NEURAL INFORMATION PROCESSING SYSTEMS

Jury-and-Judge Chain-of-Thought for Uncovering Toxic Data in 3D Visual Grounding

Kaixiang Huang, Qifeng Zhang, Jin Wang, Jingru Yang, Yang Zhou

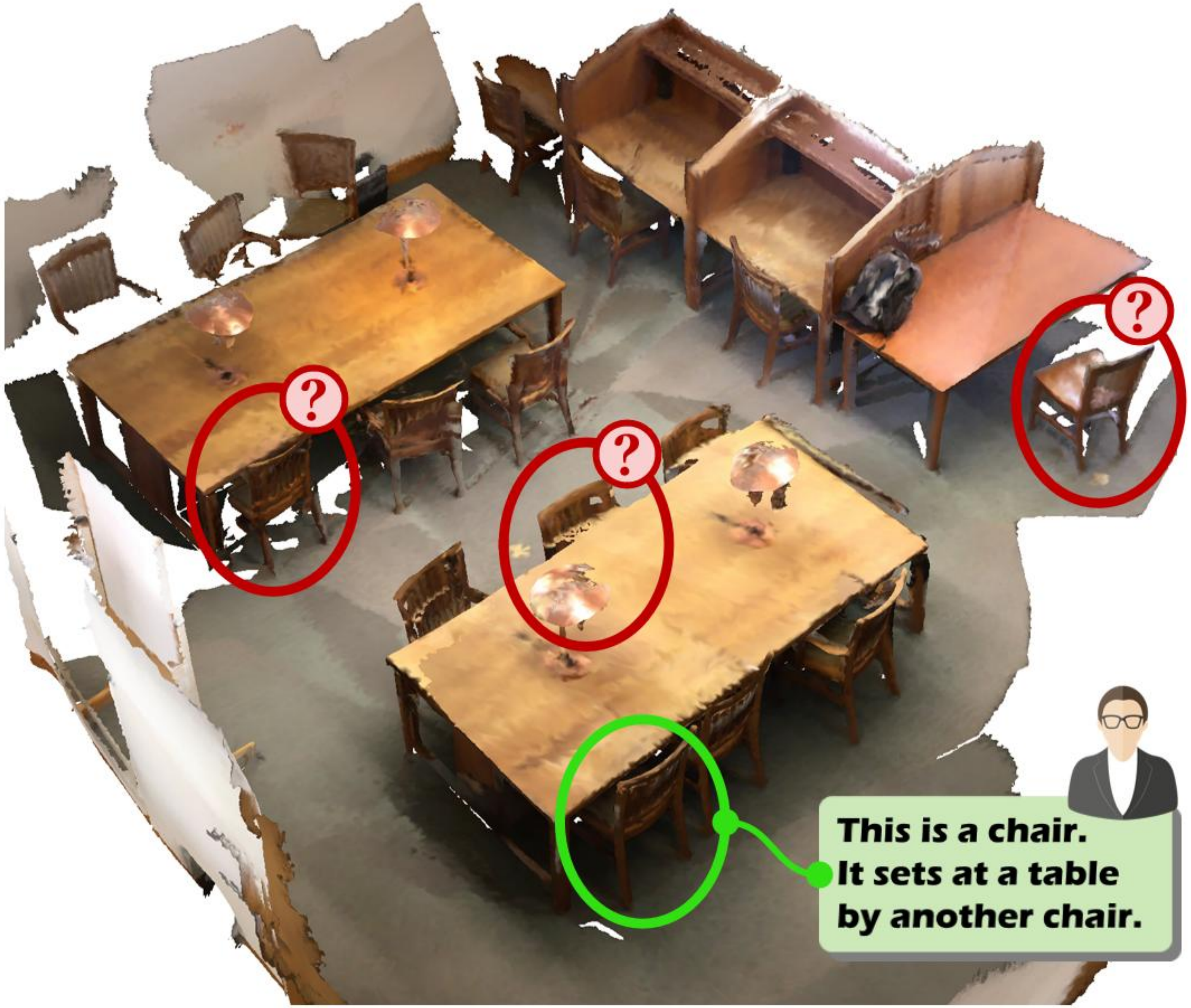
Huan Yu, Guodong Lu, Shengfeng He

ZHEJIANG UNIVERSITY 1897

CARNEGIE MELLON UNIVERSITY PITTSBURGH PENNSYLVANIA 1900

SINGAPORE MANAGEMENT UNIVERSITY

Motivation



Background

3D Visual Grounding Data Requires:

- Each annotation **uniquely** corresponds to **one object** in the scene.

Bad Annotations

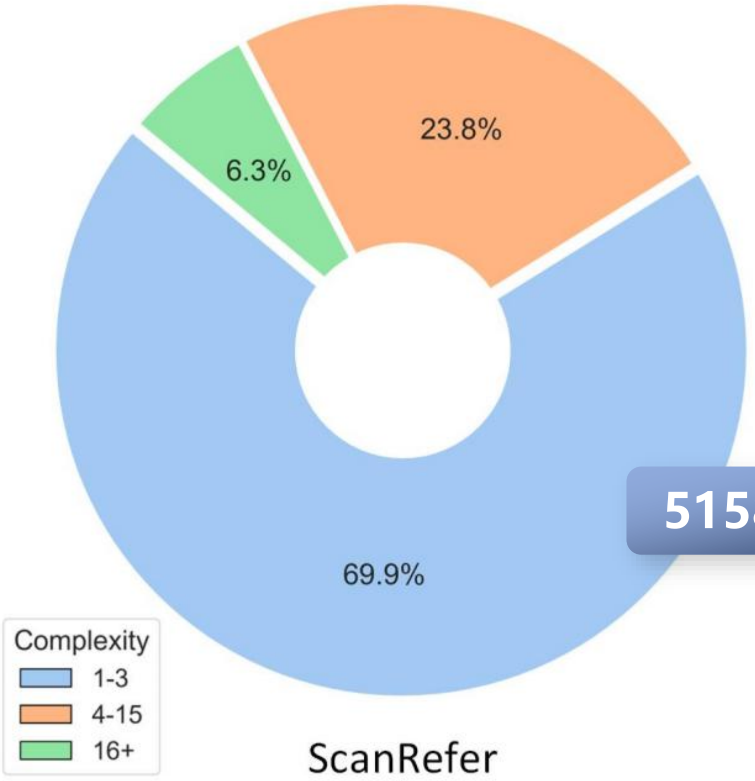
However, the annotation process of 3DVG is difficult and requires sustained focus. Annotators need to extract clues from **sparse 3D point clouds** and **disjointed 2D frames**.

↓

This leads to a sharp increase in the risk of **annotation errors**.

Motivation

So **how serious** is the data problem in 3DVG now?



