NeurIPS'23 [Shanghai Jiao Tong]

Symbol-LLM:

Leverage Language Models for Symbolic System in Visual Human Activity Reasoning

[Project, Dataset, Code, Supplementary]

2025.05.08, @ CM Lab, Dept. of CSIE, NTU

Presenter: **Shih-Yu Lai**

Outline

- Motivation
  - Contributions
  - Limitations
- Related Works
- Method
  - Framework
  - Definitions
  - Instantiating the Symbolic System
  - Reasoning with Visual Inputs
- Experiment
- Conclusion and Discussion
- Feedbacks
- (Supplementary)

Abstract	0.5 pages
Introduction	1.5 pages
Related Work	0.5 pages
Method	4.25 pages
Experiment	3 pages
Conclusion and Discussion	0.25 pages

Preliminary

Symbol-LLM: Leverage Language Models for Symbolic System in Visual Human Activity Reasoning

Xiaoqian Wu **Yong-Lu Li*** **Jianhua Sun** **Cewu Lu***
Shanghai Jiao Tong University
{enlighten, yonglu_li, gothic, lucewu}@sjtu.edu.cn

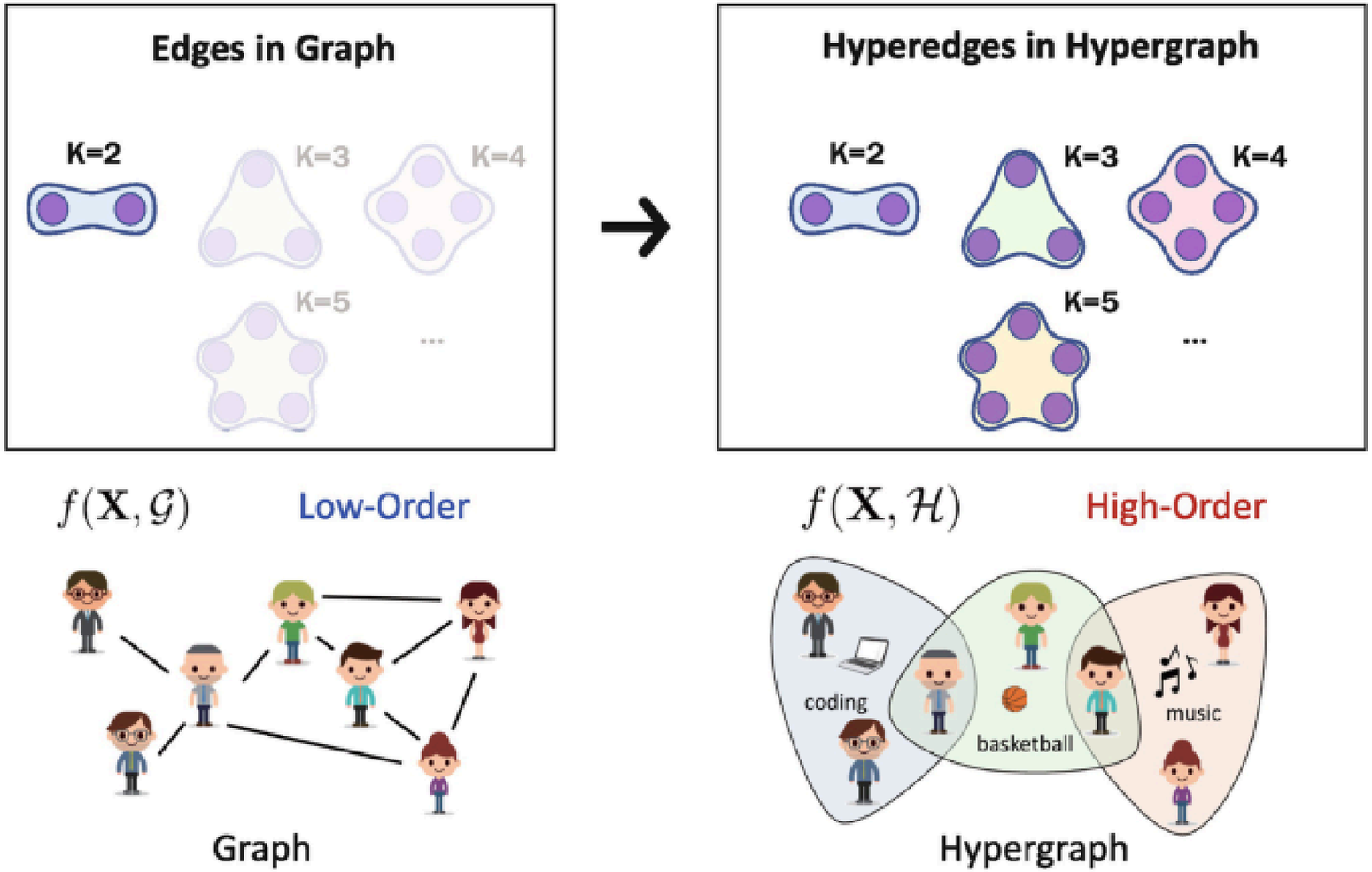
~~  ~~

~~**Symbol-LLM: Towards Foundational Symbol-centric Interface For Large Language Models**~~

~~**Fangzhi Xu[1,2], Zhiyong Wu[2]*, Qiushi Sun[3], Siyu Ren[4], Fei Yuan[2], Shuai Yuan[5], Qika Lin[3], Yu Qiao[2], Jun Liu[1]**~~
~~[1]Xi'an Jiaotong University [2]Shanghai Artificial Intelligence Laboratory~~
~~[3]National University of Singapore [4]Shanghai Jiao Tong University~~
~~[5]Hong Kong University of Science and Technology~~
~~Leo981106@stu.xjtu.edu.cn, wuzhiyong@pjlab.org.cn, qiushisun@u.nus.edu~~

Hypergraph

- A hypergraph is a generalization of a graph in which **an edge can join any number of vertices**.

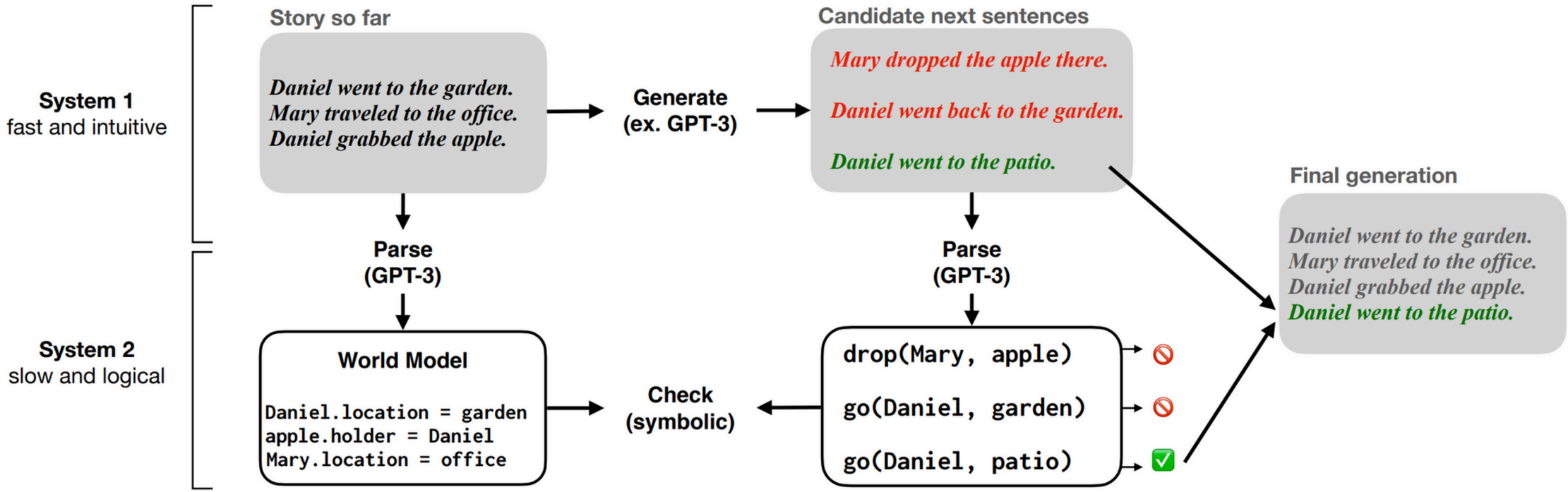


Dai, Q., Gao, Y. (2023). Mathematical Foundations of Hypergraph. In: Hypergraph Computation. Artificial Intelligence: Foundations, Theory, and Algorithms. Springer, Singapore. https://doi.org/10.1007/978-981-99-0185-2_2

