Seminar

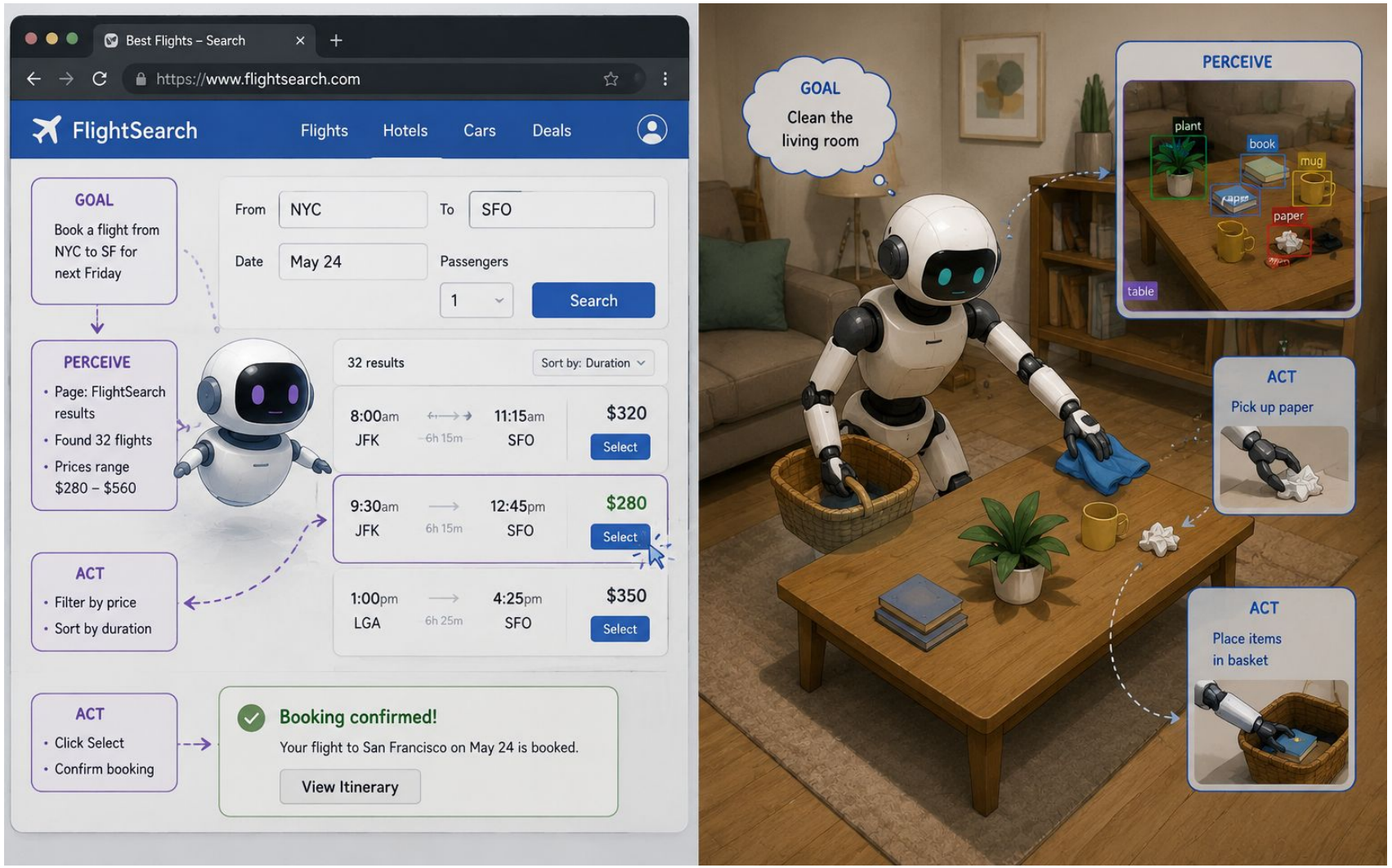
May 28, 2026

Gyeom Hwangbo

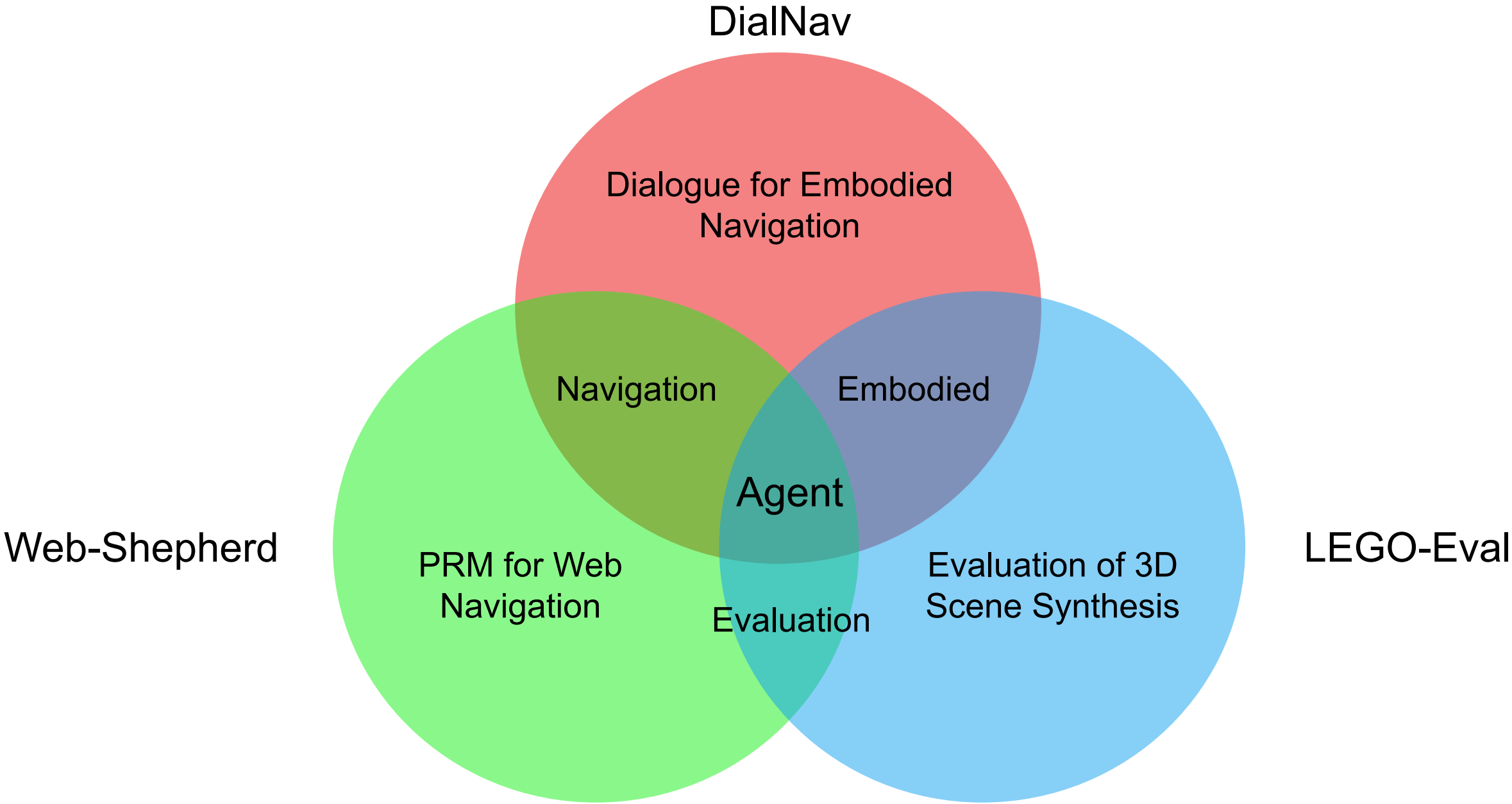
POSTECH, MOAI Lab

AI Agents

"Perceives its environment, reasons, and takes actions to achieve a goal"

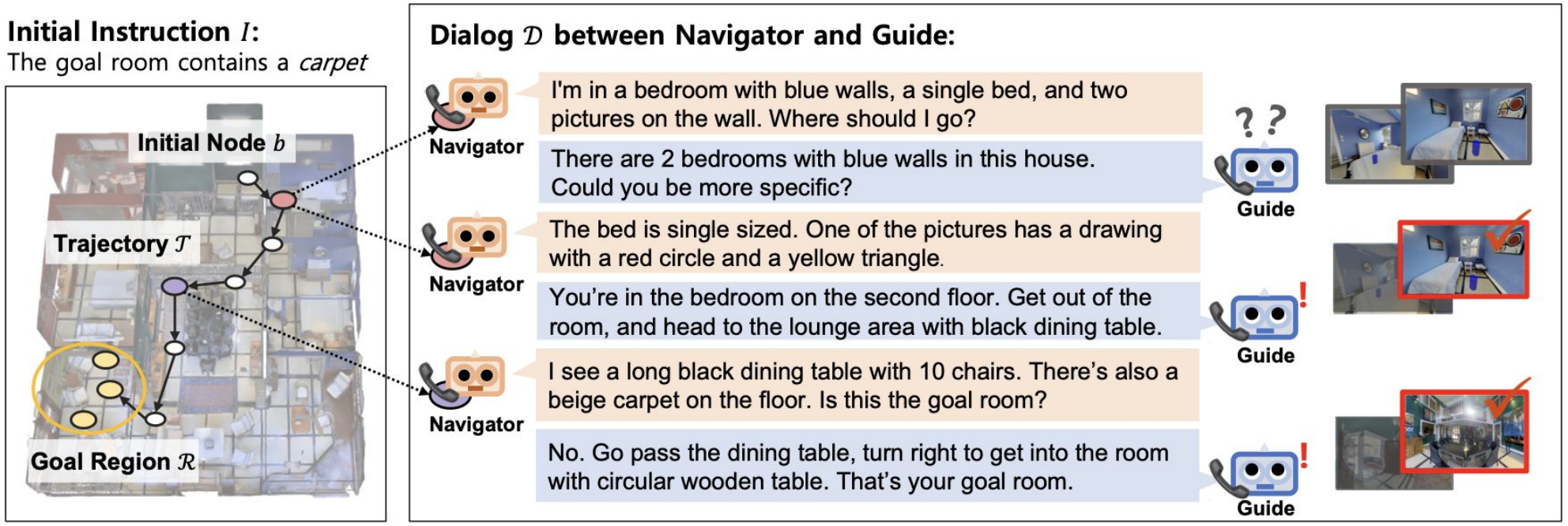


AI Agents


DialNav
Dialogue for Embodied Navigation
Navigation
Embodied
Agent
Web-Shepherd
PRM for Web Navigation
Evaluation
Evaluation of 3D Scene Synthesis
LEGO-Eval

DialNav: Multi-turn Dialog Navigation with a Remote Guide

ICCV 2025 Poster



Vision Dialogue Navigation

DialNav

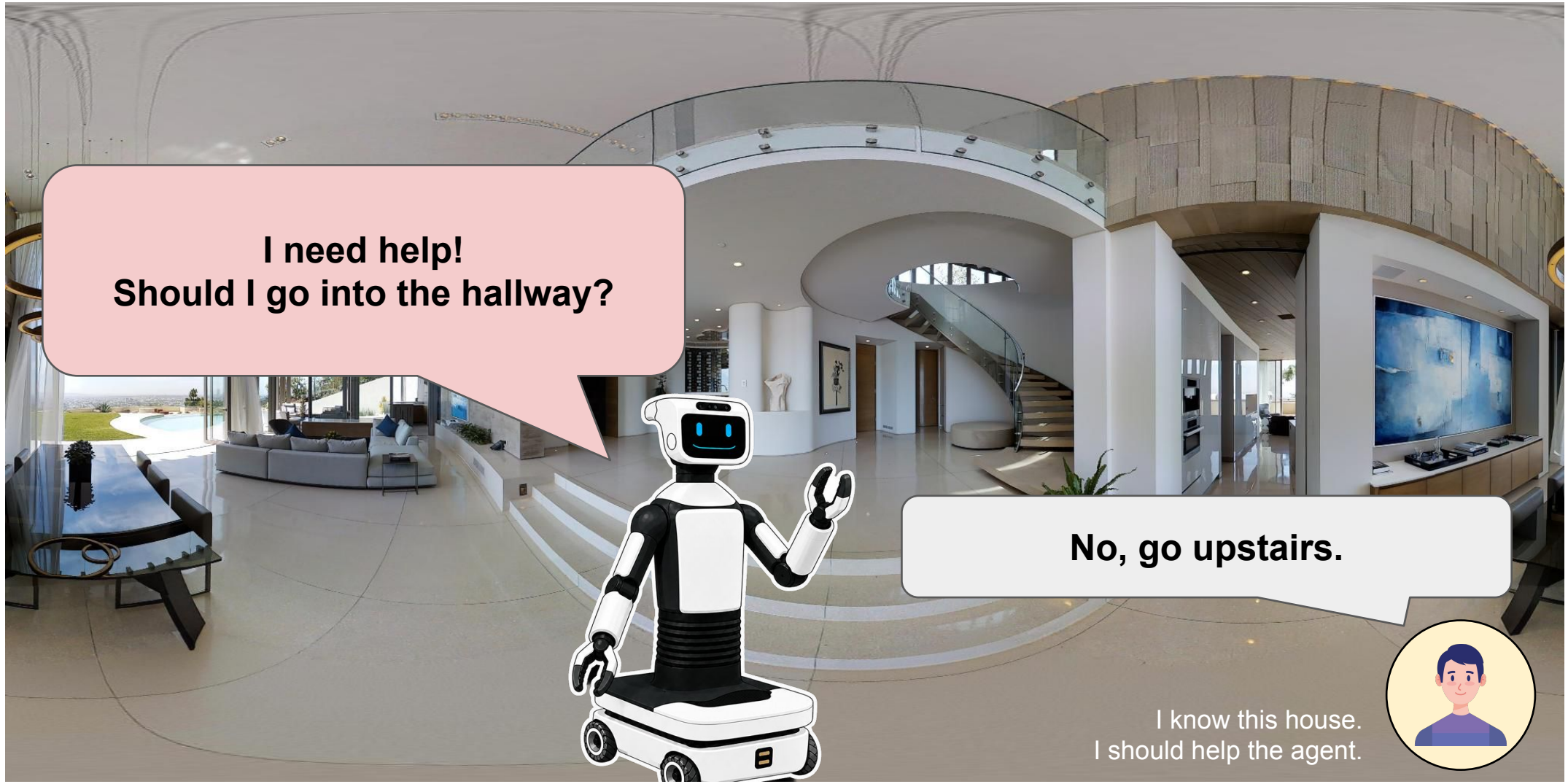
“The goal room contains a mat”



Vision Dialogue Navigation

DialNav

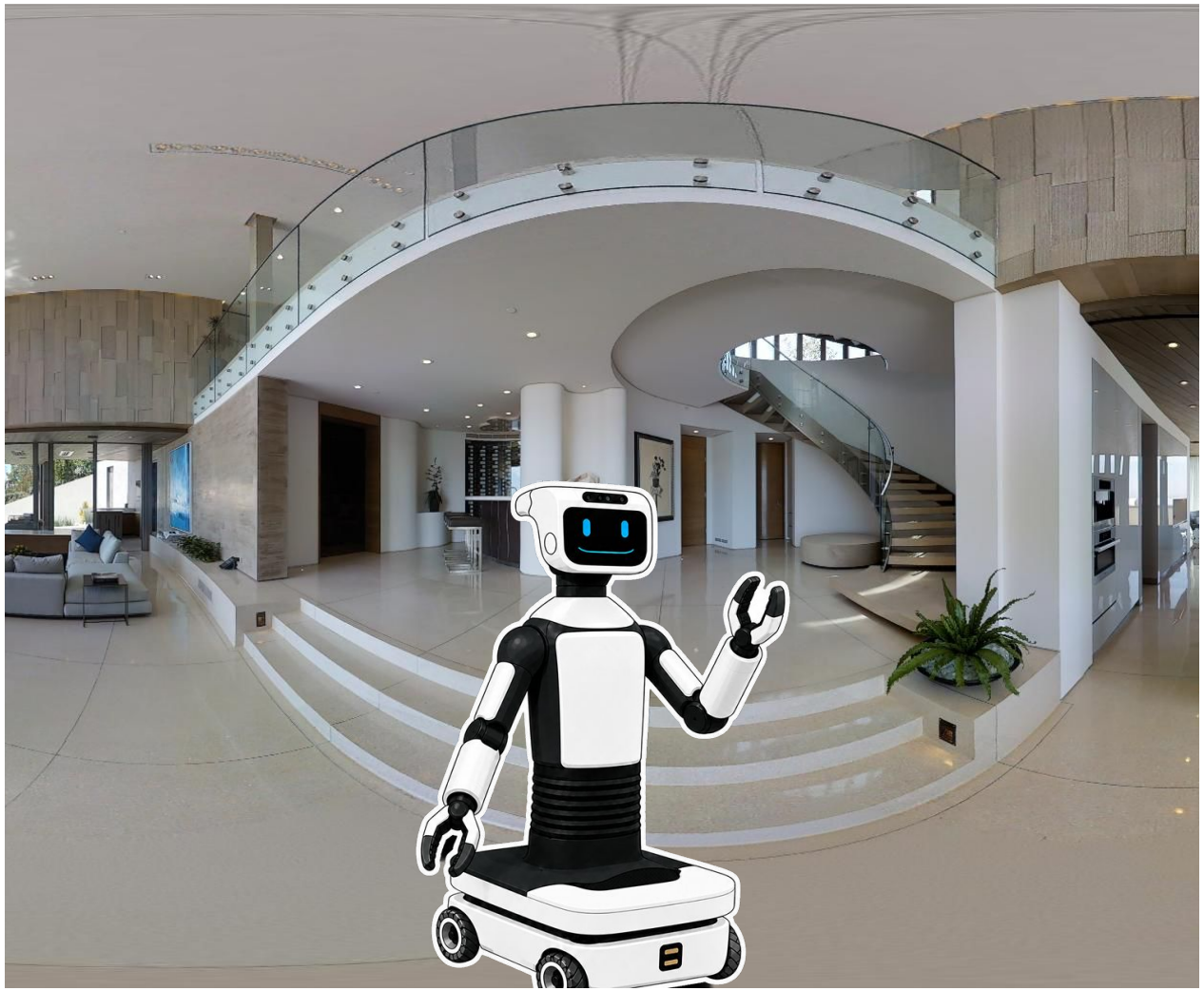
"The goal room contains a mat"



Vision Dialogue Navigation

DialNav

"The Oracle knows the current location of the Navigator"



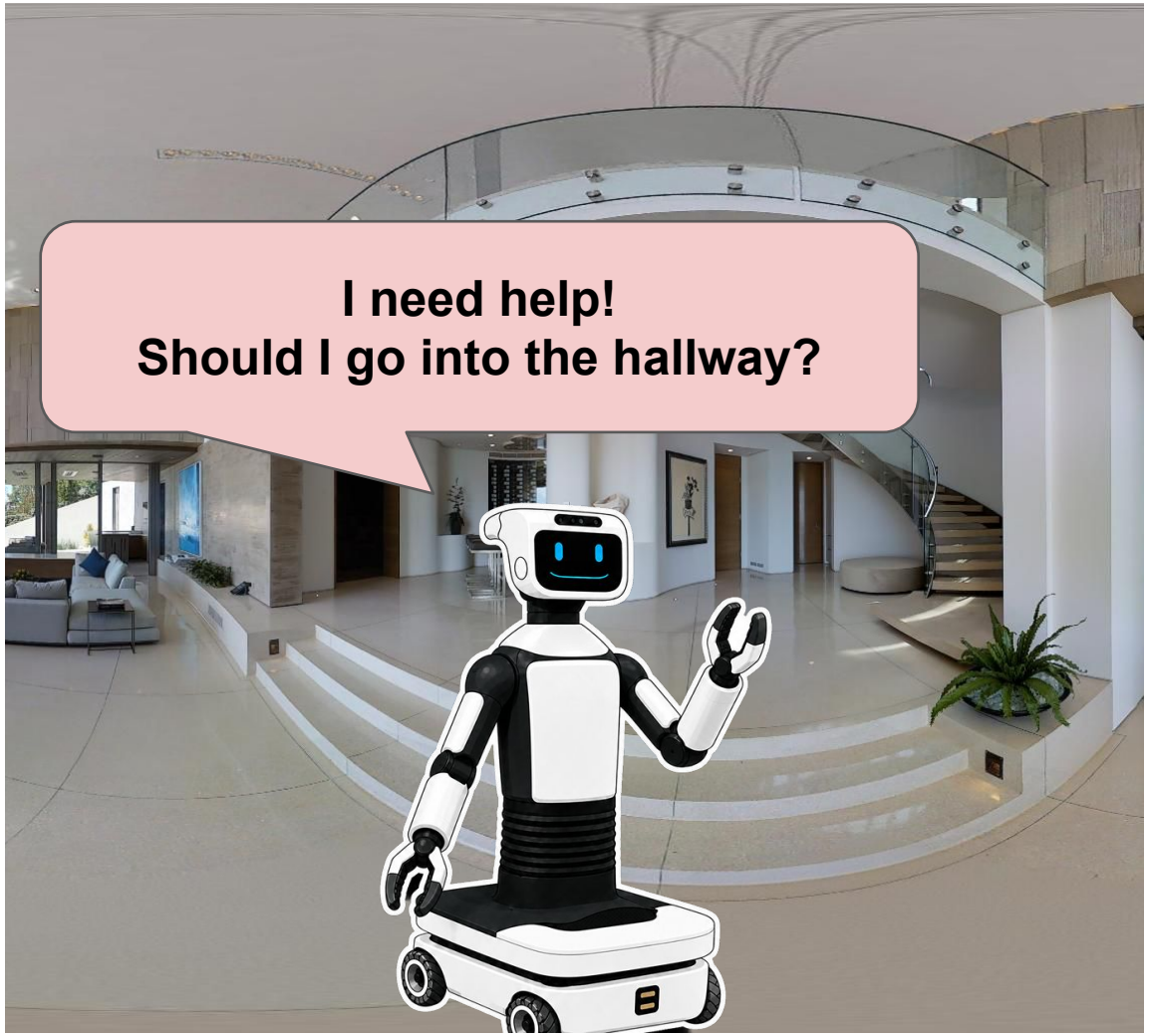
내 손바닥 안이야...