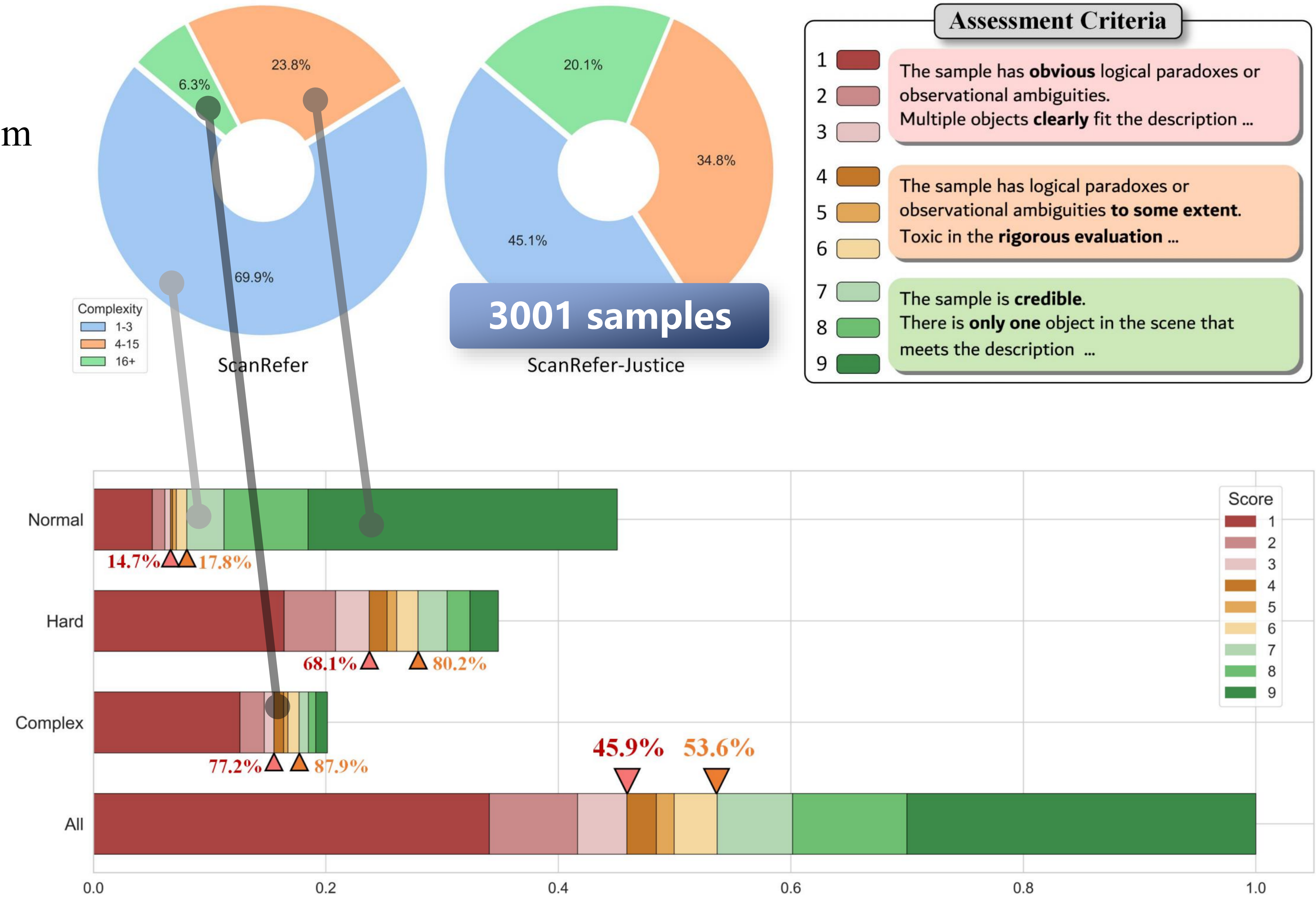
51583 samples

Motivation

So **how serious** is the data problem in 3DVG now?

- As the complexity of the scene increases, the incidence of annotation errors made by annotators **gradually rises**.

- High-confidence annotations for some complex scenes are **even below 50%.**



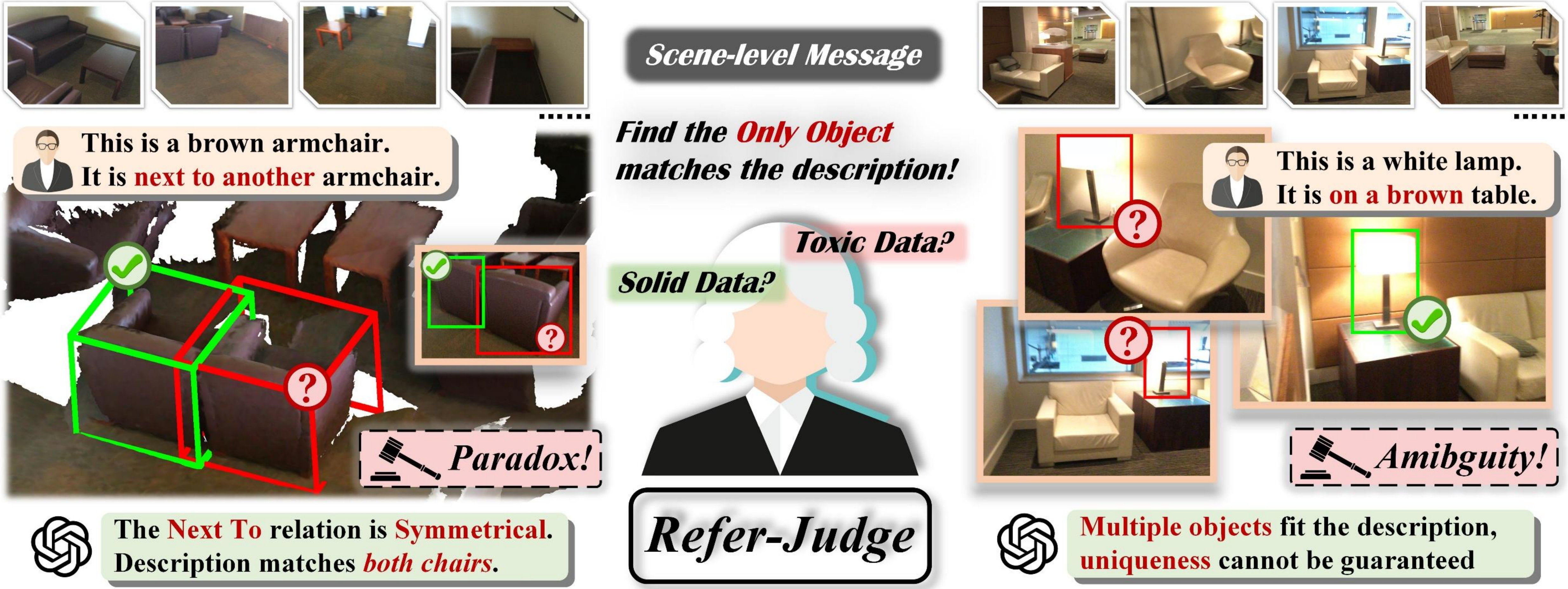
Assessment Criteria

1, 2, 3: The sample has **obvious** logical paradoxes or observational ambiguities. Multiple objects **clearly** fit the description ...