賴澍雨 Shih-Yu Lai

human reasoning in cognitive systems

- **System-1**: intuitive, associative
- **System-2**: deliberative, logical



- **logical program**

[**Dual-System**, Nye et al. , NeurIPS 2021]

Motivation

System-2 reasoning: knowledge + rules

- Human intelligence derives the answer from **reasoning in a physical symbolic system to avoid intuitive errors**.
- As a critical component in building AI systems, **visual activity understanding** urgently needs System-2 reasoning.
- **Symbol-LLM:**
  - Human knowledge: the symbols should have **broad semantic coverage** to express various activity contexts.
  - Reasonable rules: **rational and unambiguous**.
  - **symbol-rule loop** and **entailment check strategy.**
- Given an image, **symbols** are first extracted from **visual inputs**,
  then **activity semantics** are deduced via **fuzzy logic** calculation based on **rules**.

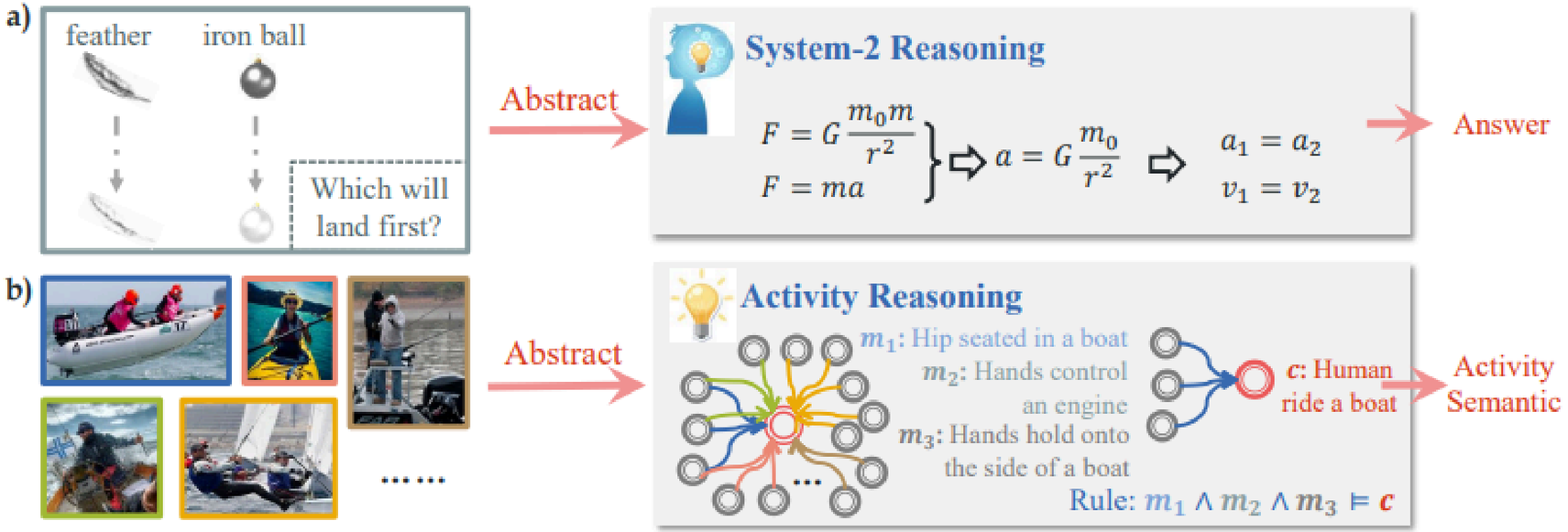
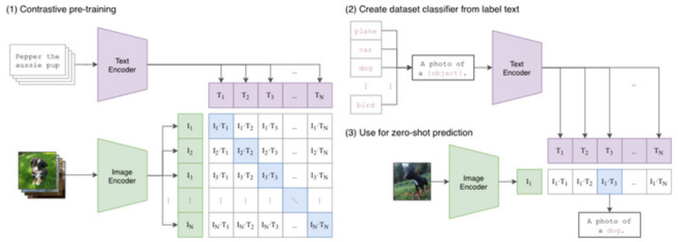


Figure 1: System-2 reasoning in **a)** solving physical questions and **b)** understanding human activities. They both involve a symbolic system implying human commonsense.

visual activity understanding with System-1 reasoning

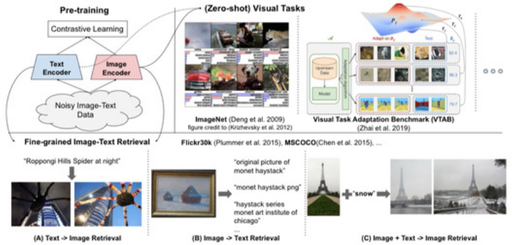
- Mapping between visual inputs and natural language.
  - **System-1** methods **depend on large-scale visual data**, thus suffering from diminishing marginal utility.
  - It **fails in explainability** and **generalization** due to **black-box computing**.
  - Thus, it is important to **integrate System-2 reasoning**.

contrastive learning + zero-shot



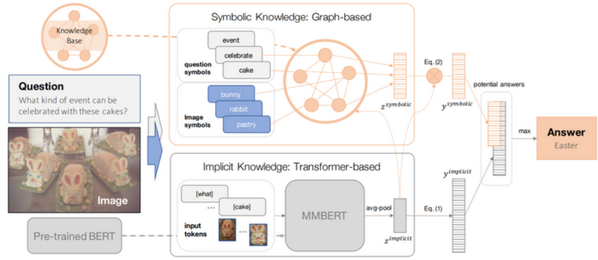
[**CLIP**, Radford et al., ICML 2021]

multi-modal alignment w/ high-quality data



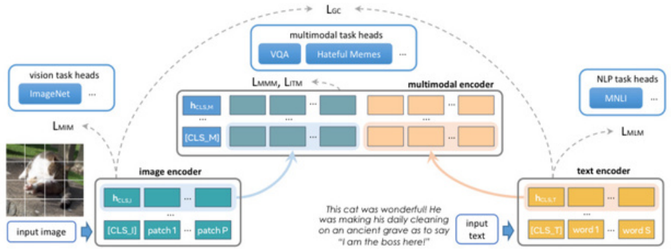
[**ALIGN**, Jia et al., ICML 2021]

non-symbolic form knowledge embed. (weights)



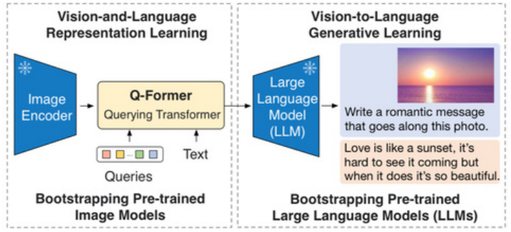
[**KRISP**, Marino et al. , CVPR 2021]

