Declarative Knowledge Distillation from Large Language Models for Visual Question Answering Datasets

Thomas Eiter, Jan Hadl, Nelson Higuera, Johannes Oetsch

Presented by Yun-Ze Li
2024.11.05

Authors

Thomas Eiter

Jan Hadl

Nelson Higuera

Johannes Oetsch

Outlines

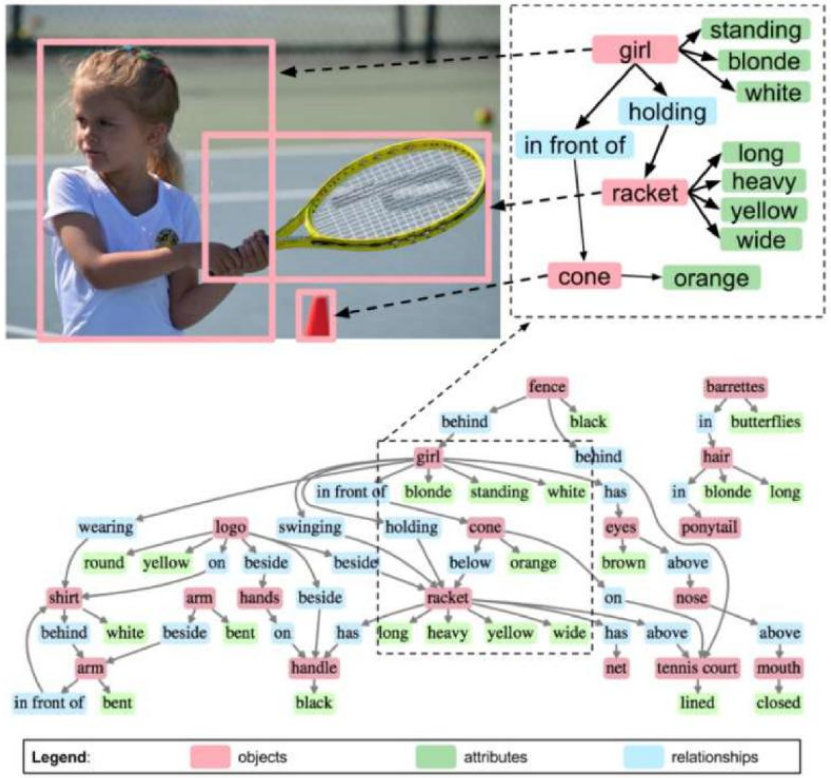
- Promblem Setting of VQA
- Neurosymbolic approaches
- Knowledge Distillation with LLM
- Experiments

Visual question answering(VQA)

VQA: Answering a question about a visual scene

- Not just **understand** vision and text.
- Requires the ability to follow complex chains of **reasoning** operations.

Visual question answering(VQA)

