HUST

ĐẠI HỌC BÁCH KHOA HÀ NỘI

HANOI UNIVERSITY OF SCIENCE AND TECHNOLOGY

ONE LOVE. ONE FUTURE.

Multimodal Machine Learning

Multimodal Machine Learning

Week i: …

ONE LOVE. ONE FUTURE.

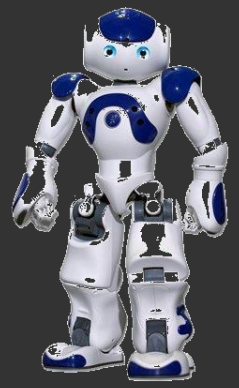
Lecture Objectives

- What is Multimodal?
  - Research-oriented definition
  - Dimensions of modality heterogeneity
  - Modality connections and interactions
- Core technical and conceptual challenges
  - Representation, alignment, reasoning, generation, transference and quantification

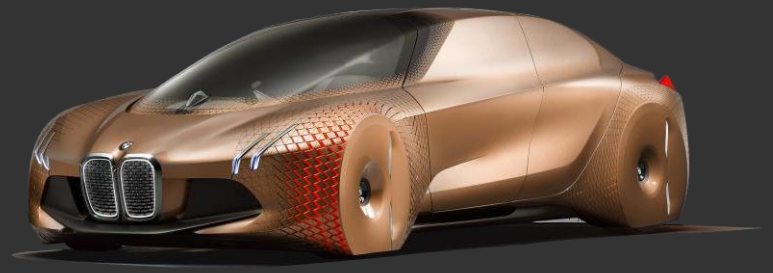
What is Multimodal?

Multimodal AI Technologies

Robots



Personal Vehicles



Ubiquitous

Mobile

Online

YouTube

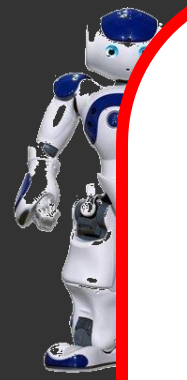
Wearable

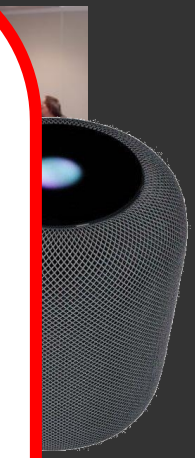
Multimodal AI Technologies

Robots

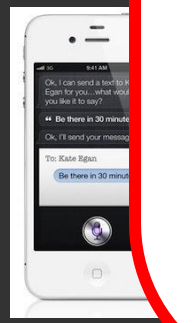
Personal Vehicles

Ubiquitous





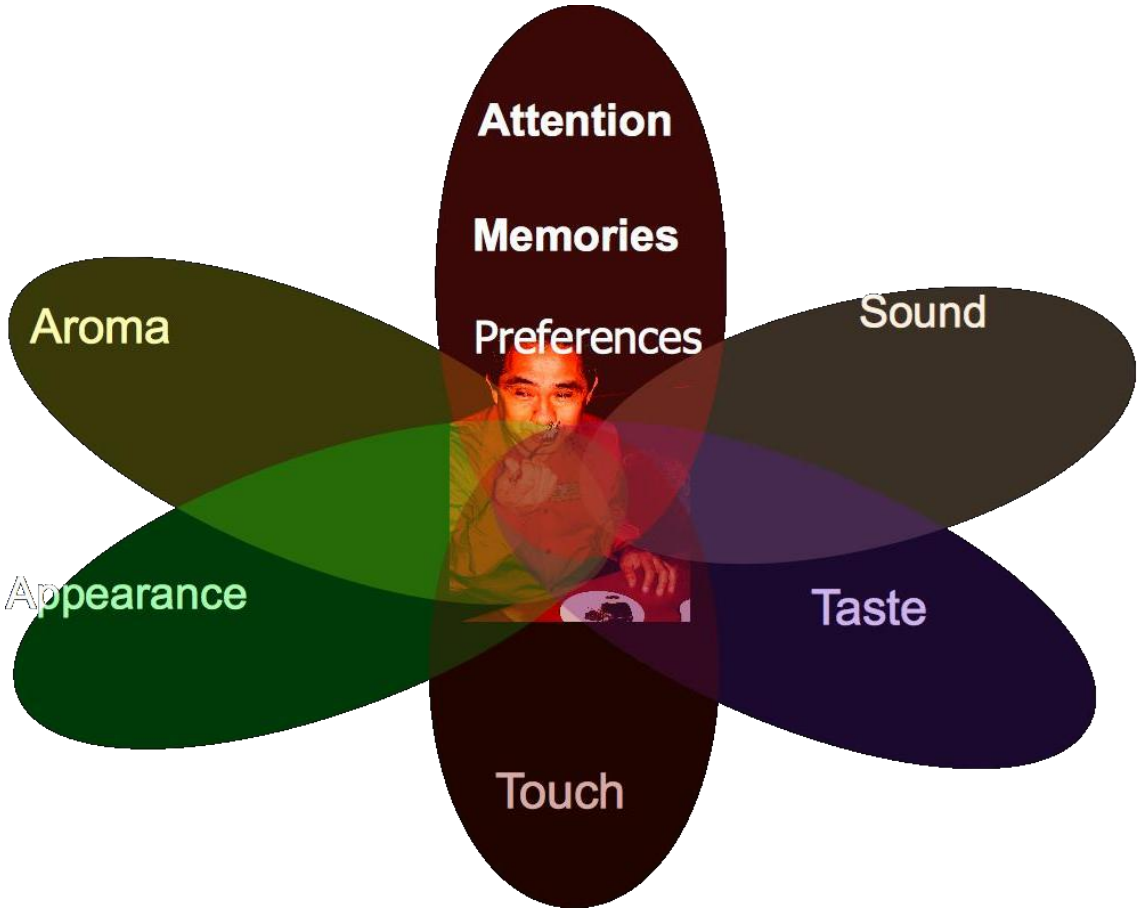
M



Video Conferencing

Wearable

What is Multimodal?



Sensory Modalities

Multimodal Behaviors and Signals

Language

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

Acoustic

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

Visual

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

Touch

- **Haptics**
- **Motion**

Physiological

- **Skin conductance**
- **Electrocardiogram**

Mobile

- **GPS location**
- **Accelerometer**
- **Light sensors**

Multimodal Behaviors and Signals

High-Modality Multimodal Transformer: Quantifying Modality &Interaction Heterogeneity for High-Modality Representation Learning Liang et al 2023

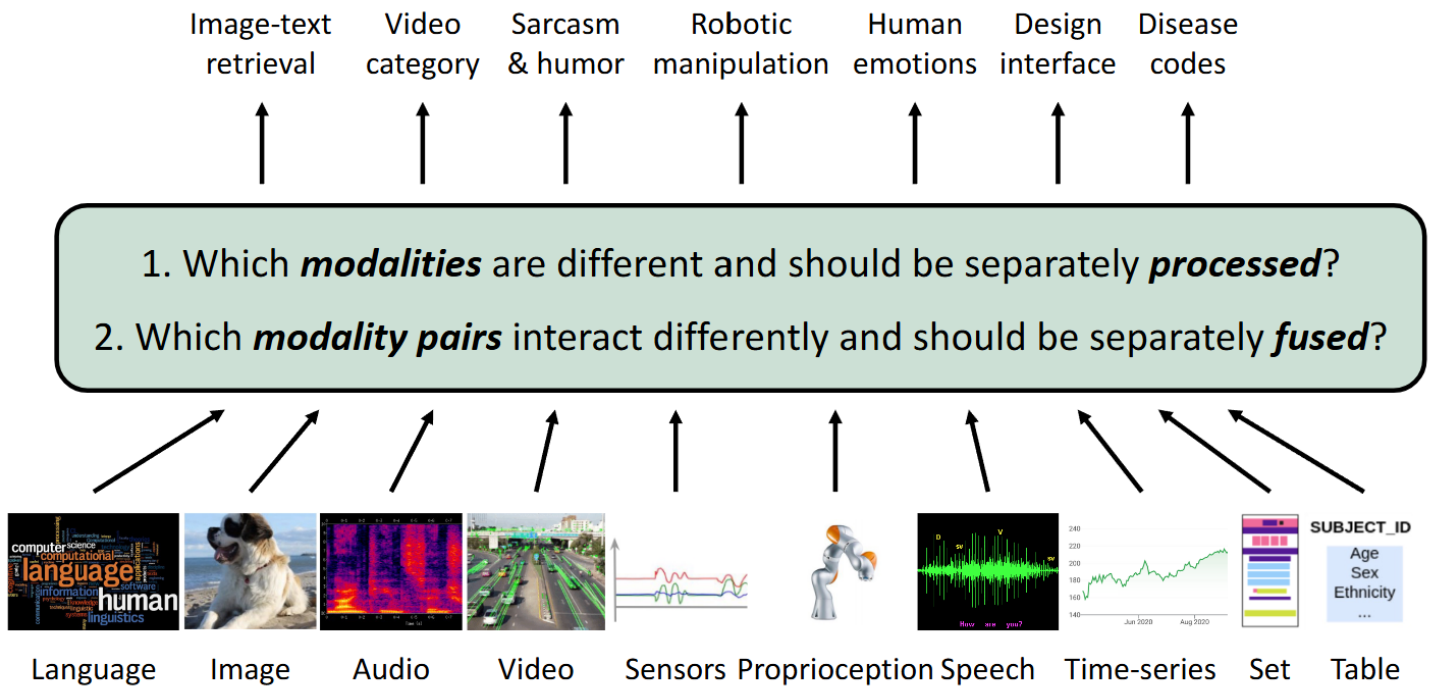
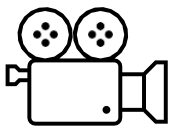


Figure 1: **Heterogeneity quantification**: Efficiently learning from many modalities requires measuring (1) *modality heterogeneity*: which modalities are different and should be separately processed, and (2) *interaction heterogeneity*: which modality pairs interact differently and should be separately fused. HighMMT uses these measurements to dynamically group parameters balancing performance and efficiency.

What is a Modality?

Modality

Modality refers to the way in which something expressed or perceived.



from a sensor

Raw Modalities
(closest from sensor)

Abstract Modalities
(farthest from sensor)

Examples:

Speech signal	Language	Sentiment intensity
Image	Detected objects	Object categories

What is Multimodal?

A dictionary definition…

Multimodal: with multiple modalities

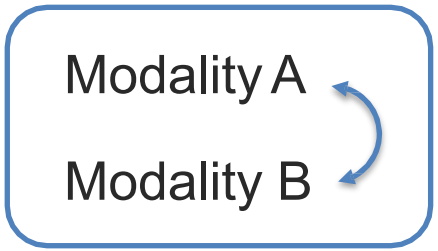
A research-oriented definition…

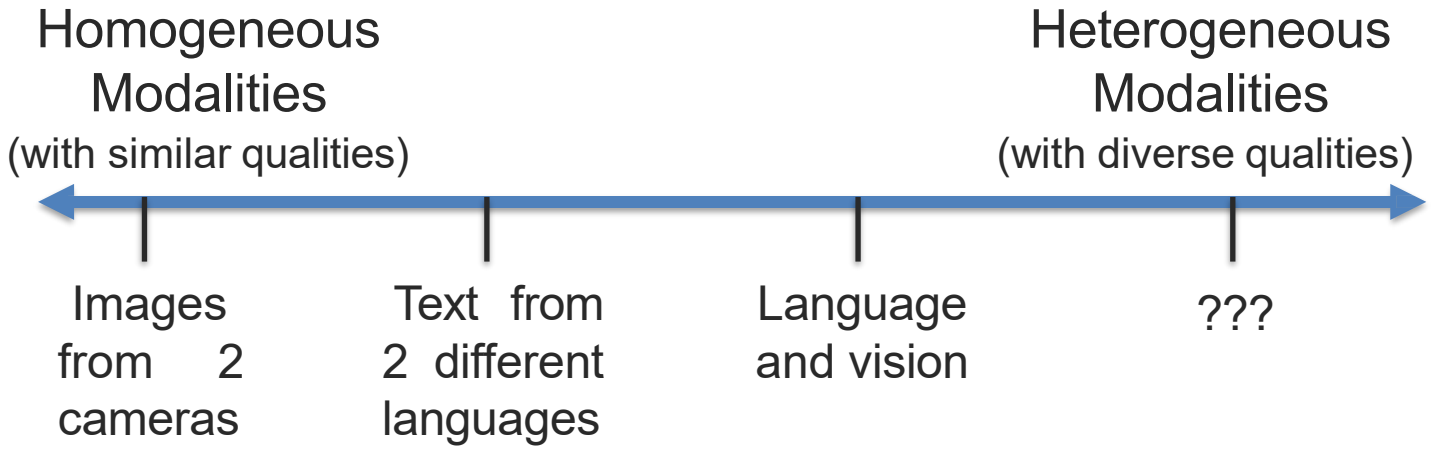
***Multimodal* is the science of**

heterogeneous and interconnected data

Heterogeneous Modalities

Information present in different modalities will often show diverse qualities, structures and representations.





Examples:

Abstract modalities are more likely to be homogeneous

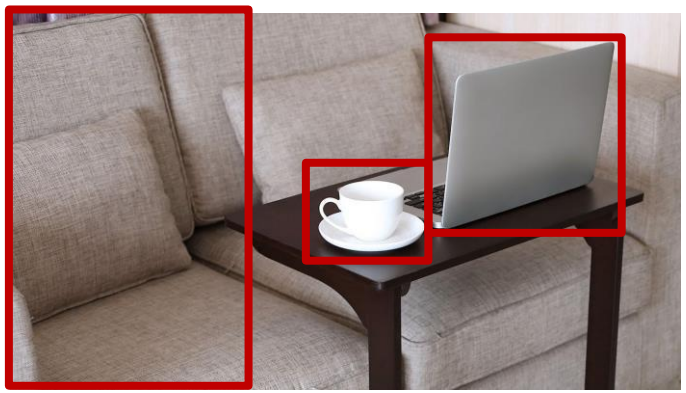
Dimensions of Heterogeneity

Information present in different modalities will often show diverse qualities, structures, and representations.

A teacup on the right of a laptop in a clean room.

Dimensions of Heterogeneity

Information present in different modalities will often show diverse qualities, structures, and representations.



*A **teacup** on the **right** of a **laptop** in a **clean room**.*

1. **Element representations:** discrete, continuous, granularity

▲ {  ,  ,  }

● *{teacup, right, laptop, clean, room}*

Dimensions of Heterogeneity

"Cross-Modal Discrete Representation Learning" by Alexander H. Liu et al

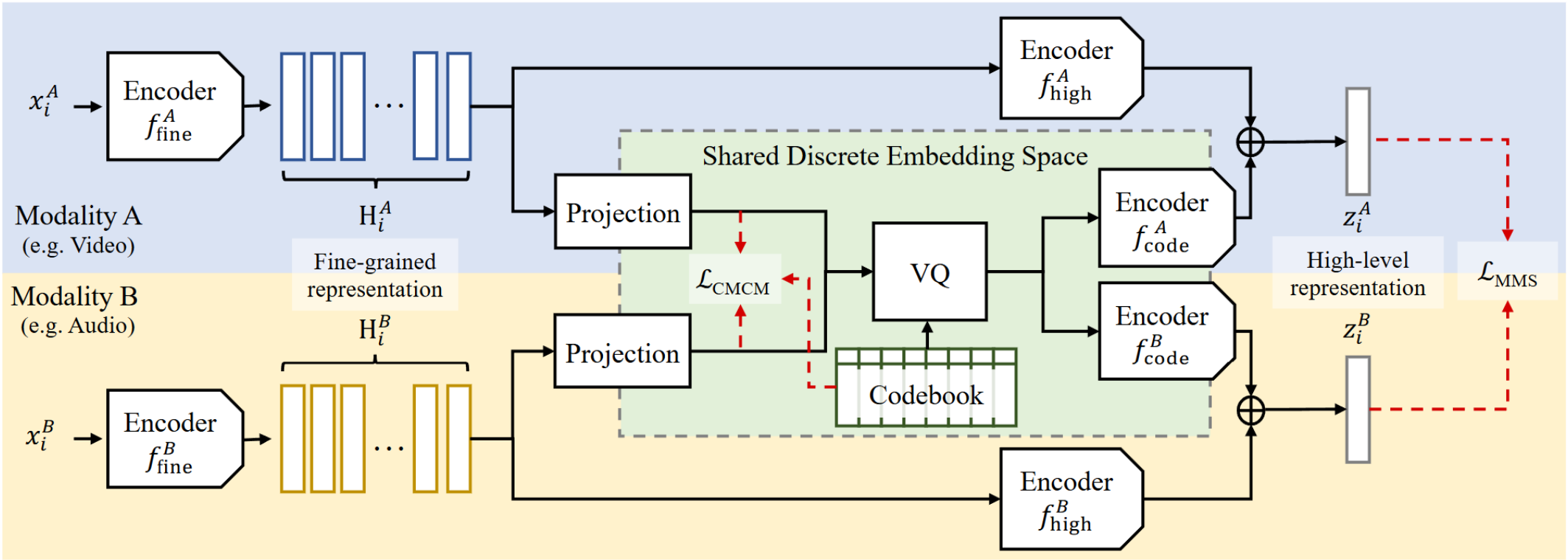


Figure 1: An overview of the proposed framework. The proposed shared discrete embedding space (green region, described in Section 2.2) is based on a cross-modal representation learning paradigm (blue and yellow regions, described in Section 2.1). The proposed Cross-Modal Code Matching $\mathcal{L}_{\text{CMCM}}$ objective is detailed in Section 2.3 and Figure 2.

Dimensions of Heterogeneity

2 **Element distributions:** density, frequency

objects per image

words per minute

•Different modalities have different **densities and frequencies** of information.

- Text: Words per minute.
- Vision: Objects per image.
- Audio: Frequencies of sound waves.

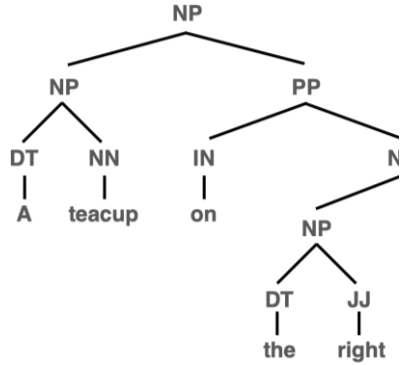
Example

•**Speech signals** are **dense** (continuous waveforms).

•**Text documents** are **sparse** (symbolic, limited vocabulary).

Dimensions of Heterogeneity

Information present in different modalities will often show diverse qualities, structures, and representations.



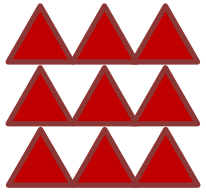
...

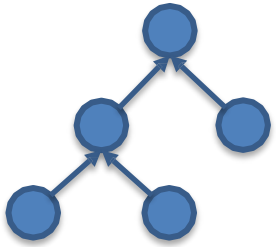
Latent (implicit)

A teacup on the right …

Explicit (observable)

3 **Structure:** temporal, spatial, hierarchical, latent, explicit





Dimensions of Heterogeneity

Structural Differences

•Data can be **temporal, spatial, hierarchical, latent, or explicit**:

- **Temporal**: Speech signals evolve over time.
- **Spatial**: Objects in an image have spatial relationships.
- **Hierarchical**: Sentences follow syntax trees.
- **Latent**: Hidden features extracted by deep learning models.
- **Explicit**: Directly observable structures.

Example

•**Speech:** Time-series structure (temporal).

•**Image:** Object layout in a scene (spatial).

•**Text:** Syntactic dependency trees (hierarchical).

A teacup on the right …

Dimensions of Heterogeneity

Information present in different modalities will often show diverse qualities, structures, and representations.

A teacup on the right of a laptop in a clean room.

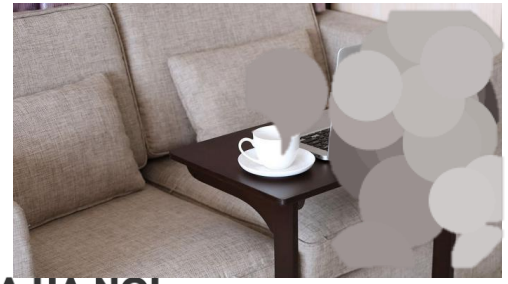
4 **Information:** abstraction, entropy

Dimensions of Heterogeneity

Information present in different modalities will often show diverse qualities, structures, and representations.

A teacup on the right of a laptop in a clean room.

5 **Noise:** uncertainty, signal-to-noise ratio, missing data



teacup → teacip

right → rihjt

Dimensions of Heterogeneity

Information present in different modalities will often show diverse qualities, structures, and representations.

A teacup on the right of a laptop in a clean room.

6 **Relevance:** task relevance, context dependence

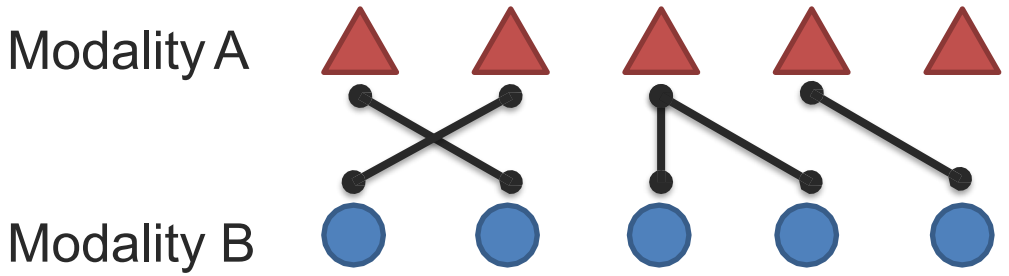
→ recreational
→ living room
→ right-handed

A teacup on the right of a laptop in a clean room. → workspace
→ study room

Interconnected Modalities

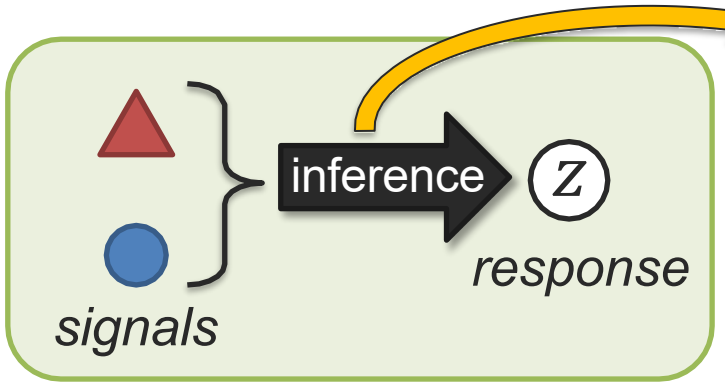
① Modality connections

Modalities are often related and share commonality



② Modality interactions

Modality elements often interact during inference



Interactions happen during inference!

