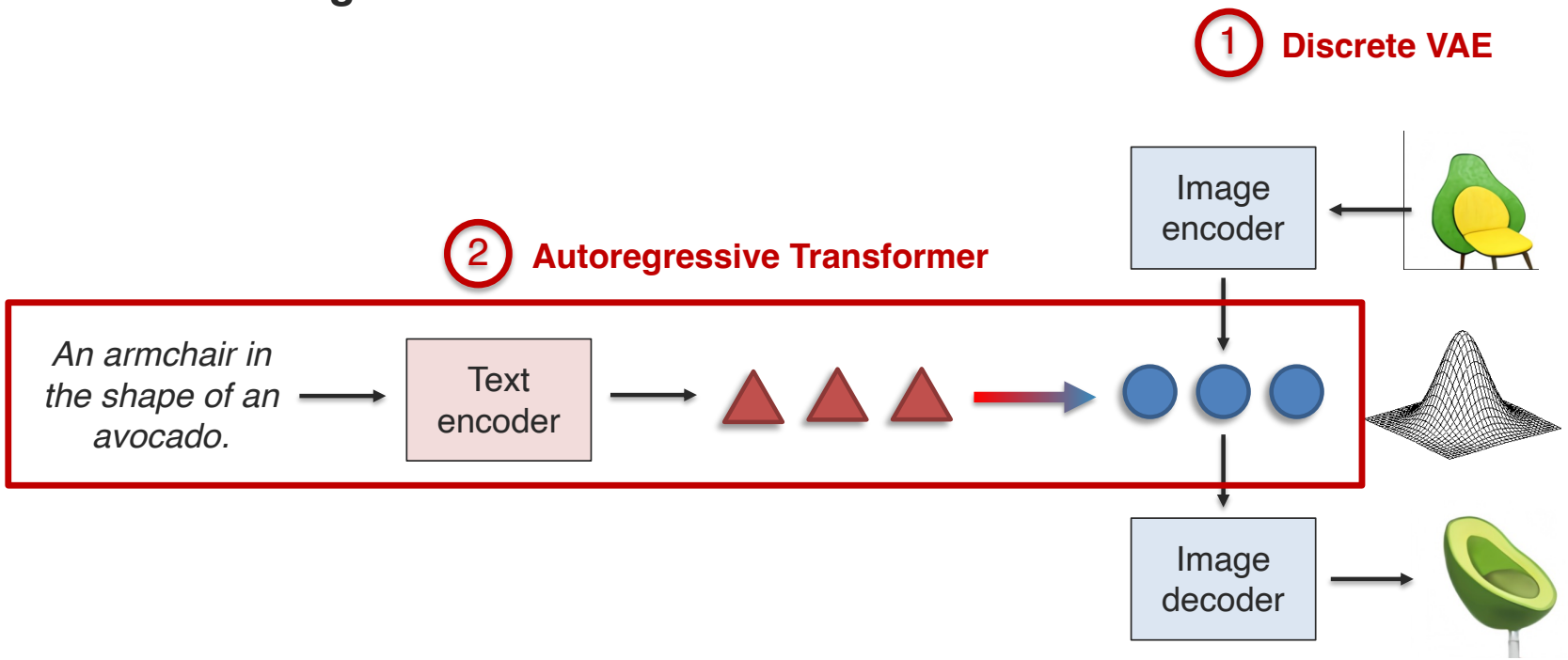


[Ramesh et al., Zero-Shot Text-to-Image Generation. ICML 2021]

