



Language
Technologies
Institute

Carnegie
Mellon
University

Multimodal Machine Learning

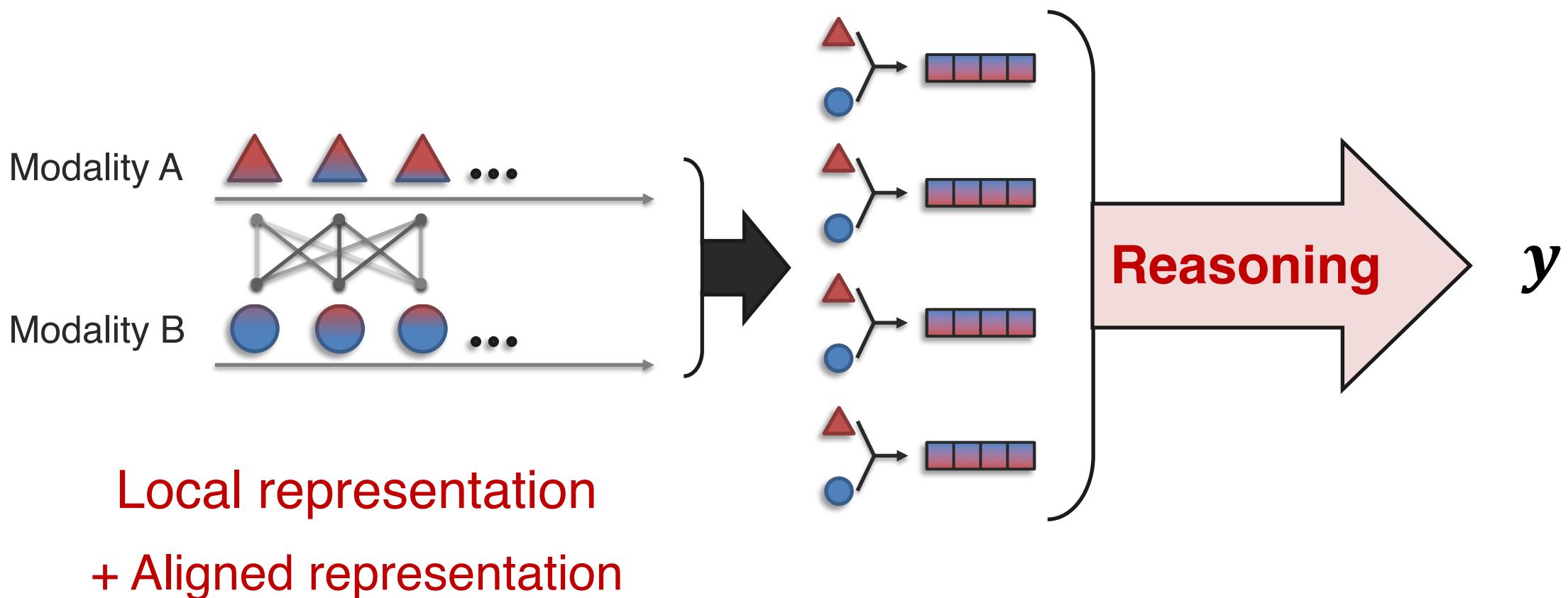
Lecture 6.2: Reasoning 1 Structure + Compositionality

Paul Liang

** Original course co-developed with Tadas Baltrusaitis.
Spring 2021 edition taught by Yonatan Bisk*

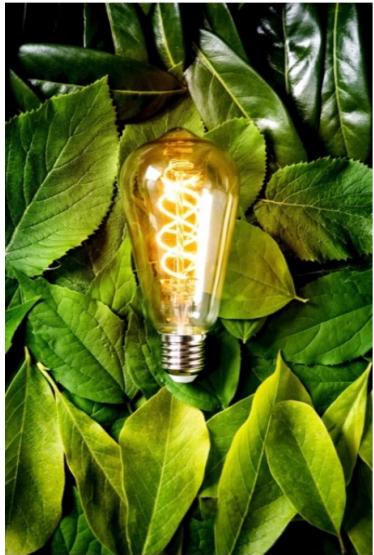
Reasoning

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



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]

The Challenge of Compositionality

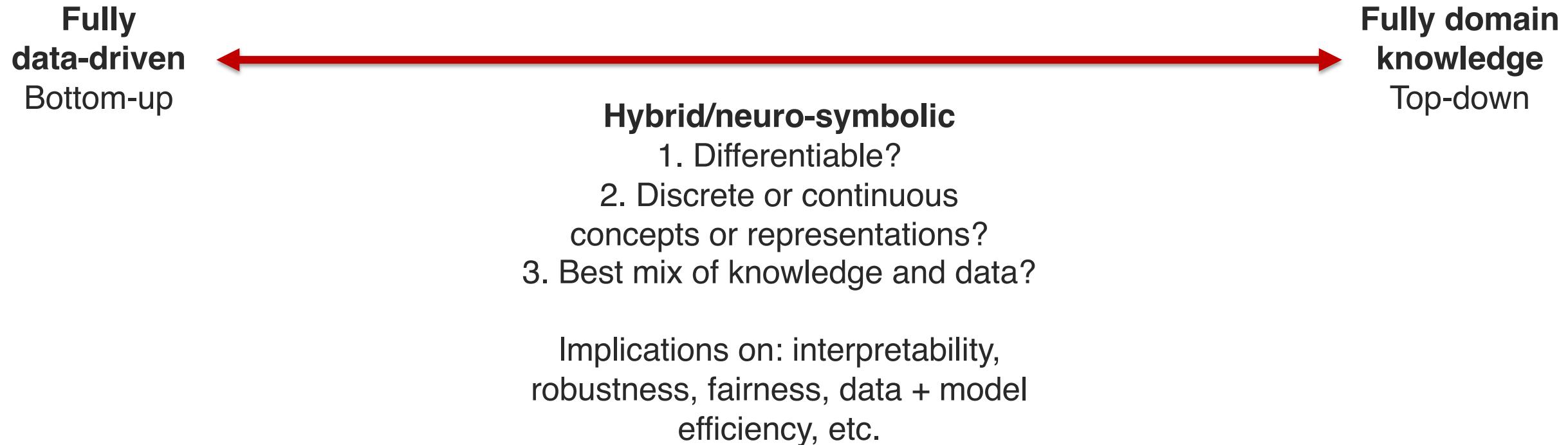
Definition: Combining knowledge, usually through multiple inferential steps, exploiting multimodal alignment and problem structure.



A horse riding an astronaut.

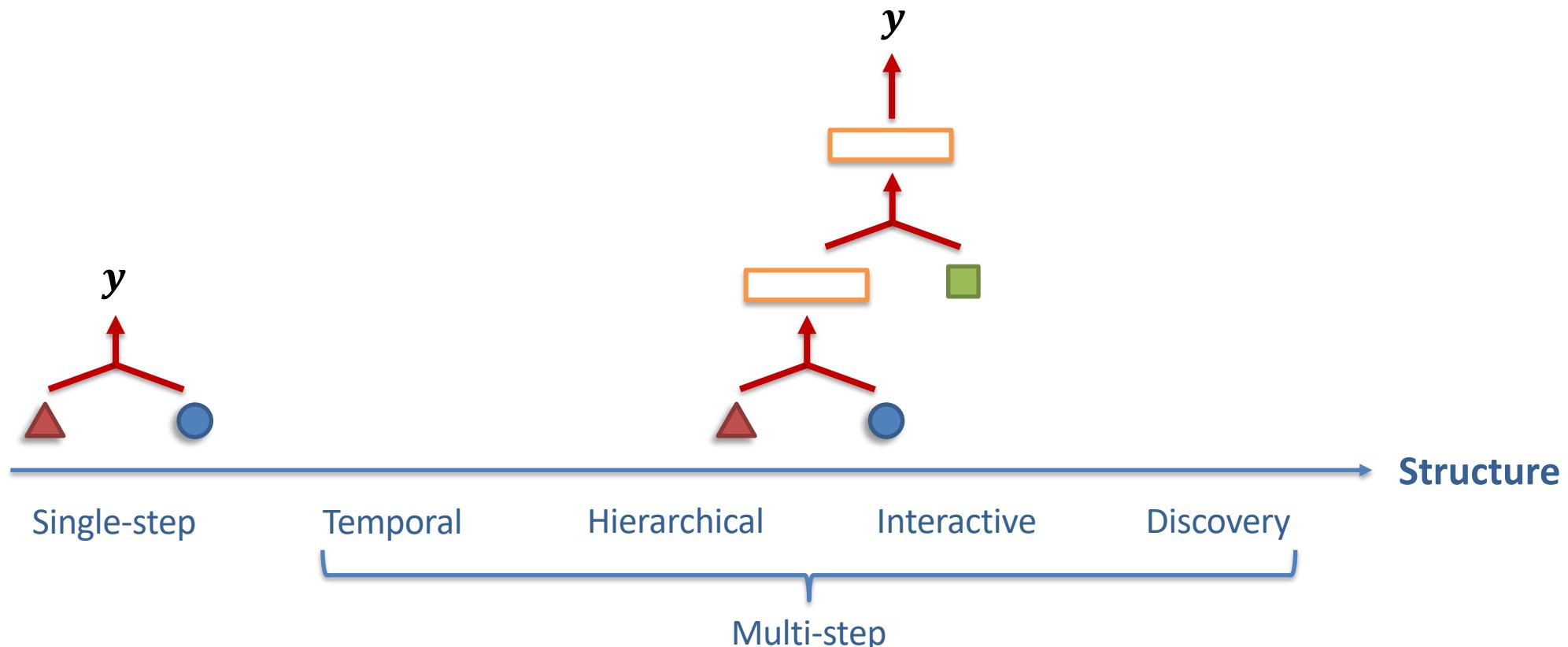
[Saharia et al., Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. arXiv 2022]

Many Debates Surrounding Reasoning



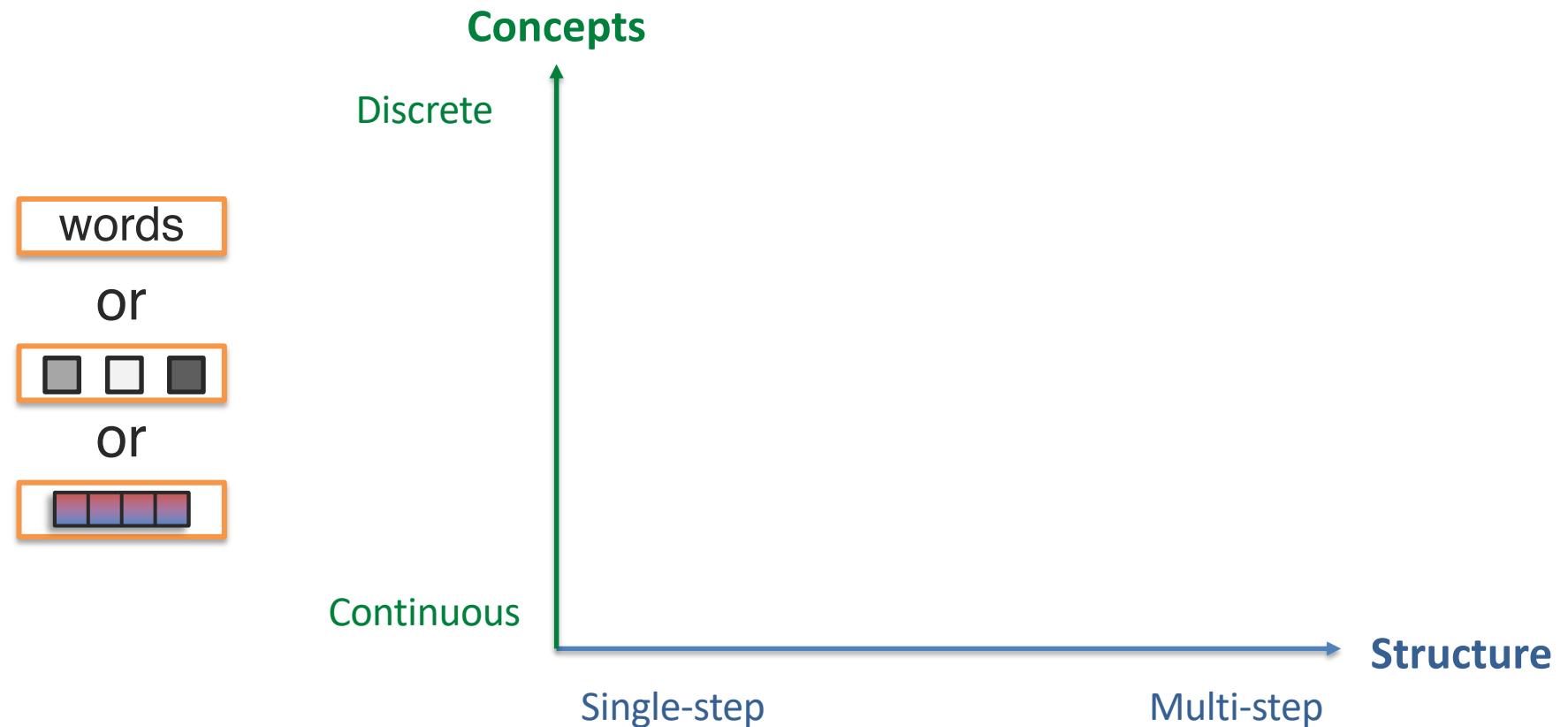
Sub-Challenge 3a: Structure Modeling

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



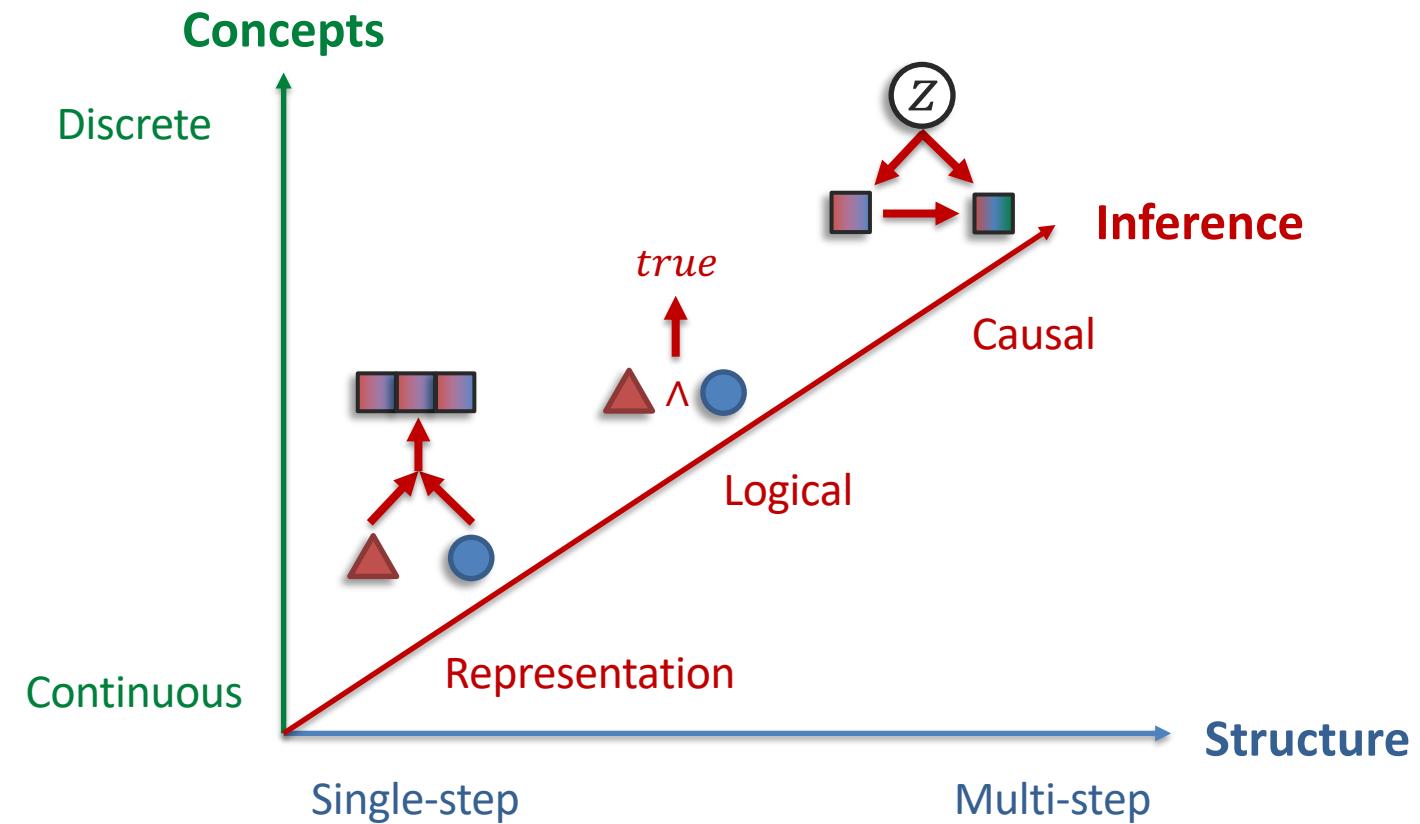
Sub-Challenge 3b: Intermediate Concepts

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



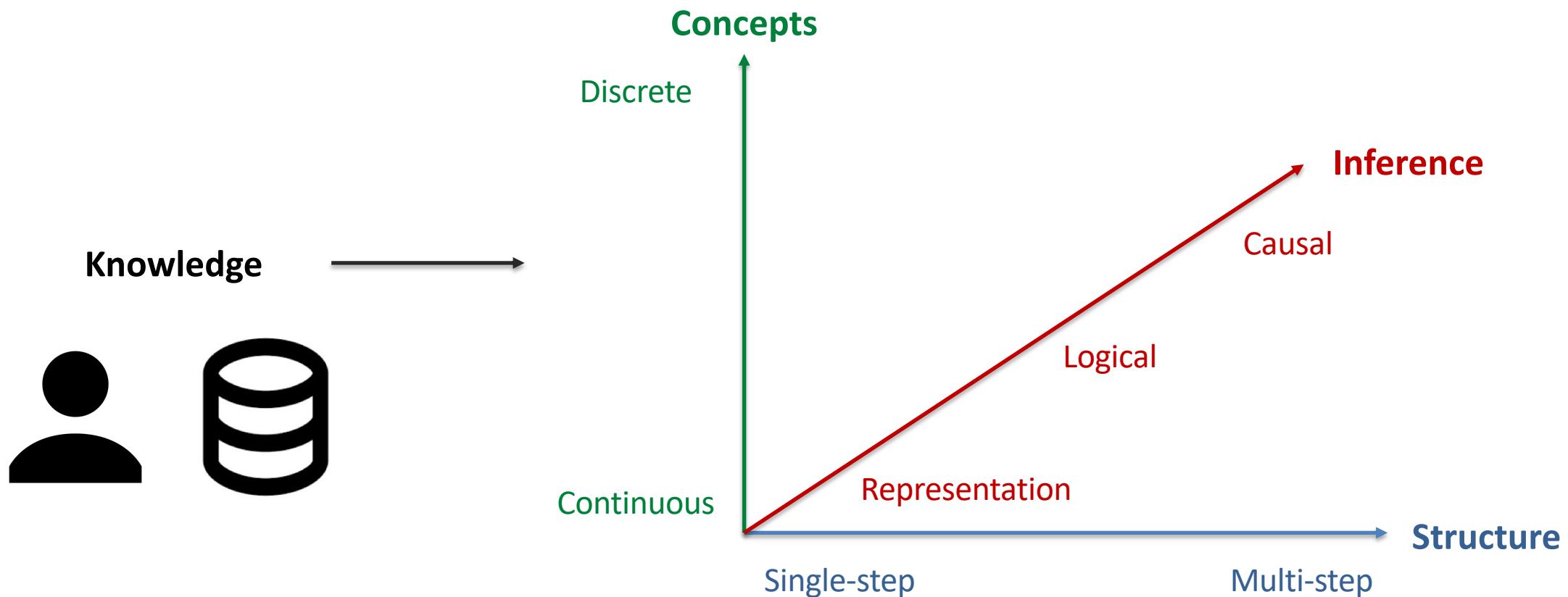
Sub-Challenge 3c: Inference Paradigm

Definition: How increasingly abstract concepts are inferred from individual multimodal evidences.



Sub-Challenge 3d: External Knowledge

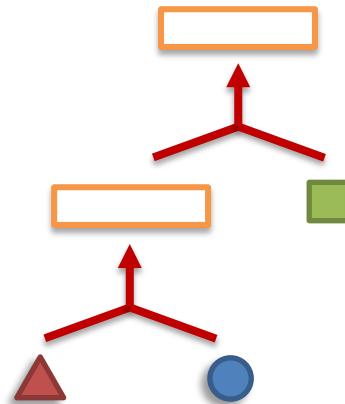
Definition: Leveraging external knowledge in the study of structure, concepts, and inference.



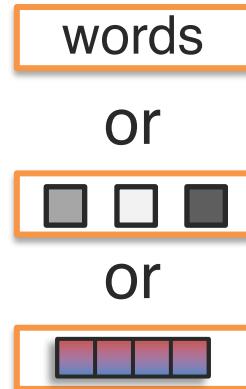
Reasoning

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

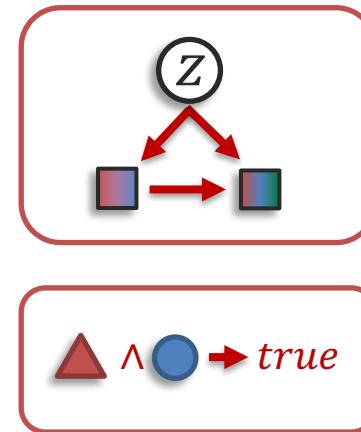
(A) Structure modeling



(B) Intermediate concepts



(C) Inference paradigm



(D) External knowledge



Roadmap for Next 3 Lectures

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

(A) Structure modeling

(B) Intermediate concepts

(C) Inference paradigm

(D) External knowledge

Today

Temporal
Hierarchical

Continuous

Next Tuesday

Interactive
Discovery

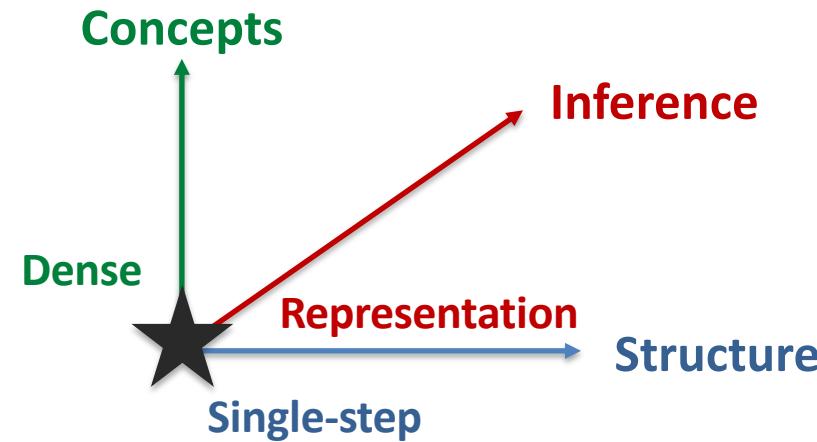
Discrete

Next Thursday

Causal
Logical

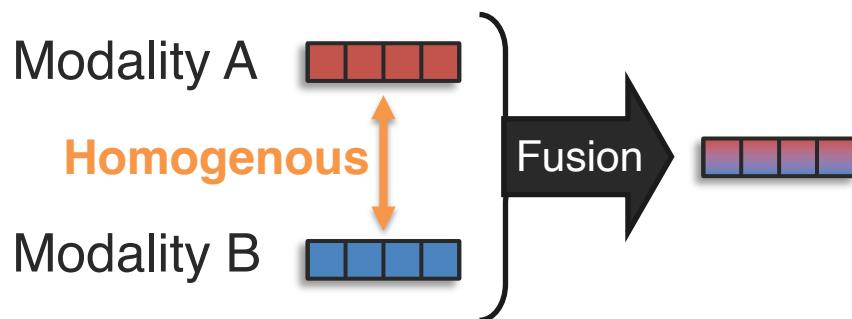
Knowledge
Commonsense

Reasoning

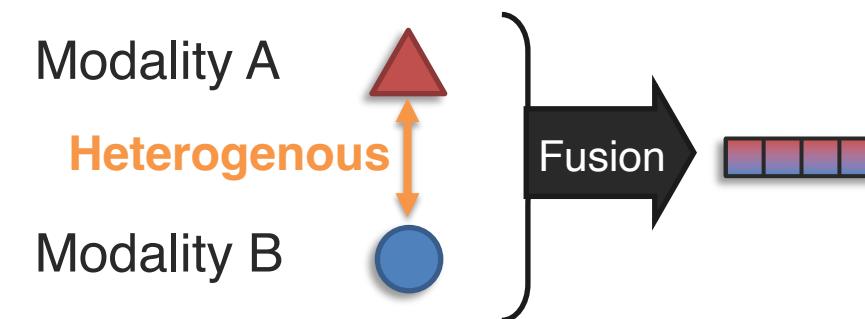


Recall representation fusion!