Unified Transformer



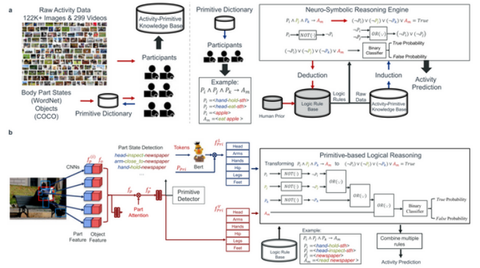
[**Flava**, Singh et al., CVPR 2022]

Frozen Models + Q-Former

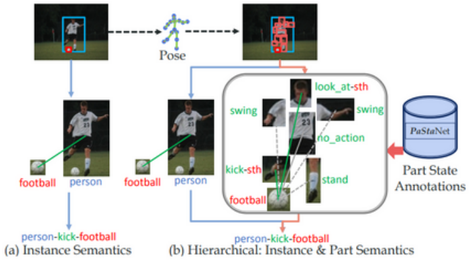


[**Blip-2**, Li et al., ICML 2023]

human body activity reasoning



[**HAKE**, Li et al., TPAMI 2022]



[**Pastanet**, Li et al., CVPR 2022]

Contributions

- Point out that the current System-2 reasoning suffers from a defective symbolic system with hand-crafted symbols and limited, ambiguous rules.
  - To overcome the defects, the authors propose a novel **symbolic system with broad-coverage symbols and rational rules.**

- Propose Symbol-LLM to **instantiate** it and show how it helps **visual reasoning**.

- In extensive **activity understanding tasks**, the method shows superiority in **explainability**, **generalization**, and **data efficiency.**

Limitations

- Rules generated from language models mostly boost System-1, but sometimes bring **language bias**
  - e.g., the person only decides to buy an orange despite apples in the background.

- The approximation is **mainly based on pre-trained LLMs** and can be improved with more elaborate designs
  - e.g., human-in-the-loop, customized LLMs with higher-quality knowledge.

- **System-2 reasoning** may be boosted if **integrated into the System-1 training process**.

- **Instance-level activities** and **temporal encoding** are beyond our scope since the main focus is on an overall framework to formulate and construct a novel symbolic system.

Related works

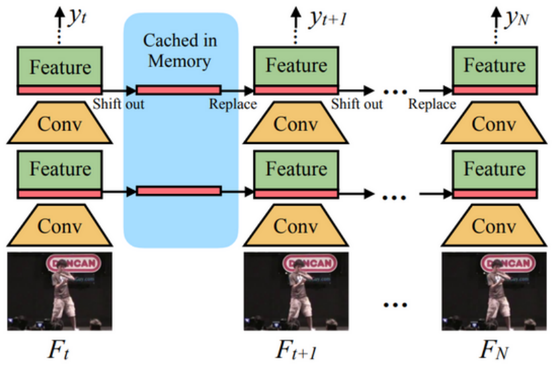
Related works (1/2)

1. Activity Understanding

- **complex visual patterns** and **long-tail data distribution**

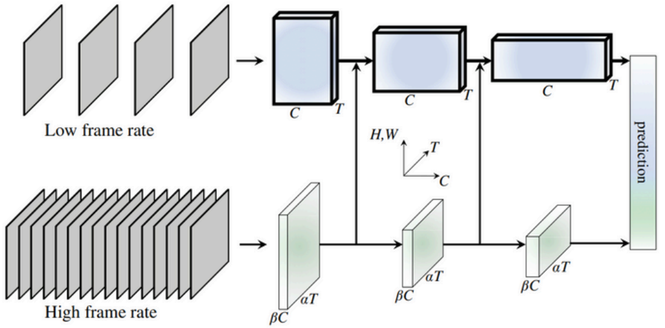
- methods:
  - ✅**image**
  - video
  - skeleton

Video Understanding



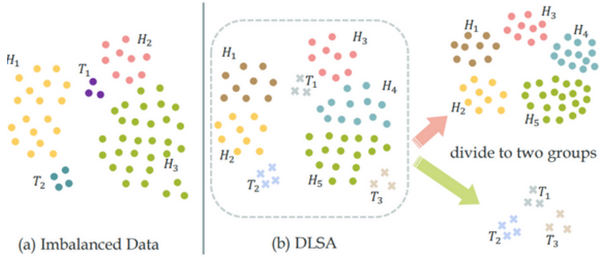
[**TSM**, Lin et al. ICCV 2019]

Video Recognition



[**SlowFast**, Feichtenhofer et al. ICCV 2019]

Imbalance for Long-tailed Image Recognition



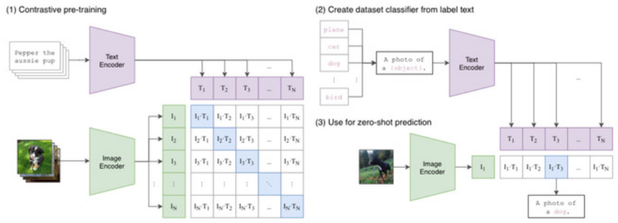
[**DLSA**, Xu et al. ECCV 2022]

2. Vision-Language Models

- Trained with **web-scale image-text pairs**

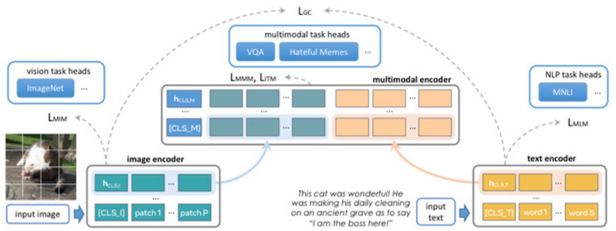
- zero-shot predictions on various recognition tasks, including human activities.

contrastive learning + zero-shot



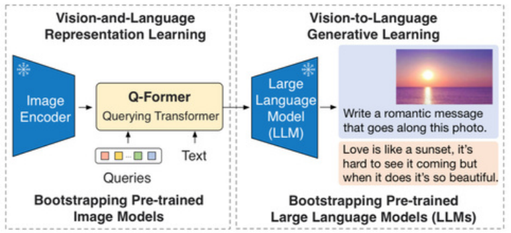
[**CLIP**, Radford et al., ICML 2021]

Unified Transformer



[**Flava**, Singh et al., CVPR 2022]

Frozen Models + Q-Former

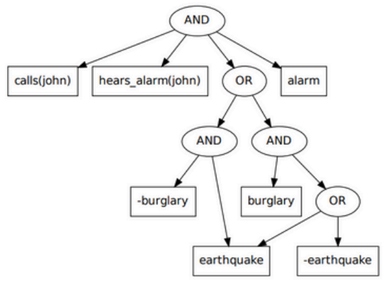


[**Blip-2**, Li et al., ICML 2023]

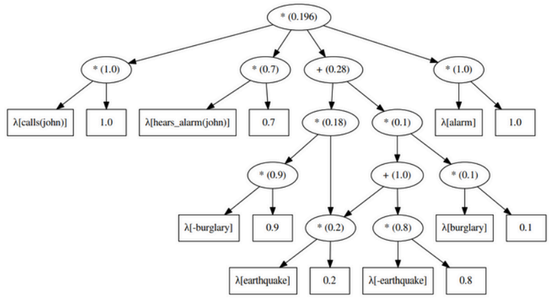
Related works (2/2)

3. Probabilistic Programming

- Follow the principles of the standard **first-order logic (FoL)**
  - e.g.,logical connectives $\wedge$ $\vee$.
- In vision reasoning tasks, the symbols are **not known to be True/False**.
  - **Predict the symbol probabilities**: neuro-symbolic



(a) d-DNNF



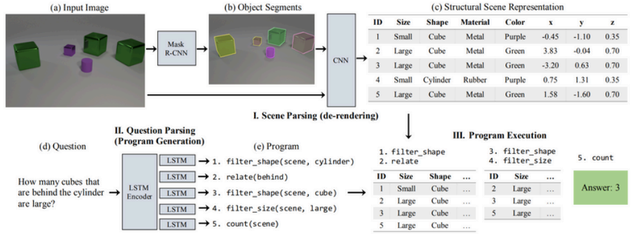
(b) arithmetic circuit

[Fierens et al., Theory and Practice of Logic Programming 2015]

