APPLIED ARTIFICIAL INTELLIGENCE INSTITUTE

DEAKIN UNIVERSITY

Tutorial at KDD, August 14th 2021

From Deep Learning to Deep Reasoning

Part B: Reasoning over unstructured and structured data

Truyen Tran, Vuong Le, Hung Le and Thao Le
{truyen.tran,vuong.le,thai.le,thao.le}@deakin.edu.au

https://bit.ly/37DYQn7

Agenda

- **Cross-modality reasoning, the case of vision-language integration.**
- Reasoning as set-set interaction.
- Relational reasoning
- Temporal reasoning
  - Video question answering.

Learning to Reason formulation

- Input:
  - A knowledge context C
  - A query q
- Output: an answer satisfying

$$\tilde{a} = \arg\max_{a \in \mathbb{A}} \mathcal{P}_\theta \left(a \mid C, q \right)$$

- C can be
  - structured: knowledge graphs
  - unstructured: text, image, sound, video

Q: "What affects her mobility?"

Q: Is it simply an optimization problem like recognition, detection or even translation?

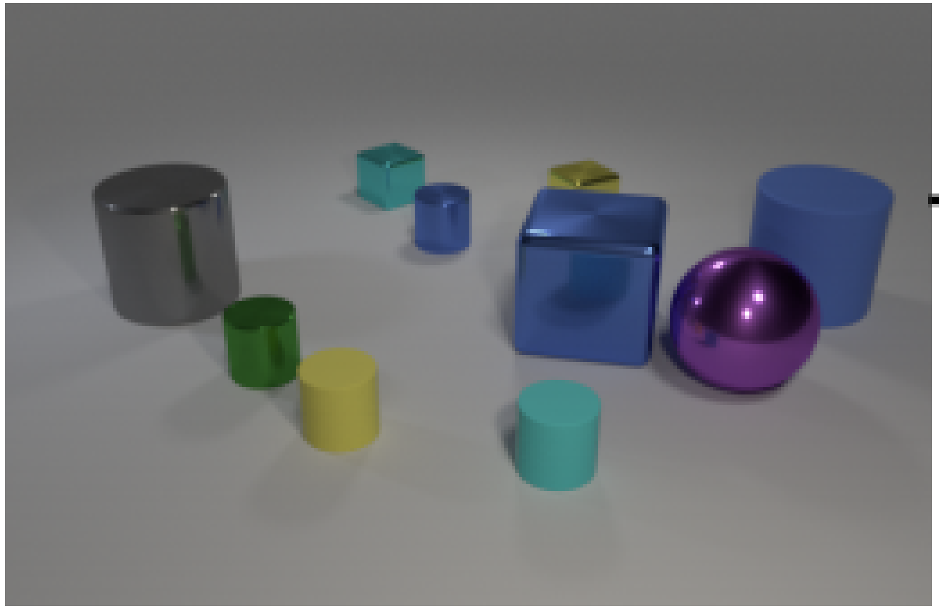
→ No, because the logics from C, q into a is more complex than other solved optimization problems

→ We can solve (some parts of) it with good structures and inference strategies

A case study: Image Question Answering

$$\tilde{a} = \arg\max_{a \in \mathbb{A}} \mathcal{P}_\theta\left(a \mid C, q\right)$$

- Realization
  - C: visual content of an image
  - q: a linguistic question
  - a: a linguistic phrase as
    the answer to q regarding K
- Challenges
  - Reasoning through facts and logics
  - Cross-modality integration



How many tiny yellow matte things are to the right of the purple thing in the front of the small cyan shiny cube?

Image QA: Question types

Open-ended
- Is this a vegetarian pizza?
- What is the red thing in the photo?

Multi-choice

(Q) What is the red thing in the photo?

(A) (1) capsicum (2) beef
(3) mushroom (4) cheese

Counting
- How many slices of pizza are there?

(VQA, Agrawal et al., 2015)

Image QA datasets

(VQA, Agrawal et al., 2015)

(Q) What is in the picture?
(Q) Is this a vegetarian pizza?

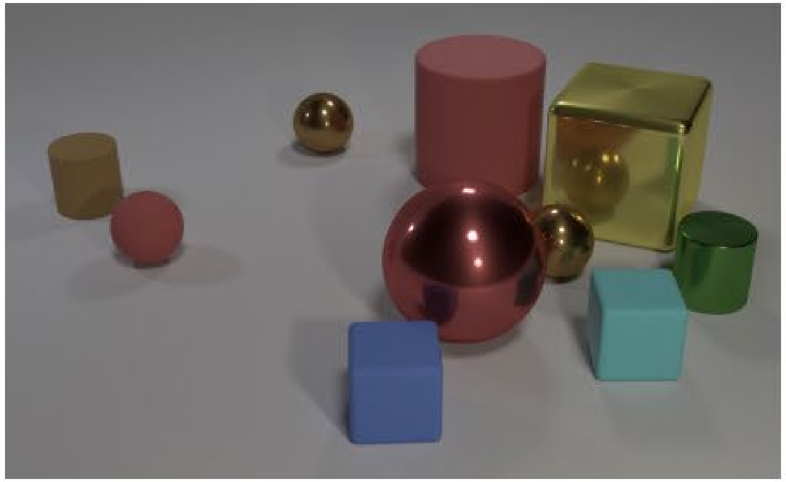
Perception

(GQA, Hudson et al., 2019)

(Q) What is the brown animal sitting inside of?
(Q) Is there a bag to the right of the green door?

Relational reasoning

(CLEVR, Johnson et al., 2017)



(Q) How many objects are either small cylinders or metal things?
(Q) Are there an equal number of large things and metal spheres?

Multi-step reasoning

The two main themes in Image QA

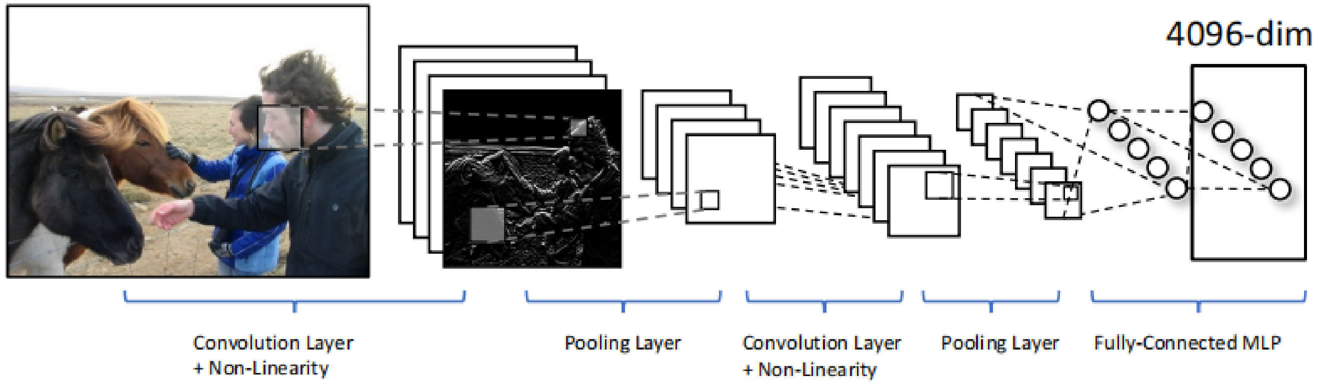
- Neuro-symbolic reasoning
  - Parse the question into a “program” of small steps
  - Learn the generic steps as neural modules
  - Use and reuse the modules for different programs
- **Compositional reasoning**
  - Extract visual and linguistic individual- and joint- representation
  - Reasoning happens on the structure of the representation
    - Sets/graphs/sequences
  - The representation got refined through multi-step compositional reasoning

Agenda

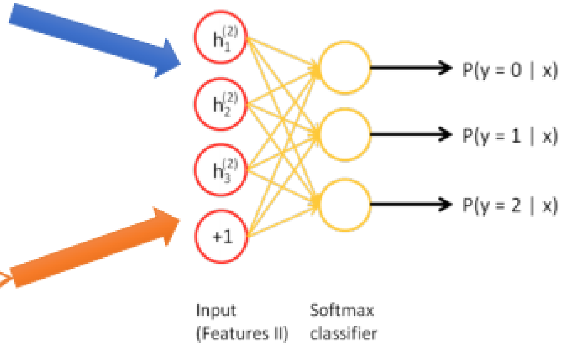
- Cross-modality reasoning, the case of vision-language integration.
- **Reasoning as set-set interaction.**
- Relational reasoning
- Temporal reasoning
  - Video question answering.

A simple approach

Image Embedding (VGGNet)

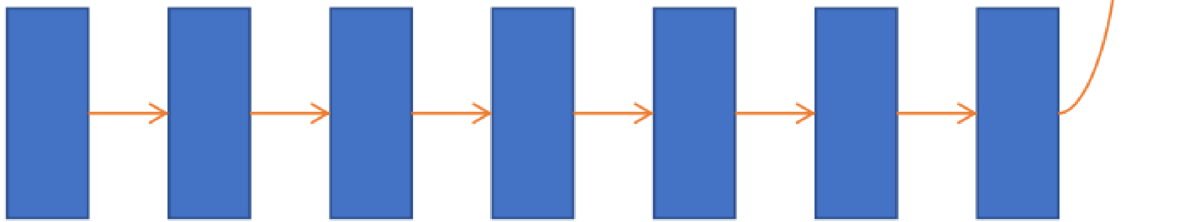


Neural Network
Softmax
over top K answers



Question Embedding (LSTM)

"How many horses are in this image?"



→ Issue: This is very susceptible to the nuances of images and questions

Agrawal et al., 2015, Slide credit: Thao Minh Le

Reasoning as set-set interaction

- C: a set of context objects

$$C = \{o_1, o_2, ..., o_n\}$$

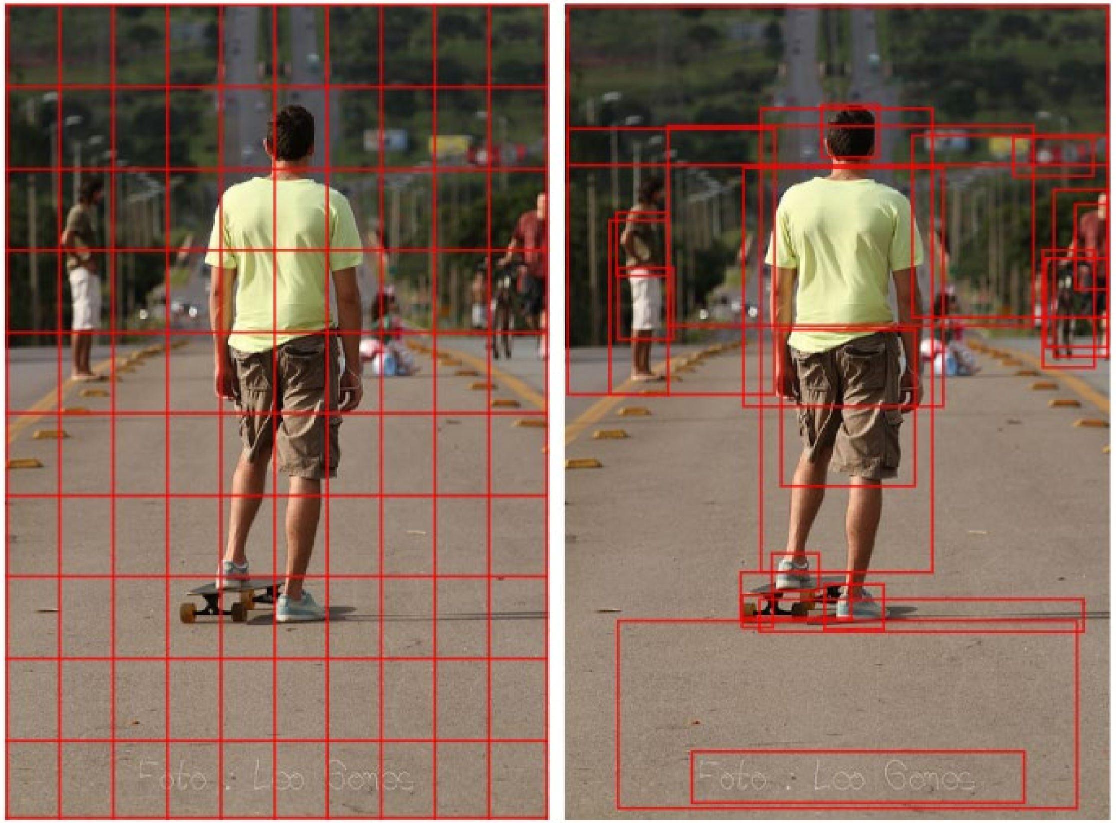
  - Faster-RCNN regions
  - CNN tubes

- q: a set of linguistic objects L.

$$L = \{w_1, w_2, ..., w_n\}$$

- biLSTM embedding of q

$$\mathbf{w}_i^q = [\overrightarrow{\mathrm{LSTM}}(\mathbf{e}_i^q); \overleftarrow{\mathrm{LSTM}}(\mathbf{e}_i^q)]$$



→ Reasoning is formulated as the interaction between the two sets O and L for the answer a

Set operations

- Reducing operation (eg: sum/average/max)

$$\mathbf{c} \quad = \quad h_{\boldsymbol{\theta}}\left(\{\mathbf{o}_1, \mathbf{o}_2, .., \mathbf{o}_N\}\right)$$

- Attention-based combination (Bahdanau et al. 2015)

$$\mathbf{c} \quad = \quad \sum_{i=1}^{N} \alpha_i \mathbf{o}_i \qquad\qquad \alpha_i \quad = \quad \frac{\exp(\mathbf{W}^o \mathbf{o}_i)}{\sum_{j=1}^{N} \exp(\mathbf{W}^o \mathbf{o}_j)}$$

- Attention weights as query-key dot product (Vaswani et al., 2017)

$$\mathbf{c} \quad = \quad \mathrm{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V}$$

→ Attention-based set ops seem very suitable for visual reasoning

Attention-based reasoning

- Unidirectional attention
  - Find relation score between parts in the context C to the question q:

$$s_i = f(\mathbf{q}, \mathbf{w}_j^c)$$

Op

$$s_i = \tanh(\mathbf{W}^c \mathbf{w}_i^c + \mathbf{W}^{\mathbf{q}} \mathbf{q})$$

- $$s_i = \mathbf{q}^\top \mathbf{W}^s \mathbf{w}_i^c$$ Hermann et al. (2015)
- Chen et al. (2016)

  - Normalized by tion weights

$$\alpha_i = \frac{\exp(\mathbf{W} s_i)}{\sum_j \exp(\mathbf{W} s_j)}$$

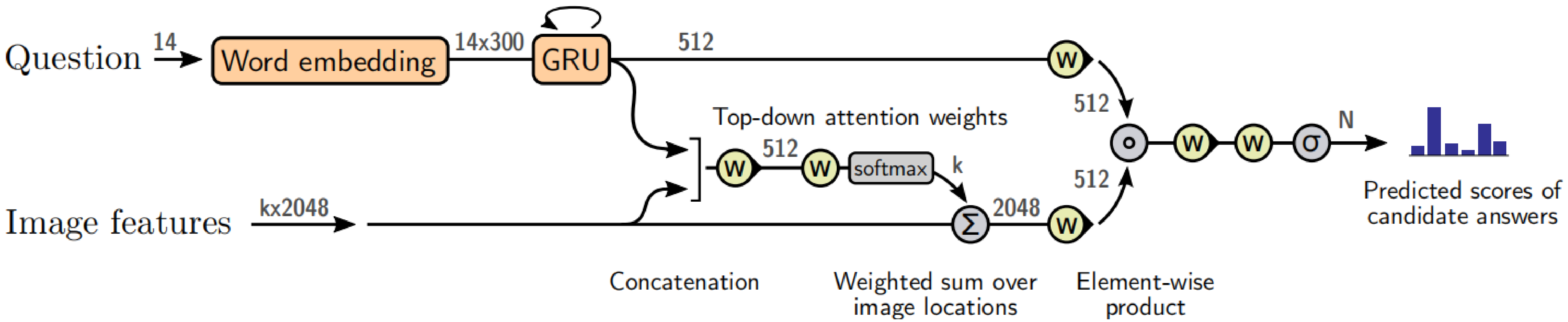
$$\mathbf{i} = \sum_i \alpha_i \mathbf{w}_i^c$$

  - Attended context vector:

→ We can now extract information from the context that is "relevant" to the query

Bottom-up-top-down attention (Anderson et al 2017)

- Bottom-up set construction: Choosing Faster-RCNN regions with high class scores
- Top-down attention: Attending on visual features by question



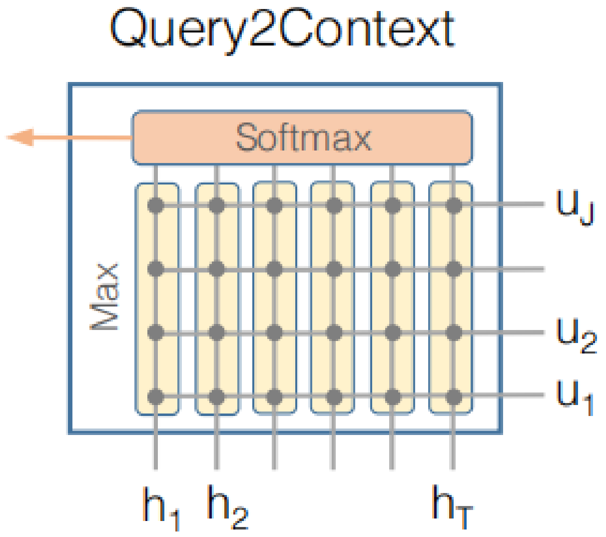
→ Q: How about attention from vision objects to linguistic objects?