Basic fusion:



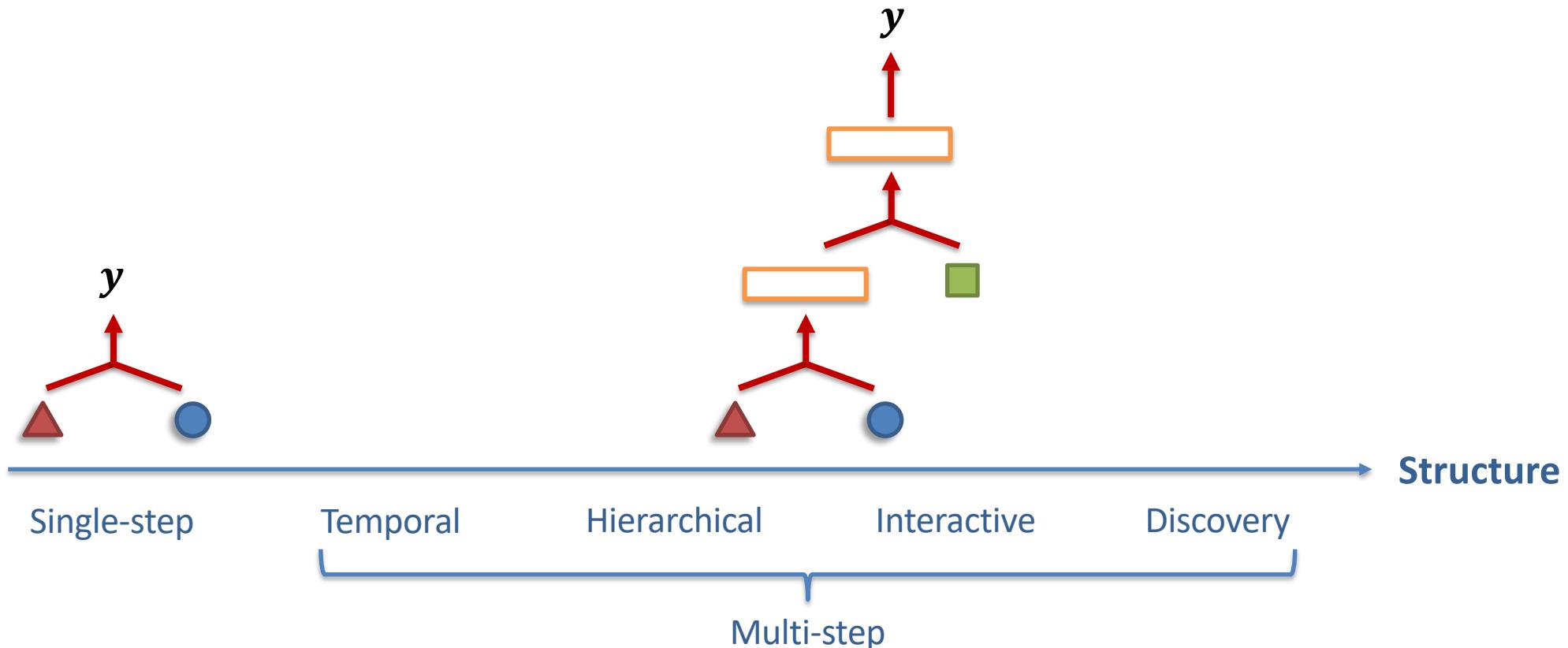
Complex fusion:



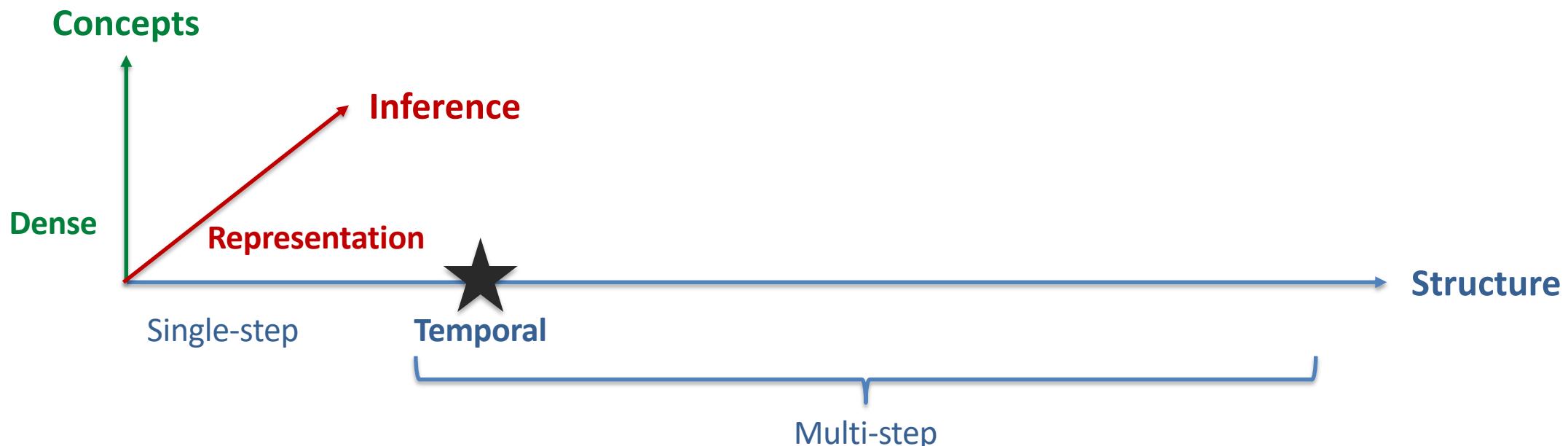
Ideas also apply here, but can we be explicitly interpretable and robust?

Sub-Challenge 3a: Structure Modeling

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



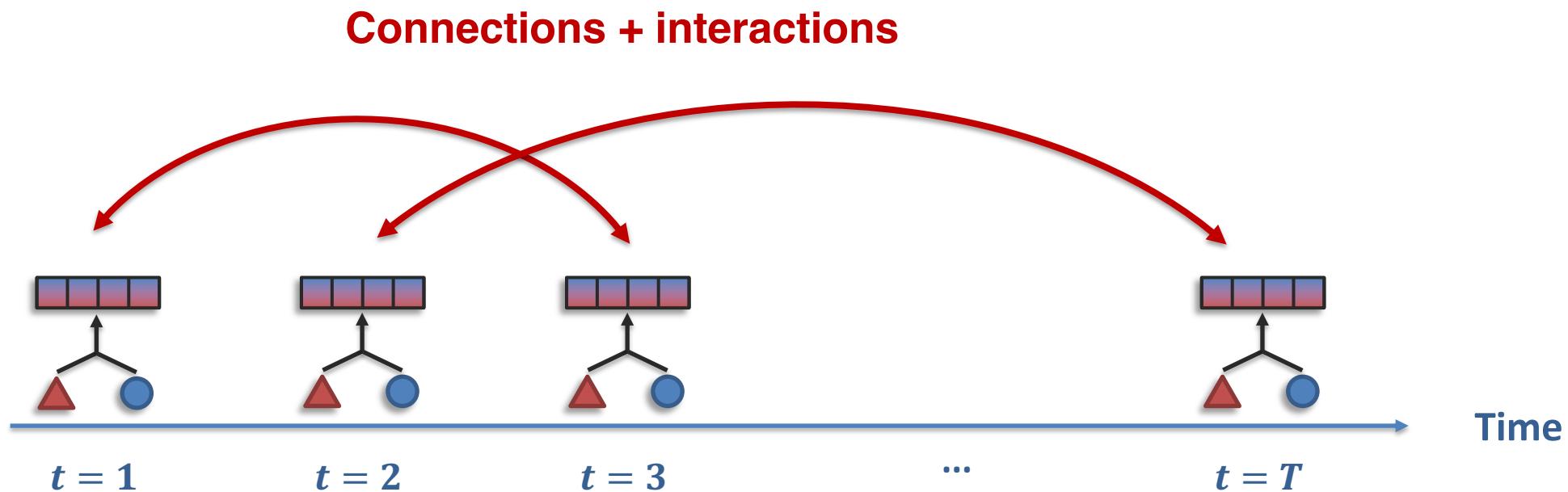
Sub-Challenge 3a: Structure Modeling



Temporal Structure

Temporal structure in multi-view sequences

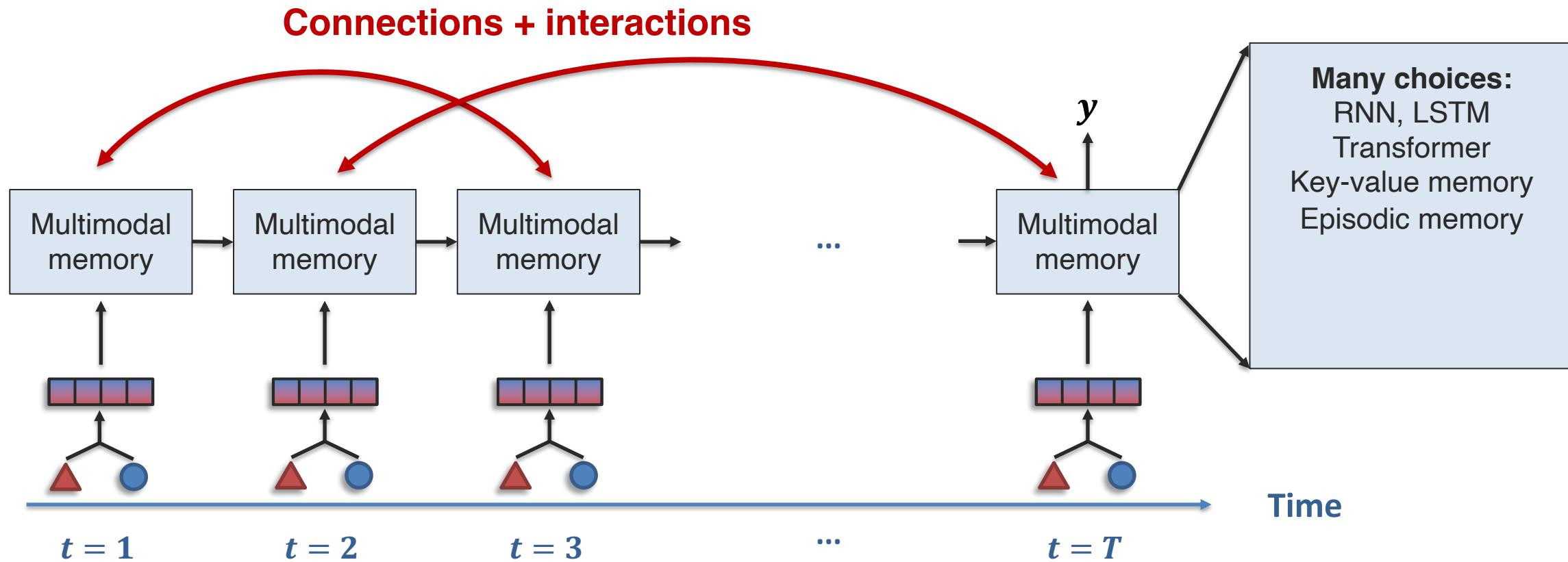
How can we capture cross-modal interactions across time?



Temporal Structure

Temporal structure in multi-view sequences

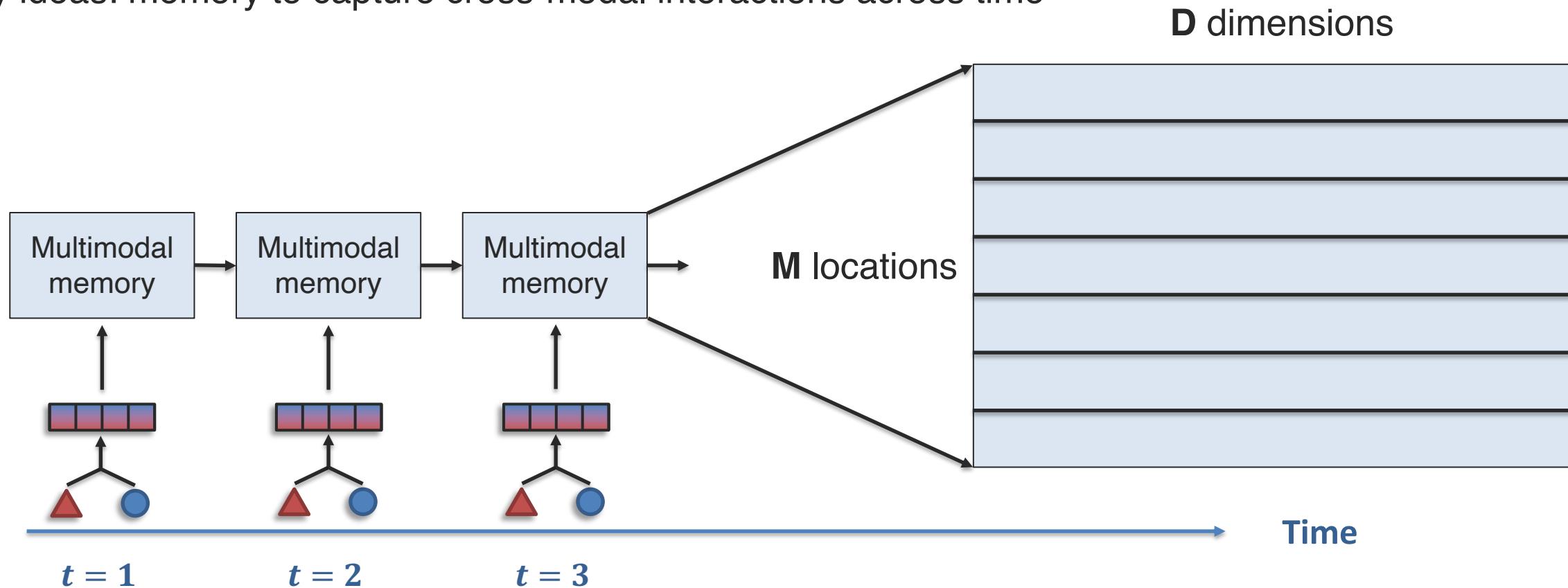
Key ideas: memory to capture cross-modal interactions across time



Temporal Structure

Temporal structure in multi-view sequences

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



Temporal Structure

Temporal structure in multi-view sequences

①

Input

Coordination function measuring **similarity** between input and memory to weight input:

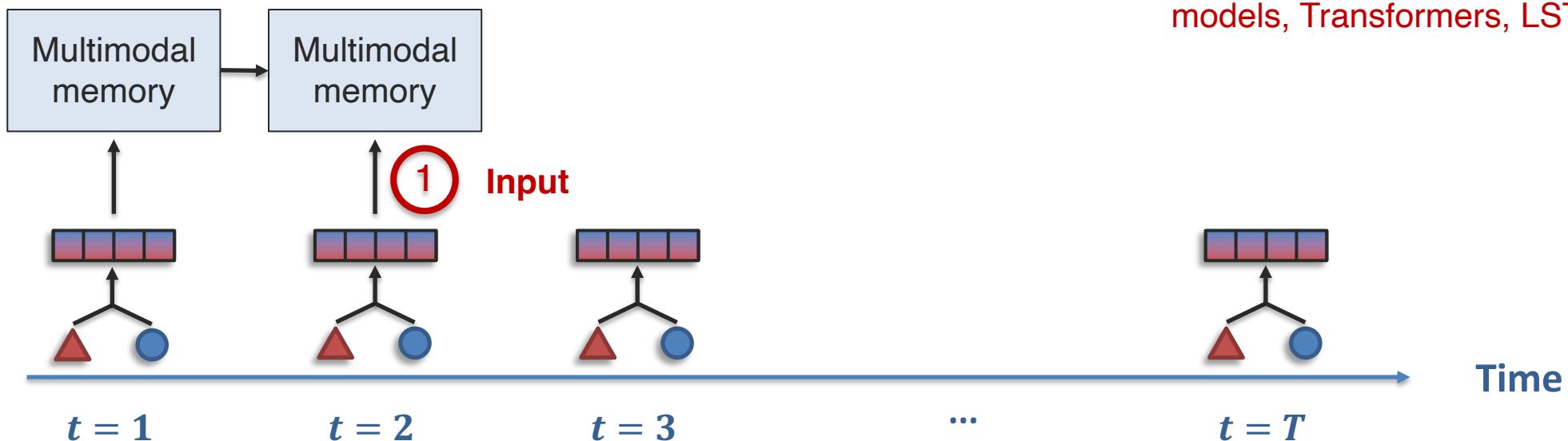
$$\mathbf{w}_t = \text{sim}(\mathbf{x}_t, \mathbf{M}_t) = \mathbf{M}_t \mathbf{x}_t$$

$$\text{Input} = \mathbf{w}_t \mathbf{x}_t^T$$

Input is M x D

Normalized vector of M entries
-> weights over M memory locations

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



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

Temporal Structure

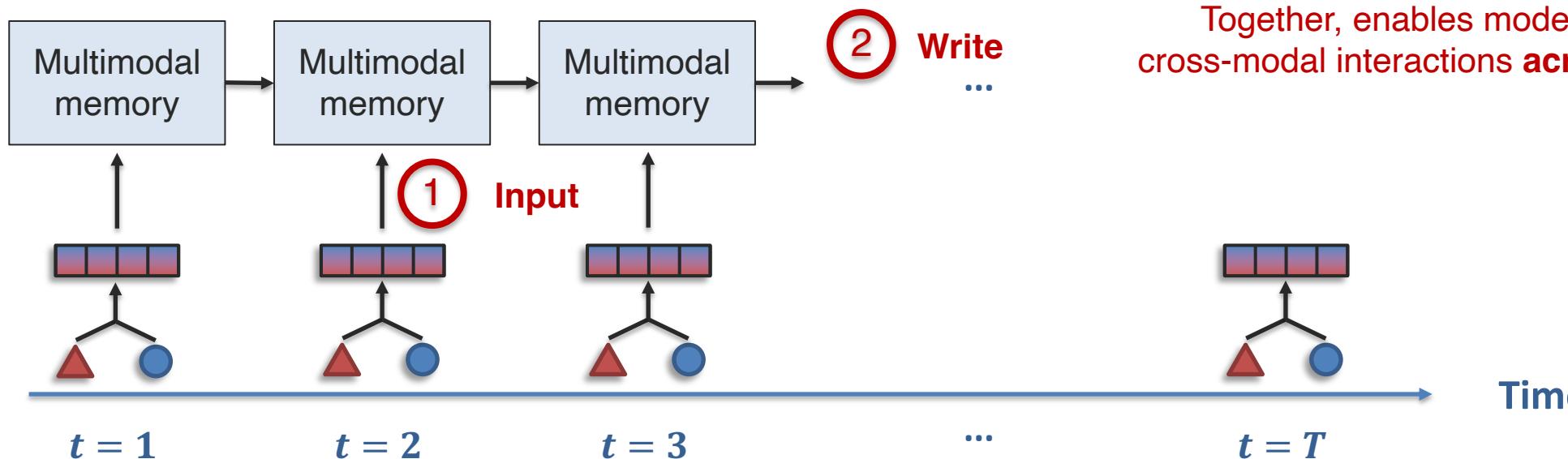
Temporal structure in multi-view sequences

② Write

Weighted function to write new addition into memory

$$\mathbf{M}_{t+1} = (1 - \alpha_t) \mathbf{M}_t + \alpha_t \text{Input}$$

α_t learnable in LSTM/RNN/Transformers, or similarity function in parameterized memory.



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

Temporal Structure

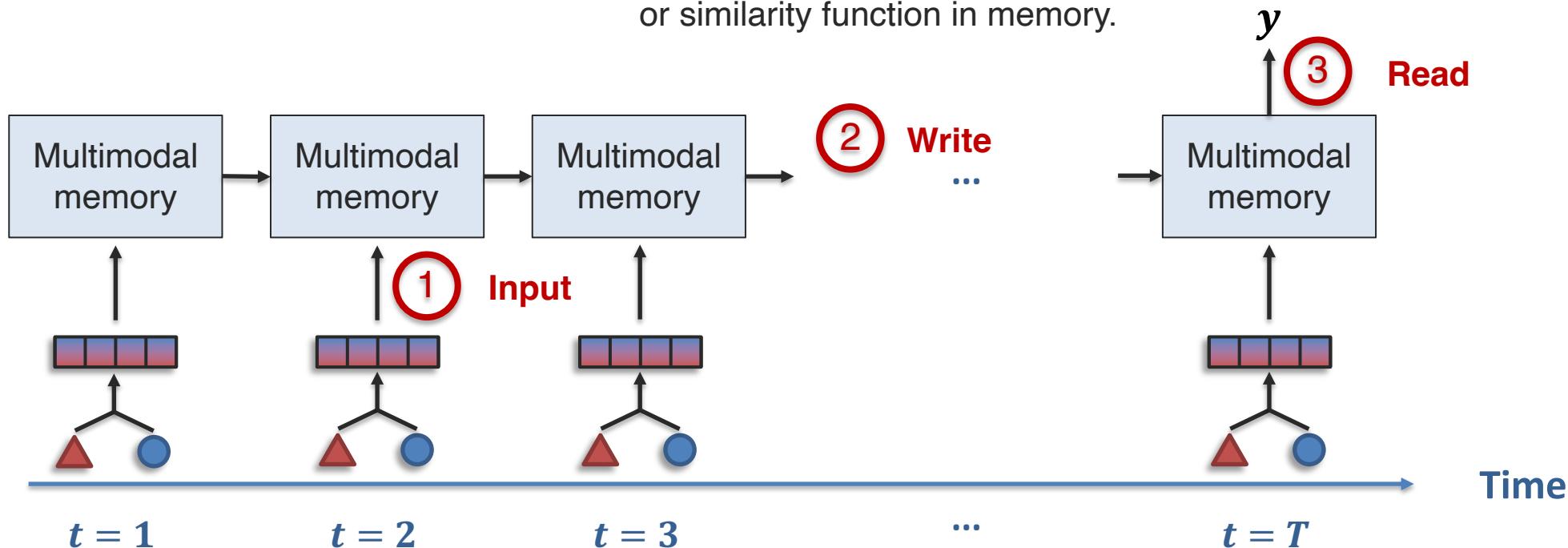
Temporal structure in multi-view sequences

③

Read

Summary function to **read** multimodal information

$$\text{Read} = \beta_T M_T \quad \beta_T \text{ learnable in LSTM/ Transformers, or similarity function in memory.}$$



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

Some Extensions

1. Input: Different addressing mechanisms

①

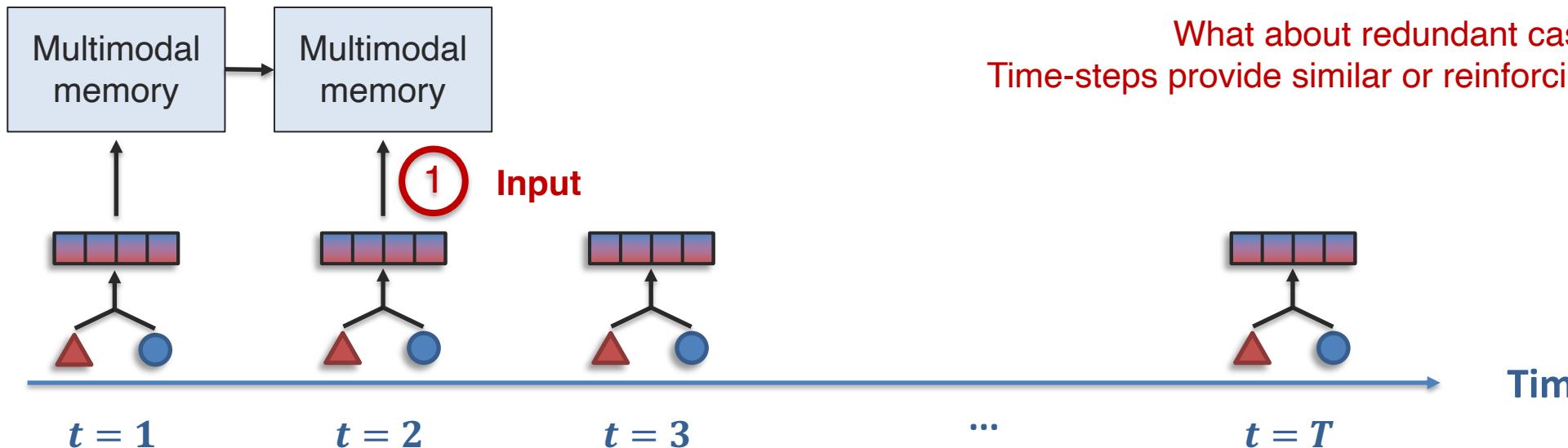
Input

Coordination function measuring **similarity** between input and memory to weight input:

$$\mathbf{w}_t = \text{sim}(\mathbf{x}_t, \mathbf{M}_t) = \mathbf{M}_t \mathbf{x}_t$$

$$\text{Input} = \mathbf{w}_t \mathbf{x}_t^T$$

Okay if different timesteps provide different information – get added to different memory cells (i.e., non-redundancy)



What about redundant case?
Time-steps provide similar or reinforcing information?