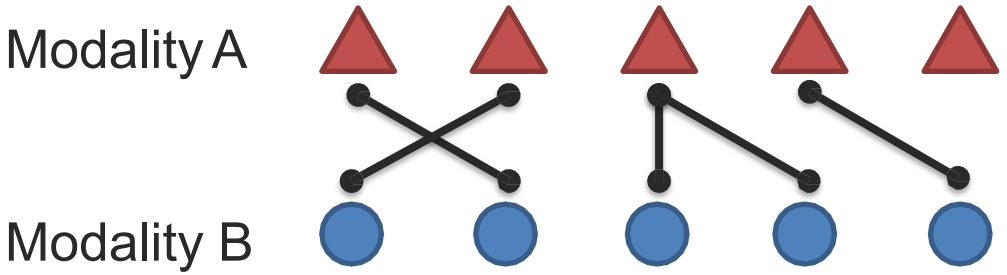
"Inference" examples:

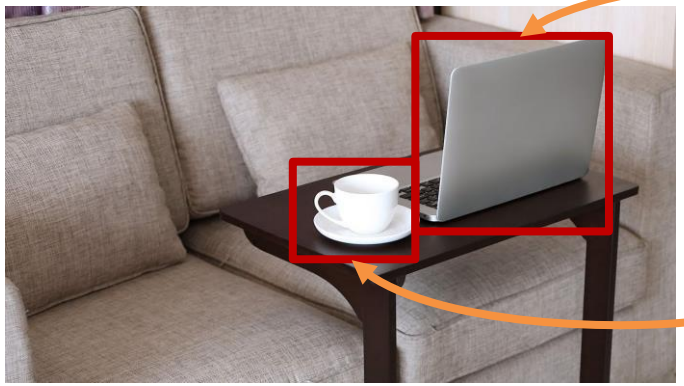
- Behavior perception
- Recognition task
- Modality translation

Interconnected Modalities

① Modality connections

Modalities are often related and share commonality



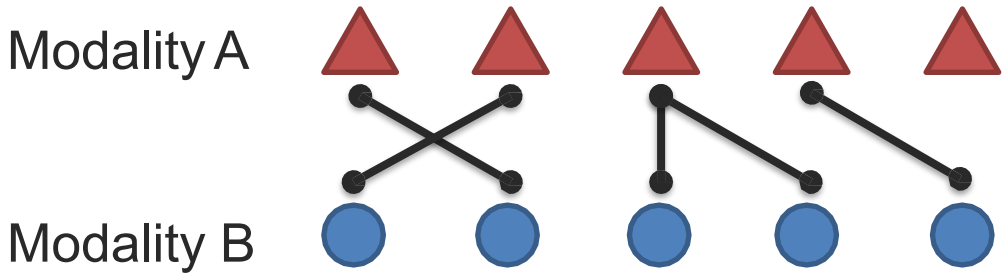


*A **teacup** on the right of a **laptop** in a clean room.*

Interconnected Modalities

① Modality connections

Modalities are often related and share commonality

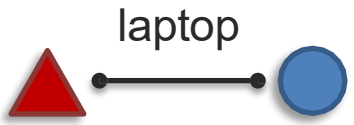


Statistical → **Semantic**

Association

e.g., correlation, co-occurrence

Correspondence

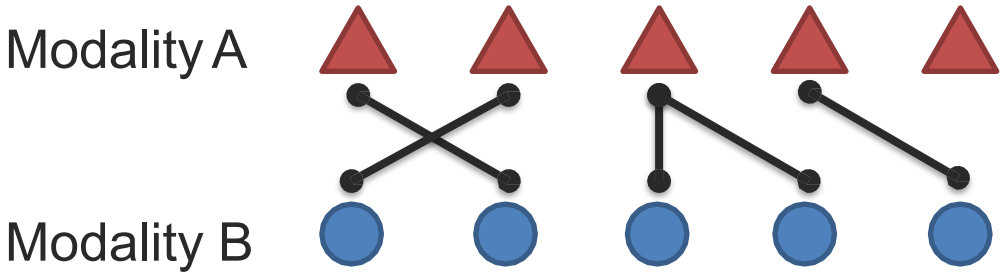


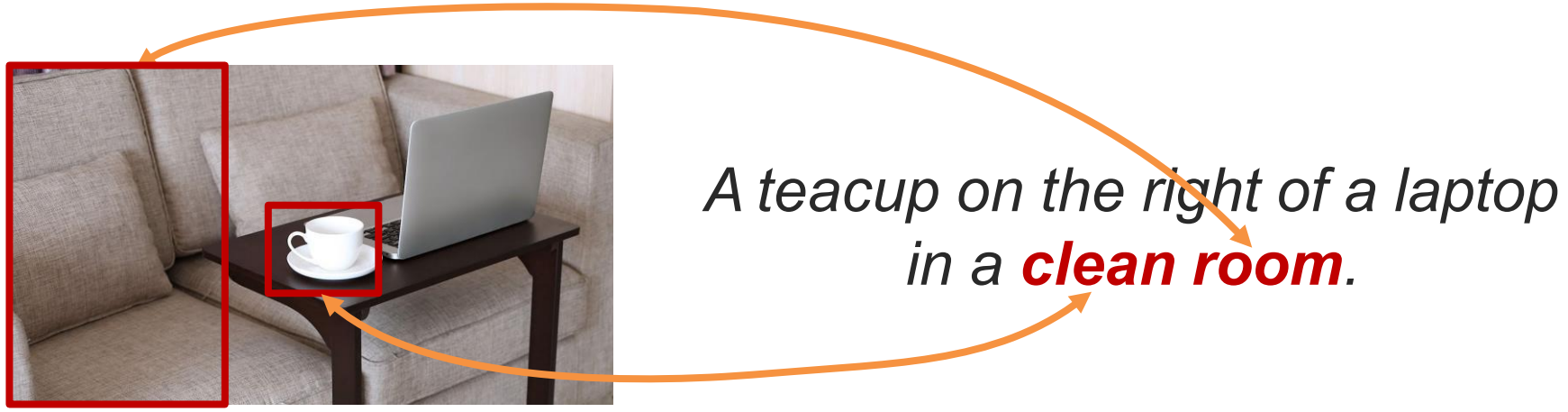
e.g., grounding

Interconnected Modalities

① Modality connections

Modalities are often related and share commonality

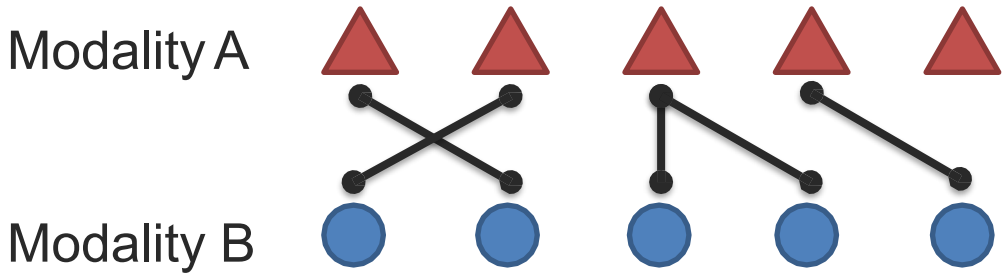




Interconnected Modalities

① Modality connections

Modalities are often related and share commonality



Statistical

Association

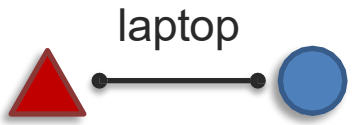
e.g., correlation, co-occurrence

Dependency

e.g., causal, temporal

Semantic

Correspondence



e.g., grounding

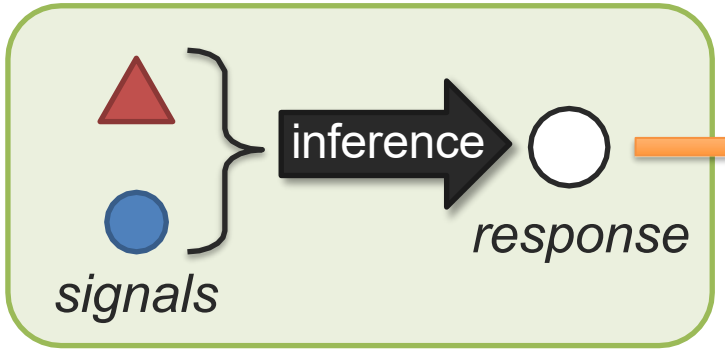
Relationship

e.g., function

Interconnected Modalities

② Modality interactions

Modality elements often interact during inference



Types of interaction responses?
(a taxonomy)

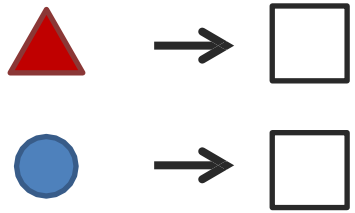
Is this indoors?

inference *Yes!*

A teacup on the right of a laptop in a clean room.

inference *Yes!*

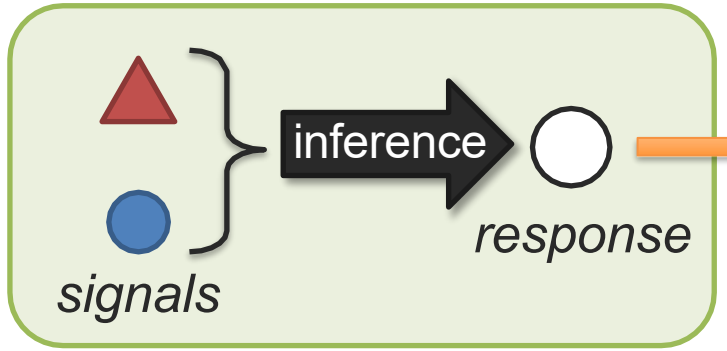
Unimodal redundancy



Interconnected Modalities

② Modality interactions

Modality elements often interact during inference

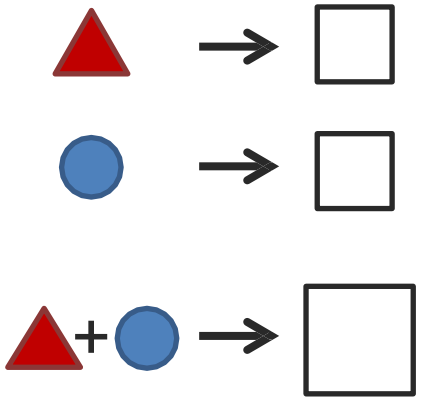


Is this indoors ?

A teacup on the right of a laptop in a clean room.

inference ***Yes!***

Unimodal redundancy

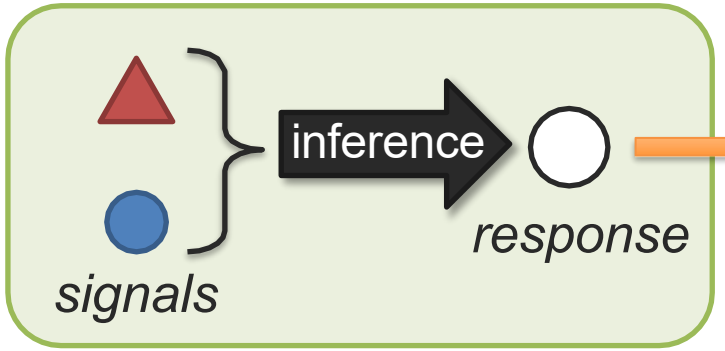


Multimodal enhancement

Interconnected Modalities

② Modality interactions

Modality elements often interact during inference



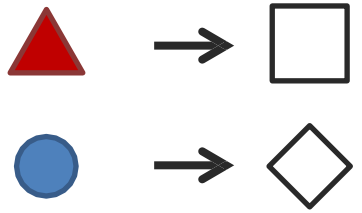
Is this a living room?

A teacup on the right of a laptop in a clean room.

inference → *Yes!*

inference → *No, probably study room.*

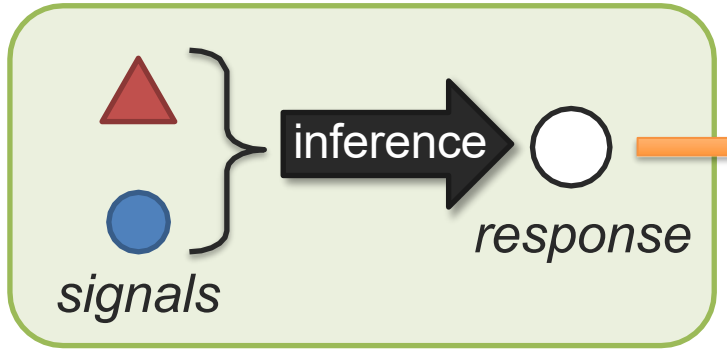
Unimodal Non-redundancy



Interconnected Modalities

② Modality interactions

Modality elements often interact during inference

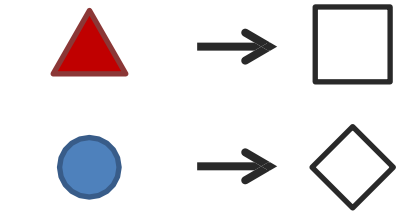


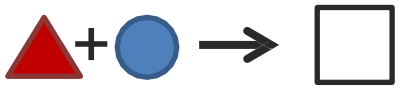
Is this a living room?

A teacup on the right of a laptop in a clean room.

inference ***Yes!***

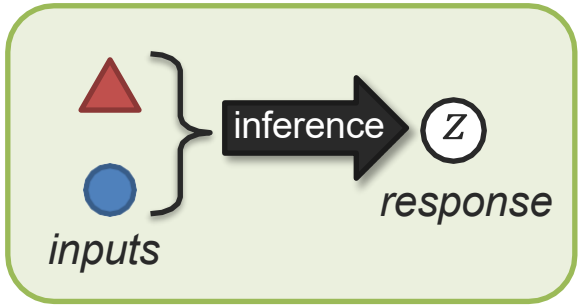
Unimodal Non-redundancy





Multimodal dominance

Taxonomy of Interaction Responses – A Behavioral Science View



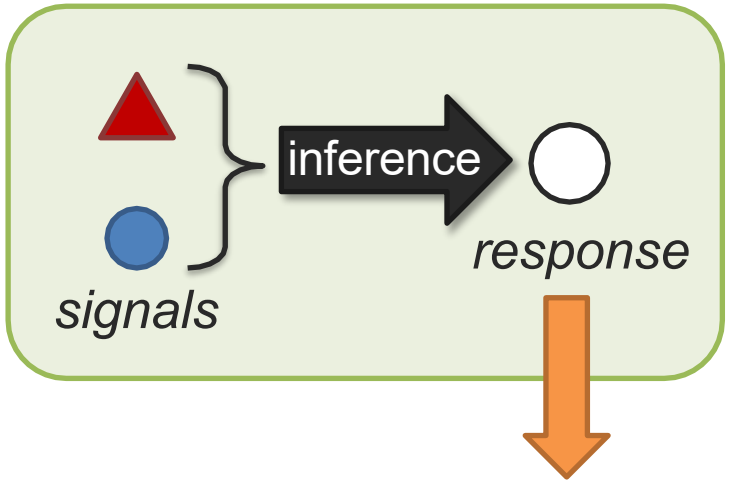
Multimodal Communication

	signal	response	signal	response	
Redundancy	a. →	□	a+b →	□	Equivalence
	b. →	□	a+b →	□	Enhancement
Nonredundancy	a →	□	a+b →	□ and ○	Independence
	b →	○	a+b →	□	Dominance
			a+b →	□ (or □)	Modulation
			a+b →	△	Emergence

Partan and Marler (2005). *Issues in the classification of multimodal communication signals*. American Naturalist, 166(2)

Dimensions of Modality Interactions

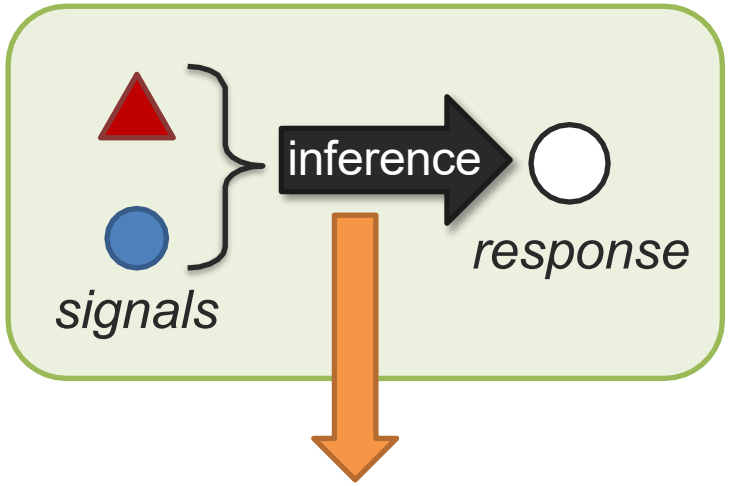
What are the dimensions for **digitally-represented** modalities?



1. Interaction Responses:
   - Redundancy
   - Non-redundancy
   - Dominance
   - Emergence…

Dimensions of Modality Interactions

What are the dimensions for **digitally-represented** modalities?

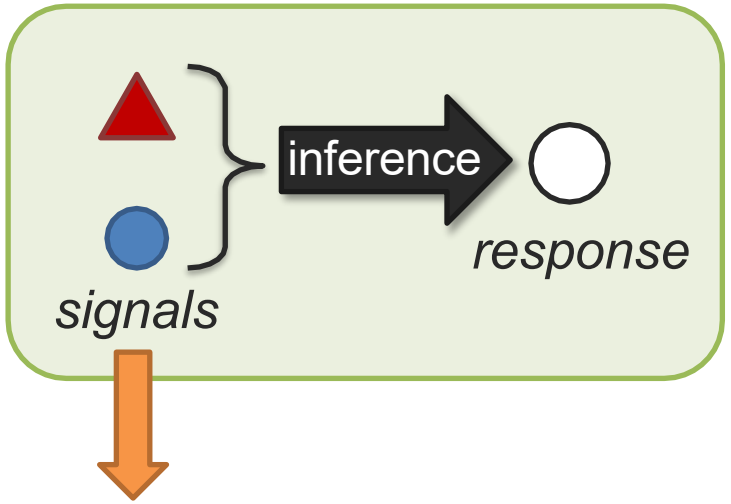


2 Interaction Mechanics:

- Additive
- multiplicative
- Nonlinear
- Causal,
- Logical, …

Dimensions of Modality Interactions

What are the dimensions for **digitally-represented** modalities?

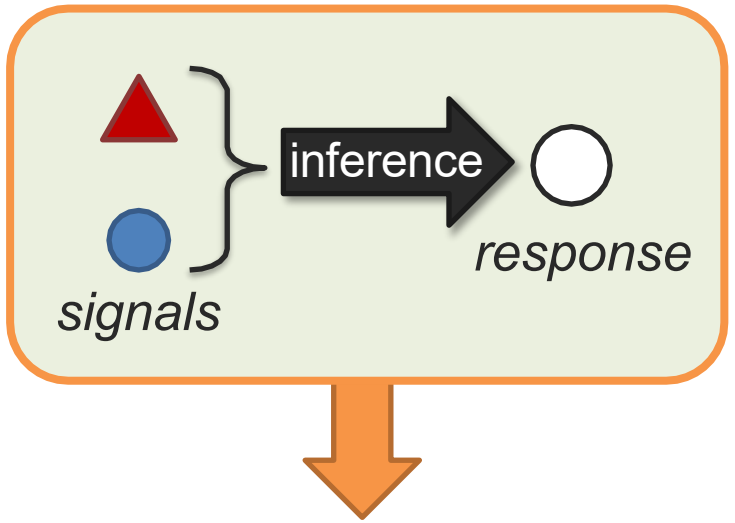


3 Input modalities:

- Unimodal
- Bimodal
- Trimodal
- High-modal, ...

Dimensions of Modality Interactions

What are the dimensions for **digitally-represented** modalities?



4 Context:

- Structure context
- Task relevance
- Context dependence
- High-modal, ...

What is Multimodal?

Multimodal is the science of

heterogeneous and interconnected data ☺

Multimodal Machine Learning

What is Multimodal Machine Learning?

Multimodal Machine Learning (ML) 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**

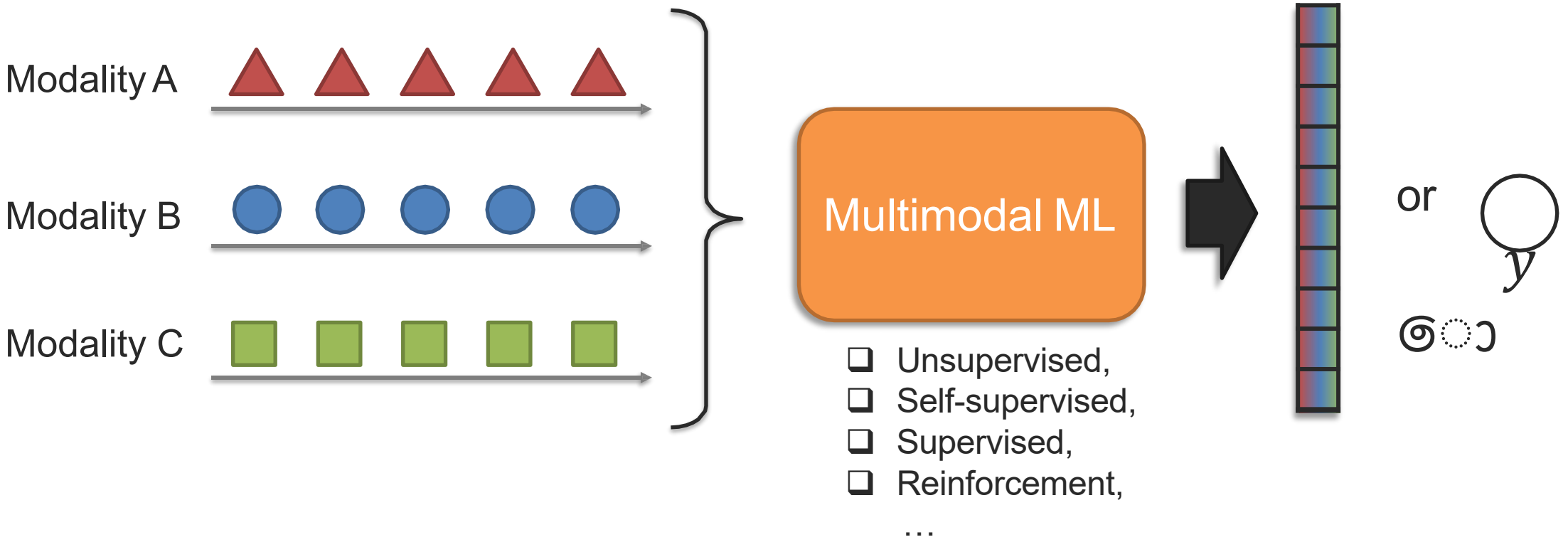
Multimodal Machine Learning

Language I really like this tutorial

Vision 

Acoustic 

Multimodal Machine Learning



Multimodal Machine Learning

*What are the **core multimodal technical challenges,***

understudied in conventional machine learning?

Multimodal Technical Challenges – Surveys, Tutorials and Courses

2016

Multimodal Machine Learning: A Survey and Taxonomy

Tadas Baltrusaitis, Chaitanya Ahuja and Louis-Philippe Morency

(Arxiv 2017, IEEE TPAMI journal, February 2019)

https://arxiv.org/abs/1705.09406

Tutorials: CVPR 2016, ACL 2016, ICMI 2016, …

Graduate-level courses:

Multimodal Machine learning (11th edition)

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

Advanced Topics in Multimodal Machine learning

https://cmu-multicomp-lab.github.io/adv-mmml-course/spring2022/

2022

Fundamentals of Multimodal ML: A Taxonomy & Open Challenges

Paul Liang, Amir Zadeh and Louis-Philippe Morency

☑ 6 core challenges
☑ 50+ taxonomic classes
☑ 600+ referenced papers

Tutorials: CVPR 2022, NAACL 2022, …

Updated graduate-level course:

Multimodal Machine learning (12th edition)

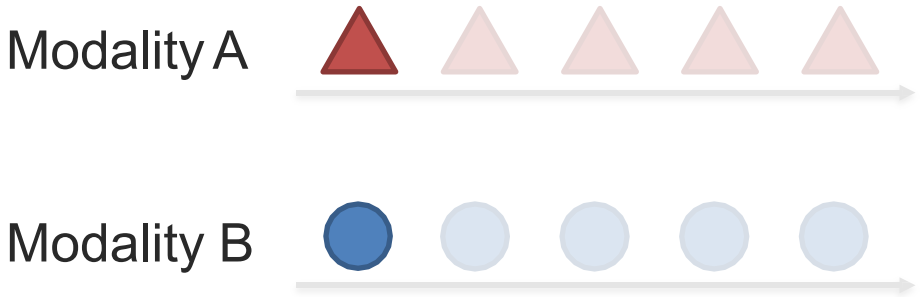
Fall 2022 semester

Challenge 1: Representation

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

This is a **core challenge** in multimodal learning because different modalities (text, images, audio, etc.) have **diverse structures** and need to be unified in a meaningful way.

Individual elements:



It can be seen as a "local" representation

or

representation using holistic features

Challenge 1: Representation

Example 1: Vision-Language Representation (CLIP)

- **Task**: Match images with text descriptions.
- **Problem**: Text and images have different feature spaces.
- **Solution**: CLIP (Contrastive Language–Image Pretraining) learns a **shared latent space** where related text and images are close together.

Learning Transferable Visual Models From Natural Language Supervision
Radford 2021

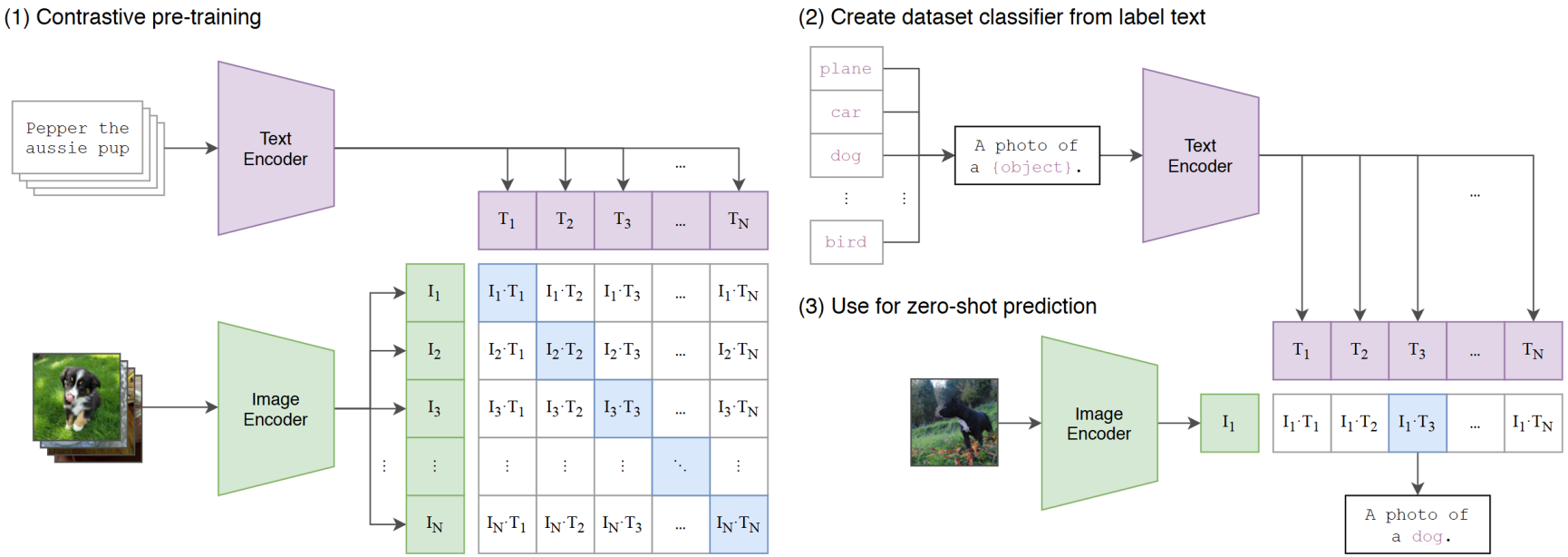


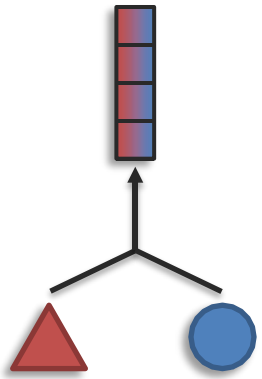
Figure 1. Summary of our approach. While standard image models jointly train an image feature extractor and a linear classifier to predict some label, CLIP jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) training examples. At test time the learned text encoder synthesizes a zero-shot linear classifier by embedding the names or descriptions of the target dataset's classes.