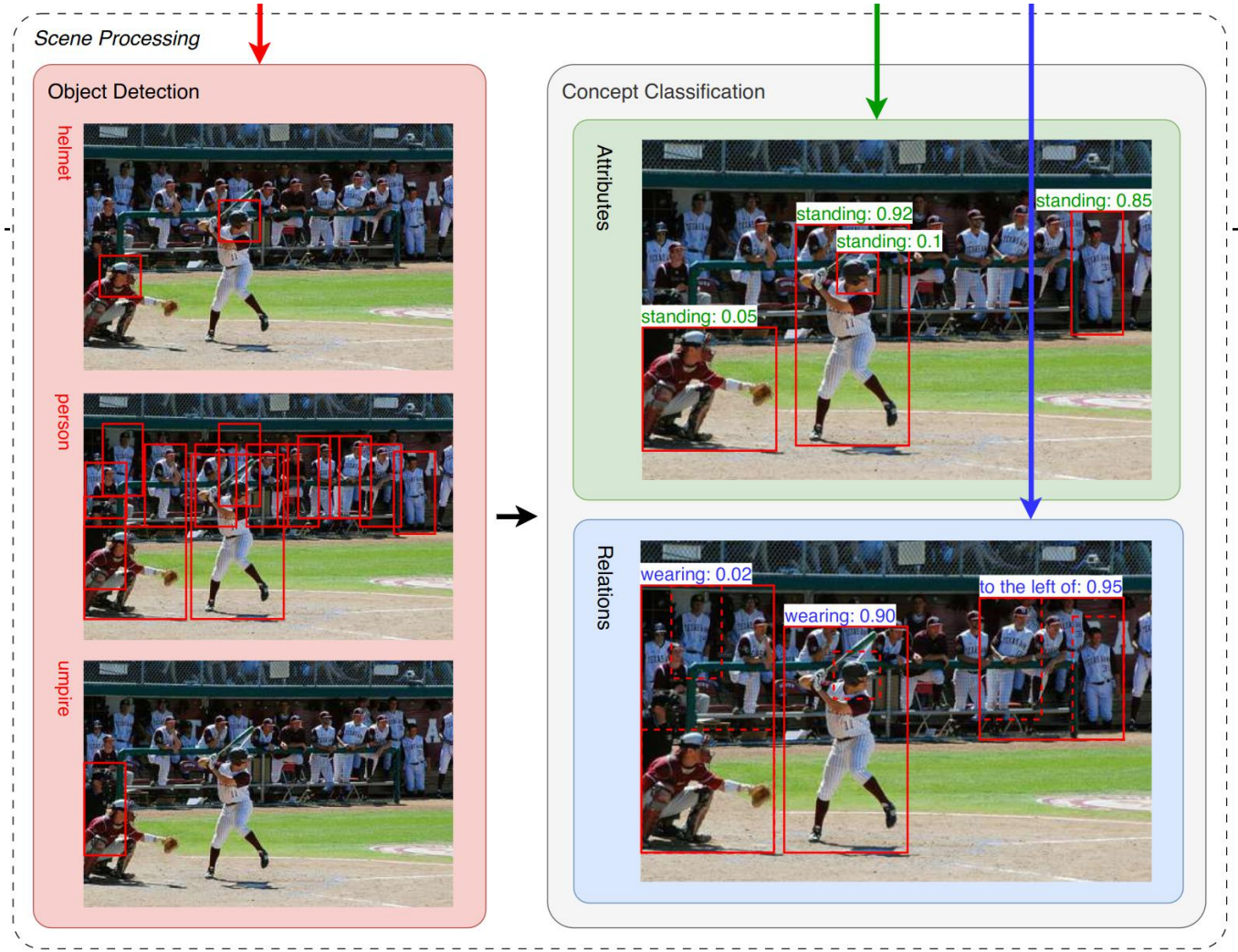


GQA Datasets

Neurosymbolic approaches(GS-VQA)

Input:

Symbolic Representation (ASP Representation)



has_attr(o1, pose, standing).

object(O).
has_attr(O, Class, Attr).
has_rel(O1, Rel, O2).

Neurosymbolic approaches(GS-VQA)

Is the umpire to the right or to the left of the standing person that is wearing a helmet?

Question Encoding

```
scene(0).
select(1, 0, helmet).
relate(2, 1, person, wearing, subject).
filter_any(3, 2, standing).
choose_rel(4, 3, umpire, to_the_left_of,
to_the_right_of, subject).
end(4).
```

Scene Encoding

```
object(o1).
has_obj_weight(o1, 1971).
has_attr(o1, class, person).
has_attr(o1, class, baseball_player).
has_attr(o1, name, baseball_player).

has_attr(o1, hposition, middle).
has_attr(o1, vposition, middle).

{has_attr(o1, pose, standing)}.
:~ has_attr(o1, pose, standing). [83]
:~ not has_attr(o1, pose, standing). [2525]

{has_rel(o1, wearing, o2)}.
:~ has_rel(o1, wearing, o2). [105]
:~ not has_rel(o1, wearing, o2). [2302]
```

ASP Theory

```
state(TO,ID) :- select(TO, TI, CLASS), state(TI,
ID), has_attr(ID, class, CLASS).

state(TO,ID) :- filter_any(TO, TI, VALUE), state(TI,
ID), has_attr(ID, ATTR, VALUE).

...
```

0: object in the scene
1: Select the object with class = helmet in scene 0
2: Find the person with relation = wearing in 1
3: Filter out the person with attr = standing in 2
...