[Graves et al., Neural Turing Machines. arXiv 2014]

Some Extensions

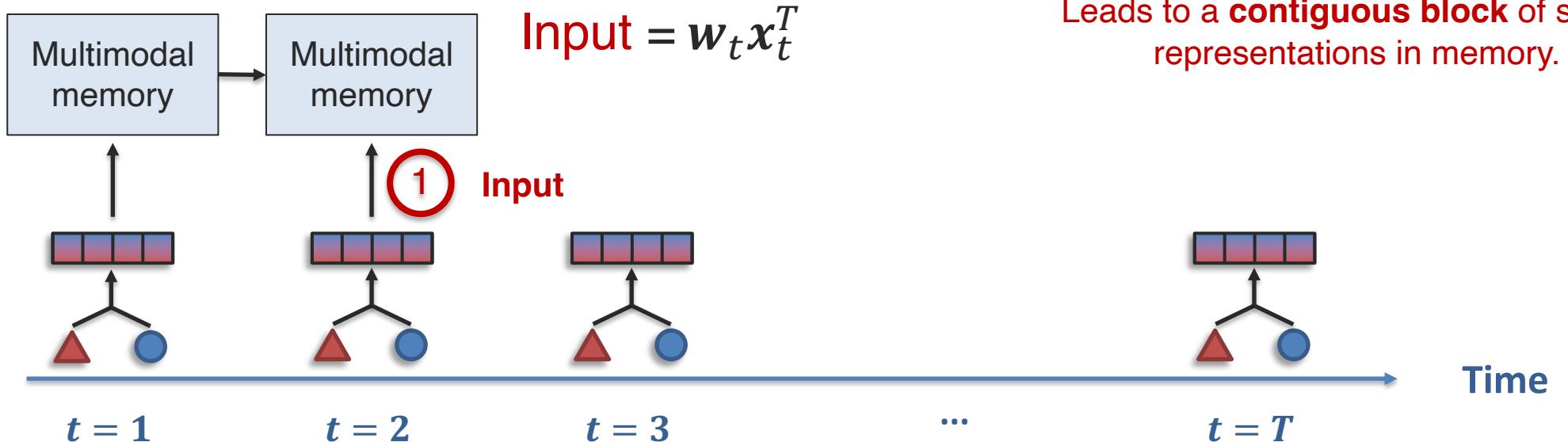
1. Input: Different addressing mechanisms – by location

① **Input**

Coordination function measuring **similarity** between input and memory to weight input, while also keeping previous input indices into account:

$$\mathbf{w}_t = \text{sim}(\mathbf{x}_t, \mathbf{M}_t) = \mathbf{M}_t \mathbf{x}_t$$

$$\mathbf{w}_t = \text{rotate}(\mathbf{w}_t, \mathbf{w}_{t-1})$$



Idea: take previous input indices into account, apply rotation to new indices upon repetition

Leads to a **contiguous block** of similar representations in memory.

[Graves et al., Neural Turing Machines. arXiv 2014]

Some Extensions

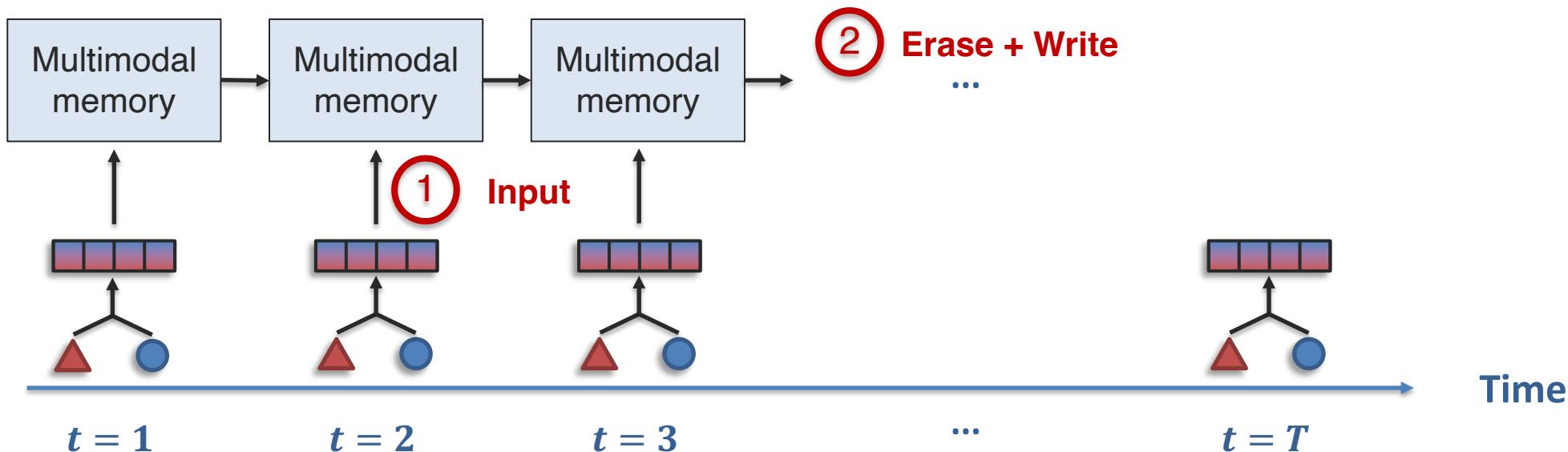
2. Writing: Including both erase and write functions

② Erase + write

$$\mathbf{M}_t = \mathbf{M}_t[1 - \alpha_t \text{Erase}]$$

Erase: learnable vector of [0,1]

$$\mathbf{M}_{t+1} = \mathbf{M}_t + \alpha_t \text{Input}$$



[Graves et al., Neural Turing Machines. arXiv 2014]

Some Extensions

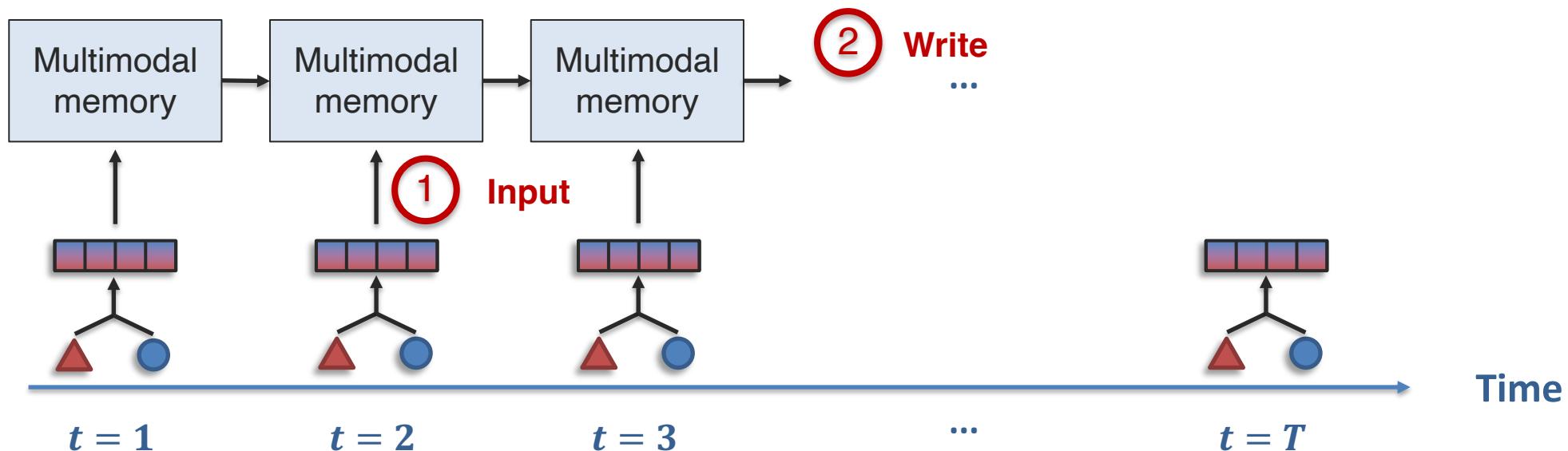
3. More exponential moving average

② Write

Write new addition into memory

$$M_{t+1} = (1 - \alpha_t)M_t + \alpha_t \text{Input}$$

Exponential moving average function
- Smooth out short-term fluctuations
- Highlight long-term trends



[Ma et al., Mega: Moving Average Equipped Gated Attention. arXiv 2022]

Some Extensions

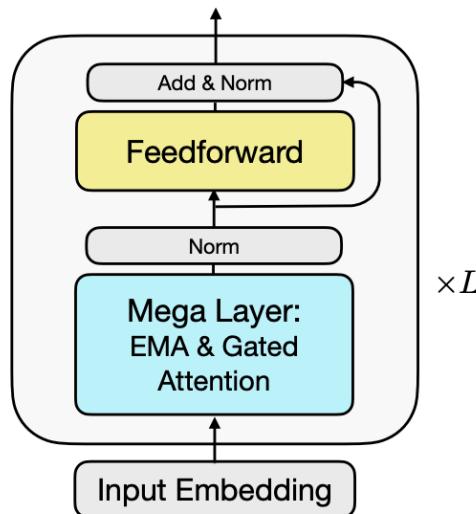
3. More exponential moving average + combine with Transformers

② Write

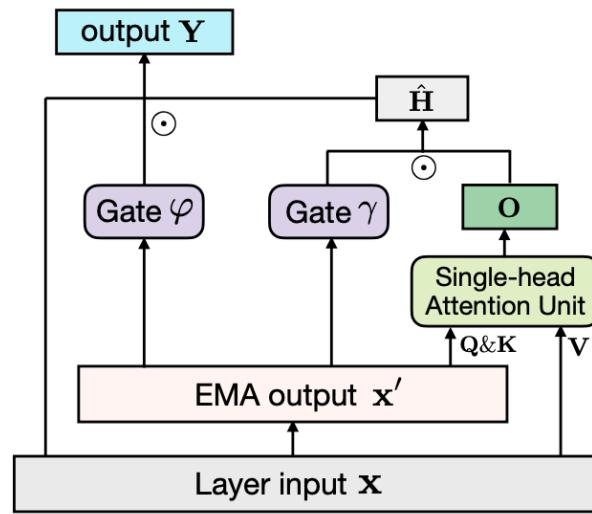
Write new addition into memory

$$\mathbf{M}_{t+1} = (1 - \alpha_t) \mathbf{M}_t + \alpha_t \text{Input}$$

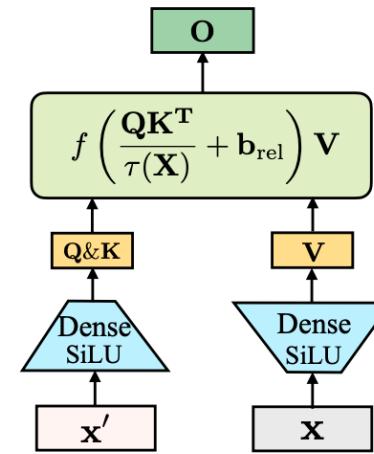
Exponential moving average function
- Smooth out short-term fluctuations
- Highlight long-term trends



(a) Mega architecture.



(b) Mega layer.



(c) Single-head attention unit.

[Ma et al., Mega: Moving Average Equipped Gated Attention. arXiv 2022]

Some Extensions

4. From recurrent to parallel convolutions

A lot of what we presented seemed to be recurrent, which may not seem easily parallelizable.

But many of these have equivalent formulations in convolutional representations.

Key idea: exponential moving average can be implemented as convolution.

② Write

Write new addition into memory

$$\mathbf{M}_{t+1} = (1 - \alpha_t) \mathbf{M}_t + \alpha_t \text{Input} \quad \mathbf{M}_T = K * [\text{Input}_1, \dots, \text{Input}_T]$$

BUT: K will be huge, the size of the entire sequence.

Many approximations, optimizations, see references below.

[Gu et al., Efficiently Modeling Long Sequences with Structured State Spaces. ICLR 2022]

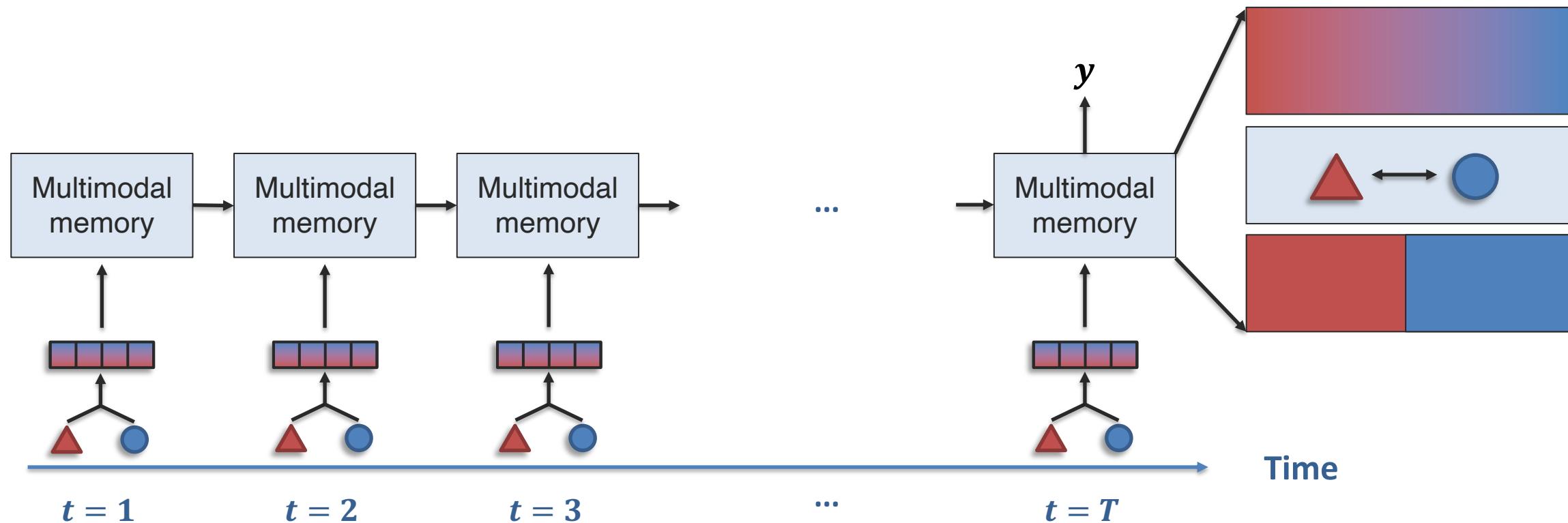
[Ma et al., Mega: Moving Average Equipped Gated Attention. arXiv 2022]

Memory for Multimodal Sequences

1. Memory + representation

We've seen early fusion of raw modalities

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

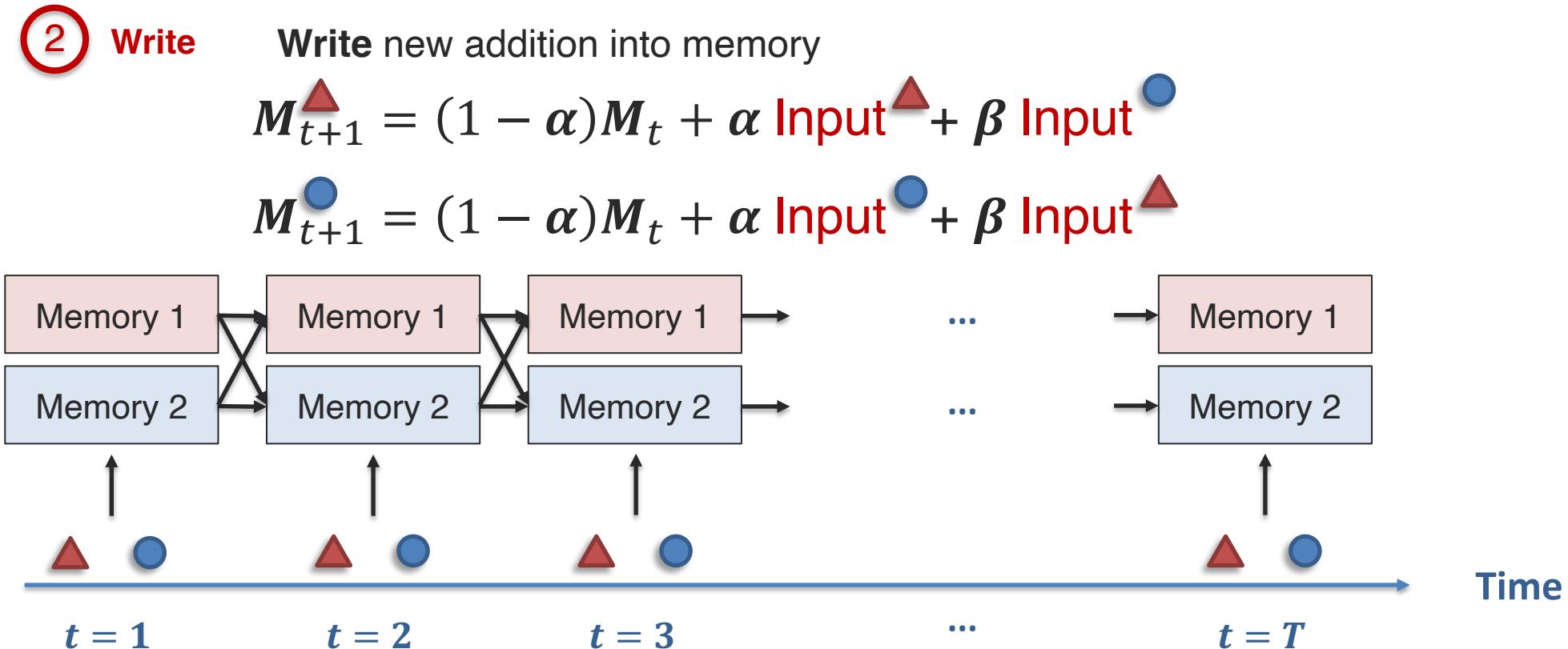


[Rajagopalan et al., Extending Long Short-Term Memory for Multi-View Structured Learning. ECCV 2016]

Memory for Multimodal Sequences

1. Memory + representation

Representation can be learned not just prior to memory but also inside memory cells



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

Memory for Multimodal Sequences

1. Memory + representation

Fully unimodal

$$x_t^{(v)} \rightarrow g_t^{(v)} \quad \alpha = 1 \quad \beta = 0$$

$$\begin{matrix} h_{t-1}^{(v)} \\ h_t^{(k)}_{k \neq v} \end{matrix}$$

(a)

$$x_t^{(v)} \rightarrow g_t^{(v)} \quad \alpha = 1/3 \quad \beta = 1/3$$

$$\begin{matrix} h_{t-1}^{(v)} \\ h_t^{(k)}_{k \neq v} \end{matrix}$$

(b)

$$x_t^{(v)} \rightarrow g_t^{(v)} \quad \alpha = 2/3 \quad \beta = 2/3$$

$$\begin{matrix} h_{t-1}^{(v)} \\ h_t^{(k)}_{k \neq v} \end{matrix}$$

(c)

$$x_t^{(v)} \rightarrow g_t^{(v)} \quad \alpha = 0 \quad \beta = 1$$

$$\begin{matrix} h_{t-1}^{(v)} \\ h_t^{(k)}_{k \neq v} \end{matrix}$$

(d)

$$x_t^{(v)} \rightarrow g_t^{(v)} \quad \alpha = 1 \quad \beta = 1$$

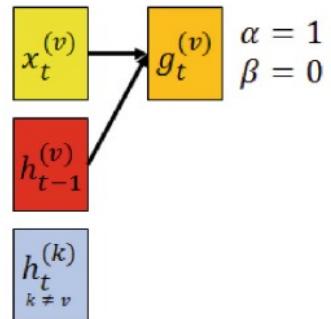
$$\begin{matrix} h_{t-1}^{(v)} \\ h_t^{(k)}_{k \neq v} \end{matrix}$$

(e)

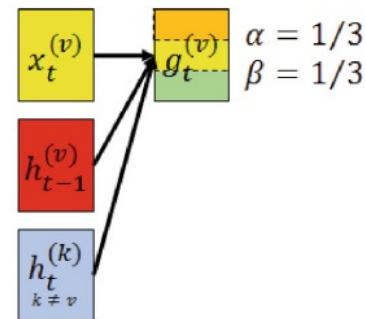
[Rajagopalan et al., Extending Long Short-Term Memory for Multi-View Structured Learning. ECCV 2016]

Memory for Multimodal Sequences

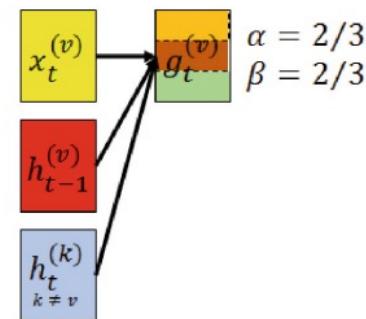
1. Memory + representation



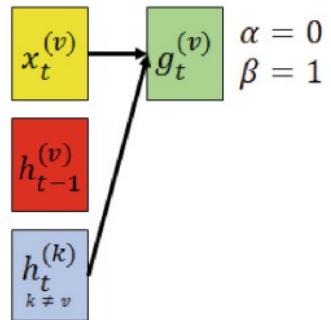
(a)



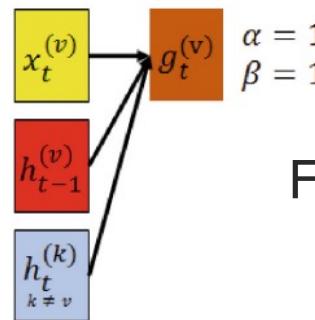
(b)



(c)



(d)



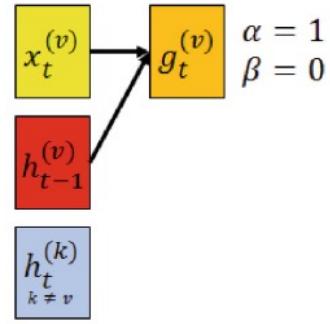
(e)

Fully fusion

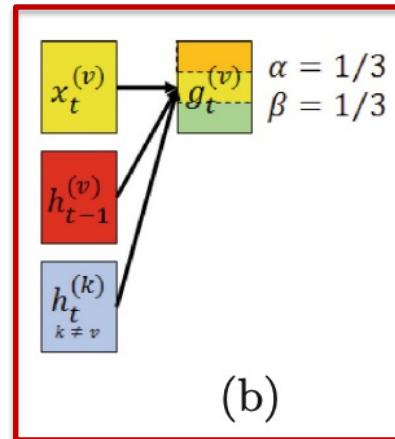
[Rajagopalan et al., Extending Long Short-Term Memory for Multi-View Structured Learning. ECCV 2016]

Memory for Multimodal Sequences

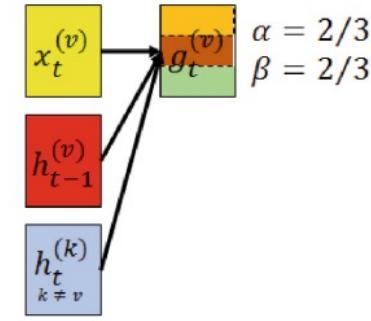
1. Memory + representation



(a)

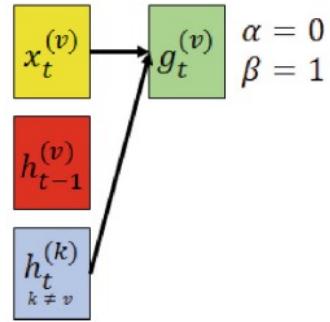


(b)

