Selective Visual Representations Improve Convergence and Generalization for Embodied-AI

embodied-codebook.github.io

Ainaz Eftekhar*, Kuo-Hao Zeng*, Jiafei Duan, Ali Farhadi
Ani Kembhavi, Ranjay Krishna



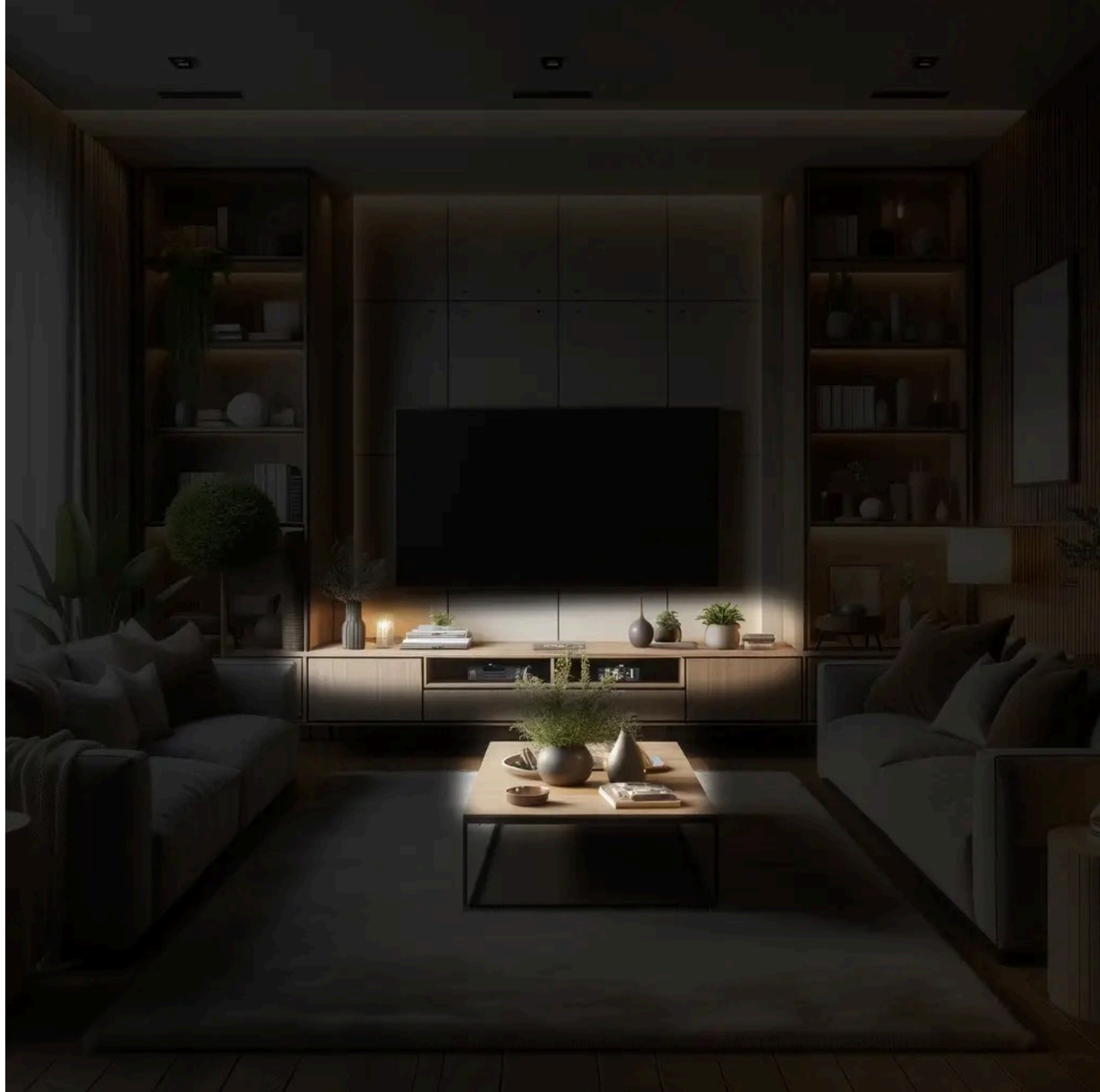



ICLR 2024 [Spotlight]

Human perception is a **selective** mechanism

Human perception is a **selective** mechanism

Find the remote control



Human perception is a **selective** mechanism

Find your books

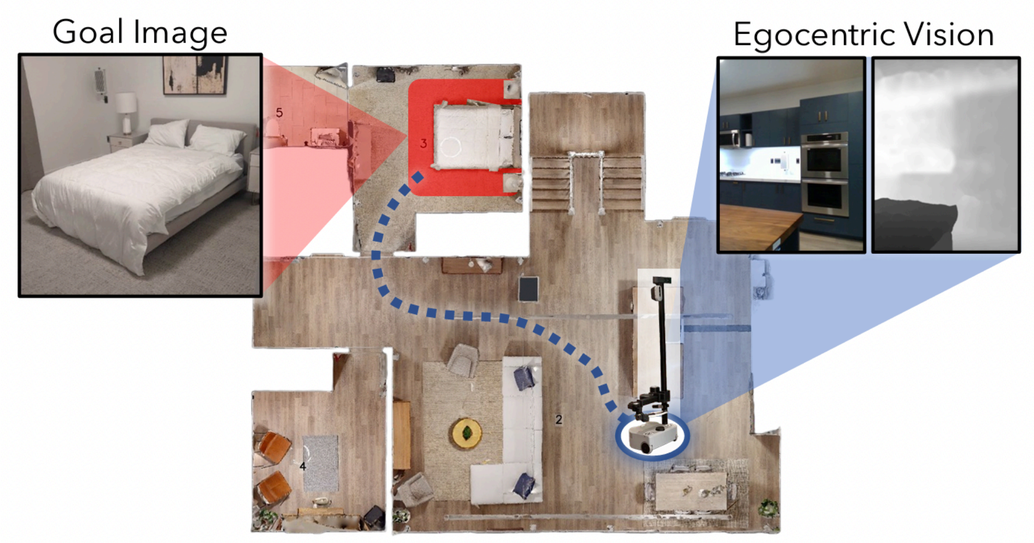
Top-down Selective Attention

Human perception is filtered based on the internal goals.

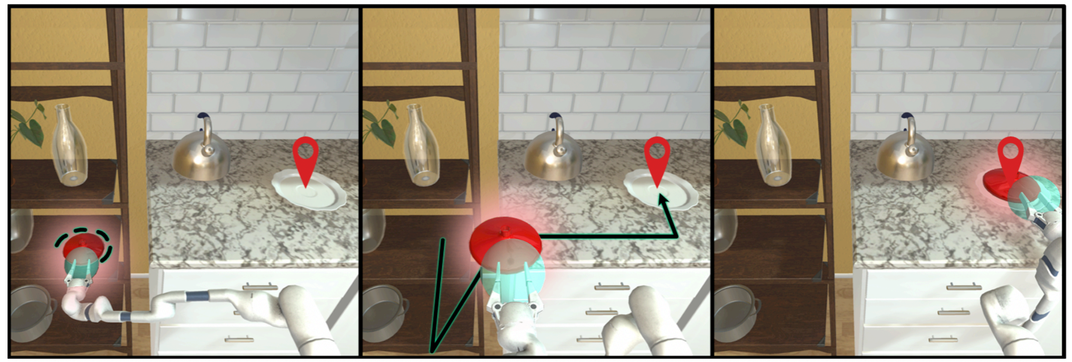
Desimone, R., & Duncan, J. (1995). Neural mechanisms of selective visual attention. Annual review of neuroscience, 18(1), 193-222.
Simons, D. J., & Chabris, C. F. (1999). Gorillas in our midst: Sustained inattentional blindness for dynamic events. perception, 28(9), 1059-1074.