I know everything about you, even when I am not with you.

Vision Dialogue Navigation

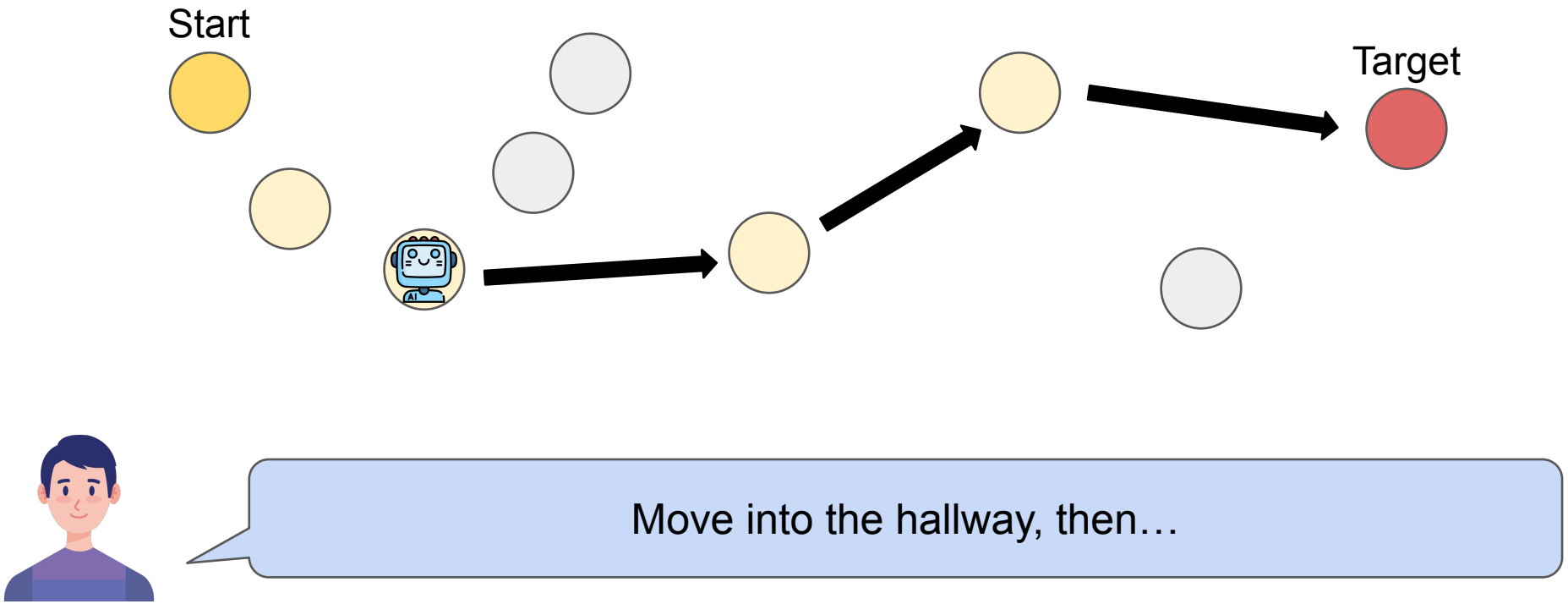
DialNav





Previous VDN Benchmark (CVDN)

The Golden Trajectory to the target destination is given to the oracle.



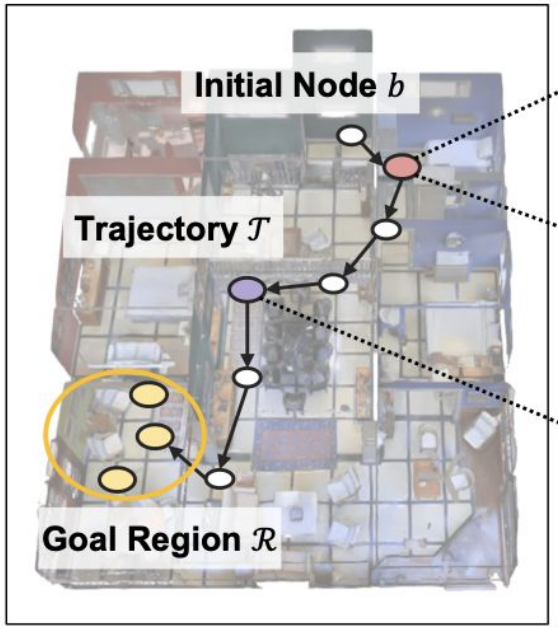
Oracle focuses on trajectory description, rather than the question answering.

DialNav Benchmark

Tasks	Subtasks					Guide
	N	*W*	*Q*	*L*	*A*	
VLN [1]	✓					✗
Just Ask [6]	✓	✓				Omniscient
HANNA [23]	✓	✓				Omniscient
VDN [31]	✓	✓	✓		✓	Omniscient
DialNav(Ours)	✓	✓	✓	✓	✓	Remote

DialNav Benchmark

Initial Instruction I:
The goal room contains a *carpet*



Dialog $\mathcal{D}$ between Navigator and Guide:

Navigator: I'm in a bedroom with blue walls, a single bed, and two pictures on the wall. Where should I go?

Guide: There are 2 bedrooms with blue walls in this house. Could you be more specific?



Navigator: The bed is single sized. One of the pictures has a drawing with a red circle and a yellow triangle.

Guide: You're in the bedroom on the second floor. Get out of the room, and head to the lounge area with black dining table.



Navigator: I see a long black dining table with 10 chairs. There's also a beige carpet on the floor. Is this the goal room?

Guide: No. Go pass the dining table, turn right to get into the room with circular wooden table. That's your goal room.



DialNav Benchmark: RAIN Dataset

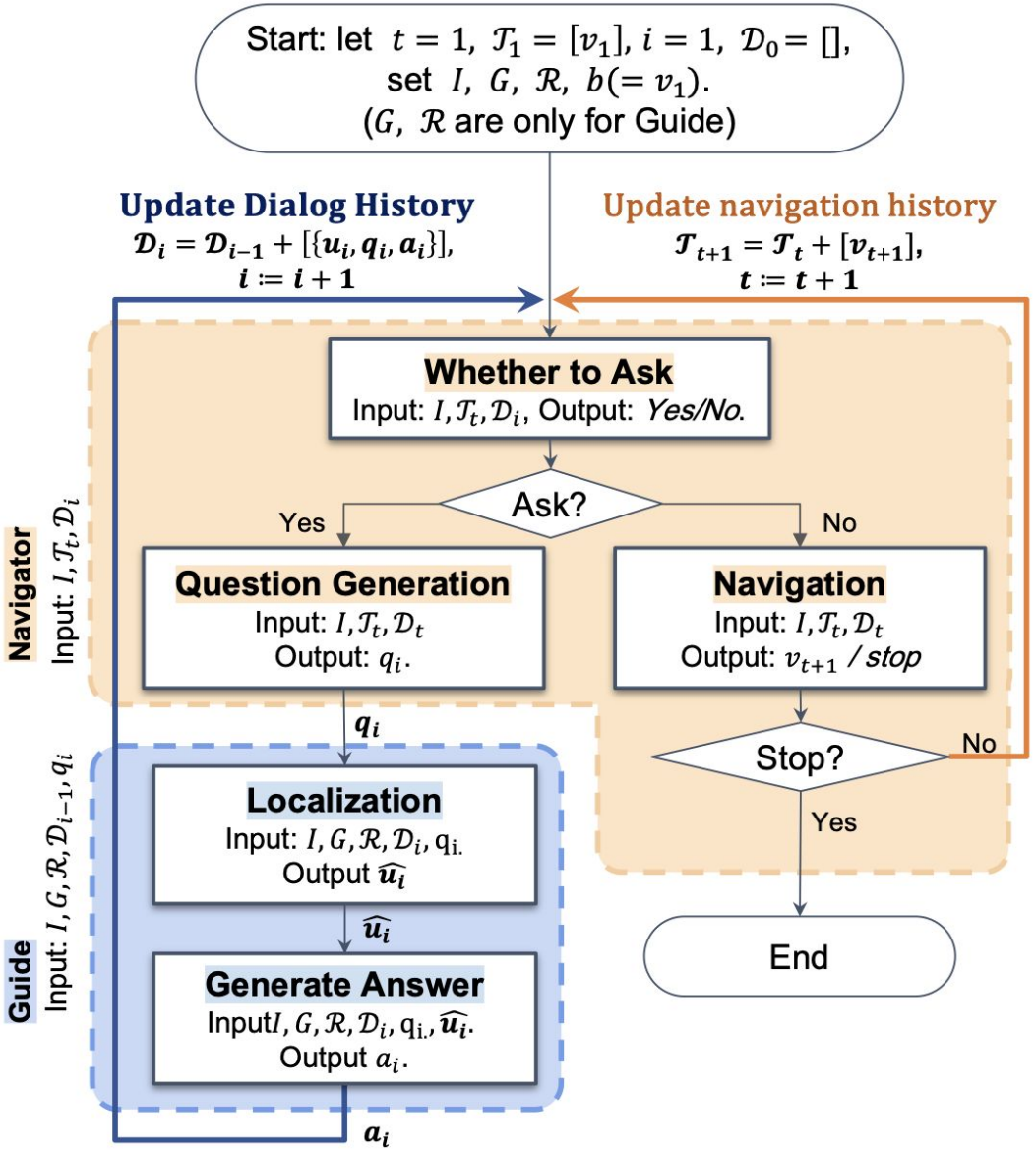
Q:	Think I'm at the room what you've noticed me. On my left side, there is a table made of transparent glass and on my right side, there is a Queen side bed. Also with the two doors opened to the balcony.
A:	That's right. Go out the left door to the balcony, you'll see two black stools. That balcony is your target room.

(a) RAIN

Q:	Should I exit this room?
A:	Yes, go out blue door.

(b) CVDN

DialNav: Baseline Model



DialNav: Quantitative Results

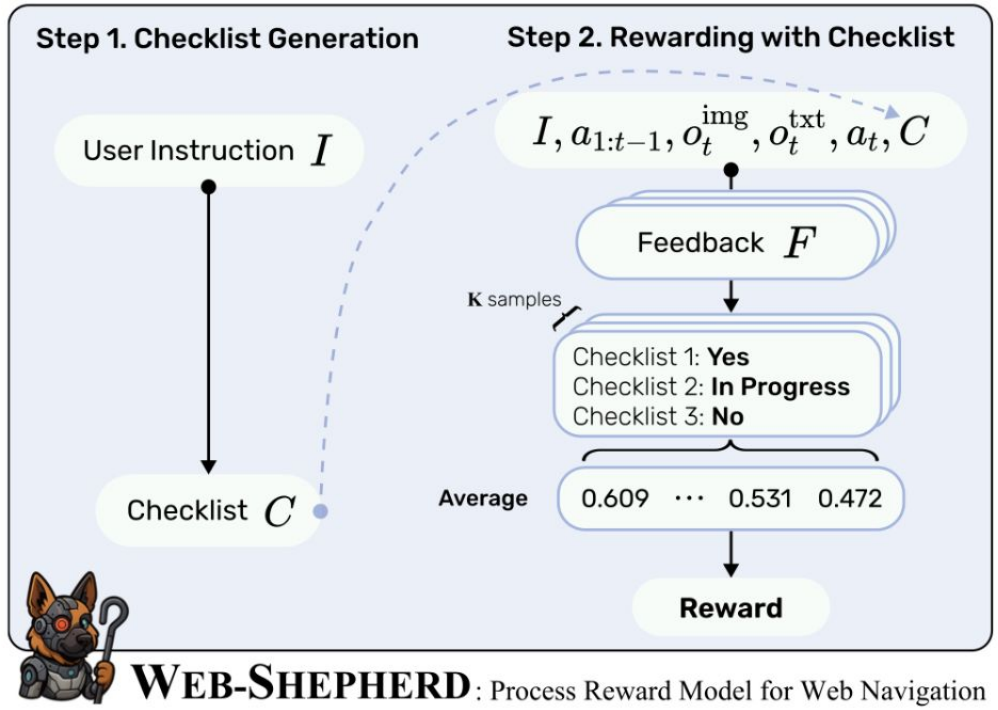
Setup	SR↑	OSR↑	SPL↑	NE↓	NSC	DTC	LE↓
Val Seen							
(1) Nav. only	18.2	28.6	17.3	14.5	19.6	-	-
(2) +Dialog	27.0	34.5	25.4	11.5	16.4	1.9	22.6
(3) +GT Loc.	**31.4**	**43.3**	**28.6**	**9.9**	18.3	1.7	-
Val Unseen							
(4) Nav. only	15.4	**33.3**	10.3	14.9	20.8	-	-
(5) +Dialog	13.9	20.7	10.0	16.3	14.2	2.9	21.9
(6) +GT Loc.	**19.8**	29.0	**15.5**	**12.7**	16.4	2.6	-
Test							
(7) Nav. only	12.7	**28.6**	8.0	14.8	22.8	-	-
(8) +Dialog	11.9	19.0	8.2	17.3	15.4	2.9	22.9
(9) +GT Loc.	**17.3**	23.7	**12.2**	**14.2**	17.4	2.7	-

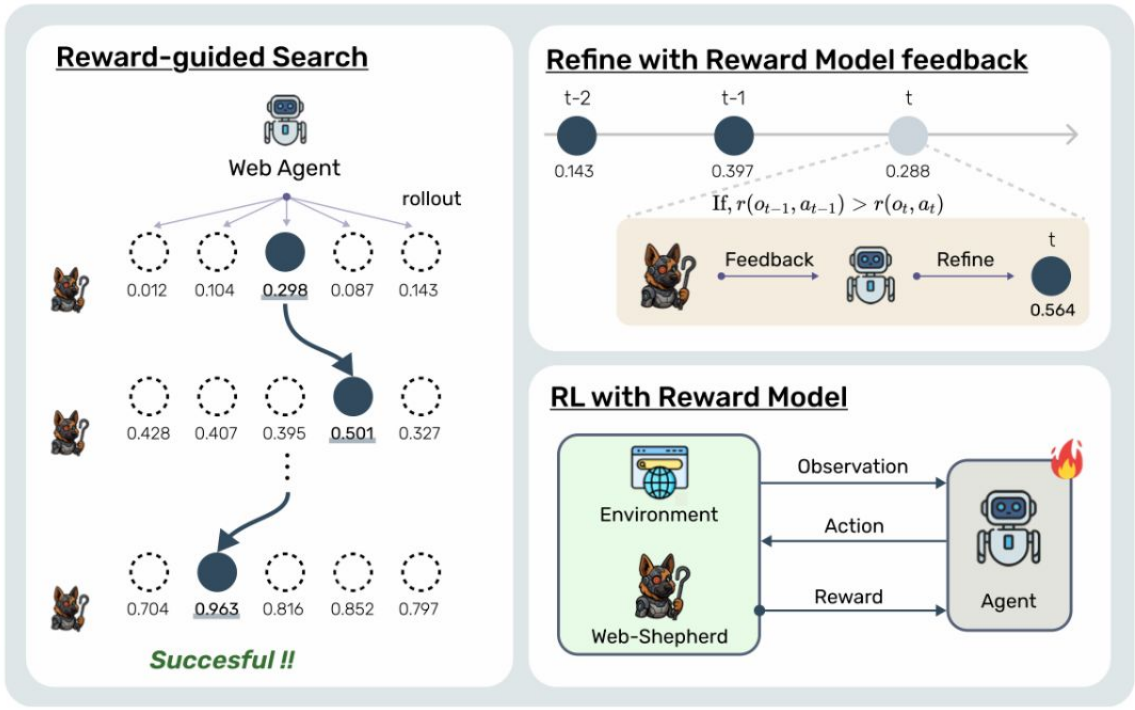
DialNav: Quantitative Results

Setup	SR↑	OSR↑	SPL↑	NE↓	NSC	DTC	LE↓
Val Seen							
(1) Nav. only	18.2	28.6	17.3	14.5	19.6	-	-
(2) +Dialog	27.0	34.5	25.4	11.5	16.4	1.9	22.6
(3) +GT Loc.	**31.4**	**43.3**	**28.6**	**9.9**	18.3	1.7	-
Val Unseen							
(4) Nav. only	15.4	**33.3**	10.3	14.9	20.8	-	-
(5) +Dialog	13.9	20.7	10.0	16.3	14.2	2.9	21.9
(6) +GT Loc.	**19.8**	29.0	**15.5**	**12.7**	16.4	2.6	-
Test							
(7) Nav. only	12.7	**28.6**	8.0	14.8	22.8	-	-
(8) +Dialog	11.9	19.0	8.2	17.3	15.4	2.9	22.9
(9) +GT Loc.	**17.3**	23.7	**12.2**	**14.2**	17.4	2.7	-