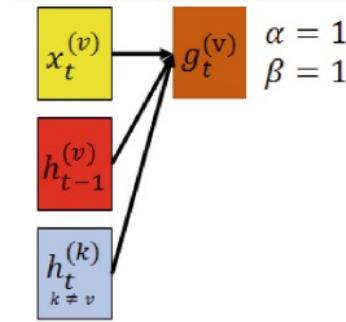


(c)

Hybrid
fusion +
fission



(d)



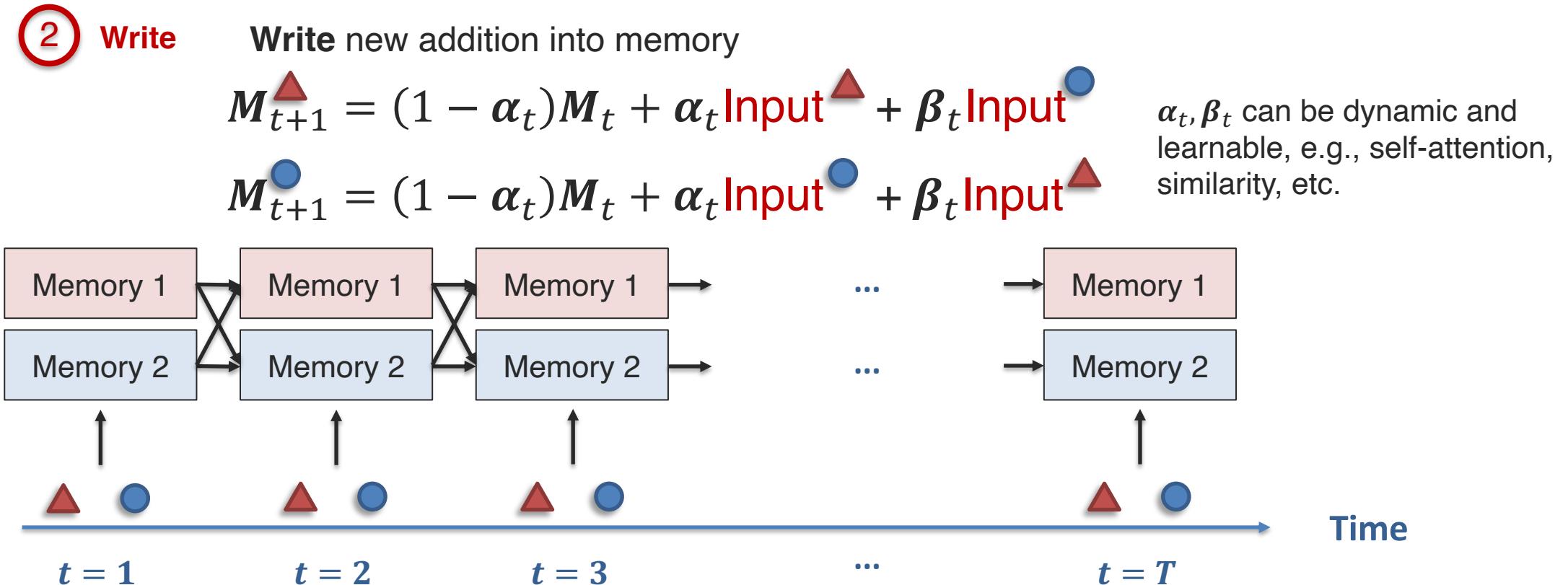
(e)

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

Memory for Multimodal Sequences

1. Memory + representation

Representation can be learned not just prior to memory but also inside memory cells

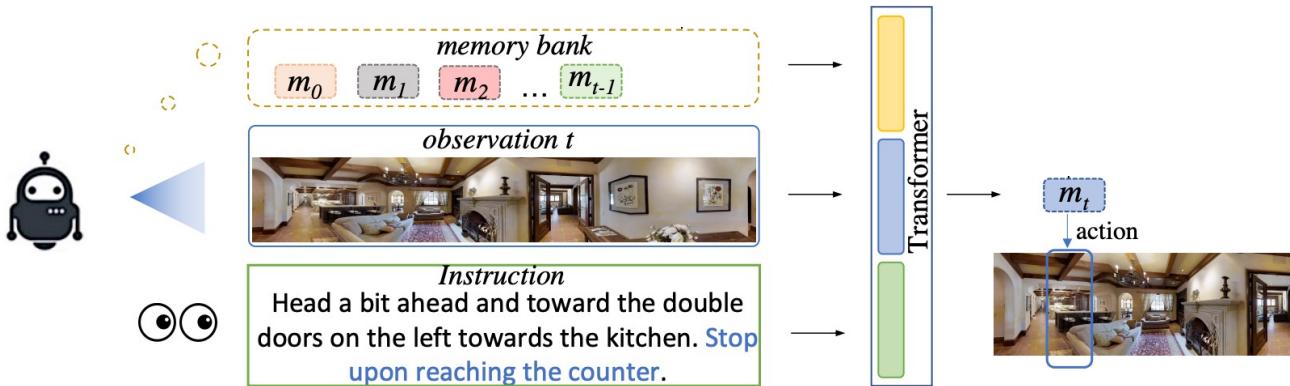


[Zadeh et al., Memory Fusion Network for Multi-view Sequential Learning. AAAI 2018]

Memory for Multimodal Sequences

2. Memory + aligned contextualized representations

Where have I visited previously?



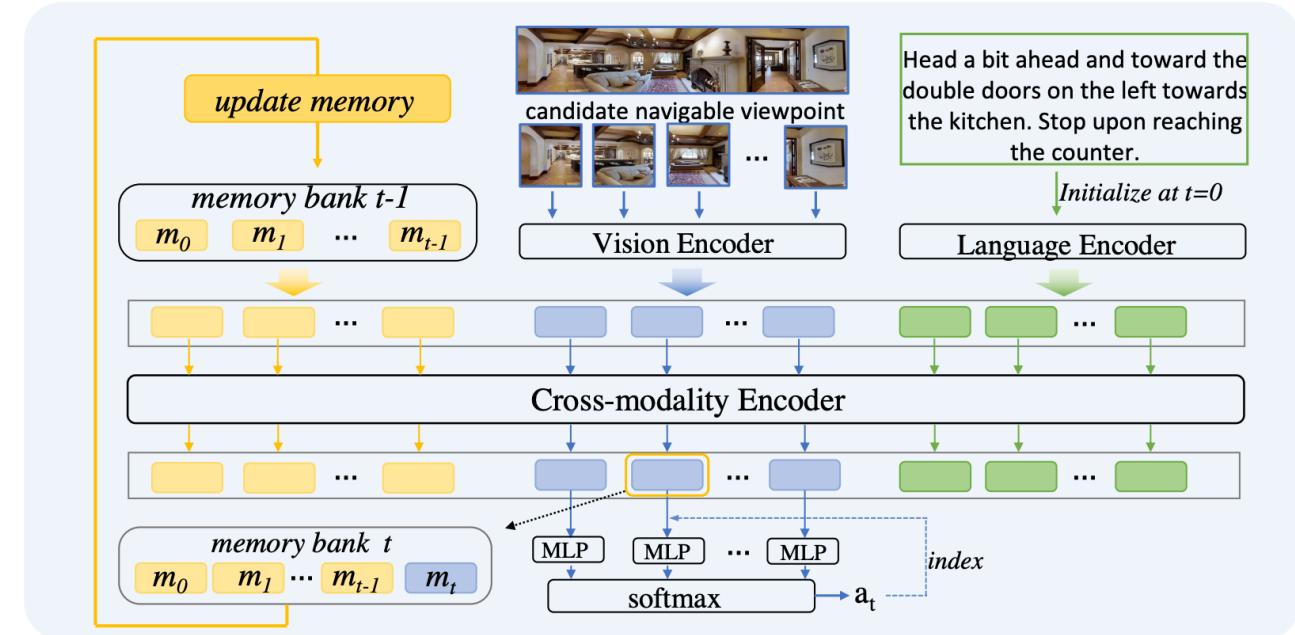
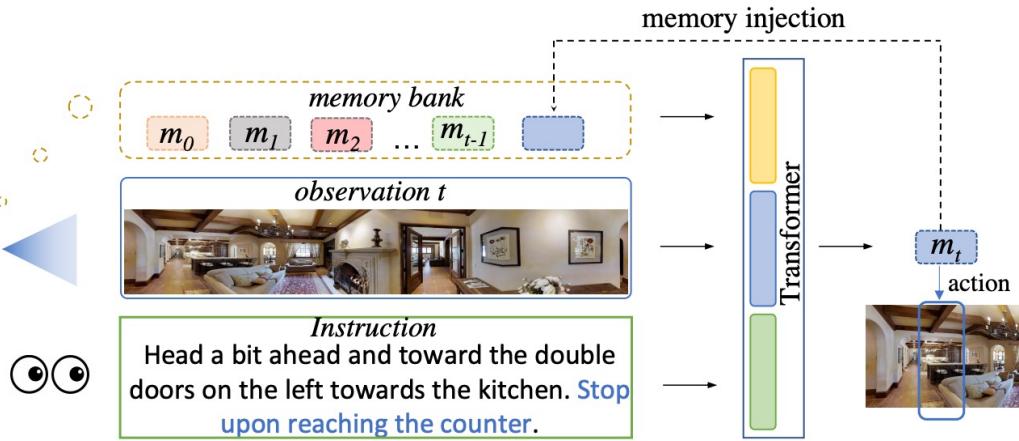
[Chen et al., History Aware Multimodal Transformer for Vision-and-Language Navigation. NeurIPS 2021]

[Lin et al., Multimodal Transformer with Variable-length Memory for Vision-and-Language Navigation. ECCV 2022]

Memory for Multimodal Sequences

2. Memory + aligned contextualized representations

Where have I visited previously?



+ Contextualized representations
+ Memory mechanisms

[Chen et al., History Aware Multimodal Transformer for Vision-and-Language Navigation. NeurIPS 2021]

[Lin et al., Multimodal Transformer with Variable-length Memory for Vision-and-Language Navigation. ECCV 2022]

Use Cases

1. Tasks involving repeating input data in a specific way

Direct copy

$51867 \rightarrow 51867$

Repeat copy

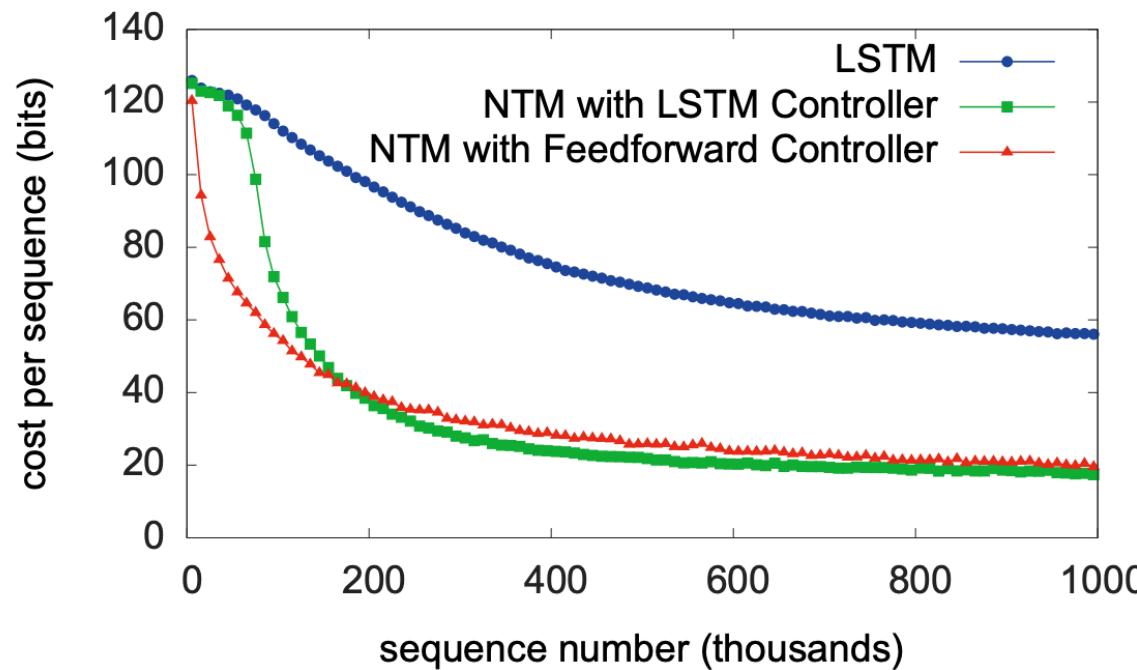
$51867, 3 \rightarrow 51867\ 51867\ 51867$

Associative recall

$51867, 8 \rightarrow 6$
 $1 \rightarrow 8$
 $6 \rightarrow 7$

Sorting

$51867 \rightarrow 15678$



Much better than methods without memory

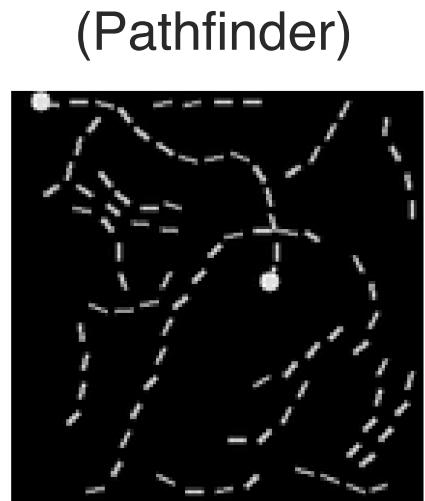
[Graves et al., Neural Turing Machines. arXiv 2014]

Use Cases

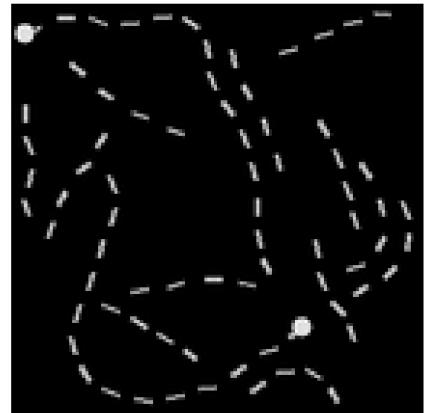
2. Extremely long-range sequences

- List operations: mean, max, sum, etc
- Long document classification and retrieval
- Image classification via sequence of pixels
- Pathfinder
- (Generating) long speech signals

$T = 2K$
 $T = 4K$
 $T = 1K$
 $T = 1K$
 $T = 128K$



(a) A positive example.



(b) A negative example.

[Gu et al., Efficiently Modeling Long Sequences with Structured State Spaces. ICLR 2022]

[Goel et al., It's Raw! Audio Generation with State-Space Models. ICML 2022]

Use Cases

3. Changing information across time

Can be implemented by explicitly writing and reading from memory, in contrast to fully neural models which are typically uncontrollable

Input	Year	Uniform	Temporal
X is the chair of Federal Reserve System.	2019	Janet L. Yellen	Jerome Powell
Nigel Farage is a member of the _X_.	2019	UK Independence Party	Brexit Party
Mark Sanford holds the position of _X_.	2017	Governor of South Carolina	United States representative
X is the head of the government of New York City.	2016	Michael Bloomberg	Bill de Blasio
X is the head coach of Real Madrid CF.	2015	Zinedine Zidane	Carlo Ancelotti
Theresa May holds the position of _X_.	2014	Prime Minister of Great Britain	Home Secretary
Peyton Manning plays for _X_.	2014	Indianapolis Colts	Denver Broncos
X is the head of the government of United Kingdom.	2011	Theresa May	David Cameron
Marissa Mayer works for _X_.	2011	Yahoo	Google
Rahm Emanuel holds the position of _X_.	2010	Mayor of Chicago	White House Chief of Staff

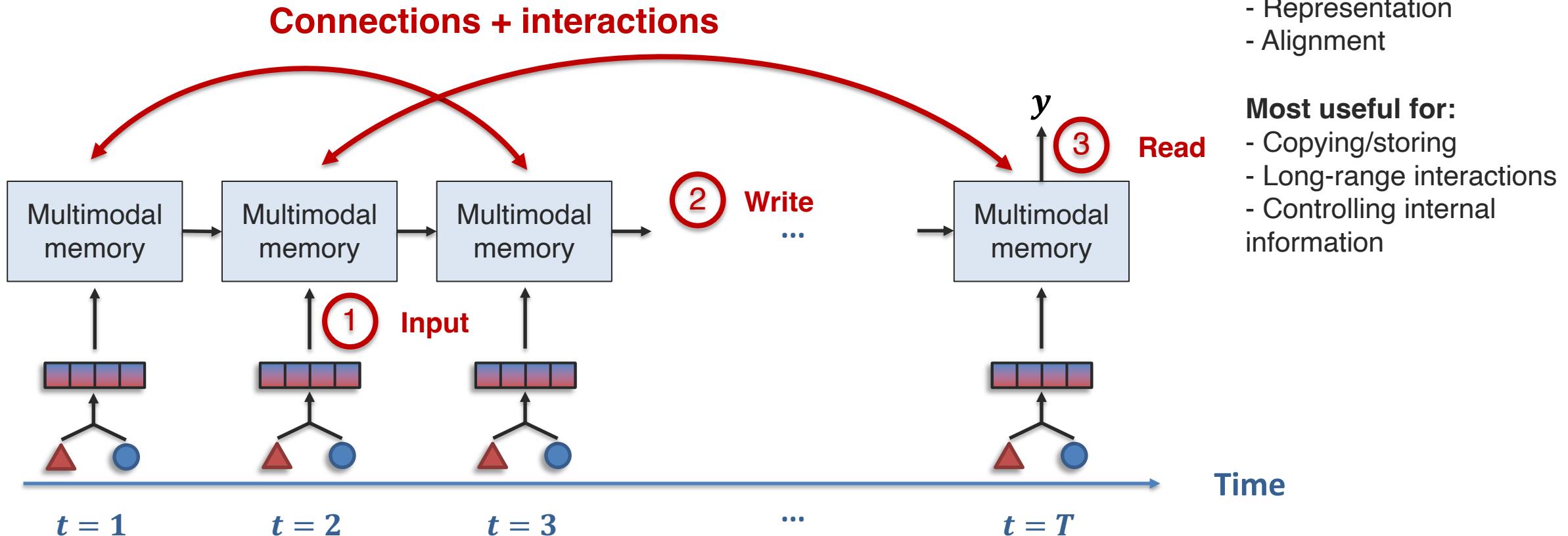
[Wu et al., Memorizing Transformers. ICLR 2022]

[Dhingra et al., Time-Aware Language Models as Temporal Knowledge Bases. TACL 2022]

Key Takeaways

Temporal structure in multi-view sequences

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



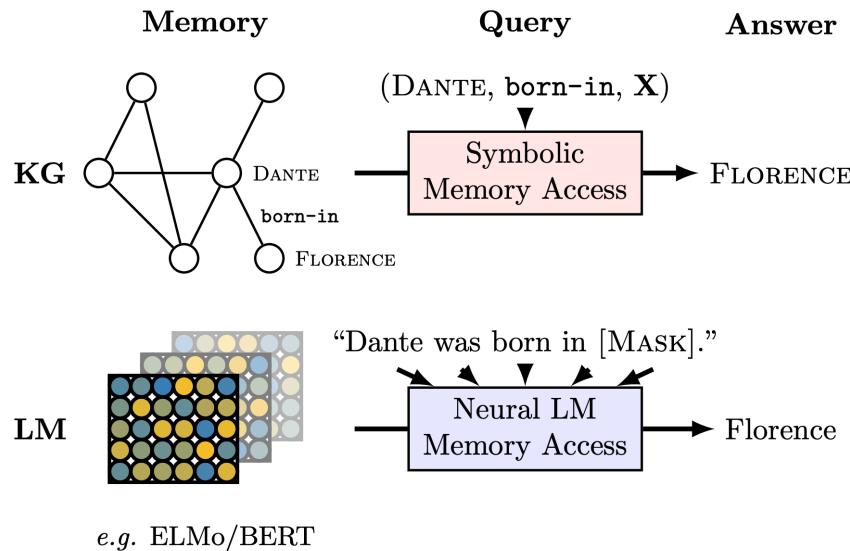
Combine with:
- Representation
- Alignment

Most useful for:
- Copying/storing
- Long-range interactions
- Controlling internal information

[Liang et al., Foundations and Recent Trends in Multimodal Machine Learning: Principles, Challenges, and Open Questions. arXiv 2022]

Open Challenges

1. Long-range multimodal sequences: good benchmarks with interactions across a long range.
2. To what extent do pre-trained models already capture memory (i.e., memorize and enable retrieval), vs explicit memory mechanisms?
3. More, see <https://cmu-multicomp-lab.github.io/adv-mmml-course/spring2022/schedule/>



2/25 Week 6: Memory and long-term interactions [synopsis]

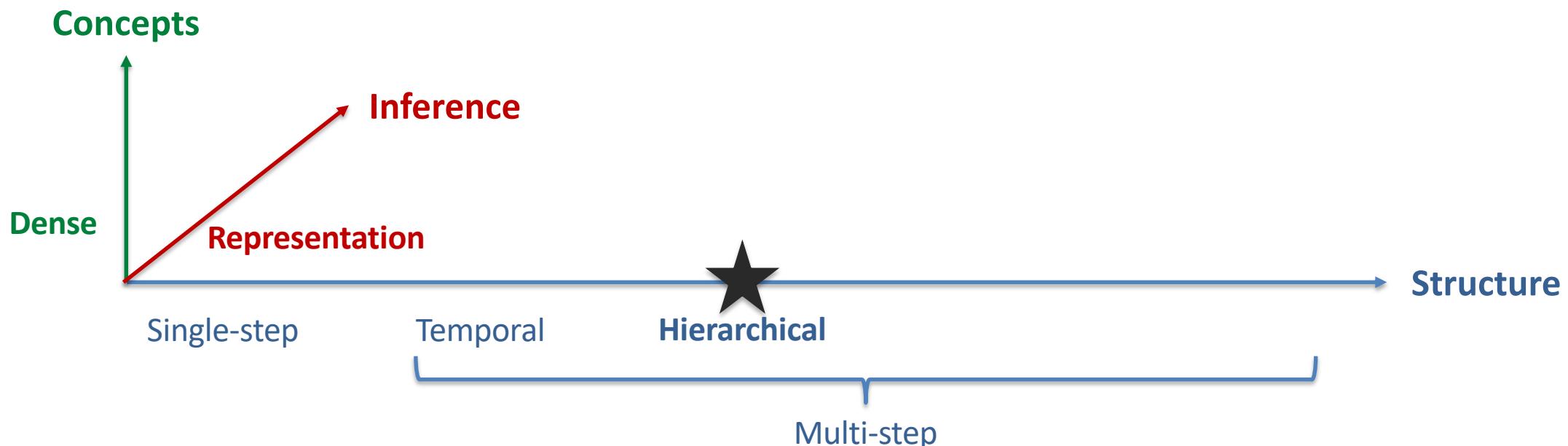
- What are the scenarios in which memory for long-term interactions is required in multimodal tasks, where data comes from heterogeneous sources? What could be a taxonomy of long-range cross-modal interactions that may need to be stored in memory?
- What are certain methods of parametrizing memory in unimodal models that may be applied for multimodal settings, and the various strengths/weaknesses of each approach?
- How should we model long-term cross-modal interactions? How can we design models (perhaps with memory mechanisms) to ensure that these long-term cross-modal interactions are captured?
- What are the main advantages of explicitly building memory-based modules into our architectures, as compared to the large-scale pre-training methods/Transformer models discussed in week 4? Do Transformer models already capture memory and long-term interactions implicitly?
- To what extent do we need external knowledge when performing reasoning, specifically multimodal reasoning? What type of external knowledge is likely to be needed to succeed in multimodal reasoning?
- A related topic is multimodal summarization: how to summarize the main events from a long multimodal sequence. How can we summarize long sequences while keeping cross-modal interactions? What is unique about multimodal summarization?