Embodied-AI agents have **goal-driven** behaviors

Image-Goal Navigation



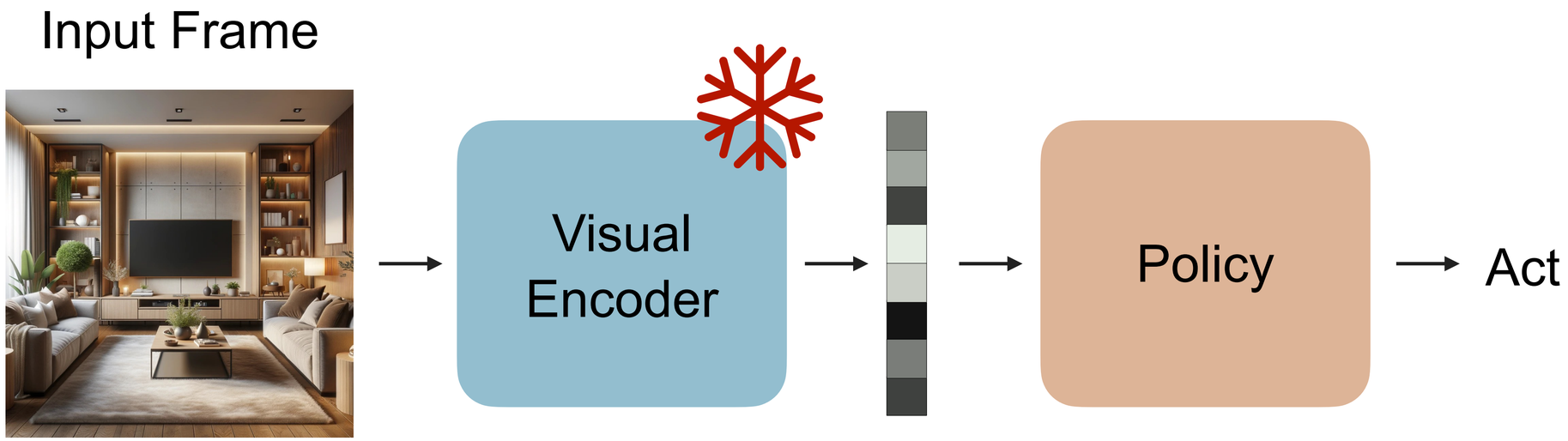
Mobile Manipulation



ManipulaTHOR, Ehsani et al., CVPR 2021

Navigating to Objects Specified by Images, Krantz et al., CVPR 2023

Embodied-AI agents use **general-purpose** visual backbones



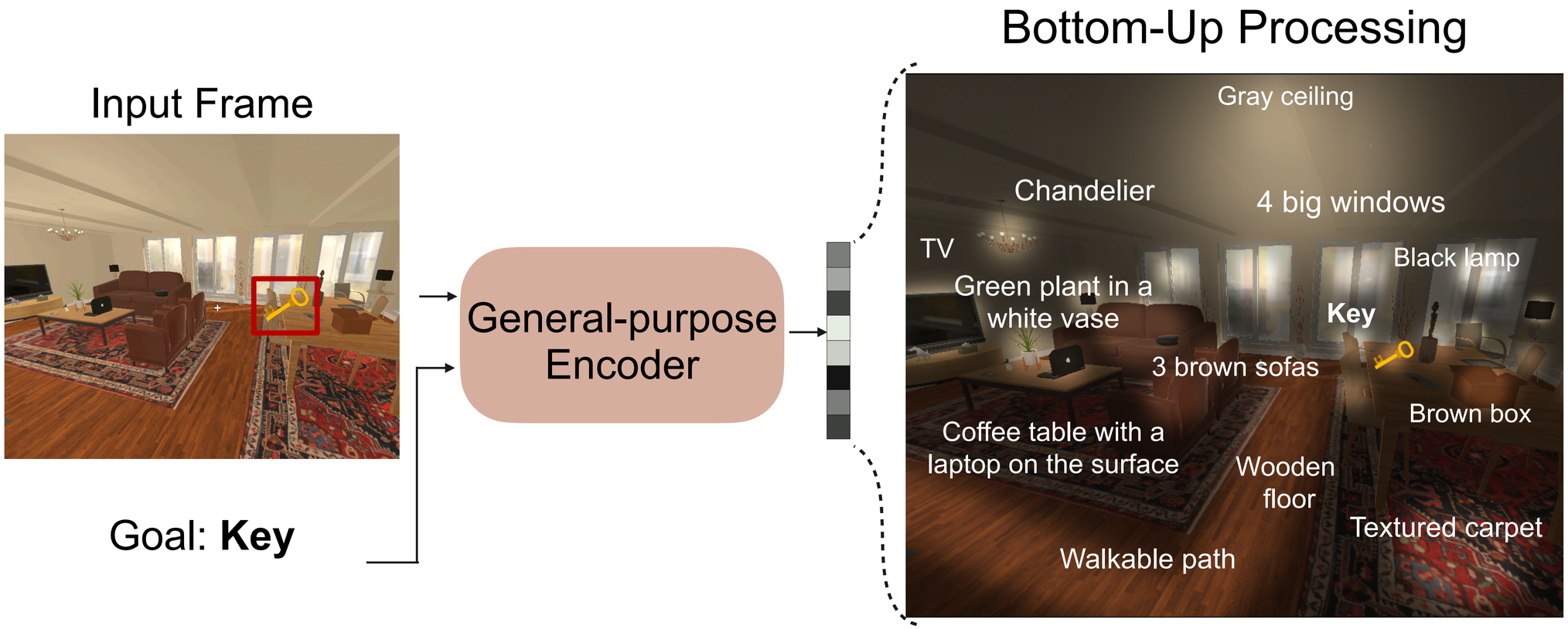
EmbCLIP, Khandelwal et al., CVPR 2022

SPOC, Ehsani et al., CVPR 2024

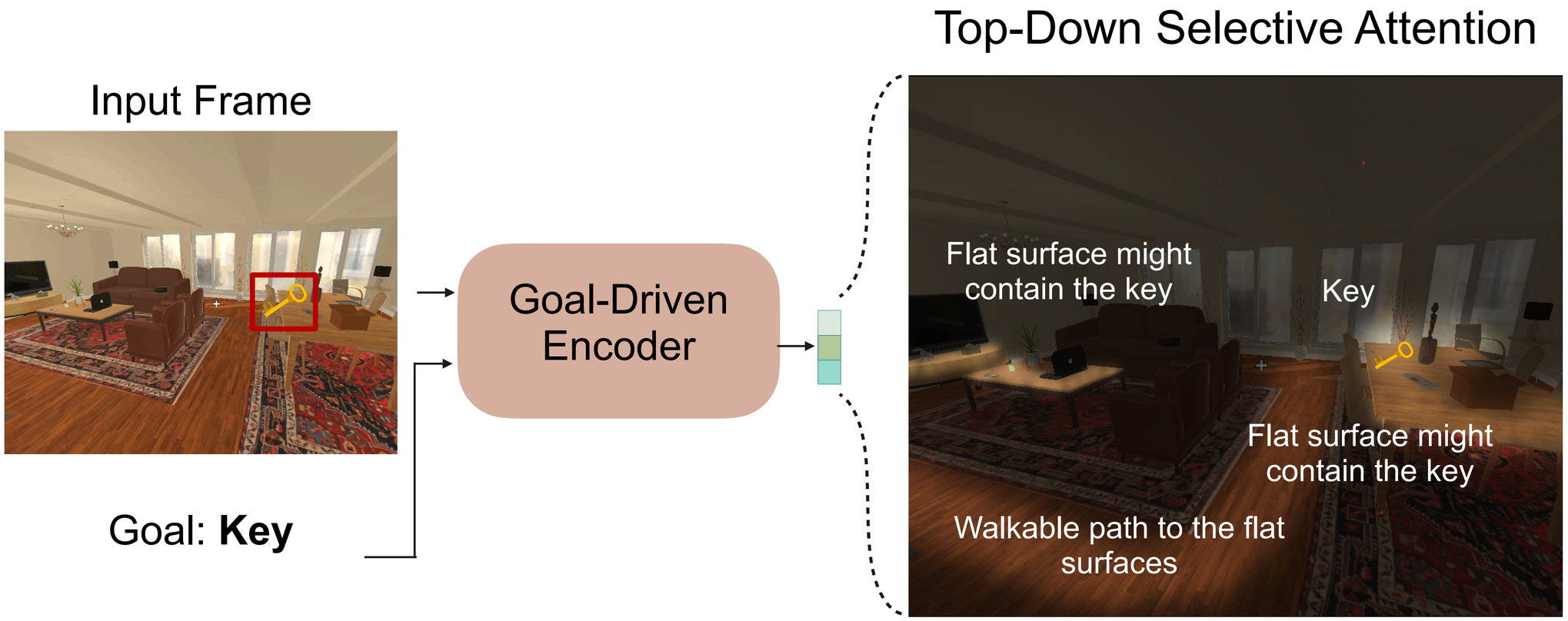
Standard visual encoders capture **general-purpose** scene information

Task: Find the key

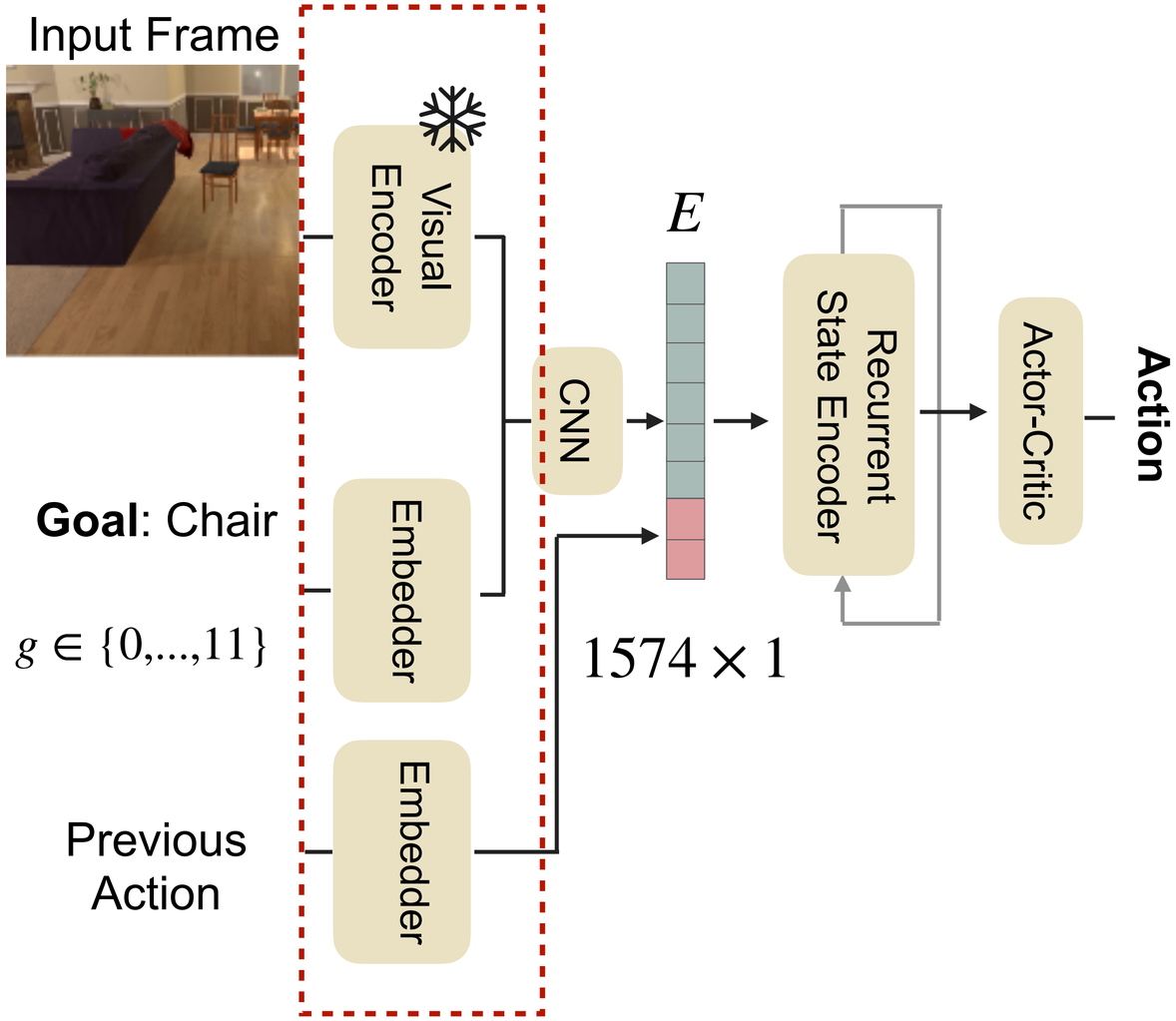
Standard visual encoders capture **general-purpose** scene information