4, 5, 6: The sample has logical paradoxes or observational ambiguities **to some extent**. Toxic in the **rigorous evaluation** ...

7, 8, 9: The sample is **credible**. There is **only one** object in the scene that meets the description ...

Refer-Judge



Scene-level Message

Find the Only Object matches the description!

This is a brown armchair. It is next to another armchair.

Paradox!

The Next To relation is Symmetrical. Description matches *both chairs*.

Solid Data?

Toxic Data?

Refer-Judge

This is a white lamp. It is on a brown table.

Amibguity!

Multiple objects fit the description, uniqueness cannot be guaranteed

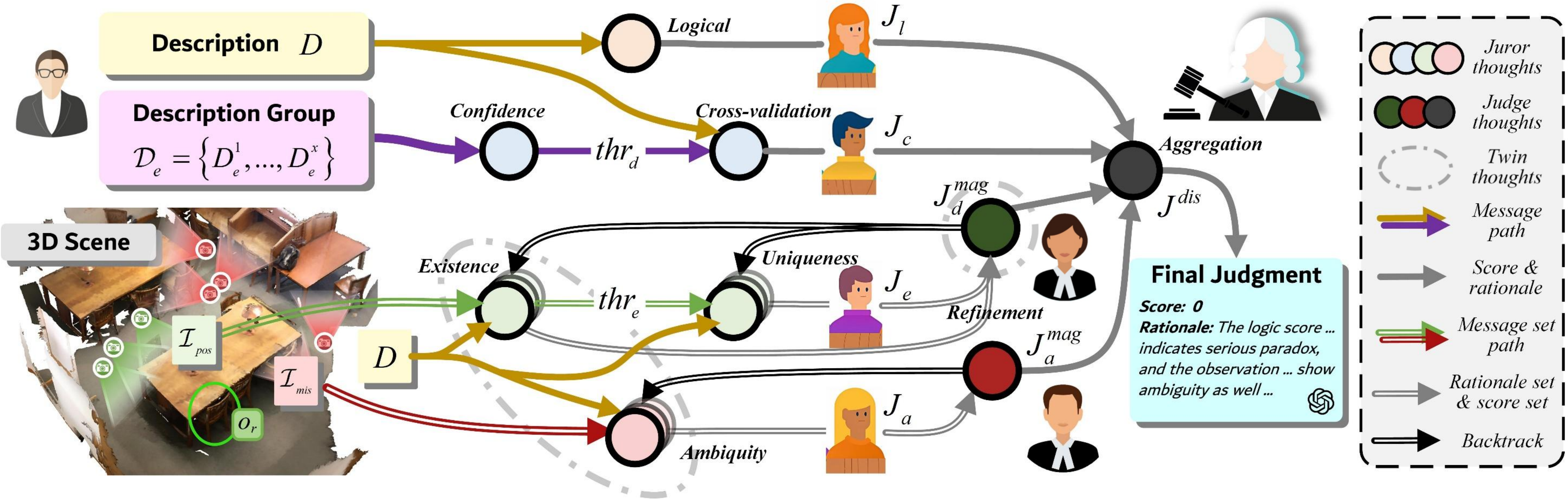
Toxic samples arise from **two sources**:

- ➢ ***Logical paradoxes***
- ➢ ***Referential ambiguities***

Impacts model **training**

Affects algorithm **evaluation**

Refer-Judge



Final Judgment

***Score:* 0**

Rationale: *The logic score ... indicates serious paradox, and the observation ... show ambiguity as well ...*

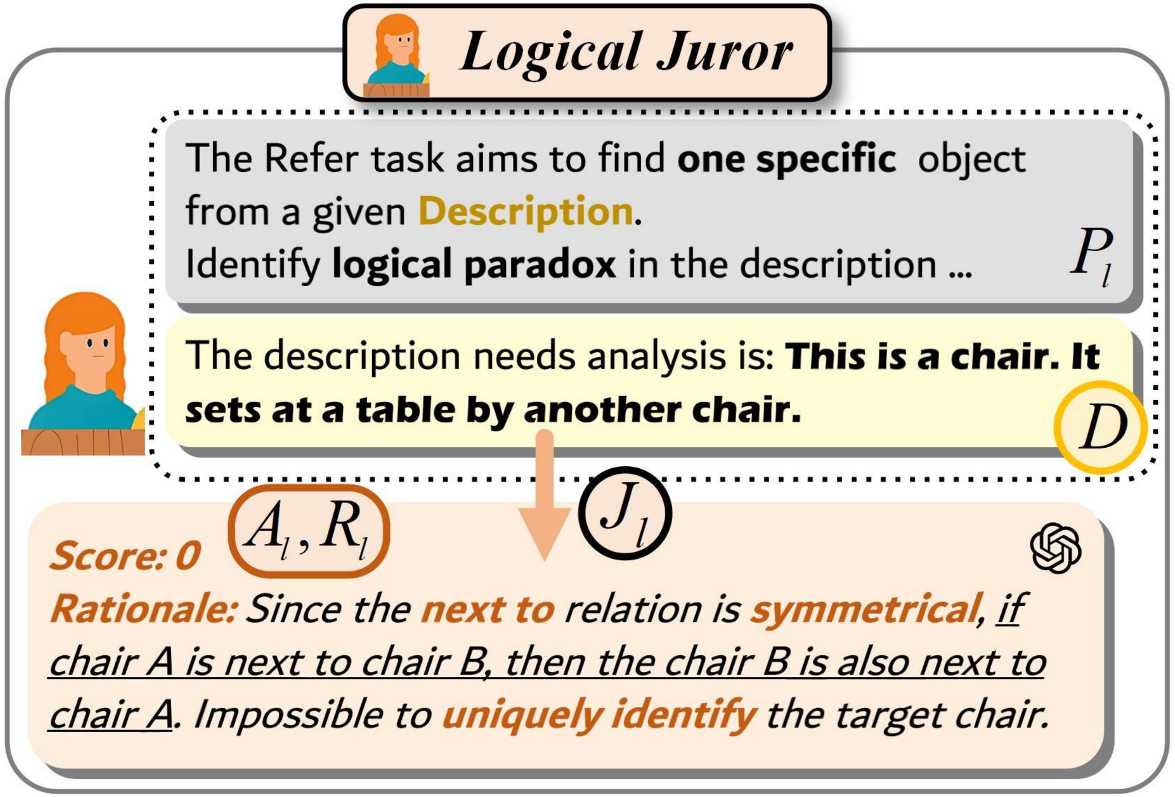
Refer-Judge

- **Hypergraph-of-Thoughts**
- **Agentic framework**
- **Deliberation-based structured analysis**