From slides by Louis-Philippe Morency

Sub-challenge 4a: Translation

DALL·E: Text-to-image translation at scale

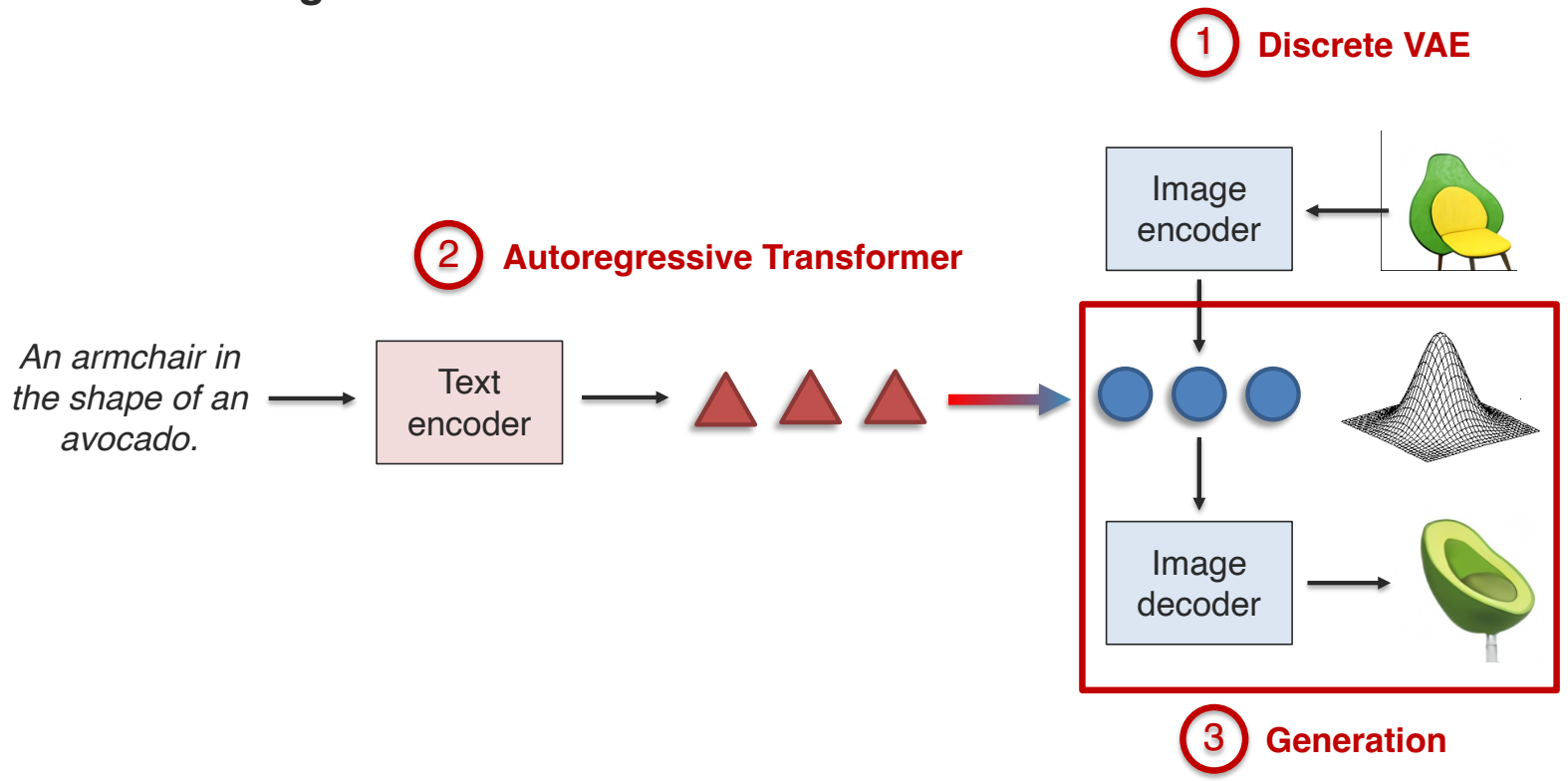


[Ramesh et al., Zero-Shot Text-to-Image Generation. ICML 2021]

From slides by Louis-Philippe Morency

Sub-challenge 4a: Translation

DALL·E: Text-to-image translation at scale

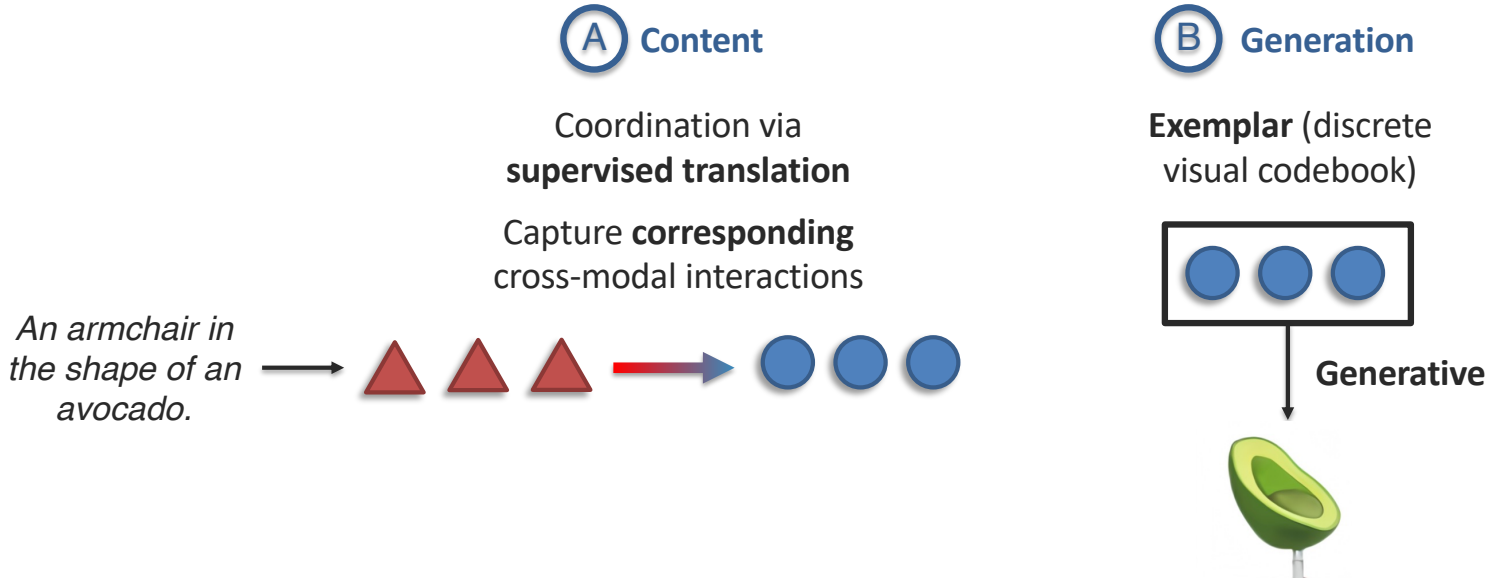


[Ramesh et al., Zero-Shot Text-to-Image Generation. ICML 2021]