Challenge 1: Representation

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

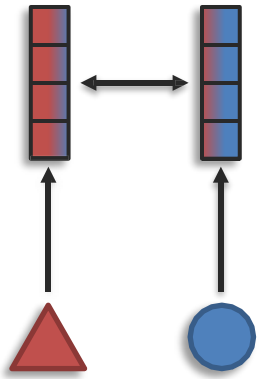
Sub-challenges:

Fusion



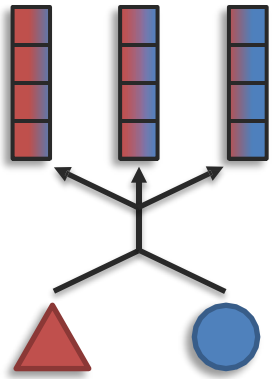
modalities **>** # representations

Coordination



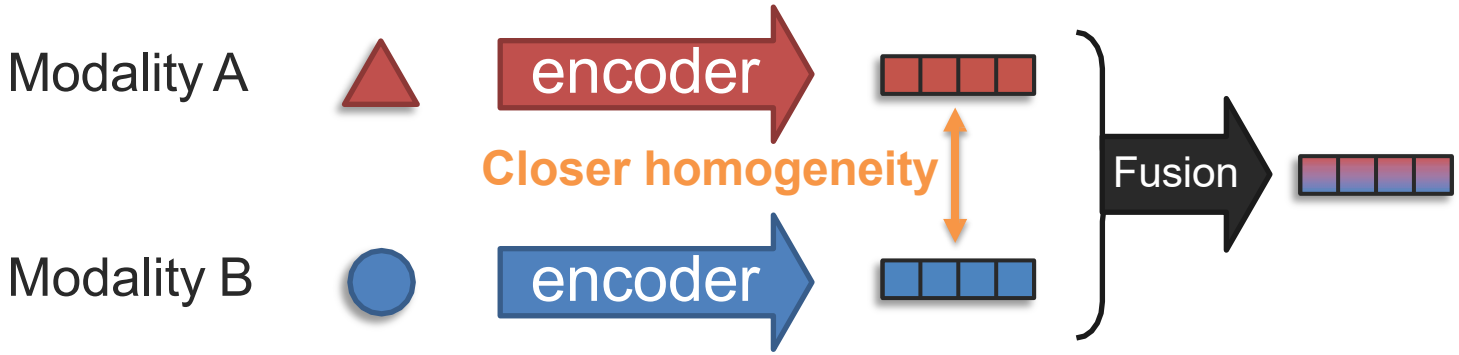
modalities **=** # representations

Fission

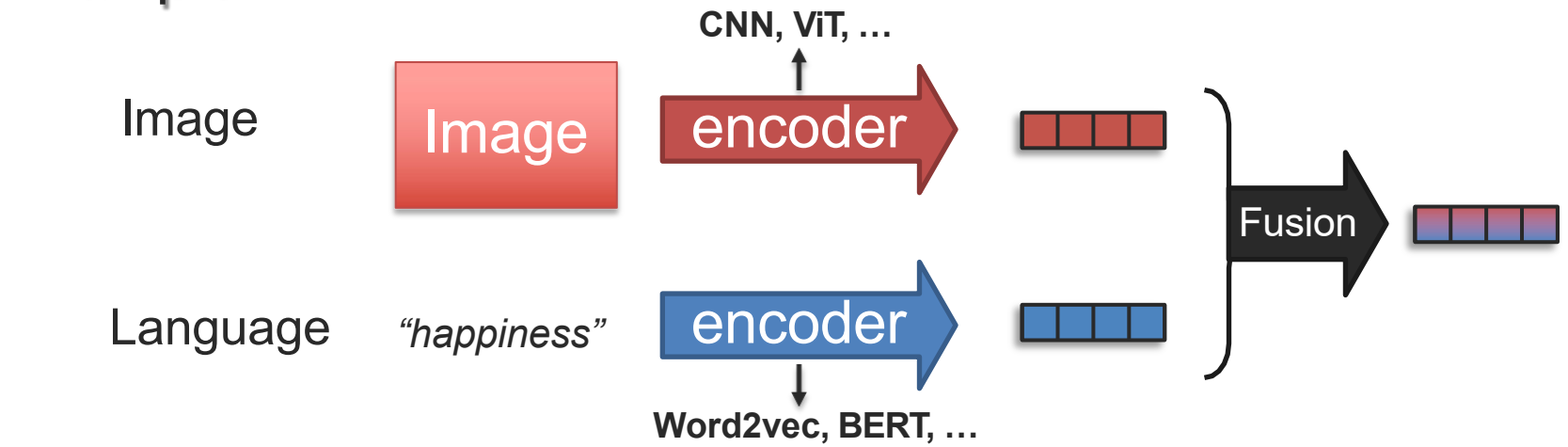


modalities **<** # representations

Fusion with Unimodal Encoders

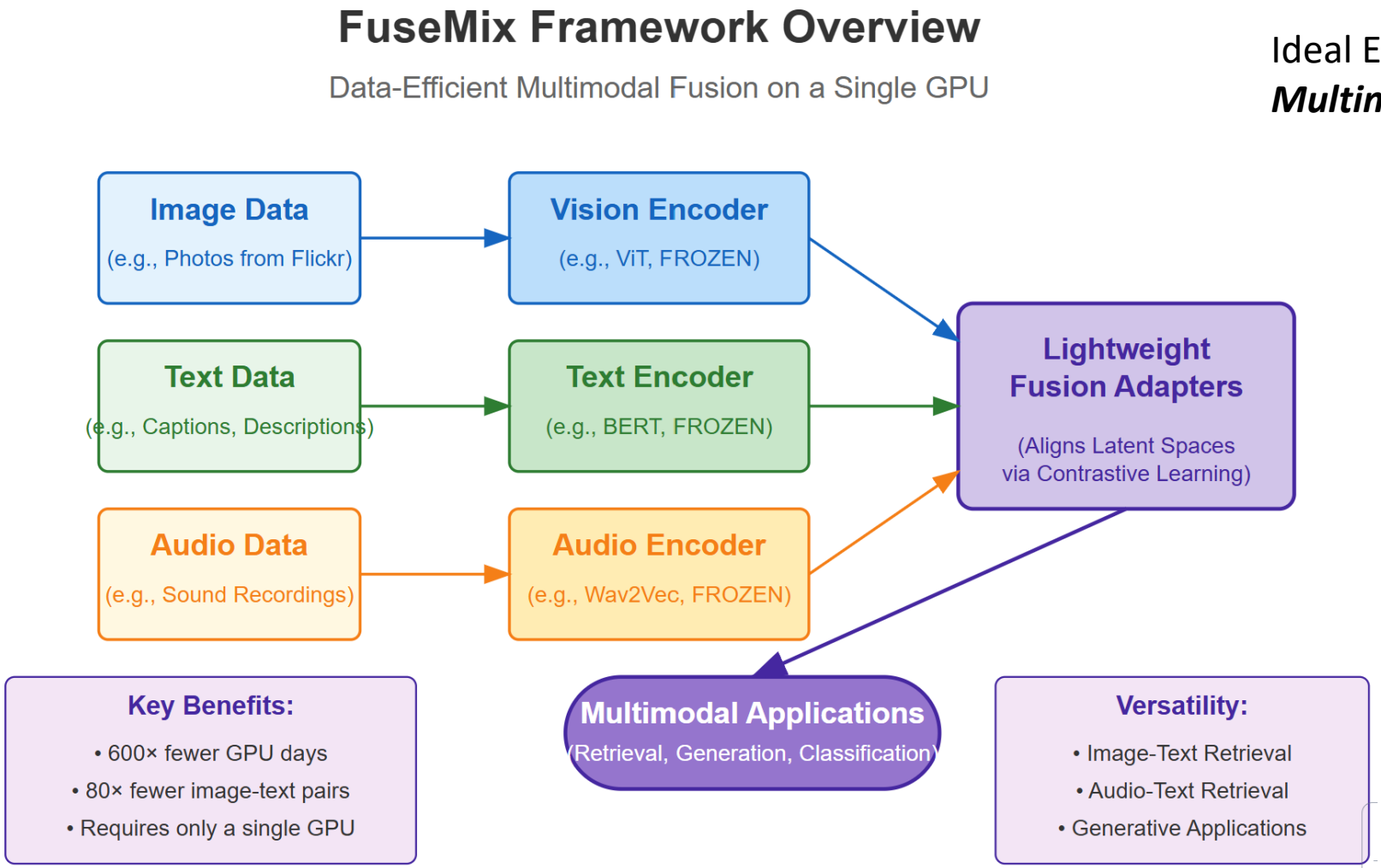


Example:



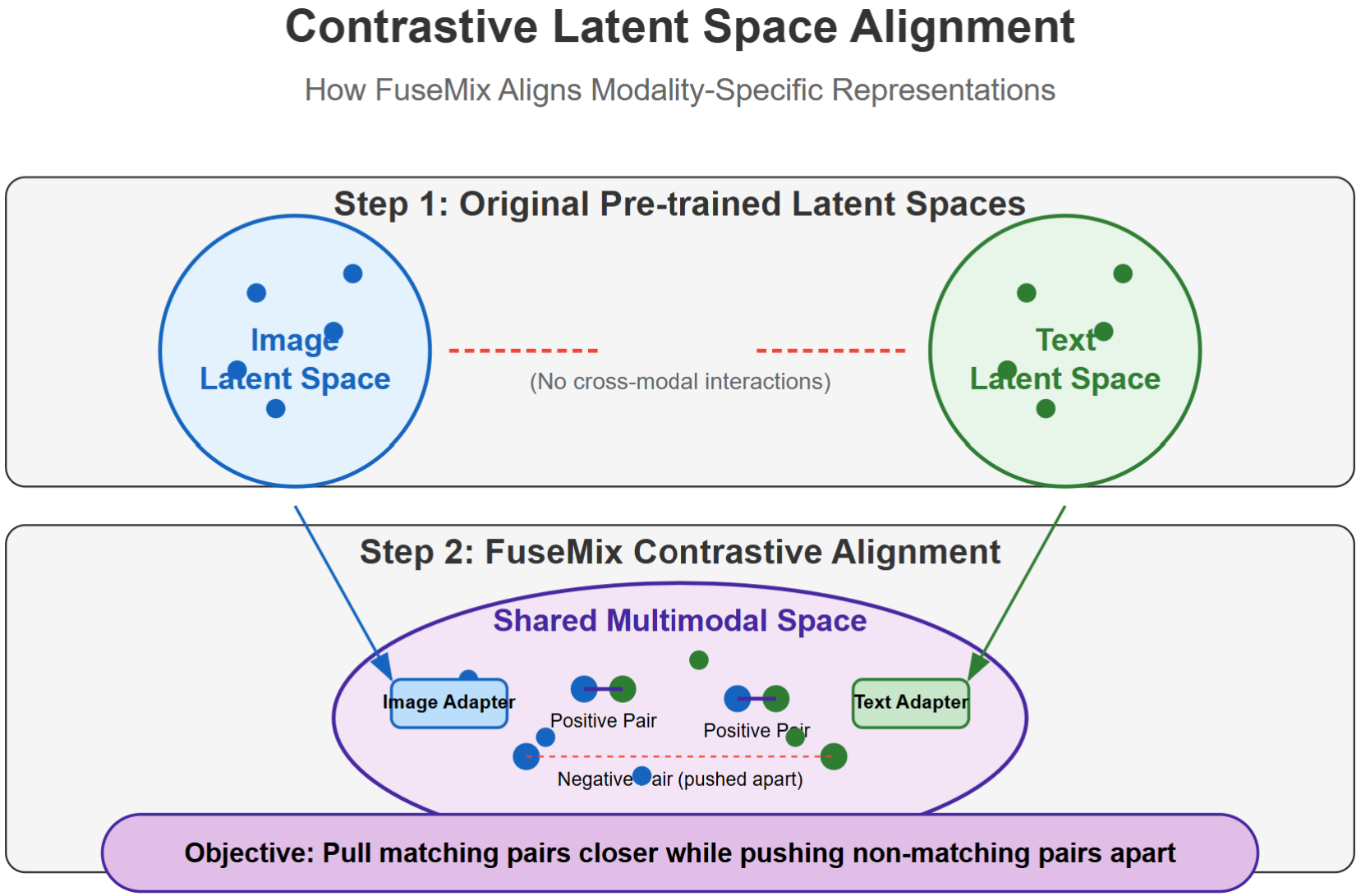
- Unimodal encoders can be jointly learned with fusion network, or pre-trained

Example paper



Ideal Example Paper: "***Data-Efficient Multimodal Fusion on a Single GPU***"

Example paper



Ideal Example Paper: "Data-Efficient Multimodal Fusion on a Single GPU"

Challenge 1: Representation

Dynamic Multimodal Fusion 2023

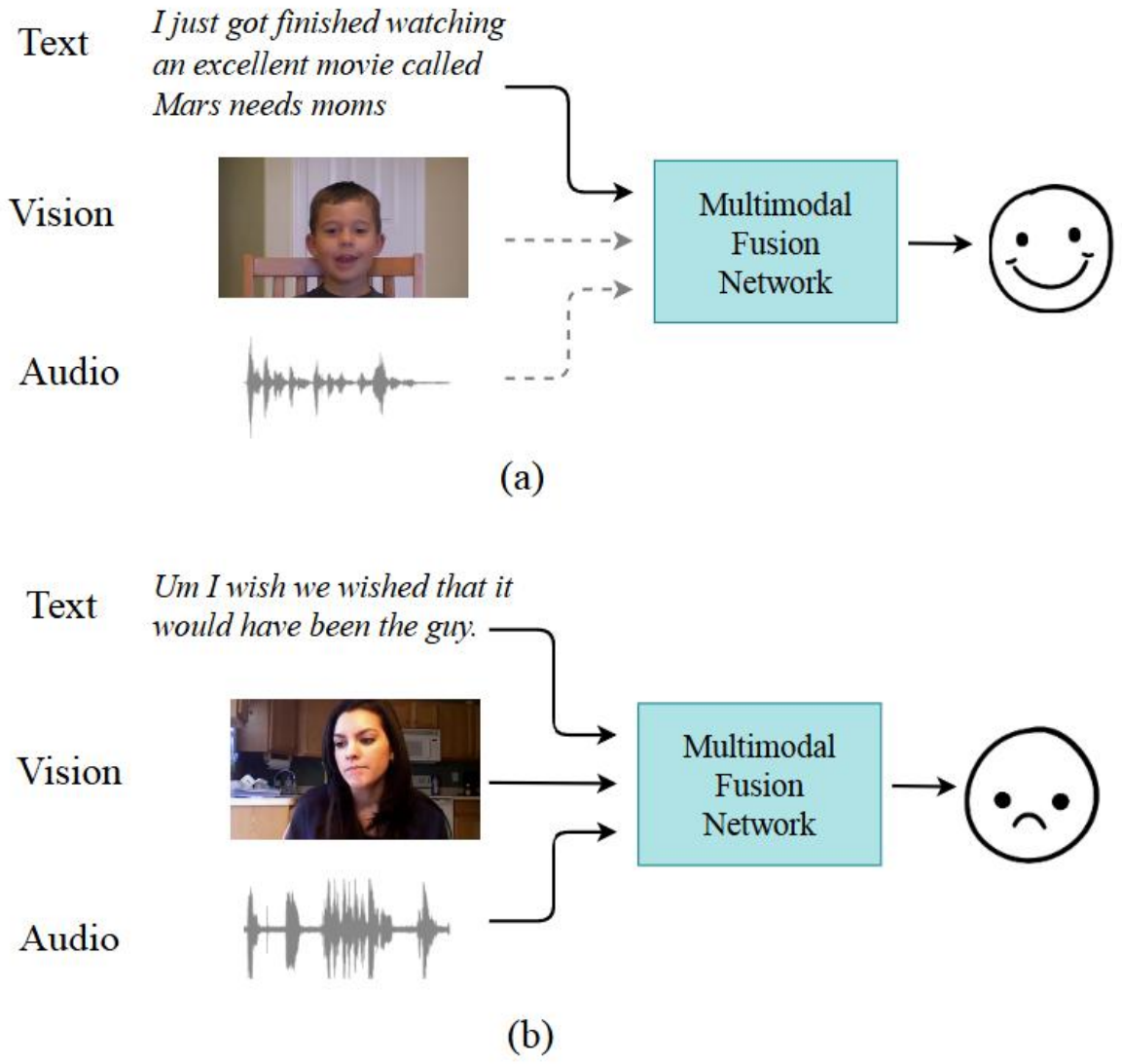
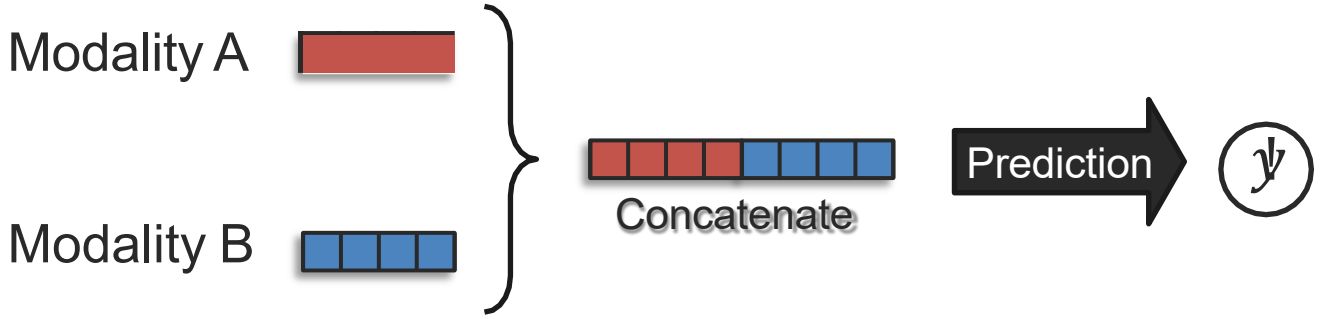


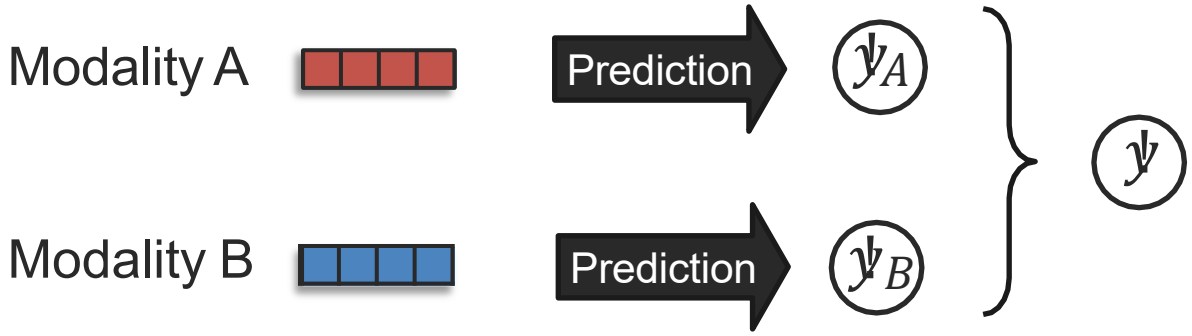
Figure 1. Two examples in CMU-MOSEI [51] for emotion recognition. Figure (a) shows an "easy" multimodal instance as using textual information is sufficient to predict emotions correctly (this is a positive emotion). Figure (b) shows a "hard" example where all three modalities are required to make correct predictions (this is a negative emotion). While static multimodal fusion networks process "hard" and "easy" inputs identically, we propose *dynamic instance-wise inference* that can achieve computational savings for "easy" examples and preserve representation power for "hard" instances. For (a), DynMM only activates the text path and skips paths corresponding to the other two modalities, thus leading to computational efficiency.

Early and Late Fusion – A historical View

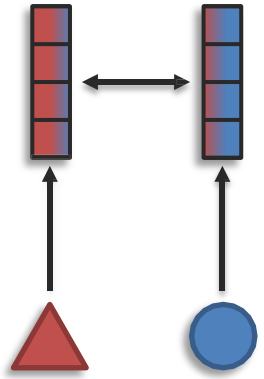
Early fusion:



Late fusion:

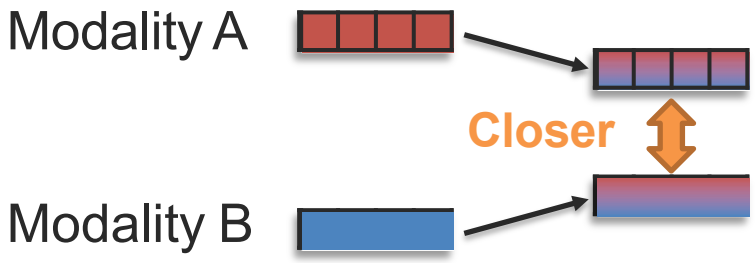


Sub-Challenge 1b: Representation Coordination

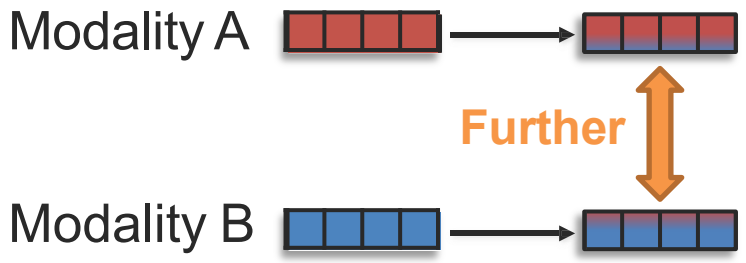


Definition: Learn multimodally-contextualized representations that are coordinated through their cross-modal interactions

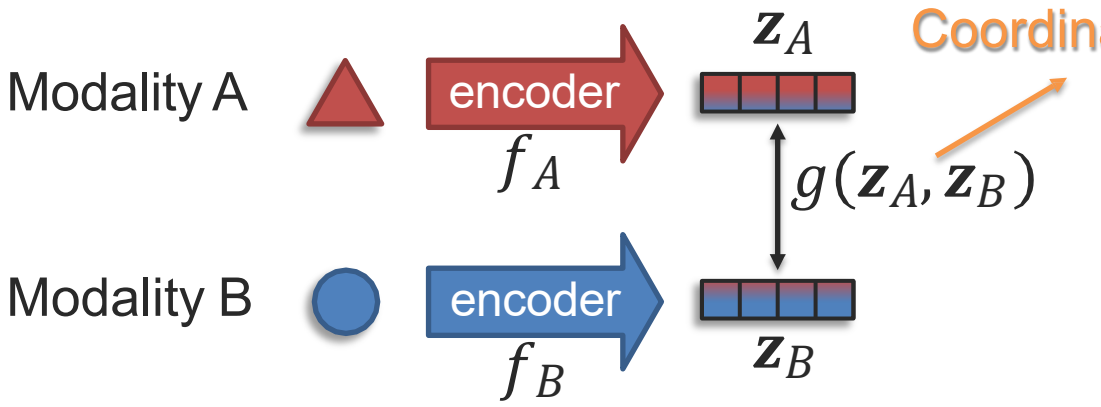
Strong Coordination:



Partial Coordination:



Coordination Function



Learning with coordination function:

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

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

➡ Requires paired data

Examples of coordination function:

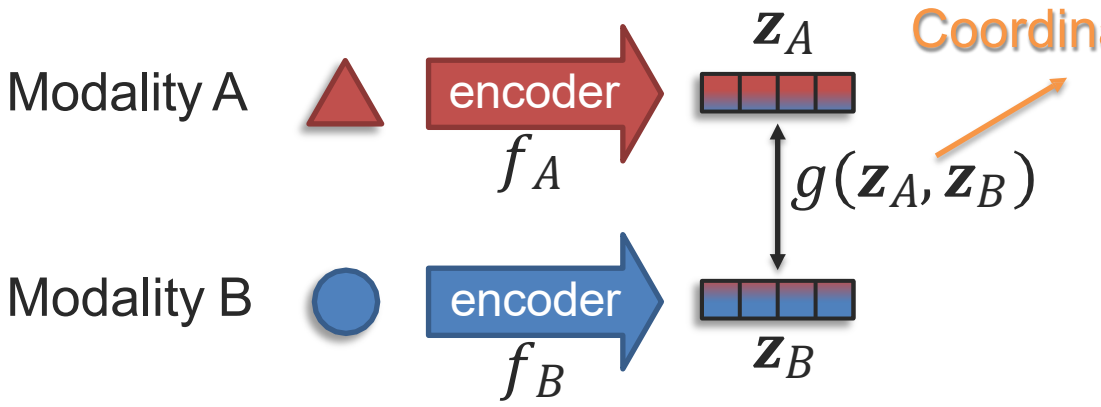
1 Cosine similarity:

$$g(z_A, z_B) = \frac{z_A \& z_B}{\|z_A\| \|z_B\|}$$

Strong coordination!