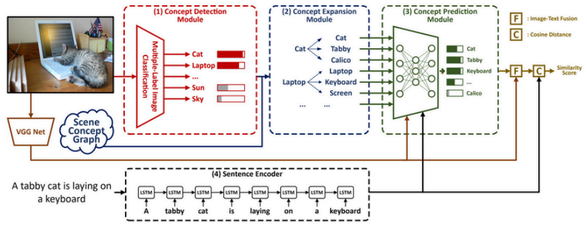
4. Neuro-Symbolic Reasoning

- Provides a way of combining System-1 and System-2 learning,
- where **knowledge** is represented in **symbolic form** and **learning and reasoning** are computed by a **neural network**.

visual question answering

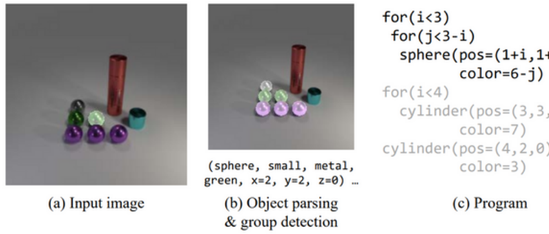


[**Neural-Symbolic VQA**, Yi et al., NeurIPS 2018]



[**KASCE**, Shi et al., IJCAI 2019]

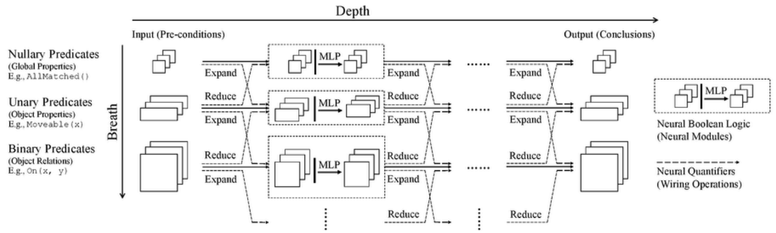
scene understanding



(a) Input image (b) Object parsing & group detection (c) Program

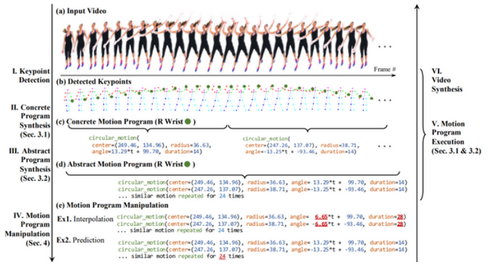
[**Learning to Describe Scenes with Programs**, Liu et al., ICLR 2019]

decision-making



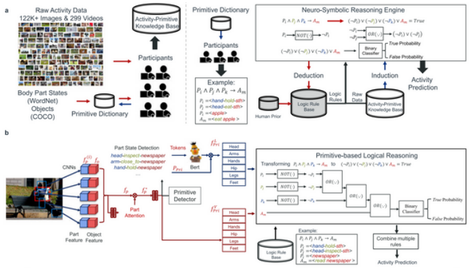
[**Neural logic machines**, Dong et al., ICML 2019]

motion programming



[**motion2prog**, Kulal et al., CVPR 2021]

human body activity reasoning



[**HAKE**, Li et al., TPAMI 2022]

Method

Framework

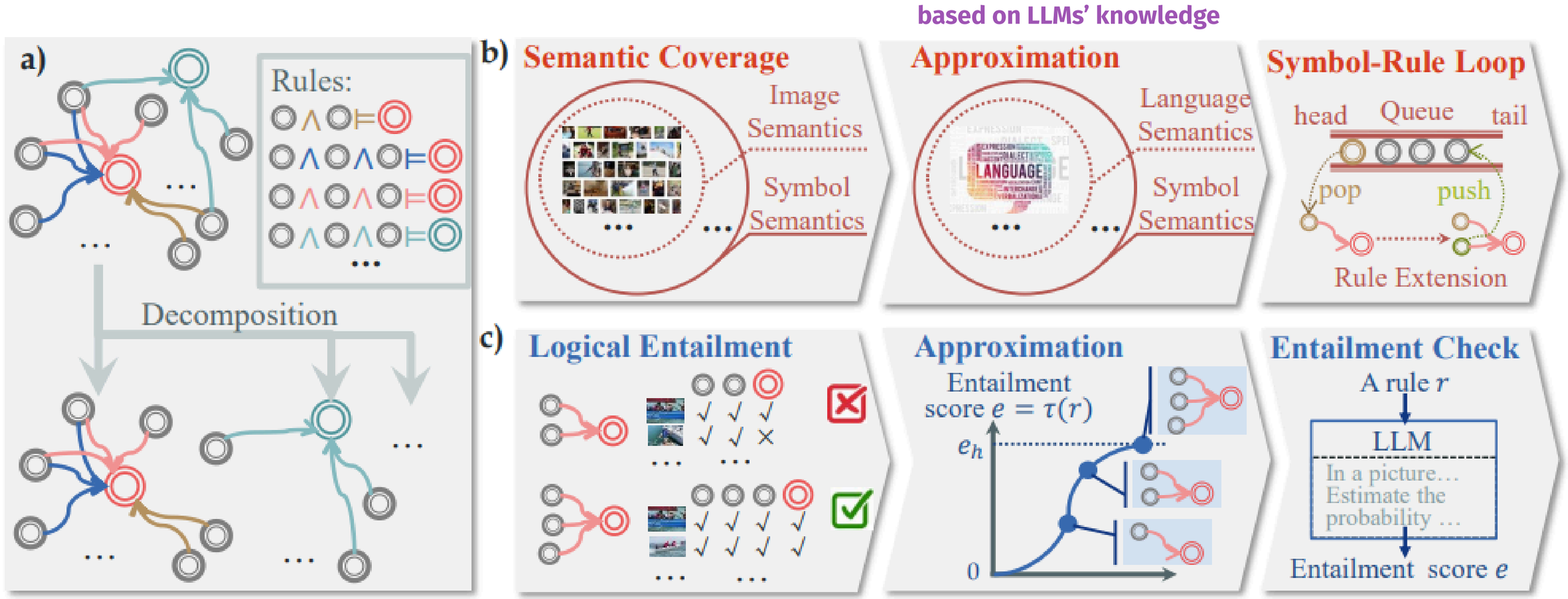


Figure 2: Our activity symbolic system. **a)** Structure and decomposition of the symbolic system (Sec. 3.1.1). **b)** Semantic coverage (Sec. 3.1.2). It can be approximated based on LLMs' knowledge and achieved via the symbol-rule loop (Sec. 3.2.1). **c)** Logical entailment (Sec. 3.1.2). It can be approximated based on an entailment scoring function and achieved by entailment check (Sec. 3.2.2).

