香港科技大學(廣州)
THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

Tencent AI Lab

Prioritized Semantic Learning for Zero-shot Instance Navigation

Xinyu Sun[1*] Lizhao Liu[3*] Hongyan Zhi Ronghe Qiu[1] Junwei Liang[1,2†]

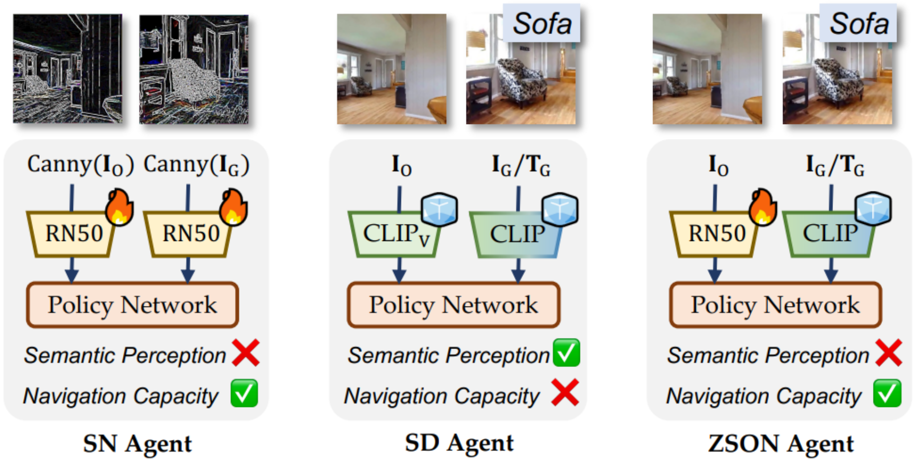
[1] AI Thrust, The Hong Kong University of Science and Technology (Guangzhou)
[2] Department of Computer Science and Engineering, The Hong Kong University of Science and Technology
[3] Tencent AI Lab, Shenzhen, China

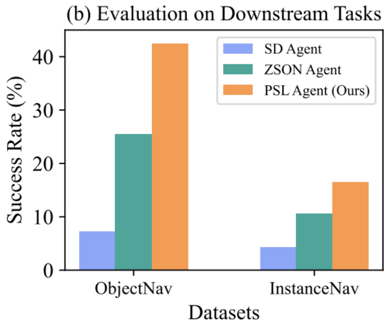
https://github.com/XinyuSun/PSL-InstanceNav

EUROPEAN CONFERENCE ON COMPUTER VISION
MILANO 2024

ECCV 2024

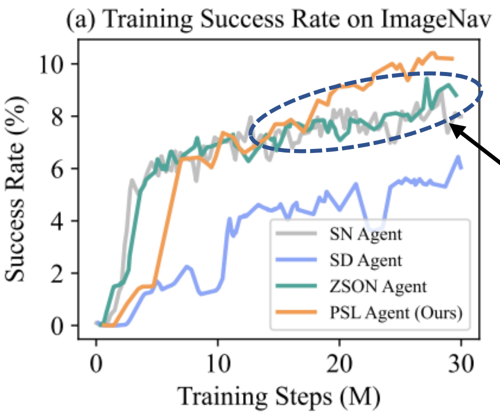
Motivation



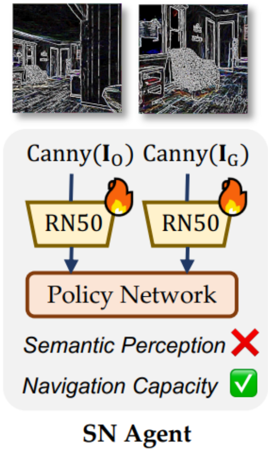


Existing methods can not posses Semantic Perception and Navigation Capacity **simultaneously**.

Motivation

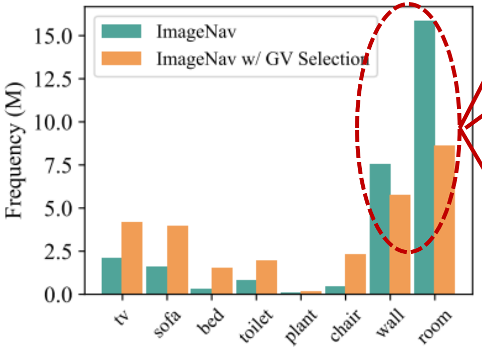


The SN agent can achieve high **ImageNav** success rate without perceiving semantic clues.



Existing **ImageNav** pre-training dataset is not suitable for training a **semantic navigation agent**.