➡ For normalized inputs (e.g., $z_A - z_A$) , equivalent to *Pearson correlation coefficient*

Coordination Function



Learning with coordination function:

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

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

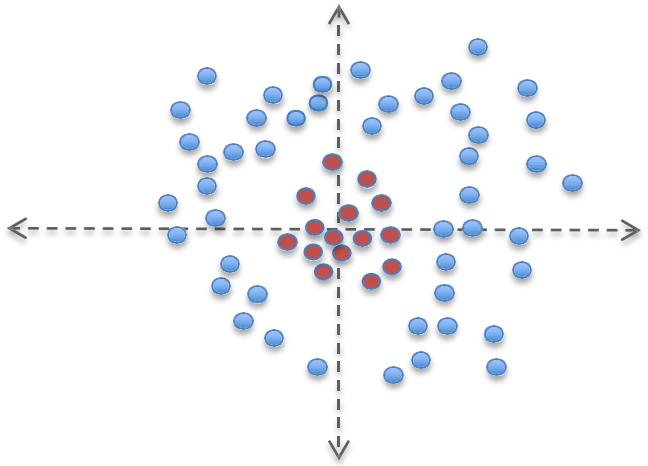
All these examples bring relatively strong coordination between modalities

Kernel Function

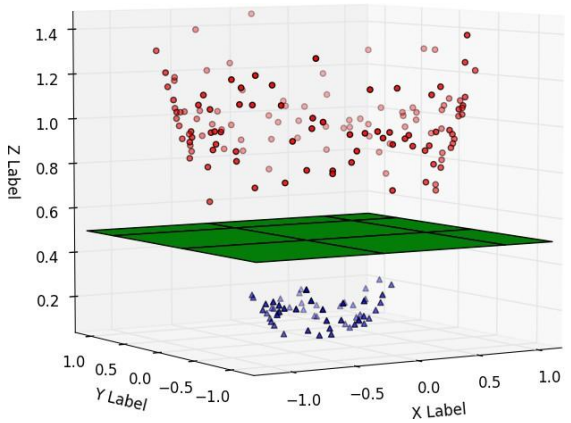
A kernel function: Acts as a similarity metric between data points

$$K(\boldsymbol{x}_i, \boldsymbol{x}_\mathrm{j}) = \phi(\boldsymbol{x}_i)^T \phi(\boldsymbol{x}_\mathrm{j}) = \langle \phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}_\mathrm{j}) \rangle$$

$\phi(\boldsymbol{x})$ can be high-dimensional space!



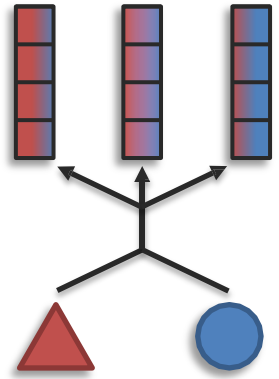
Not linearly separable in $\boldsymbol{x}$ space



Same data, but now linearly separable in $\phi(\boldsymbol{x})$ space

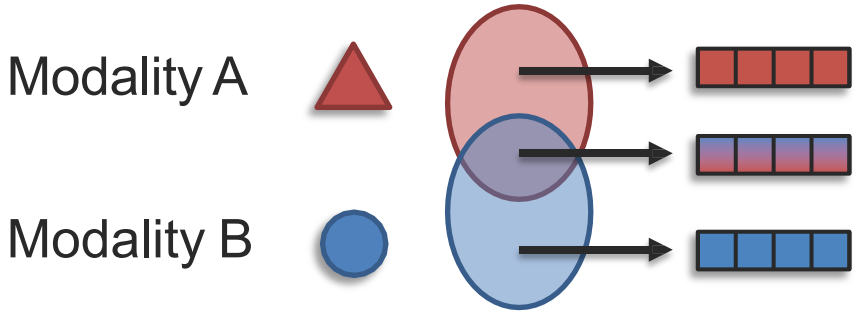
Radial Basis Function (RBF) Kernel : $K(x_i, x_\mathrm{j}) = \exp -\frac{1}{2\sigma^2} \|x_i - x_\mathrm{j}\|^2$

Sub-Challenge 1c: Representation Fission

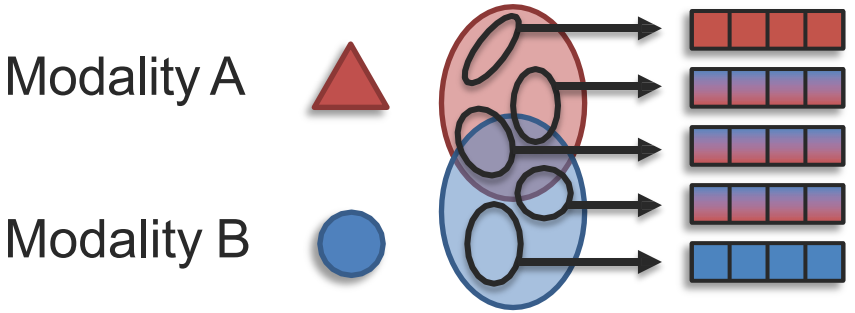


Definition: Learning a new set of representations that reflects multimodal internal structure such as data factorization or clustering

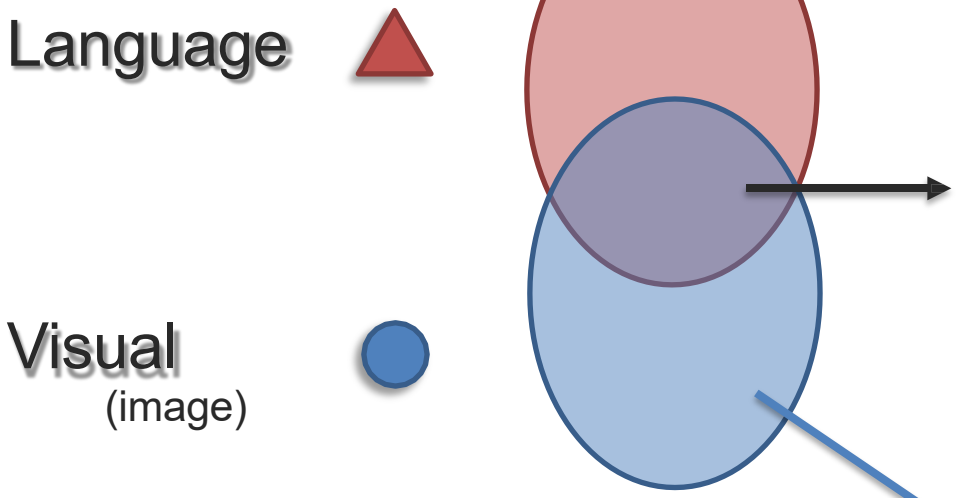
Modality-level fission:



Fine-grained fission:



Modality-Level Fission



Information primarily in language modality

- Syntactic structure
- Vocabulary, morphology
- …

Information in both modalities

- Described people, objects, actions
- Illustrative gestures, motion
- …

Information primarily in visual modality

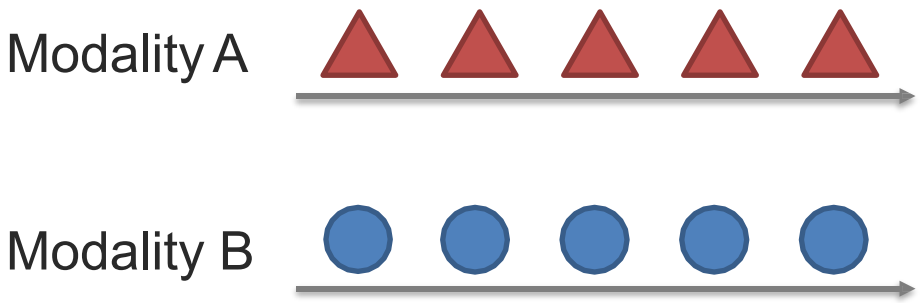
- Texture, visual appearance
- Depth, perspective, motion
- …

Challenge 2: Alignment

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

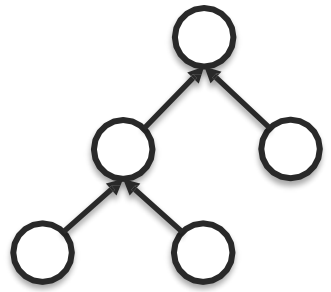
➡ Most modalities have internal structure with multiple elements

Elements with temporal structure:



Other structured examples:

Spatial



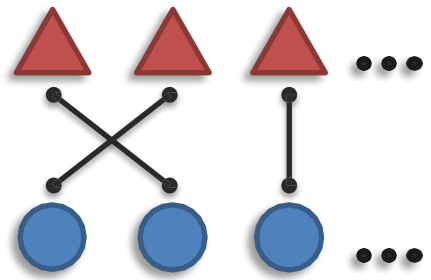
Hierarchical

Challenge 2: Alignment

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

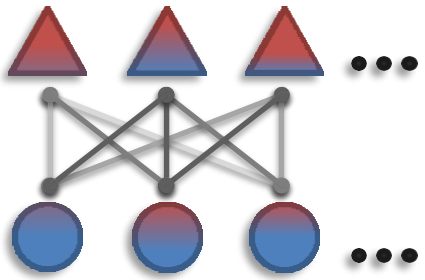
Sub-challenges:

Connections



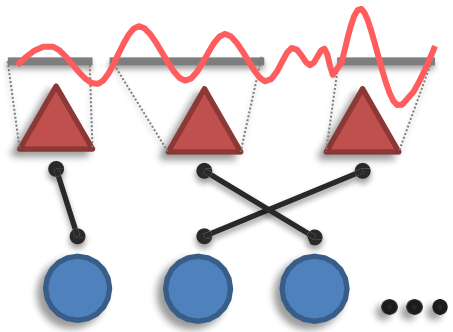
Explicit alignment
(e.g., grounding)

Aligned Representation



Alignment + representation
(aka, contextualized representation)

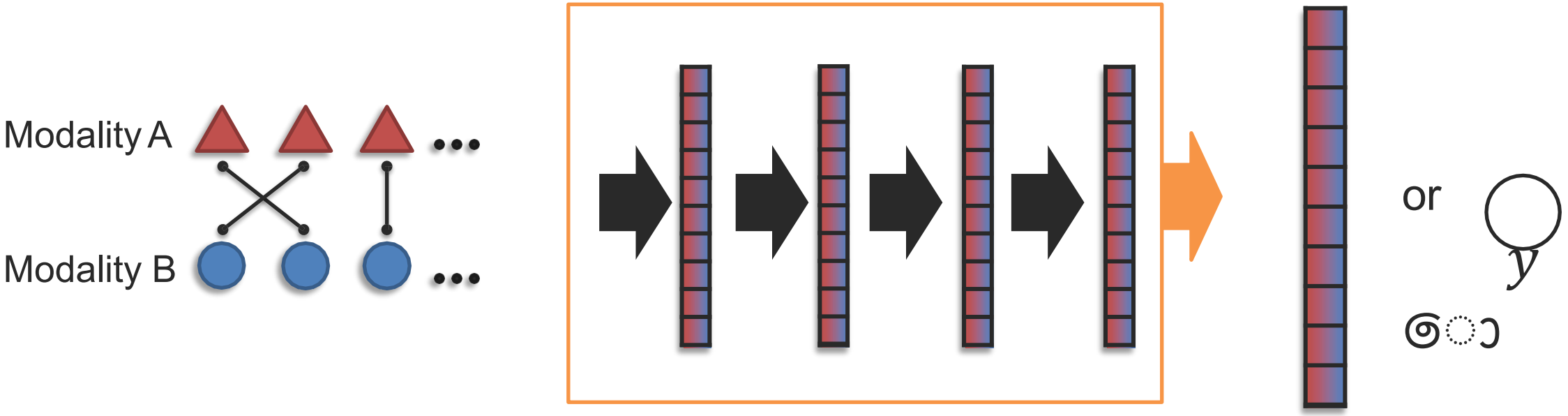
Elements



Segmentation of individual elements

Challenge 3: Reasoning

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



Challenge 3: Reasoning

"Learning to Reason: End-to-End Module Networks for Visual Question Answering"

Andreas et al., NeurIPS 2016

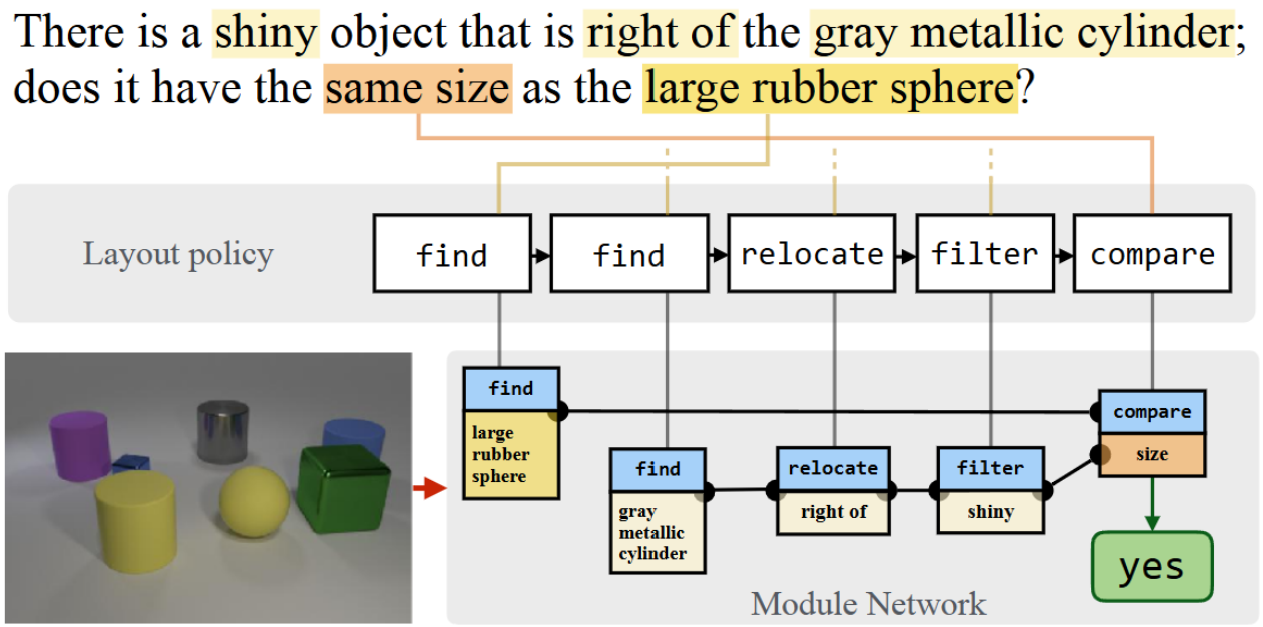
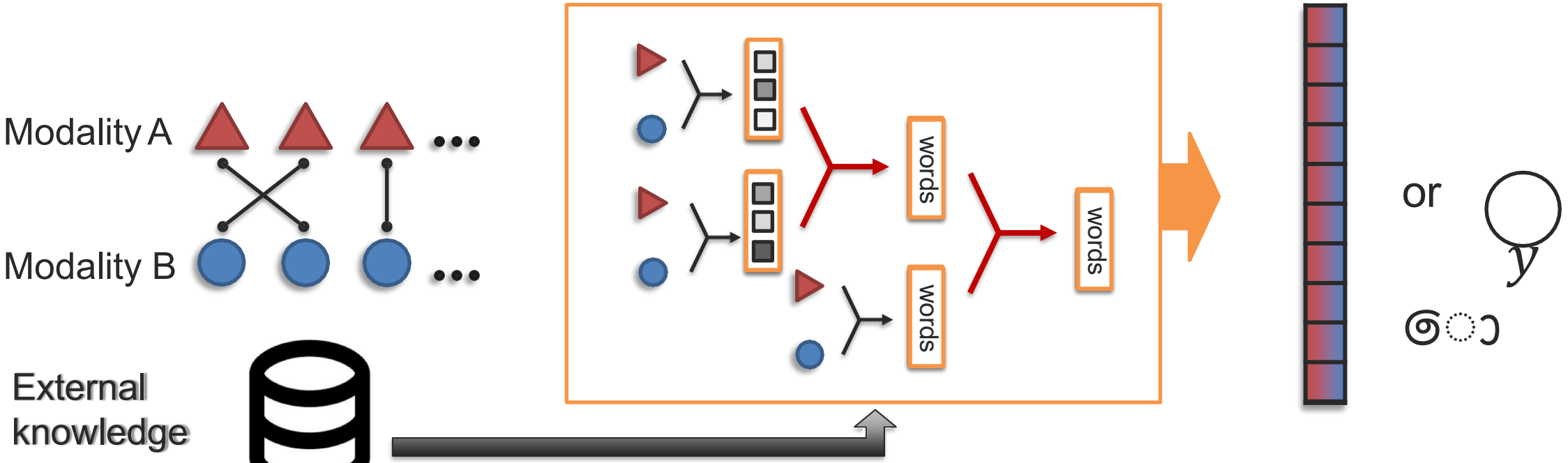


Figure 1: For each instance, our model predicts a computational expression and a sequence of attentive module parameterizations. It uses these to assemble a concrete network architecture, and then executes the assembled neural module network to output an answer for visual question answering. (The example shows a real structure predicted by our model, with text attention maps simplified for clarity.)

Challenge 3: Reasoning

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



Challenge 3: Reasoning

Q: Which American president is associated with the stuffed animal seen here?

A: Teddy Roosevelt

"OK-VQA: A Visual Question Answering Benchmark Requiring External Knowledge"

Marino et al., CVPR 2019

Outside Knowledge

Another lasting, popular legacy of Roosevelt is the stuffed toy bears—teddy bears—named after him following an incident on a hunting trip in Mississippi in 1902.

Developed apparently simultaneously by toymakers ... and named after President Theodore "Teddy" Roosevelt, the teddy bear became an iconic children's toy, celebrated in story, song, and film.

At the same time in the USA, Morris Michtom created the first teddy bear, after being inspired by a drawing of Theodore "Teddy" Roosevelt with a bear cub.

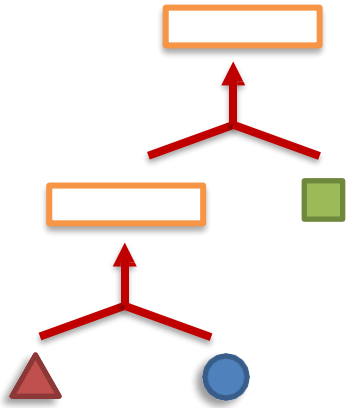
Figure 1: We propose a novel dataset for visual question answering, where the questions require external knowledge resources to be answered. In this example, the visual content of the image is not sufficient to answer the question. A set of facts about teddy bears makes the connection between teddy bear and the American president, which enables answering the question.

Challenge 3: Reasoning

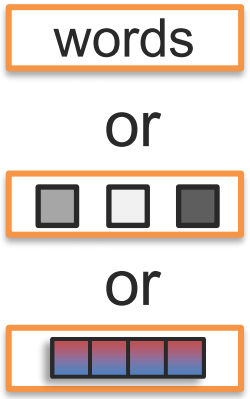
Definition: Combining knowledge, usually through multiple inferential steps, exploiting multimodal alignment and problem structure

Sub-challenges:

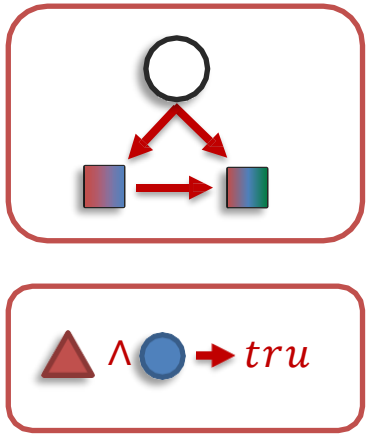
Structure Modeling



Intermediate concepts



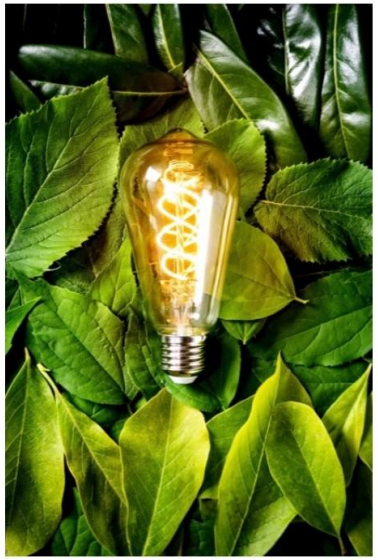
Inference Paradigm



External Knowledge

The Challenge of Compositionality

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



(a) some plants surrounding a lightbulb

(b) a lightbulb surrounding some plants

CLIP, ViLT, ViLBERT, etc.
All random chance

Compositional Generalization
to novel combinations outside of training data

1. Structure: <subject> <verb> <object>
2. Concepts: ‘plants’, ‘lightbulb’
3. Inference: ‘surrounding’ – spatial relation
4. Knowledge: from humans!

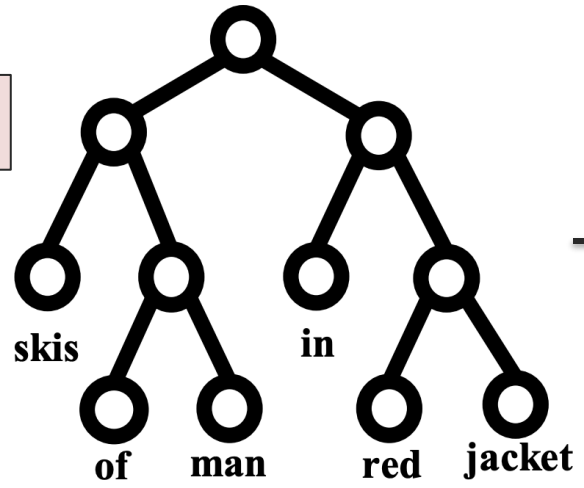
[Thrush et al., Winoground: Probing Vision and Language Models for Visio-Linguistic Compositionality. CVPR 2022]

Hierarchical Structure

Leverage syntactic structure of language

Parse

Skis of man in red jacket.

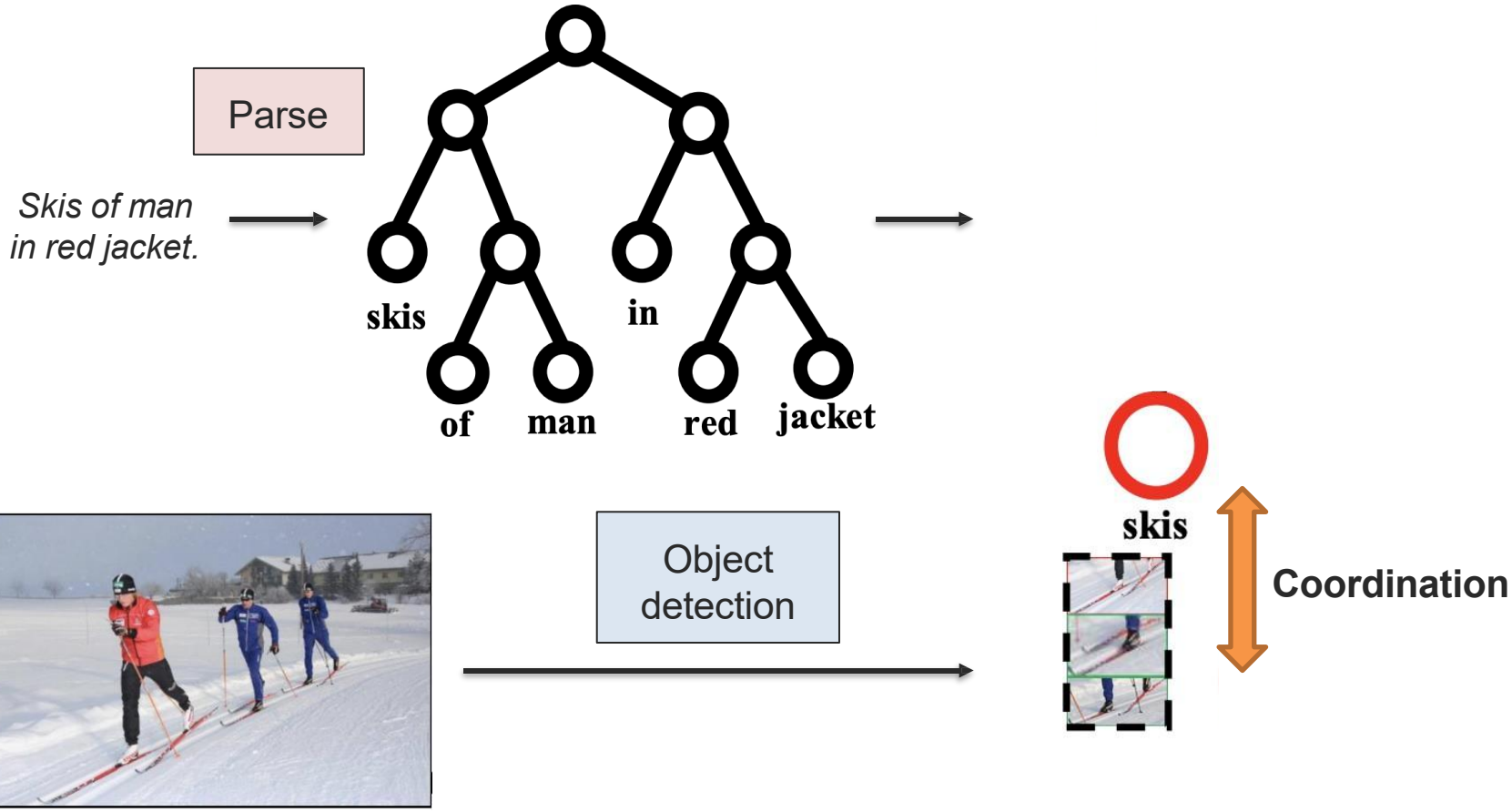


Object detection

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

Hierarchical Structure

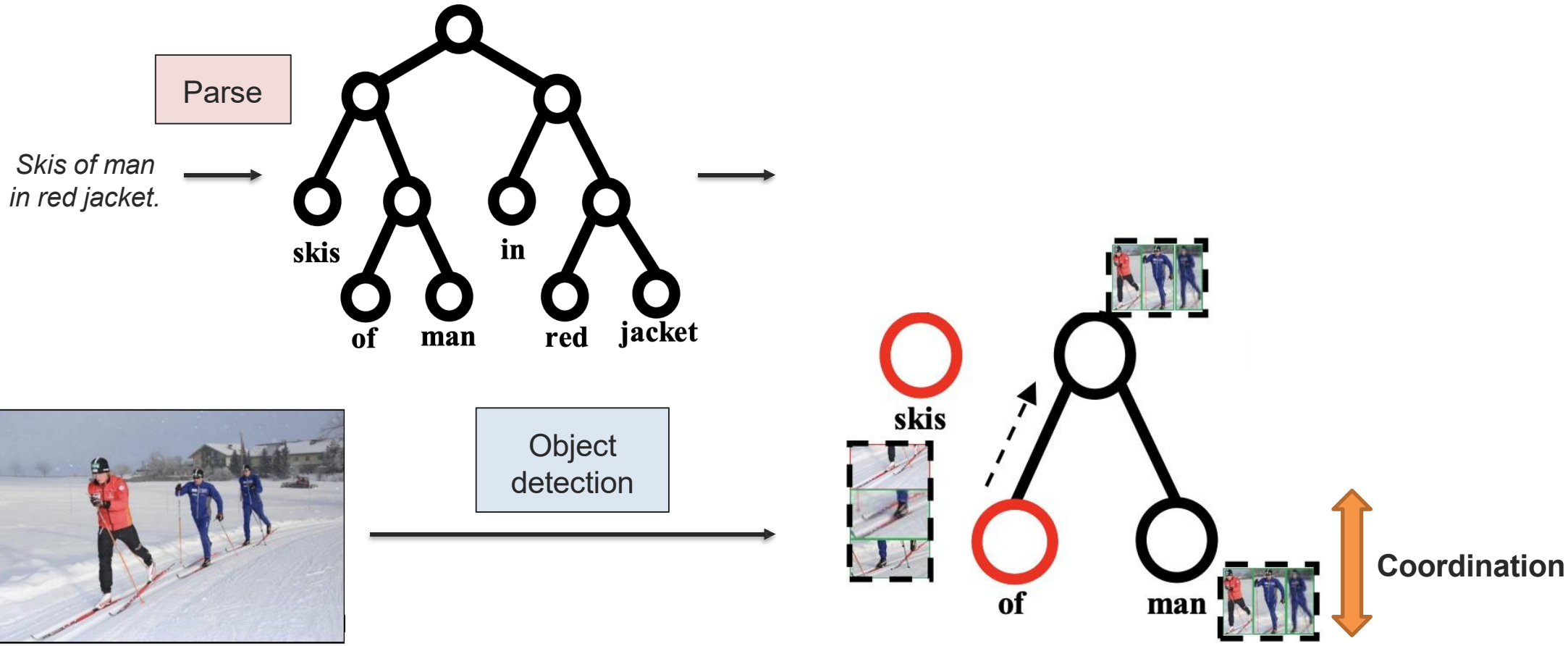
Leverage syntactic structure of language