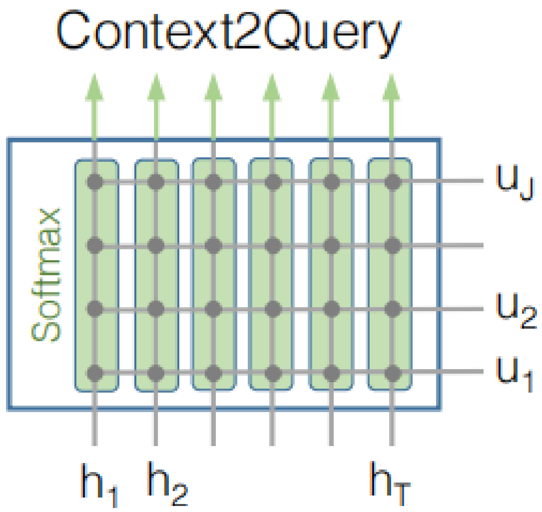
Bi-directional attention

- Question-context similarity measure

$$s_i = f(\mathbf{q}, \mathbf{w}_j^c)$$

- Question-guided context attention
  - Softmax across columns
- Context-guided question attention
  - Softmax across rows





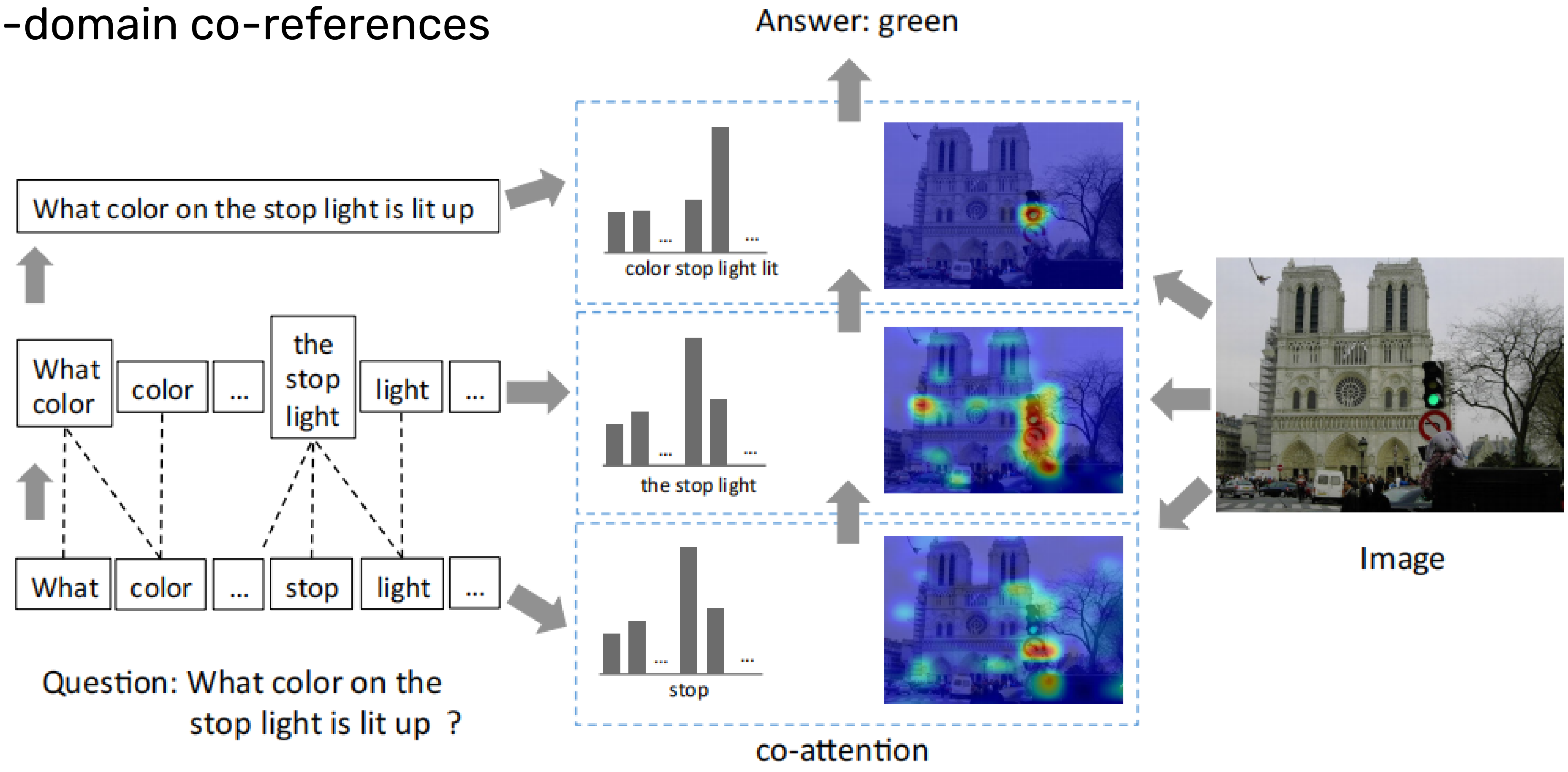
→ Q: Probably not working for image qa where single words does not have the co-reference with a region?

Dynamic coattention networks for question answering (Seo et al., ICLR 2017)

Hierarchical co-attention for ImageQA

- The co-attention is found on a word-phrase-sentence hierarchy

→ better cross-domain co-references



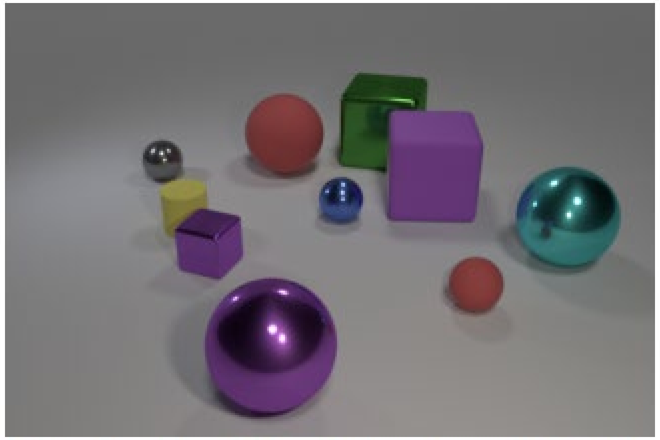
→ Q: Can this be done on text qa as well?

→ Q: How about questions with many reasoning hops?

Multi-step compositional reasoning

- Complex question need multiple hops of reasoning
- Relations inside the context are multi-step themselves
- Single shot of attention won't be enough
- Single shot of information gathering is definitely not enough

→ Q: How to do multi-hop attentional reasoning?

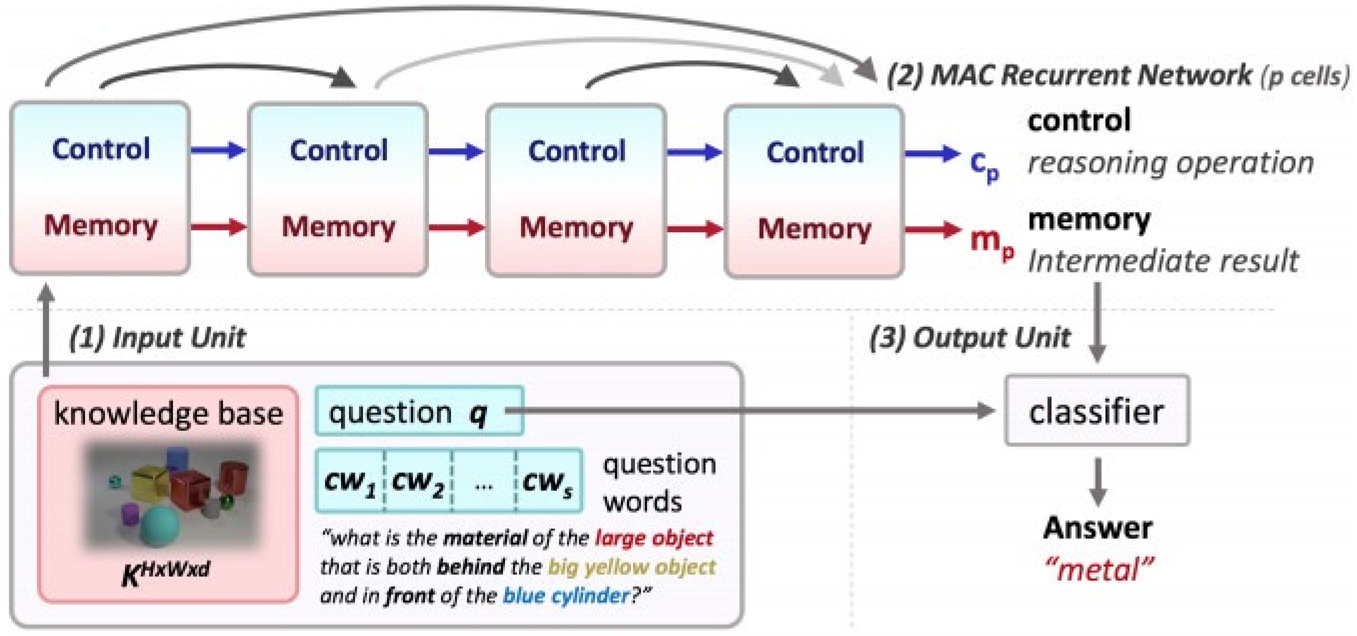


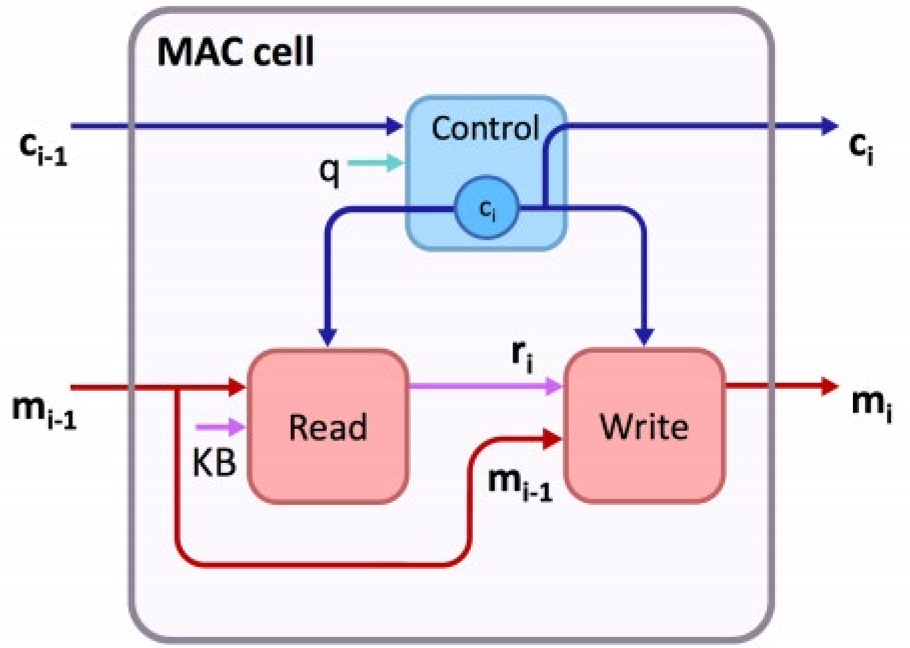
Q: *Do **the block** in **front** of **the tiny yellow cylinder** and **the tiny thing** that is to the **right** of **the large green shiny object** have the **same color**?* **A:** *No*

Figure: Hudson and Manning – ICLR 2018

Multi-step reasoning - Memory, Attention, and Composition (MAC Nets)

- Attention reasoning is done through multiple sequential steps.
- Each step is done with a recurrent neural cell
- *What is the key differences to the normal RNN (LSTM/GRU) cell?*
  - *Not a sequential input, it is sequential processing on static input set.*
  - *Guided by the question through a controller.*

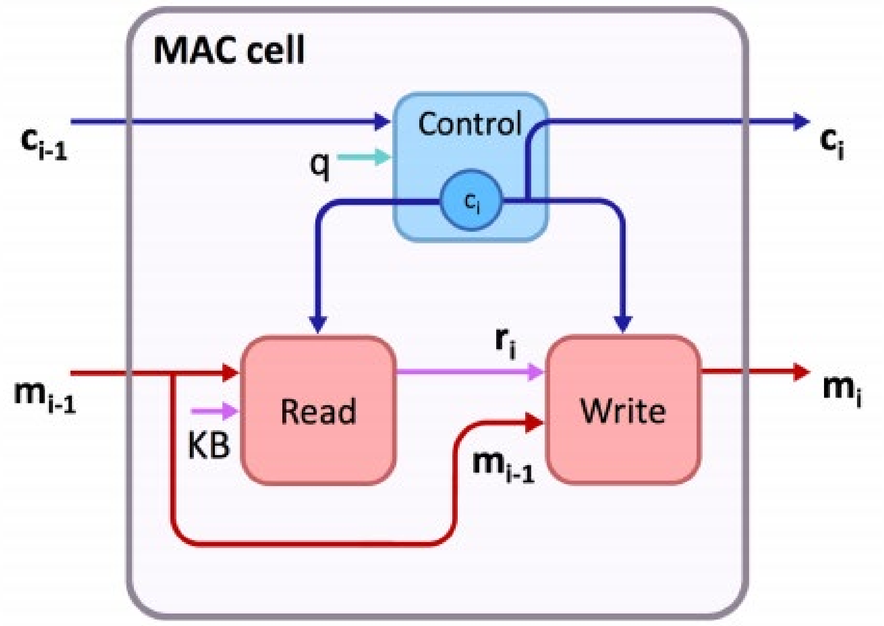




MAC network, Hudson and Manning – ICLR 2018

Multi-step attentional reasoning

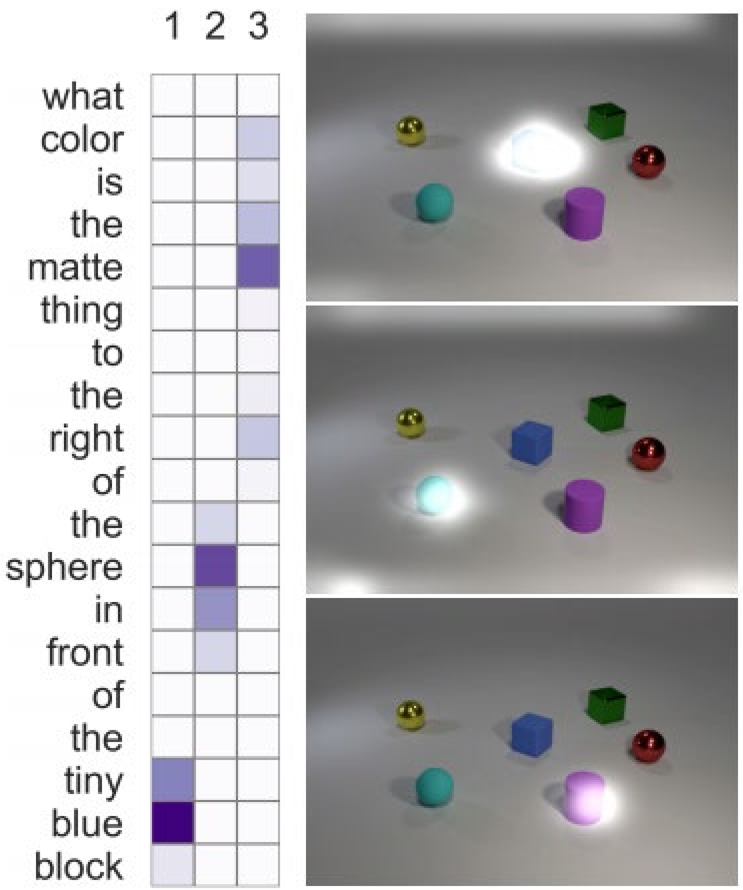
- At each step, the controller decide what to look next
- After each step, a piece of information is gathered, represented through the attention map on question words and visual objects
- A common memory kept all the information extracted toward an answer



MAC network, Hudson and Manning – ICLR 2018

Multi-step attentional reasoning

- Step 1: attends to the ***"tiny blue block"***, updating ***m*1**
- Step 2: look for ***"the sphere in front"*** ***m*2**.
- Step3: traverse from the cyan ball to the final objective – ***the purple cylinder***,



Reasoning as set-set interaction – a look back

- C : a set of context objects

$$C = \{o_1, o_2, ..., o_n\}$$

- q: a set of linguistic objects

$$L = \{w_1, w_2, ..., w_n\}$$

- Reasoning is formulated as the interaction between the two sets O and L for the answer a

Q:What is the brown animal sitting inside of?