Jury-and-Judge

- **Jury：** multi-faceted evaluations for scene-level data
- **Judge：** self-consistency evaluation and corroborative refinement
- No task-specific fine-tuning or external perception

Refer-Judge



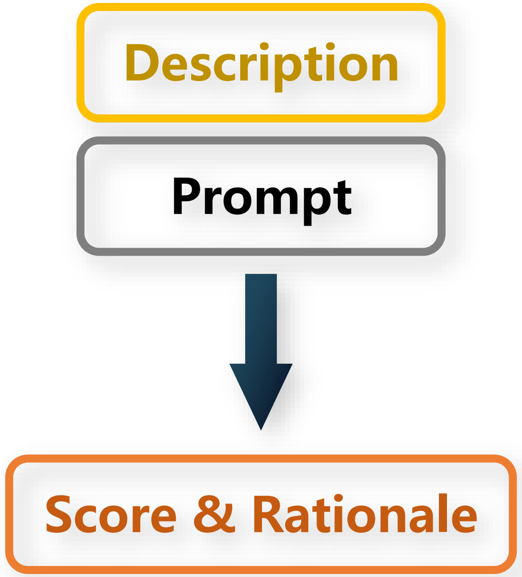
Logical Juror

The Refer task aims to find **one specific** object from a given **Description**.
Identify **logical paradox** in the description ...

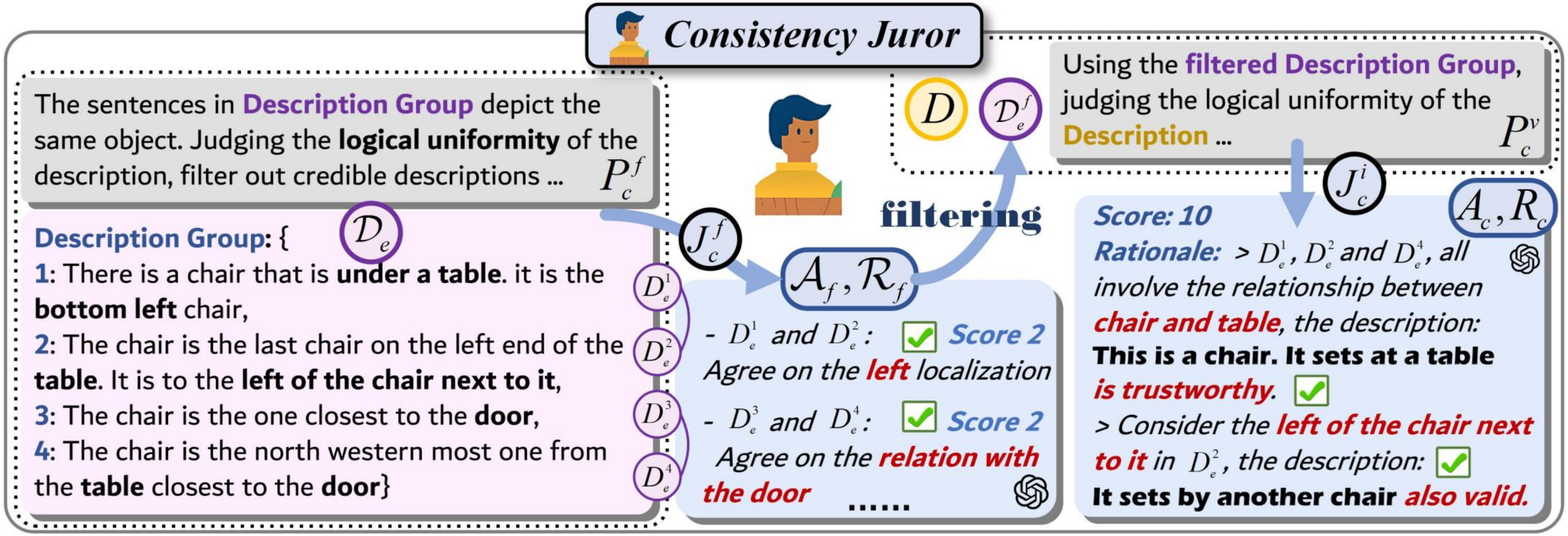
The description needs analysis is: **This is a chair. It sets at a table by another chair.**

Score: 0

Rationale: *Since the* ***next to*** *relation is* ***symmetrical****, if chair A is next to chair B, then the chair B is also next to chair A. Impossible to* ***uniquely identify*** *the target chair.*



Refer-Judge



The sentences in Description Group depict the same object. Judging the **logical uniformity** of the description, filter out credible descriptions ... P_c^f

Description Group: { D_e

1: There is a chair that is **under a table**. it is the **bottom left** chair, D_e^1

2: The chair is the last chair on the left end of the **table**. It is to the **left of the chair next to it**, D_e^2

3: The chair is the one closest to the **door**, D_e^3

4: The chair is the north western most one from the **table** closest to the **door**} D_e^4

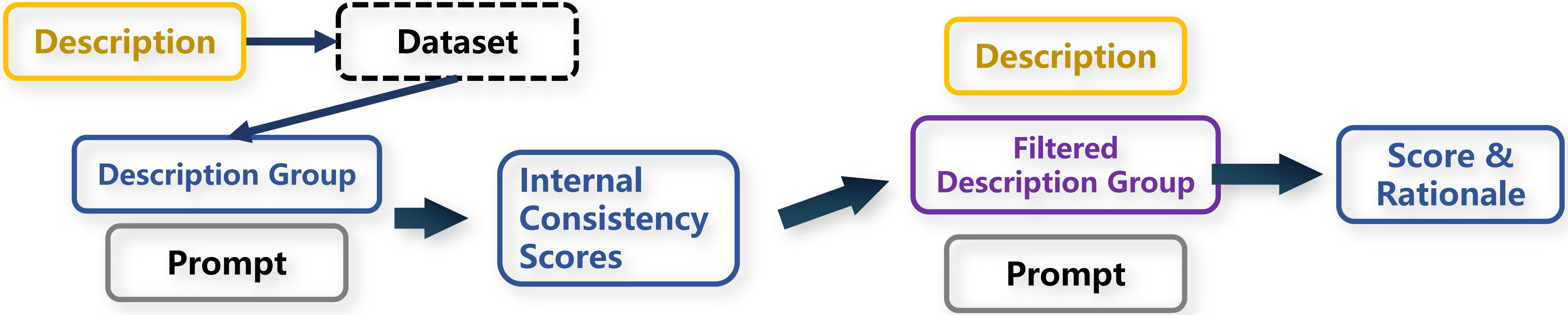
- D_e^1 and D_e^2: ✓ *Score 2*
  Agree on the ***left*** *localization*
- D_e^3 and D_e^4: ✓ *Score 2*
  Agree on the ***relation with the door*** ……

Using the **filtered Description Group**, judging the logical uniformity of the **Description** ... P_c^v