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

Hierarchical Structure

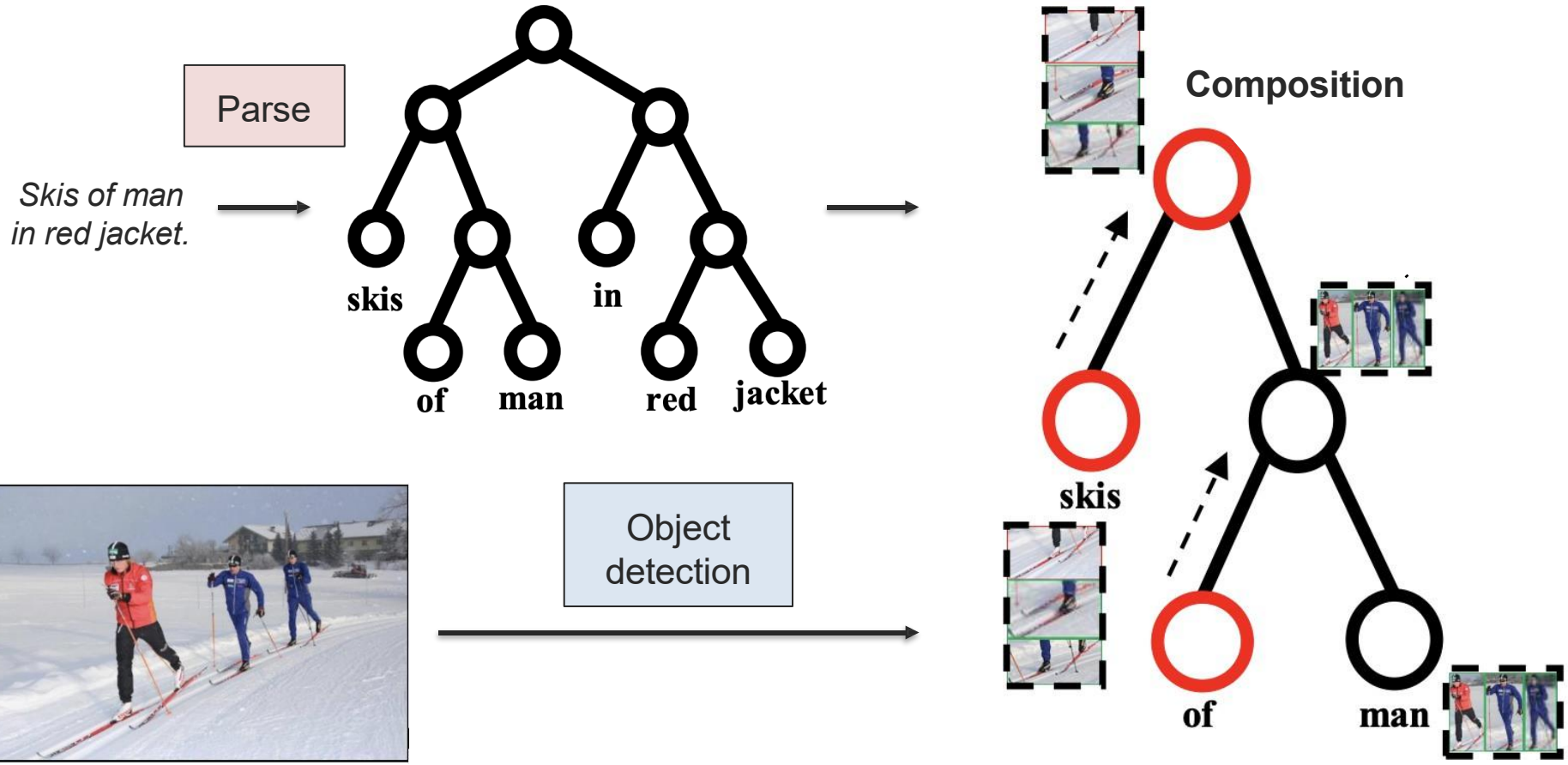
Leverage syntactic structure of language



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

Hierarchical Structure

Leverage syntactic structure of language



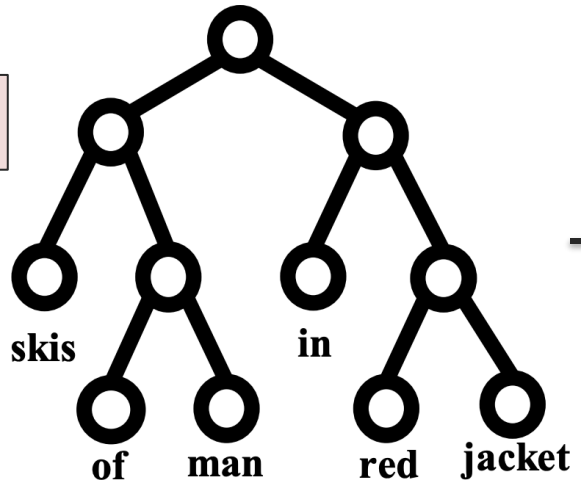
[Hong et al., Learning to Compose and Reason with Language Tree Structures for Visual Grounding. IEEE TPAMI 2019]

Hierarchical Structure

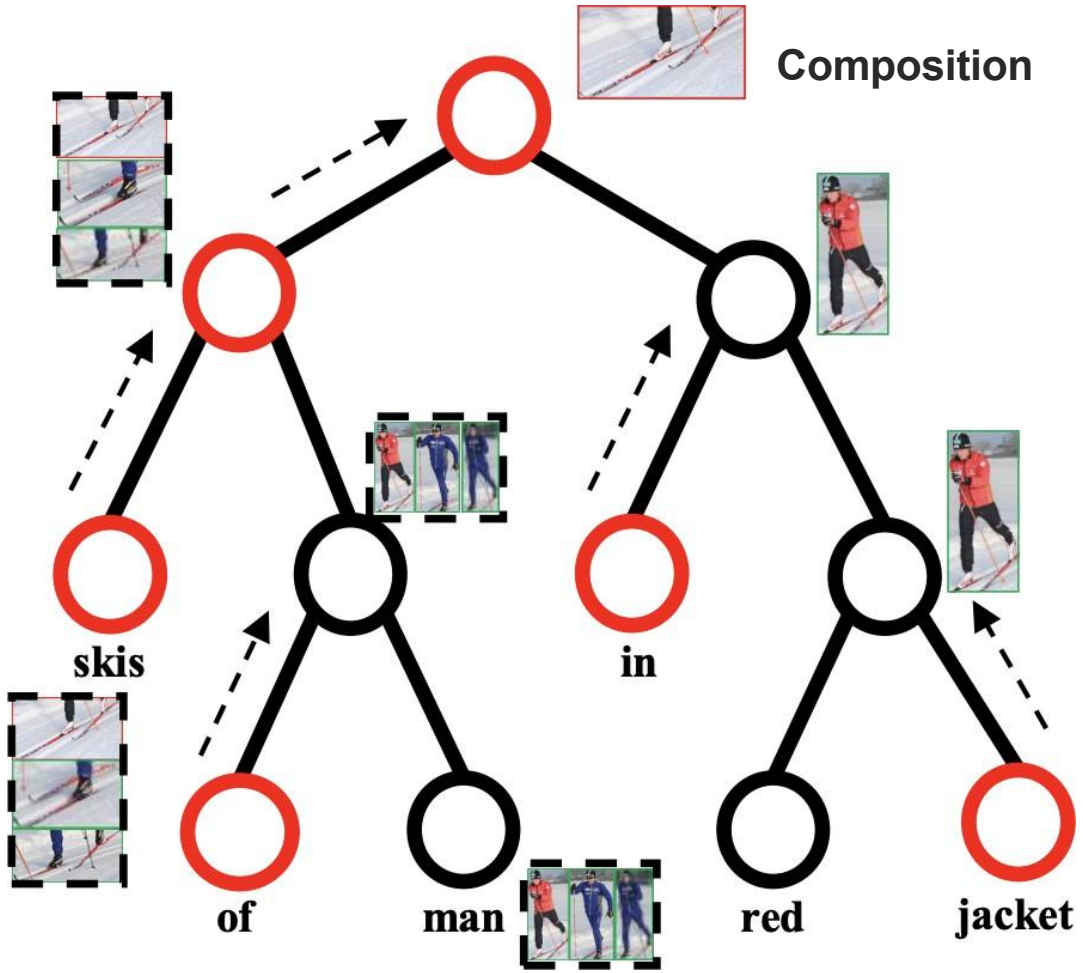
Leverage syntactic structure of language

Skis of man in red jacket.

Parse



Object detection



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

Hierarchical Structure

Leverage syntactic structure of language

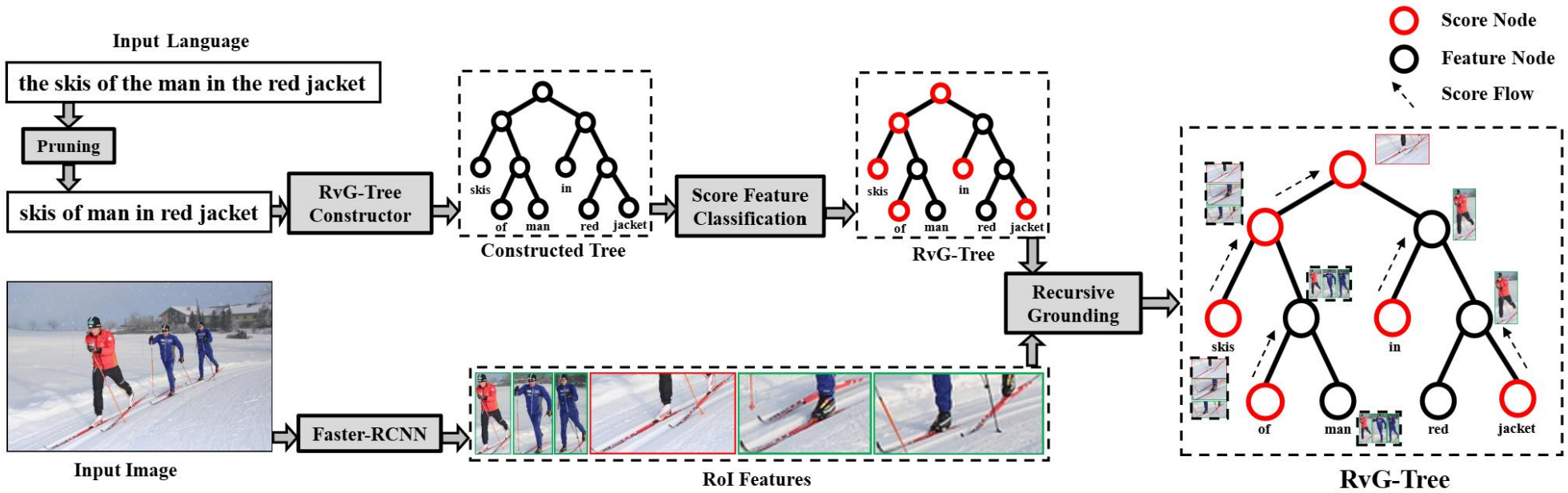


Fig. 2: The overview of using RvG-Tree for natural language grounding. Given a natural language sentence as input, we first prune the sentence and then construct RvG-Tree (Section 3.2). Then, we have a score-feature classifier (Section 3.3) to determine each node as the "score node" or "feature node", where the score node returns the recursive score and the feature node returns the feature (Section 3.3). The final score of the root node is accumulated recursively in a bottom-up fashion (Section 3.3) and the visual region with the highest score is considered as the grounding result. Note that all the nodes can be visualized by the corresponding confidence scores and only qualitative regions are visualized.

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

Module Network V2: End-to-End Learning

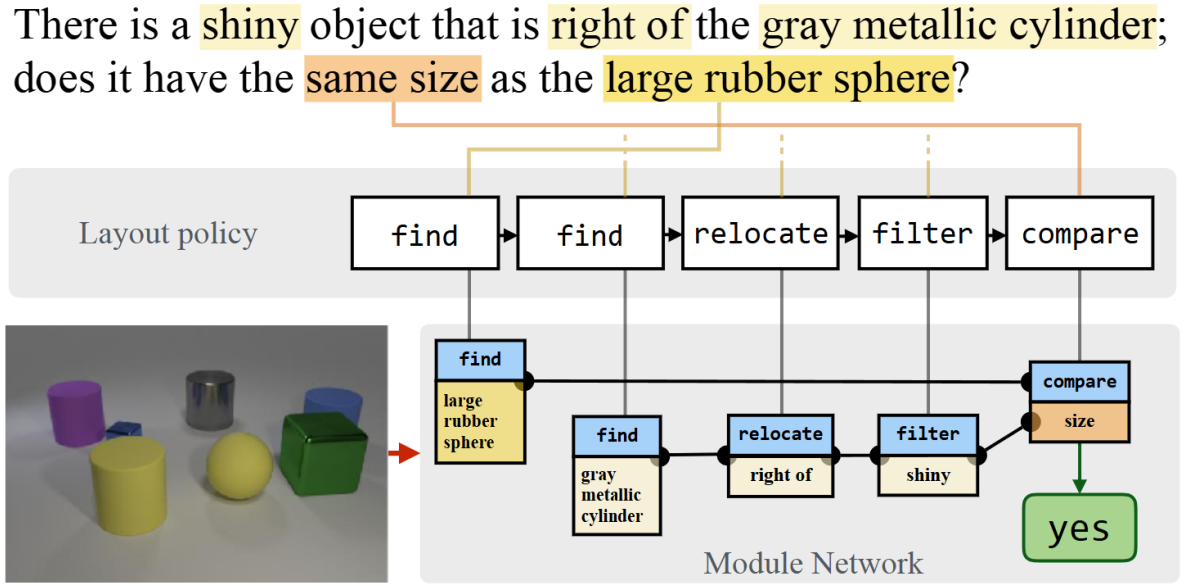
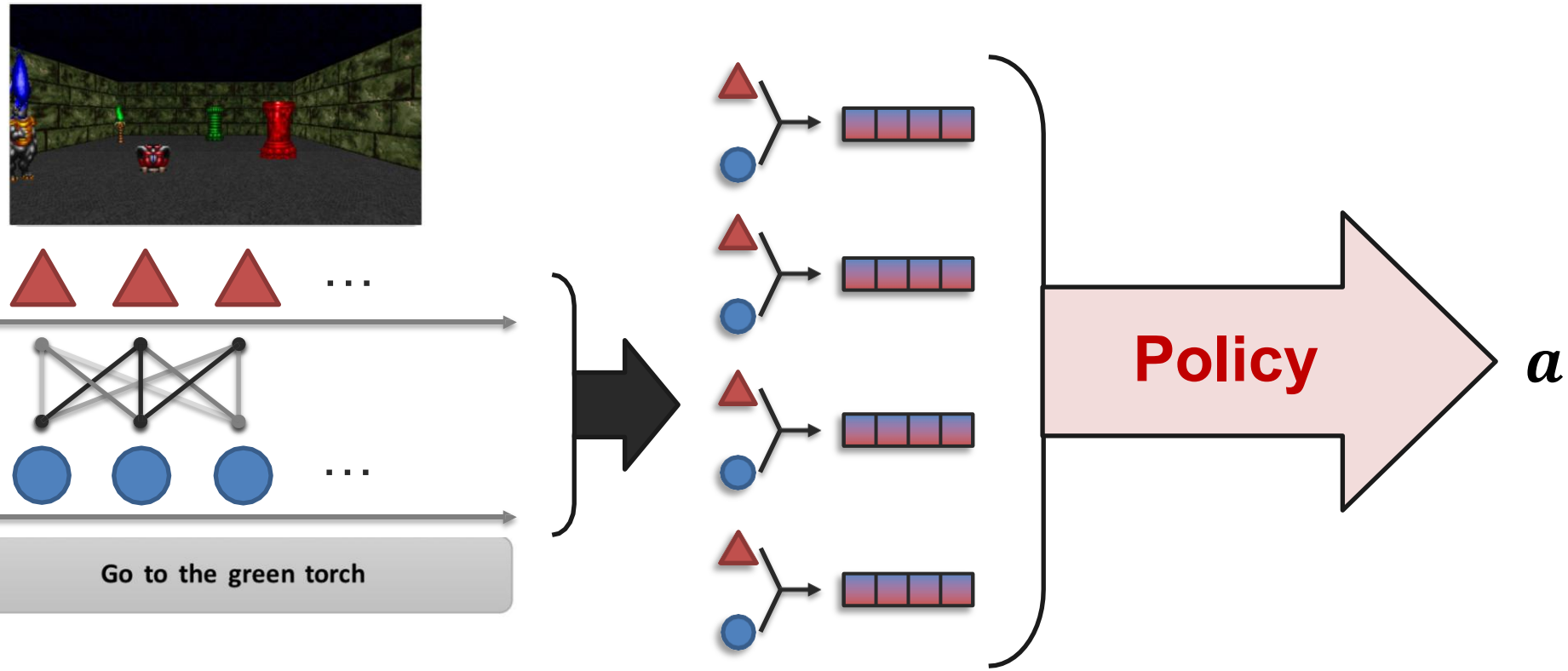


Figure 1: For each instance, our model predicts a computational expression and a sequence of attentive module parameterizations. It uses these to assemble a concrete network architecture, and then executes the assembled neural module network to output an answer for visual question answering. (The example shows a real structure predicted by our model, with text attention maps simplified for clarity.)

Interactive Structure

Structure defined through interactive environment

Main difference from temporal - actions taken at previous time steps affect future states



Challenge 3: Reasoning

Final Summary

Sub-Challenge	Definition	Concrete Example
Structure Modeling	Learning structured relationships	Scene graphs for VQA, Graph-based summarization
Intermediate Concepts	Extracting feature representations	Sentiment analysis, Lip-reading phonemes
Inference Paradigm	Logical or probabilistic reasoning over modalities	VQA (Vision-Text fusion), Commonsense AI
External Knowledge	Using external databases or knowledge graphs	Wikipedia-augmented Chatbots, COMET-based causal reasoning

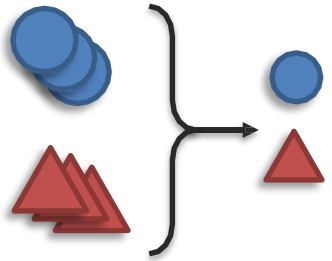
Challenge 4: Generation

Definition: **Learning a generative process** to produce new data while maintaining **cross-modal interactions, structure, and coherence**.

Covers tasks such as **summarization, translation, and creative generation** across multiple modalities.

Sub-challenges:

Summarization

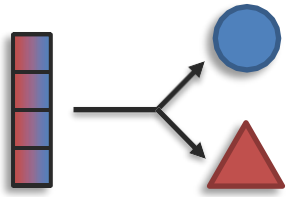


Reduction

Translation

Maintenance

Creation



Expansion

Information:
(content)

Sub-challenge 4a: 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 4a: Summarization

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

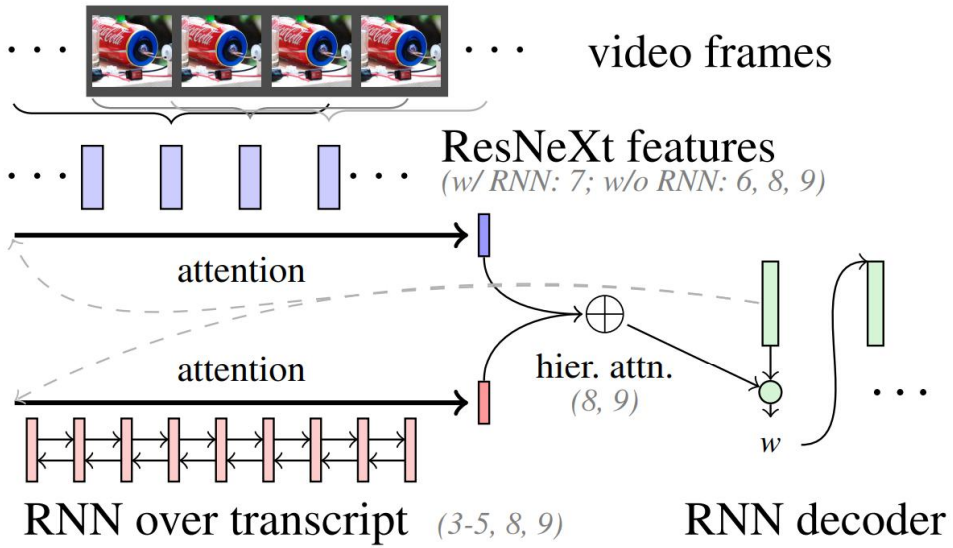


Figure 2: Building blocks of the sequence-to-sequence models, gray numbers in brackets indicate which components are utilized in which experiments.

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

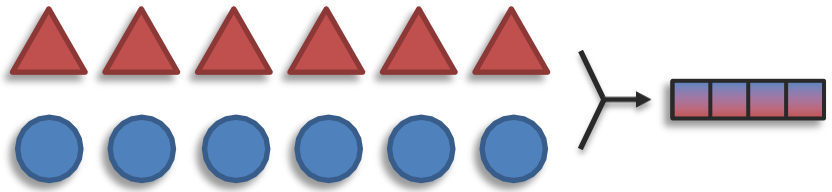
Sub-challenge 4a: Summarization

Video summarization

A Content

Fusion via
joint representation

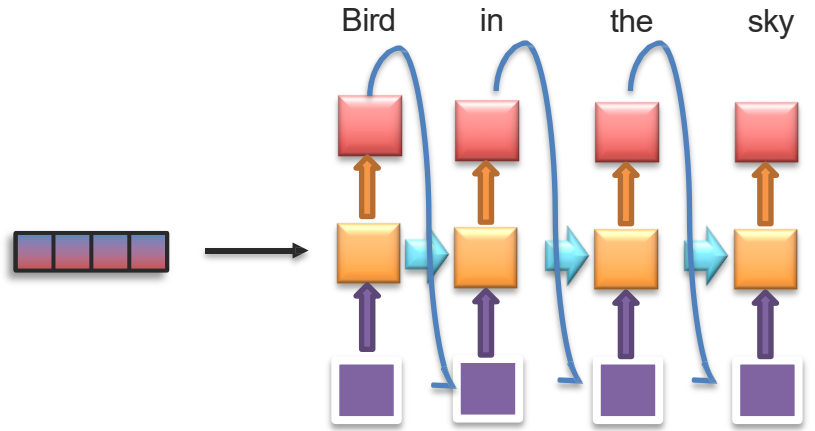
Capture **complementary**
cross-modal interactions



B Generation

Generative ≈ abstractive summarization

Exemplar ≈ extractive summarization



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

Sub-challenge 4b: Translation

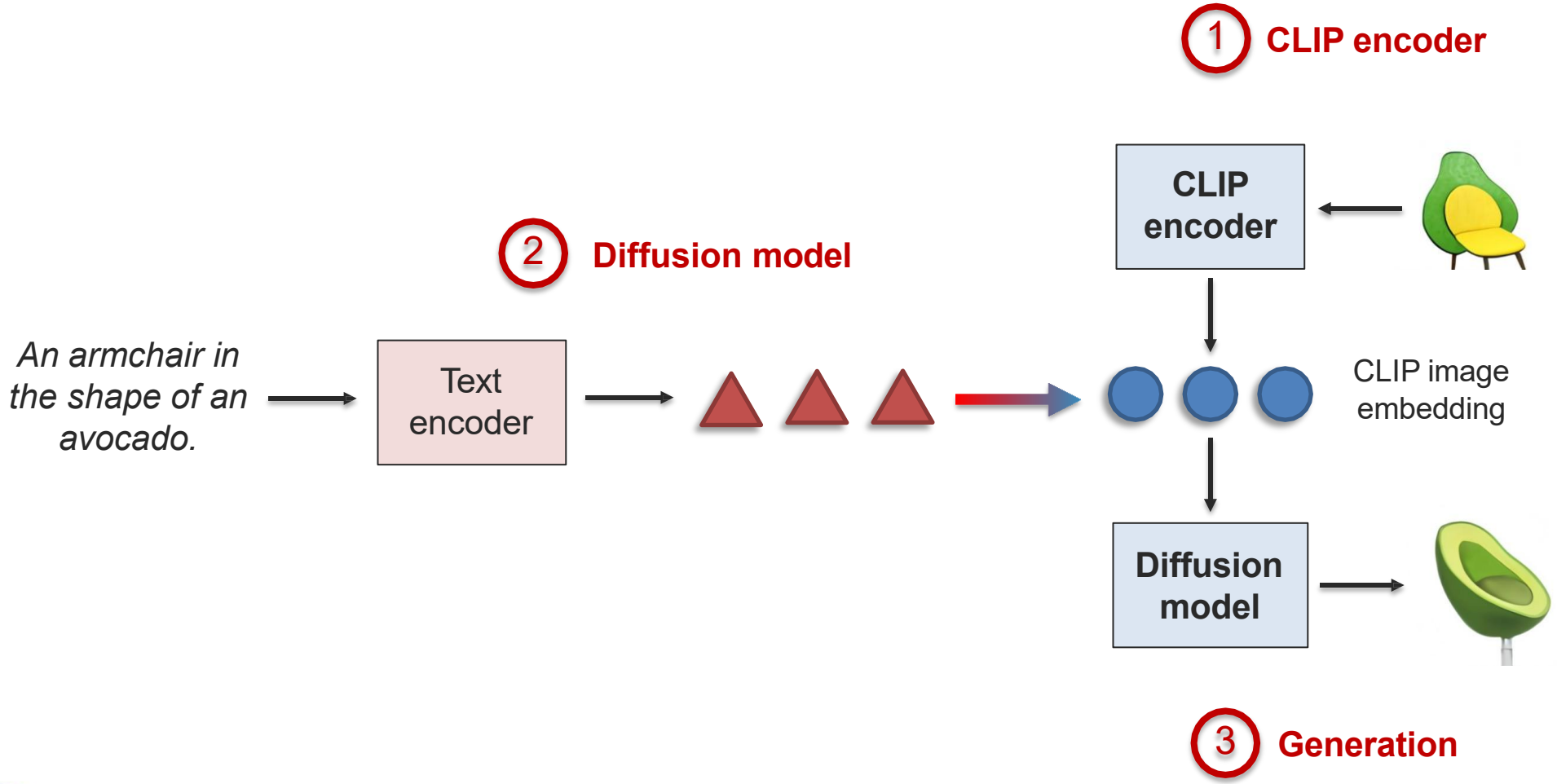
Definition: Translating from one modality to another and keeping information content while being consistent with cross-modal interactions.