→ Q: Set-set interaction falls short for questions about *relations between objects*

Agenda

- Cross-modality reasoning, the case of vision-language integration.
- Reasoning as set-set interaction.
- **Relational reasoning**
- Temporal reasoning
  - Video question answering.

Reasoning on Graphs

- Relational questions: requiring explicit reasoning about the relations between multiple objects

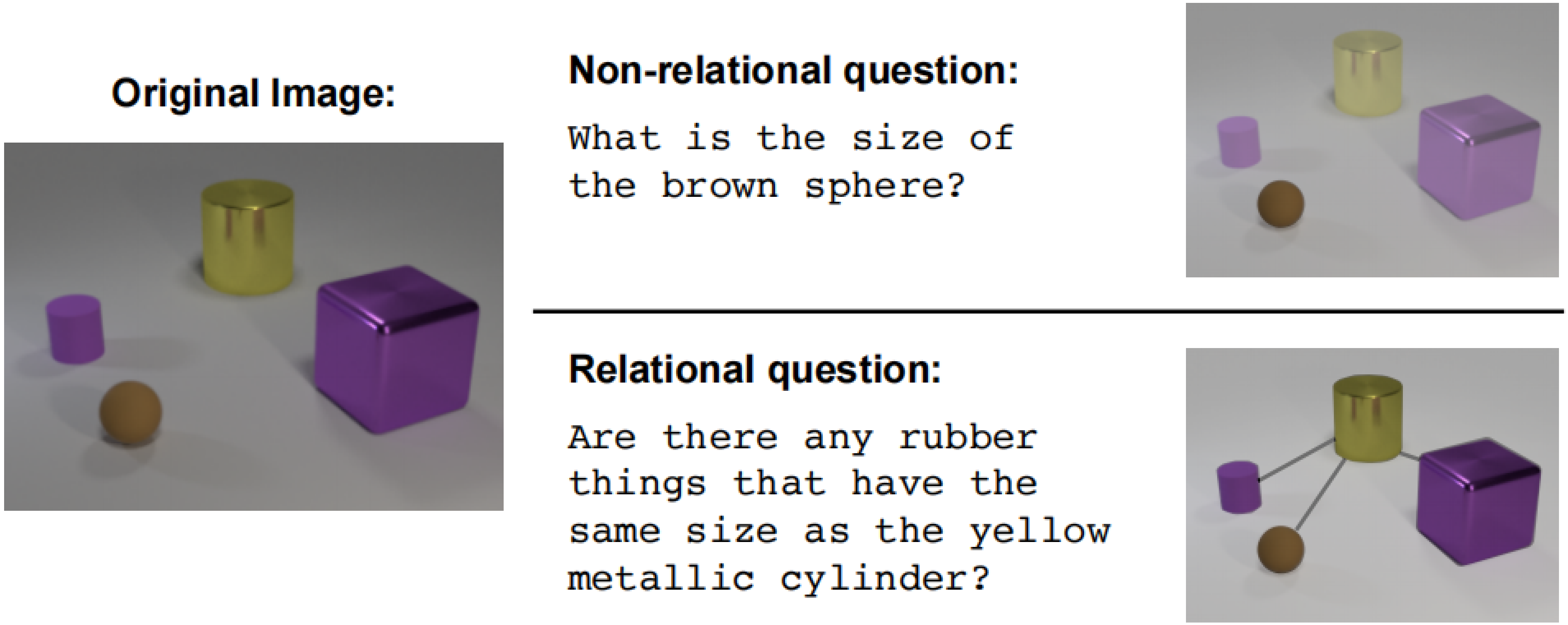
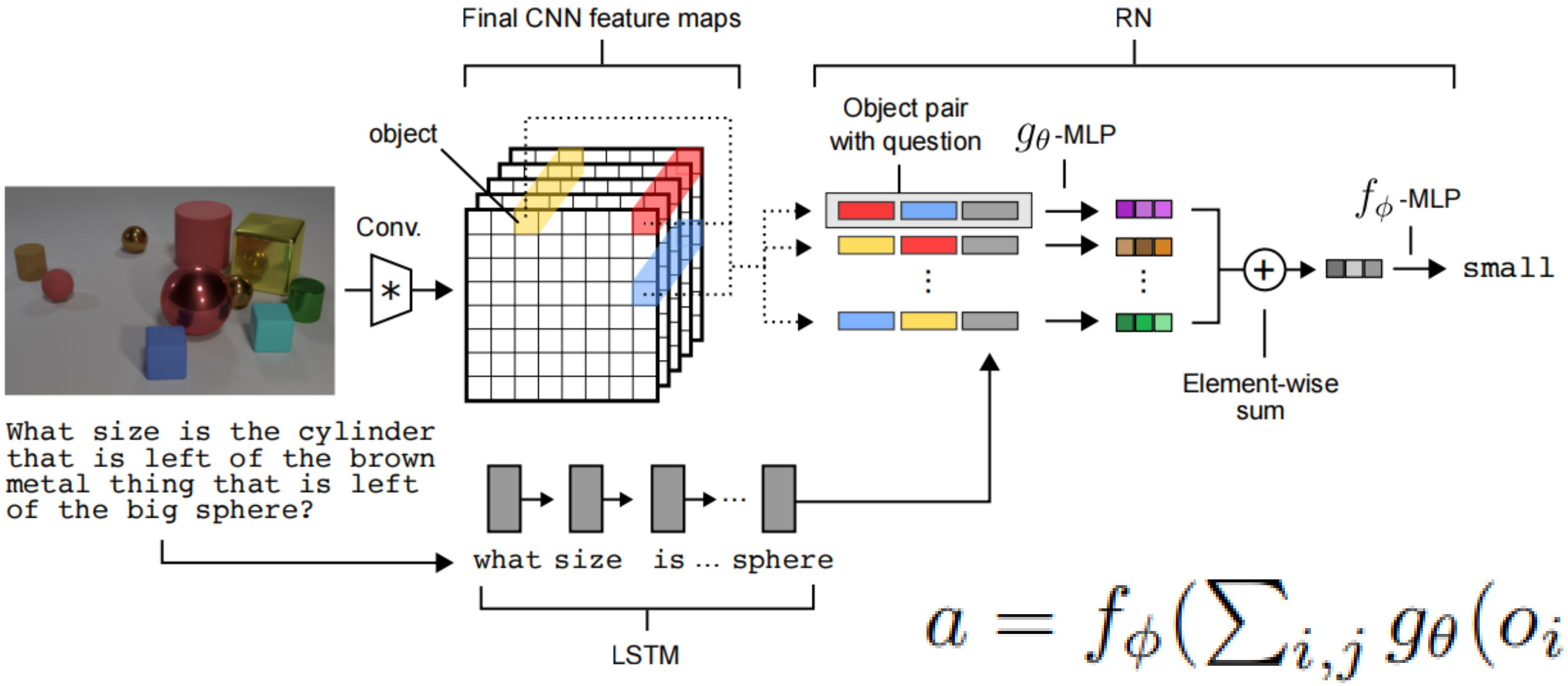


Figure credit: Santoro et al 2017

Relation networks (Santoro et al 2017)

- Relation networks $\mathrm{RN}(O) = f_\phi\left(\sum_{i,j} g_\theta(o_i, o_j)\right)$
- f_ϕ and g_θ are neural functions
- g_θ generate "relation" between the two objects
- f_ϕ is the aggregation function



$$a = f_\phi\left(\sum_{i,j} g_\theta(o_i, o_j, q)\right)$$

→ The relations here are implicit, complete, pair-wise – inefficient, and lack expressiveness

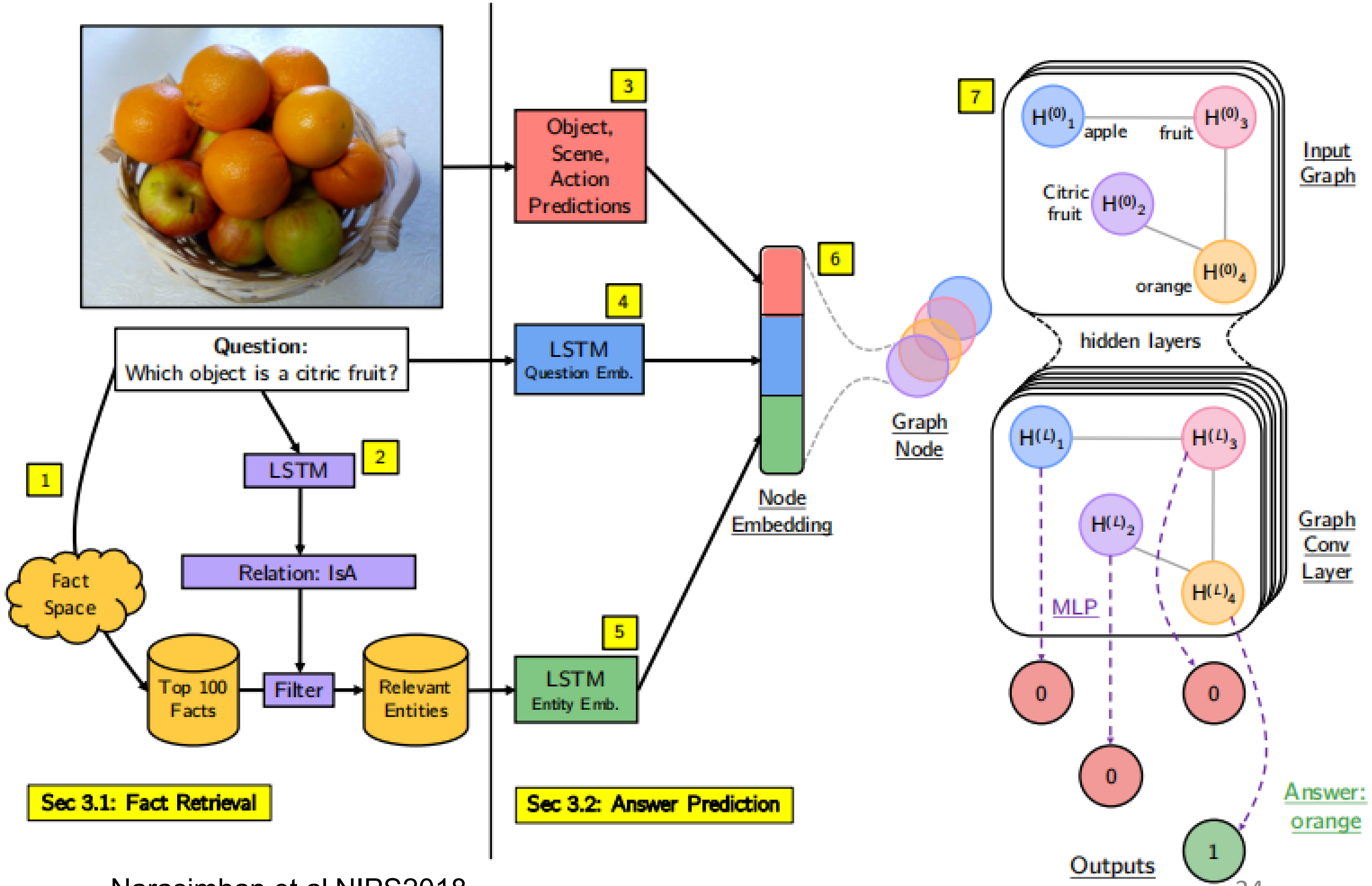
Reasoning with Graph convolution networks

- Input graph is built from image entities and question
- GCN is used to gather facts and produce answer

→ The relations are now explicit and pruned

→ But the graph building is very stiff:

- Unrecovrable if it makes a mistake?
- Information during reasoning are not used to build graphs