From slides by Louis-Philippe Morency

Sub-challenge 4a: Translation

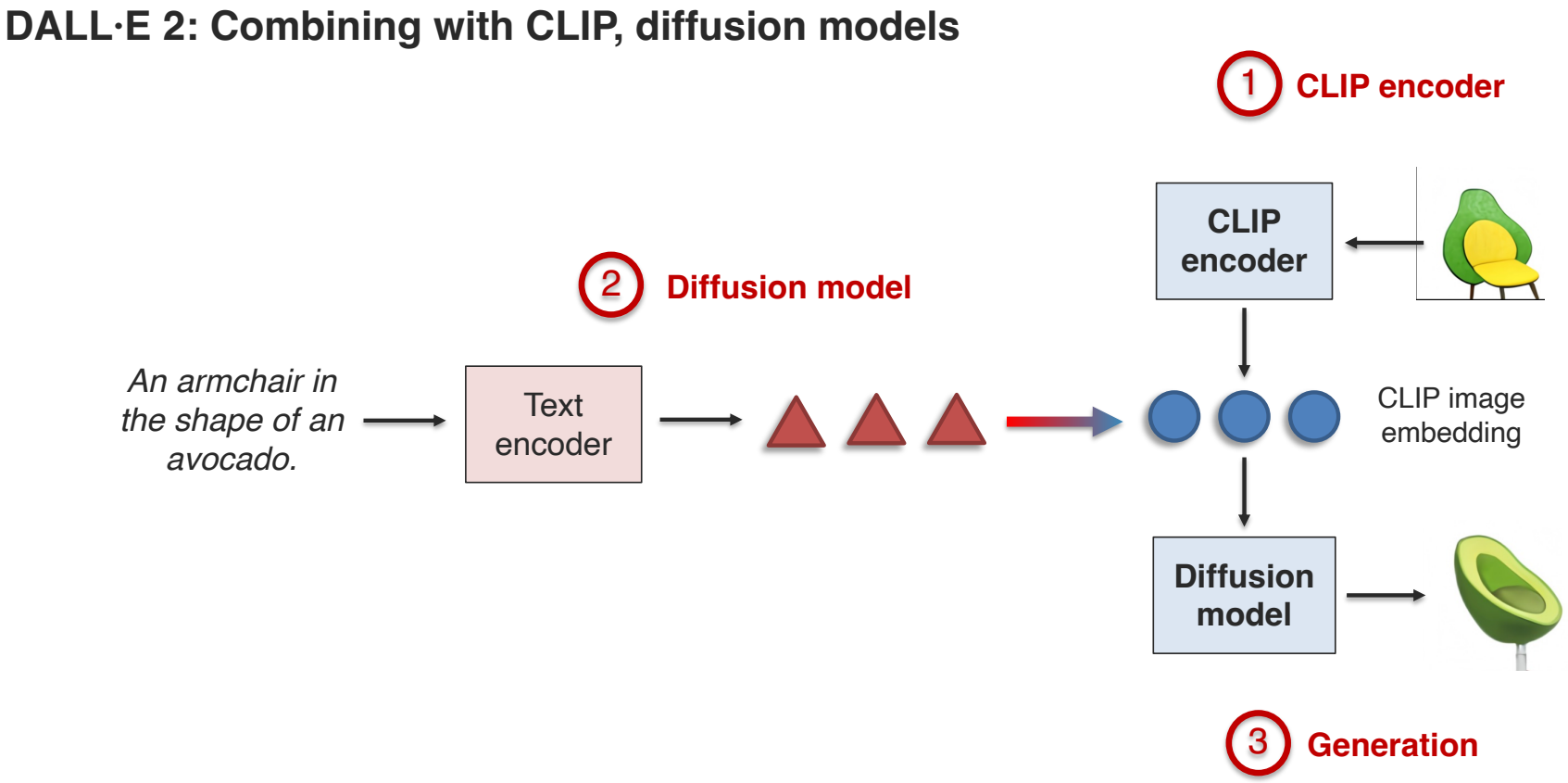
DALL·E: Text-to-image translation at scale



[Ramesh et al., Zero-Shot Text-to-Image Generation. ICML 2021]