An armchair in the shape of an avocado →

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

Sub-challenge 4b: Translation

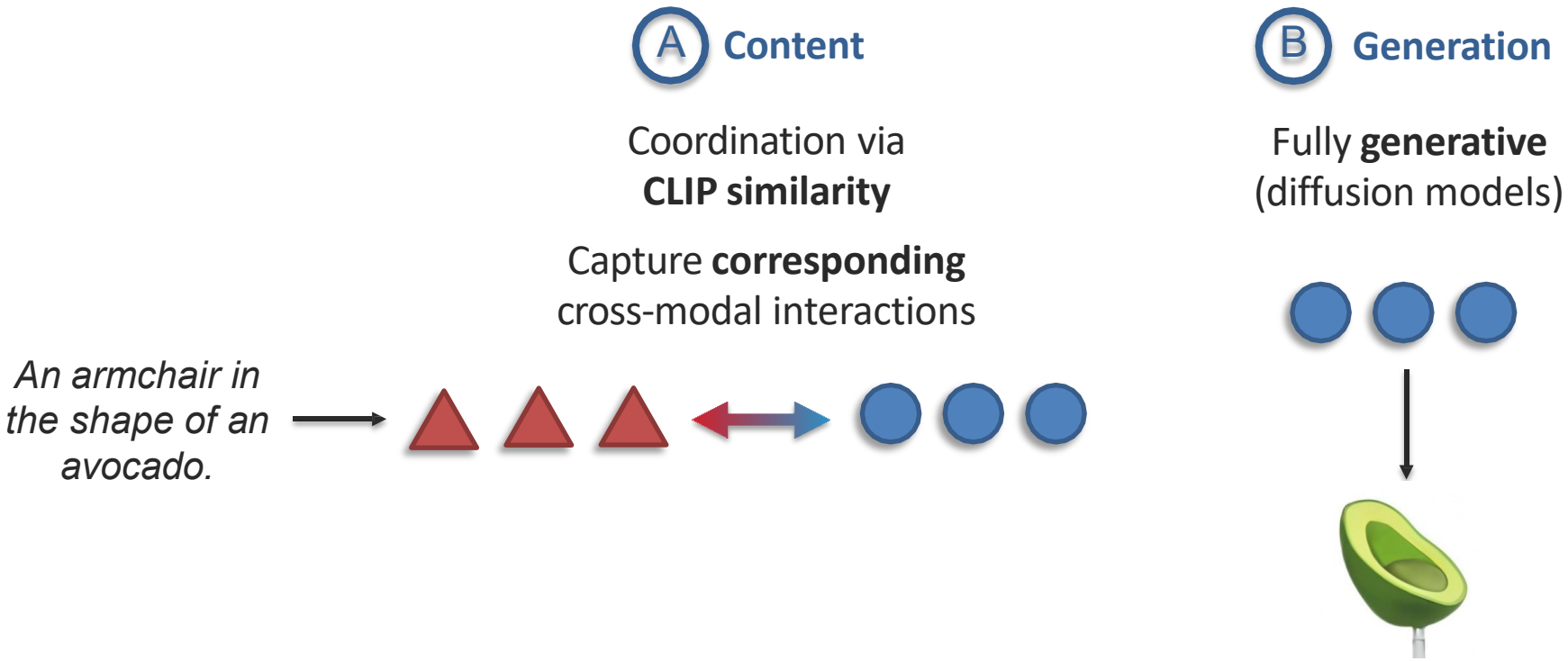
DALL·E 2: Combining with CLIP, diffusion models



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

Sub-challenge 4b: Translation

DALL·E 2: Combining with CLIP, diffusion models

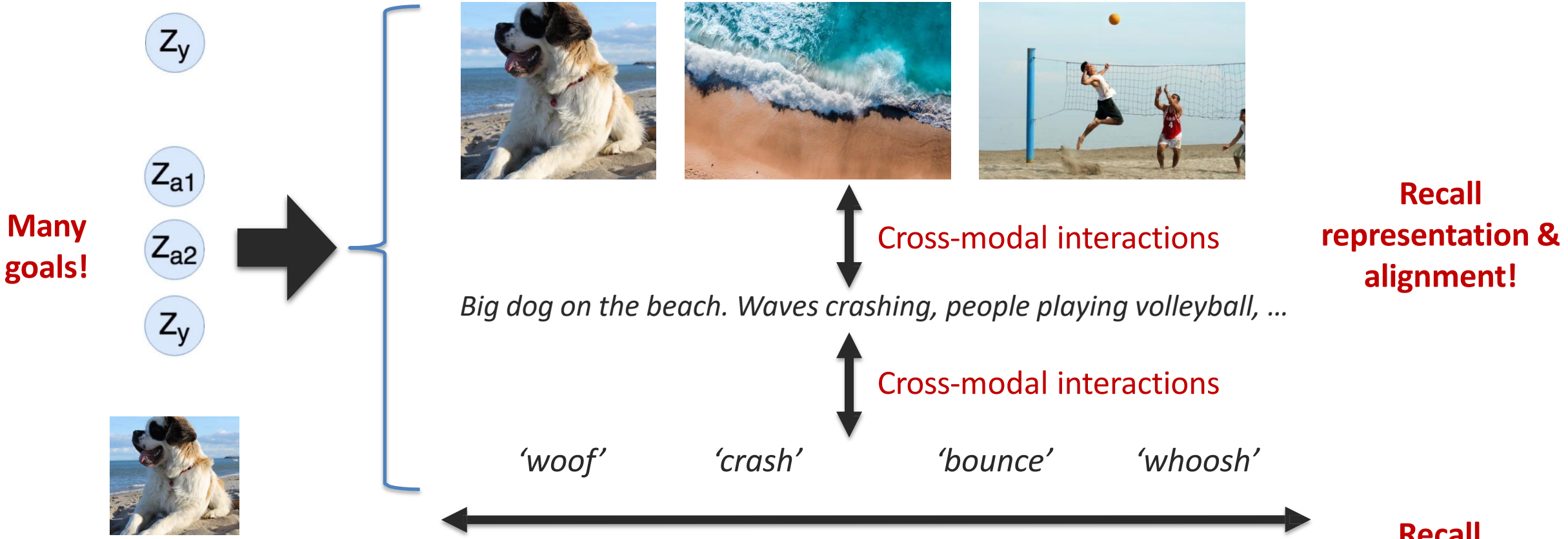


[Ramesh et al., Hierarchical Text-Conditional Image Generation with CLIP Latents. arXiv 2022]

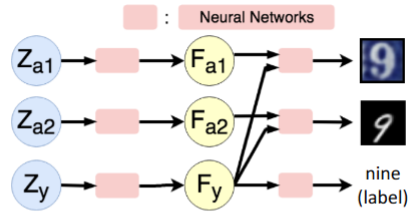
Sub-challenge 4c: Creation

Open challenges

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



Sub-challenge 4c: Creation

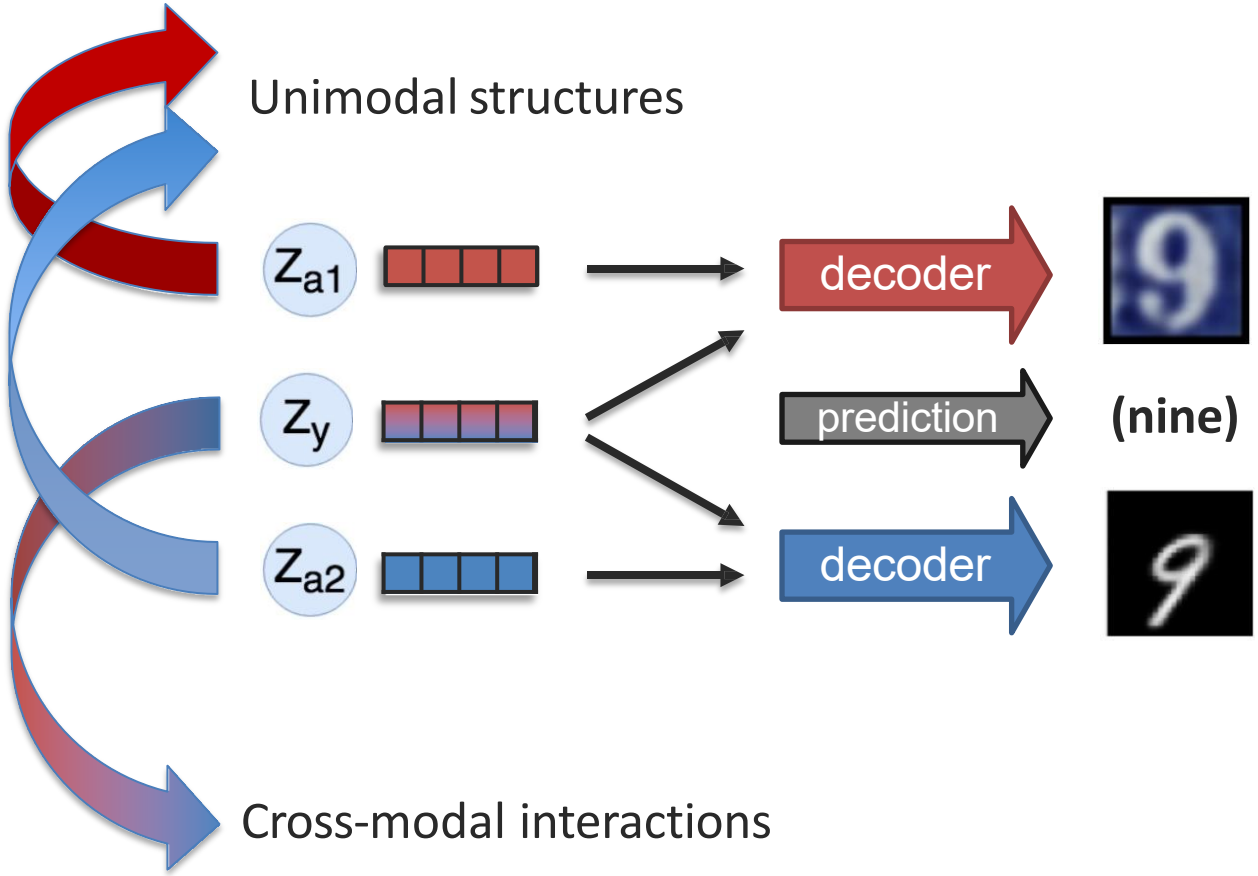


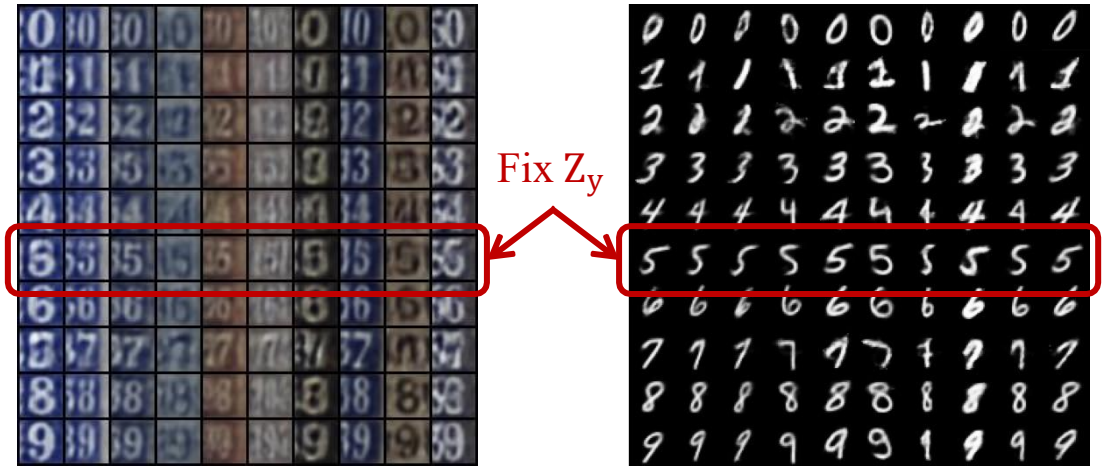
(a)

Method	UM(SVHN)	UM(MNIST)	MM	MFM
Acc.	91.84	99.01	99.20	**99.36**

(b)

Some initial attempts: factorized generation



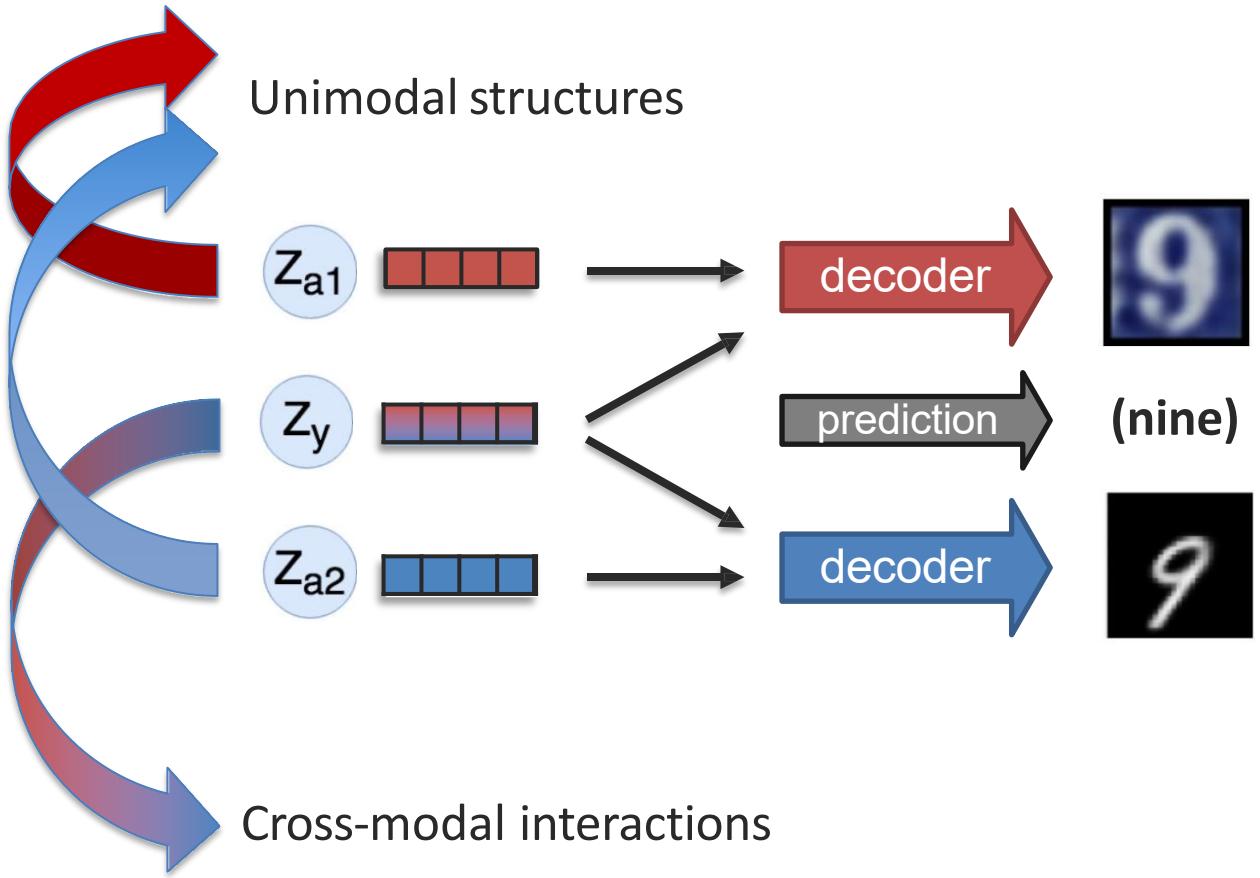


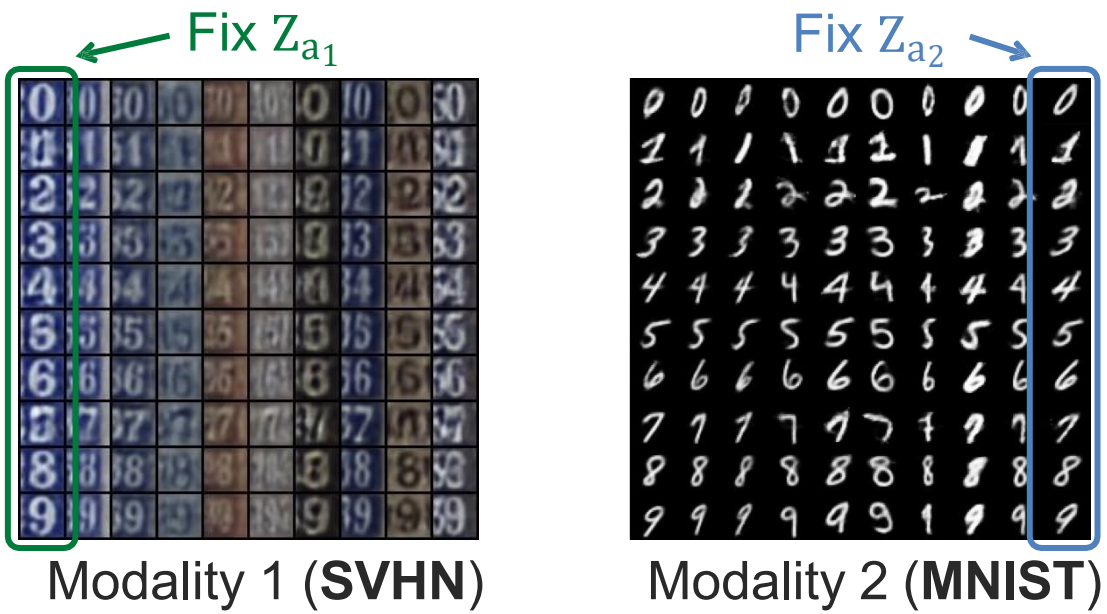
Modality 1 (**SVHN**) Modality 2 (**MNIST**)

[Tsai et al., Learning Factorized Multimodal Representations, ICLR 2019]

Sub-challenge 4c: Creation

Some initial attempts: factorized generation

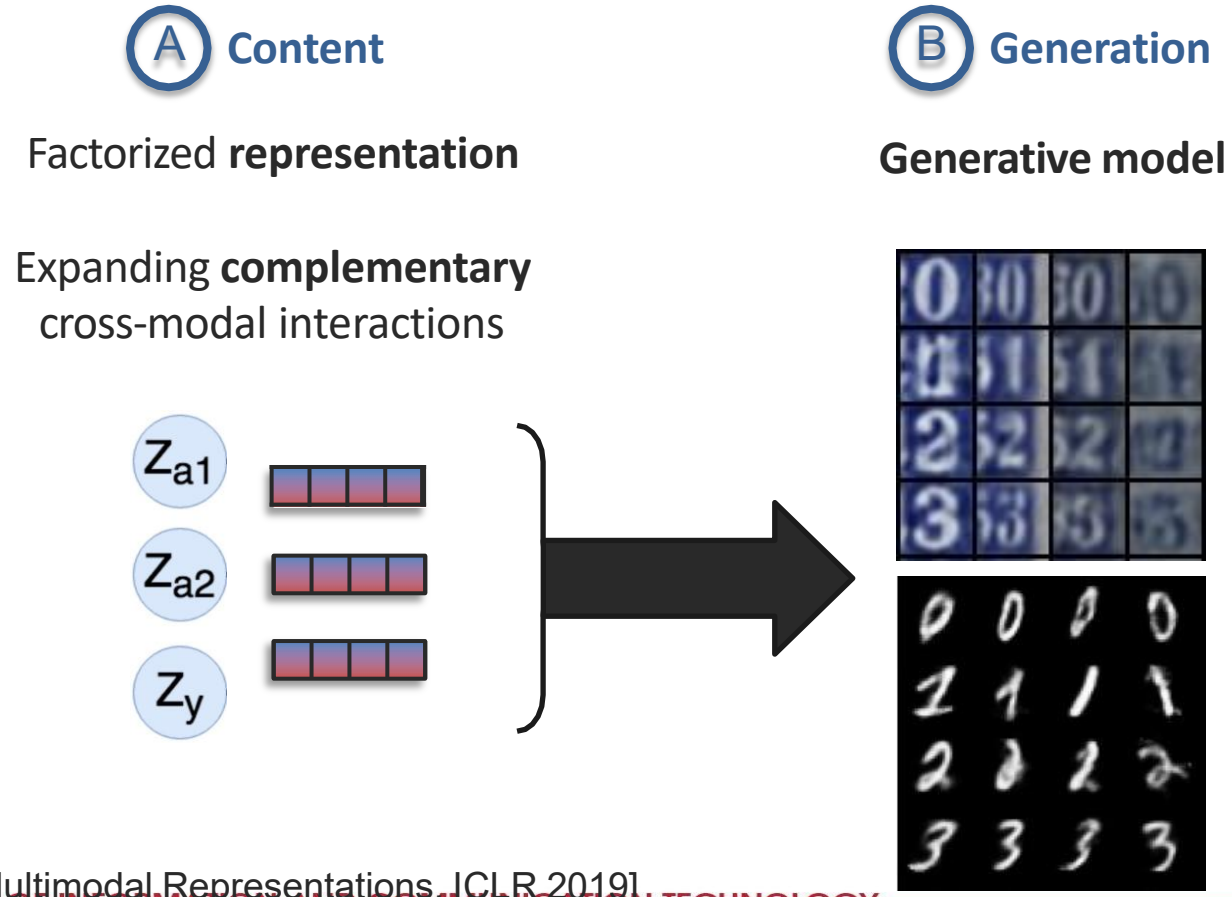




[Tsai et al., Learning Factorized Multimodal Representations, ICLR 2019]

Sub-challenge 4c: Creation

Some initial attempts: factorized generation



[Tsai et al., Learning Factorized Multimodal Representations, ICLR 2019]

Challenge 4: Generation

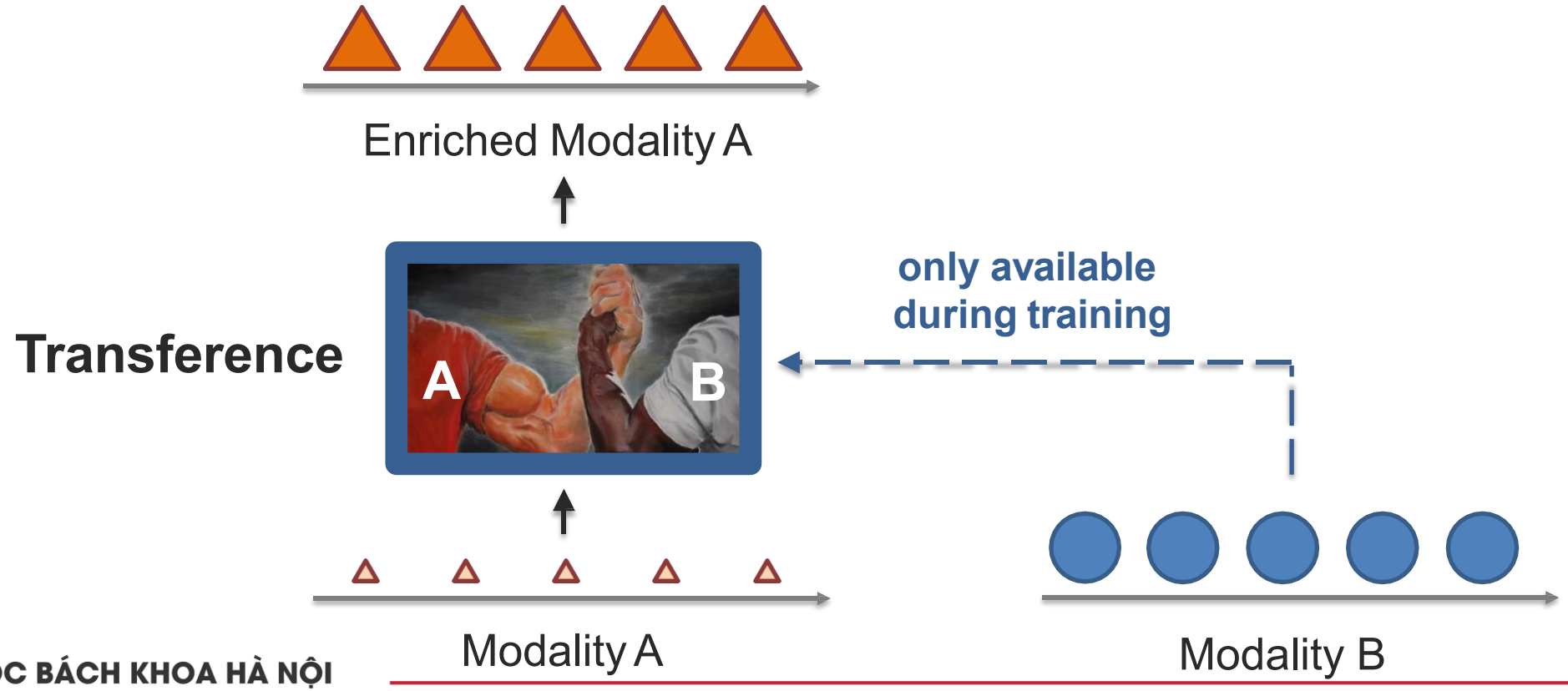
Final Summary

Sub-Challenge	Definition	Concrete Example
Summarization (Reduction)	Extract key information while removing redundancy	Video summarization, News article condensation
Translation (Preserving Meaning)	Convert data between modalities while retaining meaning	Speech-to-text, Image captioning
Creation (Expansion)	Generate new content beyond the given data	Text-to-image (DALL-E), Story generation

Challenge 5: Transference

Definition: Transferring knowledge between different modalities to help a **weaker modality** that might be **noisy, incomplete, or resource-limited**.