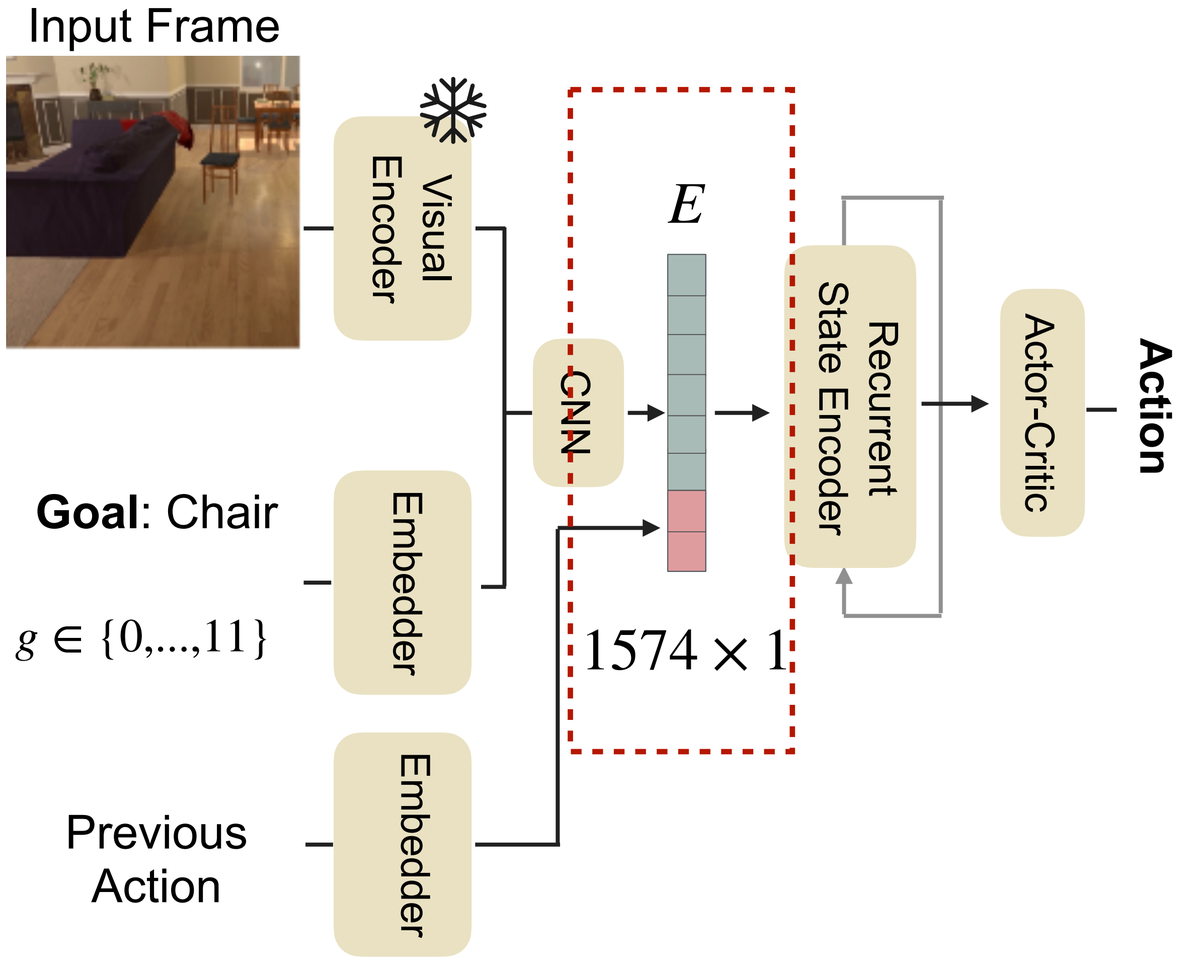
Goal-Driven Visual Encoder retains the most task-relevant information