Web-Shepherd: Advancing PRMs for Reinforcing Web Agents

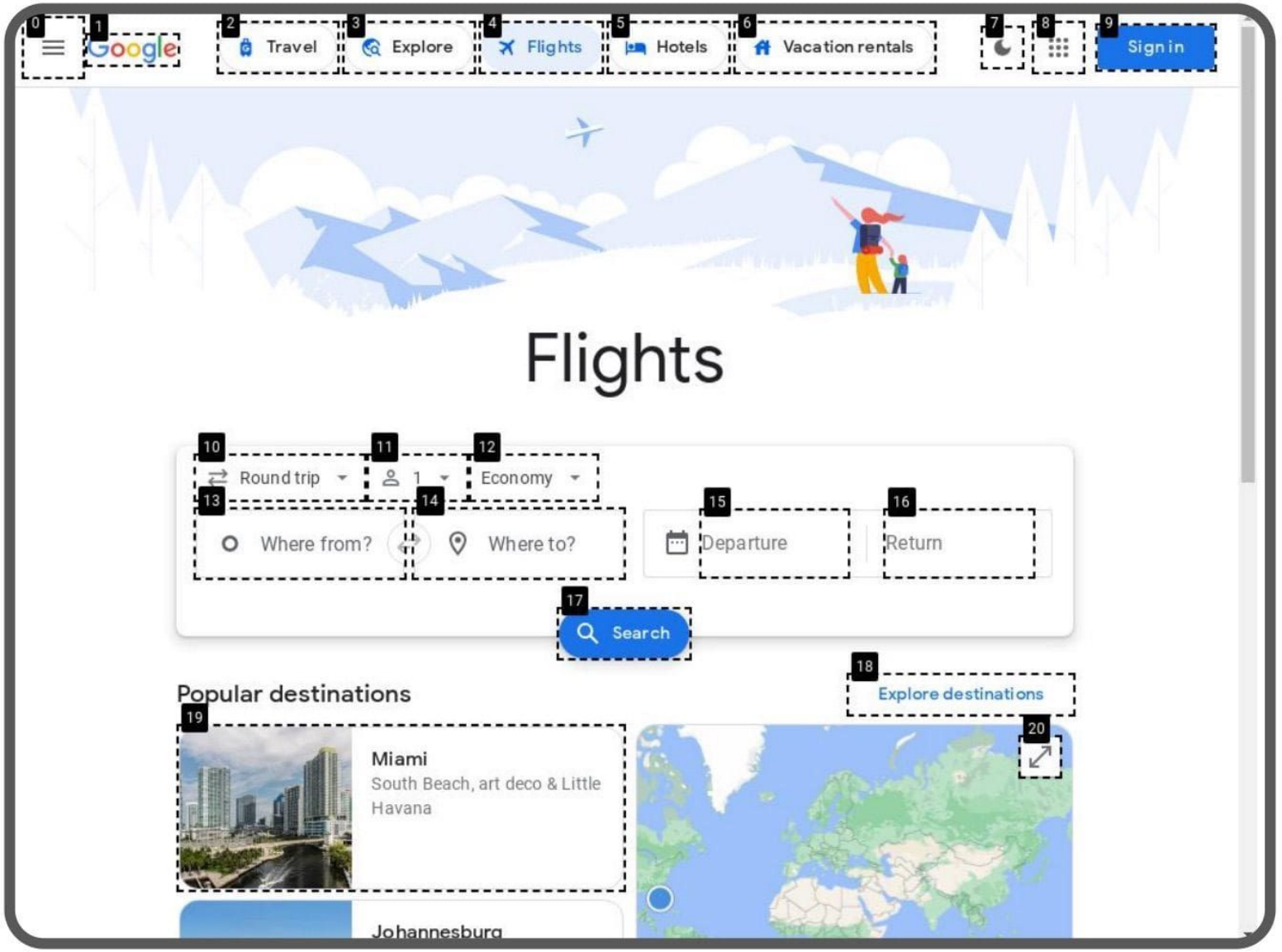
Neurips 2025 Spotlight





Web Agent Navigation

Web-Shepherd



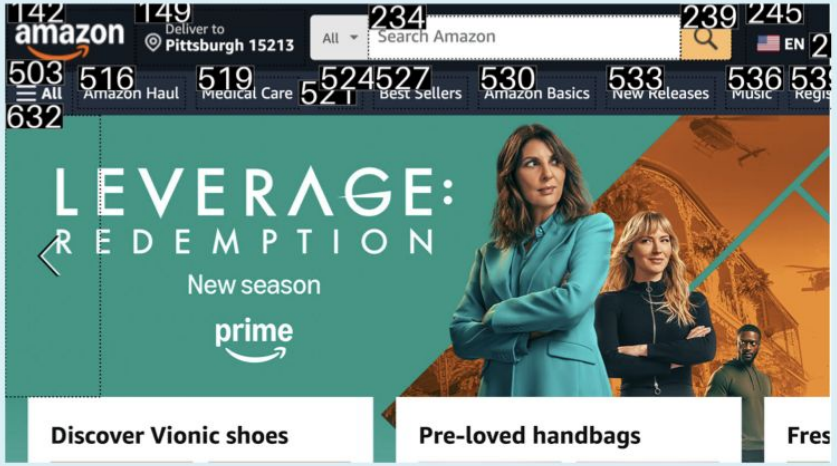


Web Agent Navigation: Inputs

User Instruction (I)

Buy Sony WH-1000XM4 in Amazon.

Observation $(o^{\text{img}}, o^{\text{txt}})$



[142] link 'Amazon', url='https://www.amazon.com/ref=nav_logo'
[149] button 'Deliver to Pittsburgh 15213\u200c'
[234] searchbox 'Search Amazon', autocomplete='list', ...
[239] button 'Go'
[245] link 'Choose a language for shopping in Amazon United ..
[251] button 'Expand to Change Language or Country'
[503] button 'Open All Categories Menu', expanded=False
...

Action (a)

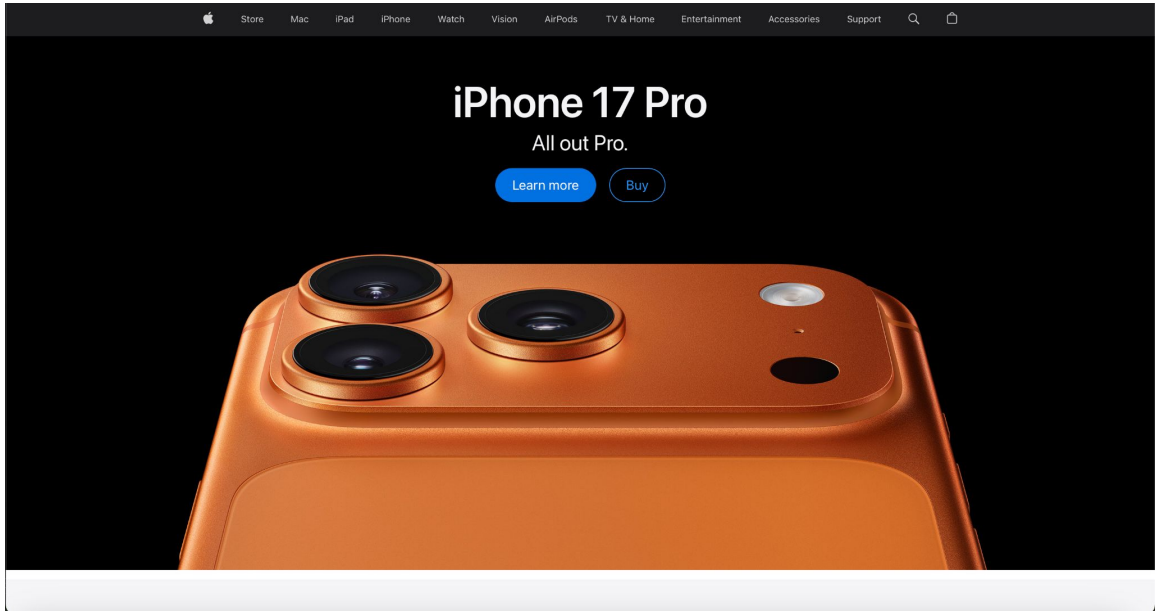
`fill(234, “Sony WH-1000XM4”)`

Web Agent Navigation: Long Horizon Task

Web-Shepherd

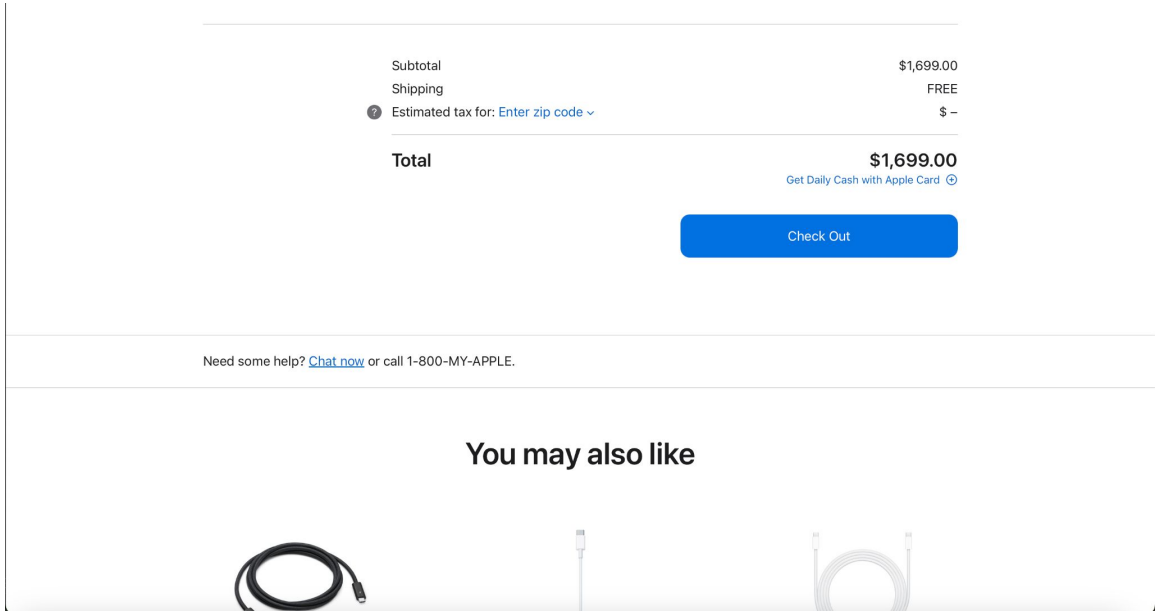
Can you purchase 2 macbook pros, one ipad, and one airpods pro 2 for me?

Timestep 0



...

Timestep t



<u>Long horizon task!</u> = Agents must reason across multiple steps and maintain goal-directed planning

Web Agent Navigation: Long Horizon Task

For LLMs to perform long-horizon navigation tasks…

- Use a reward model at test-time to perform search (Best-of-n)
- Reinforcement Learning (RLHF)

Web Agent Navigation: Long Horizon Task

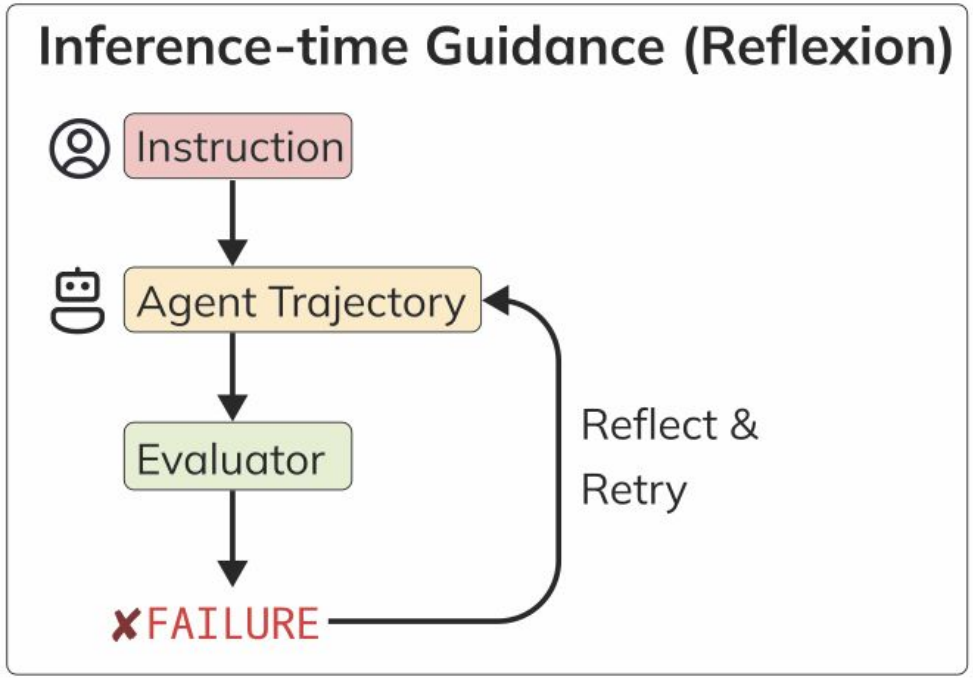
For LLMs to perform long-horizon navigation tasks…

- Use a reward model at test-time to perform search (Best-of-n)
- Reinforcement Learning (RLHF)

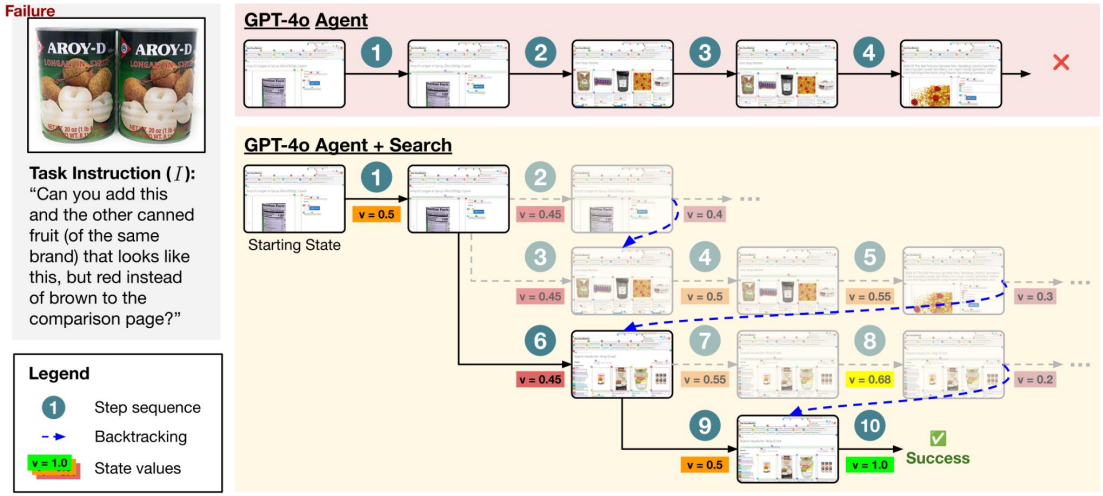
→ Reward model for web-navigation under-explored!

MLLM as Evaluator for Web Navigation

Prior works employ MLLMs as evaluators in inference-time algorithms



[1] Evaluation after action is taken for replanning

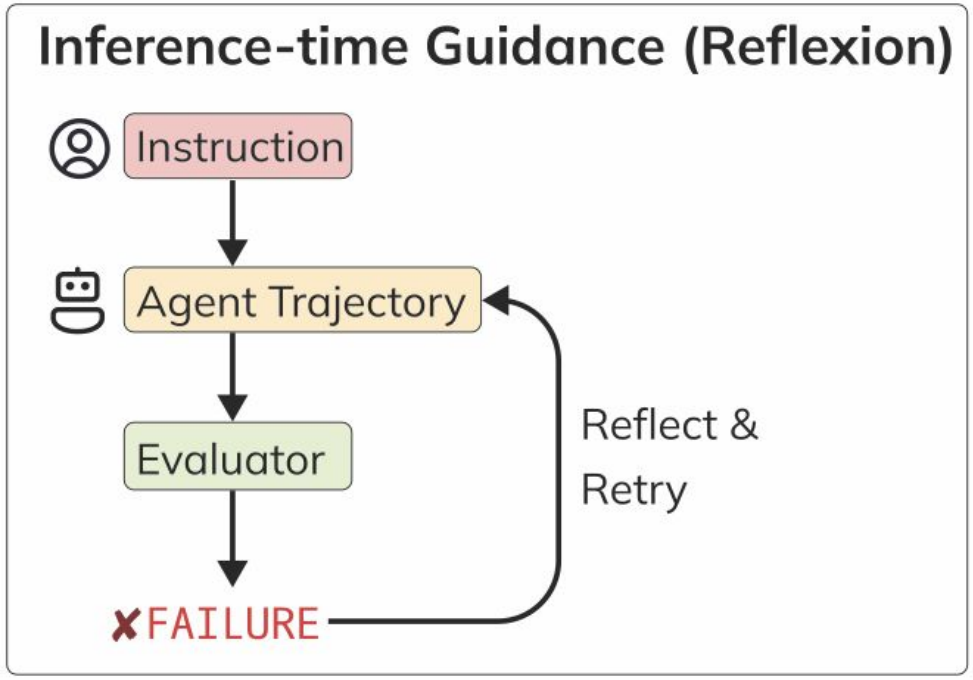


[2] Value function for tree search

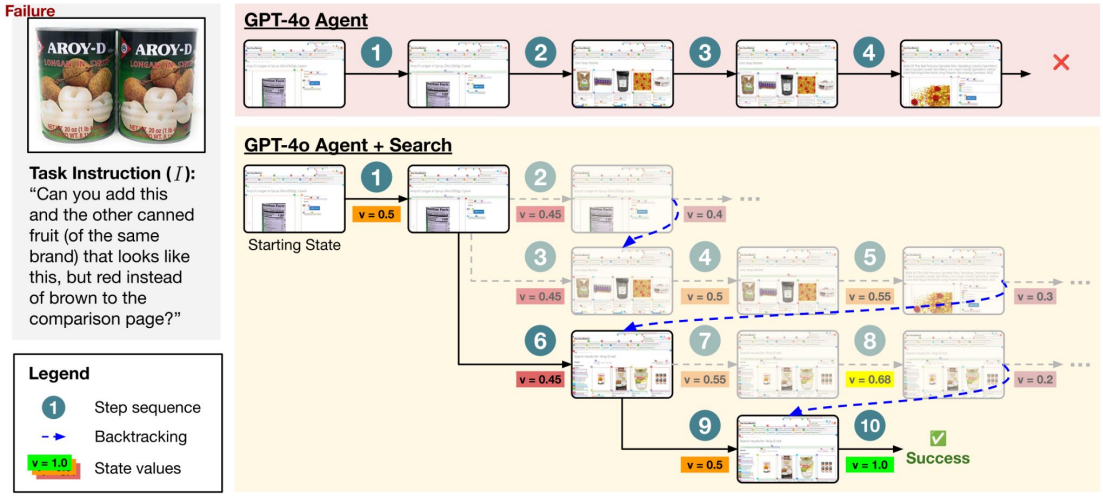
[1] Jiayi Pan, Yichi Zhang, Nicholas Tomlin, Yifei Zhou, Sergey Levine, and Alane Suhr. Autonomous evaluation and refinement of digital agents.
[2] Jing Yu Koh, Stephen McAleer, Daniel Fried, and Ruslan Salakhutdinov. Tree search for language model agents.

MLLM as Evaluator for Web Navigation

Prior works employ MLLMs as evaluators in inference-time algorithms



[1] Evaluation after action is taken for replanning



[2] Value function for tree search

Too costly!