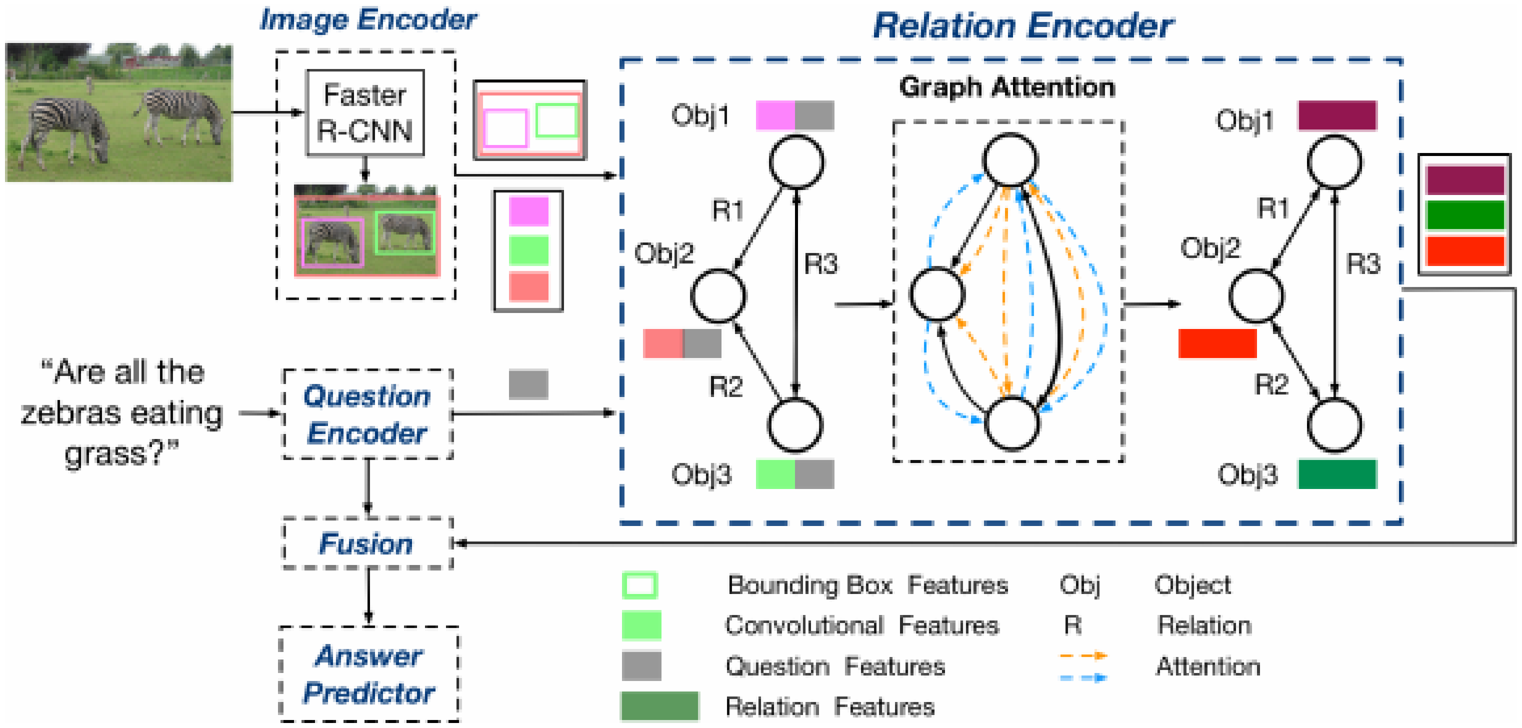
Narasimhan et.al NIPS2018

Reasoning with Graph attention networks

- The graph is determined during reasoning process with attention mechanism

→The relations are now adaptive and integrated with reasoning

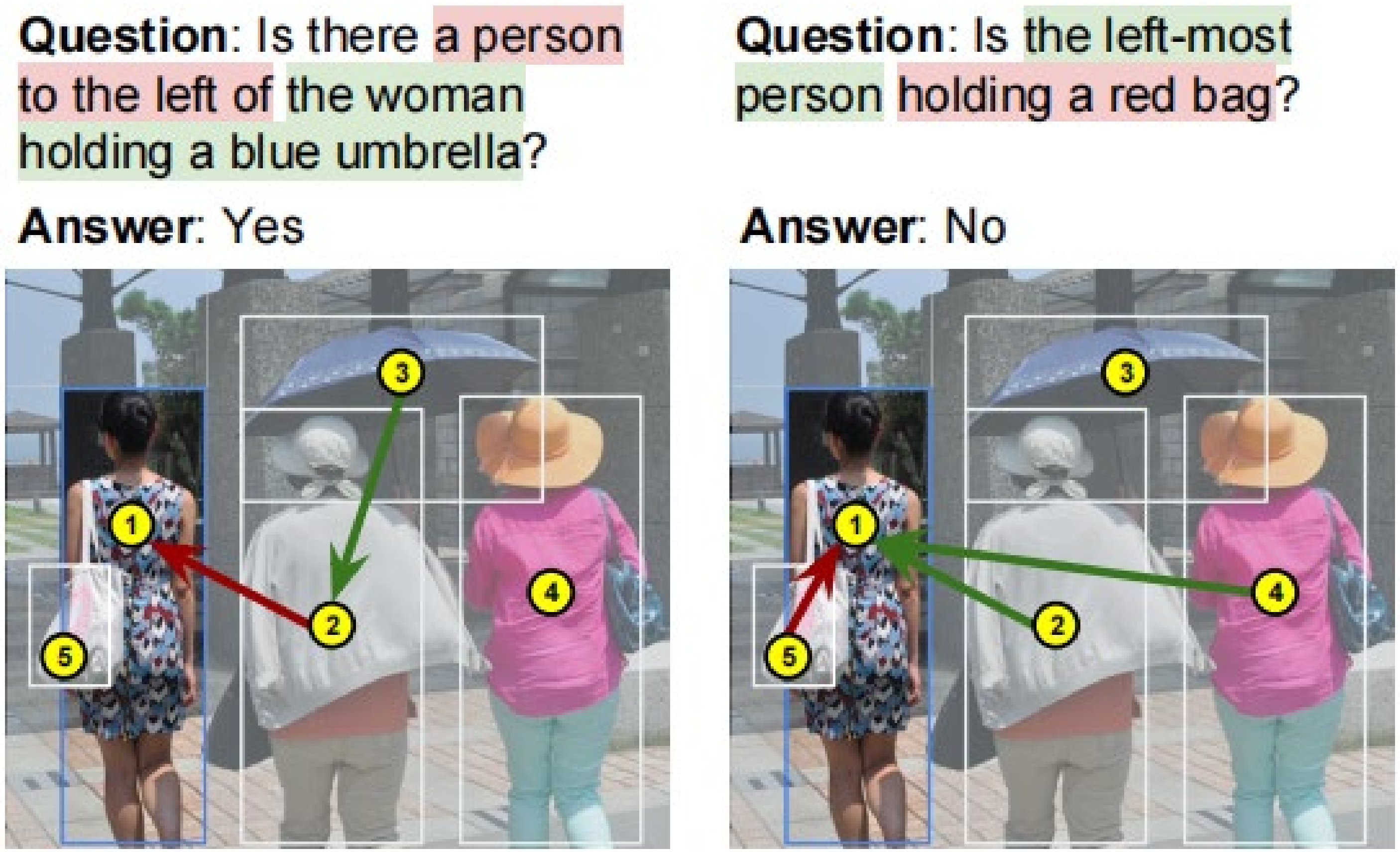
→ Are the relations singular and static?



ReGAT model, Li et.al. ICCV19

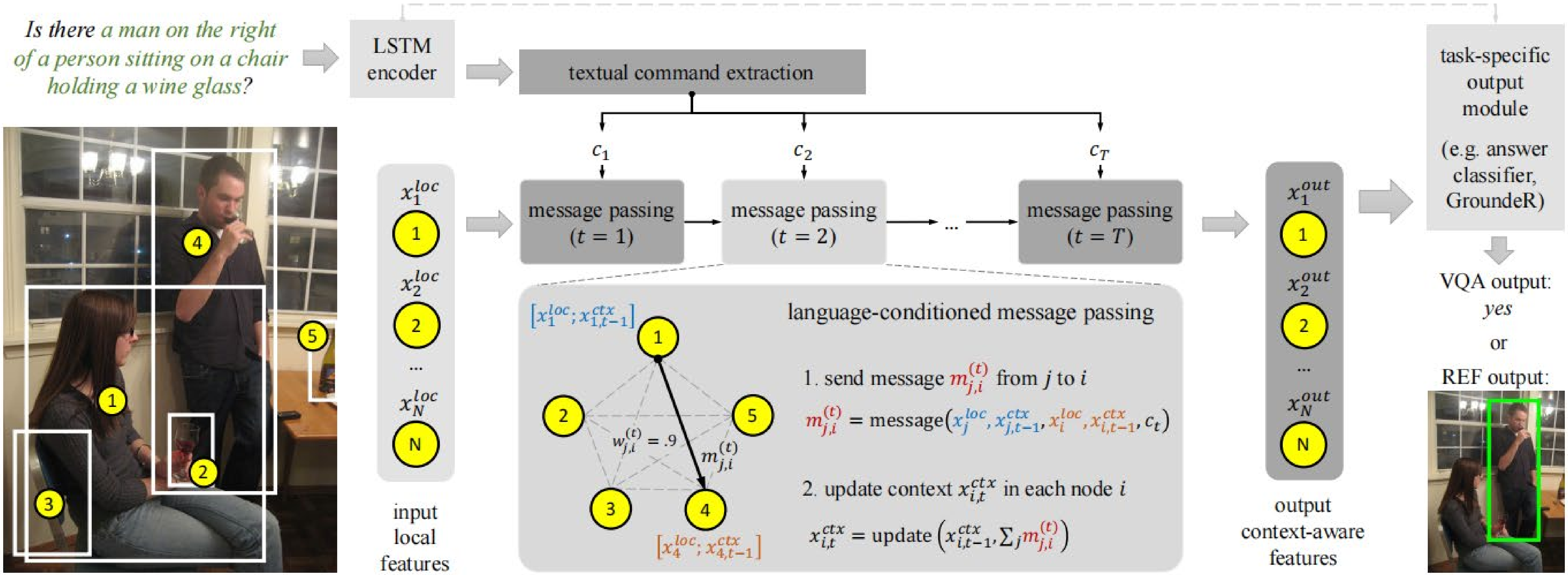
Dynamic reasoning graphs

- On complex questions, multiple sets of relations are needed
- We need not only multi-step but also multi-form structures
- Let's do multiple dynamically–built graphs!



LCGN, Hu et.al. ICCV19

Dynamic reasoning graphs

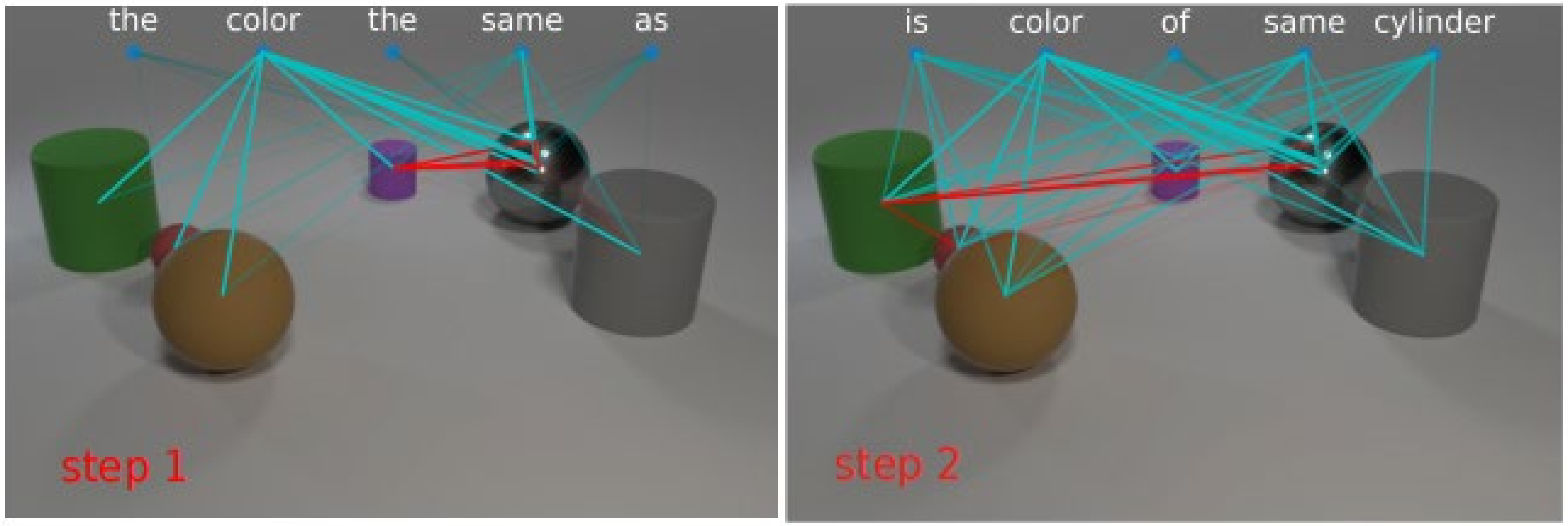


→ The questions so far act as an unstructured command in the process

→ *Aren't their structures and relations important too?*

Reasoning on cross-modality graphs

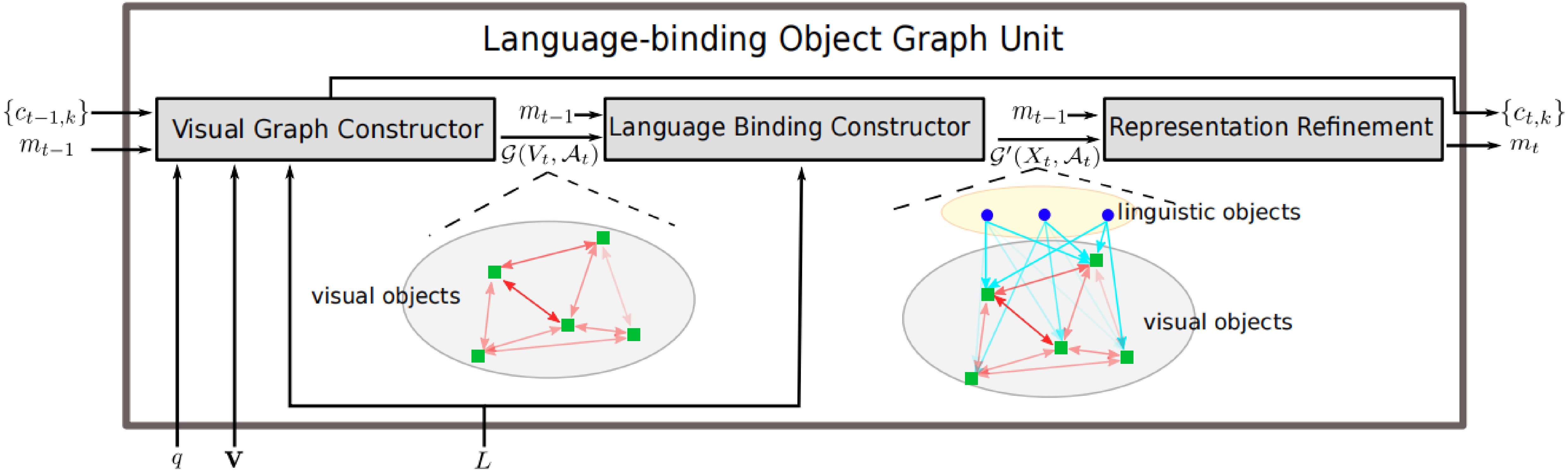
- Two types of nodes: Linguistic entities and visual objects
- Two types of edges:
  - Visual
  - Linguistic-visual binding *(as a fuzzy grounding)*
- Adaptively updated during reasoning



LOGNet, T.M Le et.al. IJCAI2020

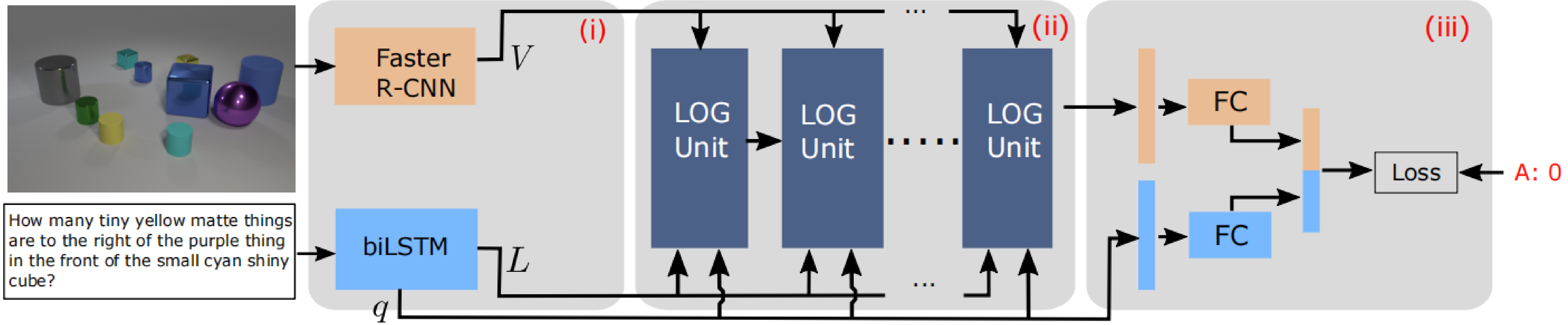
Language-binding Object Graph (LOG) Unit

- Graph constructor: build the dynamic vision graph
- Language binding constructor: find the dynamic L-V relations

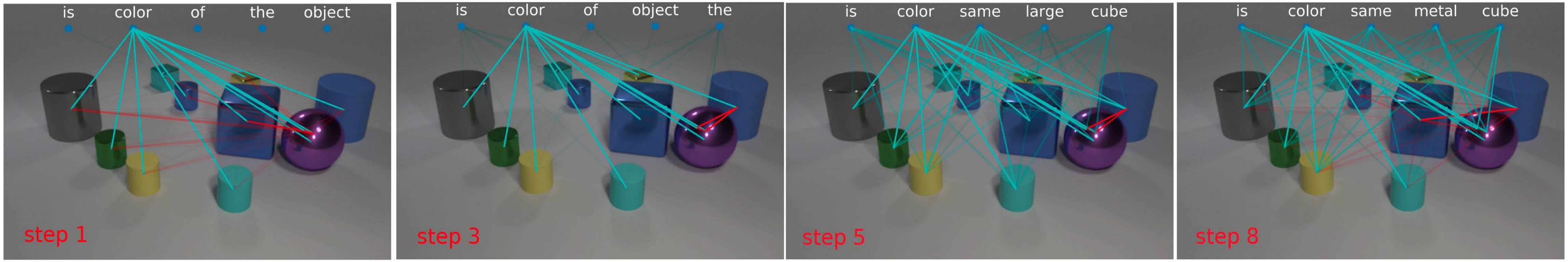


LOGNet: multi-step visual-linguistic binding

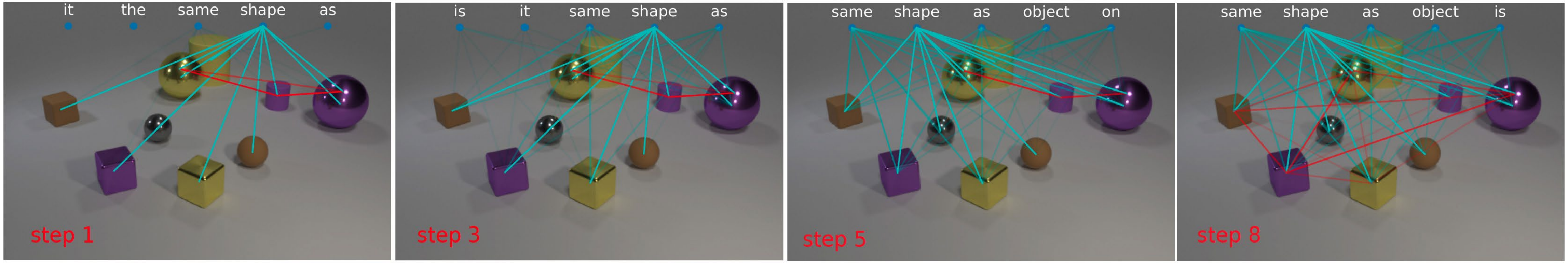
- Object-centric representation ✓
- Multi-step/multi-structure compositional reasoning ✓
- Linguistic-vision detail interaction ✓



Dynamic language-vision graphs in actions



Question: Is the color of the big matte object the same as the large metal cube?
Prediction: yes **Answer**: yes



Question: There is a tiny purple rubber thing; does it have the same shape as the brown object that is on the left side of the rubber sphere?
Prediction: no **Answer**: no

We got sets and graphs, how about sequences?