Standard Embodied-AI architectures

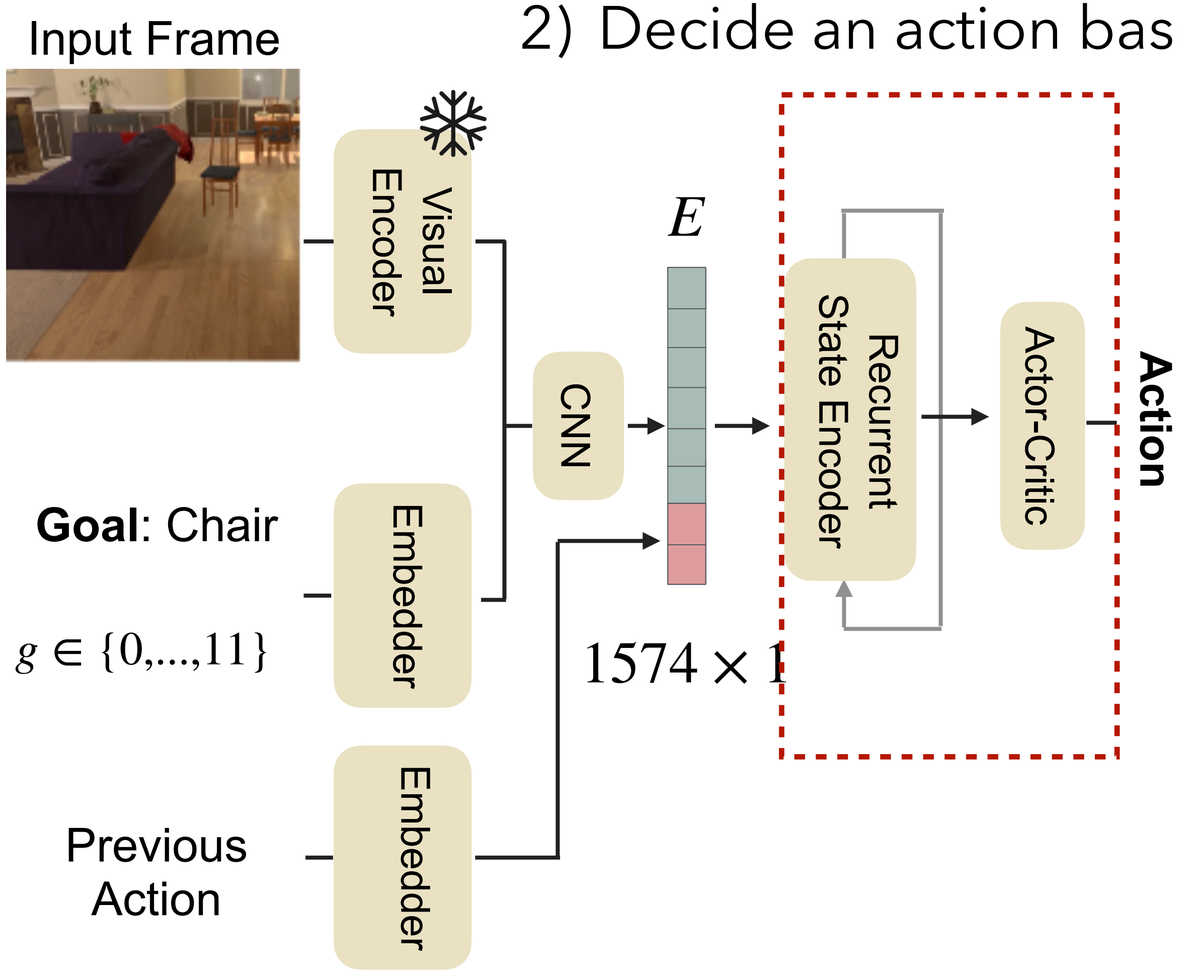


Standard Embodied-AI architectures

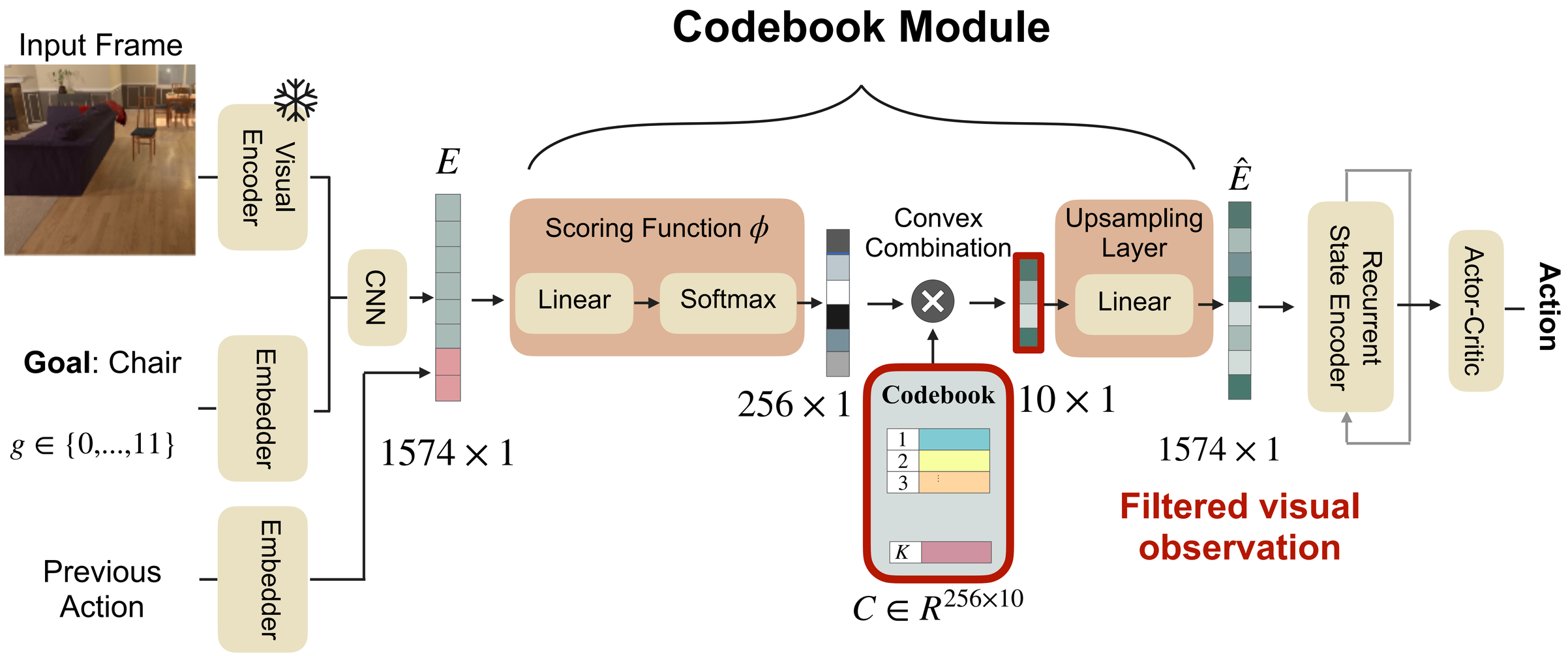


Standard Embodied-AI architectures

1) Remember what it has seen in the past
2) Decide an action based on the information

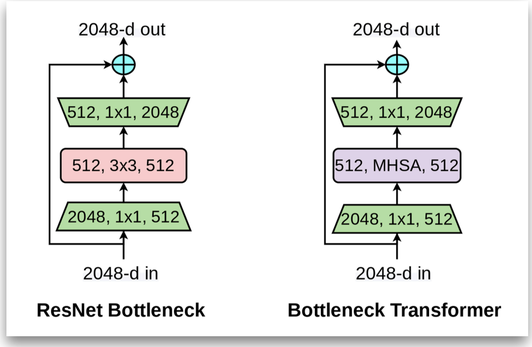


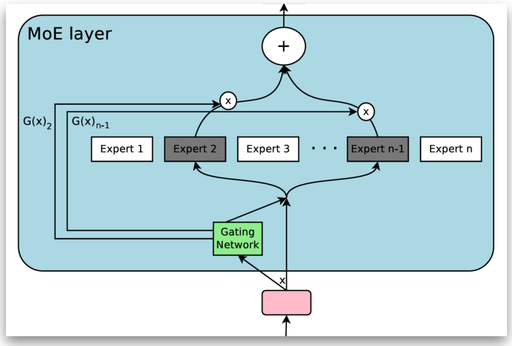
Goal-Bottlenecked Architecture



Bottleneck-based Architectures

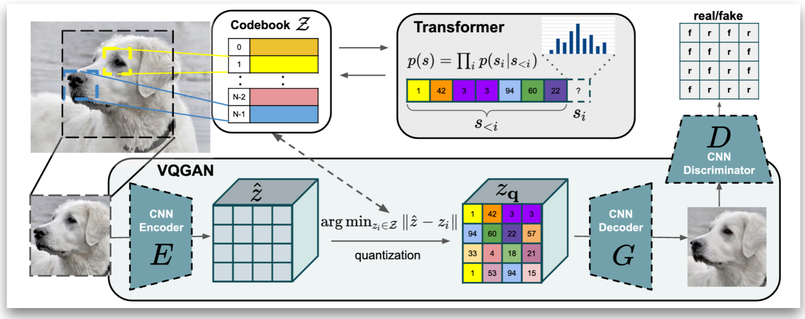
Bottleneck Transformer
Srinivas et al.
CVPR 2021



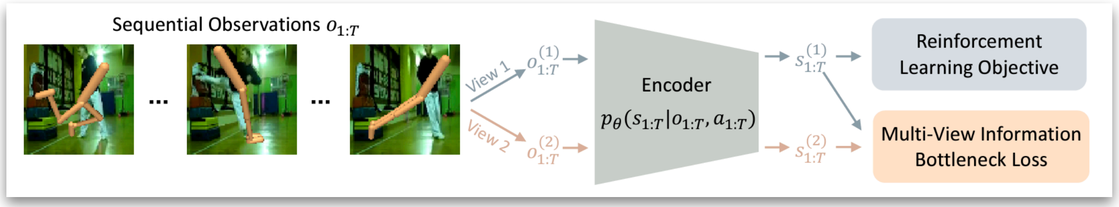


Mixture-of-Experts
Shazeer et al.
ICLR 2017

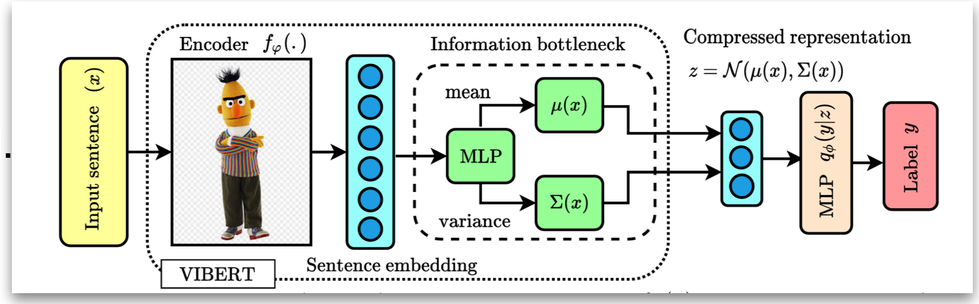
VQGAN
Esser et al.
CVPR 2021



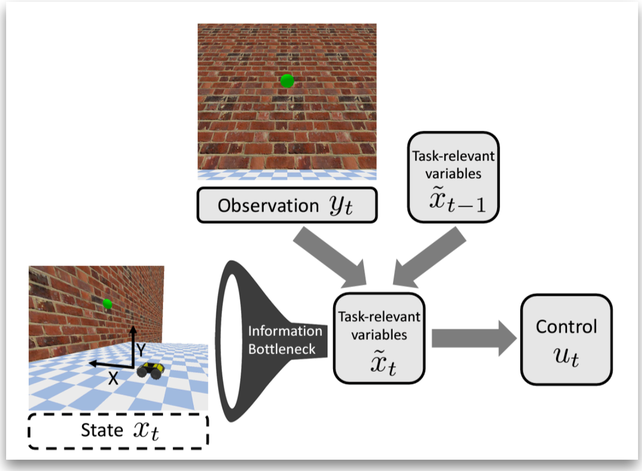
DRIBO
Fan et al.
PMLR 2022



VIBERT
Mahabadi et al.
ICLR 2021



Task-Driven Control Policies
Pacelli et al.
RSS 2020



Benefits of **codebook-bottlenecked** representations in Embodied-AI

i. Improve performance and convergence in Embodied-AI

ii. Improved agent behavior: smoother trajectories and more efficient exploration

iii. More generalizable to new visual domains

iv. Captures the most task-relevant information

v. Representation-agnostic and applicable to various visual encoders

Benefits of **codebook-bottlenecked** representations in Embodied-AI

i. **Improve performance and convergence in Embodied-AI**

ii. Improved agent behavior: smoother trajectories and more efficient exploration

iii. More generalizable to new visual domains

iv. Captures the most task-relevant information

v. Representation-agnostic and applicable to various visual encoders

Codebook-Bottlenecked Representations **Improve Performance** in Embodied-AI

Benchmark	Model	Object navigation SR(%)↑	EL↓	Curvature↓	SPL↑	SEL↑
ProcTHOR-10k (validation)	EmbCLIP	67.70	182.00	0.58	**49.00**	36.00
	+codebook	**73.72**	**136.00**	**0.23**	48.37	**43.69**
ARCHITECTHOR (0-shot)	EmbCLIP	55.80	222.00	0.49	**38.30**	20.57
	+Codebook	**58.33**	**174.00**	**0.20**	35.57	**28.31**
RoboTHOR (0-shot)	EmbCLIP	51.32	-	-	**24.24**	-
	+Codebook	**55.00**	-	-	23.65	-
AI2-iTHOR (0-shot)	EmbCLIP	70.00	121.00	0.29	**57.10**	21.45
	+Codebook	**78.40**	**86.00**	**0.16**	54.39	**26.76**

		Object displacement PU(%)↑	SR(%)↑
ManipulaTHOR	m-VOLE	81.20	59.60
	+Codebook	**86.00**	**65.10**

EmbCLIP, Khandelwal et al., CVPR 2022

Codebook-Bottlenecked Representations **Improve Performance** in Embodied-AI

Benchmark	Model	Object navigation SR(%)↑	EL↓	Curvature↓	SPL↑	SEL↑
ProcTHOR-10k (validation)	EmbCLIP	67.70	182.00	0.58	**49.00**	36.00
	+codebook	**73.72**	**136.00**	**0.23**	48.37	**43.69**
ARCHITECTHOR (0-shot)	EmbCLIP	55.80	222.00	0.49	**38.30**	20.57
	+Codebook	**58.33**	**174.00**	**0.20**	35.57	**28.31**
RoboTHOR (0-shot)	EmbCLIP	51.32	-	-	**24.24**	-
	+Codebook	**55.00**	-	-	23.65	-
AI2-iTHOR (0-shot)	EmbCLIP	70.00	121.00	0.29	**57.10**	21.45
	+Codebook	**78.40**	**86.00**	**0.16**	54.39	**26.76**

		Object displacement PU(%)↑	SR(%)↑
ManipulaTHOR	m-VOLE	81.20	59.60
	+Codebook	**86.00**	**65.10**