Definitions

Formulating the Symbolic System

Definition 1. (Directed Hypergraph [3]) *A hypergraph is a generalization of a graph in which an edge can join any number of vertices. A directed hypergraph is a pair $(\mathcal{X}, \mathcal{E})$, where $\mathcal{X}$ is a set of vertices, and $\mathcal{E}$ is a set of pairs of subsets of $\mathcal{X}$. Each of these pairs $(D, C) \in \mathcal{E}$ is a hyperedge; the vertex subset D is its domain, and C is its codomain.*

Definition 2. ($\mathcal{B}$-Graph [12]) *A $\mathcal{B}$-graph is a type of directed hypergraph with only $\mathcal{B}$-arcs. A $\mathcal{B}$-arch is a type of a hyperedge that is directed to a single head vertex, and away from all its other vertices.*

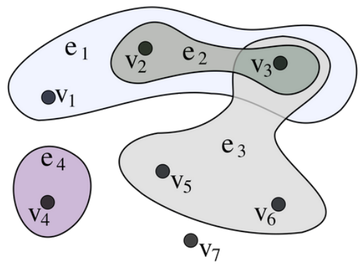
Definition 3. (Activity Symbolic System) *The activity symbolic system $(\mathcal{S}, \mathcal{R})$ is a $\mathcal{B}$-Graph, where $\mathcal{S}$ is a set of vertices/symbols, and $\mathcal{R}$ is a set of pairs of subsets of $\mathcal{S}$. Each of these pairs $r = (M, C) \in \mathcal{R}$ is a hyperedge/rule; the vertex subset $M = \{m_i\}_{i=1}^{n_r}$ is its domain/premises, and $C = \{c\}$ is its codomain/conclusion. Equivalently, a rule takes the form $r : \bigwedge_{m \in M} m \models c$. Since $|C| = 1$, r is a $\mathcal{B}$-arch.*

Definition 4. (Decomposition of Activity Symbolic System) *One conclusion c^* corresponds to one symbolic sub-system $(\mathcal{S}_{c^*}, \mathcal{R}_{c^*})$, where $\mathcal{S}_{c^*} \subseteq \mathcal{S}$, $\mathcal{R}_{c^*} \subseteq \mathcal{R}$. $\forall r = (M, C) \in \mathcal{R}_{c^*}$, $C = \{c^*\}$. $\forall s \in \mathcal{S}_{c^*} \backslash \{c^*\}$, $\exists\, r = (M, C) \in \mathcal{R}_{c^*}$, $s \in M$.*

Def. 3,4 is depicted in Fig. 2. Def. 3 is based on Def. 1,2 and analysis above. In applications, we typically judge a specific conclusion with other symbols/rules removed. It is achieved by decomposing the activity symbolic system (graph) into sub-systems (sub-graphs) in Def. 4.

Formulating the Symbolic System

- Each conclusion corresponds to **more than one rule** with varied combinations of premise symbols
  - because one activity typically has **different visual patterns**.
- **Multiple rules** for one conclusion are logically connected $\vee$.
- **Hyper-graph**: handling the **complexity of the connection**.

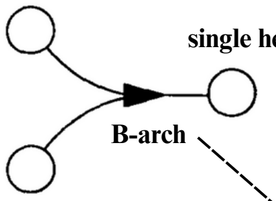


$v_i \in \mathcal{X}$ $e_j(D, C) \in \mathcal{E}$

Directed Hypergraph

rule $r : m_1 \wedge m_2 \wedge \ldots \wedge m_{n_r} \models c,$

where { premise set $M = \{m_i\}_{i=1}^{n_r}$; domain; conclusion c; codomain



B-Graph

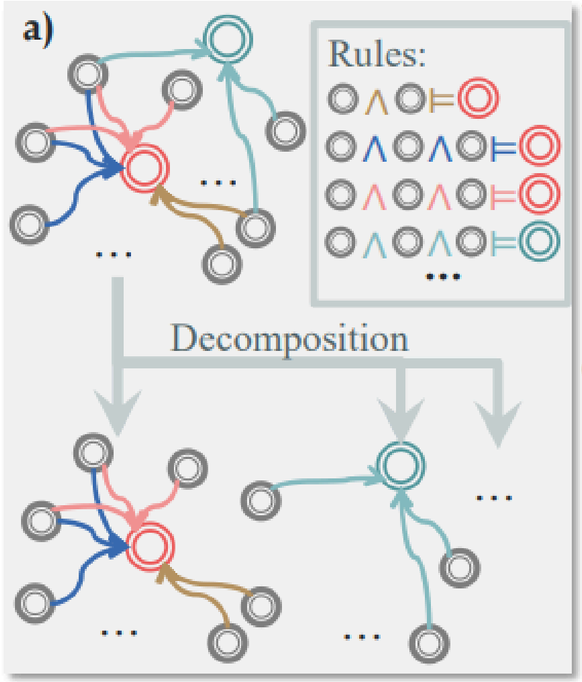
activity symbolic system $(\mathcal{S}, \mathcal{R})$

$r = (M, C) \in \mathcal{R}$ rule (hyperedge) $\longrightarrow r : \bigwedge_{m \in M} m \models c.\ |C| = 1$

$M = \{m_i\}_{i=1}^{n_r}$ premises(vertex)

$C = \{c\}$ conclusion

Activity Symbolic System



symbolic sub-system $(\mathcal{S}_{c^*}, \mathcal{R}_{c^*})$ $\longrightarrow$ *One conclusion* c^*

where $\mathcal{S}_{c^*} \subseteq \mathcal{S},\ \mathcal{R}_{c^*} \subseteq \mathcal{R}.\ \forall r = (M, C) \in \mathcal{R}_{c^*}$

Decomposition

Two Ideal Properties

Fail to **cover the complex patterns of activities** and lack **compositional generalization**.

1. Symbols should have broad **semantic coverage** to express different conditions in the activity image database.
2. Rules should satisfy **logical entailment** to add rationality and avoid ambiguity.
   i.e., the premises set should lead to the conclusion without exception.

Definition 5. **(Semantic Coverage of Activity Symbolic System)** *Given a very large-scale activity images database* $\mathcal{D} = \{(I, A_I, \mathcal{S}_I)\}$ *(I: image,* A_I*: ground-truth activities happening in* I*,* $\mathcal{S}_I$ *: ground-truth symbols happening in* I*), then* $\forall (I, A_I, \mathcal{S}_I) \in \mathcal{D}, \forall s \in \mathcal{S}_I, s \in \mathcal{S}$.

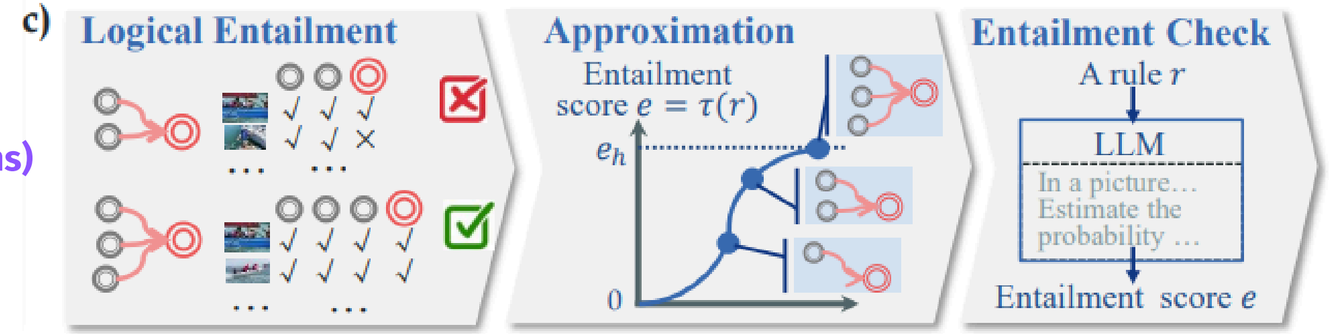
Definition 6. **(Logical Entailment of Activity Symbolic System)** $\forall r = (M, C) \in \mathcal{R}$, $M = \{m_i\}_{i=1}^{n_r}$, *we have:*