- Long Range Arena: A Benchmark for Efficient Transformers
- Large Memory Layers with Product Keys
- Dynamic Memory Networks for Visual and Textual Question Answering
- Multimodal Memory Modelling for Video Captioning
- Episodic Memory in Lifelong Language Learning
- ICON: Interactive Conversational Memory Network for Multimodal Emotion Detection
- Hybrid computing using a neural network with dynamic external memory
- History Aware Multimodal Transformer for Vision-and-Language Navigation
- Do Transformers Need Deep Long-Range Memory?
- Transformer-XL: Attentive Language Models Beyond a Fixed-Length Context
- Neural Turing Machines
- Meta-Learning with Memory-Augmented Neural Networks

[Petroni et al., Language Models as Knowledge Bases? EMNLP 2019]

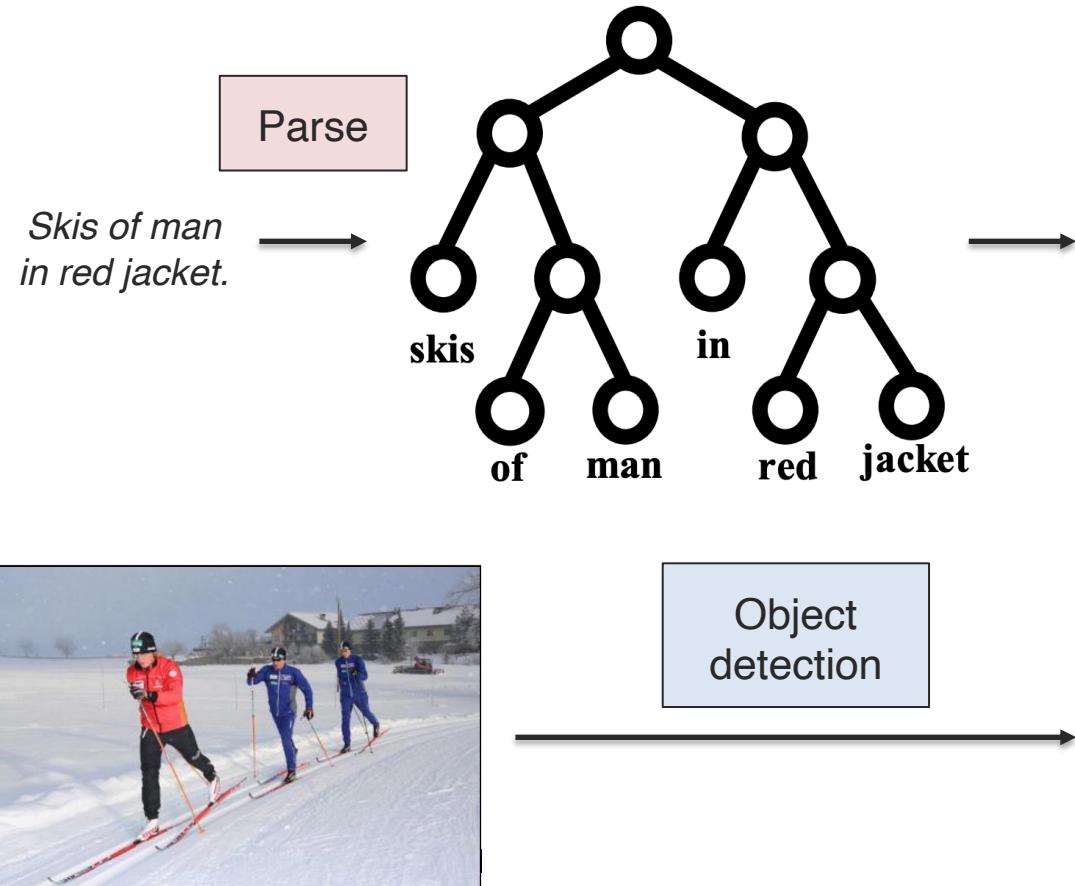
[Liang et al., Foundations and Recent Trends in Multimodal Machine Learning: Principles, Challenges, and Open Questions. arXiv 2022]

Sub-Challenge 3a: Structure Modeling



Hierarchical Structure

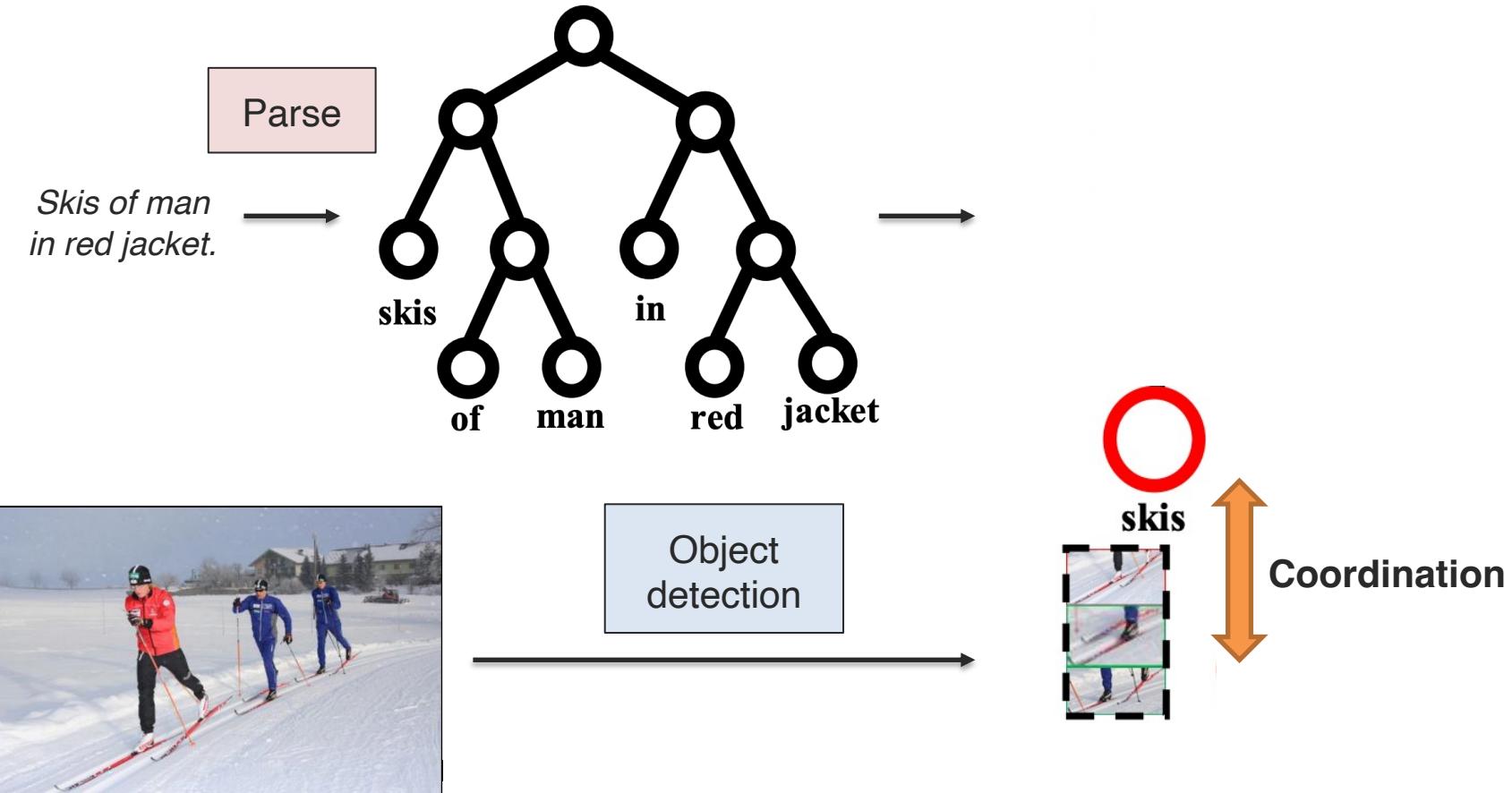
Leverage syntactic structure of language



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

Hierarchical Structure

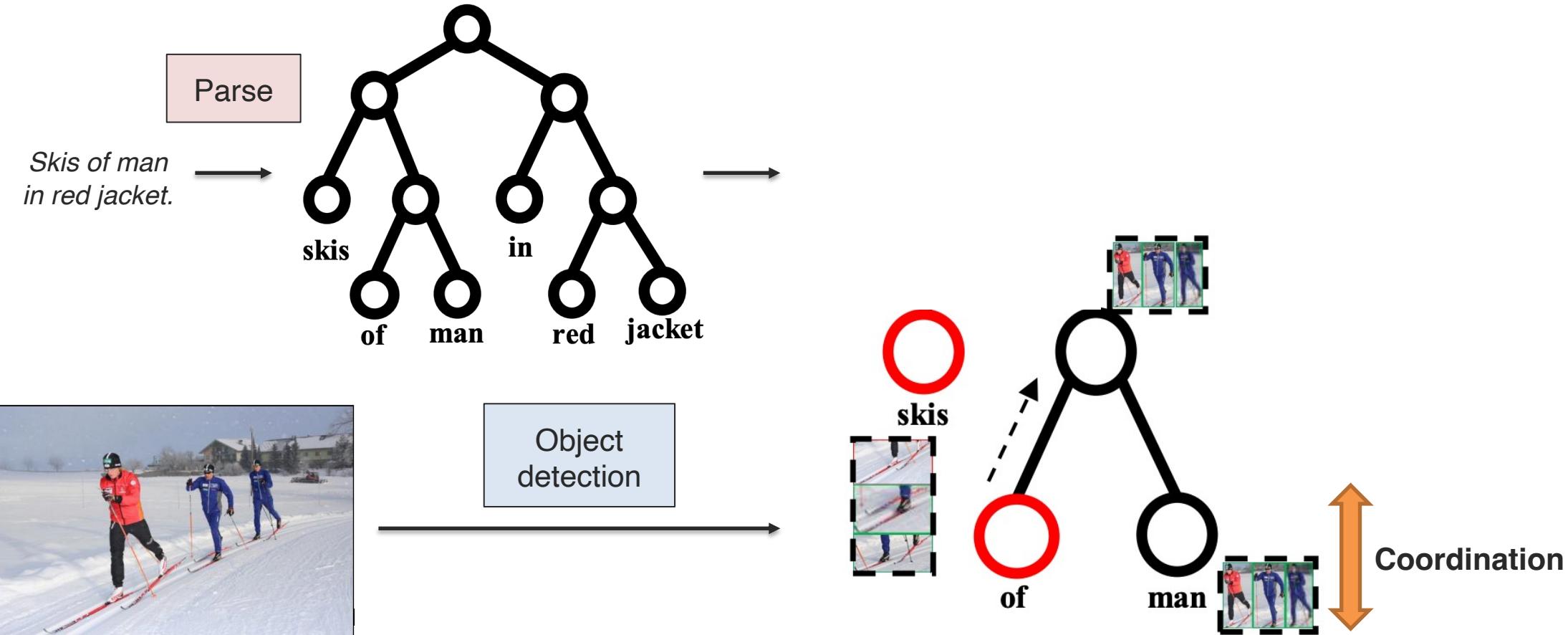
Leverage syntactic structure of language



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

Hierarchical Structure

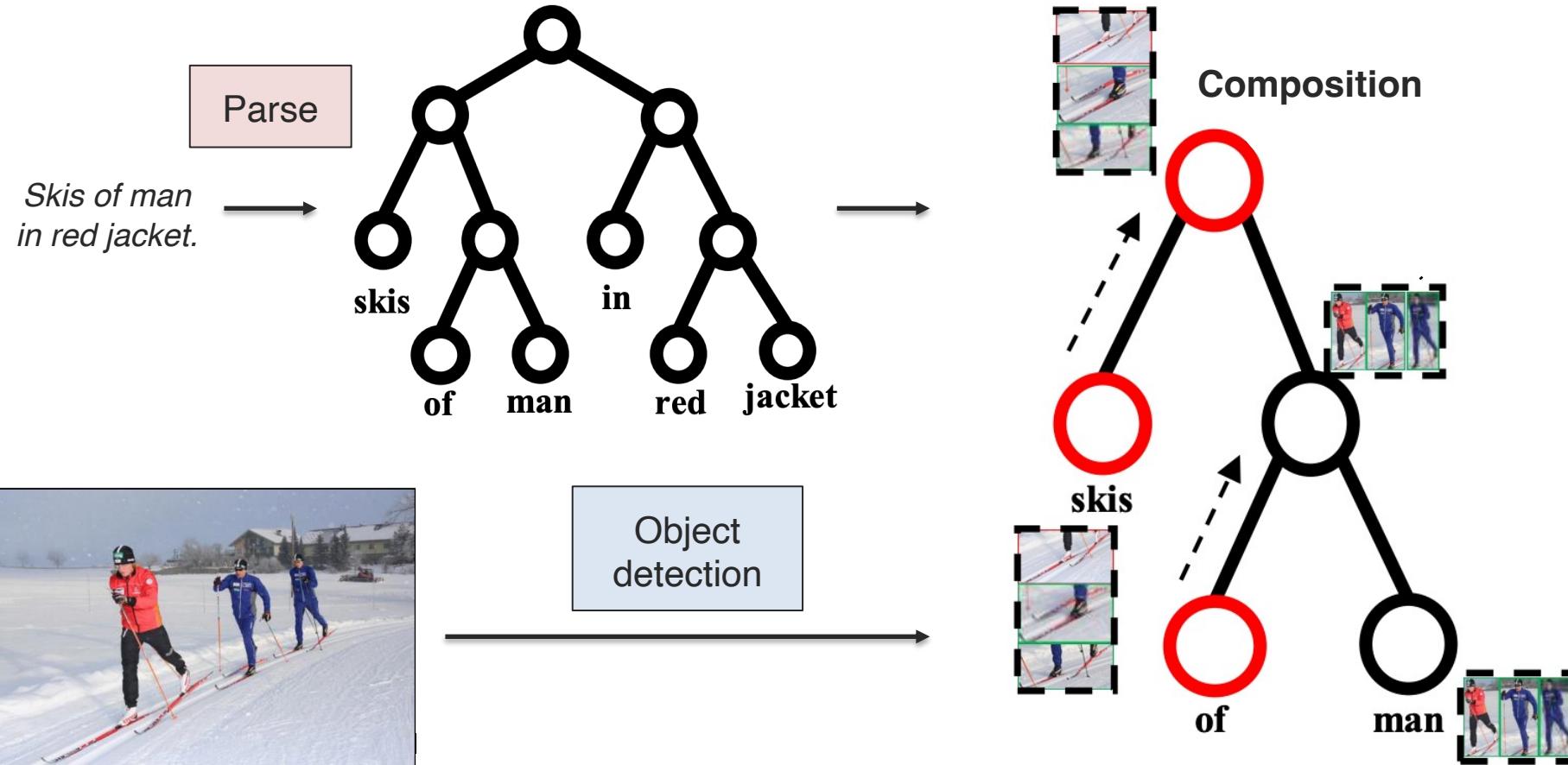
Leverage syntactic structure of language



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

Hierarchical Structure

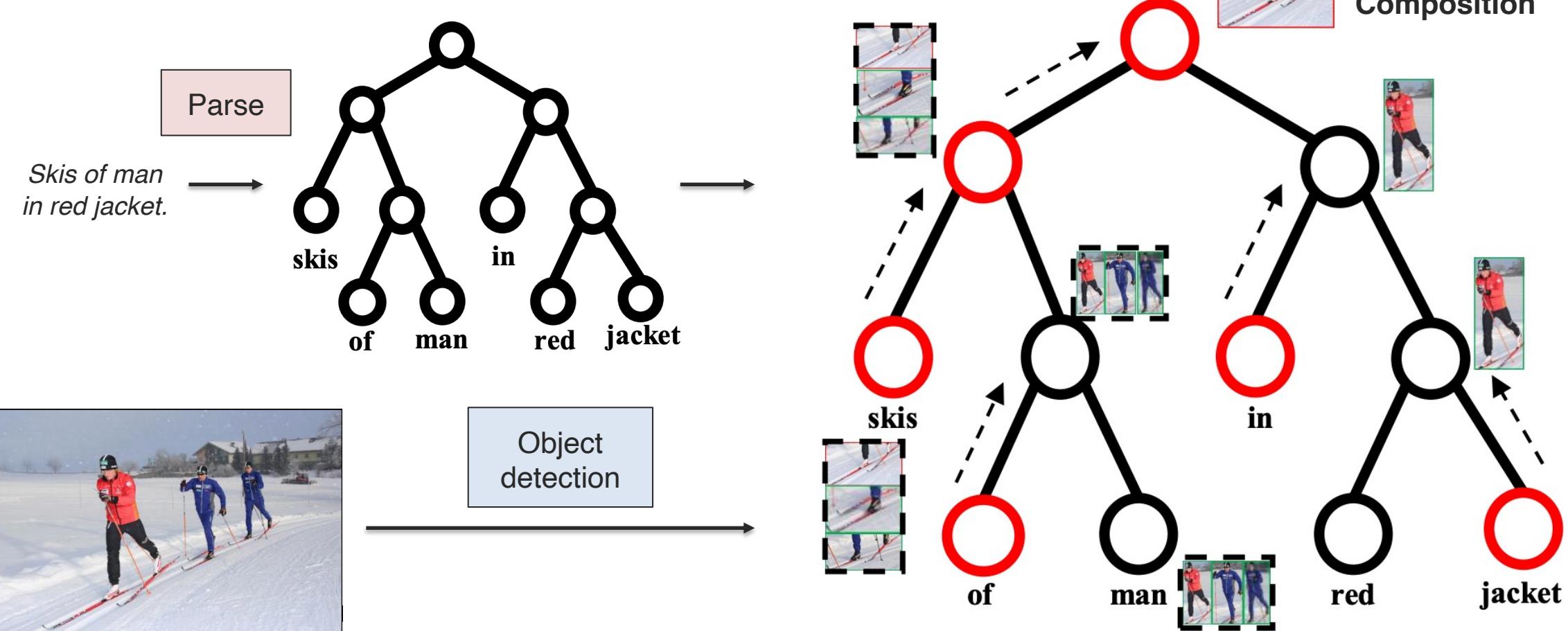
Leverage syntactic structure of language



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

Hierarchical Structure

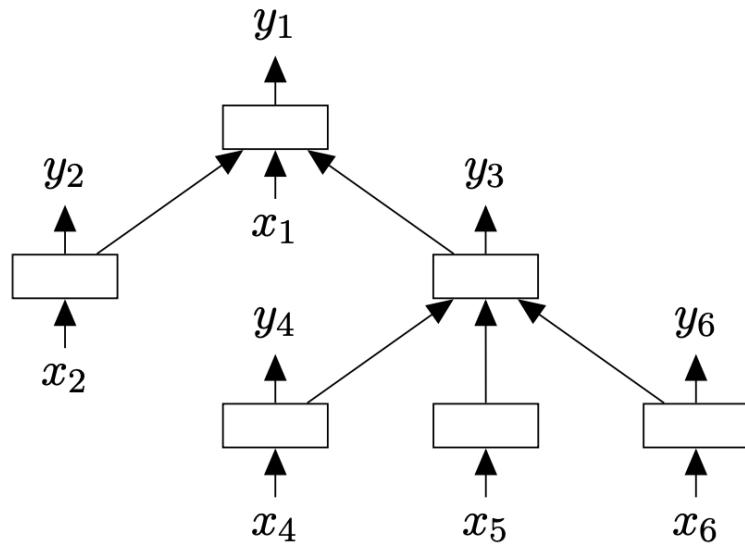
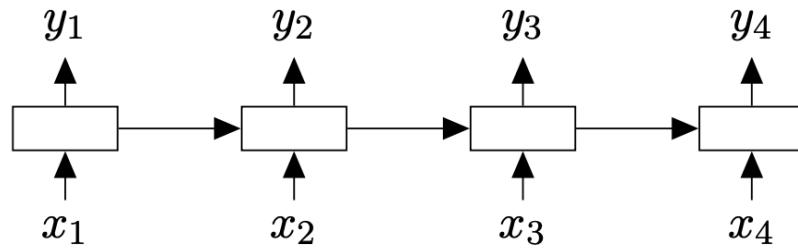
Leverage syntactic structure of language



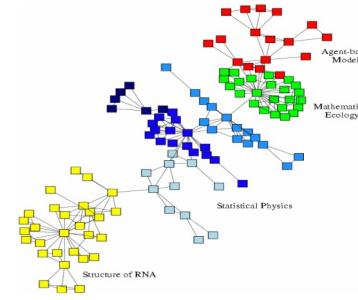
[Hong et al., Learning to Compose and Reason with Language Tree Structures for Visual Grounding. IEEE TPAMI 2019]

Tree and Graph Networks

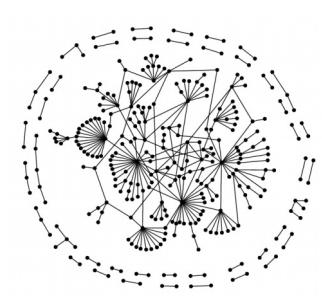
From linear chain models to tree and graph-structured models



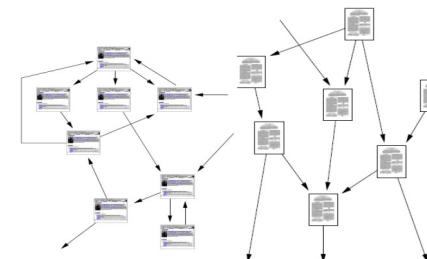
Social networks



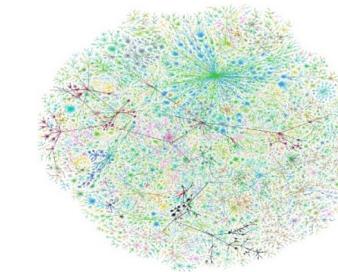
Economic networks



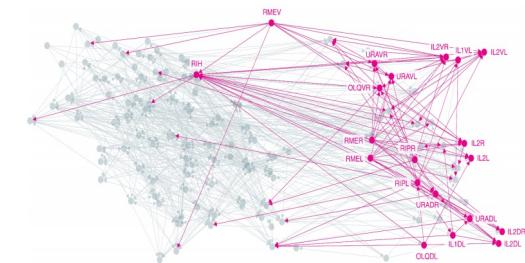
Biomedical networks



Information networks: Web & citations



Interne

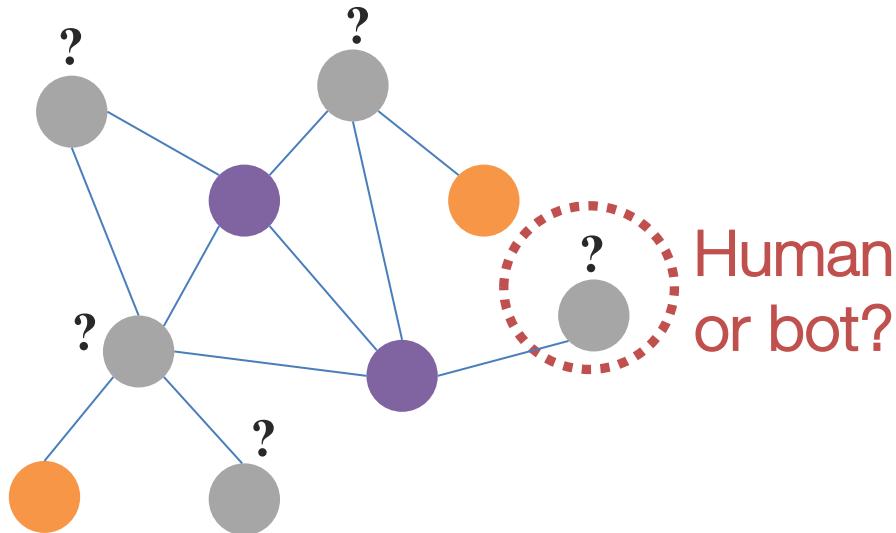


Networks of neurons

[Leskovec. Representation Learning on Networks. WWW 2018; Hamilton and Tang, Tutorial on Graph Representation Learning. AAAI 2019]

Graphs – Supervised Task

Goal: Learn from labels associated with a subset of nodes (or with all nodes)