Used in **low-resource learning scenarios**, **data-efficient AI**, and **multimodal adaptation**.



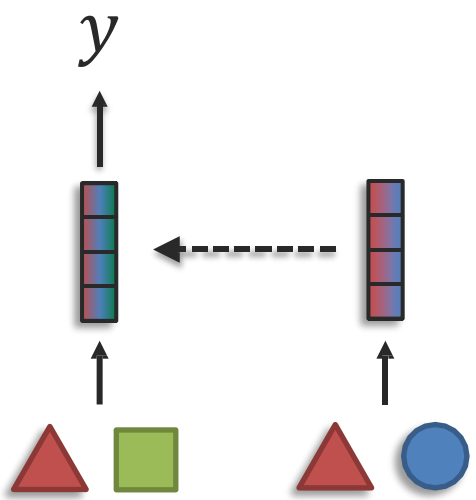
Challenge 5: Transference

Definition: **Transferring knowledge between different modalities** to help a **weaker modality** that might be **noisy, incomplete, or resource-limited**.

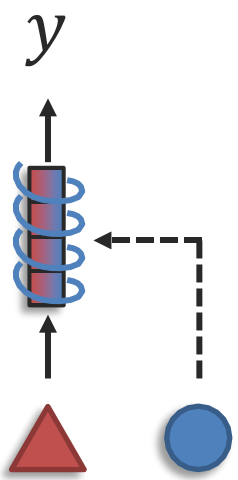
Used in **low-resource learning scenarios**, **data-efficient AI**, and **multimodal adaptation**.

Sub-challenges:

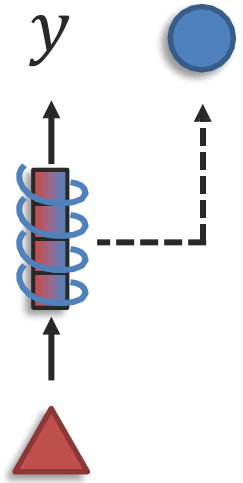
Transfer



Co-learning
via representation

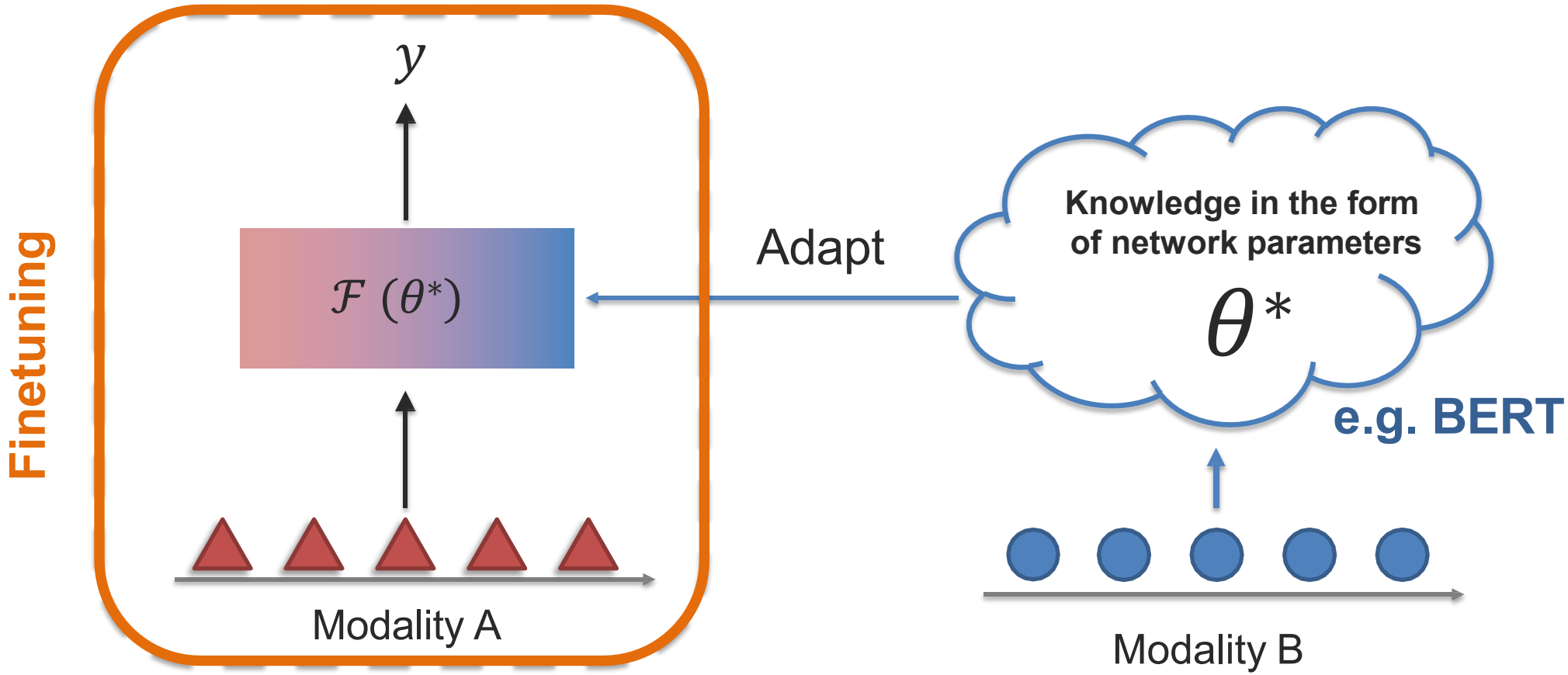


Co-learning
via generation



Sub-Challenge 5a: Transfer via Pretrained Models

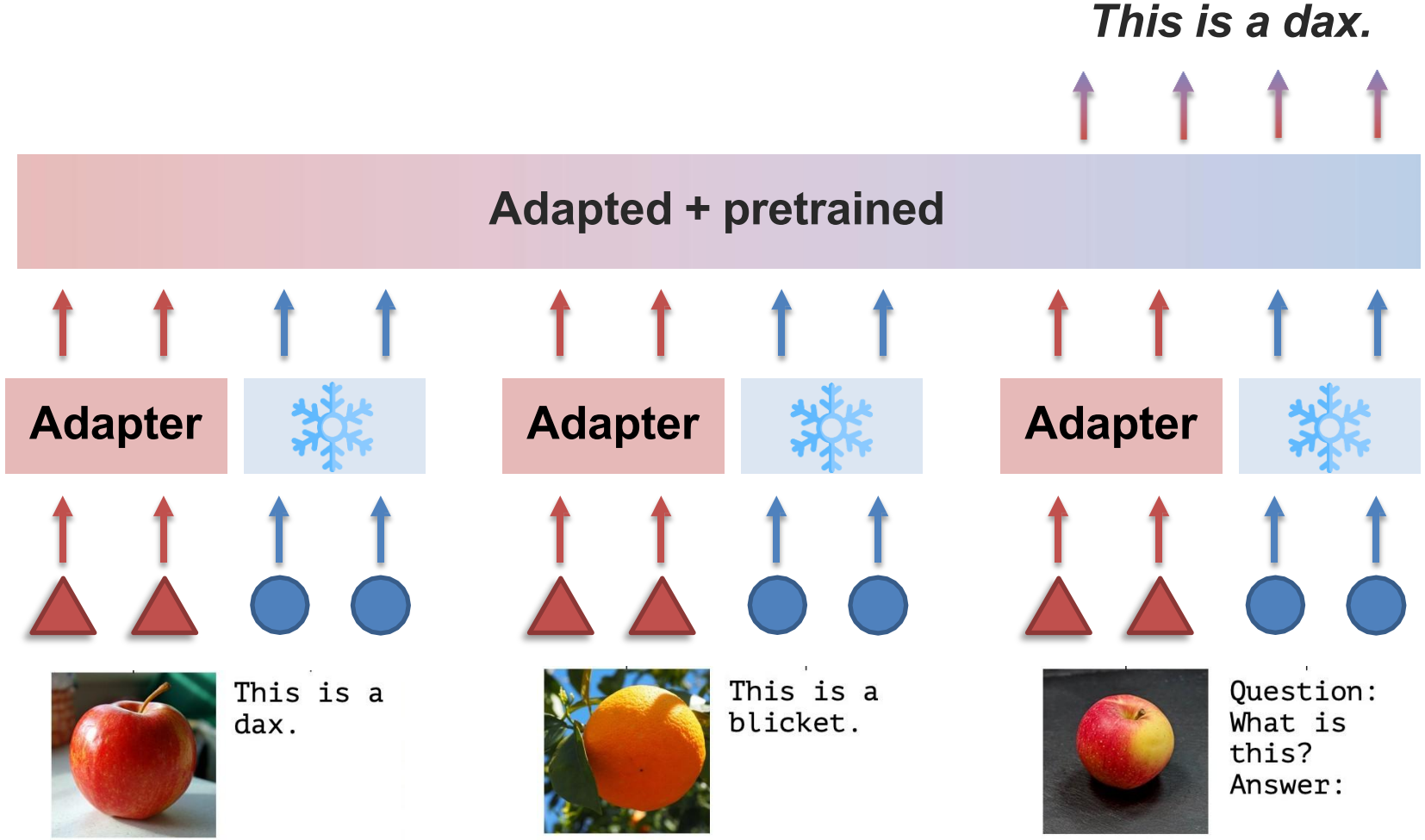
Definition: Transferring knowledge from large-scale pretrained models to downstream tasks involving the primary modality.



Sub-Challenge 5a: Transfer via Pretrained Models

Transfer via prefix tuning

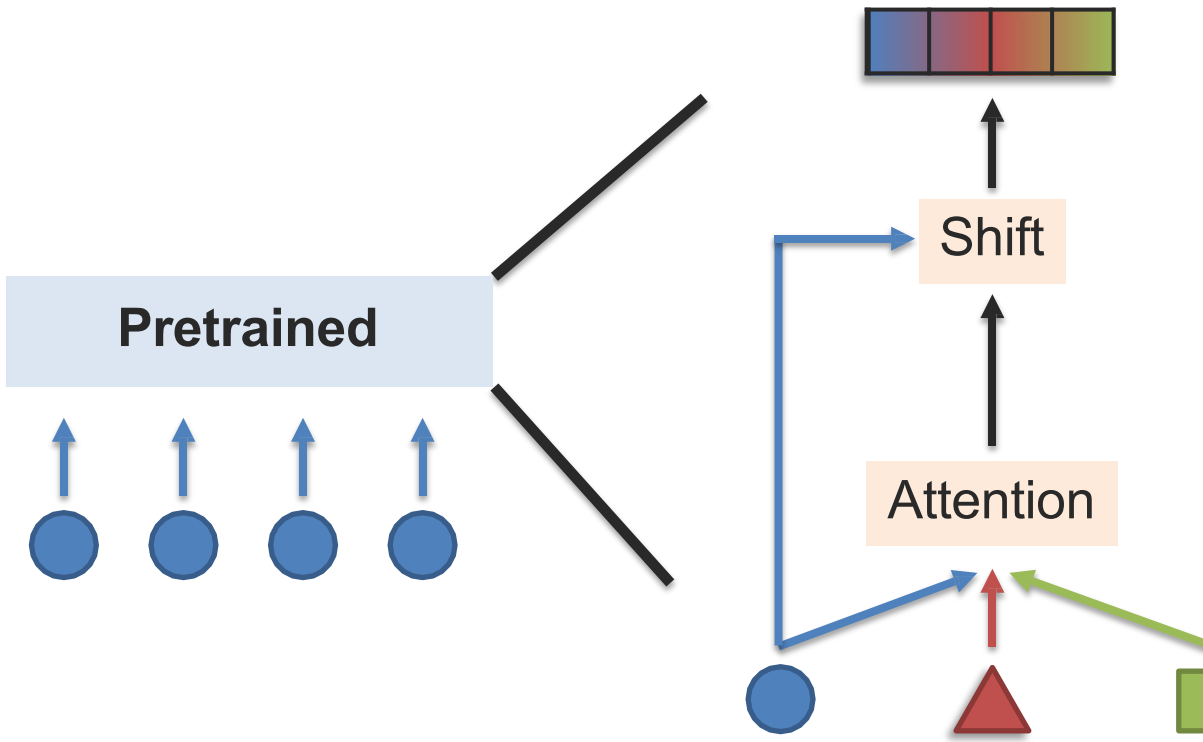
Few-shot image classification:

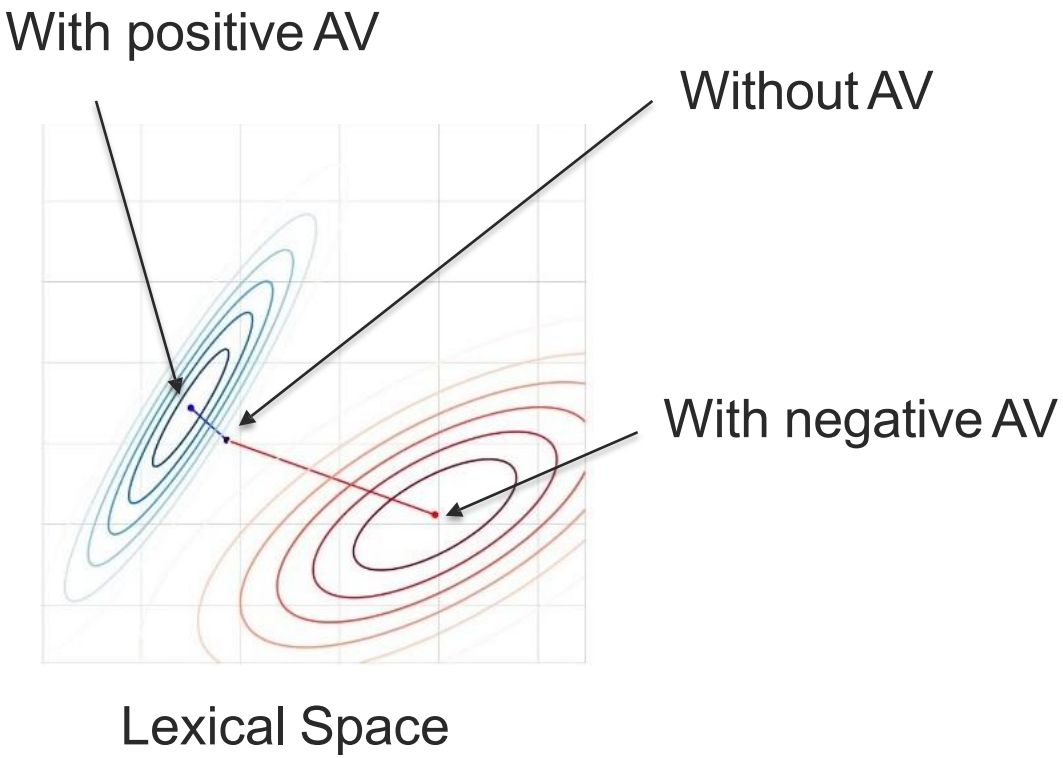


[Tsimpoukelli et al., Multimodal Few-Shot Learning with Frozen Language Models. NeurIPS 2021]

Sub-Challenge 5a: Transfer via Pretrained Models

Transfer via representation tuning

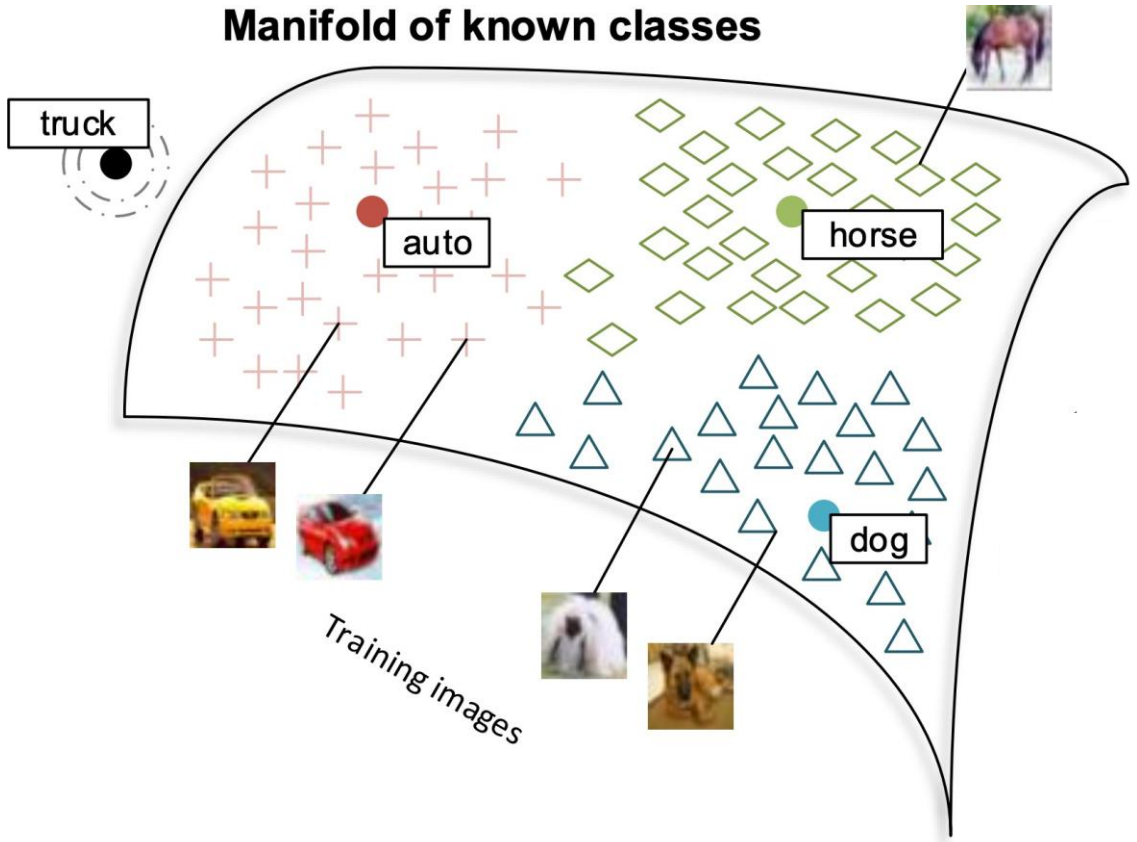




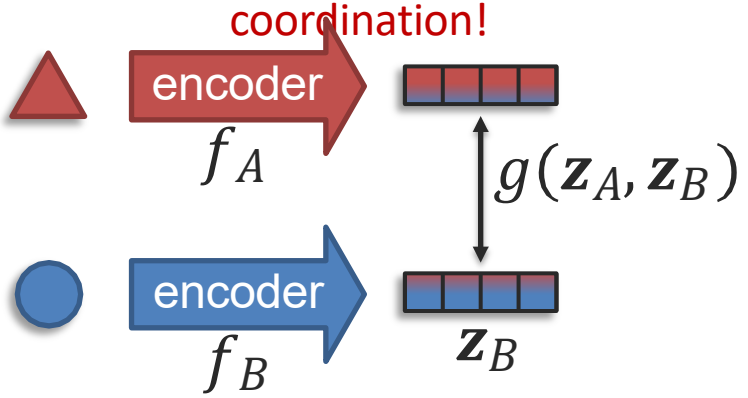
[Ziegler et al., Encoder-Agnostic Adaptation for Conditional Language Generation. arXiv 2019] [Rahman et al., Integrating Multimodal Information in Large Pretrained Transformers. ACL 2020]

Co-learning via Representation

- **Representation coordination: word embedding space for zero-shot visual classification**



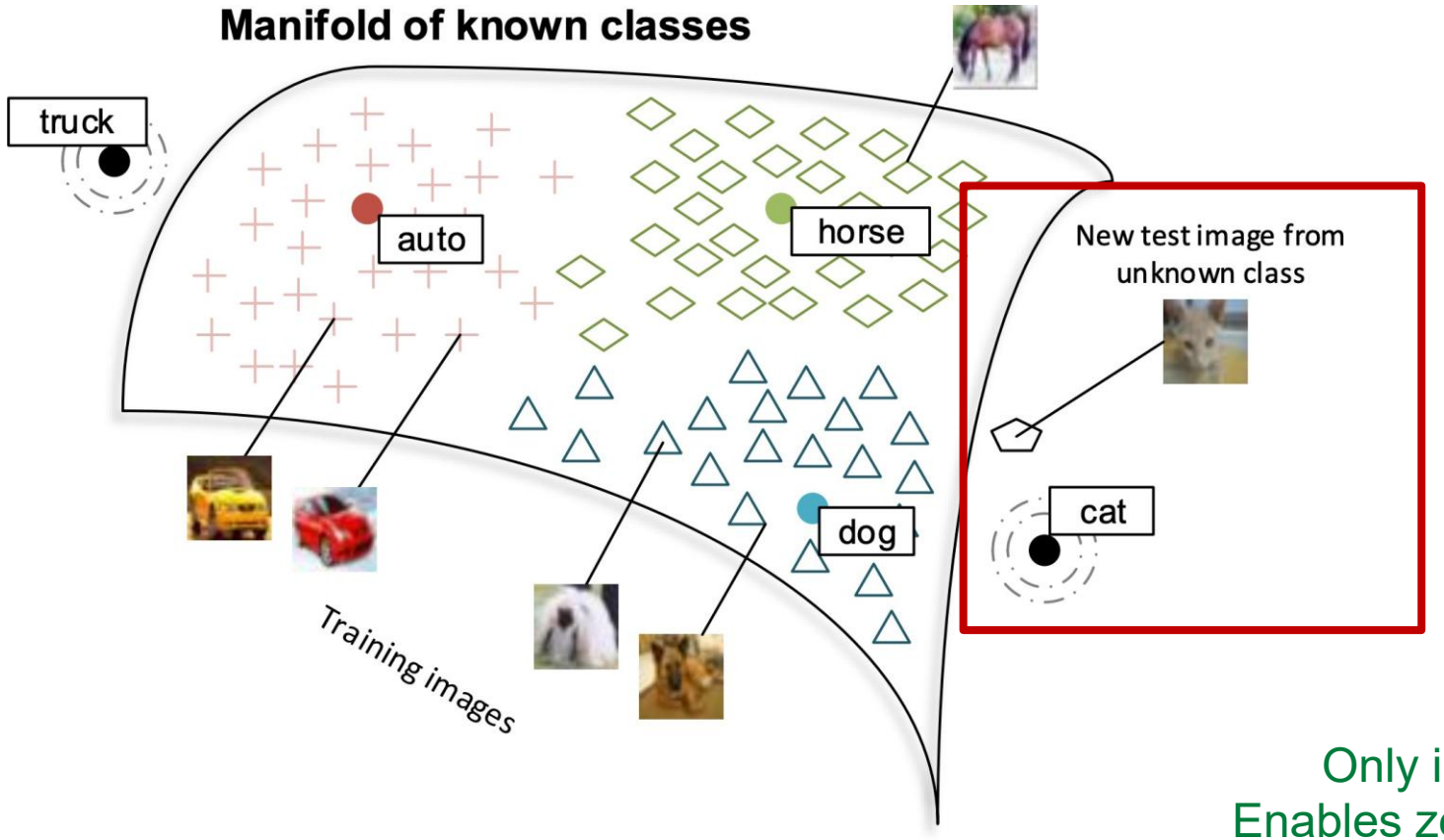
- Recall representation coordination!