PRM for Web Navigation

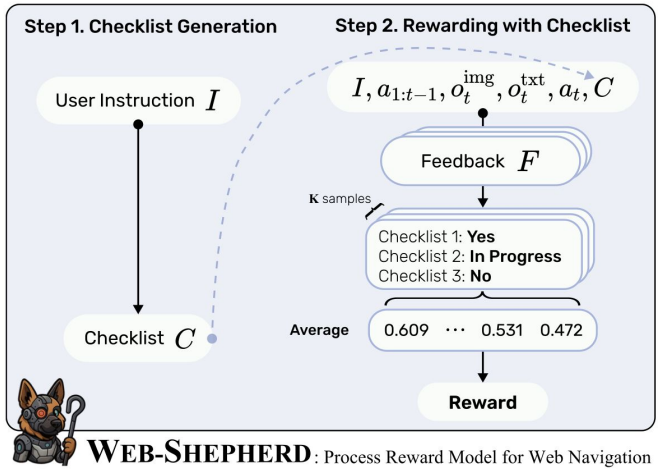
Process reward model trained for evaluating trajectories of web navigation

WebPRM Collection
(Training Data)

WebShepherd
(Process Reward Model)

WebRewardBench
(Bench for Web-nav PRM)





Model	Inputs	Checklist	Mind2Web						WebArena	
			Cross-Task		*Cross-Website*		*Cross-Domain*		*Test*	
			MRR	Acc. (traj)	MRR	Acc. (traj)	MRR	Acc. (traj)	MRR	Acc. (traj)
GPT-4o-mini	T	✗	47.5	0.0	47.6	13.5	45.4	0.8	34.4	5.0
	T	✓	63.9	5.0	66.1	12.8	63.3	12.4	60.0	15.0
	T + I	✗	48.2	0.0	49.3	0.0	49.5	0.8	38.7	0.0
	T + I	✓	58.8	2.5	64.4	5.1	63.0	4.1	53.8	5.0
GPT-4o	T	✗	56.9	5.0	55.8	2.6	59.8	3.3	59.2	15.0
	T	✓	67.4	7.5	70.3	5.1	70.2	11.6	69.7	15.0
	T + I	✗	52.5	5.0	52.2	0.0	52.8	1.7	49.7	5.0
	T + I	✓	62.4	5.0	68.1	15.4	65.1	6.6	59.7	10.0
Qwen-2.5-VL-72B	T	✗	55.7	5.0	51.8	0.0	54.2	1.7	54.6	5.0
	T	✓	59.4	0.0	62.4	0.0	57.9	1.7	52.3	5.0
	T + I	✗	50.1	2.5	47.6	0.0	49.8	0.8	43.1	0.0
	T + I	✓	52.9	2.5	53.5	2.6	52.0	2.5	47.3	0.0
WEB-SHEPHERD (3B)	T	✓	87.6	55.0	**88.0**	43.6	87.2	47.1	91.1	60.0
	T+I	✓	85.0	42.5	87.3	41.0	84.4	37.2	92.5	65.0
WEB-SHEPHERD (8B)	T	✓	**88.3**	**57.5**	87.9	**51.3**	**91.3**	**61.2**	**97.8**	**85.0**

WebPRM Collection

Instruction

Can you check the warranty period for the Apple MacBook Pro 14" MW2U3LL/A (Late 2024) 14.2" Laptop Computer in Space Black?

Observation

[131] link 'please check back regularly for restocks', url='https://www.microcenter.com/search/...'
[137] link 'Micro Center', url='https://www.microcenter.com/'
[146] textbox 'Search for product'
[147] button 'Begin Product Search'
[155] link 'Shippable Items', url='https://www.microcenter.com/'



Checklist

Checklist 1: Search for Product
- **Goal:** Enter the product name into the search bar and initiate the search to locate the product listing.

Checklist 2: Access Product Page
- **Goal:** Identify and click on the specific product listing to access detailed information about the Apple MacBook Pro 14" MW2U3LL/A.

Checklist 3: Locate Warranty Information
- **Goal:** Scroll through the product detail page to find and verify the warranty details in the "WARRANTY & SUPPORT" section.

Chosen Action

THOUGHT: The goal is to check the warranty period for the Apple ...

ACTION: fill('146', 'Apple MacBook Pro 14" MW2U3LL/A')

Rejected Action

THOUGHT: The current webpage does not directly provide information ...

ACTION: click('1976')

Feedback & Judgement

Checklist 1: Search for Product
- The agent has filled the search bar with the exact product name which ...

Checklist 2: Access Product Page
- Since the agent has not yet located ...

Checklist 3: Locate Warranty Information
- The agent's current behavior has not...

Checklist 1: **Yes**
Checklist 2: **No**
Checklist 3: **No**

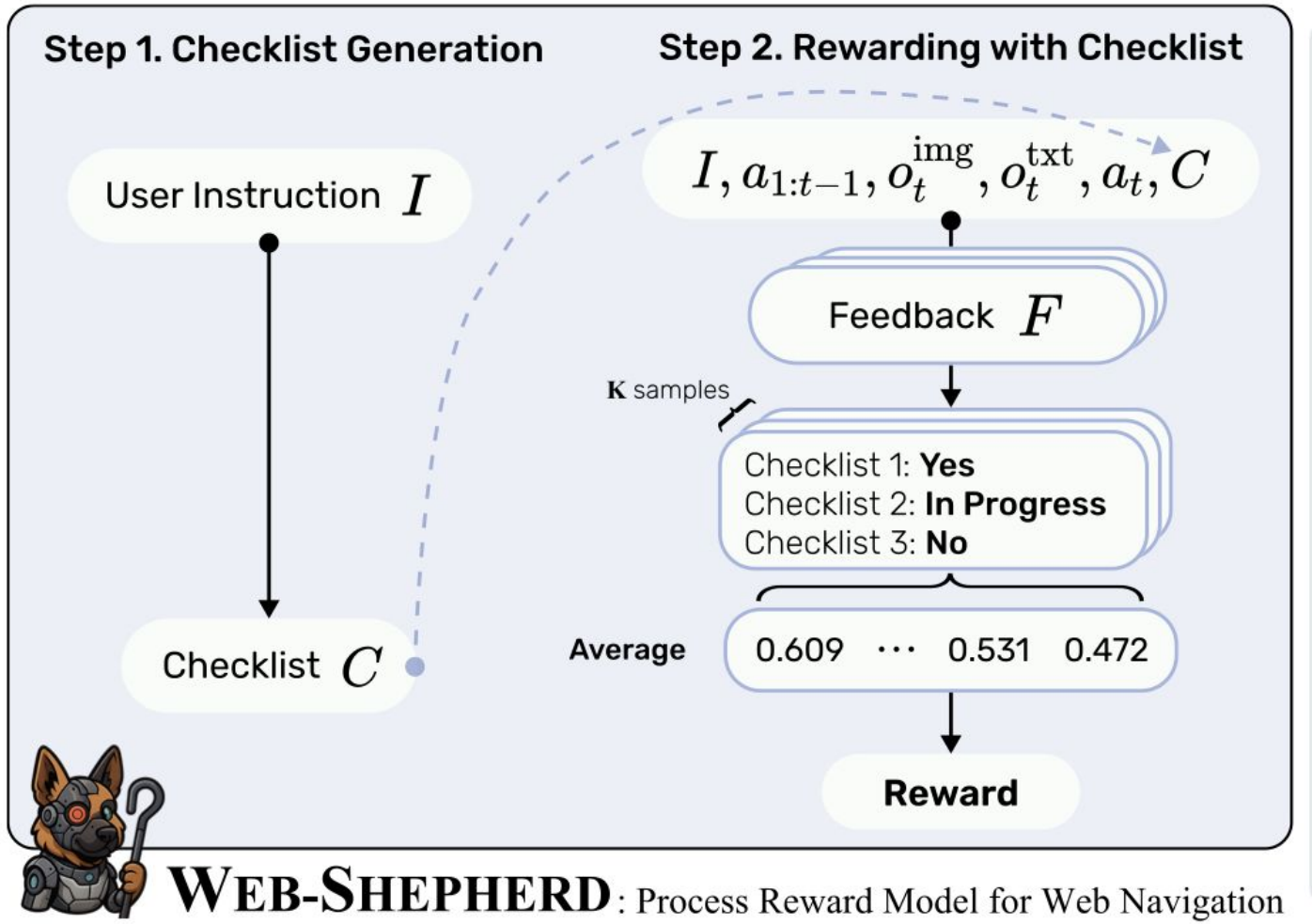
Checklist 1: Search for Product
- ... This step fails to align with the validation criteria of displaying ...

Checklist 2: Access Product Page
- Since the agent has not yet located ...

Checklist 3: Locate Warranty Information
- The agent's current behavior has not ...

Checklist 1: **No**
Checklist 2: **No**
Checklist 3: **No**

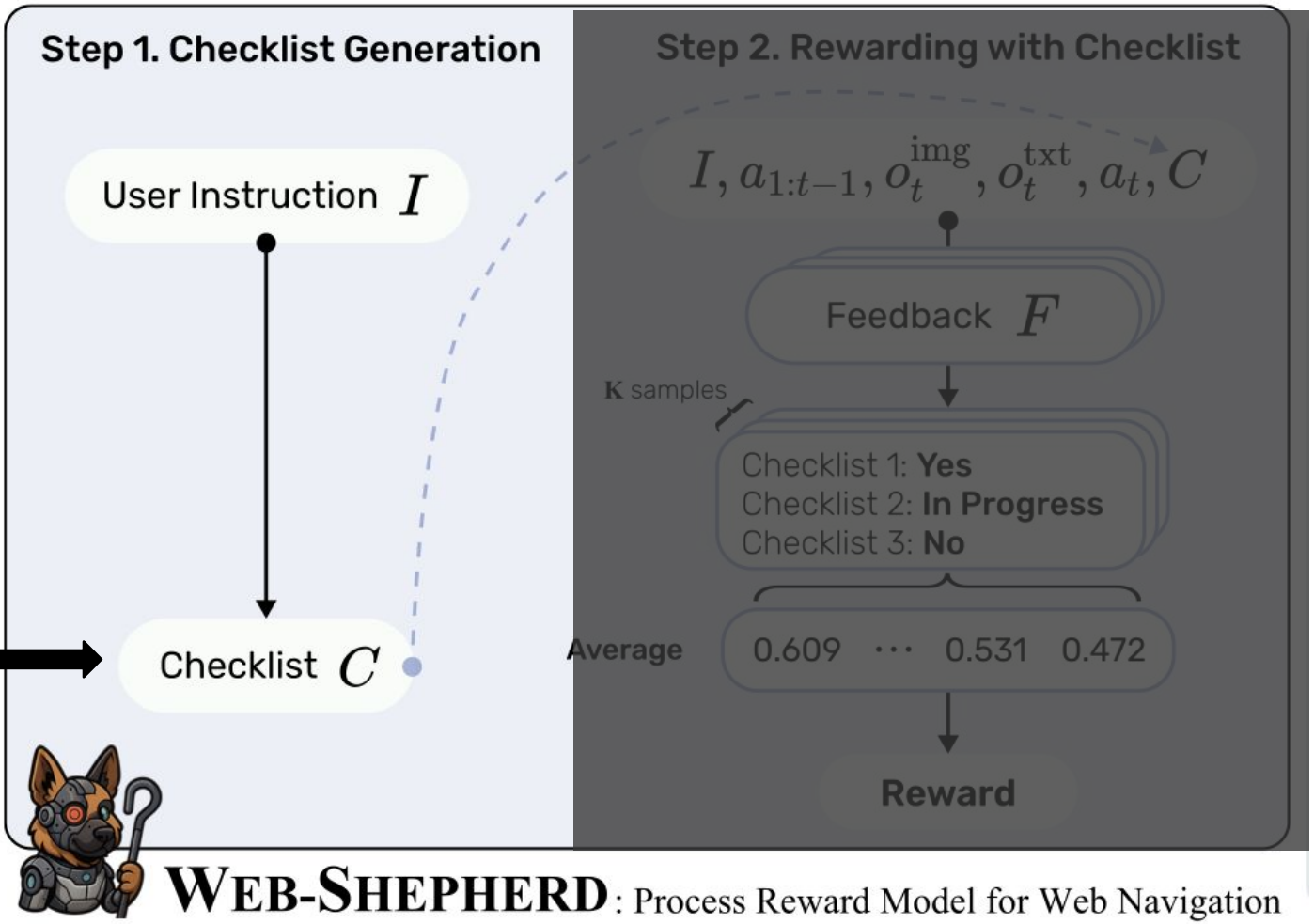
Web-Shepherd



* The model is trained to generate the corresponding feedback and judgement in auto-regressive manner.

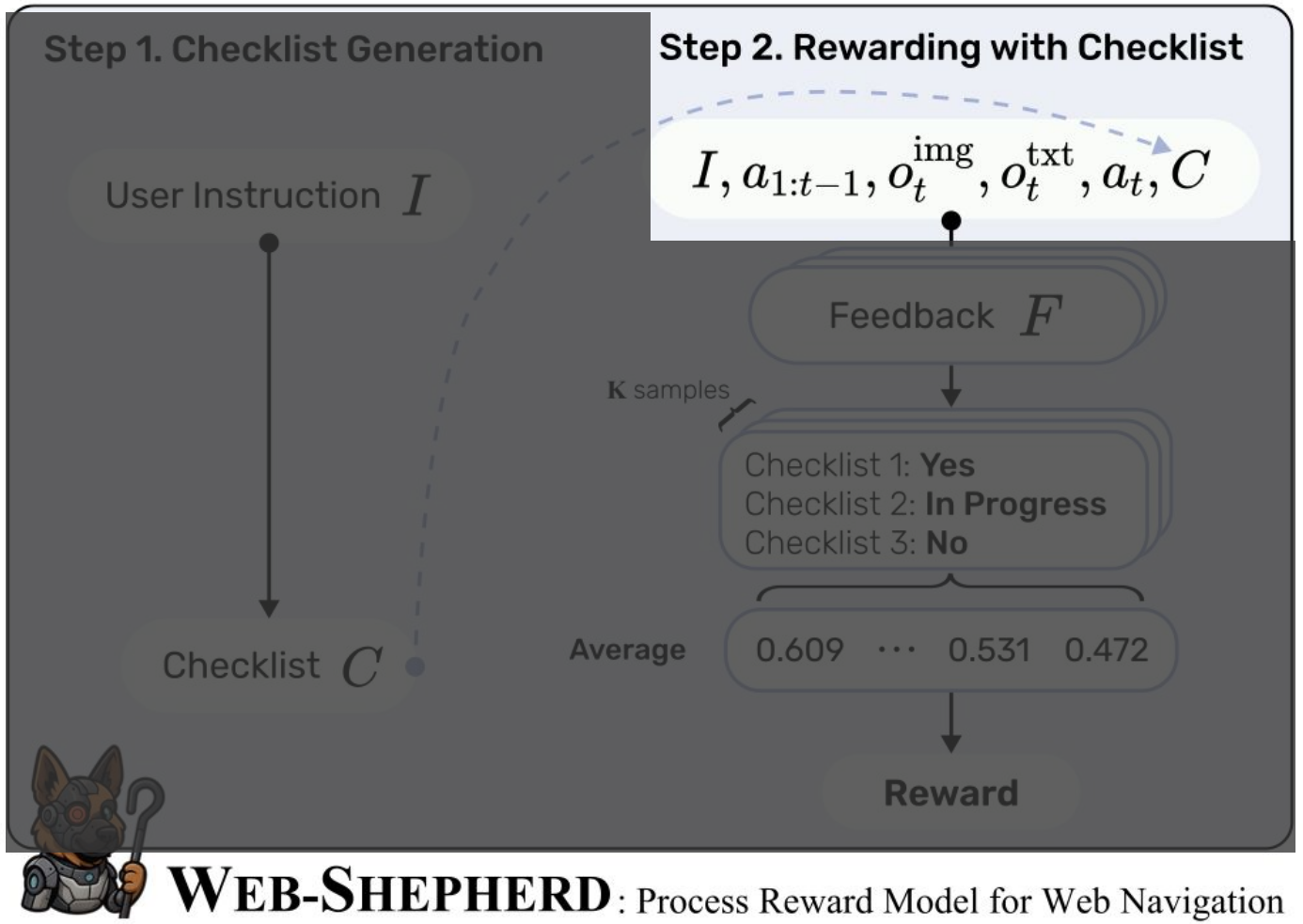
Web-Shepherd

C = (g1,g2,⋯,gn)



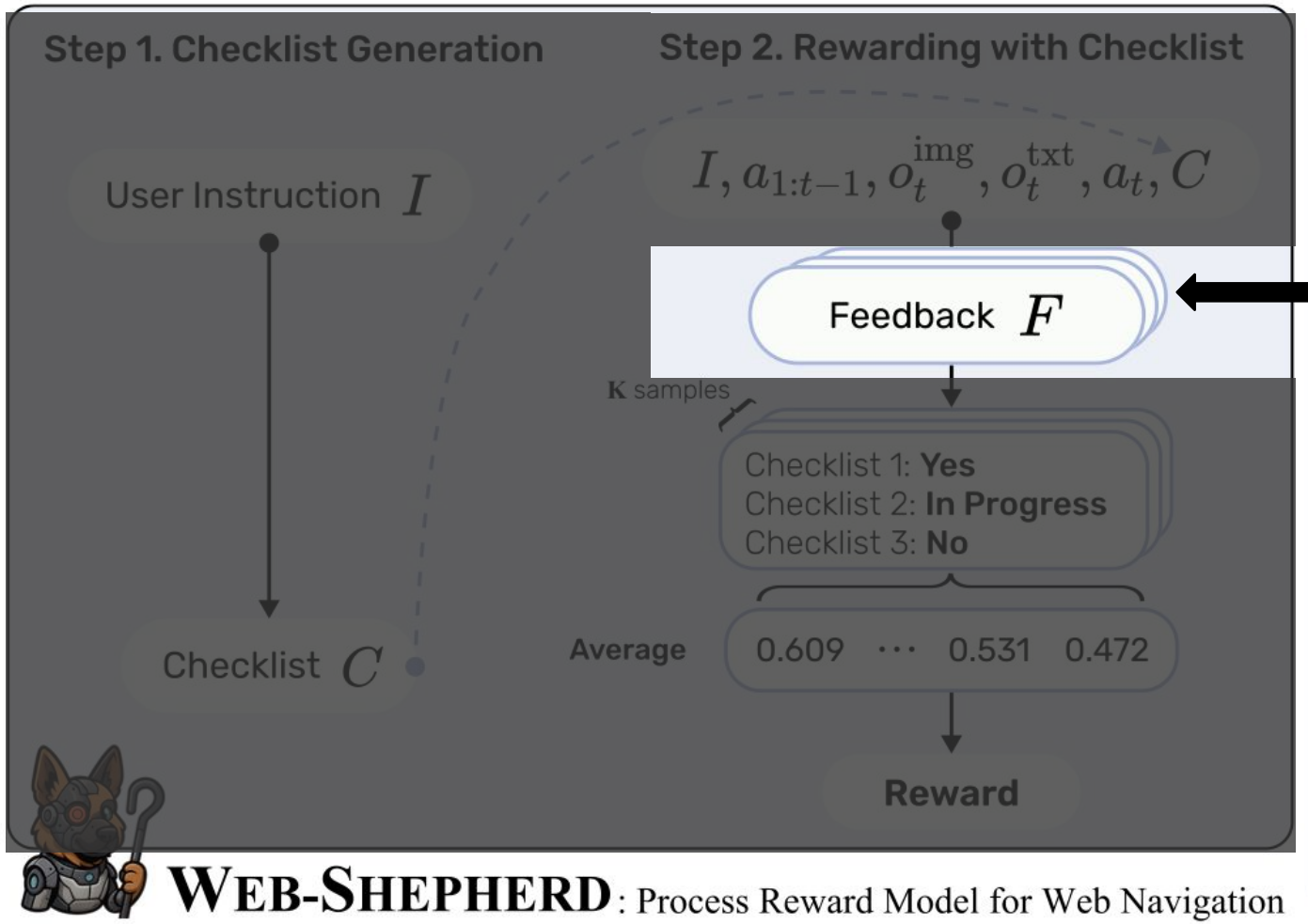
* The model is trained to generate the corresponding feedback and judgement in auto-regressive manner.

Web-Shepherd



* The model is trained to generate the corresponding feedback and judgement in auto-regressive manner.