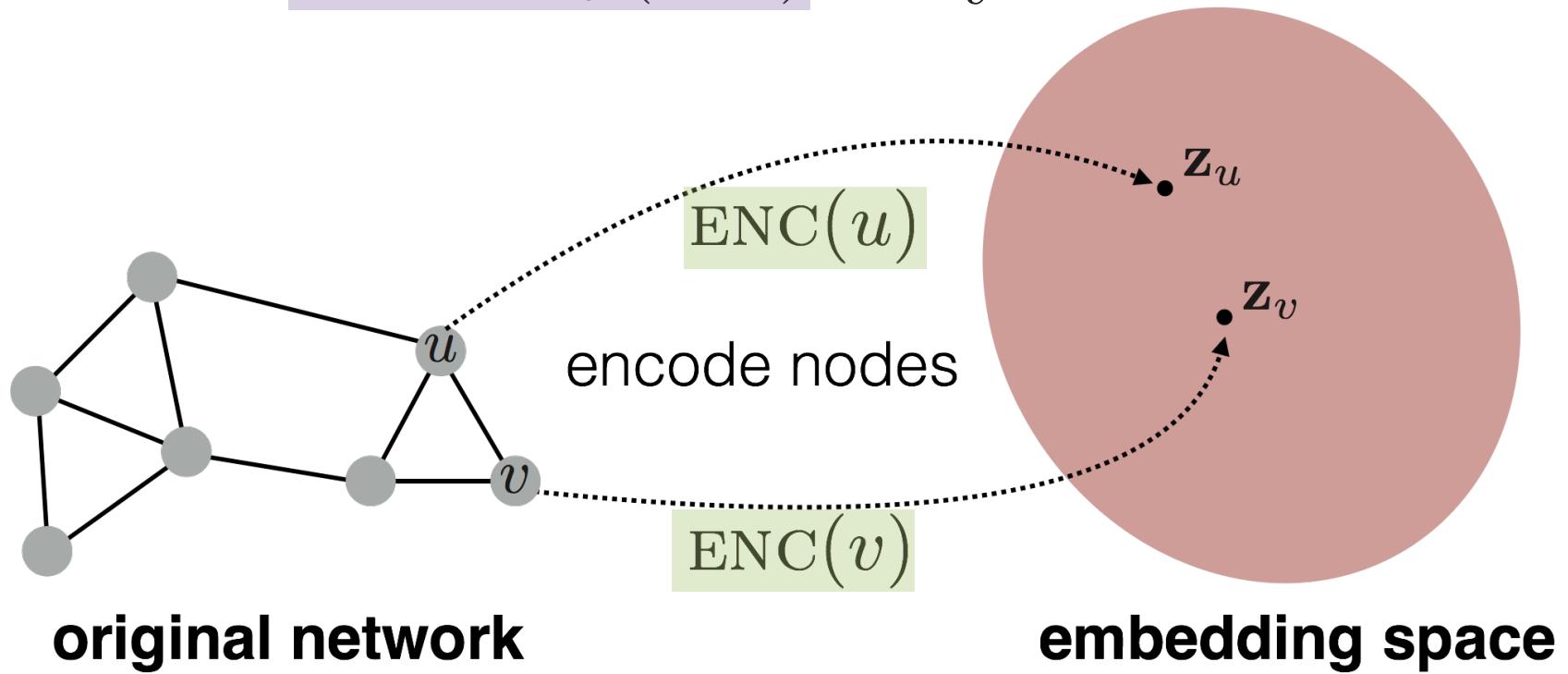
e.g., an online social network

[Leskovec. Representation Learning on Networks. WWW 2018; Hamilton and Tang, Tutorial on Graph Representation Learning. AAAI 2019]

Graphs – Unsupervised Task

Goal: Learn an embedding space where

$$\text{similarity}(u, v) \approx \mathbf{z}_v^\top \mathbf{z}_u$$



[Leskovec. Representation Learning on Networks. WWW 2018; Hamilton and Tang, Tutorial on Graph Representation Learning. AAAI 2019]

Graph Neural Nets

Assume we have a graph \mathbf{G} :

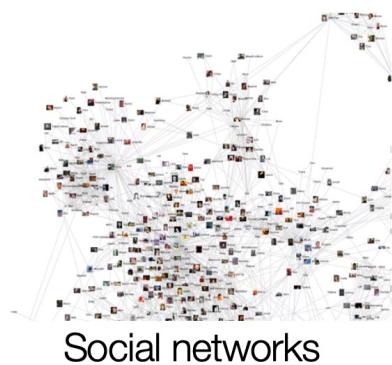
\mathbf{V} is the set of vertices

\mathbf{A} is the binary adjacency matrix

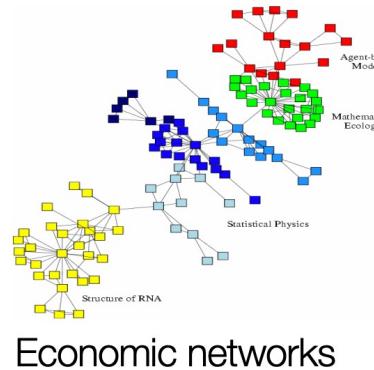
\mathbf{X} is a matrix of node features:

- Categorical attributes, text, image data
e.g. profile information in a social network
- ...

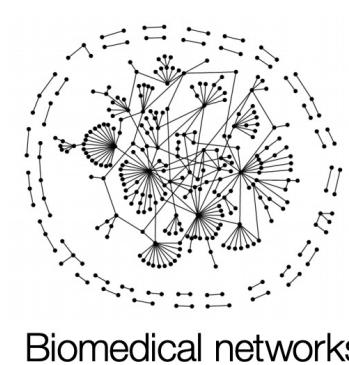
\mathbf{Y} is a vector of node labels (optional)



Social networks



Economic networks

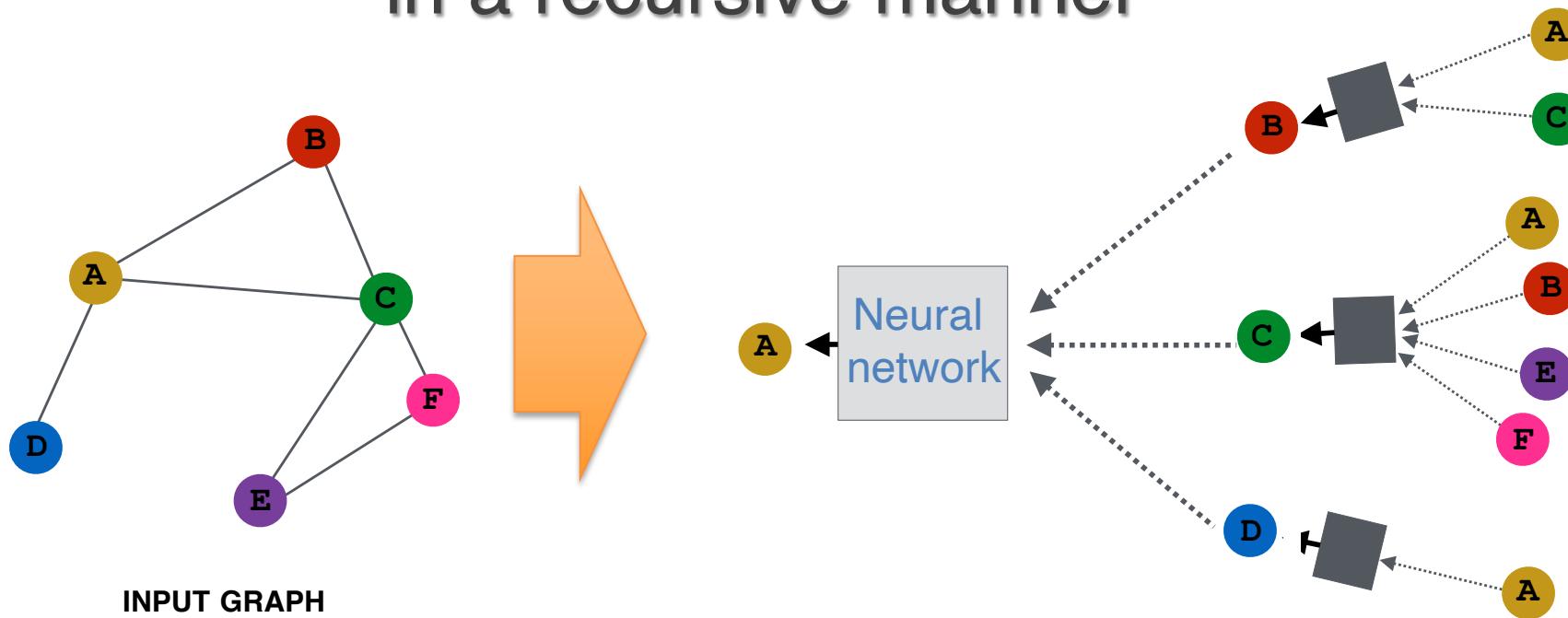


Biomedical networks

[Leskovec. Representation Learning on Networks. WWW 2018; Hamilton and Tang, Tutorial on Graph Representation Learning. AAAI 2019]

Graph Neural Nets

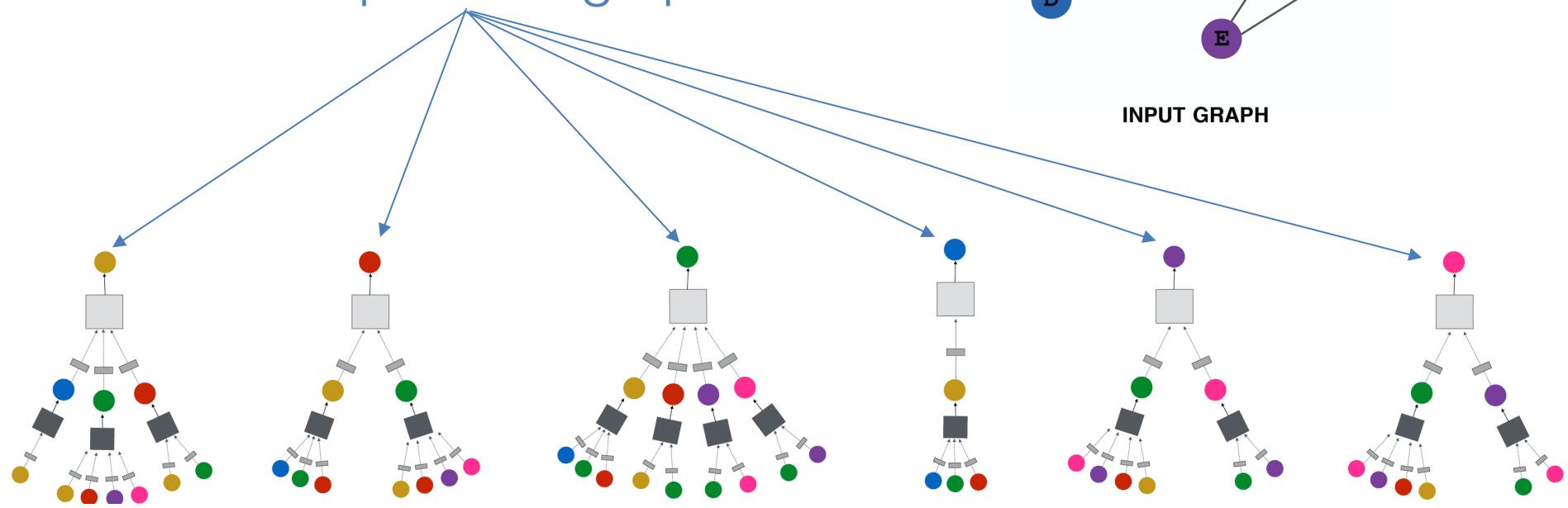
Key idea: Generate node embeddings based on local neighborhoods in a recursive manner



[Leskovec. Representation Learning on Networks. WWW 2018; Hamilton and Tang, Tutorial on Graph Representation Learning. AAAI 2019]

Graph Neural Nets

Every node defines a unique computation graph!



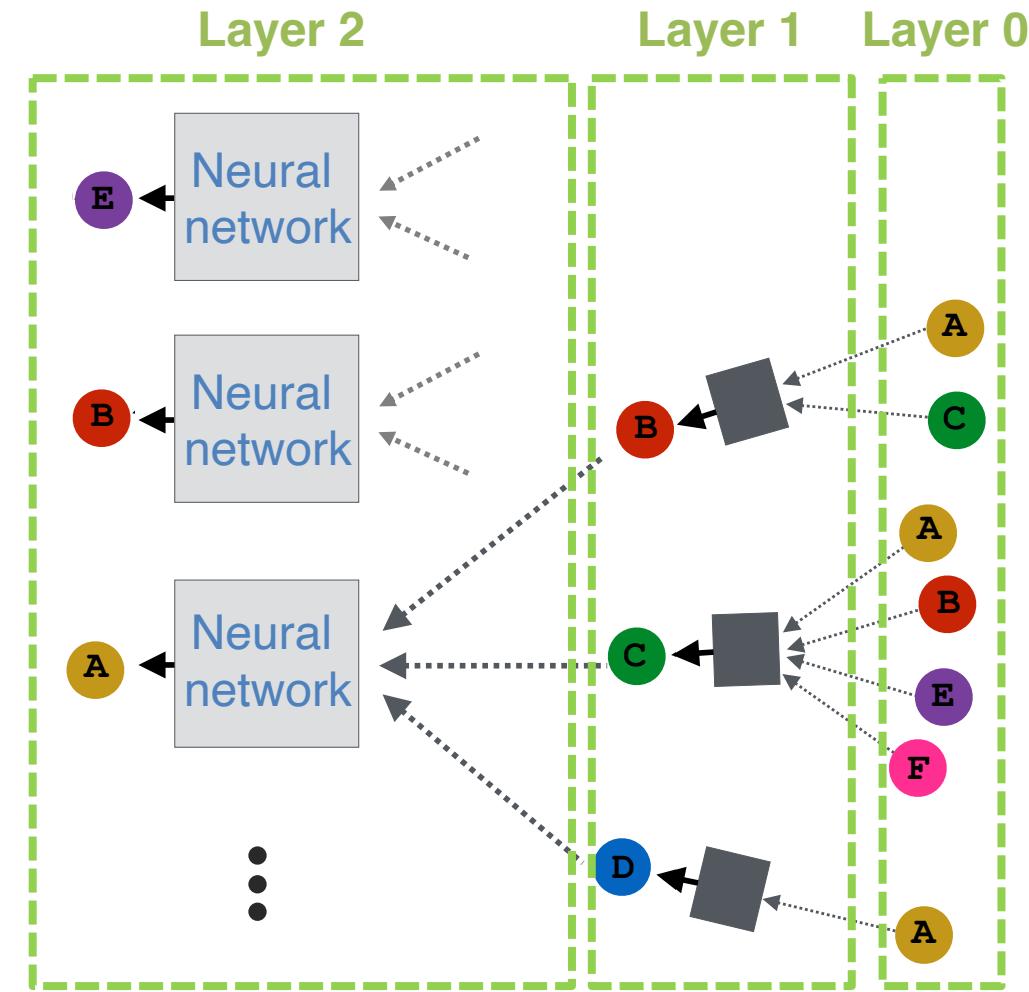
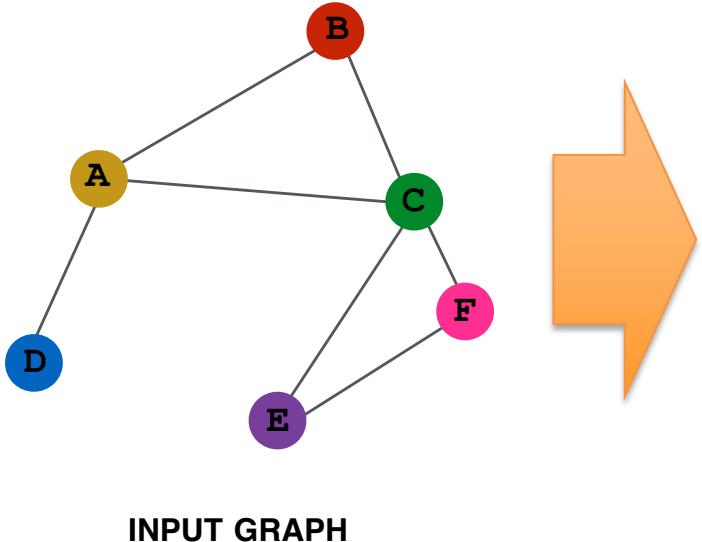
INPUT GRAPH

[Leskovec. Representation Learning on Networks. WWW 2018; Hamilton and Tang, Tutorial on Graph Representation Learning. AAAI 2019]

Graph Neural Nets

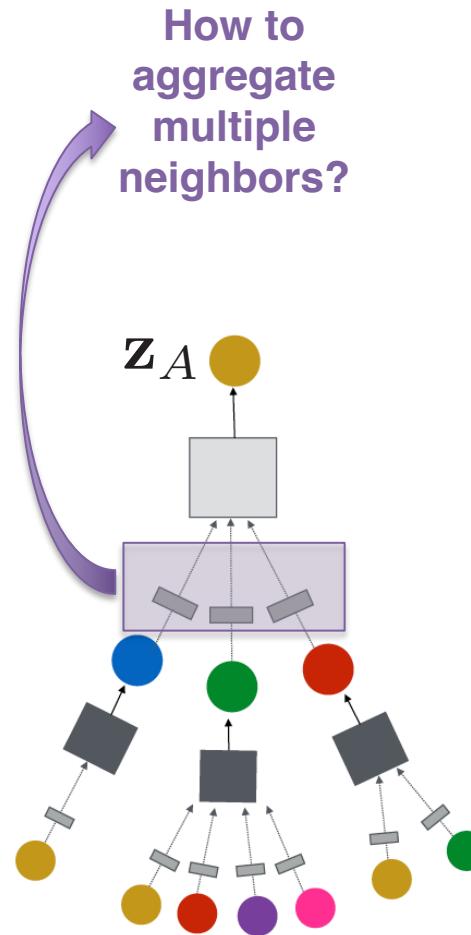
And multiple layers!

- Shared parameters within a specific layer
- “layer-0” is the input feature x_u



[Leskovec. Representation Learning on Networks. WWW 2018; Hamilton and Tang, Tutorial on Graph Representation Learning. AAAI 2019]

Graph Neural Nets – Neighborhood Aggregation



Average pooling

(Scarselli et al., 2005)

$$\mathbf{h}_v^k = \sigma \left(\mathbf{W}_k \sum_{u \in N(v)} \frac{\mathbf{h}_u^{k-1}}{|N(v)|} + \mathbf{B}_k \mathbf{h}_v^{k-1} \right)$$

Different weights
for neighbors
and self

K is num layers

Graph Convolution Network

(Kipf et al., 2017)

Same weights

$$\mathbf{h}_v^k = \sigma \left(\mathbf{W}_k \sum_{u \in N(v) \cup v} \frac{\mathbf{h}_u^{k-1}}{\sqrt{|N(u)||N(v)|}} \right)$$

Different
normalization

It can be efficiently implemented

Graph Attention Network

(Velickovic et al., 2018)

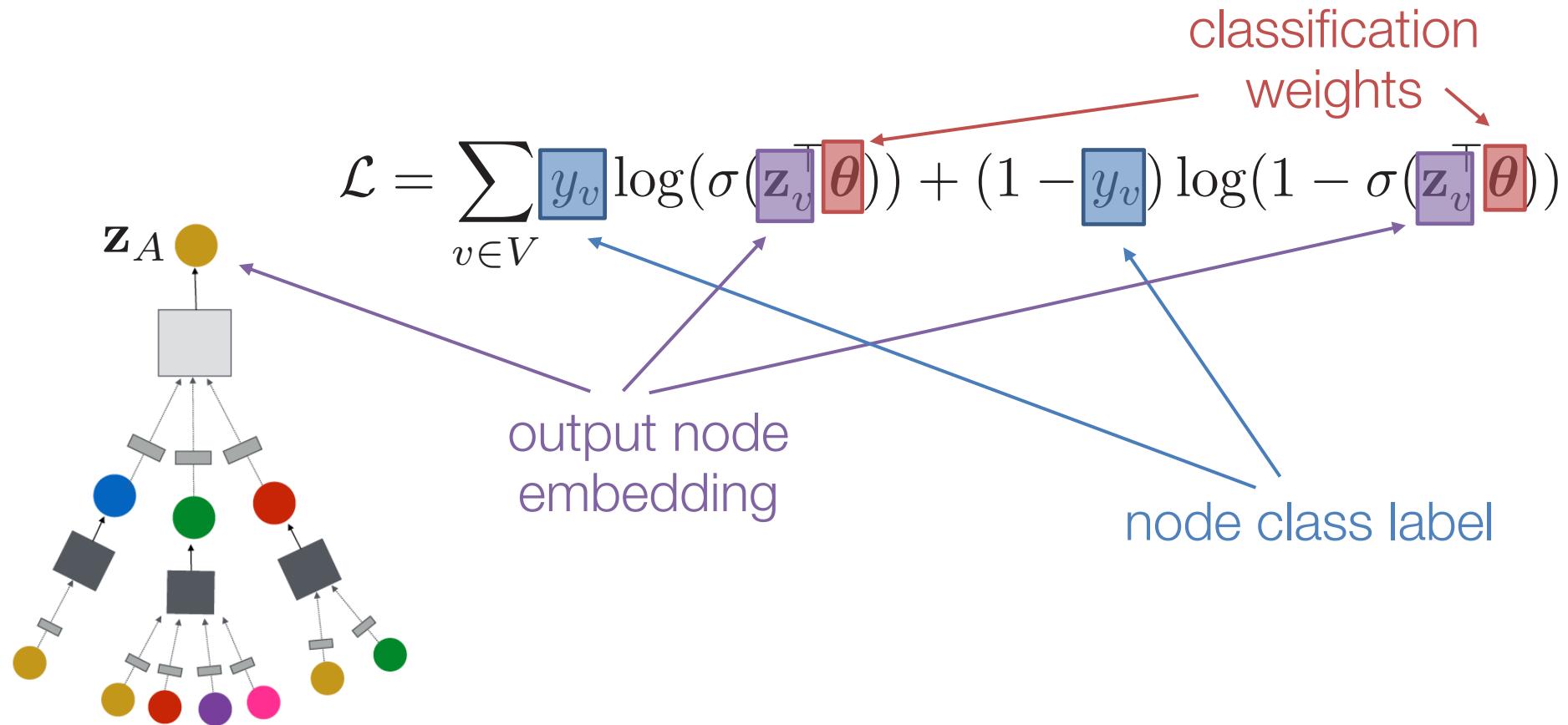
$$\mathbf{h}_v^k = \sigma \left(\mathbf{W}_k \sum_{u \in N(v) \cup v} \frac{\alpha_{uv} \mathbf{h}_u^{k-1}}{\sqrt{|N(u)||N(v)|}} \right)$$

Attention
weights

Very similar to a self-attention transformer

[Leskovec. Representation Learning on Networks. WWW 2018; Hamilton and Tang, Tutorial on Graph Representation Learning. AAAI 2019]

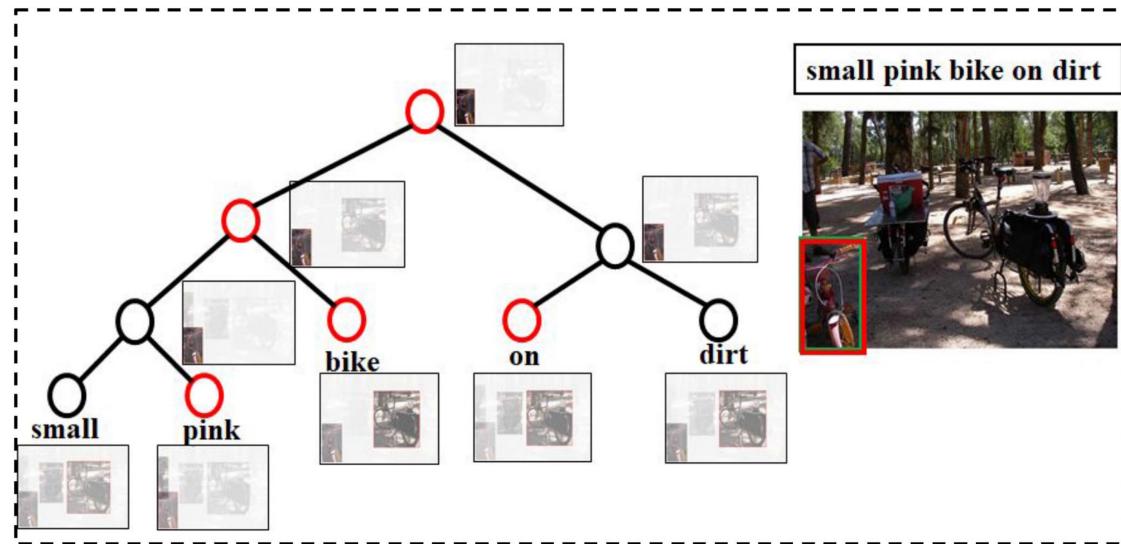
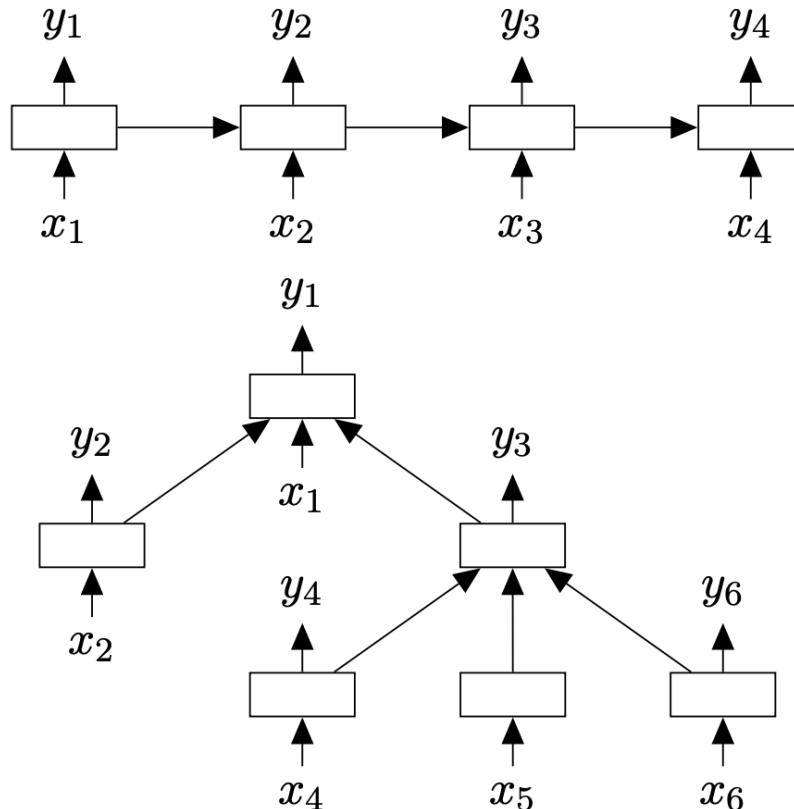
Graph Neural Nets – Supervised Training



[Leskovec. Representation Learning on Networks. WWW 2018; Hamilton and Tang, Tutorial on Graph Representation Learning. AAAI 2019]

Experiments

From linear chain models to tree models



RefCOCO

Accounting for syntactic structure also improves language-based sentiment analysis, semantic matching, question-answering, language modeling, interpreting attention scores, etc.