ASP Solver

Answer

to the left

Neurosymbolic approaches(GS−VQA)*

Question Encoding

```
scene(0).
select(1, 0, helmet).
relate(2, 1, person, wearing, subject).
filter_any(3, 2, standing).
choose_rel(4, 3, umpire, to_the_left_of,
to_the_right_of, subject).
end(4).
```

Scene Encoding

```
object(o1).
has_obj_weight(o1, 1971).
has_attr(o1, class, person).
has_attr(o1, class, baseball_player).
has_attr(o1, name, baseball_player).

has_attr(o1, hposition, middle).
has_attr(o1, vposition, middle).

{has_attr(o1, pose, standing)}.
:~ has_attr(o1, pose, standing). [83]
:~ not has_attr(o1, pose, standing). [2525]

{has_rel(o1, wearing, o2)}.
:~ has_rel(o1, wearing, o2). [105]
:~ not has_rel(o1, wearing, o2). [2302]
```

state(TO,ID):-scene(TO), object(ID)
ans(R) :- end(TO), rel(TO, R)

ASP Theory

```
state(TO,ID) :- select(TO, TI, CLASS), state(TI,
ID), has_attr(ID, class, CLASS).

state(TO,ID) :- filter_any(TO, TI, VALUE), state(TI,
ID), has_attr(ID, ATTR, VALUE).
```

state(0,ID) :- scene(0), object(ID).
state(1,ID) :- select(1), state(0,ID), has_attr(ID, class, helmet).
...
state(4,ID) :- ...