1. $\forall (I, A_I, \mathcal{S}_I) \in \mathcal{D}$, *if* $M \subset \mathcal{S}_I$, *then* $c \in A_I$; **rationality**
2. $\forall 1 \leq i \leq n_r, \exists (I, A_I, \mathcal{S}_I) \in \mathcal{D}, M \backslash \{m_i\} \subset \mathcal{S}_I$, *but* $c \notin A_I$. **exception**

large-scale activity images database

entailment :
mapping image datasets (symbols, activities) into rules (premises, conclusions)

$$if\ M \subset \mathcal{S}_I,\ then\ c \in A_I$$

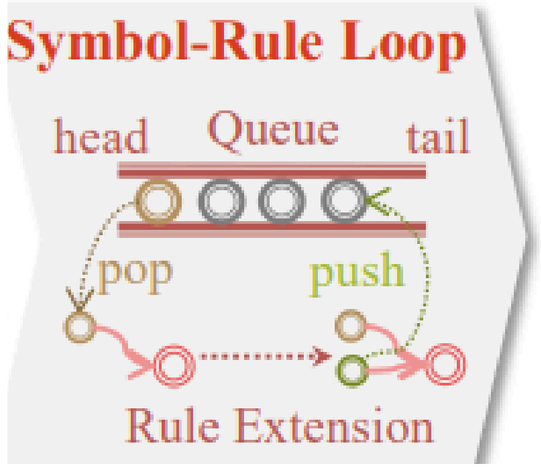


Instantiating the Symbolic System

Solving Semantic Coverage

- To instantiate the symbolic system, it is **expensive to collect massive human knowledge** via manual annotation.
- The target domain can be replaced with the semantic coverage of **pretrained LLMs**:
  - Reshaped the acquisition of human knowledge, as **natural language** carrying **commonsense knowledge.**
  - shown impressive **language reasoning** capabilities
- Question Instruction is ambiguous for LLMs to generate satisfying answers.
- Also, it is **costly** to generate **all symbols** and exhaustively query their **connections**.
- **Approximation !**

Definition 7. (Approximation of Semantic Coverage of Activity Symbolic System) *Given an LLM $\mathcal{L}$ and an activity set $\mathcal{A}$, $\forall\ A \in \mathcal{A}$, $\mathcal{L}$ implies a symbol set $\mathcal{S}_A$ as premises of A, then $\forall s \in \mathcal{S}_A, s \in \mathcal{S}$.*



Symbol-Rule Loop.

(Rule Extension)

```
In a picture, IF [<known symbols>] AND [condition] THEN [<activity>].
[condition] is one concise phrase.  The format is "<The person's
hands/arms/hip/legs/feet> <verb> <object>".  What is [condition]?
```

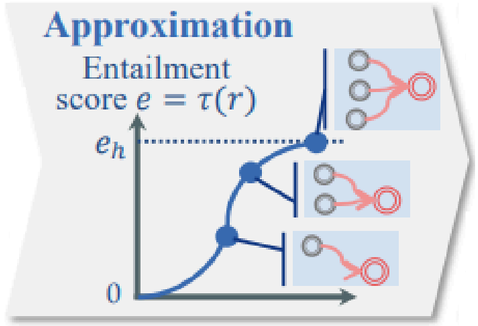
- Then, the answer for "[condition]" is the **extended symbol**.
- Also, a **candidate rule "<known symbols>∧<extended symbol> |= <activity>"** is generated.
- Thus, we get a new extended symbol from this known symbol, and they are connected with rules.
- The **new symbol can be repeatedly used as a known symbol** to generate new rules.

Instantiating the Symbolic System

Solving Logical Entailment

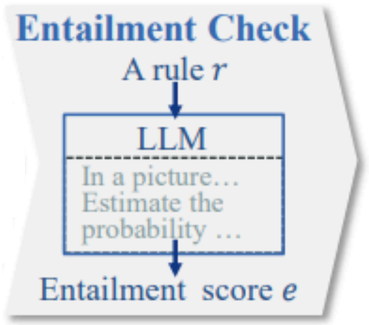
Confident Score!

- It is **costly** to verify the logical entailment of a very large-scale activity image database.
- Develop an **entailment scoring function τ (·)** from an LLM based on its knowledge and language reasoning capability.



Definition 8. **(Approximation of Logical Entailment of Activity Symbolic System)** *Given a function $\tau(\cdot)$ to measure the entailment score of a rule, and an entailment threshold e_h, then $\forall r = (M, C) \in \mathcal{R}$, $M = \{m_i\}_{i=1}^{n_r}$, we have:*

1. $\tau(\bigwedge_{m \in M} m \models c) \geq e_h$;
2. $\forall 1 \leq i \leq n_r$, $\tau(\bigwedge_{m \in M \setminus \{m_i\}} m \models c) < e_h$.



Entailment Check. As is shown in Fig. 2c, to implement the scoring function based on an LLM, we rewrite the rule r as a sentence and design the prompt as:

(Entailment Check)
`In a picture, <symbol 1>, <symbol 2> ... <symbol` n_r`>. Estimate the probability that he is <activity> at the same time. Choose from: (a) 0.1, (b) 0.5, (c) 0.7, (d) 0.9, (e) 0.95, (f) unknown.`

- Then, the output answers can be used as the **entailment score** $e_h = \tau(r)$, whose **credibility depends on the knowledge from the LLM**.
- The **answering text** is **sampled five times** per rule**,** and the **average scores** are taken as the final result to add **stability**.
- In practice, set **e_h = 0.9**.

Summarized Pipeline: Instantiating the Symbolic System

- The symbolic system is merged as $(\mathcal{S}, \mathcal{R}) = (\bigcup_{c \in \mathcal{C}} \mathcal{S}_c, \bigcup_{c \in \mathcal{C}} \mathcal{R}_c)$ from sub-systems $(\mathcal{S}_c, \mathcal{R}_c)$ with a given conclusion $\mathcal{C}$.

Algorithm 1 Instantiating the Symbolic System

Input: conclusion c, entailment threshold e_h
Output: Symbolic Sub-System $(\mathcal{S}_c, \mathcal{R}_c)$

$\mathcal{S}_c^0 \leftarrow$ Symbol_Initialization (c)
$\mathcal{S}_c^{cand}.push(\mathcal{S}_c^0)$ ▷ A queue $\mathcal{S}_c^{cand}$ stores symbols
$\mathcal{S}_c \leftarrow \{c\}, \mathcal{R}_c \leftarrow \{\}$
while not $\mathcal{S}_c^{cand}.is\_empty()$ **do**
    $m_{kno} \leftarrow \mathcal{S}_c^{cand}.pop()$ ▷ A known symbol m_{kno} is taken for rule extension
    $M = \{m_{kno}\}$ ▷ The premises set M is set for the current rule
    while Entailment_Check$(M) < e_h$ **do**
        $m_{new} \leftarrow$ Rule_Extension(M, c)
        $M = M \cup \{m_{new}\}$ ▷ A new symbol m_{new} is added
    end while
    $\mathcal{S}_c \leftarrow \mathcal{S}_c \cup M$ ▷ The symbol set $\mathcal{S}_c$ is updated
    $\mathcal{R}_c \leftarrow \mathcal{R}_c \cup \{r : \bigwedge_{m \in M} m \models c\}$ ▷ The rule set $\mathcal{R}_c$ is updated
    $\mathcal{S}_c^{cand}.push(M \backslash \mathcal{S}_c)$ ▷ M is added to $\mathcal{S}_c^{cand}$ with redundant symbols removed
end while
return $(\mathcal{S}_c, \mathcal{R}_c)$

Reasoning with Visual Inputs

- Visual information is extracted and checked based on the defined symbols and organized as a **probability distribution on the hypergraph**.
- Then, the **activity semantics can be reasoned out** based on the rules and the probability of the symbols.