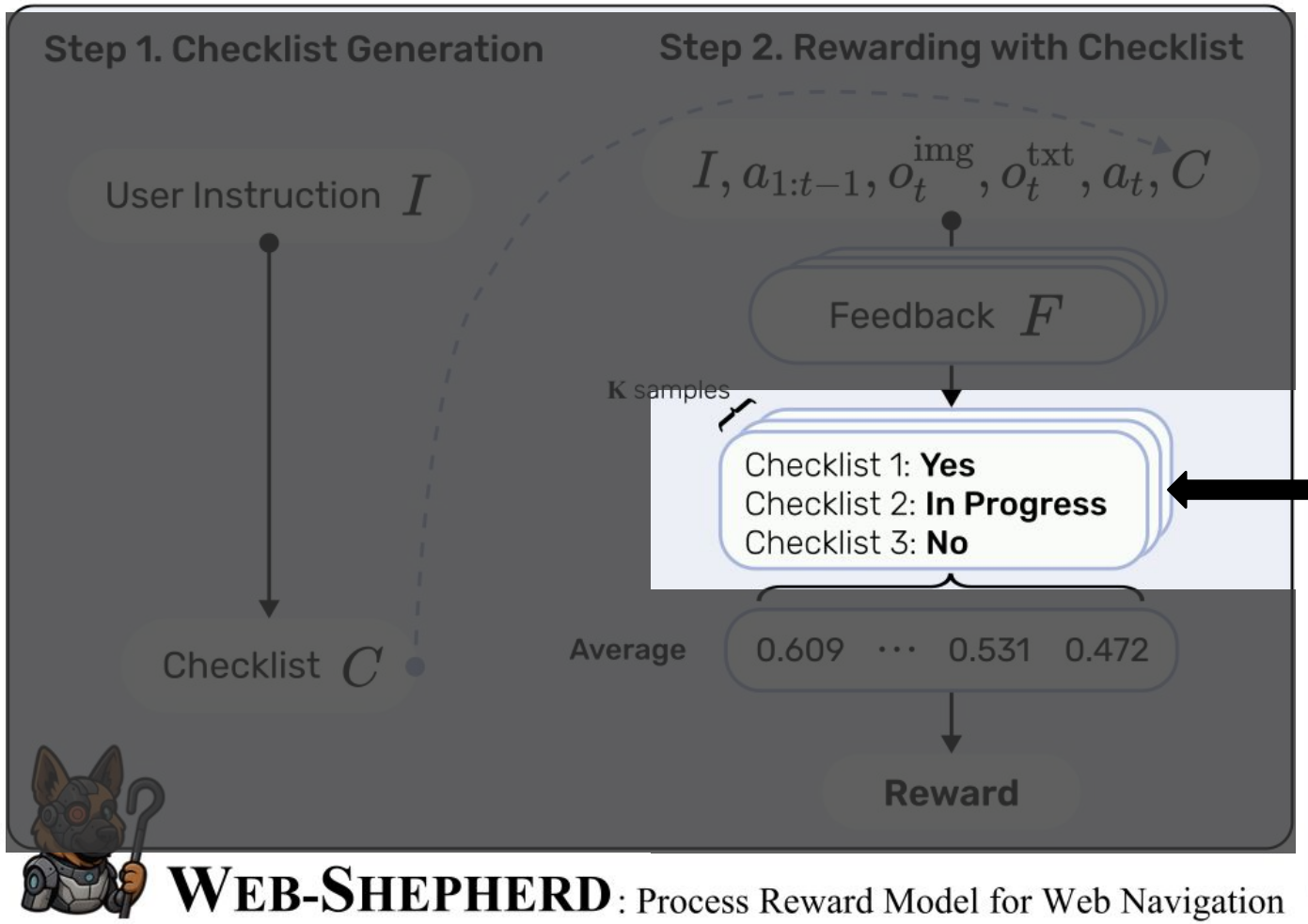
Web-Shepherd



Sample multiple output to compute the average reward

* The model is trained to generate the corresponding feedback and judgement in auto-regressive manner.

Web-Shepherd

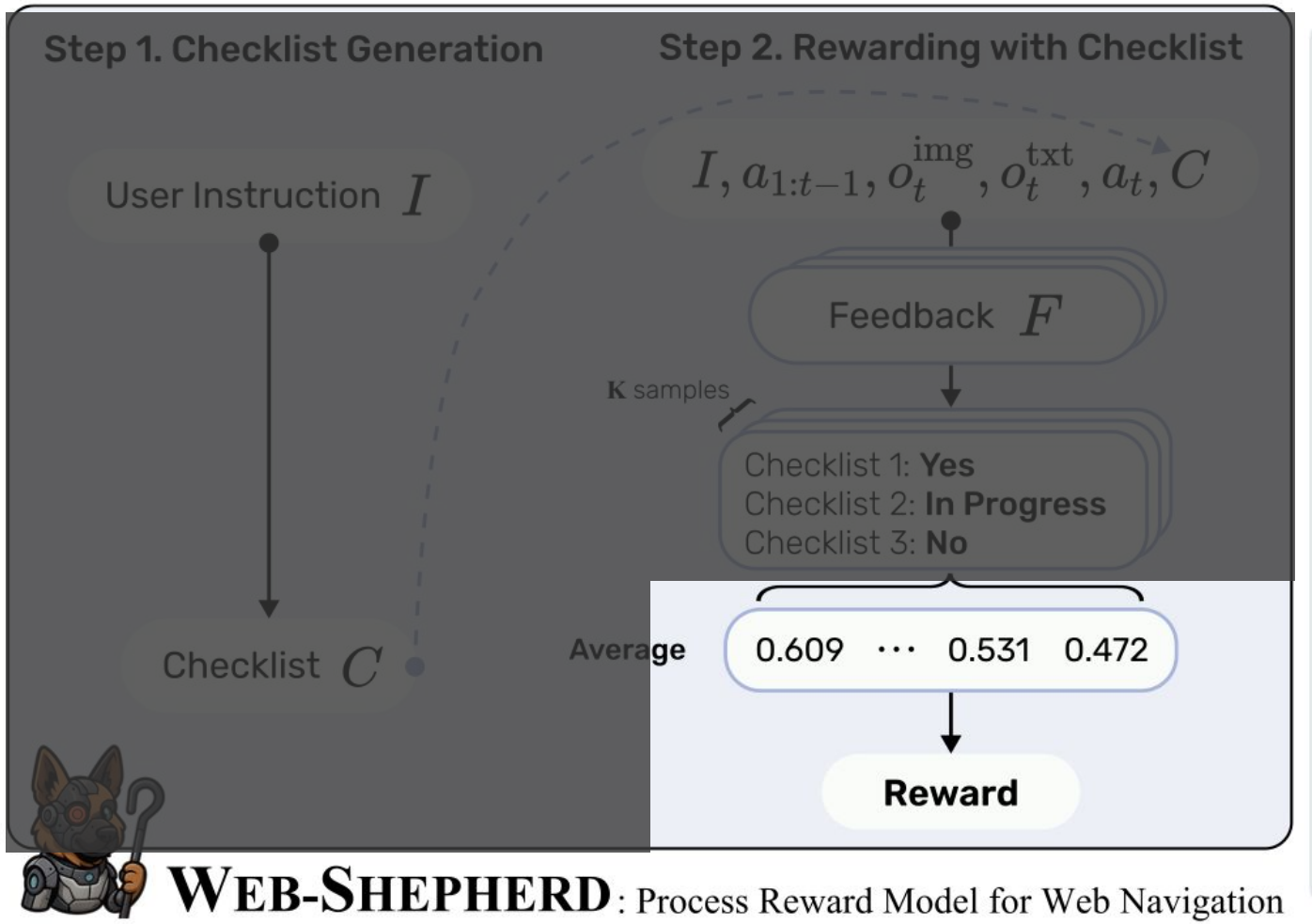


Extracts the probabilities of specific text tokens from the LM head logits to convert them into scores. (verbalizer)

The reward for sample K is computed by averaging the combined probabilities of 'Yes' and 'In Progress' across all checklist items.

* The model is trained to generate the corresponding feedback and judgement in auto-regressive manner.

Web-Shepherd



* The model is trained to generate the corresponding feedback and judgement in auto-regressive manner.

WebRewardBench

5 Answer Candidates : 1 correct answer and 4 false answers

Model	Inputs	Checklist	Mind2Web						WebArena	
			Cross-Task		*Cross-Website*		*Cross-Domain*		*Test*	
			MRR	Acc. (traj)	MRR	Acc. (traj)	MRR	Acc. (traj)	MRR	Acc. (traj)
GPT-4o-mini	T	✗	47.5	0.0	47.6	13.5	45.4	0.8	34.4	5.0
	T	✓	63.9	5.0	66.1	12.8	63.3	12.4	60.0	15.0
	T + I	✗	48.2	0.0	49.3	0.0	49.5	0.8	38.7	0.0
	T + I	✓	58.8	2.5	64.4	5.1	63.0	4.1	53.8	5.0
GPT-4o	T	✗	56.9	5.0	55.8	2.6	59.8	3.3	59.2	15.0
	T	✓	67.4	7.5	70.3	5.1	70.2	11.6	69.7	15.0
	T + I	✗	52.5	5.0	52.2	0.0	52.8	1.7	49.7	5.0
	T + I	✓	62.4	5.0	68.1	15.4	65.1	6.6	59.7	10.0
Qwen-2.5-VL-72B	T	✗	55.7	5.0	51.8	0.0	54.2	1.7	54.6	5.0
	T	✓	59.4	0.0	62.4	0.0	57.9	1.7	52.3	5.0
	T + I	✗	50.1	2.5	47.6	0.0	49.8	0.8	43.1	0.0
	T + I	✓	52.9	2.5	53.5	2.6	52.0	2.5	47.3	0.0
WEB-SHEPHERD (3B)	T	✓	87.6	55.0	**88.0**	43.6	87.2	47.1	91.1	60.0
	T+I	✓	85.0	42.5	87.3	41.0	84.4	37.2	92.5	65.0
WEB-SHEPHERD (8B)	T	✓	**88.3**	**57.5**	87.9	**51.3**	**91.3**	**61.2**	**97.8**	**85.0**

MRR: The average of the reciprocal ranks of the correct answers.
Acc. (traj): The percentage of entire task trajectories where the model correctly evaluates every single step without a mistake.

Reward-Guided Trajectory Search

Trajectory Search:

5 action candidates sampled from the policy, the action that is assigned the highest reward is executed.

Policy	PRM	Checklist	Shopping	CMS	Reddit	GitLab	Map	Total	Δ
GPT-4o-mini	w/o Trajectory Search	N/A	21.74	22.86	19.05	34.38	19.35	23.64	–
	GPT-4o-mini	✗	13.04	14.29	9.52	25.00	16.13	15.75	−7.89
	GPT-4o-mini	✓	21.74	31.43	14.29	34.38	16.13	24.24	+0.60
	WEB-SHEPHERD (3B)	✓	**32.61**	37.14	19.05	34.38	32.26	32.12	+8.48
	WEB-SHEPHERD (8B)	✓	26.09	**45.71**	**23.81**	**40.62**	**35.48**	**34.55**	**+10.90**
GPT-4o	w/o Trajectory Search	N/A	23.91	31.43	28.57	**56.25**	19.35	31.52	–
	GPT-4o-mini	✓	21.74	31.43	28.57	40.62	12.90	26.67	−4.85
	WEB-SHEPHERD (3B)	✓	28.26	37.14	**47.62**	53.12	25.81	36.97	+5.45
	WEB-SHEPHERD (8B)	✓	**30.43**	**42.86**	**47.62**	46.88	**35.48**	**39.39**	**+7.87**

Reward-Guided Trajectory Search

Trajectory Search:

5 action candidates sampled from the policy, the action that is assigned the highest reward is executed.

Policy	PRM	Checklist	Shopping	CMS	Reddit	GitLab	Map	Total	Δ
GPT-4o-mini	w/o Trajectory Search	N/A	21.74	22.86	19.05	34.38	19.35	23.64	–
	GPT-4o-mini	✗	13.04	14.29	9.52	25.00	16.13	15.75	−7.89
	GPT-4o-mini	✓	21.74	31.43	14.29	34.38	16.13	24.24	+0.60
	WEB-SHEPHERD (3B)	✓	**32.61**	37.14	19.05	34.38	32.26	32.12	+8.48
	WEB-SHEPHERD (8B)	✓	26.09	**45.71**	**23.81**	**40.62**	**35.48**	**34.55**	**+10.90**
GPT-4o	w/o Trajectory Search	N/A	23.91	31.43	28.57	**56.25**	19.35	31.52	–
	GPT-4o-mini	✓	21.74	31.43	28.57	40.62	12.90	26.67	−4.85
	WEB-SHEPHERD (3B)	✓	28.26	37.14	**47.62**	53.12	25.81	36.97	+5.45
	WEB-SHEPHERD (8B)	✓	**30.43**	**42.86**	**47.62**	46.88	**35.48**	**39.39**	**+7.87**

Reward-Guided Trajectory Search

Trajectory Search:

5 action candidates sampled from the policy, the action that is assigned the highest reward is executed.

Policy	PRM	Checklist	Shopping	CMS	Reddit	GitLab	Map	Total	Δ
GPT-4o-mini	w/o Trajectory Search	N/A	21.74	22.86	19.05	34.38	19.35	23.64	–
	GPT-4o-mini	✗	13.04	14.29	9.52	25.00	16.13	15.75	−7.89
	GPT-4o-mini	✓	21.74	31.43	14.29	34.38	16.13	24.24	+0.60
	WEB-SHEPHERD (3B)	✓	**32.61**	37.14	19.05	34.38	32.26	32.12	+8.48
	WEB-SHEPHERD (8B)	✓	26.09	**45.71**	**23.81**	**40.62**	**35.48**	**34.55**	**+10.90**
GPT-4o	w/o Trajectory Search	N/A	23.91	31.43	28.57	**56.25**	19.35	31.52	–
	GPT-4o-mini	✓	21.74	31.43	28.57	40.62	12.90	26.67	−4.85
	WEB-SHEPHERD (3B)	✓	28.26	37.14	**47.62**	53.12	25.81	36.97	+5.45
	WEB-SHEPHERD (8B)	✓	**30.43**	**42.86**	**47.62**	46.88	**35.48**	**39.39**	**+7.87**

Reward-Guided Trajectory Search

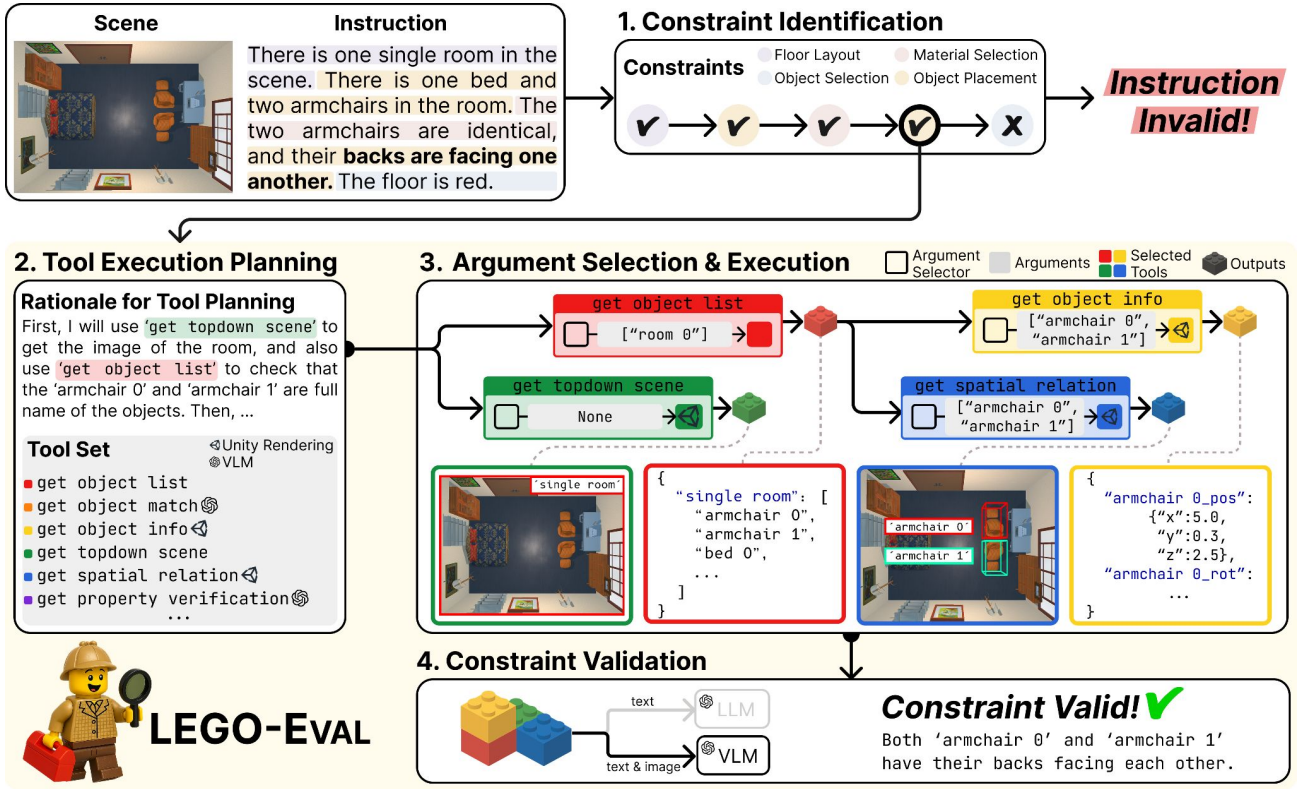
Trajectory Search:

5 action candidates sampled from the policy, the action that is assigned the highest reward is executed.

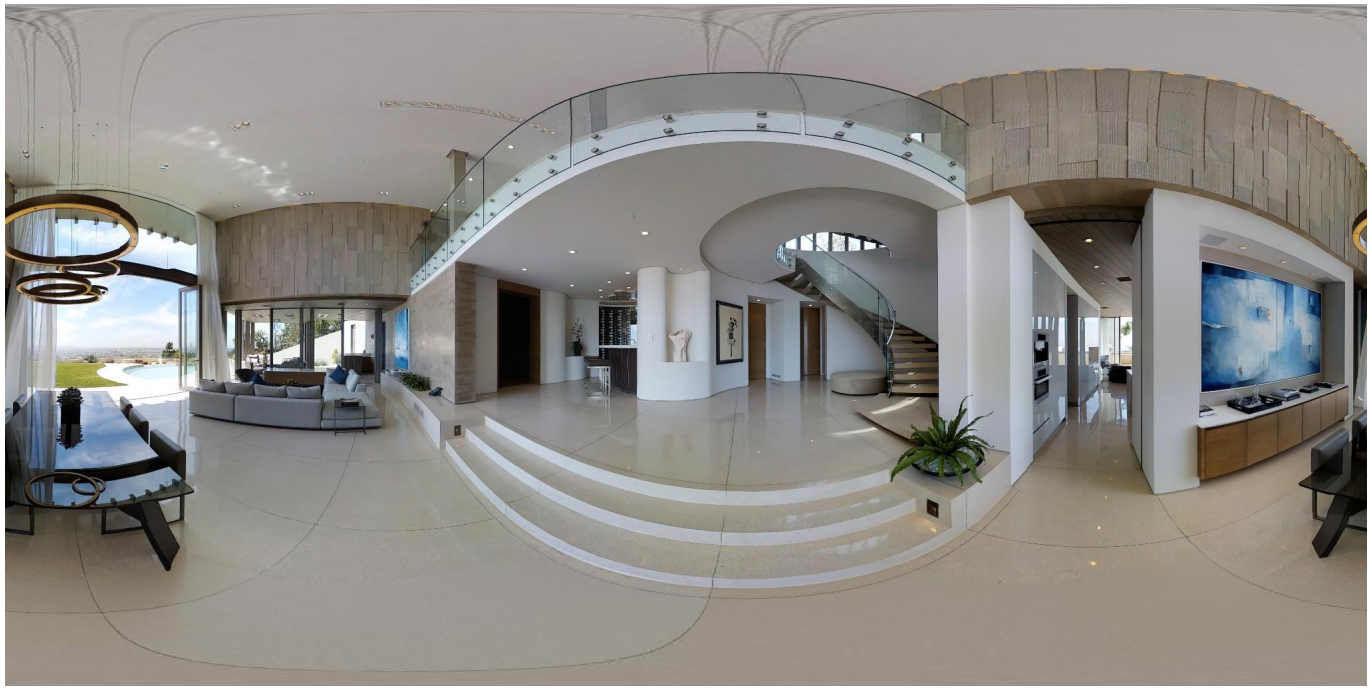
PRM	Checklist	Dashboard	Form	Knowledge	List-filter	List-sort	Menu	Service Catalog	Total	Δ
w/o Trajectory Search	N/A	50.00	0.00	10.00	0.00	5.00	25.00	2.22	9.39	–
GPT-4o-mini	✓	55.00	10.00	10.00	0.00	6.67	20.00	5.56	12.42	+3.03
Web-Shepherd (3B)	✓	57.50	**14.00**	10.00	0.00	**10.00**	10.00	**11.11**	14.85	+5.46
Web-Shepherd (8B)	✓	**65.00**	**14.00**	**20.00**	**0.00**	**10.00**	**20.00**	7.78	**15.76**	**+6.37**