- Videos pose another challenge for visual reasoning: the dynamics through time.
- Sets and graphs now becomes sequences of such.
- Temporal relations are the key factors
- The size of context is a core issue

(a) Question: What does the girl do 9 times?
Baseline: walk
HCRN: blocks a person's punch
Ground truth: blocks a person's punch

(b) Question: What does the man do before turning body to left?
Baseline: pick up the man's hand
HCRN: breath
Ground truth: breath

Agenda

- Cross-modality reasoning, the case of vision-language integration.
- Reasoning as set-set interaction.
- Relational reasoning
- **Temporal reasoning**
  - **Video question answering.**

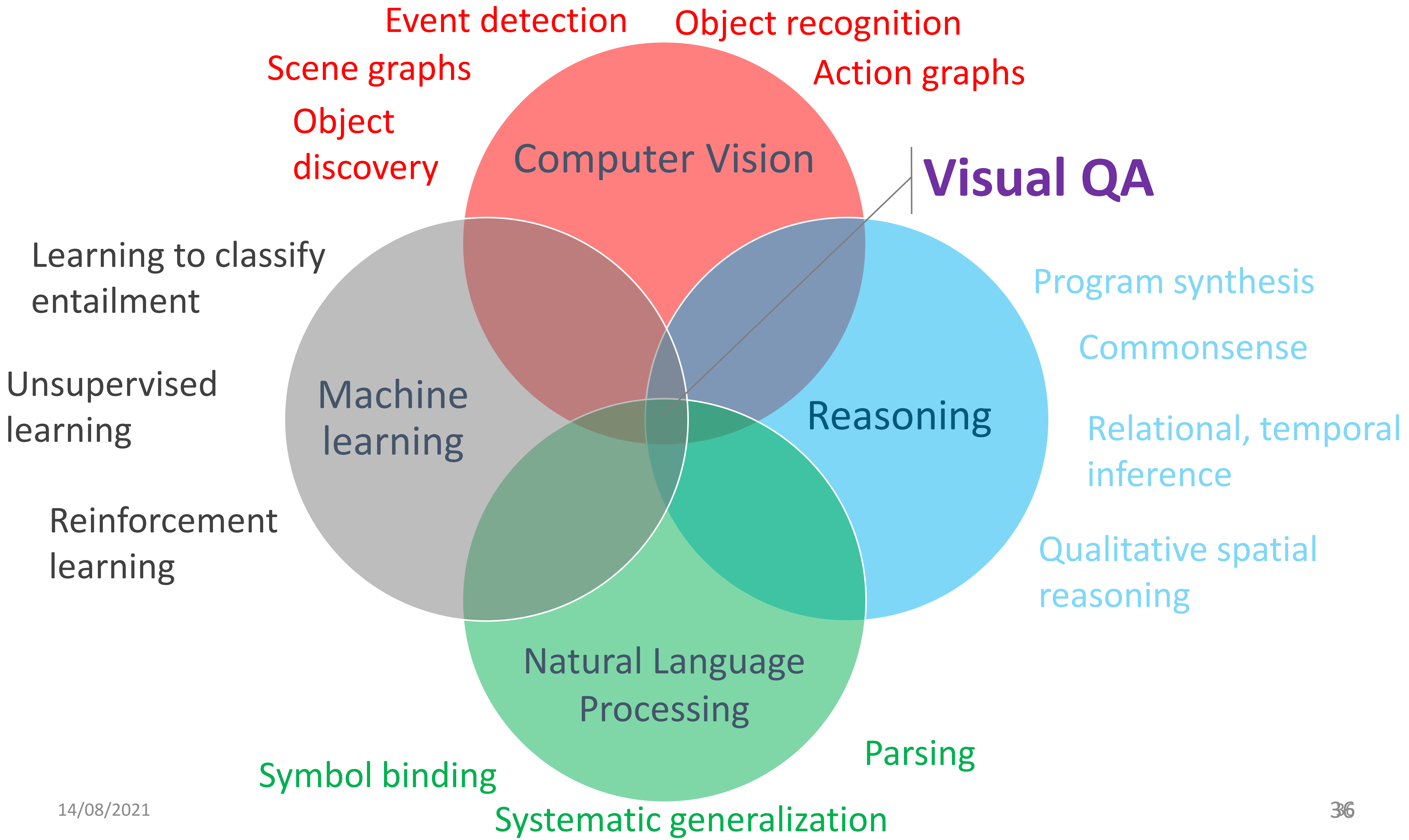
Overview

- **Goals of this part of the tutorial**
  - Understanding Video QA as a complete testbed of visual reasoning.
  - Representative state-of-the-art approaches for spatio-temporal reasoning.

Video Question Answering

Short-form Video Question Answering

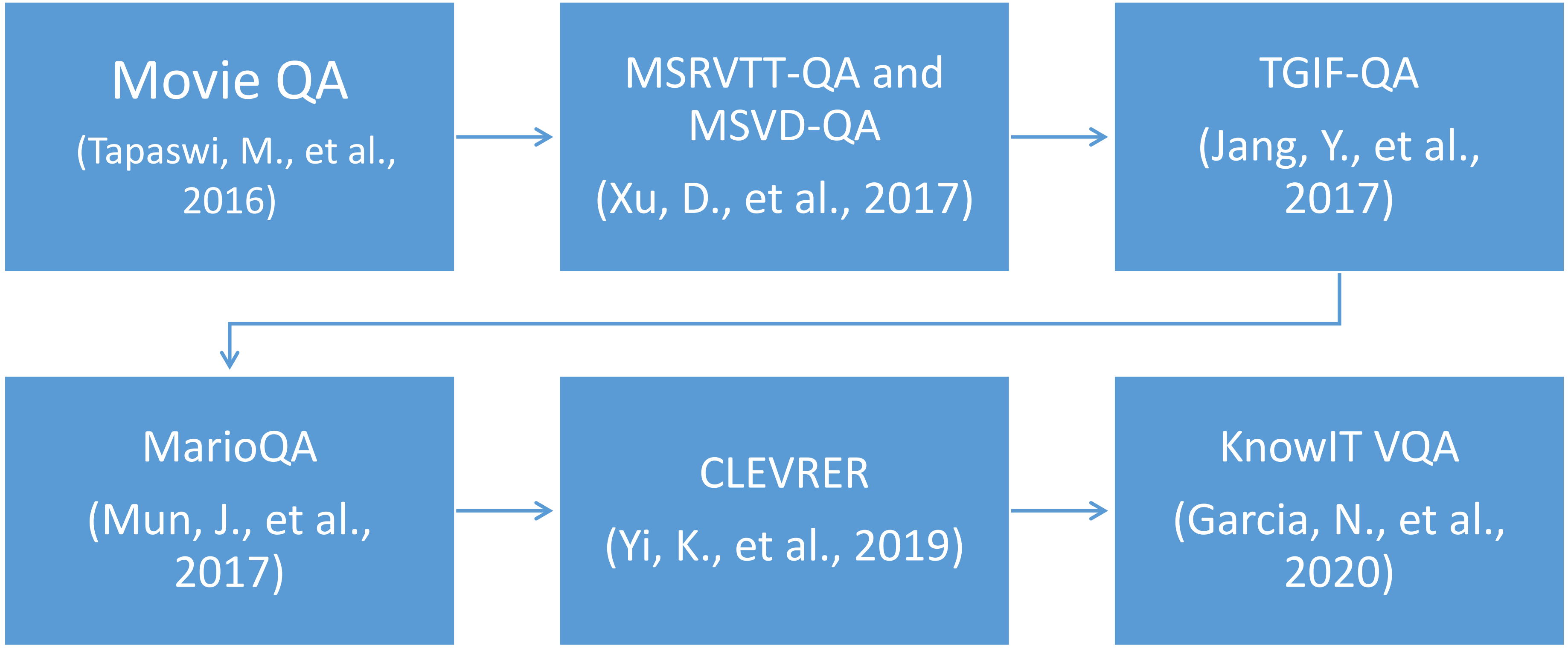
Movie Question Answering


Event detection
Object recognition
Scene graphs
Action graphs
Object discovery
Computer Vision
Visual QA
Learning to classify entailment
Program synthesis
Commonsense
Unsupervised learning
Machine learning
Reasoning
Relational, temporal inference
Reinforcement learning
Qualitative spatial reasoning
Natural Language Processing
Parsing
Symbol binding
Systematic generalization

Challenges

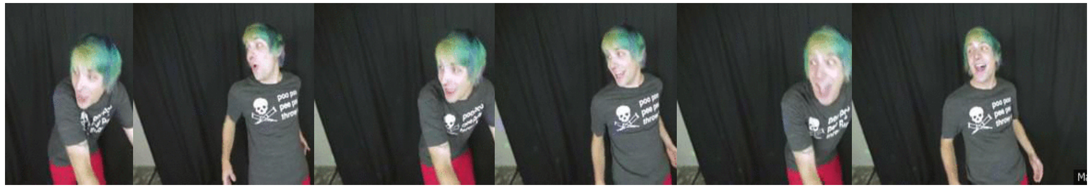
- Difficulties in data annotation.
- Content for performing reasoning spreads over space-time and multiple modalities (videos, subtitles, speech etc.)

Video QA Datasets



Video QA datasets

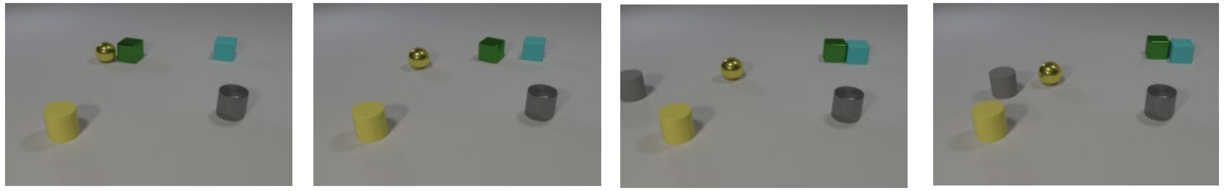
(TGIF-QA, Jang et al., 2018)



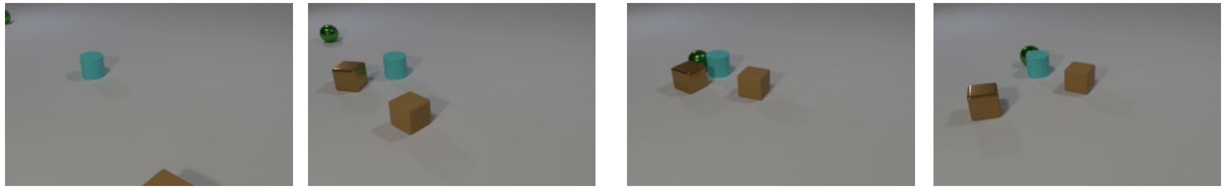
Q: What does the man do 5 times?
A: (0) step (3) bounce
(2) sway head (4) knod head
(5): move body to the front

Q: What does the man do before turing body to left?
A: (0) run a cross a ring (3) flip cover face with hand
(2) pick up the man's hand (4) raise hand
(5): breath

(CLEVRER, Yi, Kexin, et al., 2020)

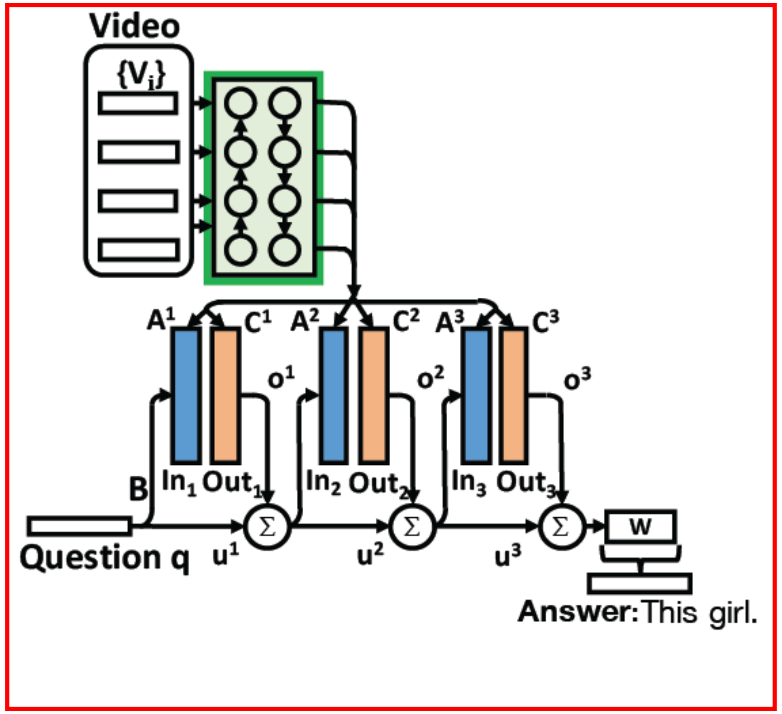


Q: What color is the last object to collide with the green cube?
A: cyan

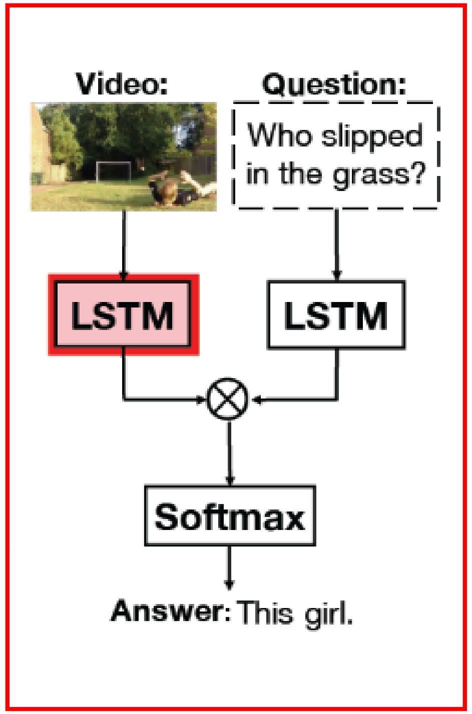


Q: Which of the following is responsible for the collision between the metal cube and the cylinder?
A: (a) The presence of the brown rubber cube
(b) The sphere's colliding with the cylinder
(c) The rubber cube's entrance
(d) The collision between the metal cube and the sphere