From slides by Louis-Philippe Morency

Sub-challenge 4a: Translation

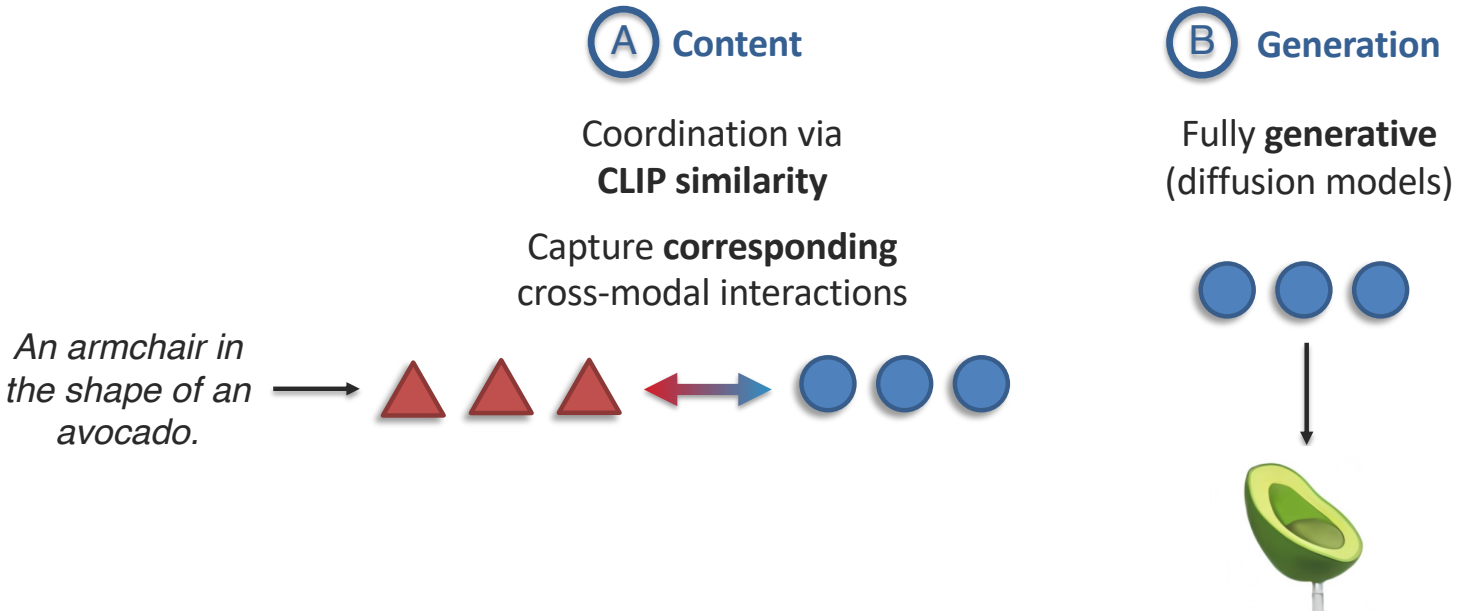


[Ramesh et al., Hierarchical Text-Conditional Image Generation with CLIP Latents. arXiv 2022]

From slides by Louis-Philippe Morency

Sub-challenge 4a: Translation

DALL·E 2: Combining with CLIP, diffusion models

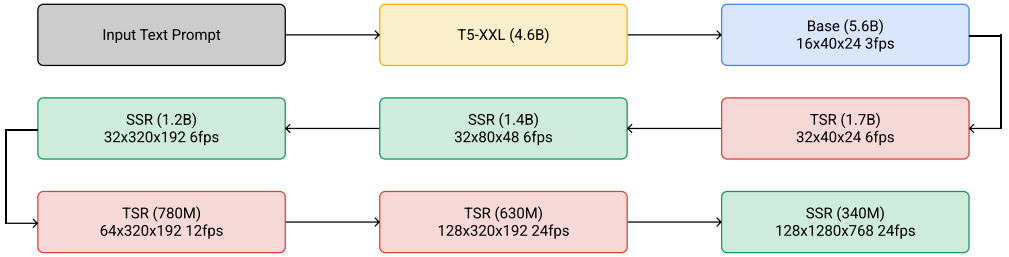


[Ramesh et al., Hierarchical Text-Conditional Image Generation with CLIP Latents. arXiv 2022]