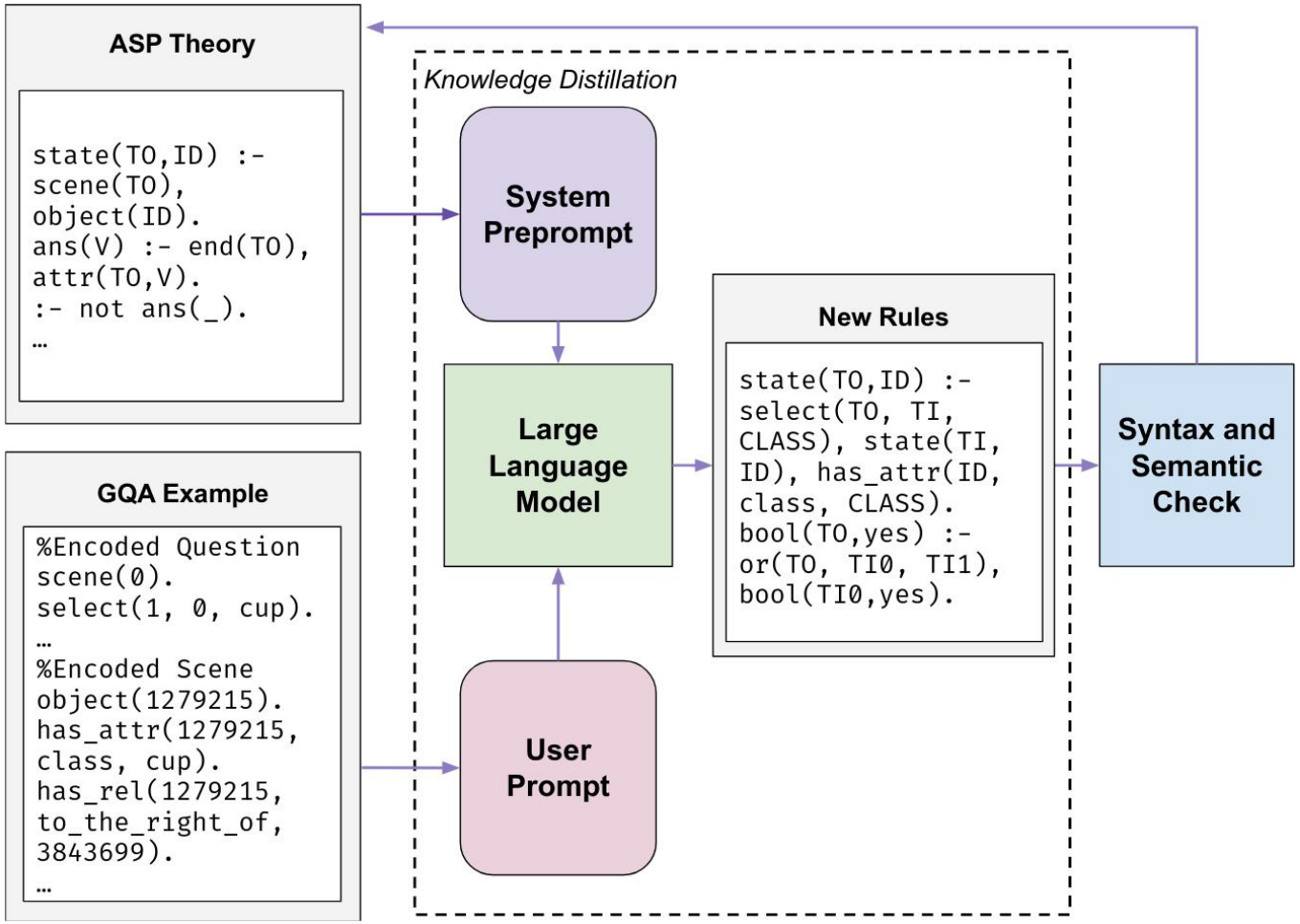
$$\xrightarrow{\text{scene}} 0 \xrightarrow{\text{select}} 1 \xrightarrow{\text{relate}} 2 \xrightarrow{\text{filter\_any}} 3 \xrightarrow{\text{choose\_rel}} 4 \xrightarrow{\text{end}} \quad \in L(A)?$$

Knowledge Distillation

Input: **ASP theory, Qustion, Scene, Ans**

each example try *r* time:

1. Prompting to get **New_Rules**
2. Solving
   a. Syntax Check(try *m* time)
   b. Semantic Check(try *m* time)
3. Regression Testing **New_Rules** on past example



Knowledge Distillation

```
Your task is to keep the ASP theory
updated with rules that allows us
to answer questions.
We provide an initial theory that
can handle some instances.
The prompt input will consist of one or
more questions in the ASP representation.

Strictly follow these guidelines:
1. Only output the new ASP Rules.
2. Do not add facts as rules.
3. New rules should be as general
as possible, i.e., have a low number of
constants and high number of variables.
4. Do not output any natural language.
```

请不断更新 ASP 理论，增加规则以回答给出问题。
我们提供了初始理论、一个或多个问题。
严格遵循以下准则：
1. 仅输出新的 ASP 规则
2. 不要将事实添加为规则
3. 新规则应尽可能通用，即常量数量少，变量数量多
4. 不要输出任何自然语言。