Codebook-Bottlenecked Representations **Improve Performance** in Embodied-AI

Benchmark	Model	Object navigation SR(%)↑	EL↓	Curvature↓	SPL↑	SEL↑
ProcTHOR-10k (validation)	EmbCLIP	67.70	182.00	0.58	**49.00**	36.00
	+codebook	**73.72**	**136.00**	**0.23**	48.37	**43.69**
ARCHITECTHOR (0-shot)	EmbCLIP	55.80	222.00	0.49	**38.30**	20.57
	+Codebook	**58.33**	**174.00**	**0.20**	35.57	**28.31**
RoboTHOR (0-shot)	EmbCLIP	51.32	-	-	**24.24**	-
	+Codebook	**55.00**	-	-	23.65	-
AI2-iTHOR (0-shot)	EmbCLIP	70.00	121.00	0.29	**57.10**	21.45
	+Codebook	**78.40**	**86.00**	**0.16**	54.39	**26.76**

		Object displacement PU(%)↑	SR(%)↑
ManipulaTHOR	m-VOLE	81.20	59.60
	+Codebook	**86.00**	**65.10**

Benefits of **codebook-bottlenecked** representations in Embodied-AI

i. Improve performance and convergence in Embodied-AI

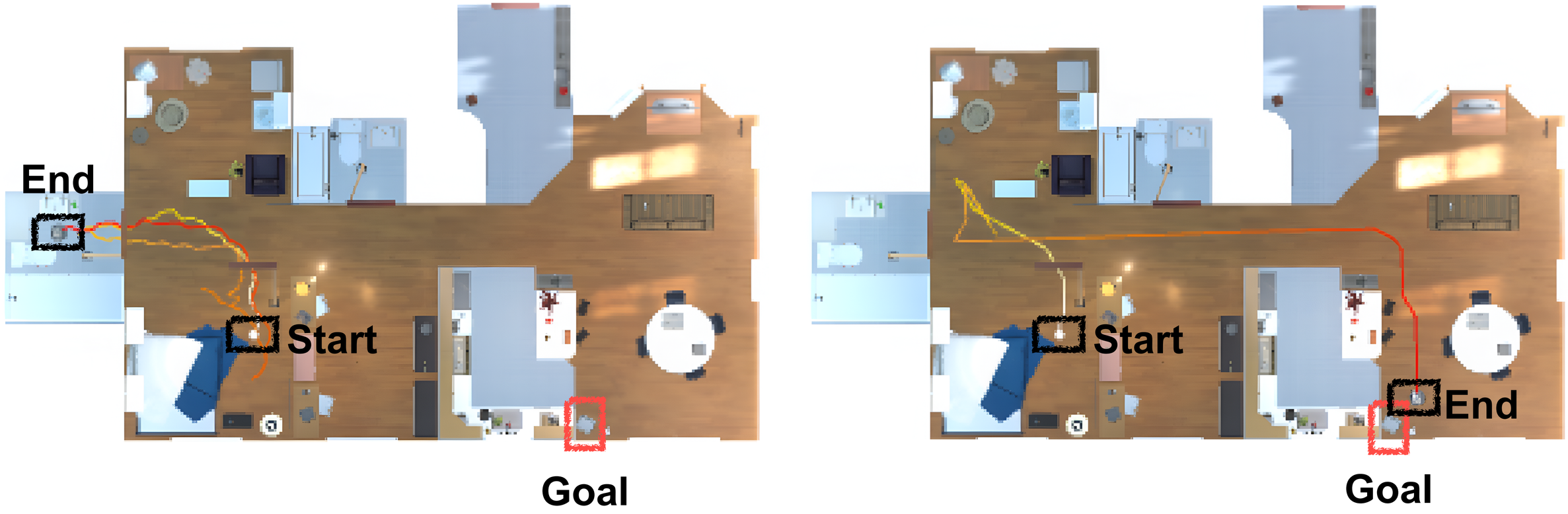
ii. **Improved agent behavior: smoother trajectories and more efficient exploration**

iii. More generalizable to new visual domains

iv. Captures the most task-relevant information

v. Representation-agnostic and applicable to various visual encoders

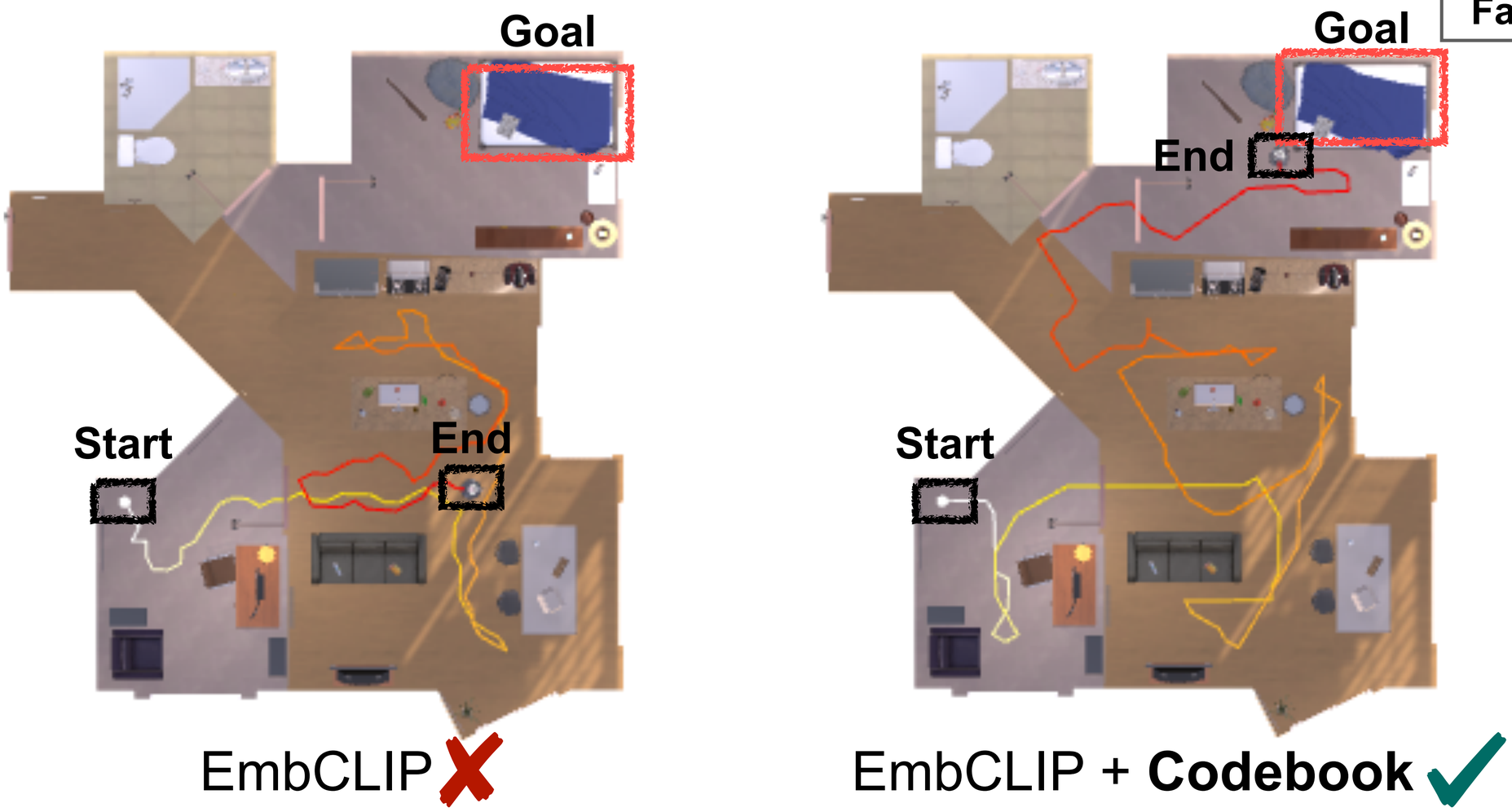
Our agent **explores more efficiently** and in **smoother trajectories**