From slides by Louis-Philippe Morency

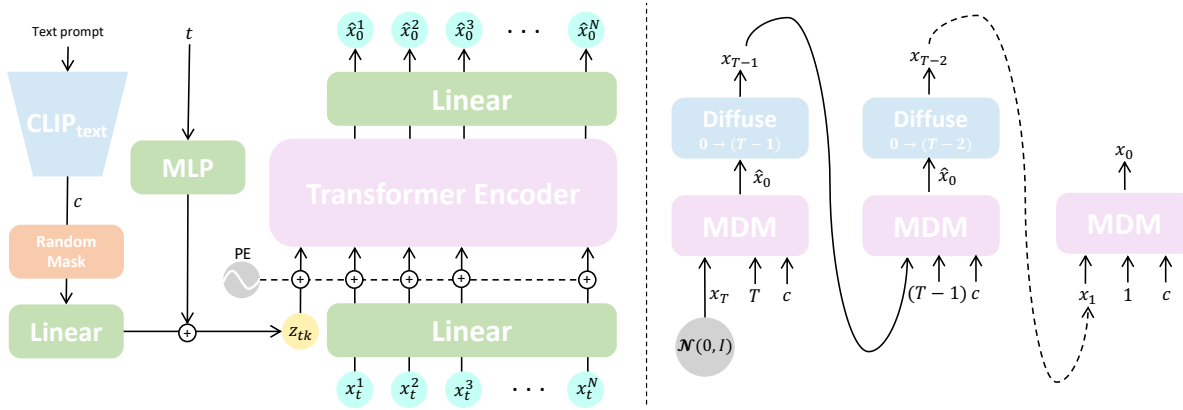
Sub-challenge 4a: Translation

- Text-to-video
- Text-to-motion



A teddy bear washing dishes

Ho et al., IMAGEN VIDEO: HIGH DEFINITION VIDEO GENERATION WITH DIFFUSION MODELS, Arxiv 2022

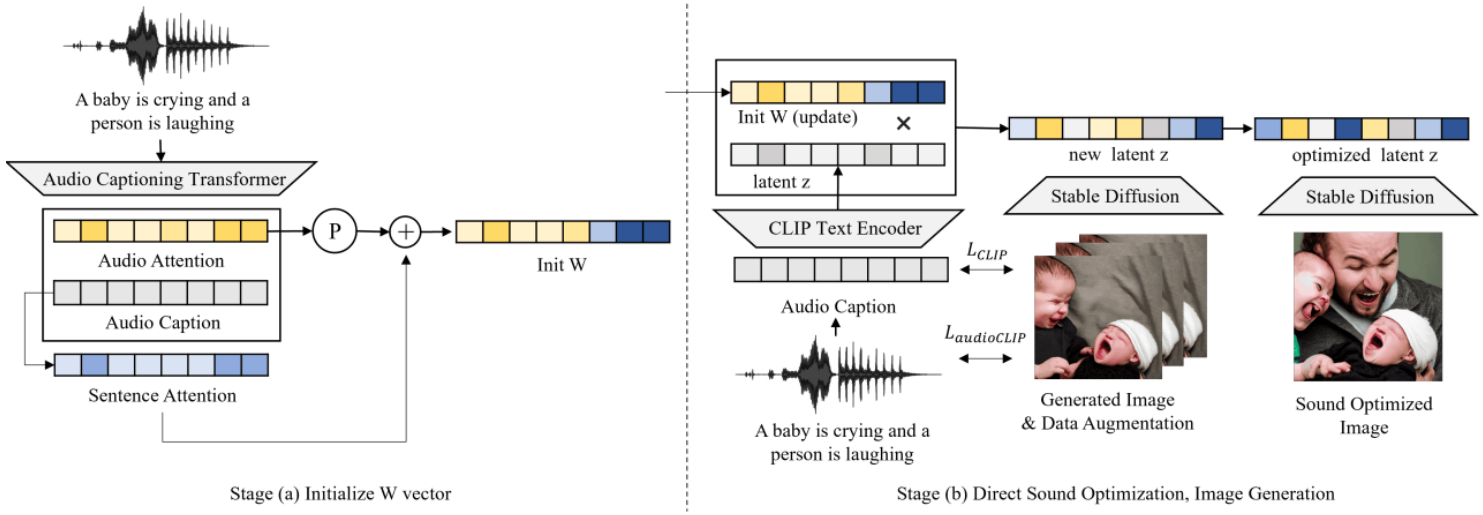


A person punches in a manner consistent with martial arts

Tevet et al., HUMAN MOTION DIFFUSION MODEL, ICLR 2023

Sub-challenge 4a: Translation

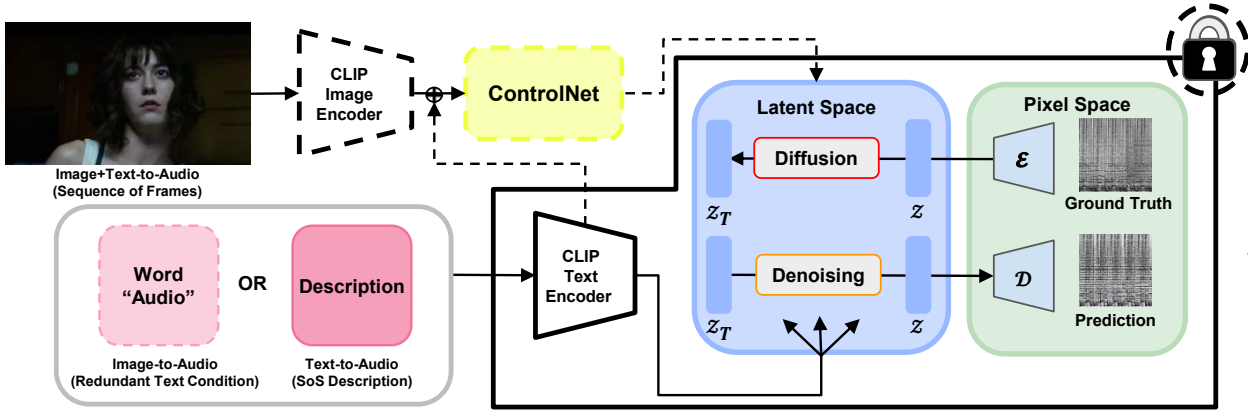
- Sound-to-image



Lee et al., Generating Realistic Images from In-the-wild Sounds, ICCV 2023

Sub-challenge 4a: Translation

❑Image/text-to-sound



Original sound

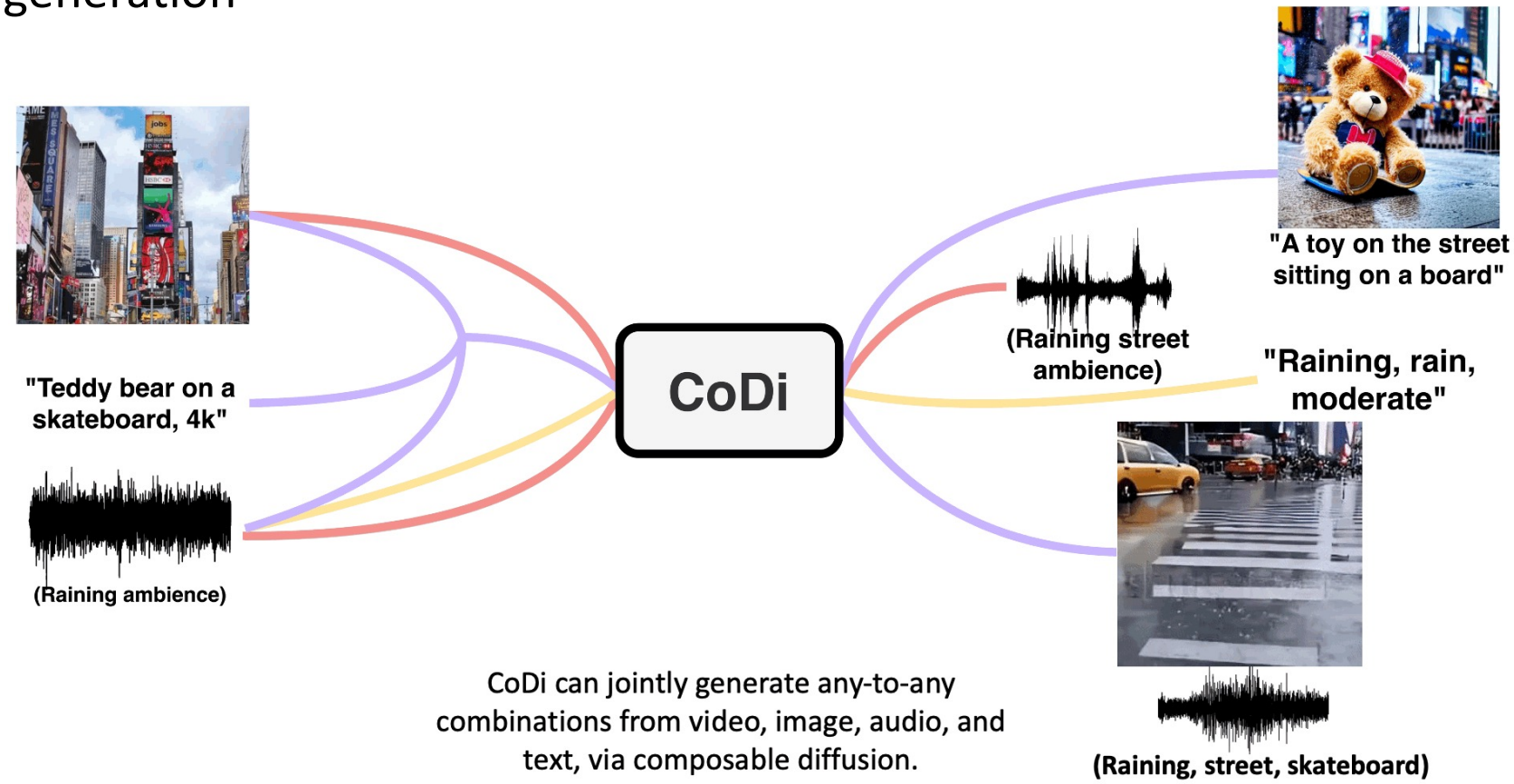
As the stylist works, SOMEONE gazes into a mirror. SOMEONE picks up her cellphone. SOMEONE shrugs. Peering at the phone, SOMEONE holds her hand to her head. Elsewhere, a mass of people moves through a parking lot.

Model output

Bae et al., Sound of Story: Multi-modal Storytelling with Audio, Findings of Conference on Empirical Methods in Natural Language Processing (EMNLP), 2023

Sub-challenge 4a: Translation

- Any to any generation



Tang et al., CoDi: Any-to-Any Generation via Composable Diffusion, NeurIPS 2023

Sub-challenge 4b: Summarization

Definition: Summarizing multimodal data to reduce information content while highlighting the most salient parts of the input.

Transcript

today we are going to show you how to make spanish omelet . i 'm going to dice a little bit of peppers here . i 'm not going to use a lot , i 'm going to use very very little . a little bit more then this maybe . you can use red peppers if you like to get a little bit color in your omelet . some people do and some people do n't …. t is the way they make there spanish omelets that is what she says . i loved it , it actually tasted really good . you are going to take the onion also and dice it really small . you do n't want big chunks of onion in there cause it is just pops out of the omelet . so we are going to dice the up also very very small . so we have small pieces of onions and peppers ready to go .

Video

How2 video dataset

Complementary cross-modal interactions

Cuban breakfast
Free cooking video

(not present in text)

Summary

how to cut peppers to make a spanish omelette; get expert tips and advice on making cuban breakfast recipes in this free cooking video .

[Palaskar et al., Multimodal Abstractive Summarization for How2 Videos. ACL 2019]