Experiments

P	Init $\setminus P$	GPT-4	GPT-3.5	Mistral
query	48.84	97.67 ± 18.05 $(89.16, 98.92)$	70.02 ± 19.36 $(48.84, 85.53)$	—
exist	86.36	99.75 ± 00.50 $(98.86, 99.98)$	87.65 ± 02.41 $(86.36, 91.95)$	89.68 ± 04.20 $(86.36, 95.66)$
or	92.18	100.0 ± 00.00 $(100.0, 100.0)$	93.03 ± 01.90 $(92.18, 96.44)$	93.20 ± 01.66 $(92.18, 96.02)$
filter	81.60	98.21 ± 00.40 $(97.49, 98.40)$	83.15 ± 03.47 $(81.60, 89.37)$	81.70 ± 00.24 $(81.60, 82.14)$
choose_attr	92.12	95.98 ± 05.37 $(88.73, 99.83)$	93.73 ± 01.36 $(92.31, 95.83)$	92.12 ± 00.01 $(92.12, 92.15)$
verify_rel	93.72	98.60 ± 01.11 $(96.73, 99.43)$	—	—
select	9.53	99.94 ± 00.07 $(99.87, 100.0)$	27.42 ± 40.01 $(9.53, 99.01)$	—
negate	98.59	98.54 ± 00.20 $(98.59, 98.74)$	—	—
relate	56.89	69.38 ± 12.50 $(56.89, 85.25)$	—	—
two_different	98.94	100.0 ± 00.00 $(100.0, 100.0)$	99.39 ± 00.55 $(98.94, 100.0)$	—
two_same	98.83	99.99 ± 00.00 $(99.99, 100.0)$	99.05 ± 00.53 $(98.83, 100.0)$	—

(a) Results for GQA.

P	Init $\setminus P$	GPT-4	GPT-3.5	Mistral
exist	79.63	99.48 ± 00.70 $(98.72, 100.0)$	87.82 ± 11.11 $(98.72, 100.0)$	80.21 ± 06.36 $(72.16, 90.02)$
unique	29.19	97.67 ± 18.05 $(89.16, 98.92)$	—	—
count	98.01	99.60 ± 00.88 $(98.01, 100.0)$	98.01 ± 01.12 $(98.01, 98.40)$	—
equal_integer	96.61	99.92 ± 00.17 $(99.61, 100.0)$	97.80 ± 01.26 $(96.61, 99.60)$	—
and	93.67	100.0 ± 00.00 $(100.0, 100.0)$	97.46 ± 03.46 $(93.67, 100.0)$	—
relate_left	84.73	100.0 ± 00.00 $(100.0, 100.0)$	96.18 ± 07.63 $(84.73, 100.0)$	94.14 ± 08.02 $(84.73, 100.0)$
filter_large	68.54	100.0 ± 00.00 $(100.0, 100.0)$	87.41 ± 17.23 $(68.54, 100.0)$	81.12 ± 17.23 $(68.54, 100.0)$
query_shape	72.23	100.0 ± 00.00 $(100.0, 100.0)$	100.0 ± 00.00 $(100.0, 100.0)$	94.44 ± 12.43 $(72.23, 100.0)$
same_color	94.79	99.36 ± 00.87 $(98.41, 100.0)$	100.0 ± 00.00 $(100.0, 100.0)$	97.07 ± 02.70 $(94.79, 100.0)$

(b) Results for CLEVR.

Table 1: Results for the knowledge distillation method when attempting to restore Init after all rules that mention a predicate P are removed.

Experiments

$s(\%)$	Init	GPT-4	GPT-3.5
10	26.57	94.67 ± 02.21 $(89.71, 95.67)$	—
20	63.54	75.56 ± 11.86 $(63.55, 90.22)$	66.14 ± 07.78 $(63.55, 89.48)$
50	7.17	47.48 ± 15.26 $(30.25, 71.64)$	24.43 ± 12.28 $(07.18, 46.88)$

(a) Results for GQA.

$s(\%)$	Init	GPT-4	GPT-3.5
10	46.61	70.76 ± 04.17 $(66.59, 75.62)$	50.30 ± 03.68 $(46.62, 53.98)$
20	9.66	44.66 ± 30.58 $(14.16, 97.30)$	32.03 ± 11.86 $(13.44, 45.70)$
50	0.0	23.90 ± 03.30 $(18.06, 27.84)$	10.19 ± 19.59 $(00.00, 49.36)$

(b) Results for CLEVR.