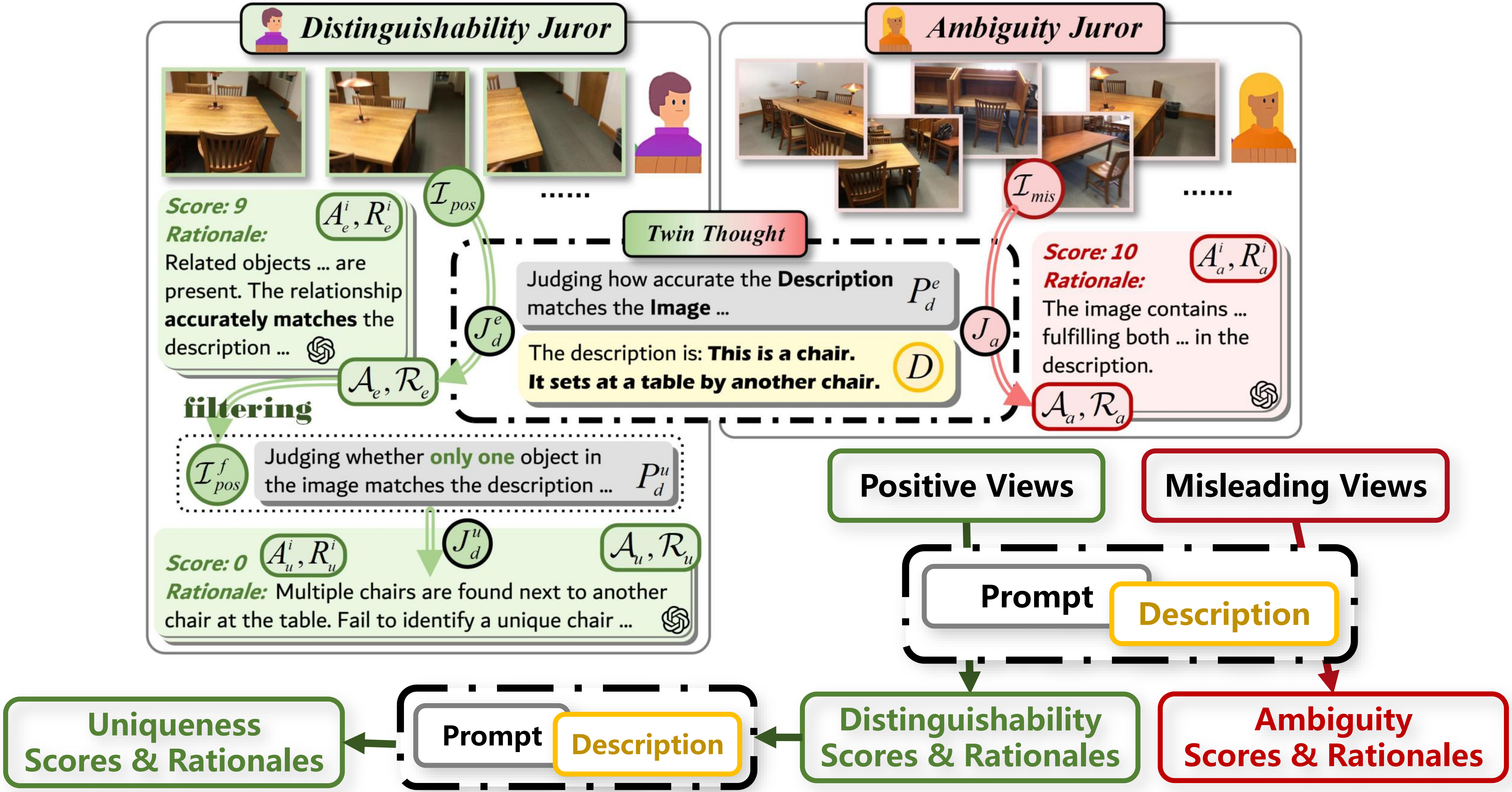
Score: 10

Rationale: *> D_e^1, D_e^2 and D_e^4, all involve the relationship between* ***chair and table****, the description:* **This is a chair. It sets at a table** ***is trustworthy.*** ✓

> Consider the ***left of the chair next to it*** *in D_e^2, the description:* ✓ **It sets by another chair** ***also valid.***



Refer-Judge



Score: 9
Rationale: Related objects ... are present. The relationship **accurately matches** the description ...

Judging how accurate the **Description** matches the **Image** ...

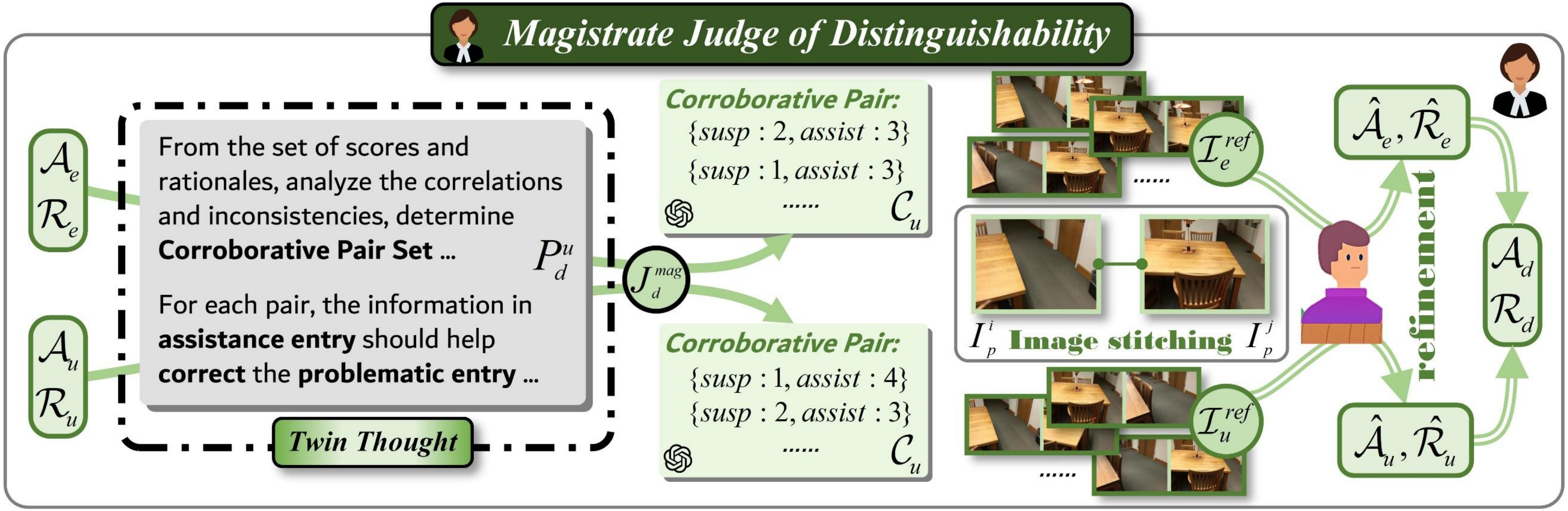
The description is: ***This is a chair. It sets at a table by another chair.***

Score: 10
Rationale: The image contains ... fulfilling both ... in the description.