EmbCLIP ✗

EmbCLIP + **Codebook** ✓

Our agent **explores more efficiently** and in **smoother trajectories**



Our agent **explores more efficiently** and in **smoother trajectories**

Success ✓
Fail ✗

EmbCLIP ✗

Our agent **explores more efficiently** and in **smoother trajectories**

Success ✓
Fail ✗

EmbCLIP + **Codebook** ✓

Benefits of **codebook-bottlenecked** representations in Embodied-AI

i. Improve performance and convergence in Embodied-AI

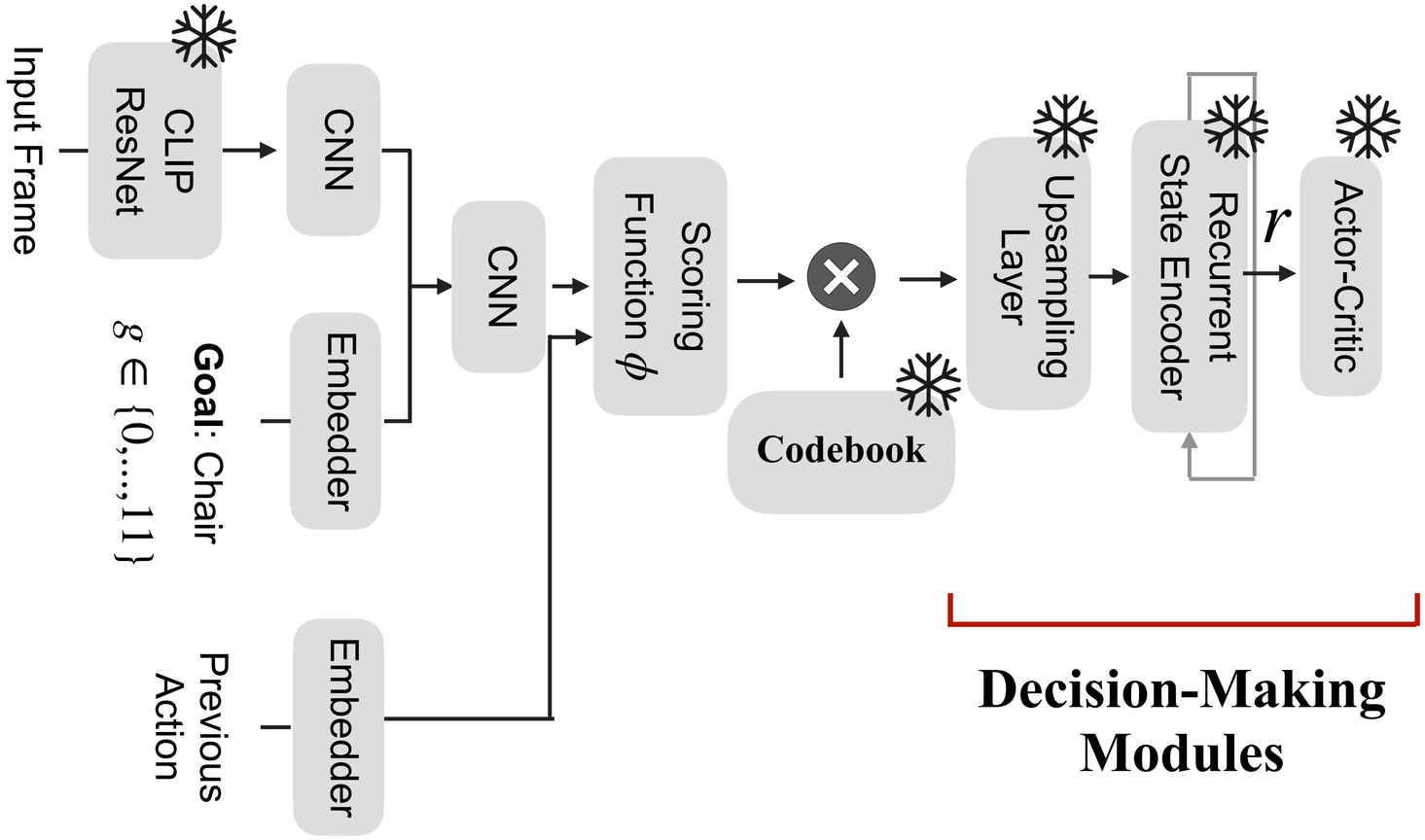
ii. Improved agent behavior: smoother trajectories and more efficient exploration

iii. More generalizable to new visual domains

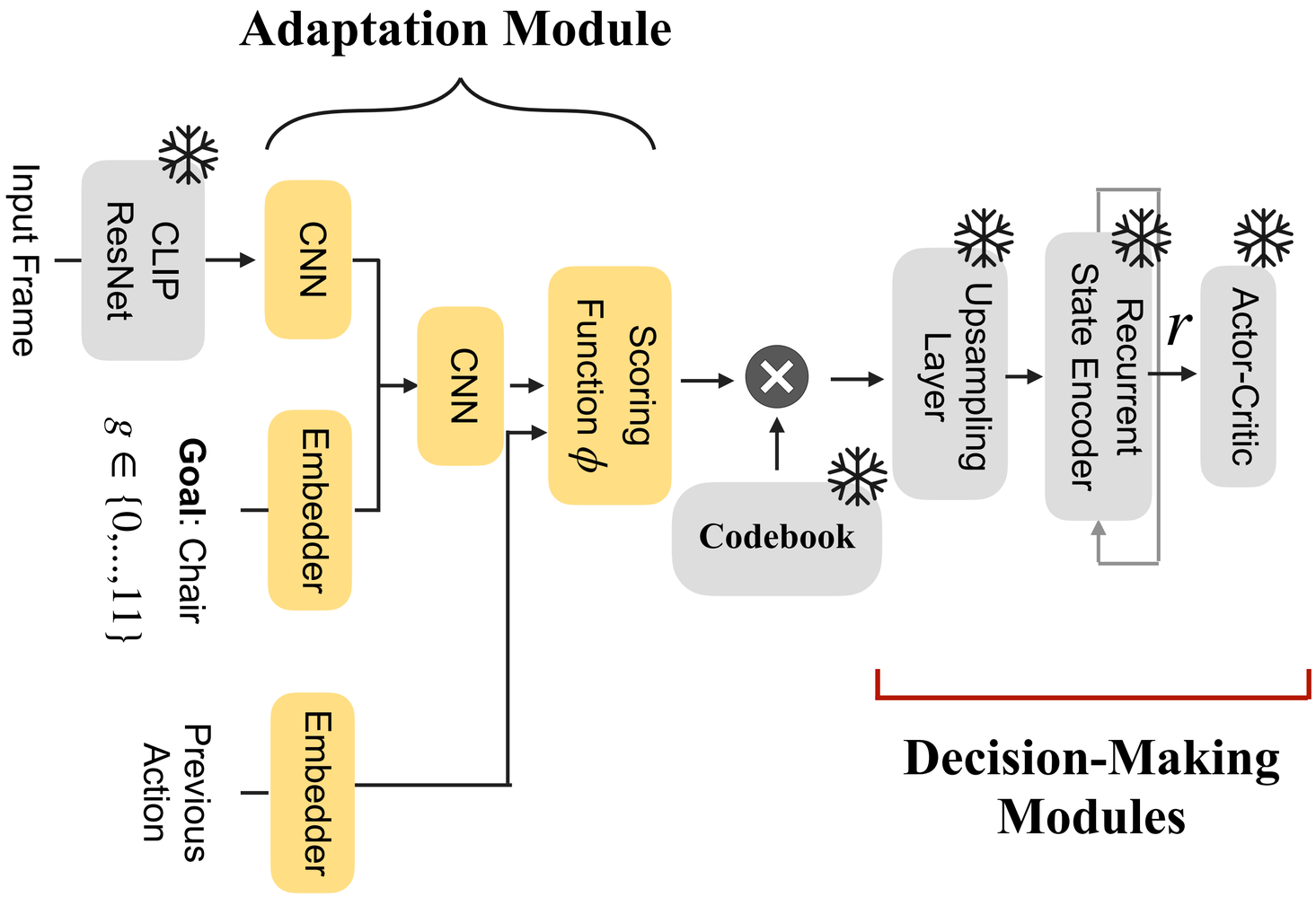
iv. Captures the most task-relevant information

v. Representation-agnostic and applicable to various visual encoders

Codebook embeddings **generalize to new visual domains**



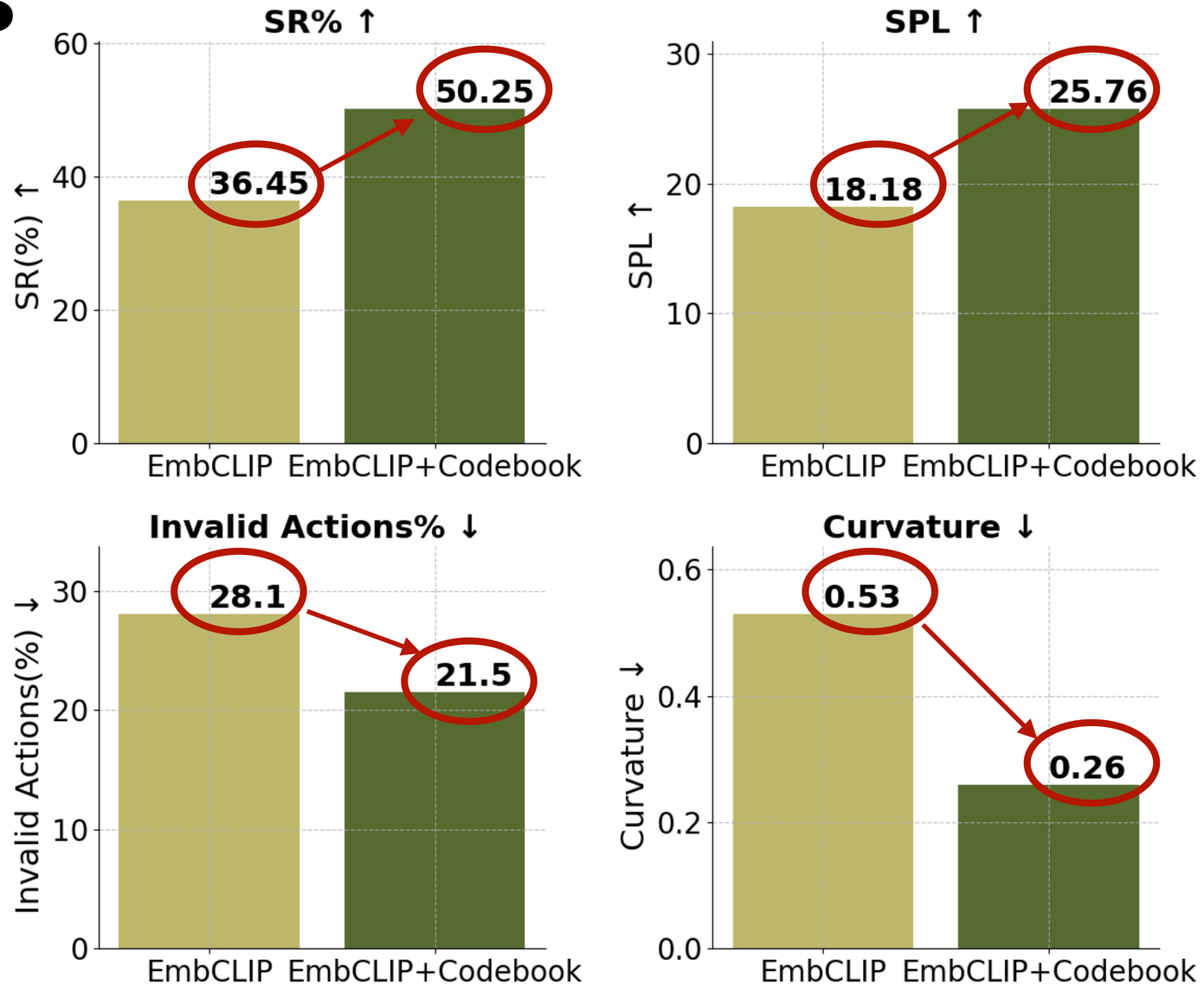
Codebook embeddings **generalize to new visual domains**