[Tai et al., Improved Semantic Representations From Tree-Structured Long Short-Term Memory Networks. ACL 2015]

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

[Wang et al., Tree Transformer: Integrating Tree Structures into Self-Attention. EMNLP 2019]

How do Graph Nets Work?

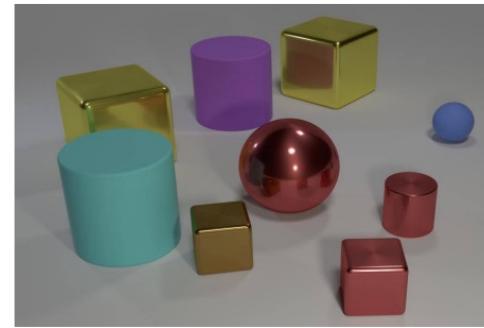
Empirically graph nets work well over less structured networks, but why?

Key idea: algorithmic alignment - link compositional structure required for task with computational structure of prediction model



Summary statistics

What is the maximum value difference among treasures?



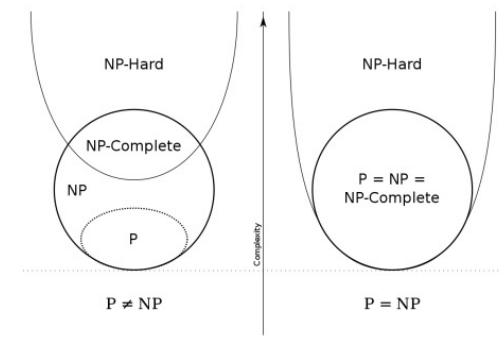
Relational argmax

What are the colors of the furthest pair of objects?



Dynamic programming

What is the cost to defeat monster X by following the optimal path?



NP-hard problem

Subset sum: Is there a subset that sums to 0?

How do Graph Nets Work?

Empirically graph nets work well over less structured networks, but why?

Key idea: algorithmic alignment - link compositional structure required for task with computational structure of prediction model

MLP $y = \text{MLP}_1(X)$

DeepSets $y = \text{MLP}_2\left(\sum_{s \in S} \text{MLP}_1(X_s)\right).$

K -layer GNN $h_s^{(k)} = \sum_{t \in S} \text{MLP}_1^{(k)}\left(h_s^{(k-1)}, h_t^{(k-1)}\right), \quad h_S = \text{MLP}_2\left(\sum_{s \in S} h_s^{(K)}\right),$

[Xu et al., What Can Neural Networks Reason About?. ICLR 2020]

How do Graph Nets Work?

Empirically graph nets work well over less structured networks, but why?

Key idea: algorithmic alignment - link compositional structure required for task with computational structure of prediction model

Many multimodal reasoning problems here:
intuitive physics, visual question answering, shortest paths



Summary statistics

What is the maximum value difference among treasures?

DeepSets



Relational argmax

What are the colors of the furthest pair of objects?

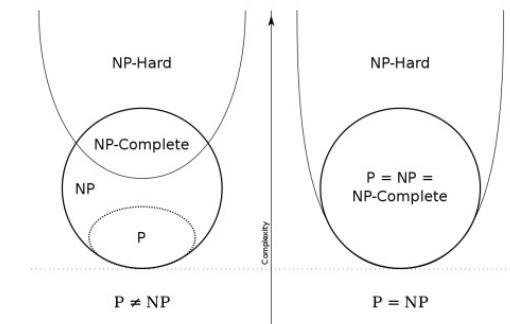
1-layer GNN



Dynamic programming

What is the cost to defeat monster X by following the optimal path?

K-layer GNN



NP-hard problem

Subset sum: Is there a subset that sums to 0?

None ☹

How do Graph Nets Work?

Empirically graph nets work well over less structured networks, but why?

Key idea: algorithmic alignment - link compositional structure required for task with computational structure of prediction model

How graph neural nets capture dynamic programming:

$$\text{distance}[1][u] = \text{cost}(s, u), \quad \text{distance}[k][u] = \min_v \{ \text{distance}[k-1][v] + \text{cost}(v, u) \},$$

Graph Neural Network

for $k = 1 \dots \text{GNN iter:}$

for u **in** $S:$

No need to learn for-loops

$$h_u^{(k)} = \sum_v \text{MLP}(h_v^{(k-1)}, h_u^{(k-1)})$$

Bellman-Ford algorithm

for $k = 1 \dots |\mathcal{S}| - 1:$

for u **in** $S:$

$$d[k][u] = \min_v d[k-1][v] + \text{cost}(v, u)$$

Learns a simple reasoning step

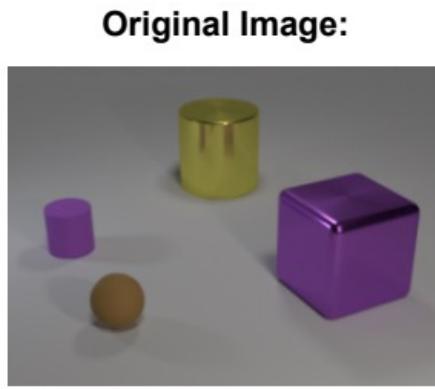
MLPs have to learn entire for loops 😞

[Xu et al., What Can Neural Networks Reason About?. ICLR 2020]

How do Graph Nets Work?

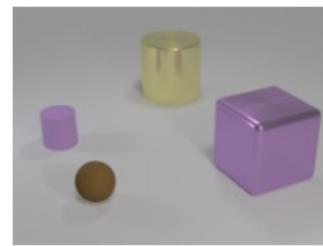
Empirically: datasets that require multiple steps of relational reasoning

1. Sudoku: number interactions, multi-step, backtracking,
2. Relational VQA: CLEVR -> Sort-of-CLEVR -> Pretty-CLEVR ('*which object is closest/k-steps away*')



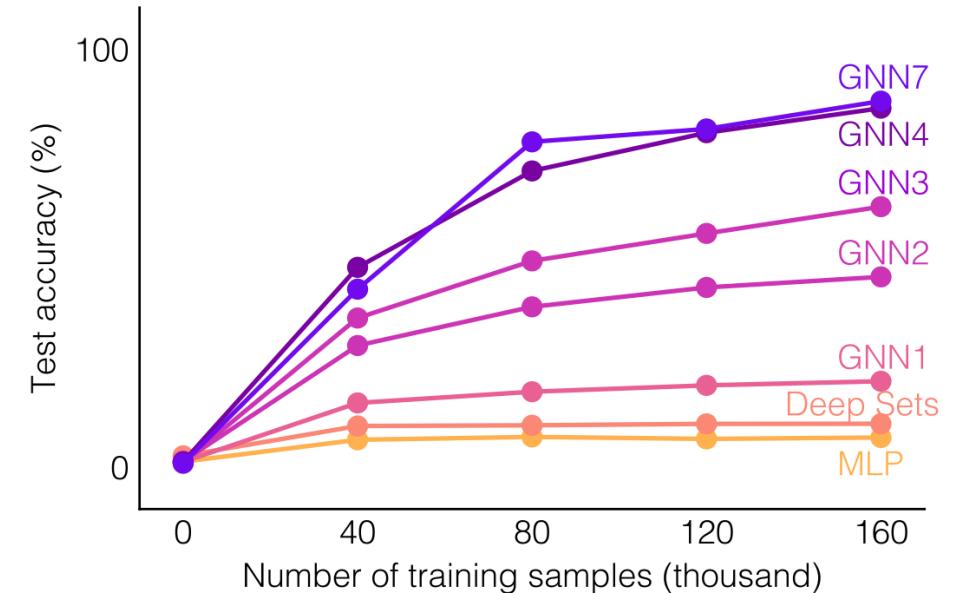
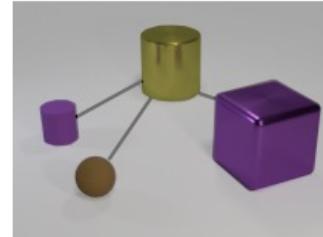
Non-relational question:

What is the size of the brown sphere?



Relational question:

Are there any rubber things that have the same size as the yellow metallic cylinder?



[Santoro et al., A Simple Neural Network Module for Relational Reasoning. NeurIPS 2017]

[Palm et al., Recurrent Relational Network. NeurIPS 2018]

[Xu et al., What Can Neural Networks Reason About?. ICLR 2020]

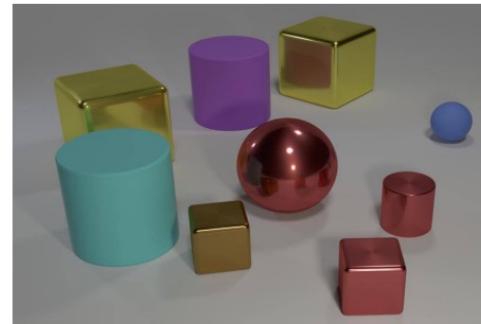
Key Takeaways & Open Challenges

1. Relations are between elements from same modality, so distances and representations are well-defined.
-> how to handle cross-modal interconnections at the same time?
2. Heterogeneous graph nets, where nodes come from different modalities.
3. Formal connections between cross-modal interactions and relational reasoning.
4. Quantifying the reasoning required by decomposing datasets into perception vs reasoning.



Summary statistics

What is the maximum value difference among treasures?



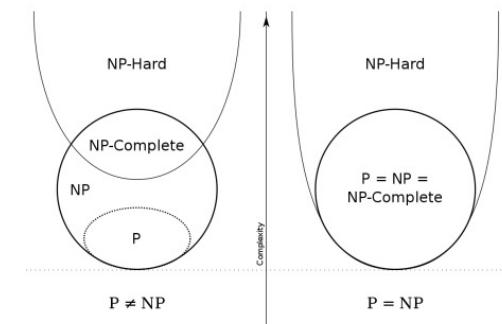
Relational argmax

What are the colors of the furthest pair of objects?



Dynamic programming

What is the cost to defeat monster X by following the optimal path?



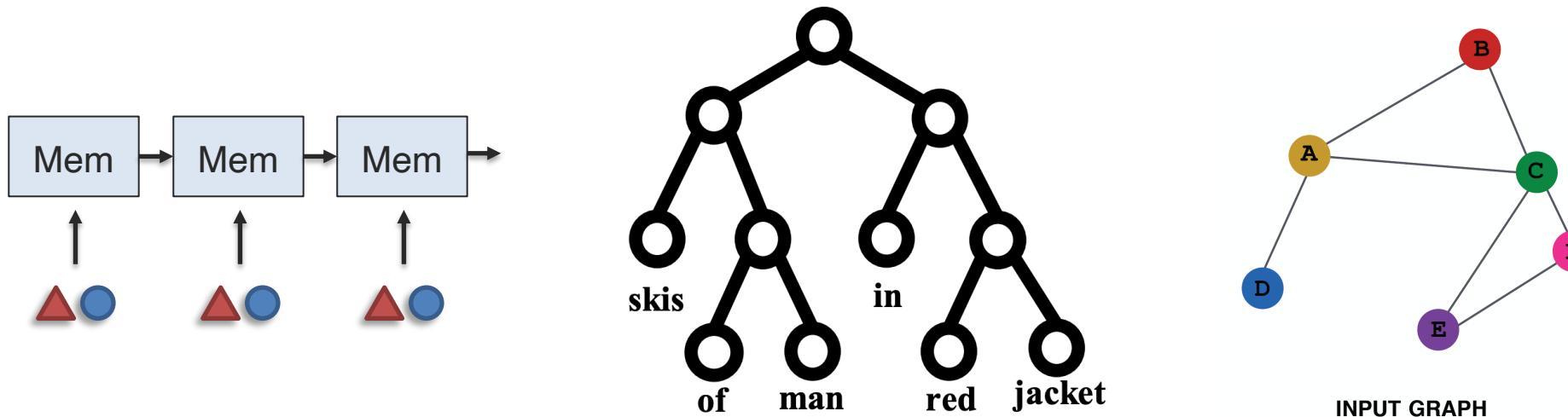
NP-hard problem

Subset sum: Is there a subset that sums to 0?

Summary

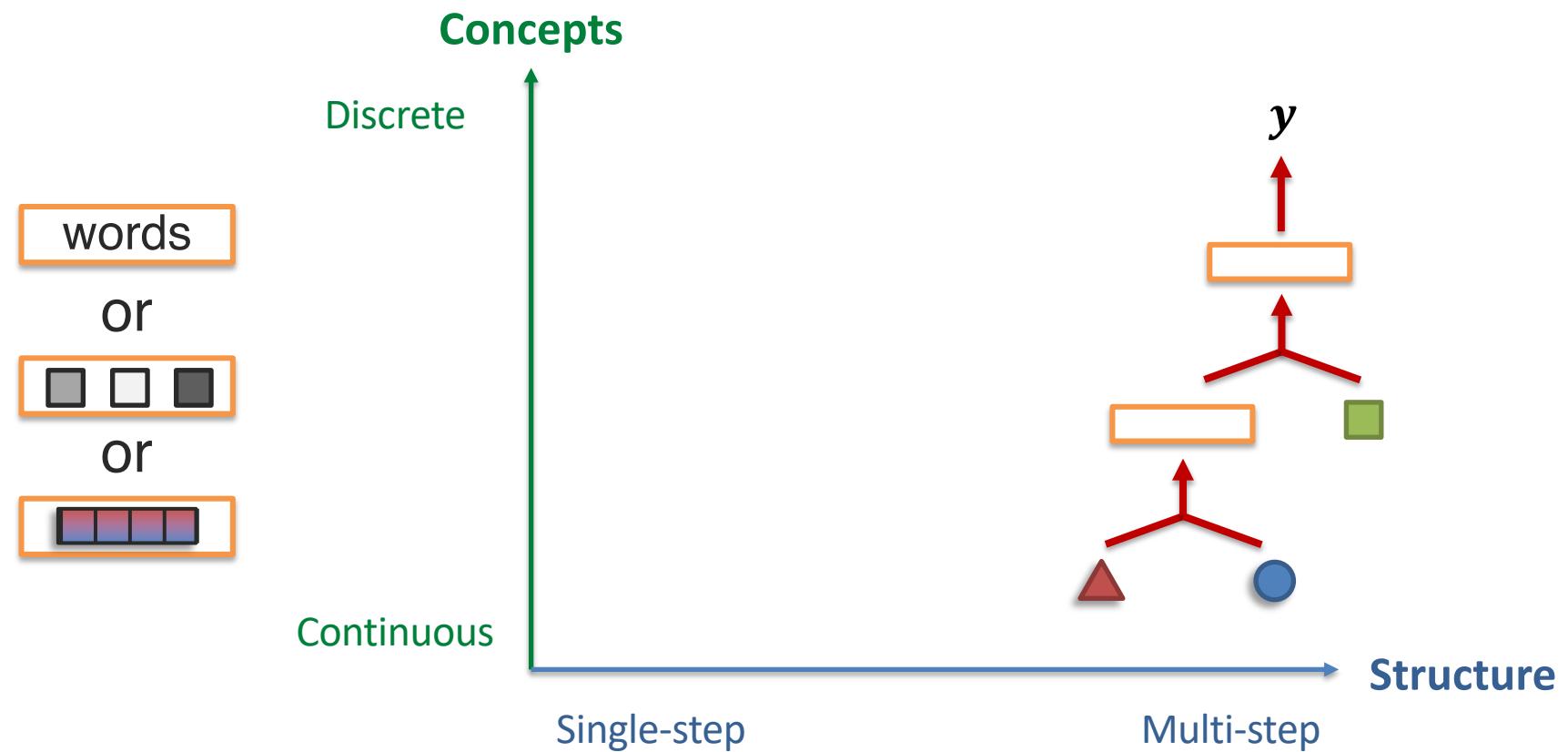
Reasoning is about compositionality, and compositionality requires knowing the structure.

In the continuous case (i.e., if structure is given or can be learned easily in a differentiable manner):



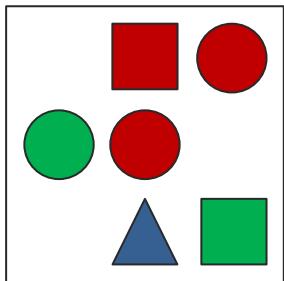
Sub-Challenge 3b: Intermediate Concepts

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



Neuro-symbolic Concepts

Hand-crafted concepts based on domain knowledge

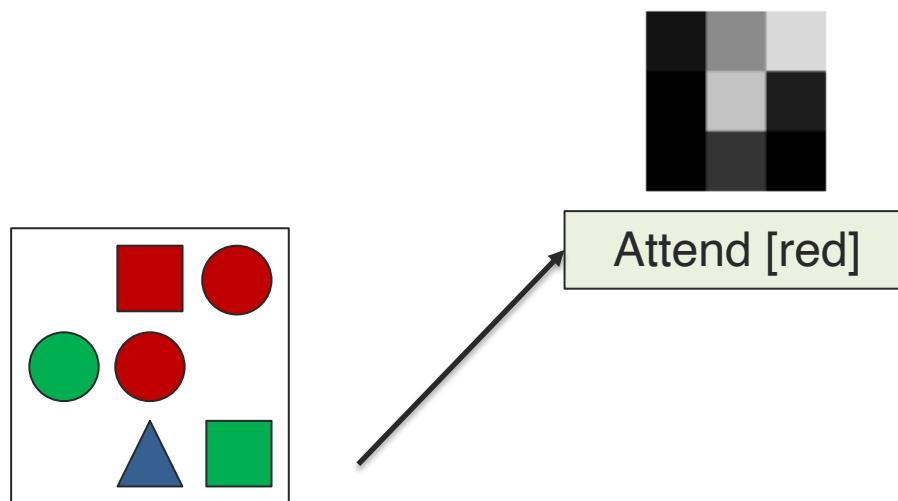


*Is there a red shape
above a circle?*

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

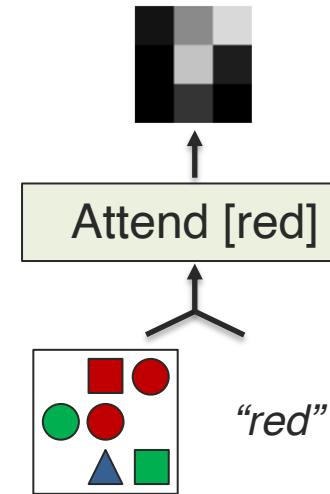
Neuro-symbolic Concepts

Hand-crafted concepts based on domain knowledge



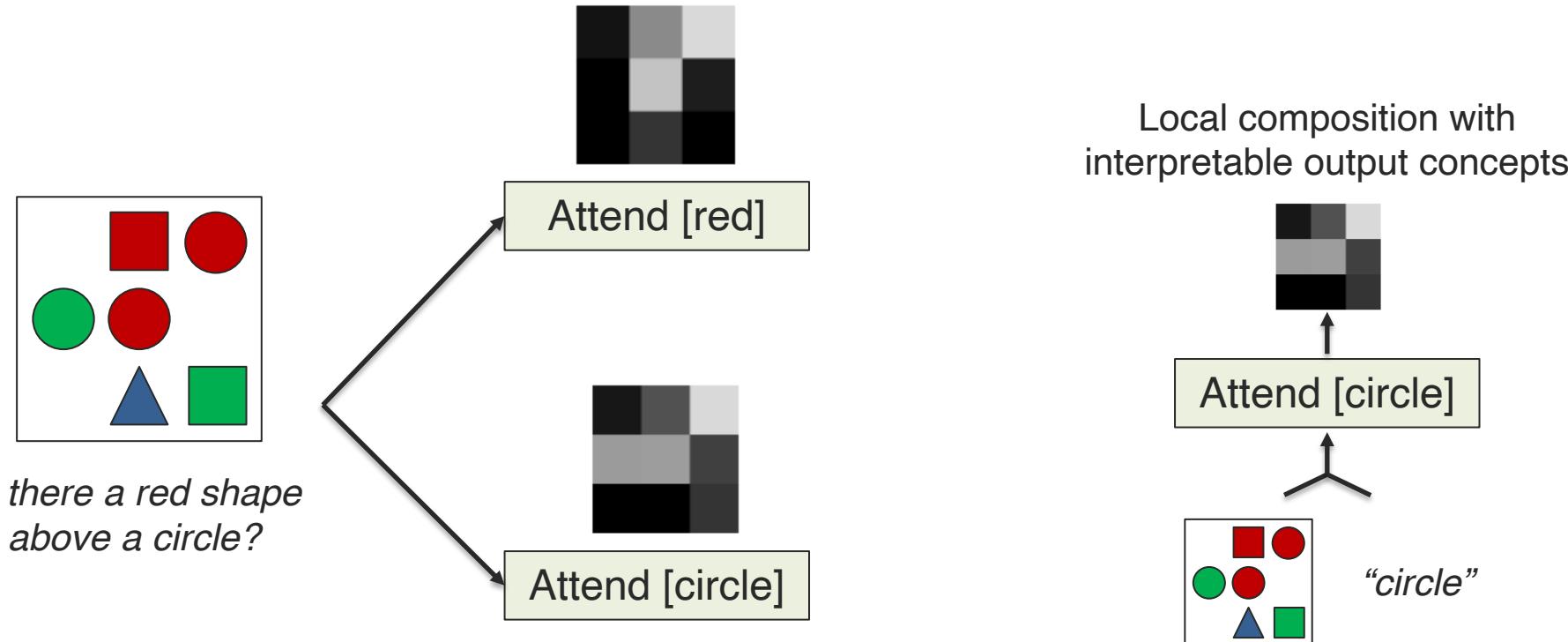
*Is there a red shape
above a circle?*

Local composition with
interpretable output concepts



Neuro-symbolic Concepts

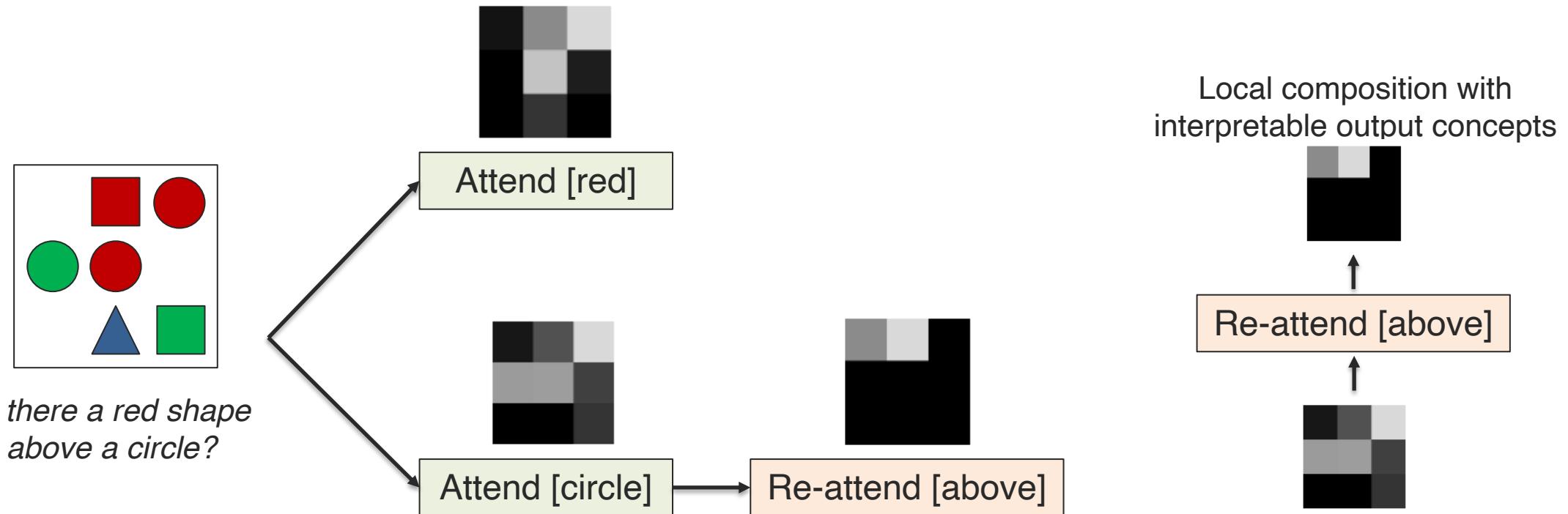
Hand-crafted concepts based on domain knowledge