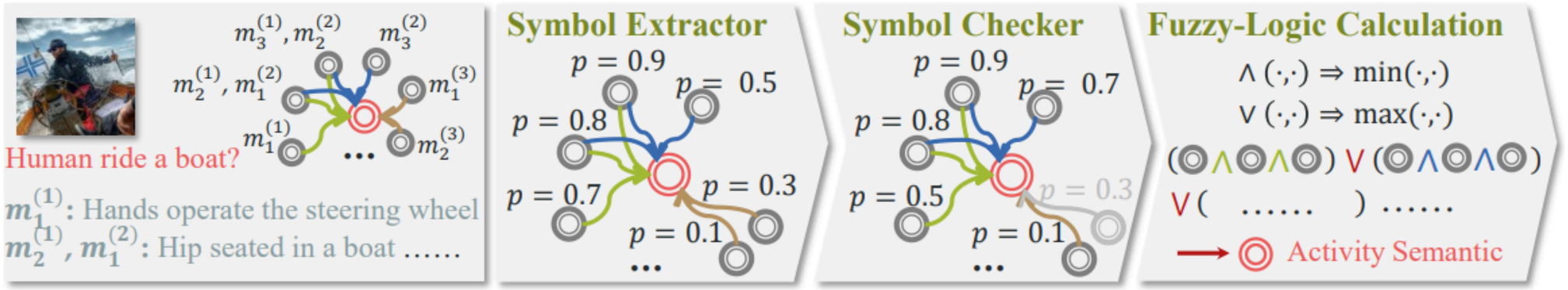


Figure 3: Visual reasoning with the proposed activity symbol system.

- **Decompose the symbol system**
  - Exclude unrelated symbols and rules

$$\mathcal{R}_c = \{r^{(j)}\}_{j=1}^{N_c}$$

$$r^{(j)} : m_1^{(j)} \wedge m_2^{(j)} \wedge \ldots m_{n_{r(j)}}^{(j)} \models c.$$

- **Extracting Visual Symbols**
  - **Text question** and query the **answer** from Existing **System-1-like VLMs**.

"Yes", "No" as $p_{y,i}^{(j)}, p_{n,i}^{(j)}$

$$p_i^{(j)} = \frac{e^{p_{y,i}^{(j)}}}{e^{p_{y,i}^{(j)}} + e^{p_{n,i}^{(j)}}}.$$

- **Reasoning**
  - **Premise symbols** within a rule are connected with ∧, while **rules for a conclusion** are connected with ∨.
  - **fuzzy logic:** $x \wedge y \rightarrow \min(x, y)$

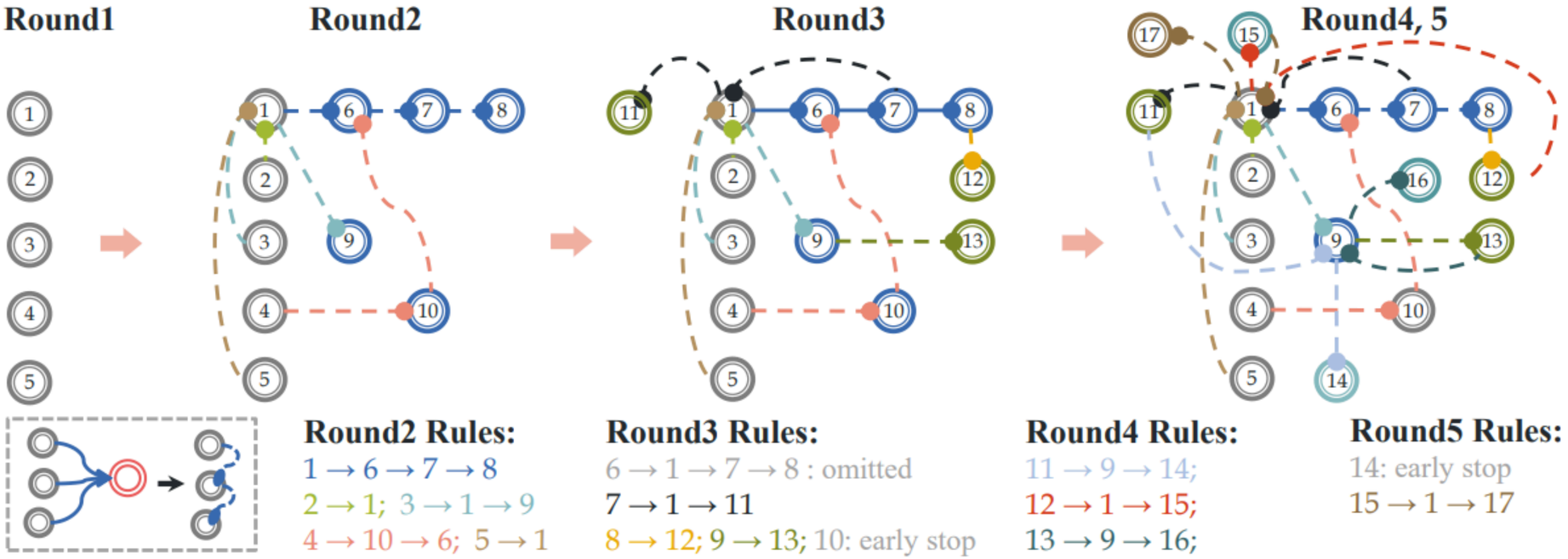
    $x \vee y \rightarrow \max(x, y)$
  - **The probability of the conclusion**

$$p_c = max_{1 \le j \le N_c}(min_{1 \le i \le n_{r(j)}} p_i^{(j)}).$$

Experiment

Symbolic System Experiment: An example

- Here, the **edge connects premises** instead of connecting premise and conclusion.



	Symbols	
Round1	(1) hold a boarding pass	(2) place luggage in overhead compartment
	(3) adjust seatbelt	(4) wave goodbye to loved ones
	(5) grip a luggage handle	
Round2	(6) walk towards the boarding gate	(7) luggage visible beside him
	(8) boarding pass is scanned by airport staff	(9) stand on the jet bridge
	(10) luggage is loaded onto the plane	
Round3	(11) reach for the airplane door handle	(12) stand in line with carry-on luggage
	(13) hold the carry-on luggage	
Round4	(14) open the airplane door	(15) move forward in the line
	(16) move towards the airplane door	
Round5	(17) airline staff checking the boarding pass	

Table 1: Generated symbols & rules for "human board an airplane". Round i: i-th symbol-rule loop.

Dataset and Metric

- **Experimental Scope**:
  - **Image-level** activity understanding across diverse tasks.

- **Benchmarks:**
  - **HICO**: Human-Object Interaction (HOI) recognition.
  - **Stanford40**: Action recognition.
  - **HAKE-Verb**: Verb recognition (newly constructed from HAKE).
  - **HAKE-PaSta**: Conditional PaSta Q-A (also newly constructed).

- **Evaluation Metrics**:
  - **mAP** (mean Average Precision) :
    - HICO, HAKE-Verb, and Stanford40.
  - **Top-1 Accuracy** :
    - HAKE-PaSta.

Entailment

- We find a **climb up in the entailment score**, implying the **effectiveness of the entailment check**.

- Once the entailment score *e* surpasses the threshold *eh* = 0.9, it is regarded as a **rule** equipped with logical entailment and **updated into the symbolic system**.