Table 2: Knowledge distillation results when attempting to restore a complete ASP theory after a percentage s of rules is randomly removed.

b	Light	Medium	Heavy
Init	0.0	0.0	6.24
1	$56,26 \pm 10.23$ $(34.54, 61.28)$	81.45 ± 05.07 $(76.86, 87.91)$	83.85 ± 02.49 $(81.38, 87.77)$
2	32.71 ± 04.31 $(25.72, 43.15)$	79.83 ± 03.42 $(75.11, 83.03)$	74.32 ± 02.91 $(75.86, 80.54)$
5	16.62 ± 05.28 $(10.51, 17.59)$	69.68 ± 31.12 $(24.18, 82.19)$	84.25 ± 04.59 $(78.93, 89.48)$
10	—	15.38 ± 12.30 $(11.62, 31.75)$	84.75 ± 04.20 $(80.64, 90.85)$

(a) Results for GQA.

b	Light	Medium	Heavy
Init	0.0	5.56	20.80
1	84.68 ± 26.42 $(38.23, 100.0)$	86.97 ± 04.35 $(83.89, 90.05)$	95.40 ± 03.83 $(91.25, 98.81)$
2	75.4 ± 33.78 $(27.84, 99.88)$	18.68 ± 04.50 $(15.67, 26.09)$	88.51 ± 04.46 $(83.37, 91.25)$
5	17.06 ± 29.55 $(00.00, 51.19)$	17.79 ± 03.00 $(15.67, 19.92)$	94.39 ± 03.71 $(91.33, 98.52)$
10	—	—	89.88 ± 09.04 $(77.68, 98.81)$

(b) Results for CLEVR.

Table 3: Results for the knowledge distillation method when using batch sizes b and the different initial theories Light, Medium and Heavy.

Experiments

2.Solving: m = 0

P	Init $\setminus P$	GPT-4	GPT-3.5	Mistral
query	48.84	99.02 ± 00.04 (99.02, 99.03)	55.90 ± 15.80 (48.84, 84.17)	—
exist	86.36	99.09 ± 01.76 (95.94, 99.97)	87.83 ± 03.29 (86.36, 93.73)	86.60 ± 0.05 (86.36, 87.57)
or	92.18	100.0 ± 00.00 (100.0, 100.0)	93.03 ± 01.90 (92.18, 96.44)	94.41 ± 3.35 (92.18, 99.73)
filter	81.60	98.56 ± 00.58 (98.47, 98.63)	82.38 ± 0.078 (81.92, 82.50)	84.99 ± 7.59 (81.60, 98.59)
choose_attr	92.12	98.65 ± 02.65 (93.91, 99.88)	95.16 ± 04.26 (92.03, 99.84)	92.08 ± 0.06 (91.98, 92.12)
verify_rel	93.72	95.74 ± 03.49 (93.72, 99.08)	94.30 ± 01.17 (93.72, 96.06)	—
select	9.53	81.69 ± 36.81 (9.53, 100.0)	28.94 ± 39.82 (9.53, 100.0)	—
negate	98.59	98.72 ± 00.02 (98.59, 99.12)	—	—
relate	56.89	58.02 ± 02.19 (57.54, 61.91)	57.29 ± 00.09 (56.89, 58.91)	—
two_different	98.94	100.0 ± 00.00 (100.0, 100.0)	—	—
two_same	98.83	99.60 ± 00.02 (99.50, 100.0)	99.09 ± 00.05 (98.83, 100.0)	—

Table 4: Ablation study for GQA: Results when attempting to restore Init after removing rules for P without mending step.