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

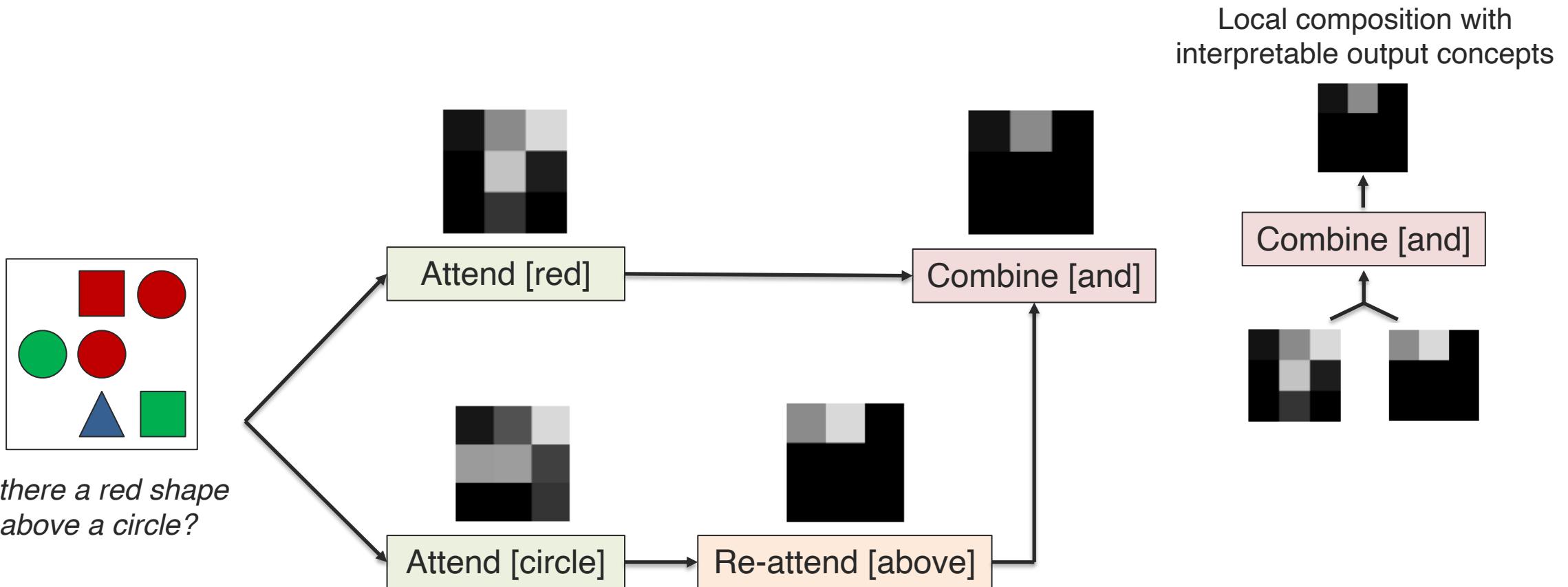
Neuro-symbolic Concepts

Hand-crafted concepts based on domain knowledge



Neuro-symbolic Concepts

Hand-crafted concepts based on domain knowledge

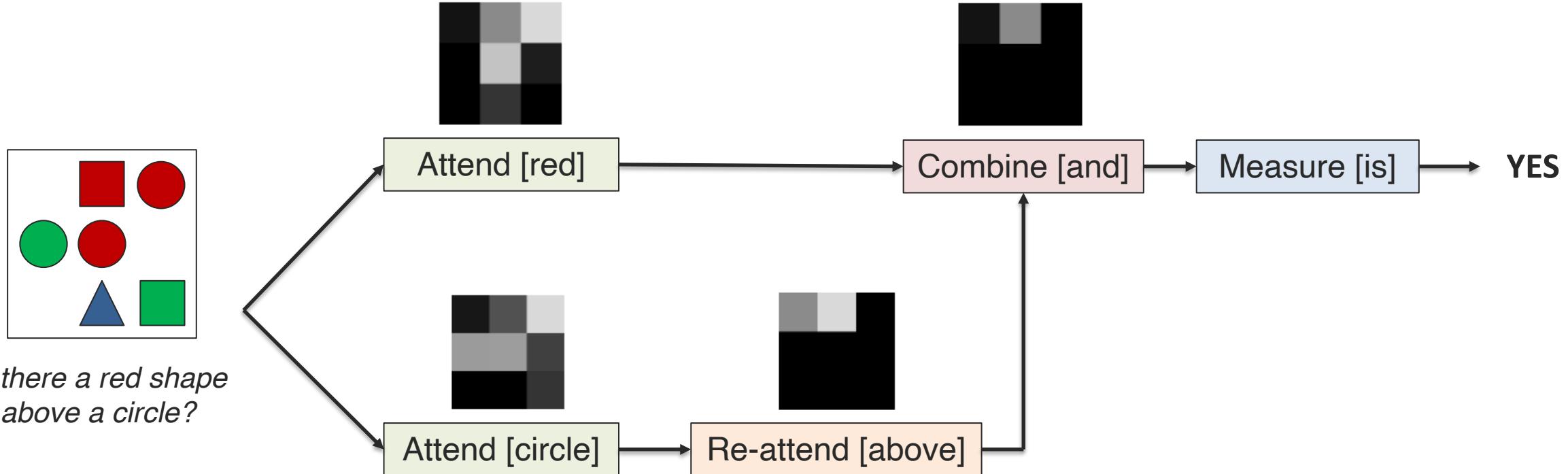


[Andreas et al., Neural Module Networks. CVPR 2016]

Neuro-symbolic Concepts

Hand-crafted concepts based on domain knowledge

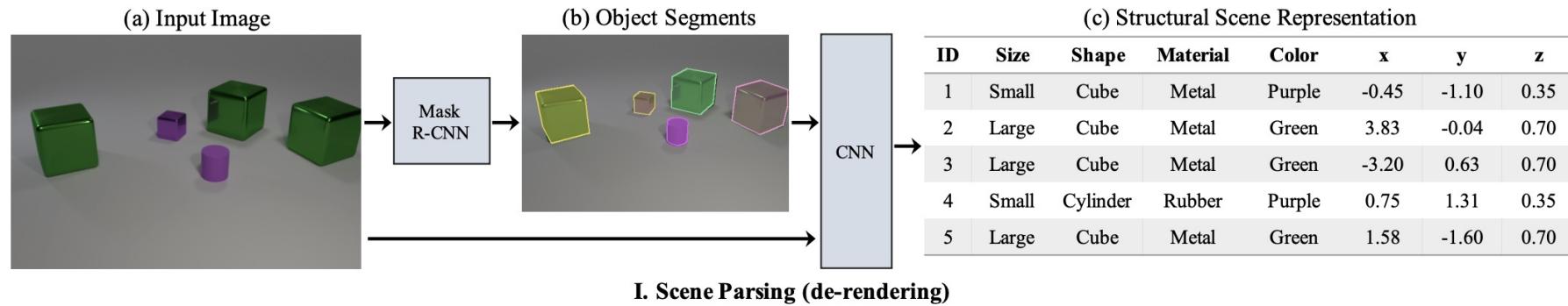
Recall structure - leverage syntactic structure of language



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

More Neuro-symbolic Concepts

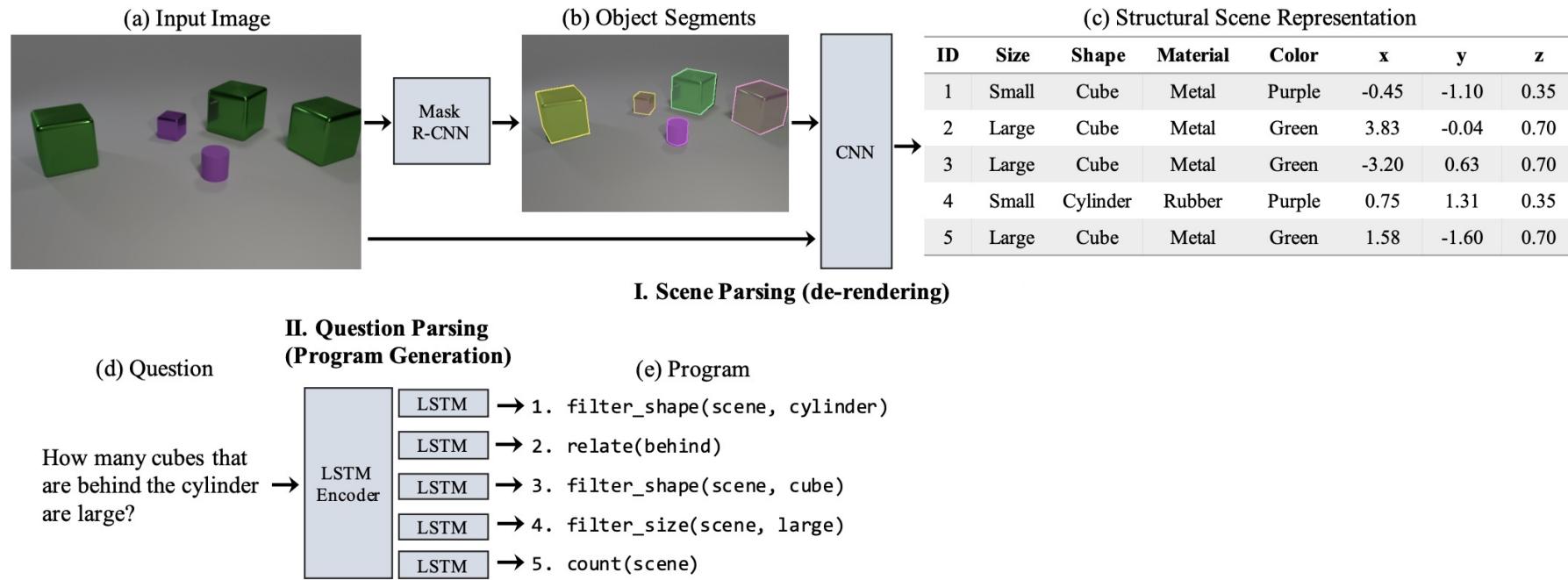
Hand-crafted concepts based on domain knowledge



[Yi et al., Neural-Symbolic VQA: Disentangling Reasoning from Vision and Language Understanding. NeurIPS 2018]

More Neuro-symbolic Concepts

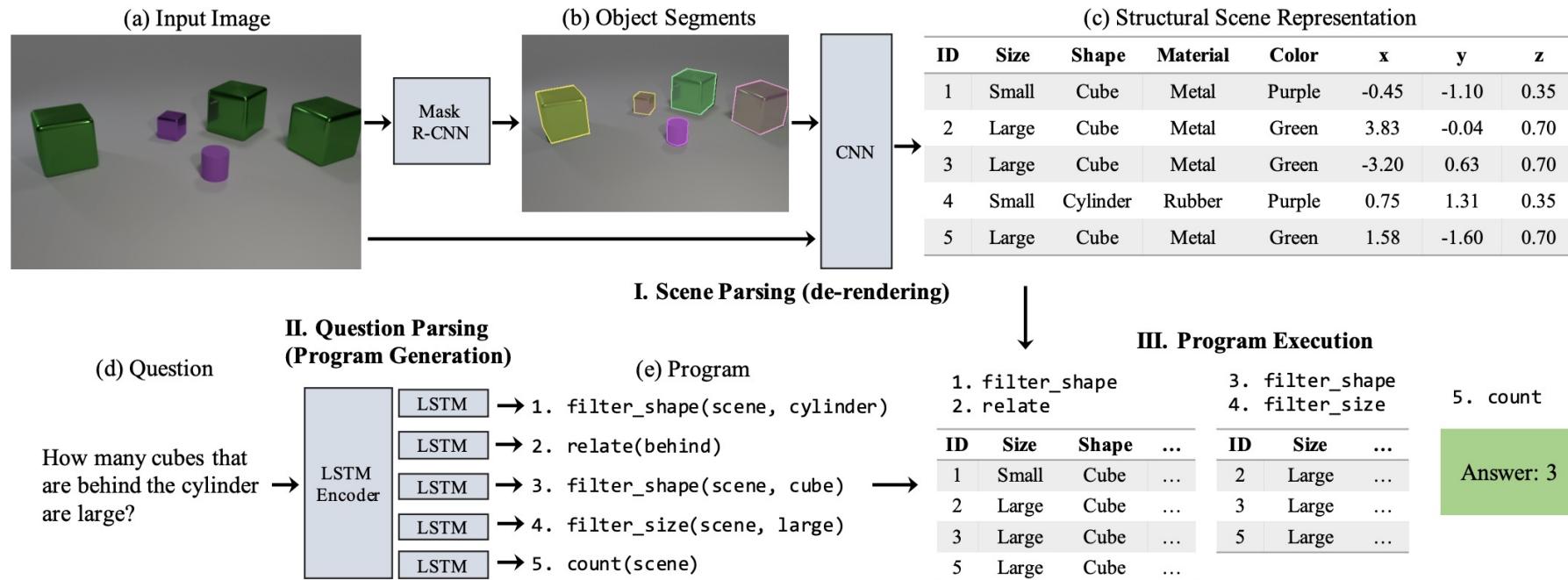
Hand-crafted concepts based on domain knowledge



[Yi et al., Neural-Symbolic VQA: Disentangling Reasoning from Vision and Language Understanding. NeurIPS 2018]

More Neuro-symbolic Concepts

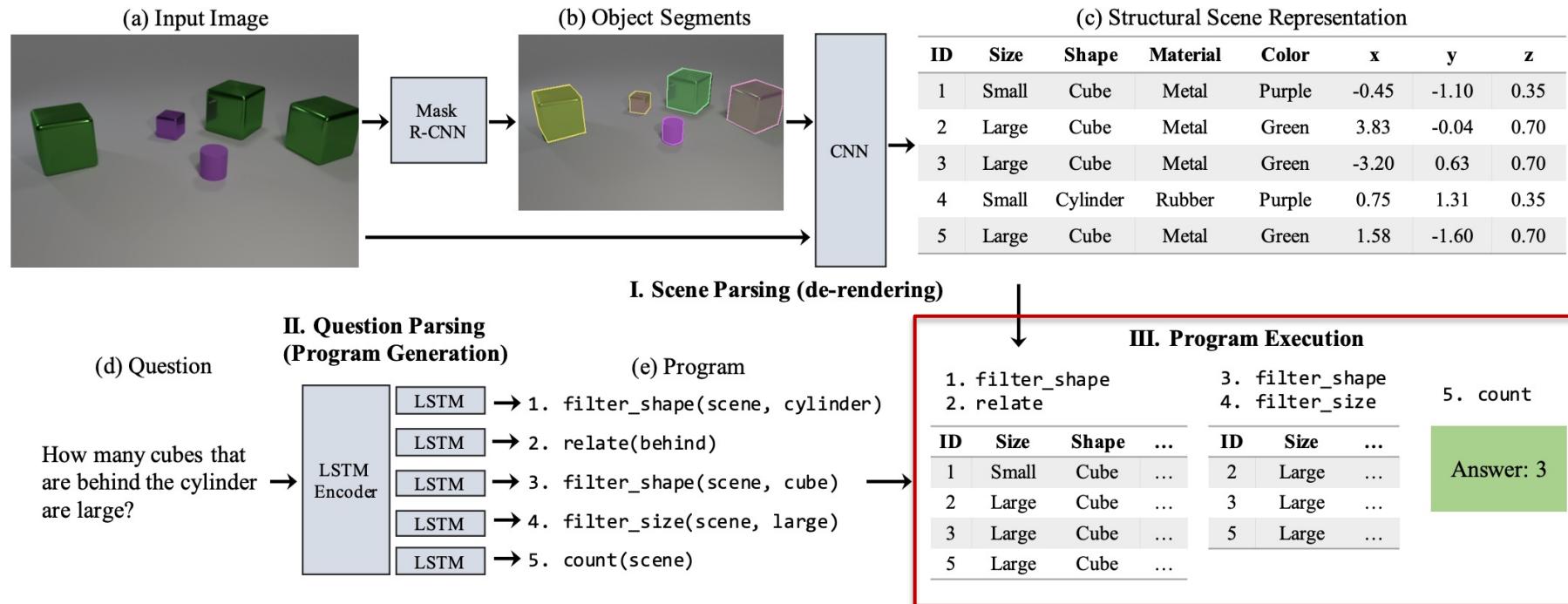
Hand-crafted concepts based on domain knowledge



[Yi et al., Neural-Symbolic VQA: Disentangling Reasoning from Vision and Language Understanding. NeurIPS 2018]

More Neuro-symbolic Concepts

Hand-crafted concepts based on domain knowledge



Pros:

- Robust (either it works or it doesn't)
- Data-efficient
- Human-interpretable

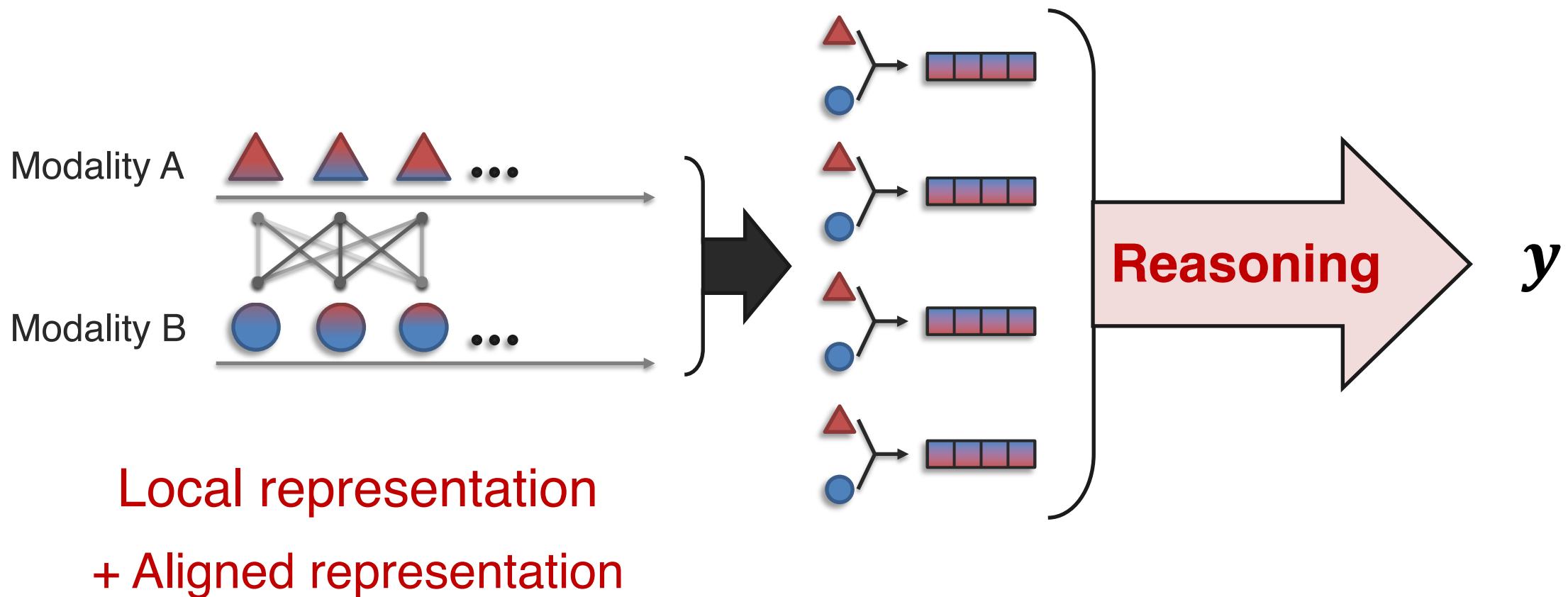
Cons:

- More engineered, specialized models
- Sometimes not fully differentiable (structure or concepts)
- Sometimes not perfect compatible with large-scale pre-training

[Yi et al., Neural-Symbolic VQA: Disentangling Reasoning from Vision and Language Understanding. NeurIPS 2018]

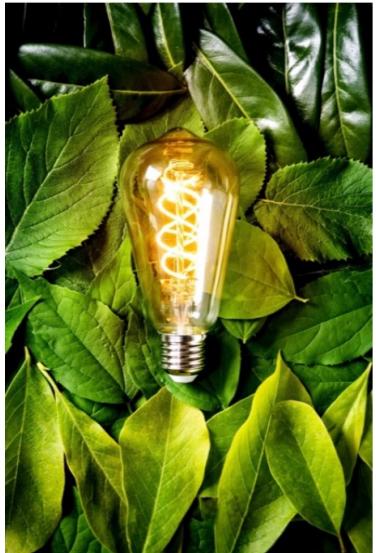
Summary: Reasoning Part 1

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



Summary: Reasoning Part 1

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

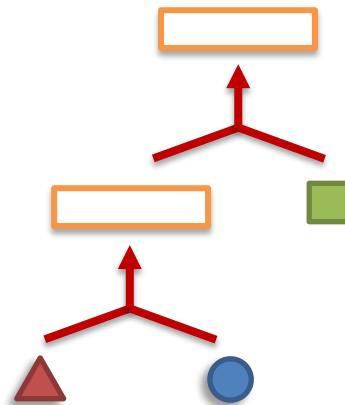
1. Structure: <subject> <verb> <object>
2. Concepts: ‘plants’, ‘lightbulb’
3. Inference: ‘surrounding’ – spatial relation
4. Knowledge: from humans!

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

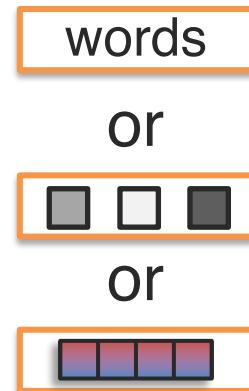
Summary: Reasoning Part 1

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

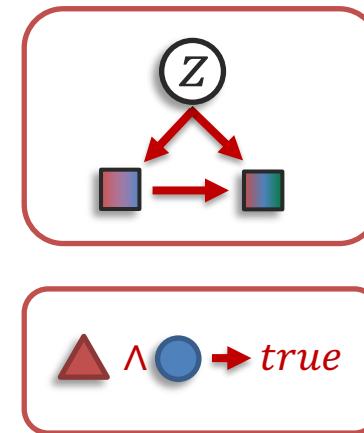
(A) Structure modeling



(B) Intermediate concepts



(C) Inference paradigm



(D) External knowledge



Summary: Reasoning Part 1

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

- (A) Structure modeling
- (B) Intermediate concepts
- (C) Inference paradigm
- (D) External knowledge

Today

Temporal
Hierarchical

Continuous

Dense or attention maps

Tree and graph networks

Memory and temporal networks

*Structure is given or can be learned easily in a differentiable manner.
* In the continuous case.

Roadmap for Next 3 Lectures

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

(A) Structure modeling

(B) Intermediate concepts

(C) Inference paradigm

(D) External knowledge

Today

Temporal
Hierarchical

Continuous

Next Tuesday

Interactive
Discovery

Discrete

Next Thursday

Causal
Logical

Knowledge
Commonsense