LEGO-Eval: Towards Fine-Grained Evaluation on Synthesizing 3D Embodied Environments with Tool Augmentation

Submitted to EMNLP 2026



3D Environment for Embodied Agents



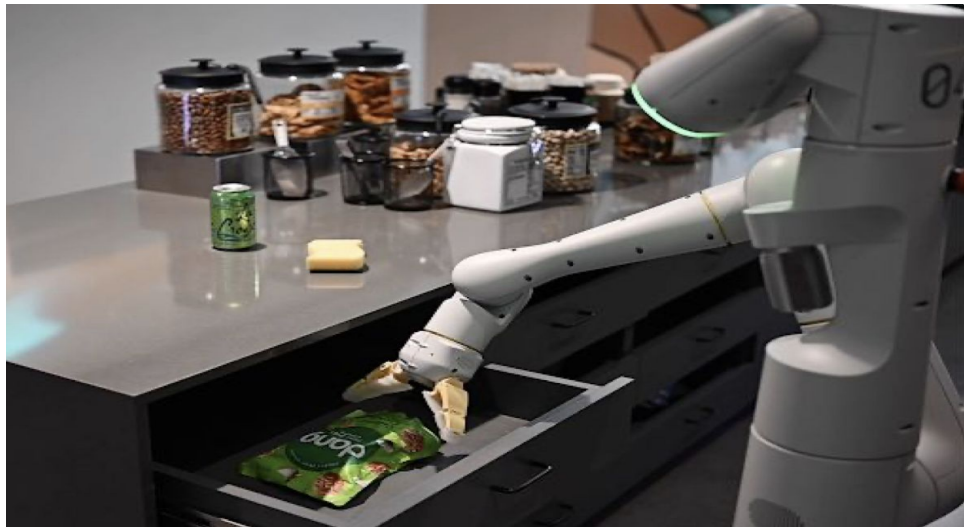
Matterport3D



AI2Thor

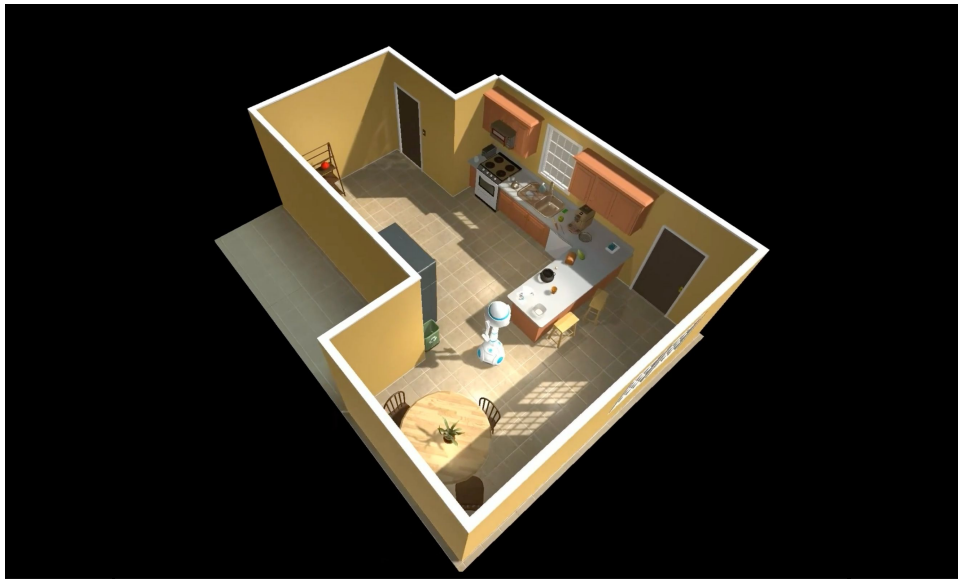
Simulators are Necessary for Embodied Training

Training in Real World



- High cost
- Time consuming
- Hard to reproduce same environment

Training in Simulator



- Relatively low cost
- Less time consuming
- Easy to reproduce same environment

Limitations of Existing Simulators



Limitation: Scale of provided scenes

- Generalizability of agents

- Limited tasks that can be trained

Limitations of Existing Simulators



Limitation: Scale of provided scenes

- Generalizability of agents

- Limited tasks that can be trained

→ **Too costly to create human-authored scenes!**

3D Environment Synthesis for Embodied Agents

Limitation: Scale of provided scenes

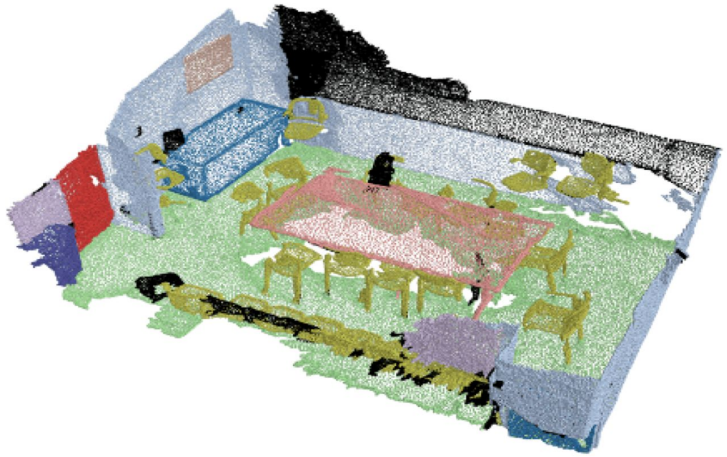
⇩

Text-guided automatic scene synthesis

Mesh



PCD



Scene Graph


```
"objects": [
  {
    "assetId": "08566f5b460242859256de7a04f33c87",
    "id": "sofa-0 (living room)",
    "kinematic": true,
    "position": {
      "x": 1.5,
      "y": 0.40328151618058866,
      "z": 2.302312041320852
    },
    "rotation": {
      "x": 0,
      "y": 180,
      "z": 0
    },
    "material": null,
    "roomId": "living room",
```

3D Environment Synthesis for Embodied Agents

Limitation: Scale of provided scenes

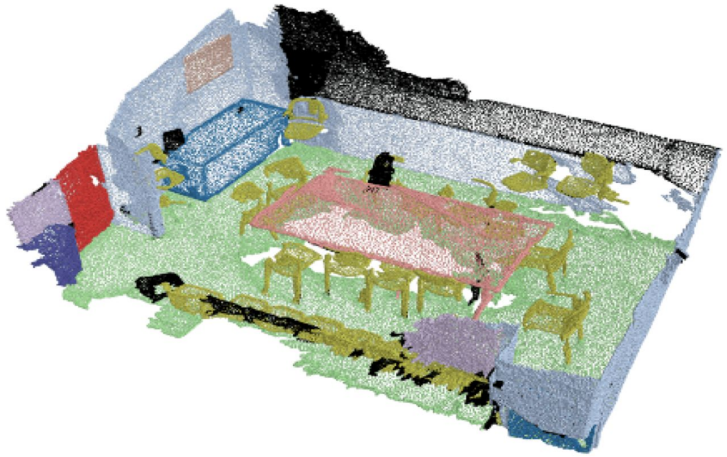
⇩

Text-guided automatic scene synthesis

Mesh



PCD



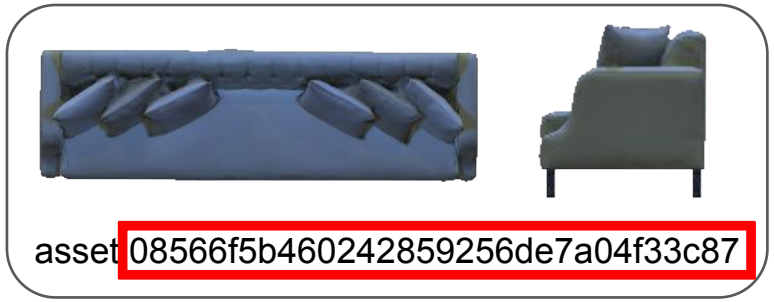
Scene Graph


```
"objects": [
  {
    "assetId": "08566f5b460242859256de7a04f33c87",
    "id": "sofa-0 (living room)",
    "kinematic": true,
    "position": {
      "x": 1.5,
      "y": 0.40328151618058866,
      "z": 2.302312041320852
    },
    "rotation": {
      "x": 0,
      "y": 180,
      "z": 0
    },
    "material": null,
    "roomId": "living room",
```

Scene Graphs for Training Embodied Agents

Scene saved as Json → **Json to Scene** → **Scene Running on Simulator**


```
"objects": [
  {
    "assetId": "08566f5b460242859256de7a04f33c87",
    "id": "sofa-0 (living room)",
    "kinematic": true,
    "position": {
      "x": 1.5,
      "y": 0.40328151618058866,
      "z": 2.302312041320852
    },
    "rotation": {
      "x": 0,
      "y": 180,
      "z": 0
    },
    "material": null,
    "roomId": "living room",
```



Importance of Text-guided 3D Scene Synthesis

Why do we need text-guided?

- To generate scenes appropriate for training certain tasks
- To generate scenes that mirror the real-world scenes [1]

1) We believe that providing highly detailed descriptions of real-world scenes is one effective way to generate scenes that closely mirror real-world environments, alongside other approaches such as using images of real-world environments.

Importance of Text-guided 3D Scene Synthesis

Why do we need text-guided?

- To generate scenes appropriate for training certain tasks
- To generate scenes that mirror the real-world scenes

→ **Scene - description alignment is very important!**

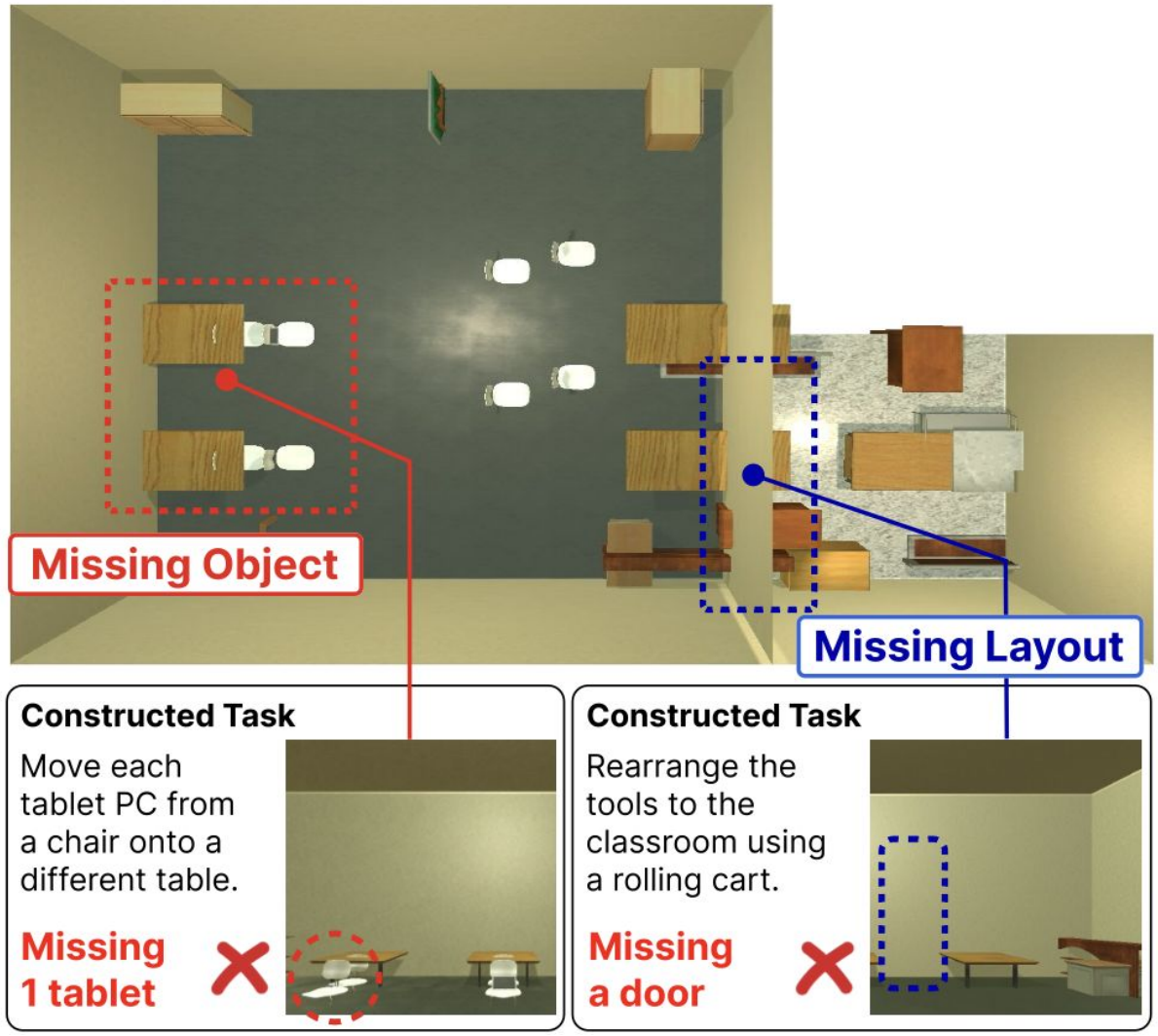
Importance of Text-guided 3D Scene Synthesis

LEGO-Eval

User Expected Scene Description

There is one science classroom and one preparation room in the school. ... **Two tablet PCs are placed on chairs in the classroom.** There are two tables attached to the wall. **The preparation room is directly connected to the classroom.** ... The rolling cart in the preparation room is placed next to the wall.

AI Generated Scene



Importance of Reliable Evaluation between Scene-Description

LEGO-Eval

Verification of scene-description alignment is important.



Importance of Reliable Evaluation between Scene-Description

LEGO-Eval

Verification of scene-description alignment is important.

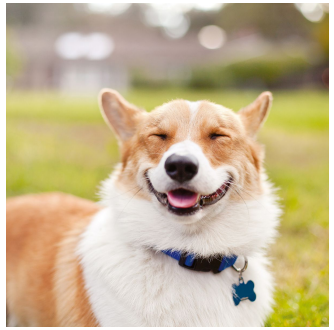
Human? ❌

Automatic? ✅



Existing Evaluation Methods

CLIP-Score



A smiling dog with gold and white furs…

Image Encoder

Text Encoder

I_1 I_2 I_3 … I_n

T_1 T_2 T_3 … T_n

Cosine Similarity

VLM-as-a-Judge



You must evaluate if the given images align with the instruction.

Instruction: It is a museum with multiple pictures…





LEGO & LEGO-Eval

LEGO
(Benchmark Dataset)

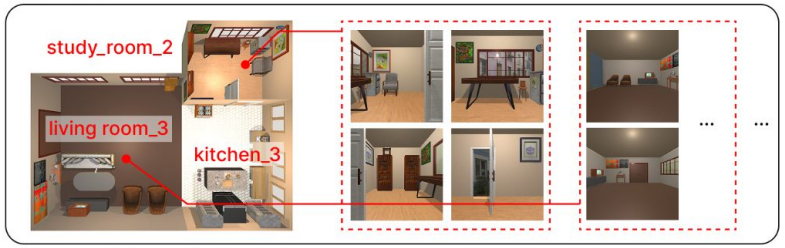
Constraint Types

Object Selection | Object Placement | Floor Layout | Material Selection

Annotated Description

There is a house with one study room, one kitchen, and one living room. The floor in the living room is light wooden. In the study room, there is a table, a clock and a bookshelf. The two bookshelves are identical. The clock in the study...

Scene



Scene Metadata

Presence

```
{
  "study room": [
    "desk-0",
    "bookshelf-0",
    "book-1|side-
     table-0",
    ...
  ]
}
```

Position/Rotation

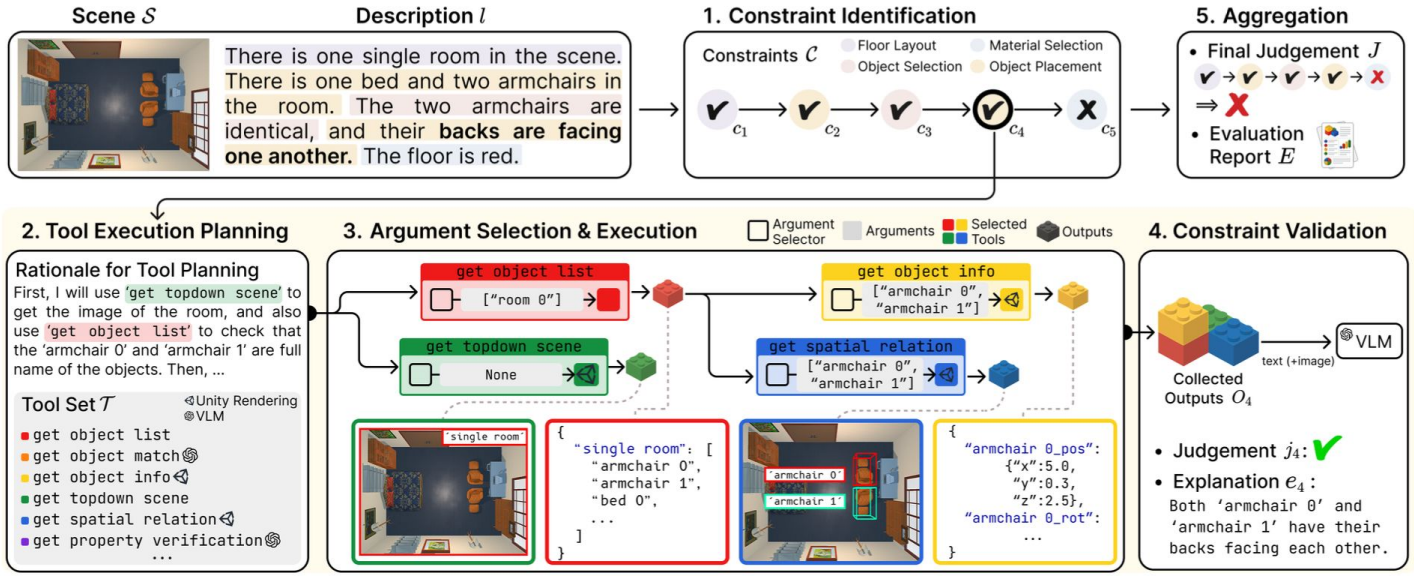
```
{
  "desk-0_pos": {
     "x":3.00,
     "y":0.35,
     "z":0.35},
  "desk-0_rot": {
     "x":0,
     "y":180,...
}
```

Relation

```
{
  "object_to_recep":
    {"obj":"book-2",
     "recep":"table-0"}
  "room_to_room":
    {"room0":"kitchen",
     "room1":"study room"}
      ...
}
```

...

LEGO-Eval
(Evaluation Method)



LEGO Dataset



Instruction

There is a cafe that has a seating area and kitchen. The seating area has light beige walls and a dark wooden floor. In the seating area, there are two tables and four chairs. The tables in the seating area are round and made of brown oak. The tables in the seating area are each surrounded by two chairs. The chairs in the seating area are white.

Constraints

1. There is a cafe that has a seating area and kitchen.
2. The seating area has light beige walls and a dark wooden floor.
3. In the seating area, there are two tables and four chairs.
4. The tables in the seating area are round and made of brown oak.
5. The tables in the seating area are each surrounded by two chairs.
6. The chairs in the seating area are white.

LEGO-Dataset

Constraint Types

Object Selection | Object Placement | Floor Layout | Material Selection

Annotated Description

There is a house with one study room, one kitchen, and one living room. The floor in the living room is light wooden. In the study room, there is a table, a clock and a bookshelf. The two bookshelves are identical. The clock in the study...