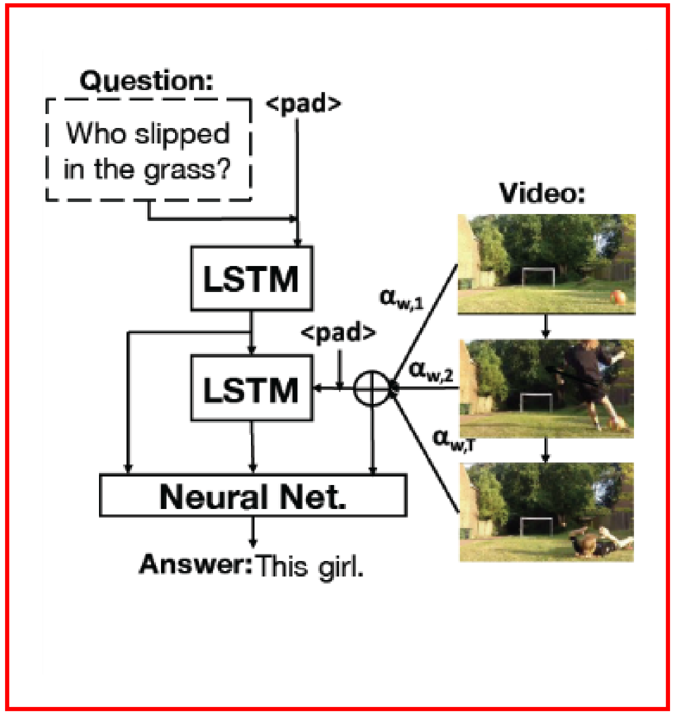
Video QA as a spatio-temporal extension of Image QA



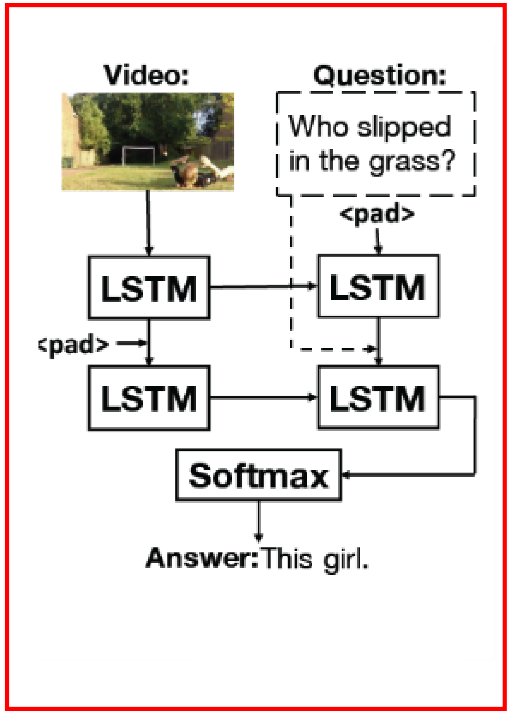
(a) Extended end-to-end memory network



(b) Extended simple VQA model



(c) Extended temporal attention model



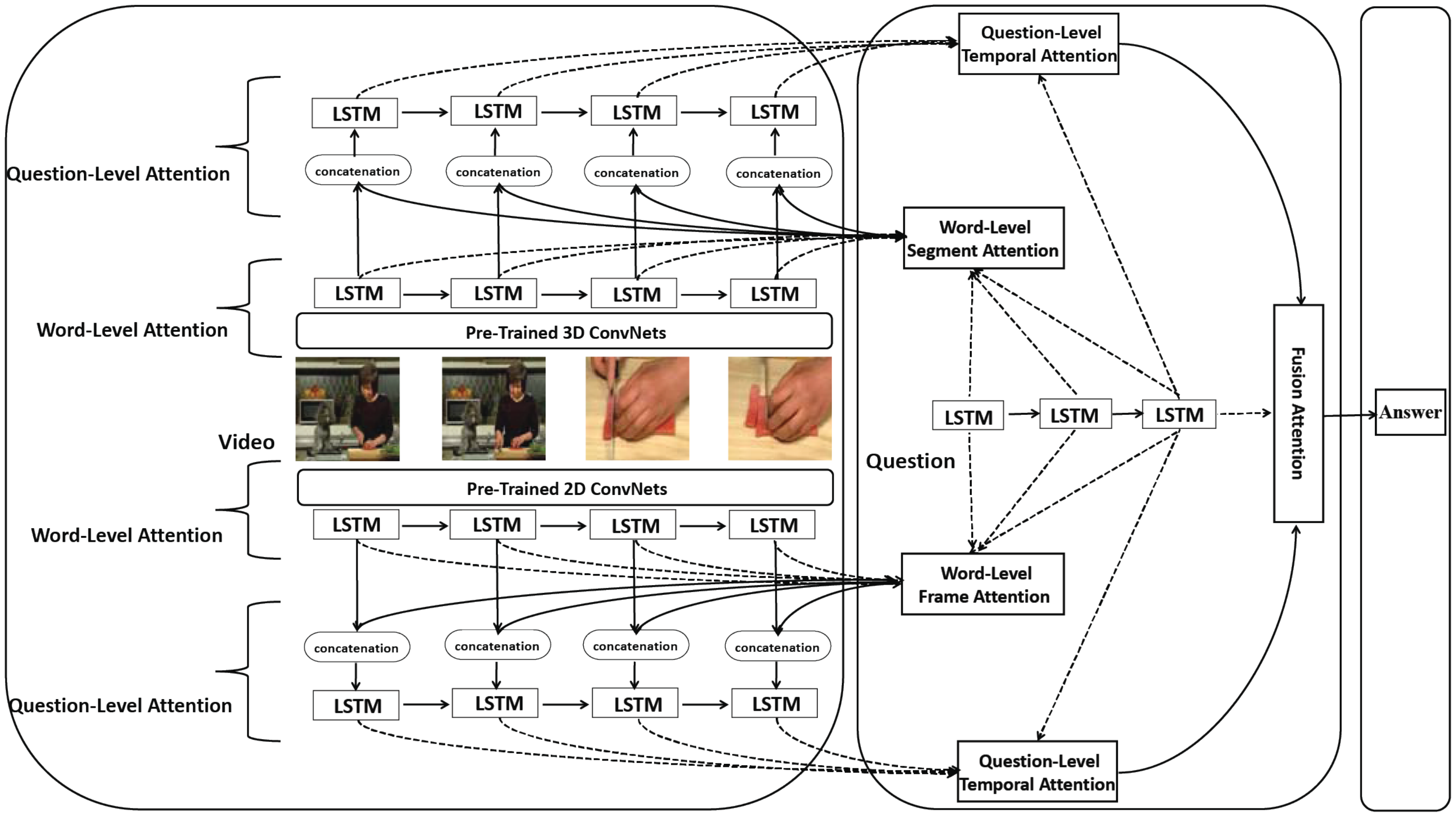
(d) Extended sequence-to-sequence model

Zeng, Kuo-Hao, et al. "Leveraging video descriptions to learn video question answering." *AAAI'17*.

Spatio-temporal cross-modality alignment

Key ideas:

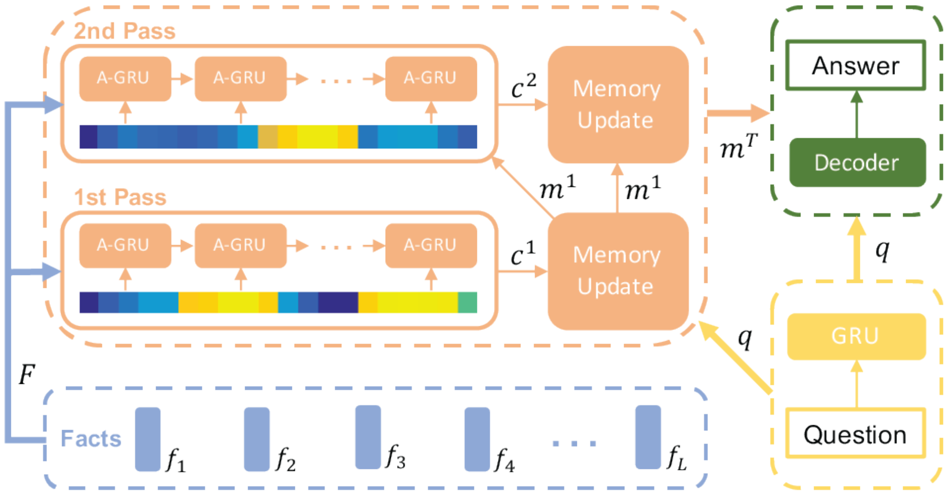
- Explore the correlation between vision and language via attention mechanisms.
- Joint representations are query-driven spatio-temporal features of a given videos.



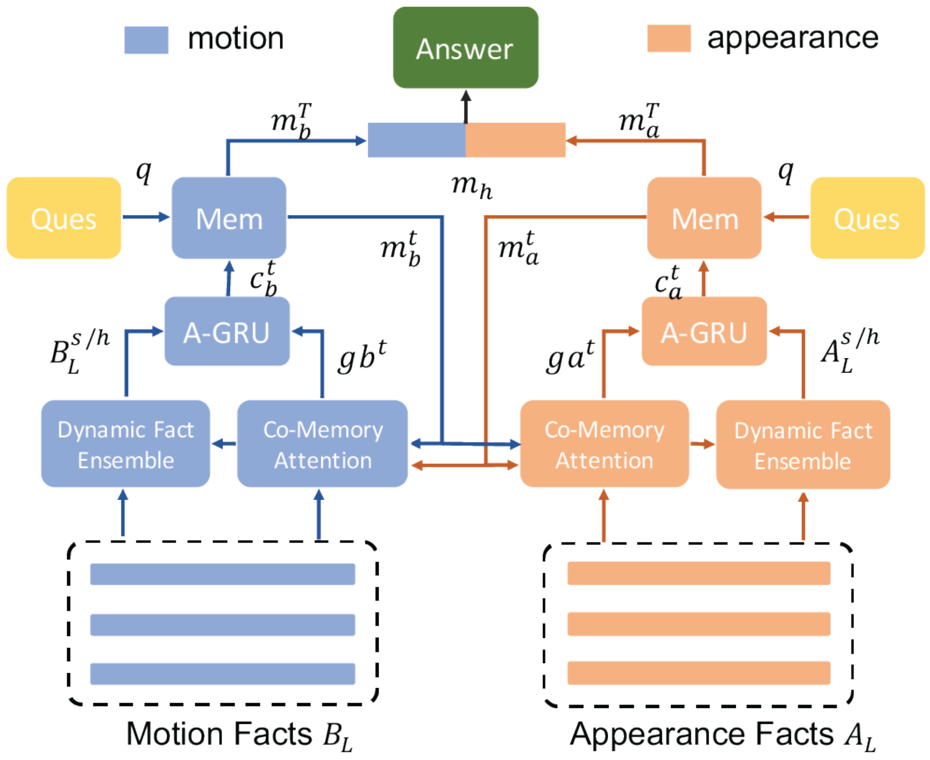
(a) Dual-Level Video Feature Extraction (b) Hierarchical Dual-Level Attention (c) Answer Prediction

Zhao, Zhou, et al. "Video question answering via hierarchical dual-level attention network learning." *ACL'*17.

Memory-based Video QA



General Dynamic Memory Network (DMN)



Co-memory attention networks for Video QA

Key ideas:

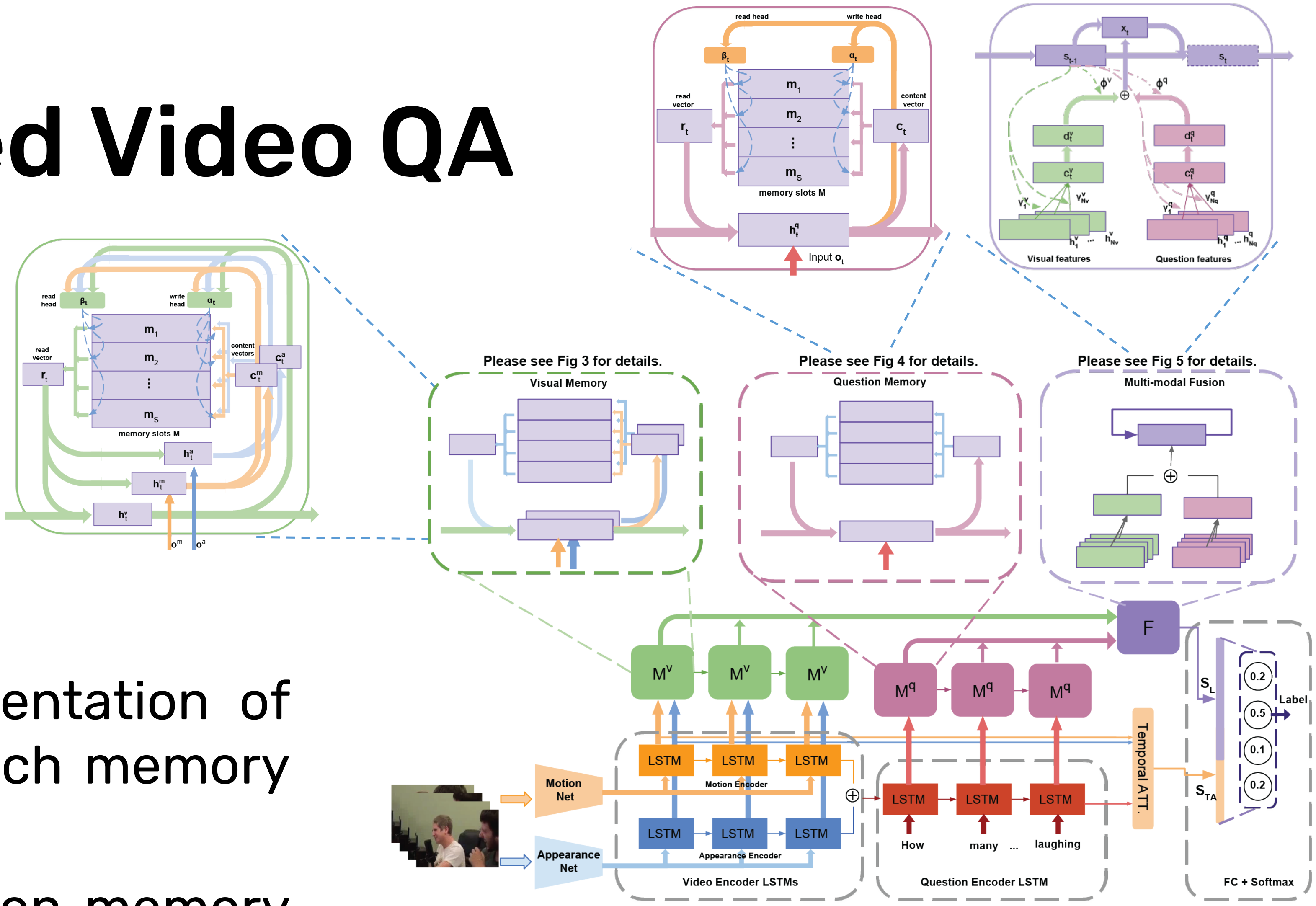
- DMN refines attention over a set of facts to extract reasoning clues.
- Motion and appearance features are complementary clues for question answering.

Gao, Jiyang, et al. "Motion-appearance co-memory networks for video question answering." *CVPR'18*.

Memory-based Video QA

Key differences:

- Learning a joint representation of multimodal inputs at each memory read/write step.
- Utilizing external question memory to model context-dependent question words.