P	Init $\setminus P$	GPT-4	GPT-3.5	Mistral
query	48.84	97.67 ± 18.05 (89.16, 98.92)	70.02 ± 19.36 (48.84, 85.53)	—
exist	86.36	99.75 ± 00.50 (98.86, 99.98)	87.65 ± 02.41 (86.36, 91.95)	89.68 ± 04.20 (86.36, 95.66)
or	92.18	100.0 ± 00.00 (100.0, 100.0)	93.03 ± 01.90 (92.18, 96.44)	93.20 ± 01.66 (92.18, 96.02)
filter	81.60	98.21 ± 00.40 (97.49, 98.40)	83.15 ± 03.47 (81.60, 89.37)	81.70 ± 00.24 (81.60, 82.14)
choose_attr	92.12	95.98 ± 05.37 (88.73, 99.83)	93.73 ± 01.36 (92.31, 95.83)	92.12 ± 00.01 (92.12, 92.15)
verify_rel	93.72	98.60 ± 01.11 (96.73, 99.43)	—	—
select	9.53	99.94 ± 00.07 (99.87, 100.0)	27.42 ± 40.01 (9.53, 99.01)	—
negate	98.59	98.54 ± 00.20 (98.59, 98.74)	—	—
relate	56.89	69.38 ± 12.50 (56.89, 85.25)	—	—
two_different	98.94	100.0 ± 00.00 (100.0, 100.0)	99.39 ± 00.55 (98.94, 100.0)	—
two_same	98.83	99.99 ± 00.00 (99.99, 100.0)	99.05 ± 00.53 (98.83, 100.0)	—

(a) Results for GQA.

Experiments

$s(\%)$	Init	GPT-4	GPT-3.5
10	26.57	83.06 ± 23.26 $(36.61, 95.80)$	—
20	63.54	55.46 ± 14.84 $(26.99, 69.84)$	38.67 ± 21.00 $(08.48, 64.96)$
50	7.17	36.38 ± 10.62 $(18.15, 48.55)$	04.21 ± 03.42 $(00.00, 7.94)$

Table 5: Ablation study for GQA: Attempting to restore theory T after s percent of rules were randomly removed without mending.

b	Light	Medium	Heavy
Init	0.0	0.0	6.24
1	$51,43 \pm 08.56$ $(42.72, 59.85)$	80.41 ± 05.34 $(74.54, 85.01)$	76.70 ± 00.76 $(75.82, 77.19)$
2	27.06 ± 09.60 $(21.09, 38.14)$	77.70 ± 05.42 $(75.25, 83.92)$	77.60 ± 03.85 $(73.41, 81.00)$
5	15.23 ± 02.28 $(12.60, 16.49)$	60.16 ± 34.17 $(21.08, 84.45)$	86.28 ± 10.84 $(27.93, 94.71)$
10	—	19.55 ± 07.41 $(13.03, 27.62)$	73.93 ± 06.50 $(66.42, 77.69)$

Table 6: Ablation study for GQA: Results when using batch sizes b and different initial theories without mending step.

Experiments

Model	Category	Accuracy
BLIP-2	end-to-end	44.7%
CodeVQA	question-symbolic	49.0%
FewVLM	end-to-end	29.3%
GS-VQA (ours)	neurosymbolic	39.5%
PnP-VQA	semi-symbolic	42.3%
ViperGPT	question-symbolic	48.1%

State−of−the−art VLM for VQA.

- **End-to-End**: End-to-end systems are those that rely solely on neural networks for computing the answer.
- **Neurosymbolic**: Neurosymbolic systems like ours are those that combine both neural networks for parsing data and symbolic execution to calculate the answers.
- **Question-Symbolic**: Such methods extract a symbolic representation from only the input question, usually in the form of some programmatic specification of the reasoning steps needed to arrive at the answer of the question.
- **Semi-Symbolic**: PnP-VQA (Tiong et al. 2022) extracts a symbolic representation of the image but does not perform its reasoning purely symbolically, hence we classify this method as semi-symbolic.

Summary

- Pros：
  - Structural representation of state：Able to use ASP to do complex reasoning
  - General Rules. A small number of rules with LLM can be applied to large datasets such as GQA.

- Cons：
  - The fact in question is given by the dataset
  - Assume that the fact generated by VLM must be correct.