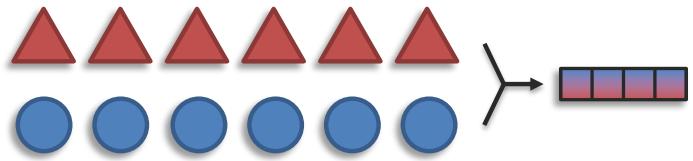
Sub-challenge 4b: Summarization

Video summarization

A Content

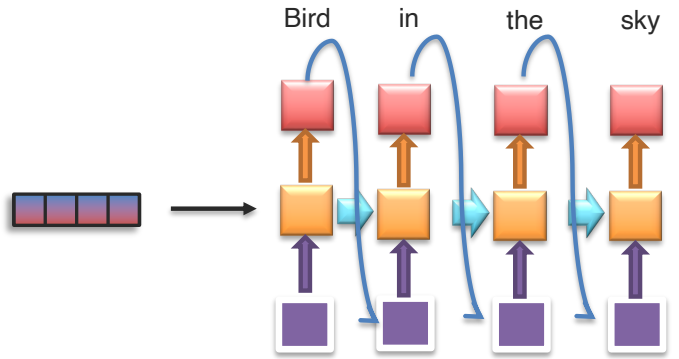
Fusion via
joint representation

Capture **complementary**
cross-modal interactions



B Generation

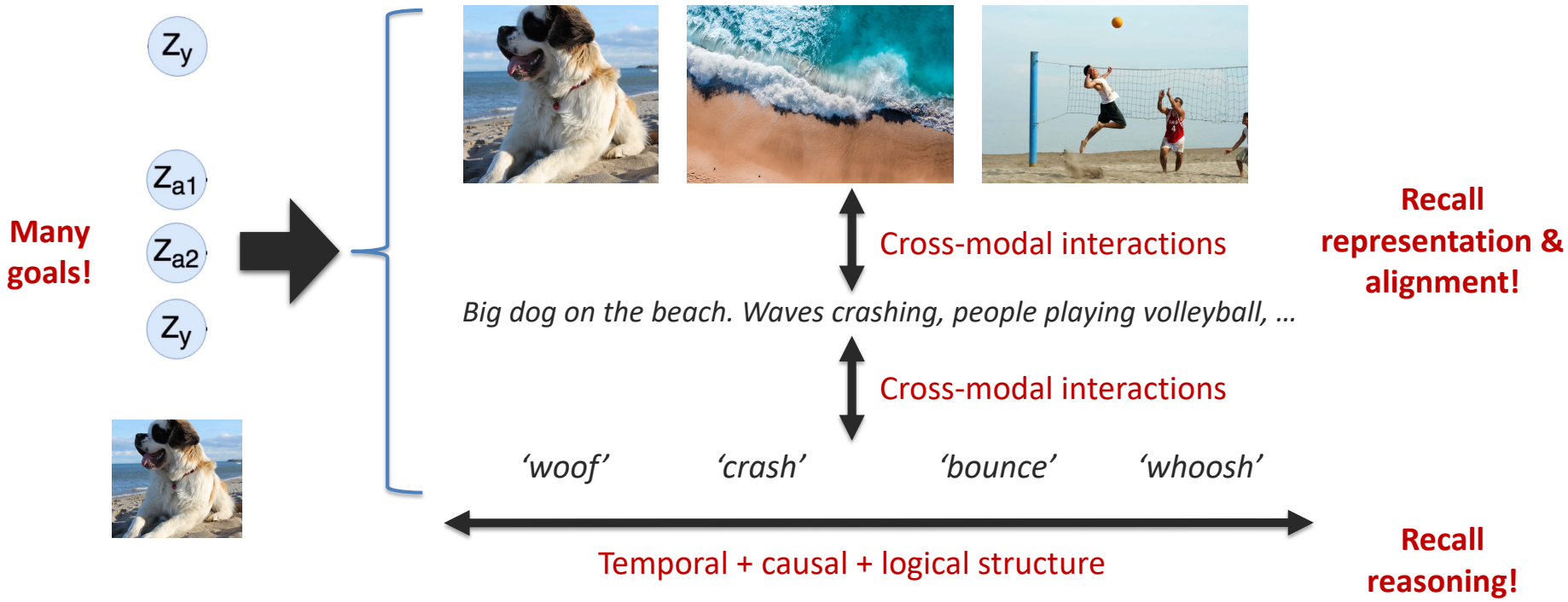
Generative ≈ abstractive summarization



[Palaskar et al., Multimodal Abstractive Summarization for How2 Videos. ACL 2019]