Judging whether **only one** object in the image matches the description ...

Score: 0
Rationale: Multiple chairs are found next to another chair at the table. Fail to identify a unique chair ...

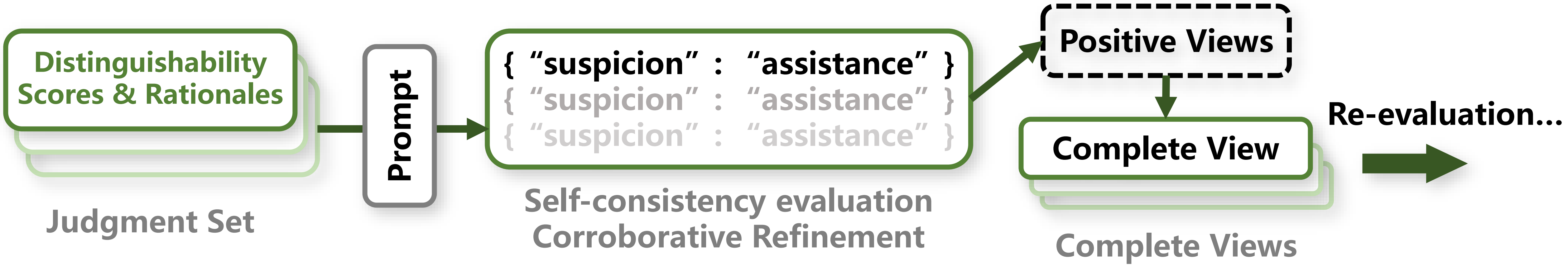
Refer-Judge



Magistrate Judge of Distinguishability

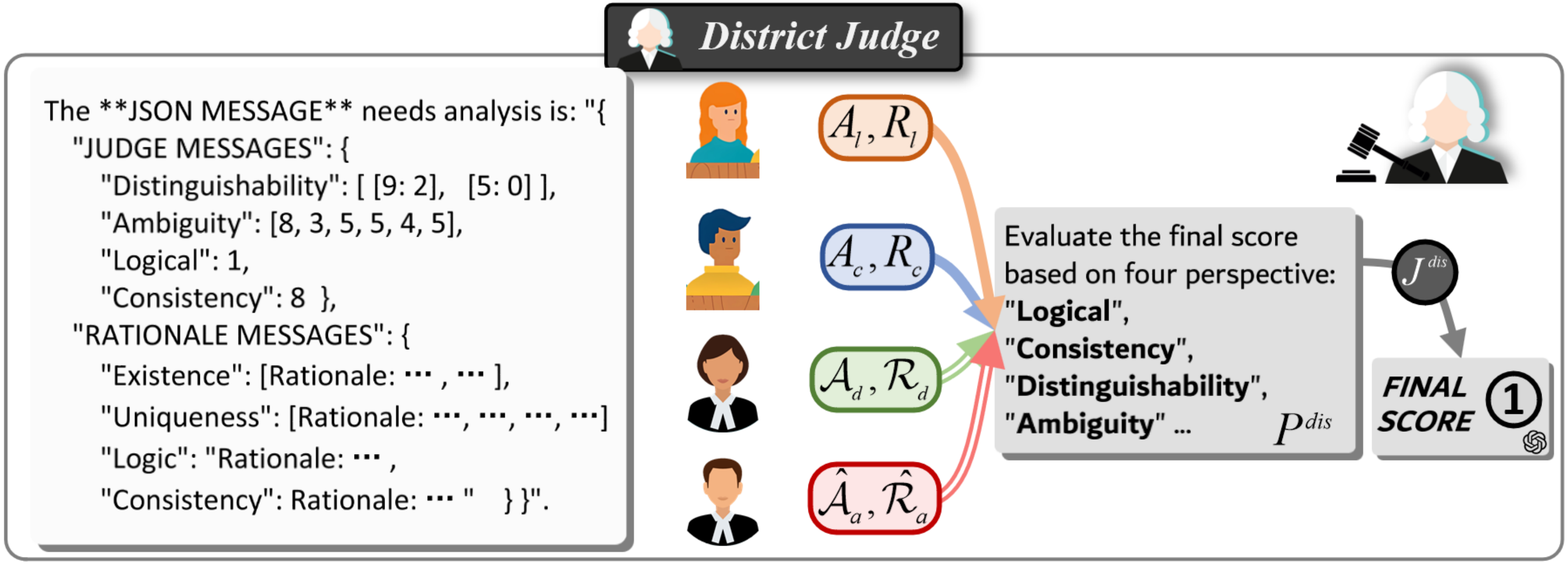
From the set of scores and rationales, analyze the correlations and inconsistencies, determine **Corroborative Pair Set ...**

For each pair, the information in **assistance entry** should help **correct** the **problematic entry ...**



Distinguishability Scores & Rationales
Judgment Set
Prompt
{ "suspicion" : "assistance" }
{ "suspicion" : "assistance" }
{ "suspicion" : "assistance" }
Self-consistency evaluation
Corroborative Refinement
Positive Views
Complete View
Complete Views
Re-evaluation...

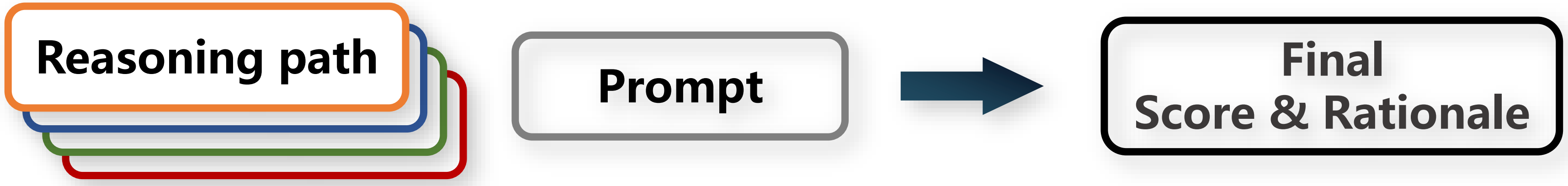
Refer-Judge



District Judge

```
The **JSON MESSAGE** needs analysis is: "{
   "JUDGE MESSAGES": {
      "Distinguishability": [ [9: 2],   [5: 0] ],
      "Ambiguity": [8, 3, 5, 5, 4, 5],
      "Logical": 1,
      "Consistency": 8  },
   "RATIONALE MESSAGES": {
      "Existence": [Rationale: ··· , ··· ],
      "Uniqueness": [Rationale: ···, ···, ···, ···]
      "Logic": "Rationale: ··· ,
      "Consistency": Rationale: ··· "    } }".
```