Codebook embeddings **generalize to new visual domains**

Pre-train: AI2THOR

Finetune: Habitat



Benefits of **codebook-bottlenecked** representations in Embodied-AI

i. Improve performance and convergence in Embodied-AI

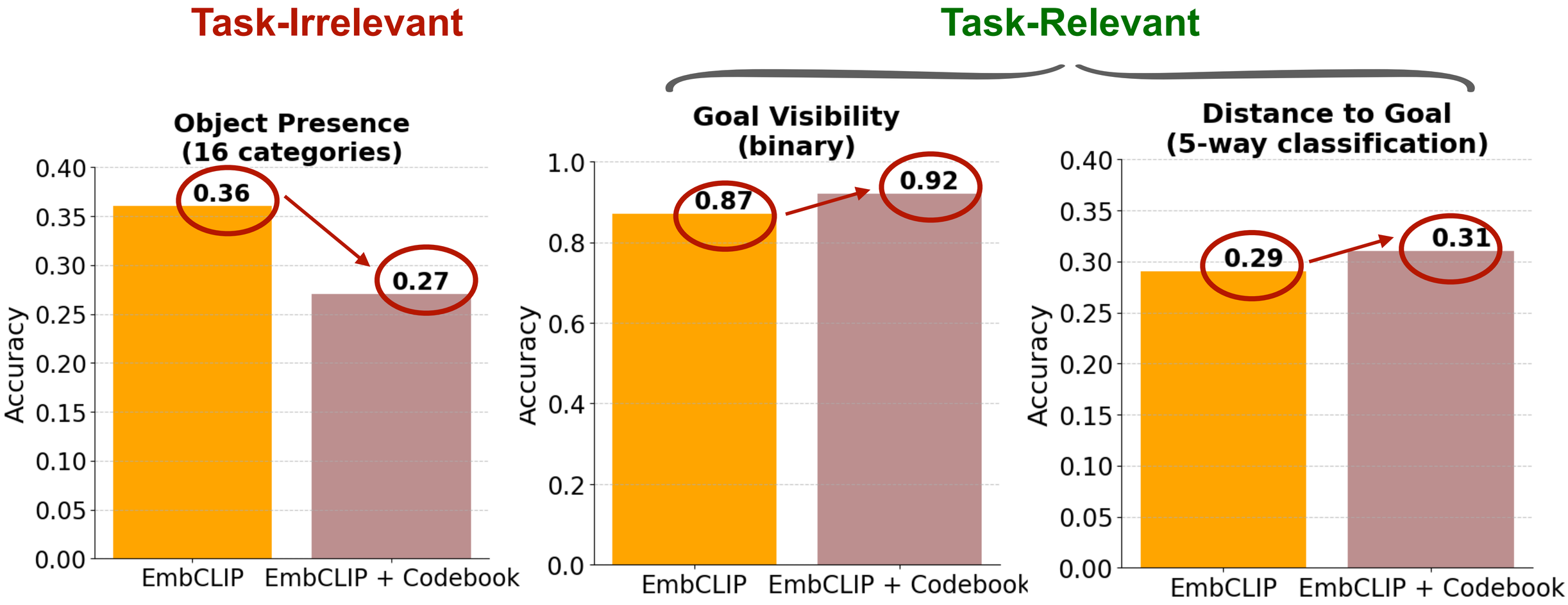
ii. Improved agent behavior: smoother trajectories and more efficient exploration

iii. More generalizable to new visual domains

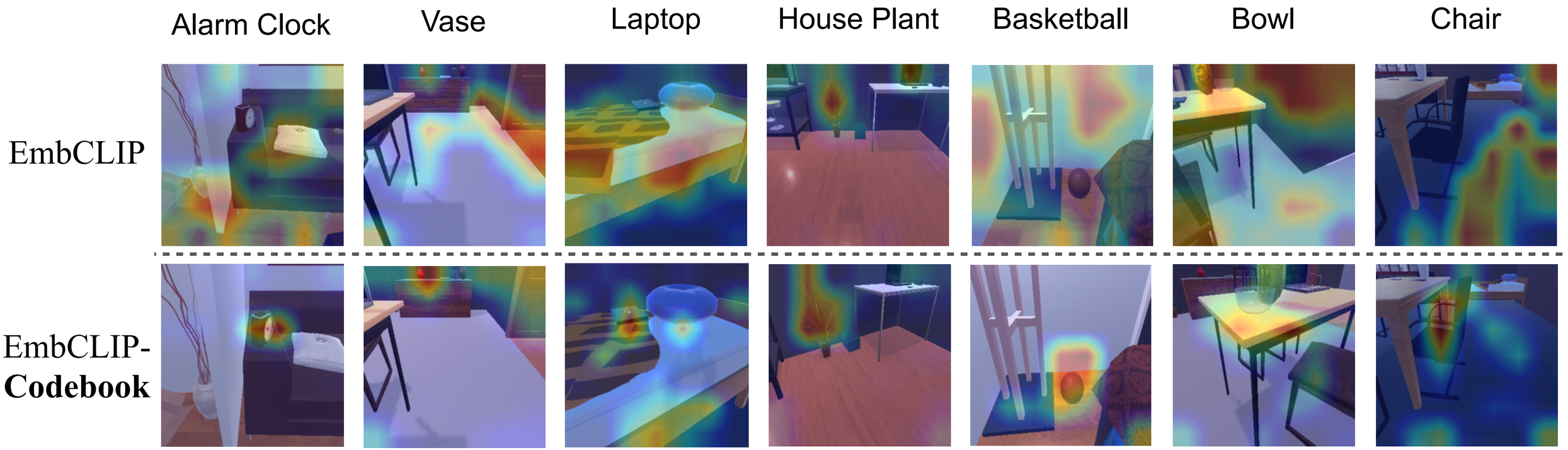
iv. Captures the most task-relevant information

v. Representation-agnostic and applicable to various visual encoders

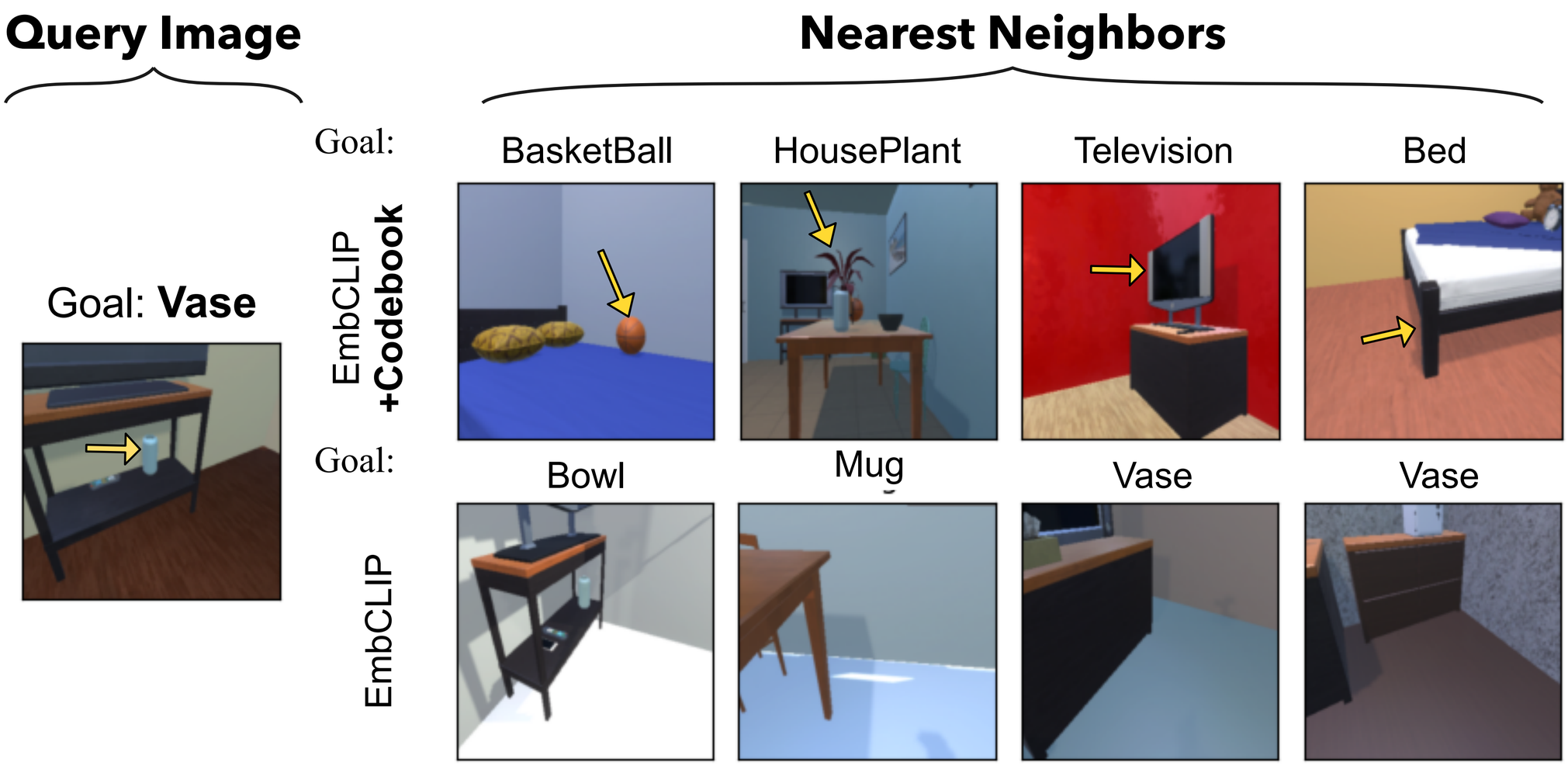
Codebook-bottlenecked embeddings retain the most **task-relevant** information