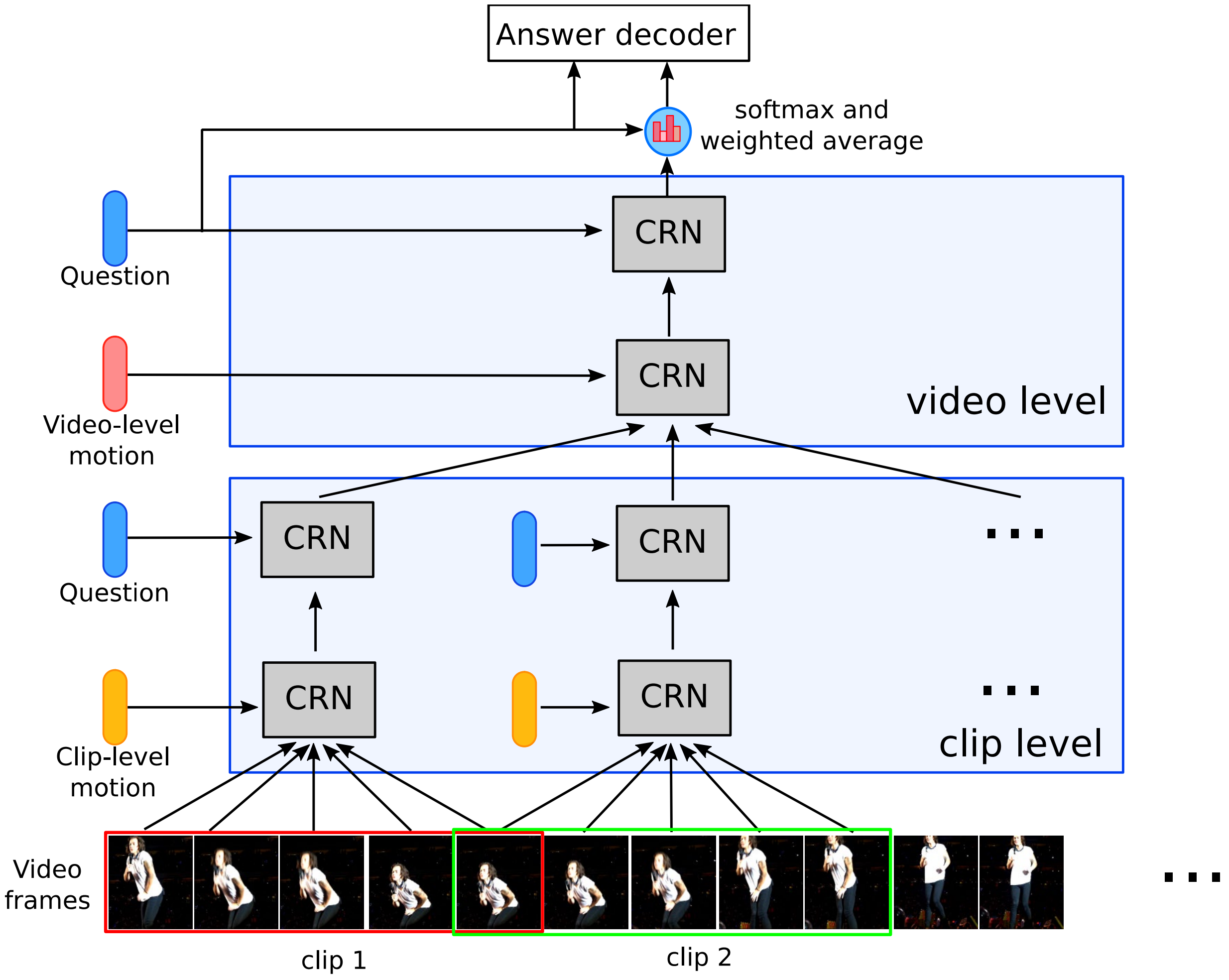


Heterogeneous video memory for Video QA

Fan, Chenyou, et al. "Heterogeneous memory enhanced multimodal attention model for video question answering." *CVPR’19.*

Multimodal reasoning units for Video QA

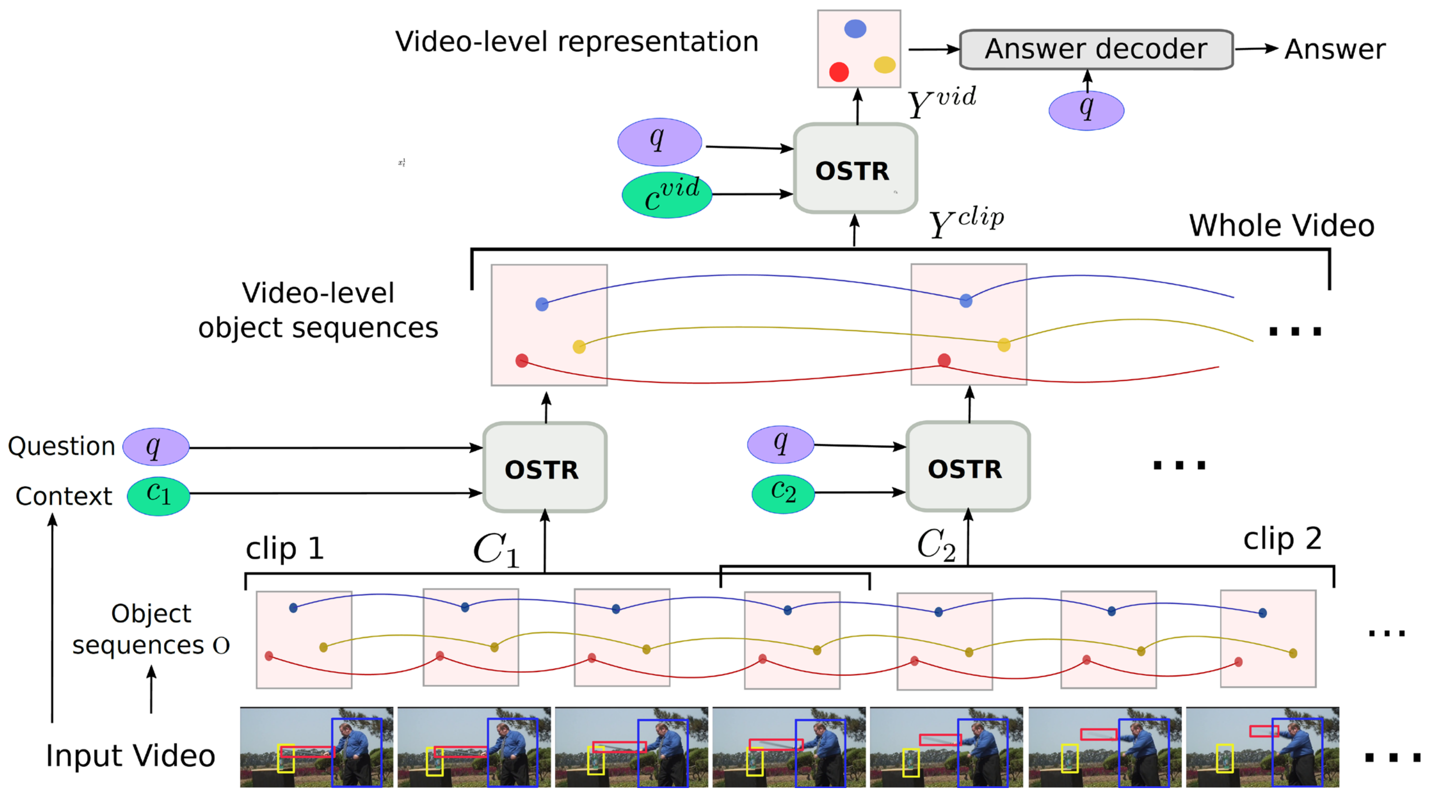
- CRN: Conditional Relation Networks.
- Inputs:
    - Frame-based appearance features
    - Motion features
    - Query features
- Outputs:
    - Joint representations encoding temporal relations, motion, query.



Le, Thao Minh, et al. "Hierarchical conditional relation networks for video question answering." CVPR'20

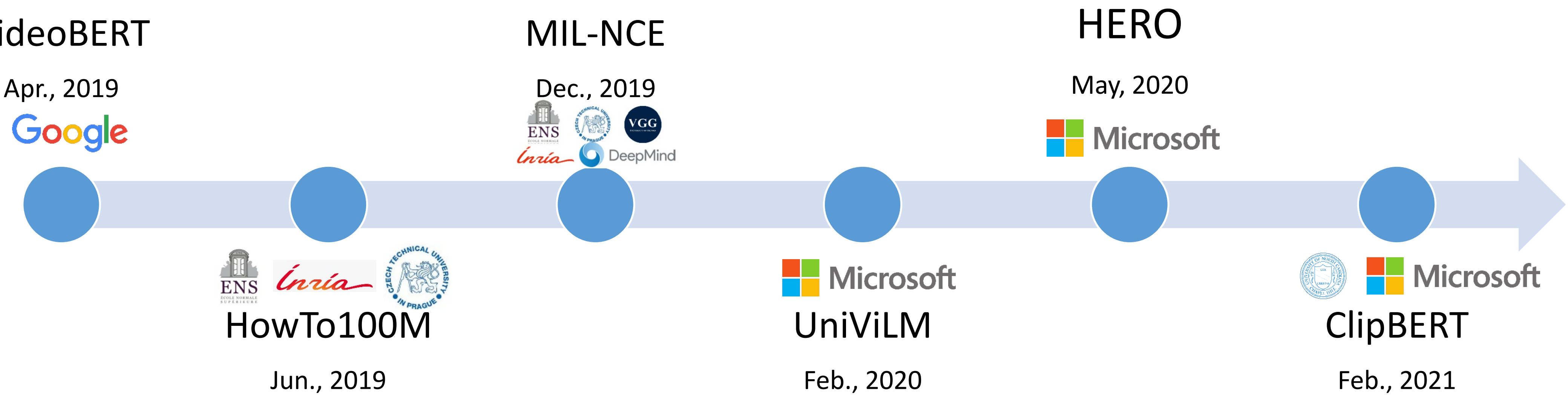
Object-oriented spatio-temporal reasoning for Video QA

- OSTR: Object-oriented Spatio-Temporal Reasoning.
- Inputs:
  - Object lives tracked through time.
  - Context (motion).
  - Query features.
- Outputs:
  - Joint representations encoding temporal relations, motion, query.



Dang, Long Hoang, et al. "Hierarchical Object-oriented Spatio-Temporal Reasoning for Video Question Answering." *IJCAI'21*

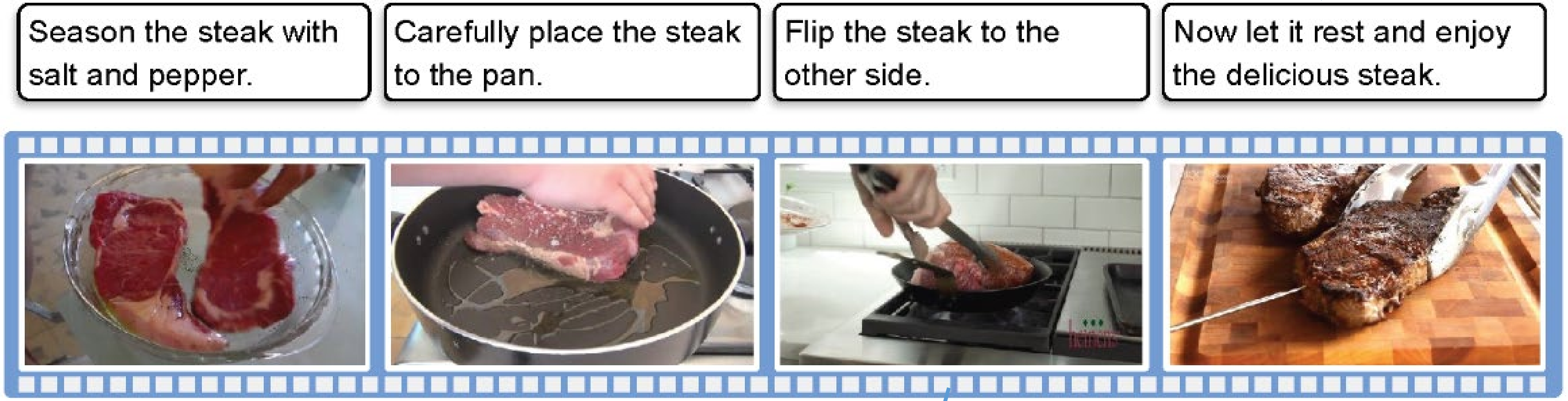
Video QA as a down-stream task of video language pre-training



VideoBERT: a joint model for video and language representation learning

- Data for training: Sample videos and texts from YouCook II.

Instructions in text given by ASR toolkit

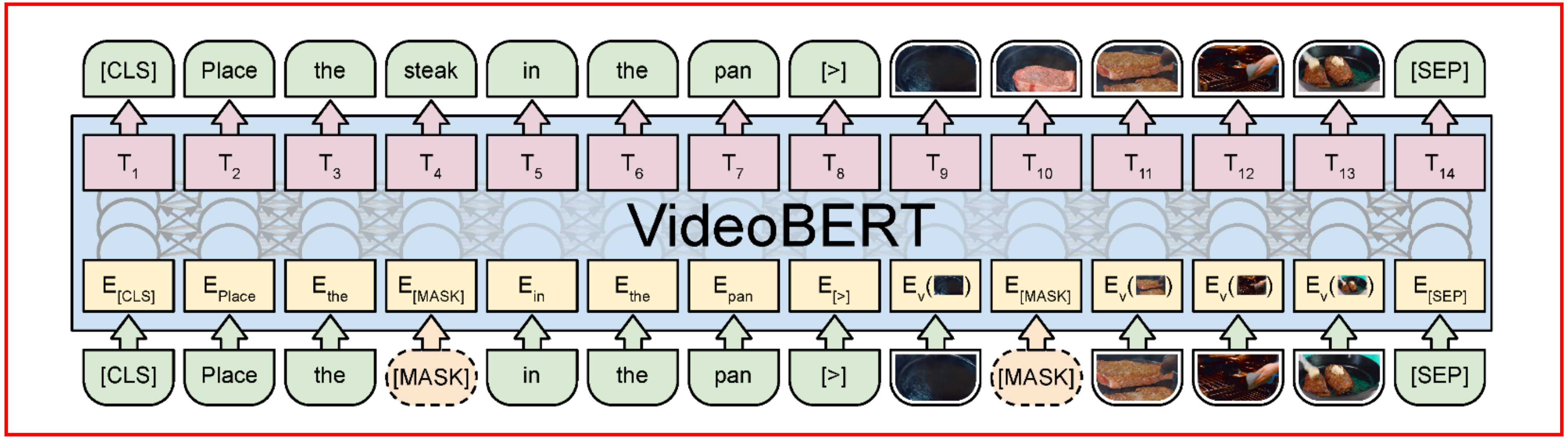


Subsampled video segments

Sun, Chen, et al. "Videobert: A joint model for video and language representation learning.“ ICCV’19.

VideoBERT: a joint model for video and language representation learning

Pre-training

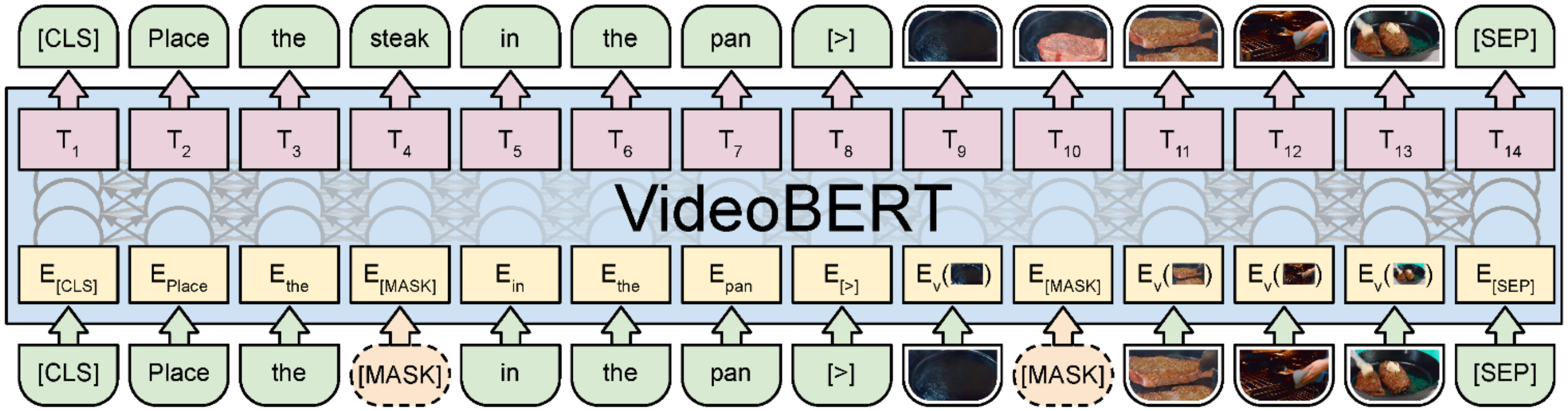


- Linguistic representations:
  - Tokenized texts into WordPieces, similar as BERT.
- Visual representations:
  - S3D features for each segmented video clips.
  - Tokenized into clusters using hierarchical k-means.