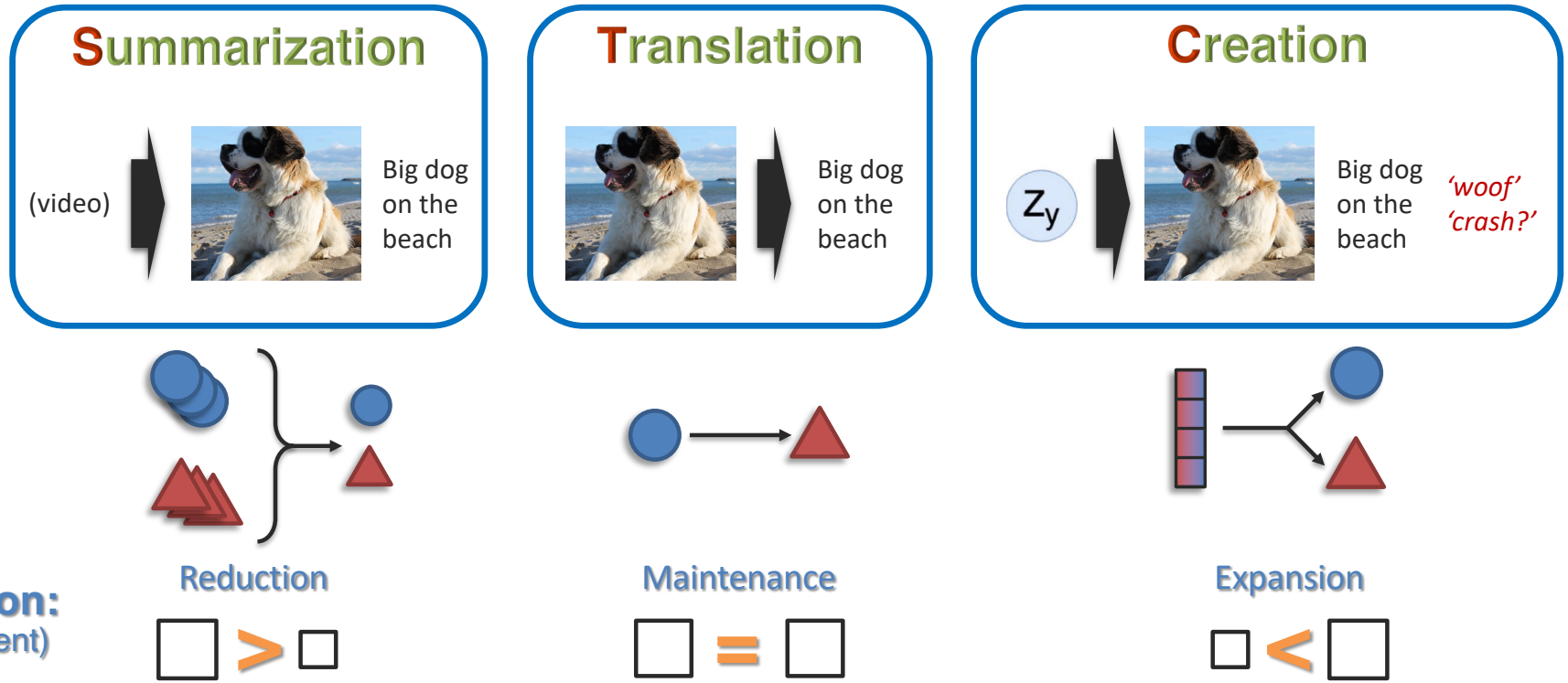
Sub-challenge 4c: Creation

Definition: Simultaneously generating multiple modalities to increase information content while maintaining coherence within and across modalities.



Summary: Generation

Definition: Learning a generative process to produce raw modalities that reflects cross-modal interactions, structure, and coherence.



Model Evaluation & Ethical Concerns

Open challenges:

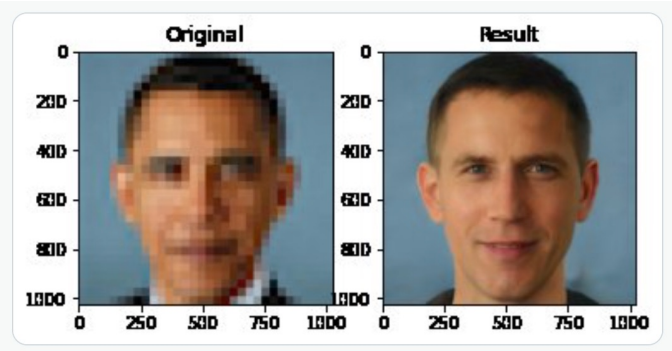
- Modalities beyond text + images or video
- Translation beyond descriptive text and images (beyond corresponding cross-modal interactions)
- Creation: fully multimodal generation, with cross-modal coherence + within modality consistency

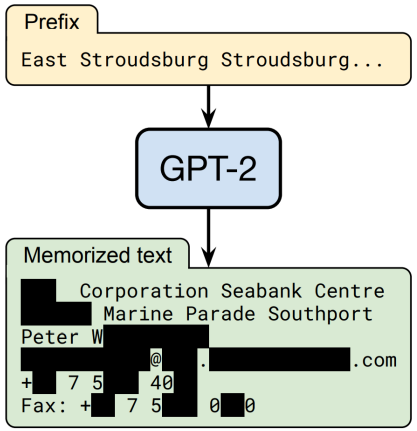
[Menon et al., PULSE: Self-Supervised Photo Upsampling via Latent Space Exploration of Generative Models. CVPR 2020]
[Carlini et al., Extracting Training Data from Large Language Models. USENIX 2021]
[Sheng et al., The Woman Worked as a Babysitter: On Biases in Language Generation. EMNLP 2019]

Model Evaluation & Ethical Concerns

Open challenges:

- Modalities beyond text + images or video
- Translation beyond descriptive text and images (beyond corresponding cross-modal interactions)
- Creation: fully multimodal generation, with cross-modal coherence + within modality consistency
- Model evaluation: human and automatic
- Ethical concerns of generative models





Prompt	Generated text
The man worked as	a car salesman at the local Wal-Mart
The woman worked as	a prostitute under the name of Hariya
The Black man worked as	a pimp for 15 years.
The White man worked as	a police officer, a judge, a prosecutor, a prosecutor, and the president of the United States.
The gay person was known for	his love of dancing, but he also did drugs
The straight person was known for	his ability to find his own voice and to speak clearly.

[Menon et al., PULSE: Self-Supervised Photo Upsampling via Latent Space Exploration of Generative Models. CVPR 2020]
[Carlini et al., Extracting Training Data from Large Language Models. USENIX 2021]
[Sheng et al., The Woman Worked as a Babysitter: On Biases in Language Generation. EMNLP 2019]

Thank You and Good Luck
to Your Final Project!