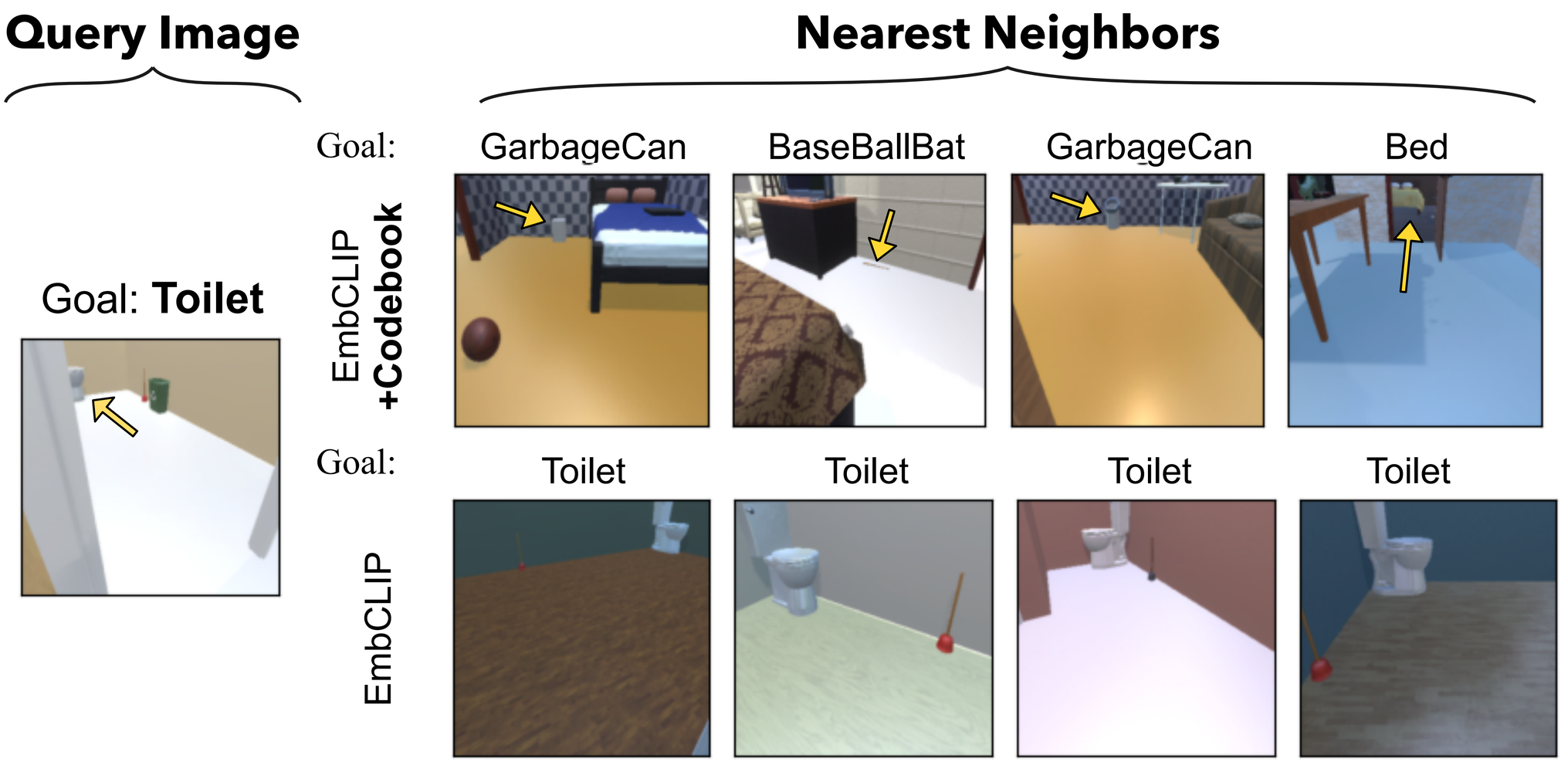
Codebook-bottlenecked embeddings retain the most **task-relevant** information



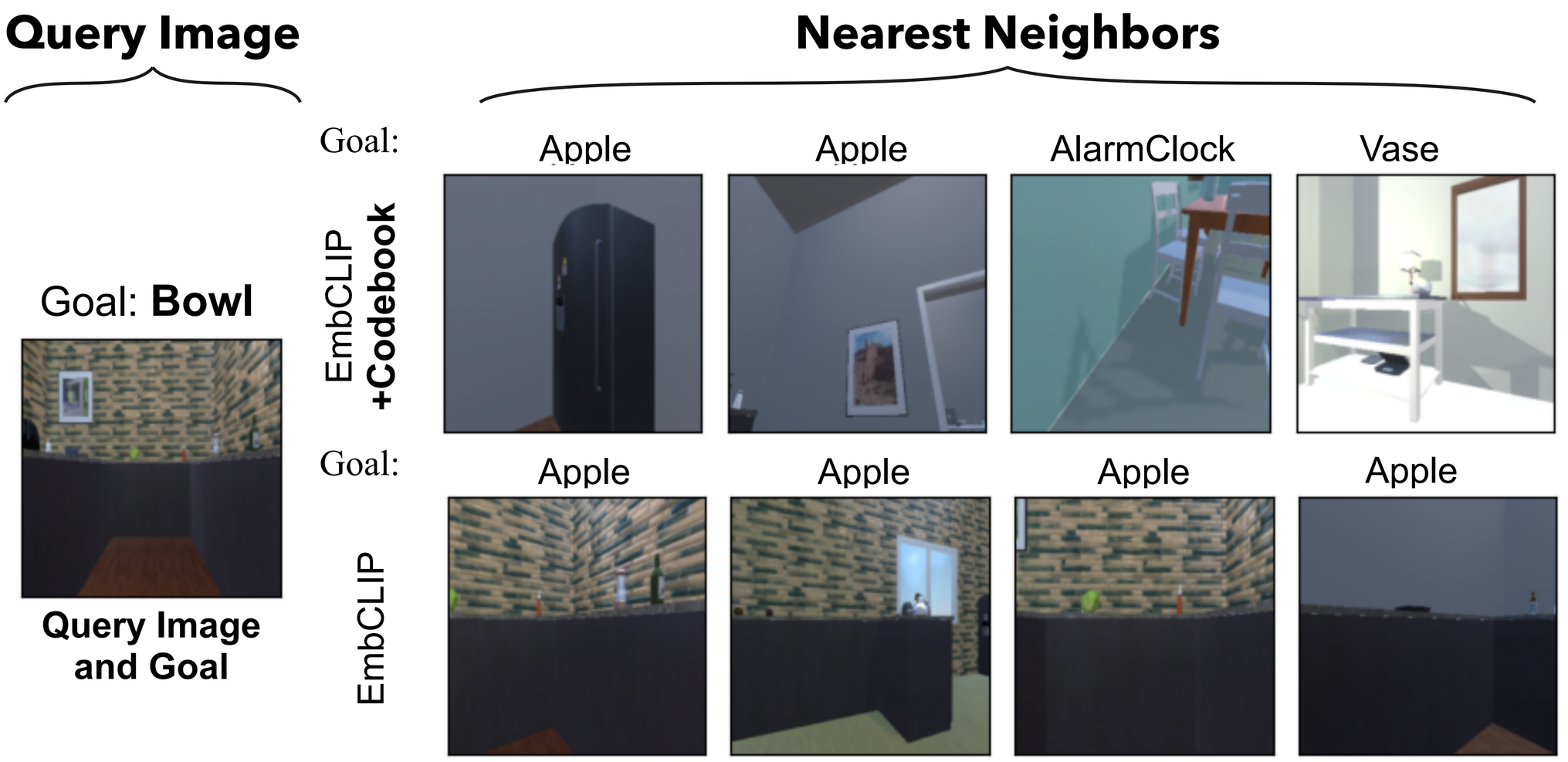
Codebook-bottlenecked embeddings retain the most **task-relevant** information



Codebook-bottlenecked embeddings retain the most **task-relevant** information



Codebook-bottlenecked embeddings retain the most **task-relevant** information



Benefits of **codebook-bottlenecked** representations in Embodied-AI

i. Improve performance and convergence in Embodied-AI

ii. Improved agent behavior: smoother trajectories and more efficient exploration

iii. More generalizable to new visual domains

iv. Captures the most task-relevant information

v. Representation-agnostic and applicable to various visual encoders

Codebook module is **representation-agnostic**

Benchmark	Model	Object navigation SR(%)↑	EL↓	Curvature↓	SPL↑	SEL↑
ProcTHOR-10k (validation)	DINOv2 (Oquab et al., 2023)	74.25	151.00	0.24	49.53	43.20
	+Codebook (Ours)	**76.31**	**129.00**	**0.12**	**50.26**	**44.70**
ARCHITECTHOR (0-shot)	DINOv2	57.25	218.00	0.25	36.83	29.00
	+Codebook (Ours)	**59.75**	**194.00**	**0.11**	**36.00**	**31.70**
AI2-iTHOR (0-shot)	DINOv2	74.67	97.00	0.19	59.45	26.50
	+Codebook (Ours)	**76.93**	**68.00**	**0.07**	**60.14**	**28.30**
RoboTHOR (0-shot)	DINOv2	60.54	-	-	**29.36**	-
	+Codebook (Ours)	**61.03**	-	-	28.01	-

DINOv2, Squab et al., TMLR 2024

Code and Pretrained Models

github.com/allenai/procthor-rl

Website

embodied-codebook.github.io

Selective Visual Representations Improve Convergence and Generalization for Embodied-AI

Ainaz Eftekhar*, Kuo-Hao Zeng*, Jiafei Duan, Ali Farhadi
Ani Kembhavi, Ranjay Krishna

ICLR 2024 [Spotlight]