Scene



Scene Metadata

Presence

```
{
  "study room": [
    "desk-0",
    "bookshelf-0",
    "book-1|side-
     table-0",
    ...
  ]
}
```

Position/Rotation

```
{
  "desk-0_pos": {
     "x":3.00,
     "y":0.35,
     "z":0.35},
  "desk-0_rot": {
     "x":0,
     "y":180,...
}
```

Relation

```
{
  "object_to_recep":
    {"obj":"book-2",
     "recep":"table-0"}
  "room_to_room":
    {"room0":"kitchen",
     "room1":"study room"}
       ...
}
```

...

LEGO-Eval



Committed Crime?



Article 1.

Article 2.

Article 3.









Article 1. ✓

Article 2. ✓

Article 3. ✓



Article 1. ✓

Article 2. ✗

Article 3. ✓



LEGO-Eval: Tools

Tool Set (21 tools)

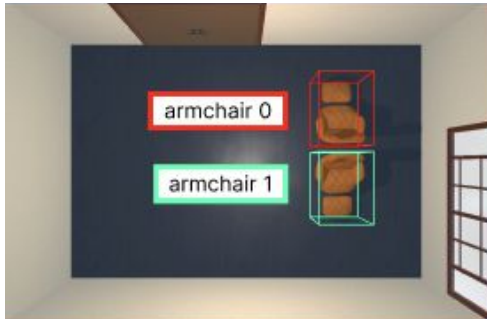
get_topdown_scene

arguments: None



get_spatial_relation

arguments:
['armchair 0', 'armchair 1']



get_object_list

arguments: ['room 0']

```
{
    "room 0":
        [ "armchair 0",
          "armchair 1",
          "bed 0",
          "green painting",
          ... ]
}
```

get_object_info

arguments:
['armchair 0', 'armchair 1']

```
{
    "armchair 0_position":
        {"x": 5.0, "y": 0.3, "z": 2.5},
    "armchair 0_coordinates":
        [[4.6, 2.0], [4.6, 3.0],
         [5.4, 3.0], [5.4, 2.0]],
    "armchair 1_position": ...,
                           ...
}
```

LEGO-Eval: Tools

Tool Set (21 tools)

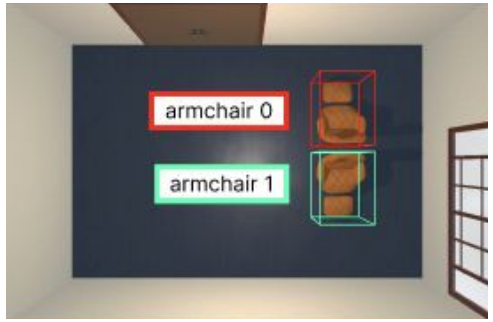
get_topdown_scene

arguments: None



get_spatial_relation

arguments:
['armchair 0', 'armchair 1']



get_object_list

arguments: ['room 0']

```
{
    "room 0":
        [ "armchair 0",
          "armchair 1",
          "bed 0",
          "green painting",
          ... ]
}
```

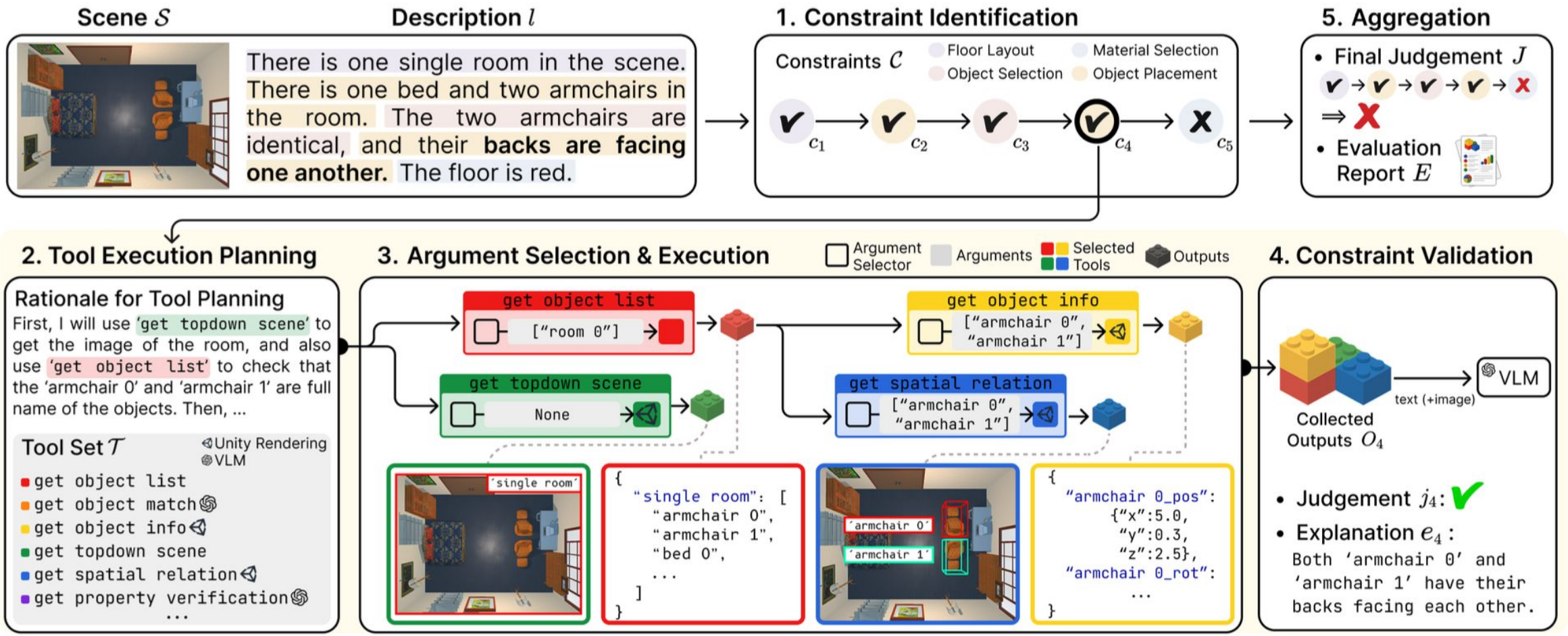
get_object_info

arguments:
['armchair 0', 'armchair 1']

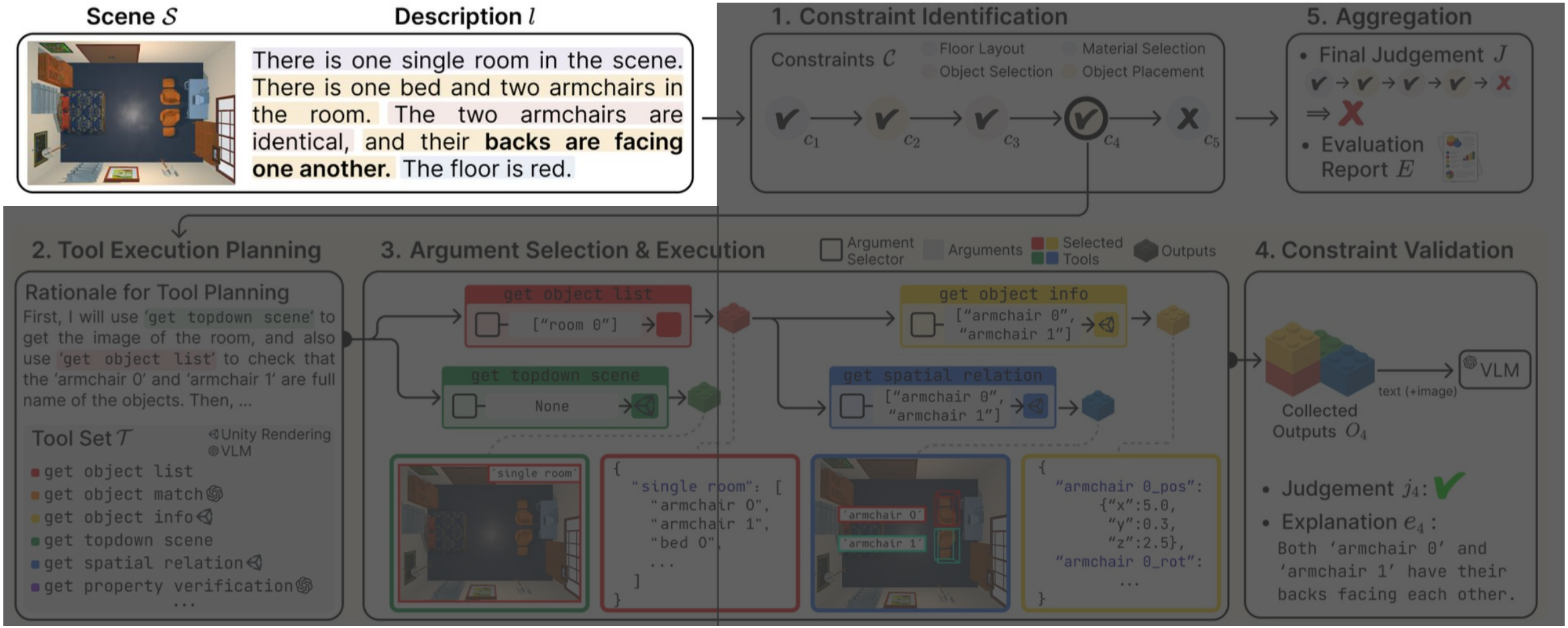
```
{
    "armchair 0_position":
        {"x": 5.0, "y": 0.3, "z": 2.5},
    "armchair 0_coordinates":
        [[4.6, 2.0], [4.6, 3.0],
         [5.4, 3.0], [5.4, 2.0]],
    "armchair 1_position": ...,
                           ...
}
```