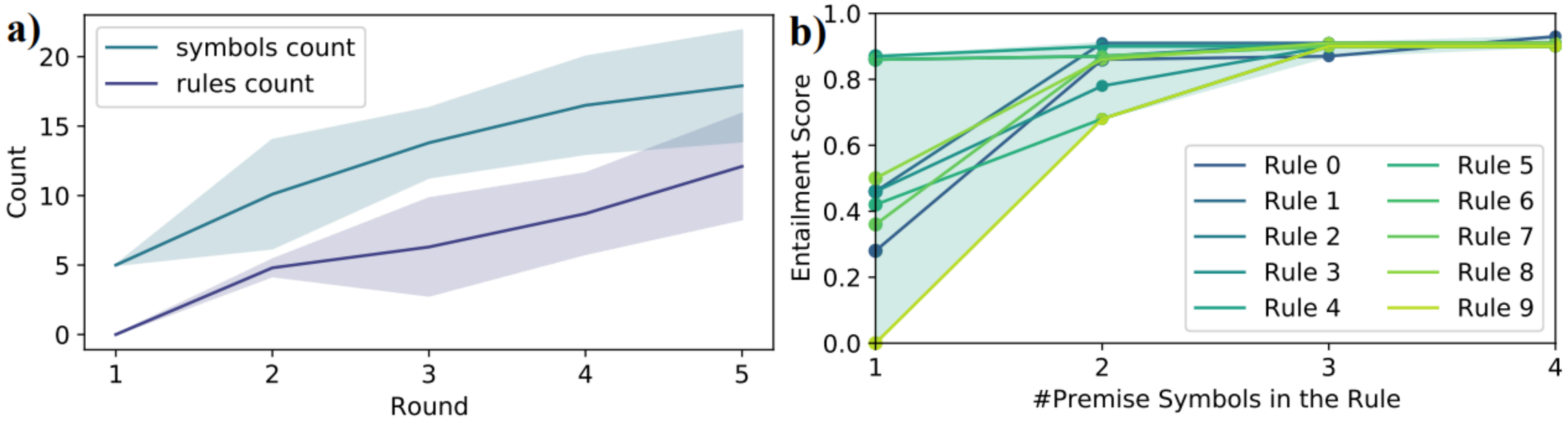


Figure 5: Statistics of the symbolic system. a) Accumulated symbols & rules count in each round. The data is from 50 randomly sampled activities. The average value and fluctuation range are shown. b) Increased entailment scores as the 10 randomly sampled rules extended with more premise symbols.

Results

- Comparing the two baselines, we find that with a frozen language model, **BLIP2** is more capable of **understanding activity semantics** and outperforms CLIP in **various zero-shot benchmarks.**

- **Implementation and Settings**
  - For visual reasoning, we use **BLIP2 ViT-g FlanT5-XL model** to **extract visual symbols.**
  - BLIP2 is a visual-language pre-trained model with a frozen large language model, thus well equipped with question-answering abilities to effectively extract symbols.

- **Performance Evaluation**
  - **SymAct (Symbol Activity) test set** (subset of the HICO test set).
  - **broader semantic coverage** than HAKE
  - **more rational rules** than HAKE
  - **Confusion**:
    - counting different image-activity pairs which share the same symbols
    - **No confusion pairs** on the test set because of the presentation ability of symbols and entailment check.

- **Robustness**
  - **Convergence**
  - Low sensitivity to **initial conditions**
  - Low sensitivity to **prompts**

Symbol probability	I	II	III	IV
m_1: talk with seller	0.92	0.87	0.07	0.16
m_2: reach for an orange	0.14	0.68	0.95	0.12
m_3: seller hand over orange	0.94	0.82	0.09	0.13
m_4: stand in front of fruit stand	0.57	0.85	0.04	0.77
m_5: place orange in a bag	0.67	0.90	0.06	0.45
m_6: pick orange from a basket	0.68	0.83	0.03	0.74
m_7: hold a bag of oranges	0.47	0.88	0.05	0.75
m_8: reach for a wallet	0.86	0.83	0.04	0.43
m_9: seller put the orange in bag	0.83	0.86	0.04	0.45
m_{10}: give money to seller	0.82	0.87	0.04	0.41

Figure 6: Visualization results. Symbols & activity predictions for "human buy orange" are shown.

Method	HICO mAP		HAKE-verb mAP		Stanford-40	HAKE-PaSta
	fine-tuned	zero-shot	fine-tuned	zero-shot	zero-shot mAP	zero-shot Acc(%)
CLIP [26]	67.12	37.08	73.82	43.92	75.68	39.36
CLIP [26]+Reasoning	**69.73**	**43.21**	**75.27**	**48.95**	**82.22**	**40.47**
BLIP2 [19]	-	50.61	-	49.47	91.85	43.81
BLIP2 [19]+Reasoning	-	**53.15**	-	**51.37**	**92.59**	**44.65**

Table 2: Results on activity benchmarks. More HICO [5] baselines are listed in supplementary.

Ablation Study

- verify the **effectiveness of integrating System-2 reasoning**.
  - Results of CLIP baseline can be either combined with reasoning results or trivially combined with other baseline results (CLIP +BLIP2).

- It verifies the **necessity of System-2 reasoning** other than trivially combining predictions from two models.
  - Find that the former is superior (48.95 mAP) to the latter (44.76 mAP), though both outperform the baseline CLIP (43.92 mAP).

- replace the symbol extractor BLIP2 with **CLIP** and find a performance fall with a **weaker ability to extract visual information.**

- reasoning **without a symbol checker** suffers from degradation due to the **negative effect of inaccurate symbol predictions**.

Method	mAP
CLIP [26]+Reasoning	**48.95**
CLIP [26]+BLIP2 [19]	44.76
CLIP [26]	43.92
rules from HAKE [20]	44.98
w/o entailment check	46.51
w/o symbol-rule loop	47.53
80% rules	48.57
50% rules	47.61
20% rules	46.22
CLIP [26] as extractor	46.09
w/o checking symbols	48.17

Table 3: Ablation studies on zero-shot HAKE [20]-Verb.

Bottleneck Analysis

- The drop from 100 to 60.79 mAP is caused by the **symbolic system not covering all samples**.
- The drop from 100 → 83.11 → 41.52 mAP reveals the performance loss due to **errors in symbol prediction**.

Symbol Prediction	Symbolic System	mAP
perfect	perfect	**100.00**
imperfect	perfect	**83.11**
perfect	imperfect	**60.79**
imperfect	imperfect	**41.52**

Table 4: System-1/2 analysis on SymAct test set.

Conclusion and Discussion

Conclusion

- Rethink the symbolic system in activity reasoning and propose a new one with **broad-coverage symbols and rational rules**.
- Enhance **System-2 reasoning** by integrating it into System-1.
- Demonstrate how to **instantiate** it and how it helps to reason with visual inputs.
- The method shows superiority in **explainability**, **generalization**, and **efficiency** in extensive experiments.

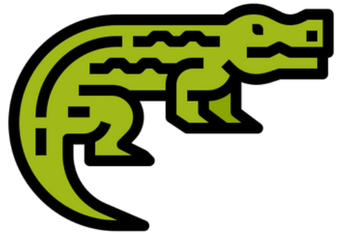
Discussion

- **Computation Cost**
  - **Symbol predictions** will increase the computational cost as a trade-off for explainability and generalization.
  - It can be eased by discovering the **hierarchical** and **reusable** nature of the symbols.

- **Broader Application**
  - Choose **activity understanding** as a good and important initial test bed because it is a difficult task with complex visual patterns and a compositional nature.
  - To verify the broader application, the authors provide some initial results on:
    - **object recognition/classification**: CIFAR-100 test set with a zero-shot setting.
    - **VCR (visual commonsense reasoning)** tasks: VCR val set.
      - BLIP2 is adopted as a baseline as it is a VQA-style task.

What are person on the right doing?
a) person on the left is taking person on the right home.
b) person on the right are on a first date.
c) person on the left is conducting a job interview with person on the right.
d) They are picking something up.

object (e.g., crocodile) is:
long, slightly curved body ∧ four short legs ∧ ... ∧ a long, muscular tail
→ crocodile



Feedbacks

Feedbacks

- Utilized symbol “∧” to connect the visual pattern elements/primitives.
  - How if “∨” represented the **uncertainty events** in the visual tasks,
    “**¬**” represented for **not happening**?
  - Would a sophisticated symbolic system make a more precise, complex modelling in a real-world application?

- Entailment scoring with a prompt set into a few probabilistic answers, would the **discrete choices** introduce some **biases** in the confident scoring?

Supplementary

Artificial Intelligence Course:

Call for Team Members!

賴澍雨 Shih-Yu Lai

- Multimodal Hypergraph
- Visual-Language Model
- Reasoning
- Scalable Inference

Thank You for Listening!