Evaluate the final score based on four perspective: **"Logical", "Consistency", "Distinguishability", "Ambiguity"** ... P^{dis}

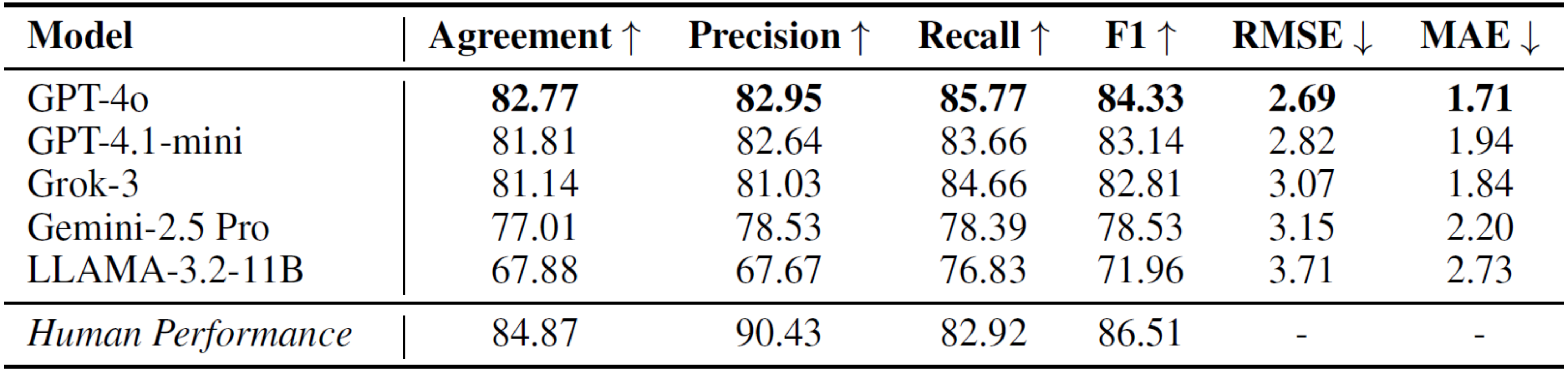


Reasoning path

Prompt

Final Score & Rationale

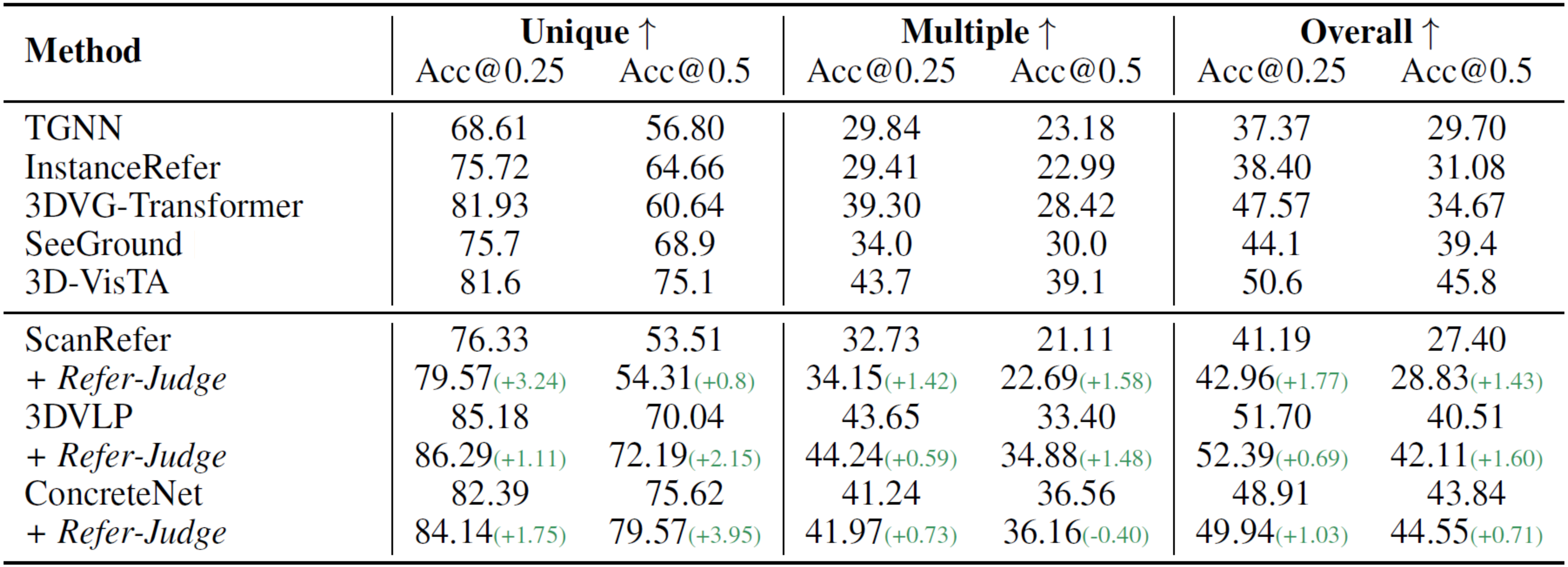
Experiment



Model	Agreement ↑	Precision ↑	Recall ↑	F1 ↑	RMSE ↓	MAE ↓
GPT-4o	**82.77**	**82.95**	**85.77**	**84.33**	**2.69**	**1.71**
GPT-4.1-mini	81.81	82.64	83.66	83.14	2.82	1.94
Grok-3	81.14	81.03	84.66	82.81	3.07	1.84
Gemini-2.5 Pro	77.01	78.53	78.39	78.53	3.15	2.20
LLAMA-3.2-11B	67.88	67.67	76.83	71.96	3.71	2.73
Human Performance	84.87	90.43	82.92	86.51	-	-

- ➢ Refer-Judge achieves **human-level judgment capability**, slightly lagging with human experts.
- ➢ The Refer-Judge algorithm can **generalize** to multiple models.
- ➢ Better base models result in better performance.