Allows to obtain and use only related information for evaluation

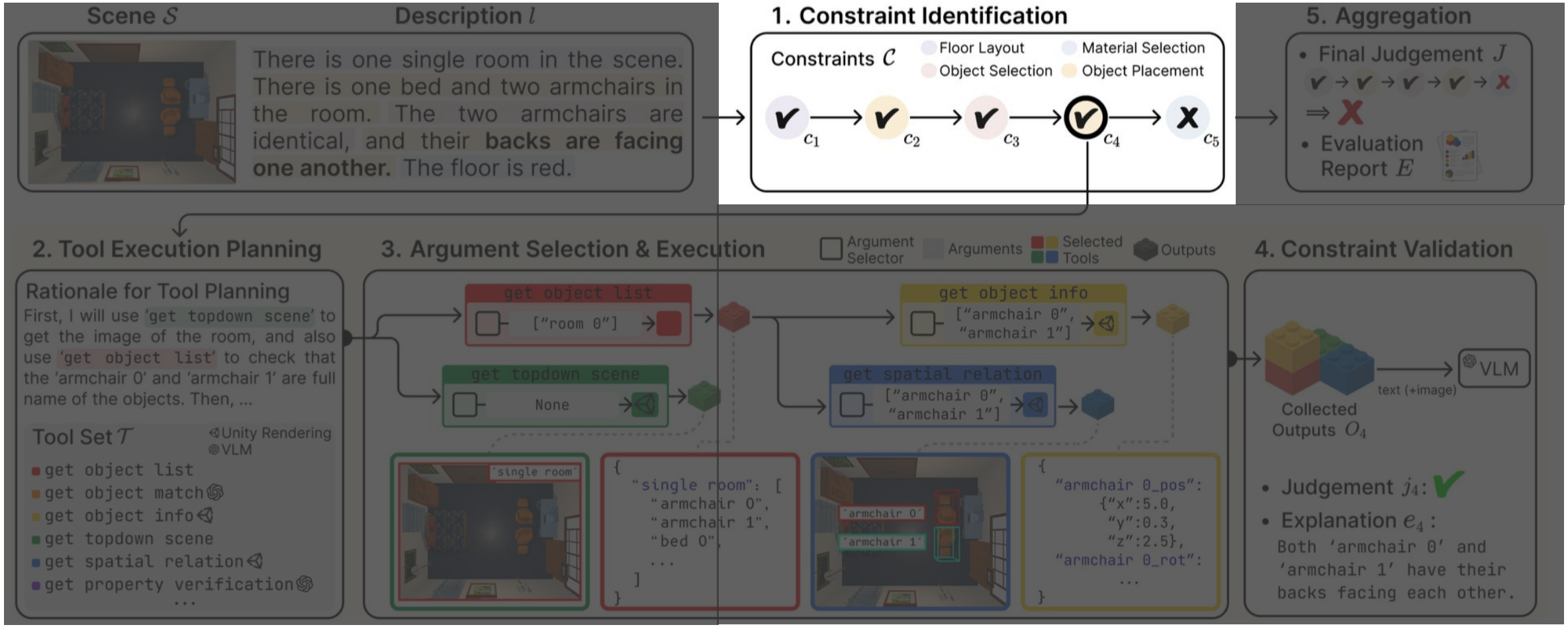
LEGO-Eval: Method Overview



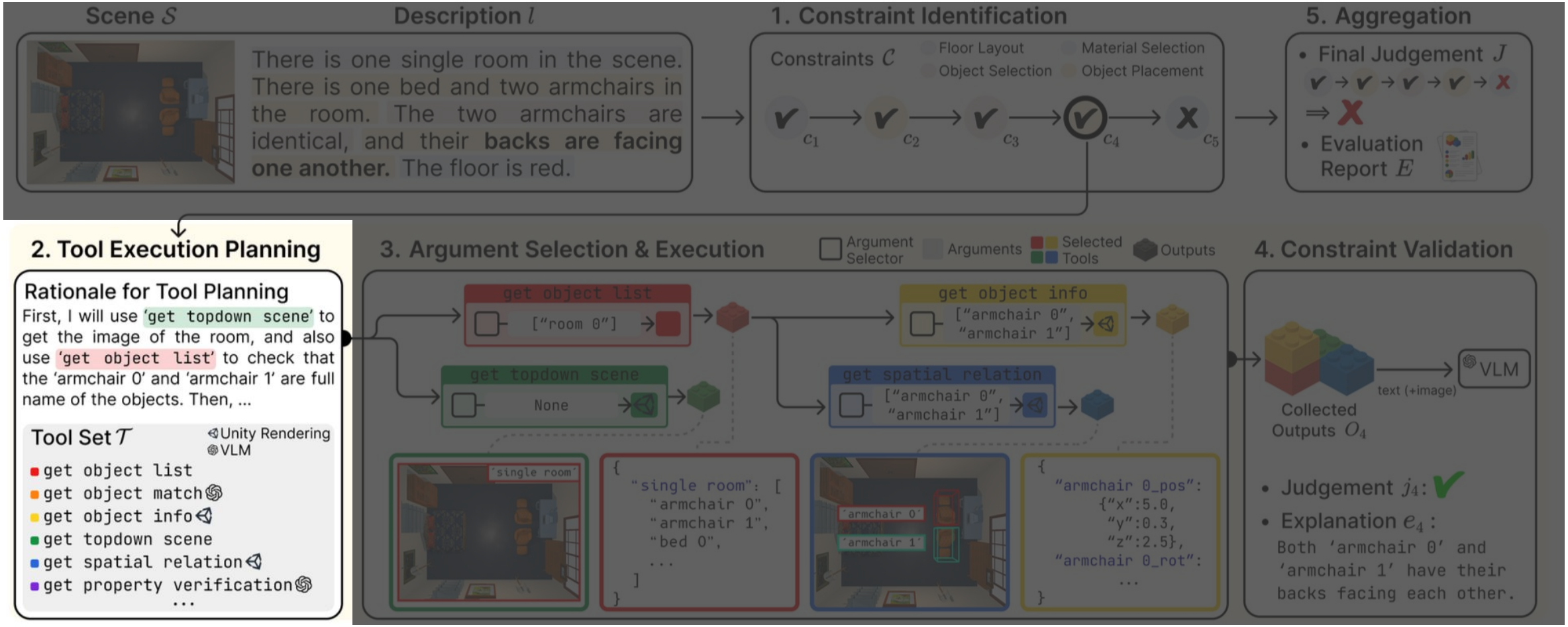
LEGO-Eval: Method



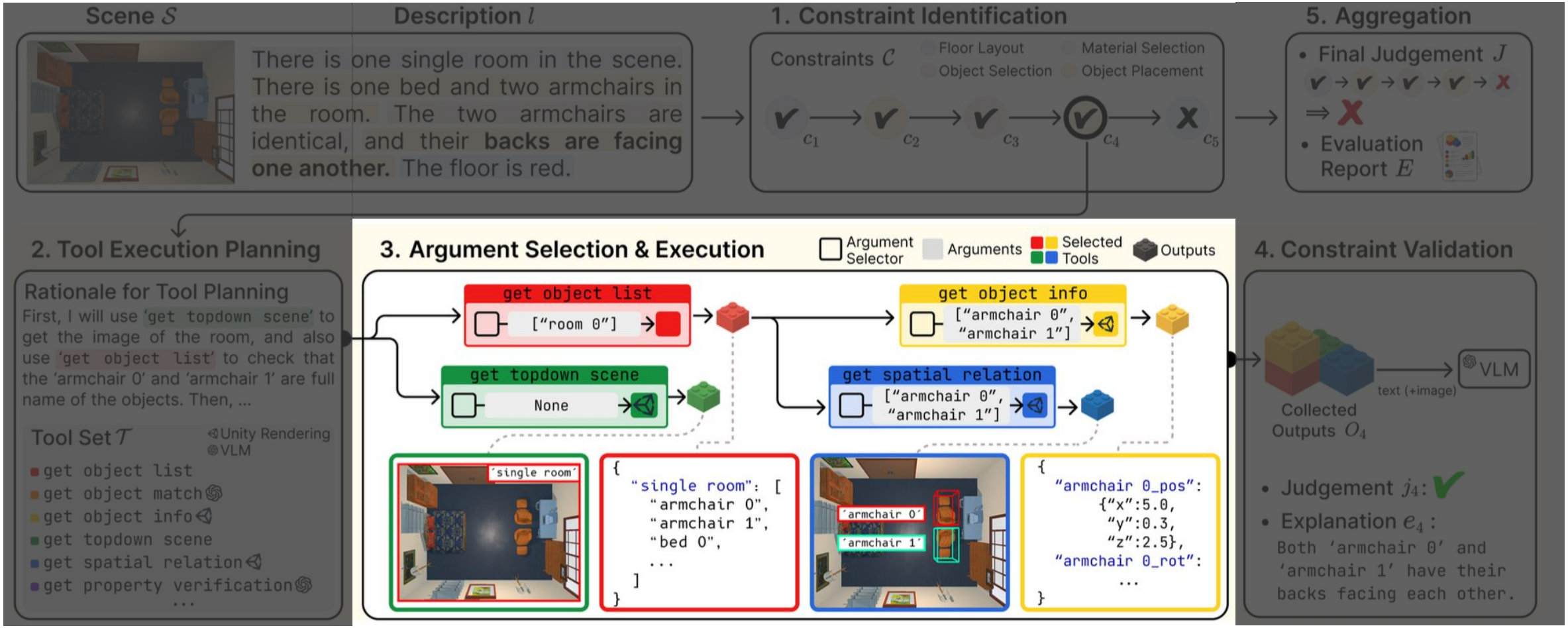
LEGO-Eval: Method



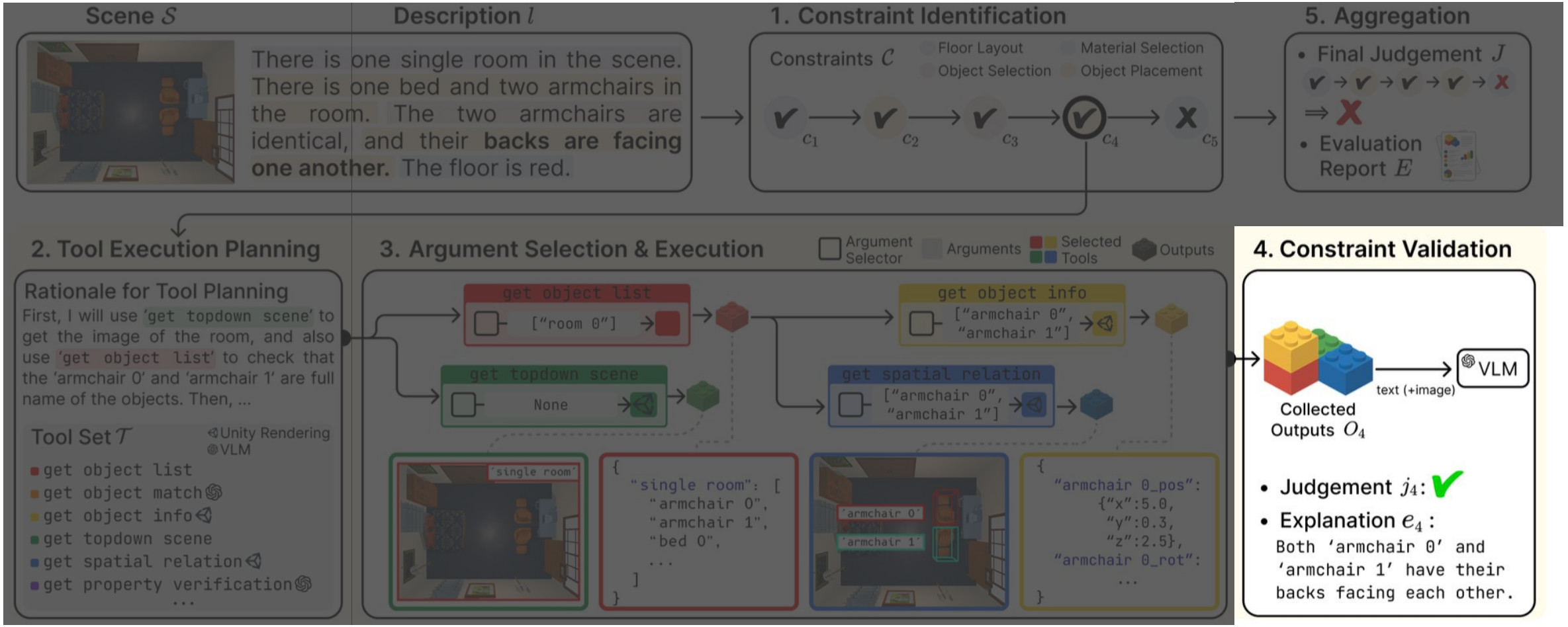
LEGO-Eval: Method



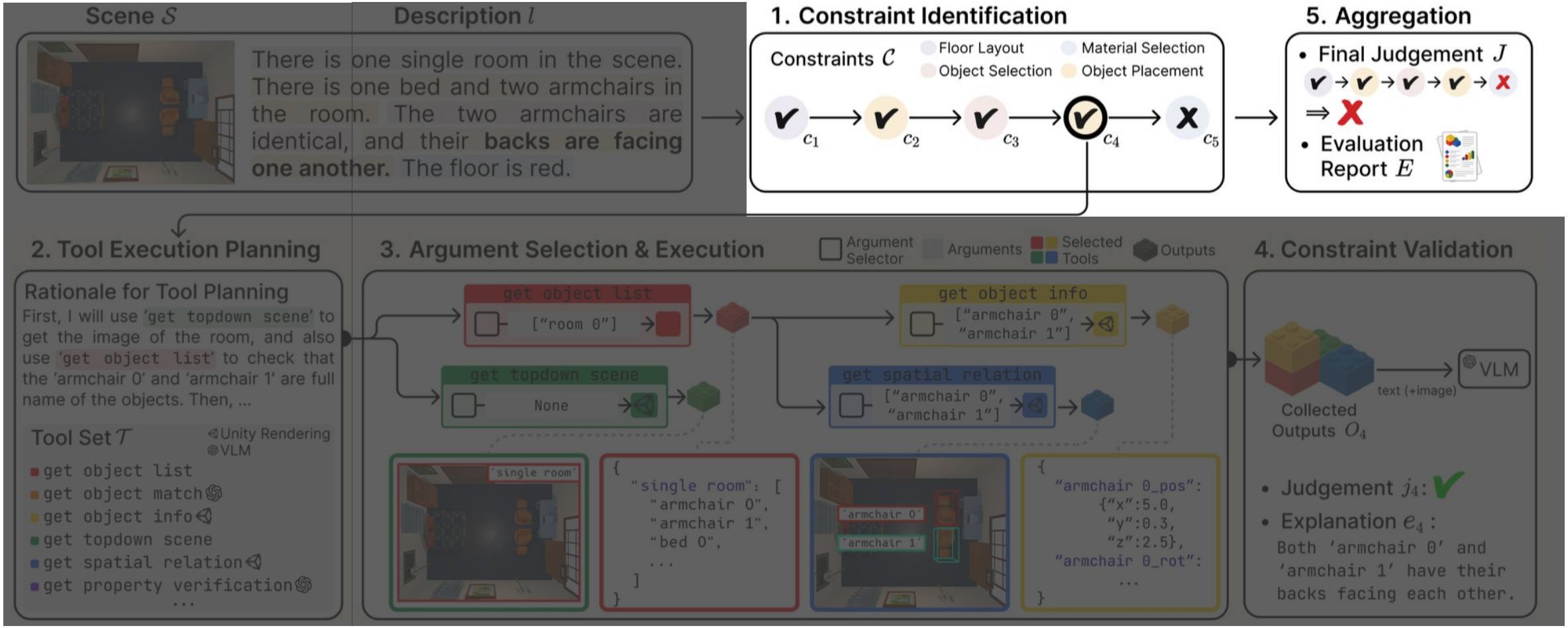
LEGO-Eval: Method



LEGO-Eval: Method



LEGO-Eval: Method



Comparisons Between Evaluation Methods

Dataset for Evaluation: Instructions from LEGO with aligned scenes and misaligned scenes

Methods	Holistic ↑				Partial ↑			
	F1	Precision	Recall	Cohen's κ	F1	Precision	Recall	Cohen's κ
CLIPScore	0.49	0.51	0.51	0.02	0.46	0.50	0.50	0.00
VLM-as-a-judge								
Gemini 2.5 Pro	0.38	0.76	0.52	0.05	0.60	0.66	0.70	0.28
GPT-4.1	0.40	0.67	0.53	0.05	0.68	0.65	0.73	0.35
GPT-5.1	0.38	0.68	0.52	0.04	0.60	0.71	0.71	0.28
SceneEval*								
GPT-4o	0.47	0.74	0.58	0.15	0.45	0.64	0.58	0.12
LEGO-EVAL								
Qwen3-VL-8B	0.67	0.74	0.69	0.38	0.71	0.74	0.69	0.42
Qwen3-VL-32B	0.67	0.79	0.70	0.39	0.80	0.80	**0.81**	0.61
GPT-4.1-mini	0.70	0.78	0.72	0.43	0.78	0.80	0.76	0.56
GPT-4.1	**0.81**	**0.84**	**0.82**	**0.63**	**0.83**	**0.86**	**0.81**	**0.66**

Holistic: Considered correct if all constraints in instruction are evaluated correctly
Partial: Considered correct if the constraint is evaluated correctly

Comparison Between Evaluation Methods

Methods	Holistic ↑				Partial ↑			
	F1	Precision	Recall	Cohen's κ	F1	Precision	Recall	Cohen's κ
CLIPScore	0.49	0.51	0.51	0.02	0.46	0.50	0.50	0.00
VLM-as-a-judge								
Gemini 2.5 Pro	0.38	0.76	0.52	0.05	0.60	0.66	0.70	0.28
GPT-4.1	0.40	0.67	0.53	0.05	0.68	0.65	0.73	0.35
GPT-5.1	0.38	0.68	0.52	0.04	0.60	0.71	0.71	0.28
SceneEval*								
GPT-4o	0.47	0.74	0.58	0.15	0.45	0.64	0.58	0.12
LEGO-EVAL								
Qwen3-VL-8B	0.67	0.74	0.69	0.38	0.71	0.74	0.69	0.42
Qwen3-VL-32B	0.67	0.79	0.70	0.39	0.80	0.80	**0.81**	0.61
GPT-4.1-mini	0.70	0.78	0.72	0.43	0.78	0.80	0.76	0.56
GPT-4.1	**0.81**	**0.84**	**0.82**	**0.63**	**0.83**	**0.86**	**0.81**	**0.66**

Comparison Between Evaluation Methods

Methods	Holistic ↑				Partial ↑			
	F1	Precision	Recall	Cohen's κ	F1	Precision	Recall	Cohen's κ
CLIPScore	0.49	0.51	0.51	0.02	0.46	0.50	0.50	0.00
VLM-as-a-judge								
Gemini 2.5 Pro	0.38	0.76	0.52	0.05	0.60	0.66	0.70	0.28
GPT-4.1	0.40	0.67	0.53	0.05	0.68	0.65	0.73	0.35
GPT-5.1	0.38	0.68	0.52	0.04	0.60	0.71	0.71	0.28
SceneEval*								
GPT-4o	0.47	0.74	0.58	0.15	0.45	0.64	0.58	0.12
LEGO-EVAL								
Qwen3-VL-8B	0.67	0.74	0.69	0.38	0.71	0.74	0.69	0.42
Qwen3-VL-32B	0.67	0.79	0.70	0.39	0.80	0.80	**0.81**	0.61
GPT-4.1-mini	0.70	0.78	0.72	0.43	0.78	0.80	0.76	0.56
GPT-4.1	**0.81**	**0.84**	**0.82**	**0.63**	**0.83**	**0.86**	**0.81**	**0.66**

Why do VLM-as-a-Judge Fail?

Method	Total	Ident.	Spa.	Attr.
VLM-as-a-judge	376	150	76	150
+ **Grounding DINO**	451	288	42	121
+ **LASER**	497	233	100	164
+ **Point-LLM**	447	203	72	172
LEGO-Eval	**119**	**53**	**20**	**46**

Identification Error: Cannot ground scene elements mentioned in the description

Spatial Error: Cannot identify spatial relations between scene elements

Attribute Error: Cannot identify attribute of scene elements

Why do VLM-as-a-Judge Fail?

Method	Total	Ident.	Spa.	Attr.
VLM-as-a-judge	376	150	76	150
+ **Grounding DINO**	451	288	42	121
+ **LASER**	497	233	100	164
+ **Point-LLM**	447	203	72	172
LEGO-Eval	**119**	**53**	**20**	**46**

Bounding Box

Scene Graph

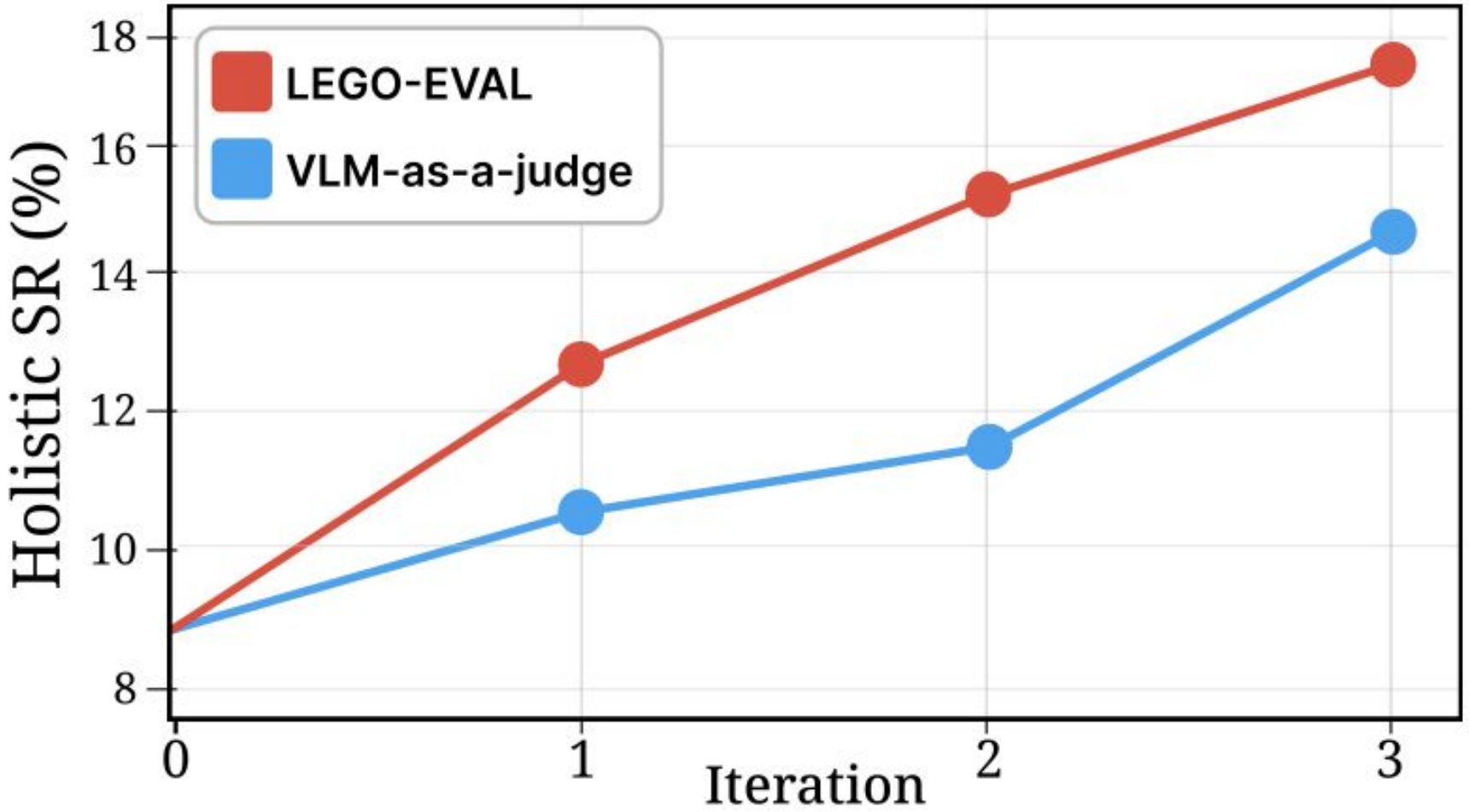
Point Cloud

Why do VLM-as-a-Judge Fail?

Methods	Holistic ↑		Partial ↑	
	F1	Cohen's Kappa	F1	Cohen's Kappa
Scene Images	0.40	0.05	0.68	0.35
SceneGraph	0.37	0.04	0.66	0.36
SceneGraph + Scene Images	0.37	0.04	0.65	0.35
LEGO-Eval (Ours)	**0.81**	**0.63**	**0.83**	**0.66**

More information does not always lead to better results.
'Which' information and 'how' to use them is equally important.

Refinement using feedback from evaluation methods



Refinement using feedback from evaluation methods

LEGO-Eval

Description

There is only one room in the scene. The room is square shaped. There are two large windows on one side of the room. There is one laptop, one flashlight, and two chairs. **The flashlight and the laptop are facing the same direction.**



VLM-as-a-Judge *Invalid* ✗

The laptop is on a side table under the TV on the right wall. The flashlight lies on the round table next to the armchair. **Since the laptop faces into the room and the flashlight faces toward the wall, their orientations do not match.**

SceneEval *Invalid* ✗

There is a laptop in the scene. There is no flashlight in the scene. Not all constrained objects are present, so the constraint is unsatisfied.

LEGO-Eval *Valid* ✓

Since neither object is present, there is no way to assess whether the flashlight and the laptop are oriented. Thus, the constraint cannot be satisfied.

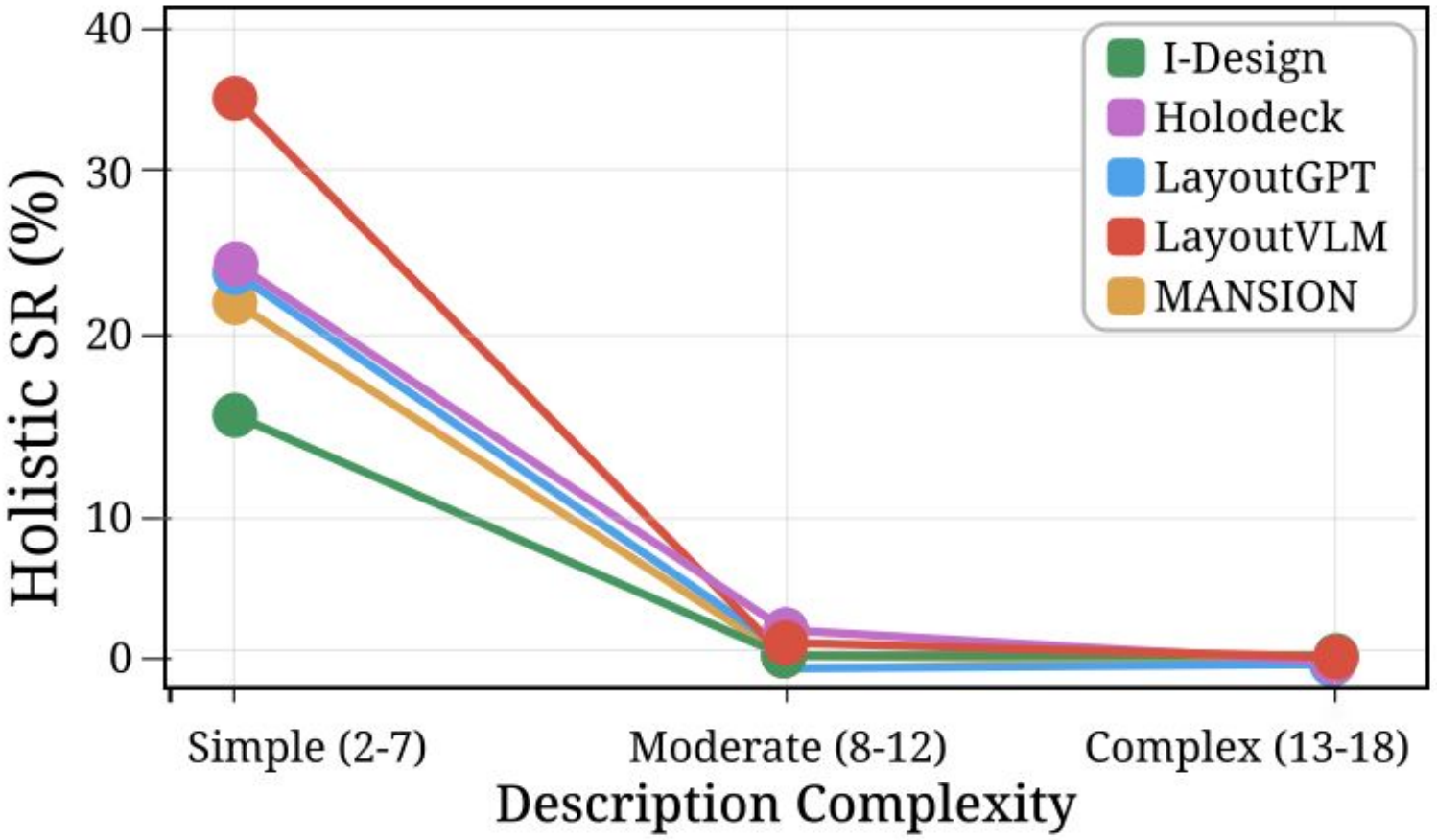
Existing methods Fail to Generate Fully Aligned Scenes

LEGO-Eval

Method	Holistic SR ↑	Partial SR ↑				
		Floor Layout	Material Selection	Object Selection	Object Placement	Avg.
I-Design	3.8	92.7	63.7	11.0	4.1	34.2
MANSION	5.4	67.7	53.9	55.7	56.1	58.2
LayoutGPT	6.9	96.0	**65.3**	40.9	37.3	55.2
Holodeck	8.4	**96.3**	61.6	46.3	43.7	58.5
LayoutVLM	**10.0**	95.6	**65.3**	**49.8**	**46.0**	**60.6**

Existing methods Fail to Generate Fully Aligned Scenes

LEGO-Eval



Questions?



Thank You

Gyeom Hwangbo

POSTECH, MOAI Lab

APPENDIX. WebPRM Collection