Sun, Chen, et al. "Videobert: A joint model for video and language representation learning." ICCV'19.

VideoBERT: a joint model for video and language representation learning

Pre-training

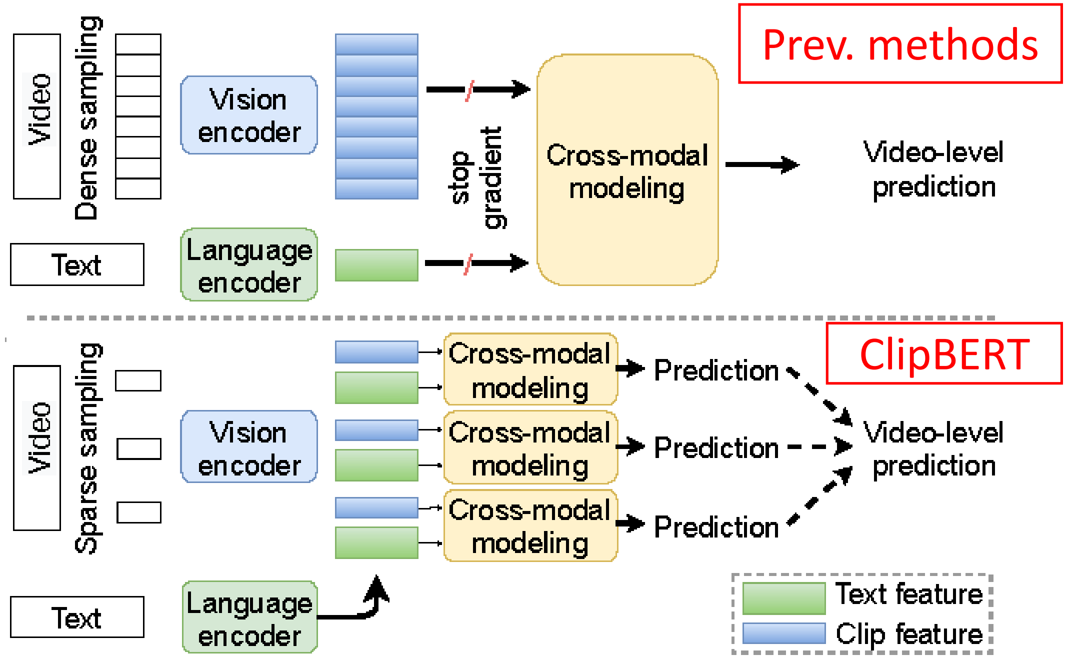


Down-stream tasks

- Video captioning
- Video question answering
- Zero-shot action classification

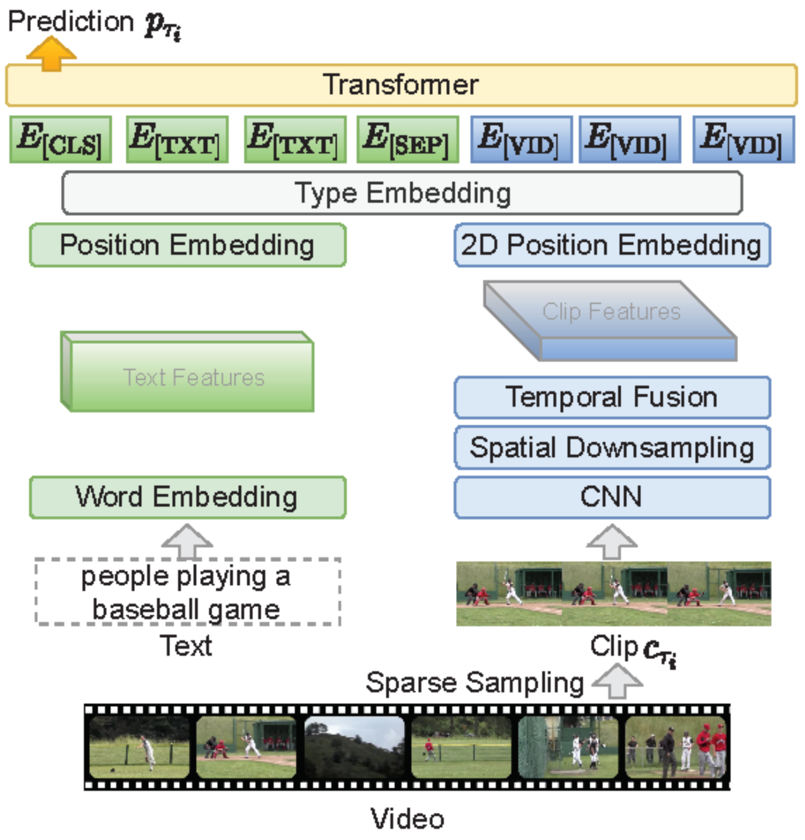
Sun, Chen, et al. "Videobert: A joint model for video and language representation learning." ICCV'19.

CLIPBERT: video language pre-training with sparse sampling



Procedure:

- Pretraining on large-scale image-text datasets.
- Finetuning on video-text tasks.



ClipBERT overview

Lei, Jie, et al. "Less is more: Clipbert for video-and-language learning via sparse sampling." *CVPR'21.*

From short-form Video QA to Movie QA

Long-term temporal relationships

Multimodal inputs

Subtitle:

00:00:0.395 --> 00:00:1.896 (Keith:) I'm not gonna stand here and let you accuse me	00:00:1.897 --> 00:00:4.210 (Keith:) of killing one of my best friends, all right?	00:00:8.851 --> 00:00:10.394 (Castle:) You hear that sound?

Question: What did Keith do when he was on the stage?

Choice 1: Keith drank beer
Choice 2: Keith played drum
Choice 3: Keith sing to the microphone
Choice 4: Keith played guitar
Choice 5: Keith got off the stage and walked out

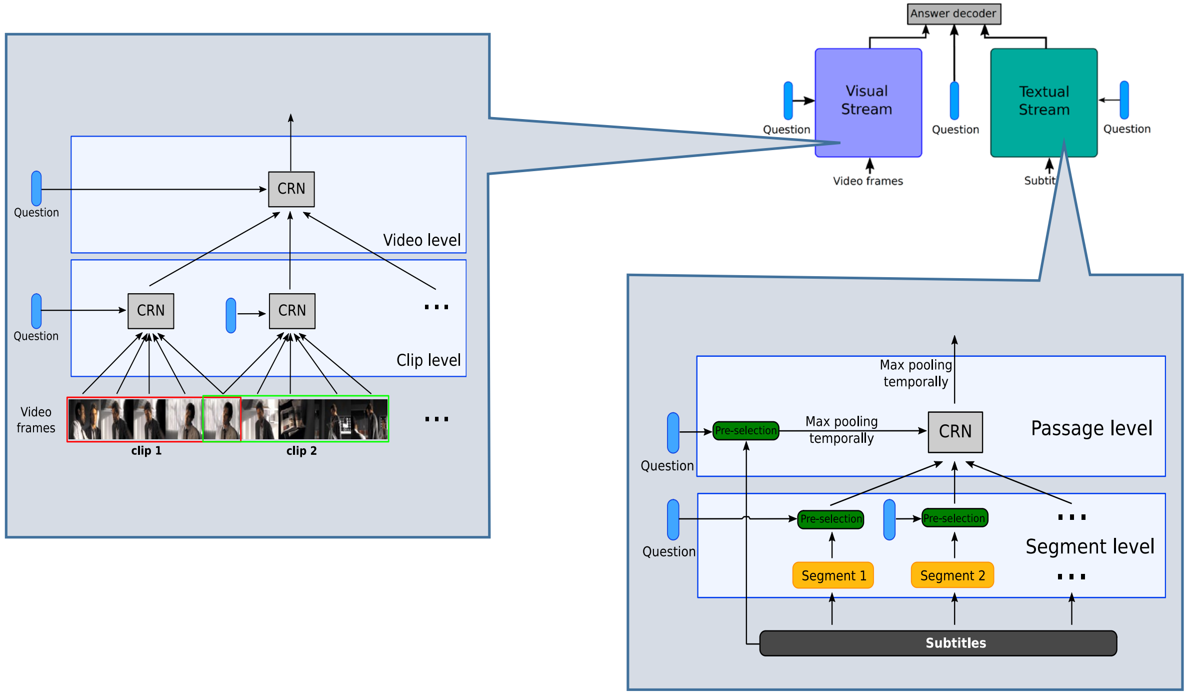
Baseline: Keith played guitar
HCRN: Keith got off the stage and walked out
Ground truth: Keith got off the stage and walked out

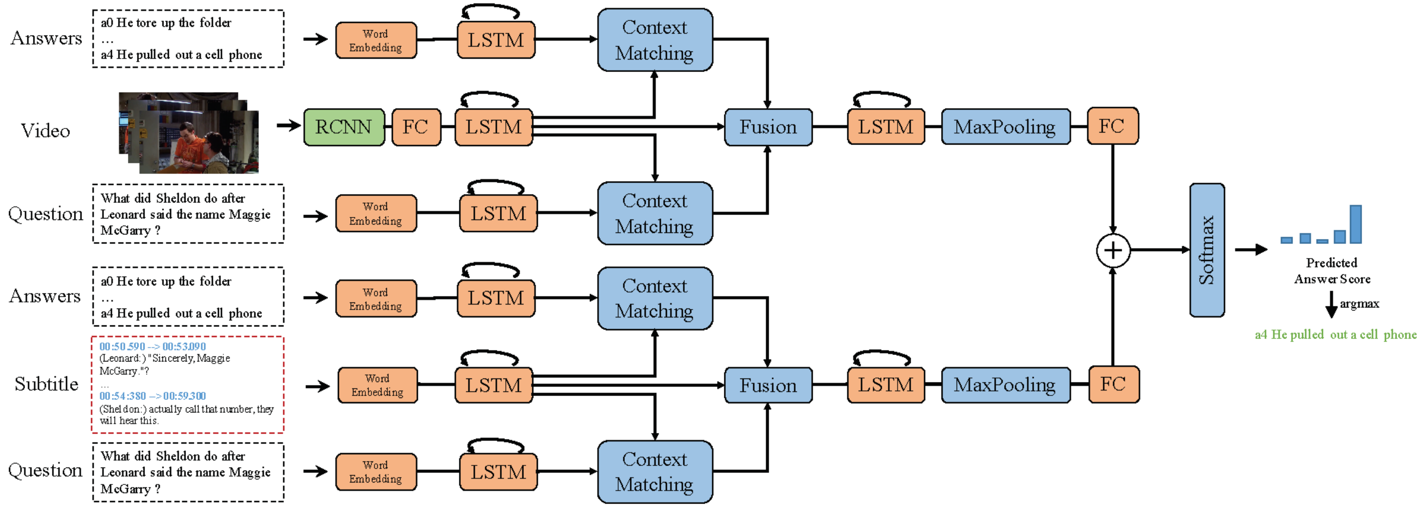
Lei, Jie, et al. "Tvqa: Localized, compositional video question answering." *EMNLP'*18.

Conventional methods for Movie QA

Question-driven multi-stream models:

- Short-term temporal relationships are less important.
- Long-term temporal relationships and multimodal interactions are key.



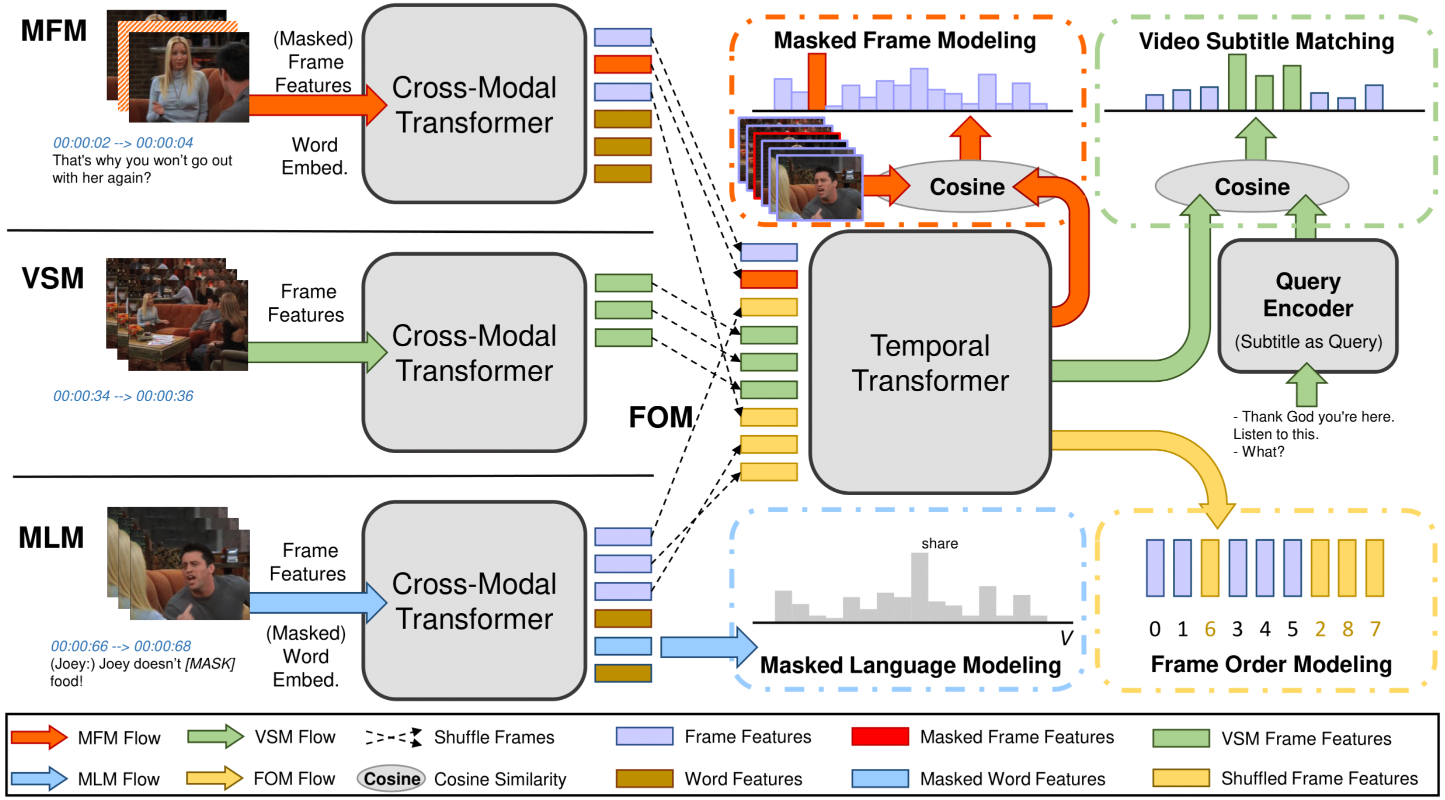


Le, Thao Minh, et al. "Hierarchical conditional relation networks for video question answering." IJCV'21.

Lei, Jie, et al. "Tvqa: Localized, compositional video question answering." *EMNLP*'18.

HERO: large-scale pre-training for Movie QA

- Pre-trained on 7.6M videos and associated subtitles.
- Achieved state-of-the-art results on all datasets.



Method \Task	TVR			How2R			TVQA	How2QA	VIOLIN	TVC			
	R@1	R@10	R@100	R@1	R@10	R@100	Acc.	Acc.	Acc.	Bleu	Rouge-L	Meteor	Cider
SOTA Baseline	3.25	13.41	30.52	2.06	8.96	13.27	70.23	-	67.84	10.87	32.81	16.91	45.38
HERO	**6.21**	**19.34**	**36.66**	**3.85**	**12.73**	**21.06**	**73.61**	**73.81**	**68.59**	**12.35**	**34.16**	**17.64**	**49.98**

Li, Linjie, et al. "Hero: Hierarchical encoder for video+ language omni-representation pre-training." *EMNLP*'20.

End of part B

https://bit.ly/37DYQn7