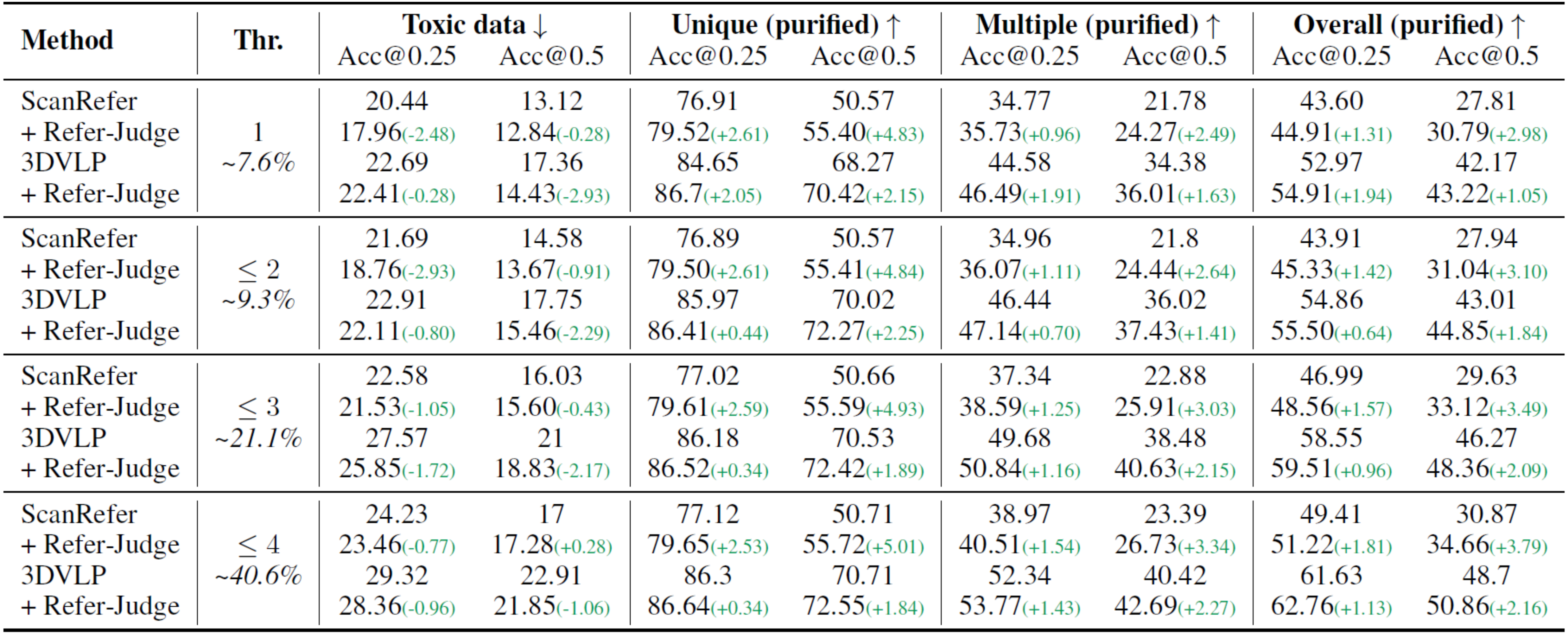
Experiment



Method	Unique ↑		Multiple ↑		Overall ↑	
	Acc@0.25	Acc@0.5	Acc@0.25	Acc@0.5	Acc@0.25	Acc@0.5
TGNN	68.61	56.80	29.84	23.18	37.37	29.70
InstanceRefer	75.72	64.66	29.41	22.99	38.40	31.08
3DVG-Transformer	81.93	60.64	39.30	28.42	47.57	34.67
SeeGround	75.7	68.9	34.0	30.0	44.1	39.4
3D-VisTA	81.6	75.1	43.7	39.1	50.6	45.8
ScanRefer	76.33	53.51	32.73	21.11	41.19	27.40
+ *Refer-Judge*	79.57(+3.24)	54.31(+0.8)	34.15(+1.42)	22.69(+1.58)	42.96(+1.77)	28.83(+1.43)
3DVLP	85.18	70.04	43.65	33.40	51.70	40.51
+ *Refer-Judge*	86.29(+1.11)	72.19(+2.15)	44.24(+0.59)	34.88(+1.48)	52.39(+0.69)	42.11(+1.60)
ConcreteNet	82.39	75.62	41.24	36.56	48.91	43.84
+ *Refer-Judge*	84.14(+1.75)	79.57(+3.95)	41.97(+0.73)	36.16(-0.40)	49.94(+1.03)	44.55(+0.71)

- After **removing the toxic data** from the ScanRefer training set, all baseline achieving **consistent improvements**.

Experiment



Method	Thr.	Toxic data ↓		Unique (purified) ↑		Multiple (purified) ↑		Overall (purified) ↑	
		Acc@0.25	Acc@0.5	Acc@0.25	Acc@0.5	Acc@0.25	Acc@0.5	Acc@0.25	Acc@0.5
ScanRefer		20.44	13.12	76.91	50.57	34.77	21.78	43.60	27.81
+ Refer-Judge	1	17.96(-2.48)	12.84(-0.28)	79.52(+2.61)	55.40(+4.83)	35.73(+0.96)	24.27(+2.49)	44.91(+1.31)	30.79(+2.98)
3DVLP	~*7.6%*	22.69	17.36	84.65	68.27	44.58	34.38	52.97	42.17
+ Refer-Judge		22.41(-0.28)	14.43(-2.93)	86.7(+2.05)	70.42(+2.15)	46.49(+1.91)	36.01(+1.63)	54.91(+1.94)	43.22(+1.05)
ScanRefer		21.69	14.58	76.89	50.57	34.96	21.8	43.91	27.94
+ Refer-Judge	≤ 2	18.76(-2.93)	13.67(-0.91)	79.50(+2.61)	55.41(+4.84)	36.07(+1.11)	24.44(+2.64)	45.33(+1.42)	31.04(+3.10)
3DVLP	~*9.3%*	22.91	17.75	85.97	70.02	46.44	36.02	54.86	43.01
+ Refer-Judge		22.11(-0.80)	15.46(-2.29)	86.41(+0.44)	72.27(+2.25)	47.14(+0.70)	37.43(+1.41)	55.50(+0.64)	44.85(+1.84)
ScanRefer		22.58	16.03	77.02	50.66	37.34	22.88	46.99	29.63
+ Refer-Judge	≤ 3	21.53(-1.05)	15.60(-0.43)	79.61(+2.59)	55.59(+4.93)	38.59(+1.25)	25.91(+3.03)	48.56(+1.57)	33.12(+3.49)
3DVLP	~*21.1%*	27.57	21	86.18	70.53	49.68	38.48	58.55	46.27
+ Refer-Judge		25.85(-1.72)	18.83(-2.17)	86.52(+0.34)	72.42(+1.89)	50.84(+1.16)	40.63(+2.15)	59.51(+0.96)	48.36(+2.09)
ScanRefer		24.23	17	77.12	50.71	38.97	23.39	49.41	30.87
+ Refer-Judge	≤ 4	23.46(-0.77)	17.28(+0.28)	79.65(+2.53)	55.72(+5.01)	40.51(+1.54)	26.73(+3.34)	51.22(+1.81)	34.66(+3.79)
3DVLP	~*40.6%*	29.32	22.91	86.3	70.71	52.34	40.42	61.63	48.7
+ Refer-Judge		28.36(-0.96)	21.85(-1.06)	86.64(+0.34)	72.55(+1.84)	53.77(+1.43)	42.69(+2.27)	62.76(+1.13)	50.86(+2.16)

- A more significant improvement in model performance can be observed on the **purified validation set**.
- The **original model** outperforms the purified model on toxic validation set (*due to toxic prior knowledge*).

Thanks for watching