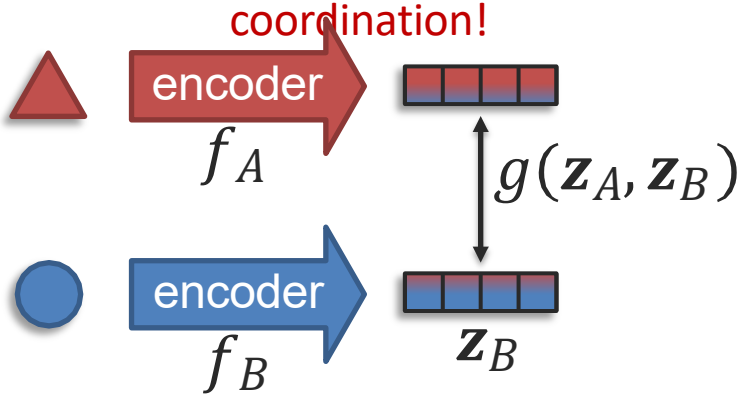
[Socher et al., Zero-Shot Learning Through Cross-Modal Transfer. NeurIPS 2013]

Co-learning via Representation

- **Representation coordination: word embedding space for zero-shot visual classification**



- Recall representation coordination!

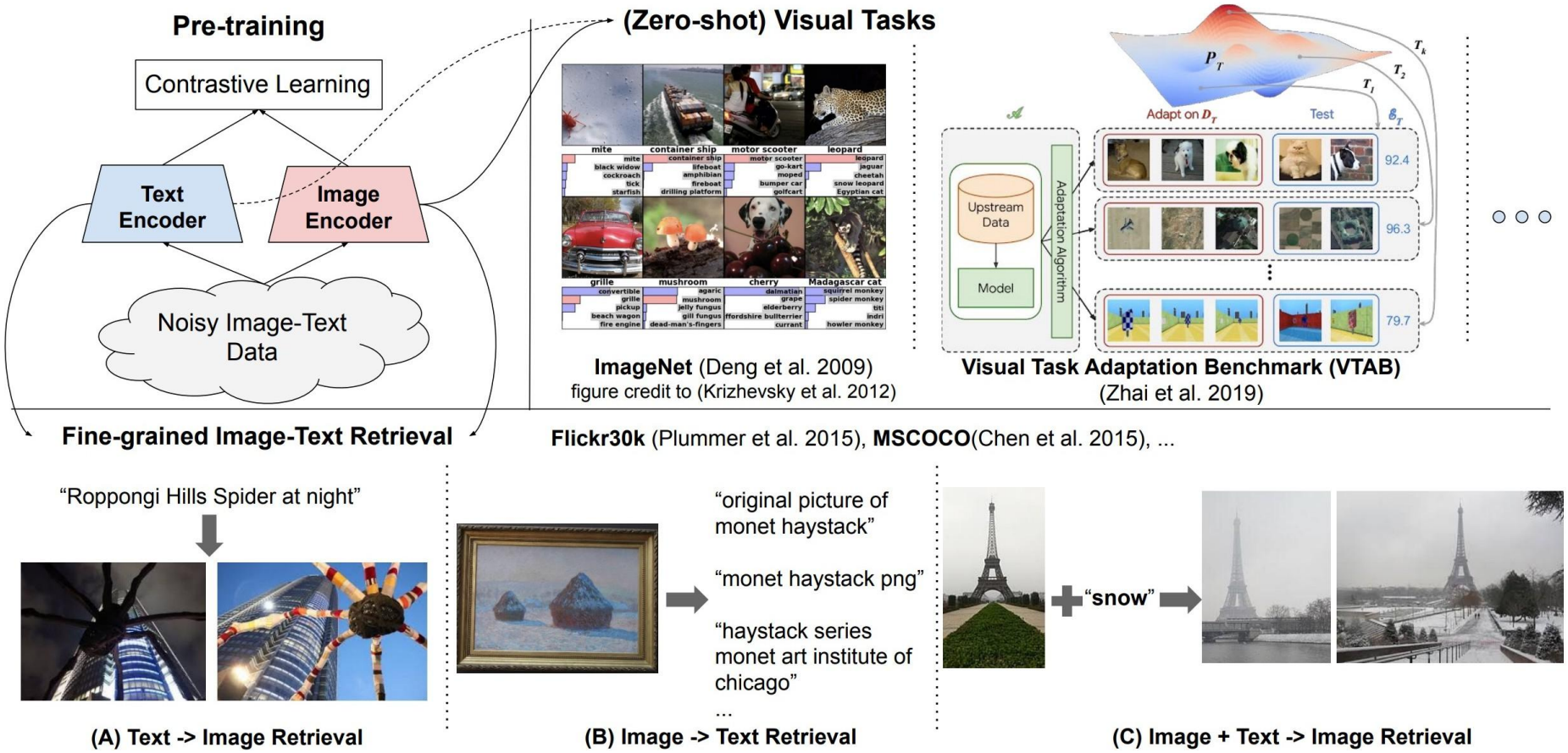


Only images used at test-time
Enables zero-shot image classification

[Socher et al., Zero-Shot Learning Through Cross-Modal Transfer. NeurIPS 2013]

Co-learning via Representation

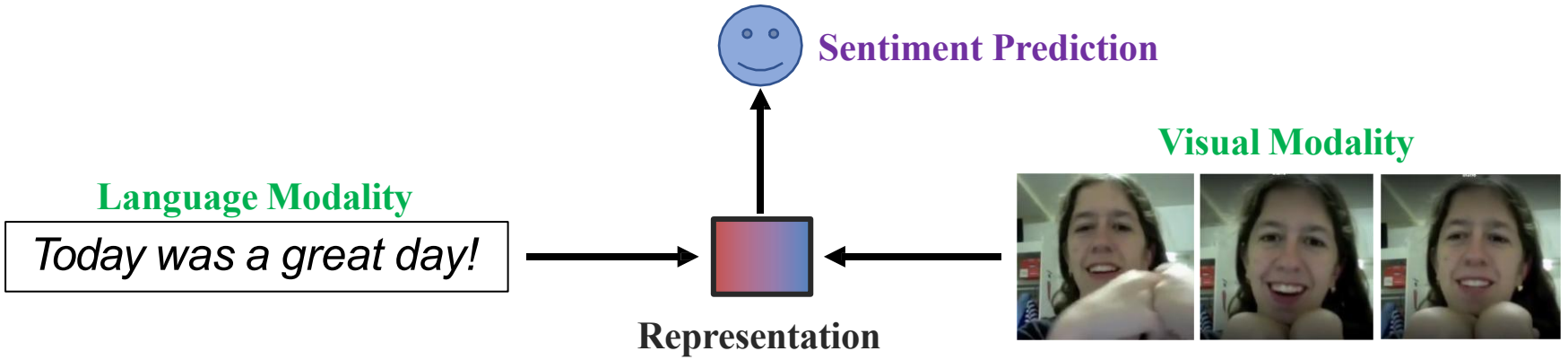
Representation coordination at scale



[Jia et al., Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision. ICML 2021]

Co-learning via Generation

Bimodal translations



Both modalities required at test time!
Sensitive to noisy/missing visual modality.

We want to leverage information from visual modality
while being robust to it during test-time.

[Pham et al., Found in Translation: Learning Robust Joint Representations via Cyclic Translations Between Modalities. AAAI 2019]

Challenge 5: Transference

Final Summary

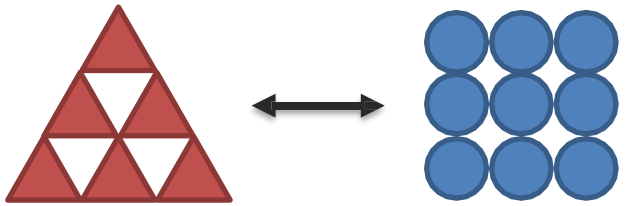
Sub-Challenge	Definition	Concrete Example
Transfer	One modality transfers knowledge to another	Text-trained models help low-resource speech recognition
Co-Learning via Representation	Learning shared representations for robustness	Lip-reading assists noisy speech recognition
Co-Learning via Generation	One modality generates missing data for another	Text-to-speech, synthetic image-text augmentation

Challenge 6: Quantification

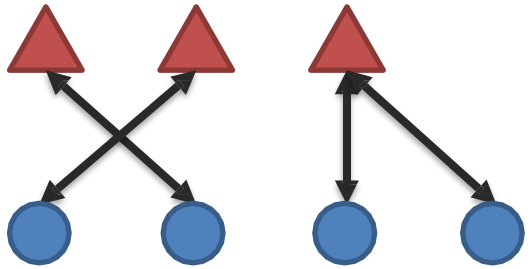
Definition: **Empirical and theoretical study** to improve understanding of **heterogeneity, cross-modal interactions, and learning dynamics** in multimodal AI. Involves **metrics, evaluation techniques, and optimization strategies** to enhance multimodal learning.

Sub-challenges:

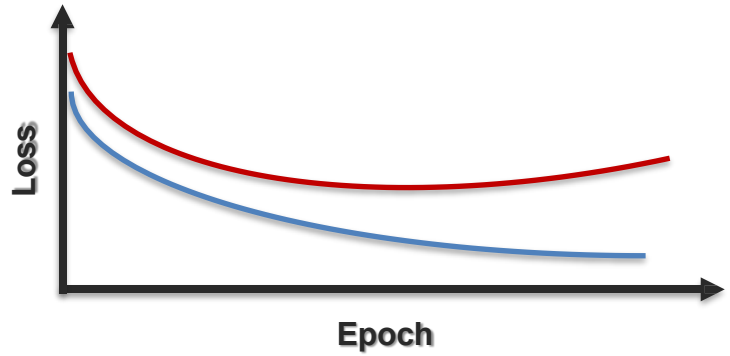
Heterogeneity



Interactions

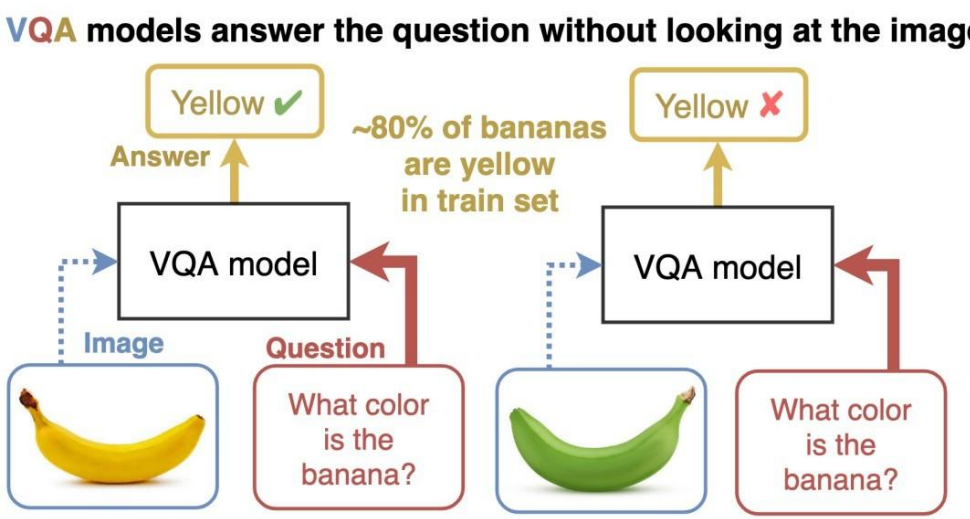


Learning



Modality Biases

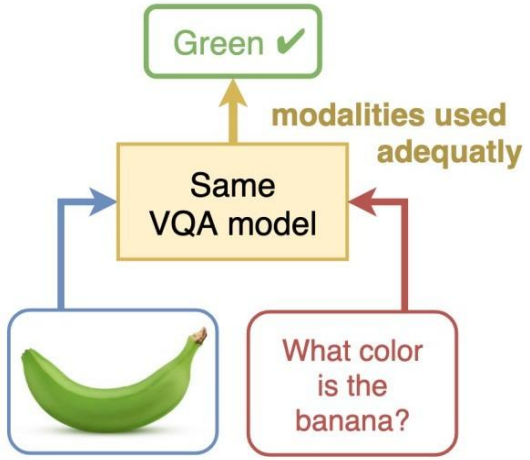
Heterogeneity in information and relevance
Unimodal biases and modality collapse



Balancing modalities

Balancing training

Not the case when trained with RUBi



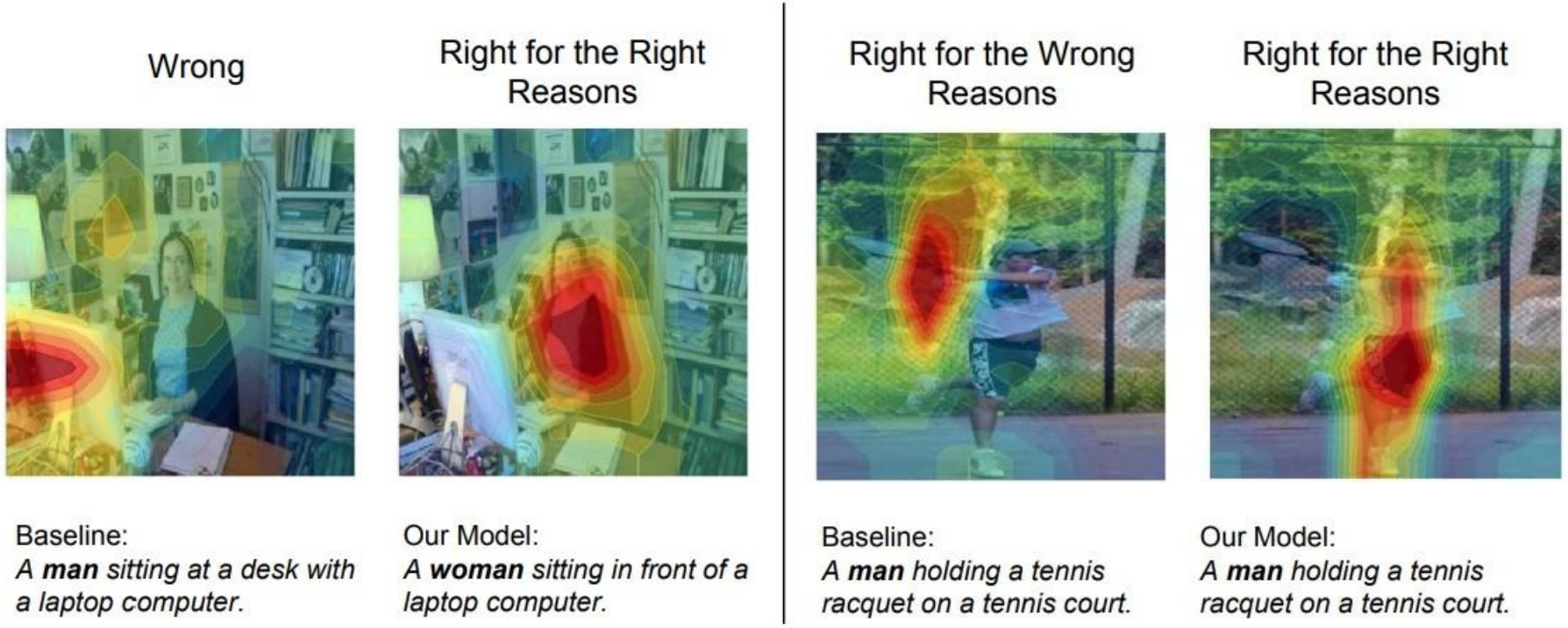
[Wu et al., Characterizing and Overcoming the Greedy Nature of Learning in Multi-modal Deep Neural Networks. ICML 2022]
[Javaloy et al., Mitigating Modality Collapse in Multimodal VAEs via Impartial Optimization. ICML 2022]

[Goyal et al., Making the V in VQA Matter: Elevating the Role of Image Understanding in Visual Question Answering. CVPR 2017]

Modality Biases

Heterogeneity in information and relevance
Fairness and social biases – unimodal social biases

Finding: Image captioning models capture spurious correlations between gender and generated actions



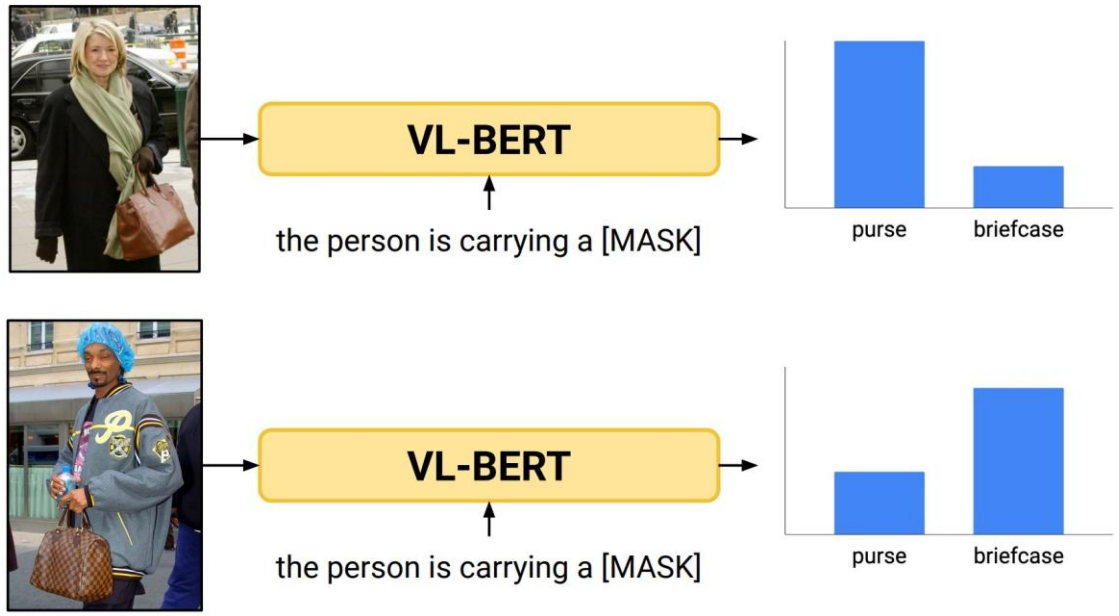
[Hendricks et al., Women also Snowboard: Overcoming Bias in Captioning Models. ECCV 2018]

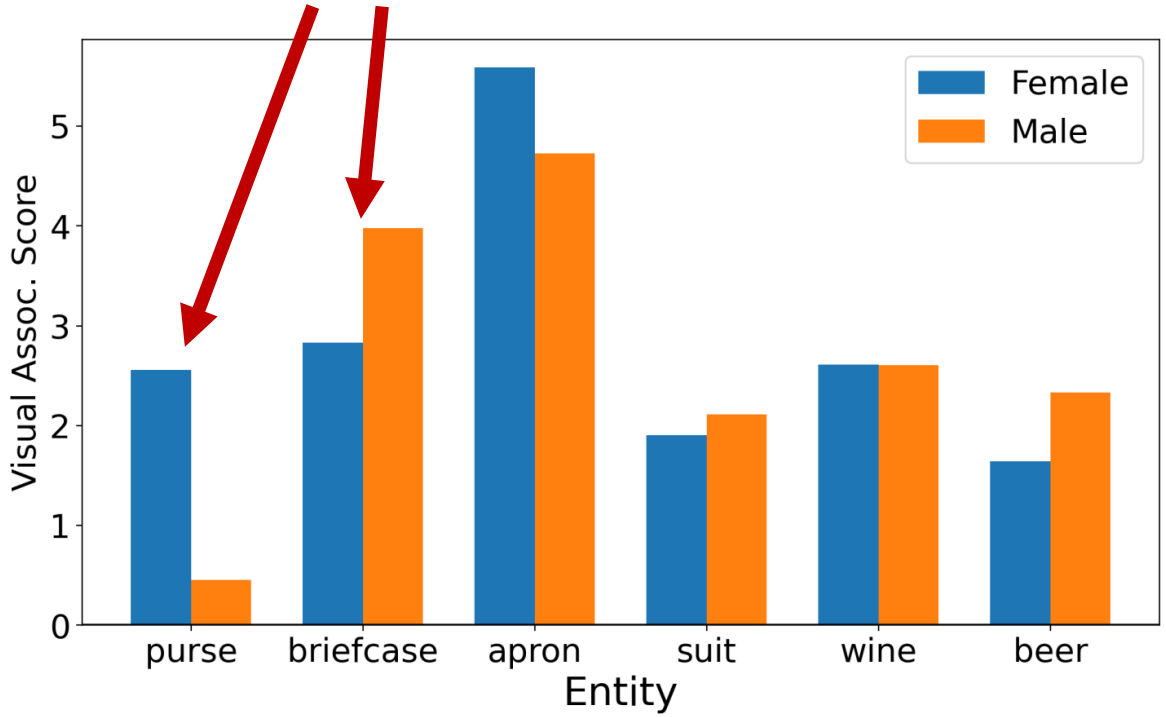
Modality Biases

Heterogeneity in information and relevance

Fairness and social biases – cross-modal interactions worsen social biases

Visual information makes model more confident in reinforcing gender stereotypes





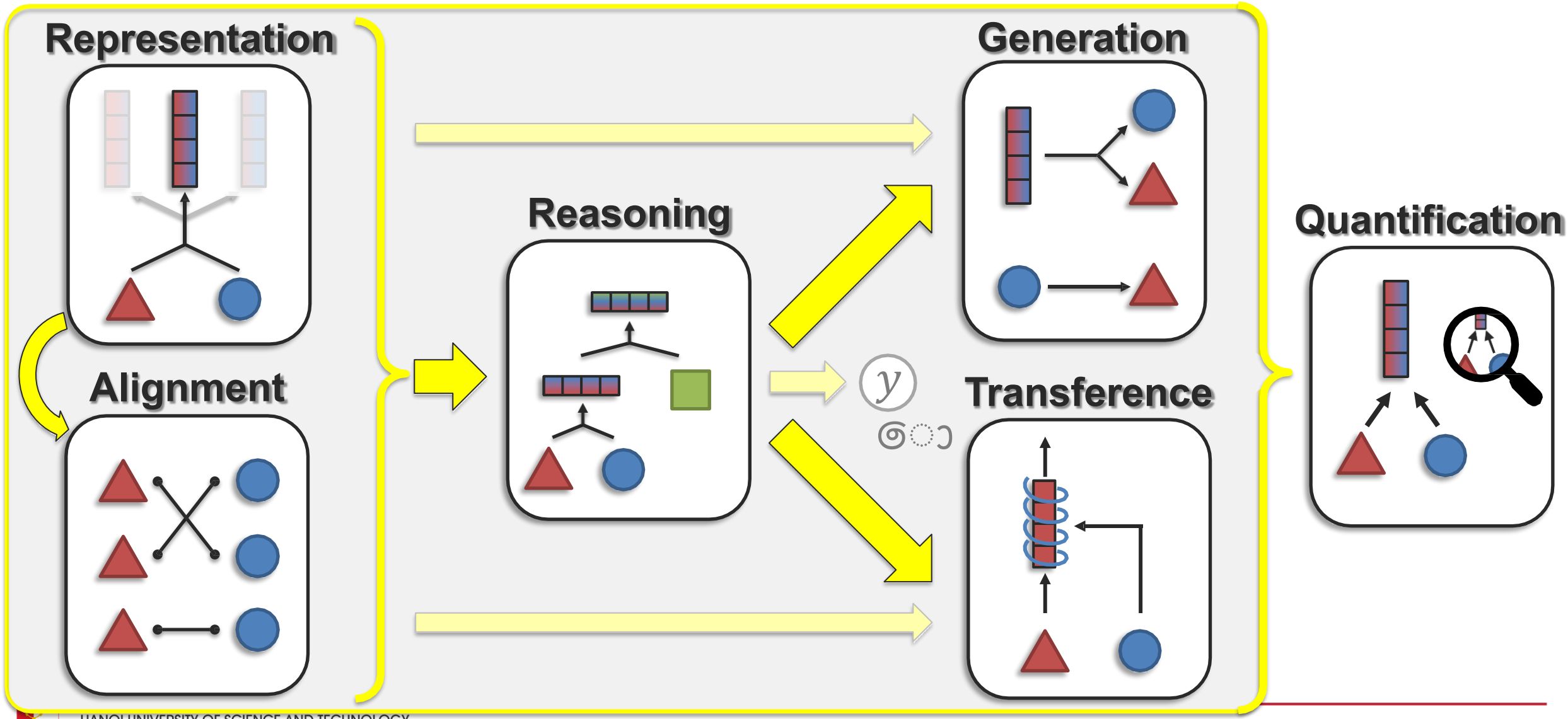
[Srinivasan and Bisk, Worst of Both Worlds: Biases Compound in Pre-trained Vision-and-Language Models. NAACL 2022]

Challenge 6: Quantification

Final Summary

Sub-Challenge	Definition	Concrete Example
Heterogeneity	Quantifying differences across modalities	Measuring **modality importance** in model performance
Interactions	Evaluating cross-modal dependencies	Attention weights in vision-language transformers
Learning	Studying optimization and generalization	Tracking **loss convergence for each modality**

Core Multimodal Challenges



Multimodal Machine Learning Models

Models for Multimodal Learning

1. **Concatenation-based Models** (e.g., Early Fusion)
2. **Attention-based Models** (e.g., Cross-modal transformers like CLIP, ViLBERT)
3. **Graph Neural Networks (GNNs)** (e.g., Graph-based multimodal learning)
4. **Contrastive Learning** (e.g., Aligning image-text embeddings)

Future of Multimodal Learning

Research Challenges

1.Multimodal Explainability

1. Understanding how models integrate different modalities.

2.Efficient & Scalable Fusion

1. Reducing computation for real-time multimodal applications.

3.Few-Shot & Zero-Shot Learning

1. Generalizing to unseen modalities without retraining.

Emerging Trends

- **GPT-4V** (multimodal ChatGPT).
- **Self-Supervised Multimodal Learning** (reducing reliance on labeled data).
- **Embodied AI** (robots using multimodal perception for interaction).

References

- Multi Modal Machine Learning, 11-777 • Fall 2023 • Carnegie Mellon University

Thanks