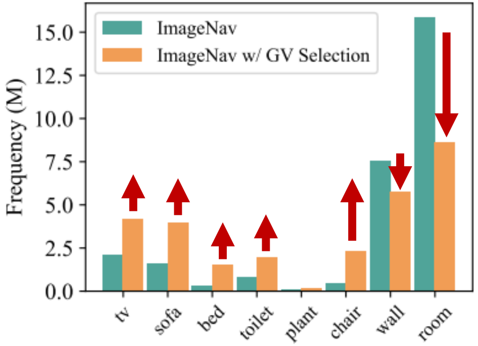
Motivation

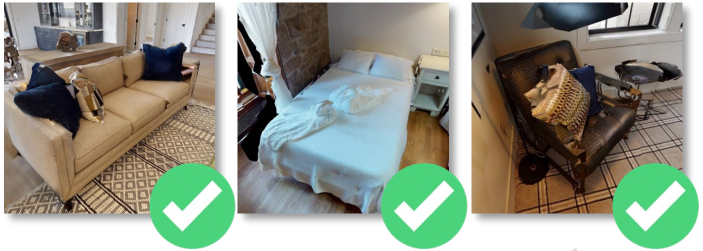


A majority of the images are meaning-less (*wall, ceiling, others…*).

The images in **ImageNav** dataset suffer from **unreasonable category distribution.**

Method: *Entropy-minimized Goal View Selection.*





Increasing the frequency of **objects** in images &&
Reducing the freq. of **meaningless regions** in images.

Method: *Entropy-minimized Goal View Selection.*

$$\omega^* = \arg\min_{\omega \in \Omega} -\frac{1}{\log(|\mathcal{C}|)} \sum_{c \in \mathcal{C}} \mathbf{p}_c \log \mathbf{p}_c, \quad \mathbf{p}_c = \mathrm{softmax}\left(g(\mathbf{v}_w, \mathbf{q}_c)\right)$$

$$g(\mathbf{a}, \mathbf{b}) = \tau \cdot \frac{\mathbf{a}^T \mathbf{b}}{||\mathbf{a}||_2 ||\mathbf{b}||_2}$$

Step1: Selecting **meaningful** images from candidates.

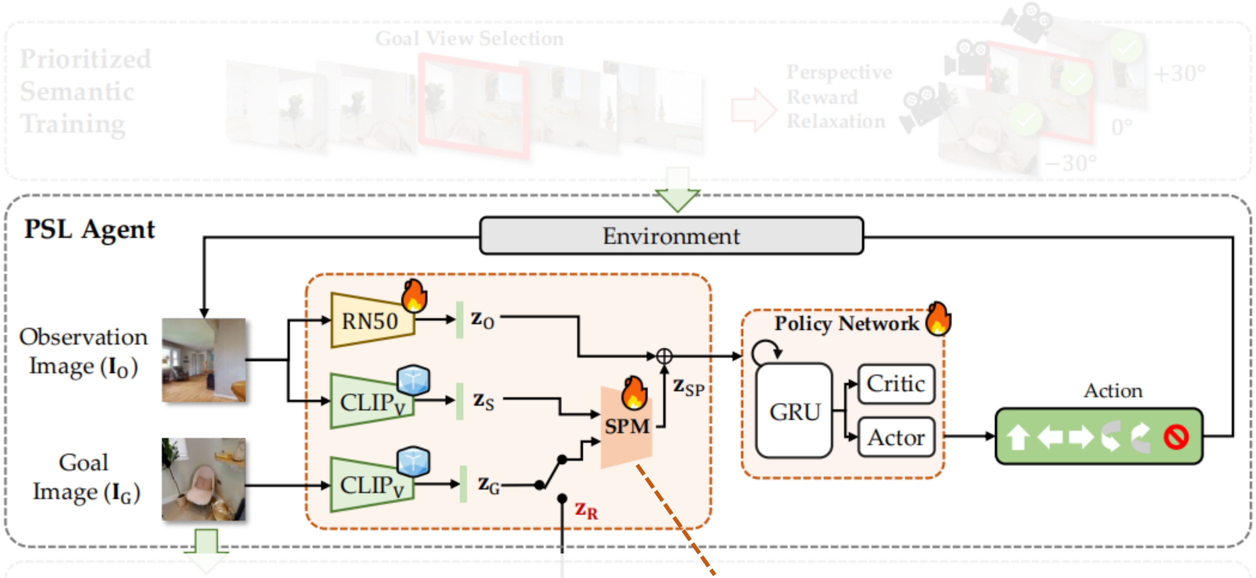
Method: *Perspective Reward Relaxation.*

$$R_t^{\text{PSL}} = \underbrace{\gamma^{suc} * \mathbb{1}\{d_t < \epsilon^d\}}_{\textit{reach the goal location or not}} + \underbrace{\gamma^{suc} * \mathbb{1}\{d_t < \epsilon^d\} * \mathbb{1}\{(\text{extract}_{\mathbf{Y}}(\mathbf{a}_t) < \epsilon^a\}}_{\textit{match the goal view or not}}$$

$$+ \underbrace{r_d(d_t, d_{t-1}) + \mathbb{1}\{d_t < \epsilon^d\} * \text{extract}_{\mathbf{Y}}(r_a(\mathbf{a}_t, \mathbf{a}_{t-1}))}_{\textit{closer to the goal or not}} - \gamma^{delay},$$

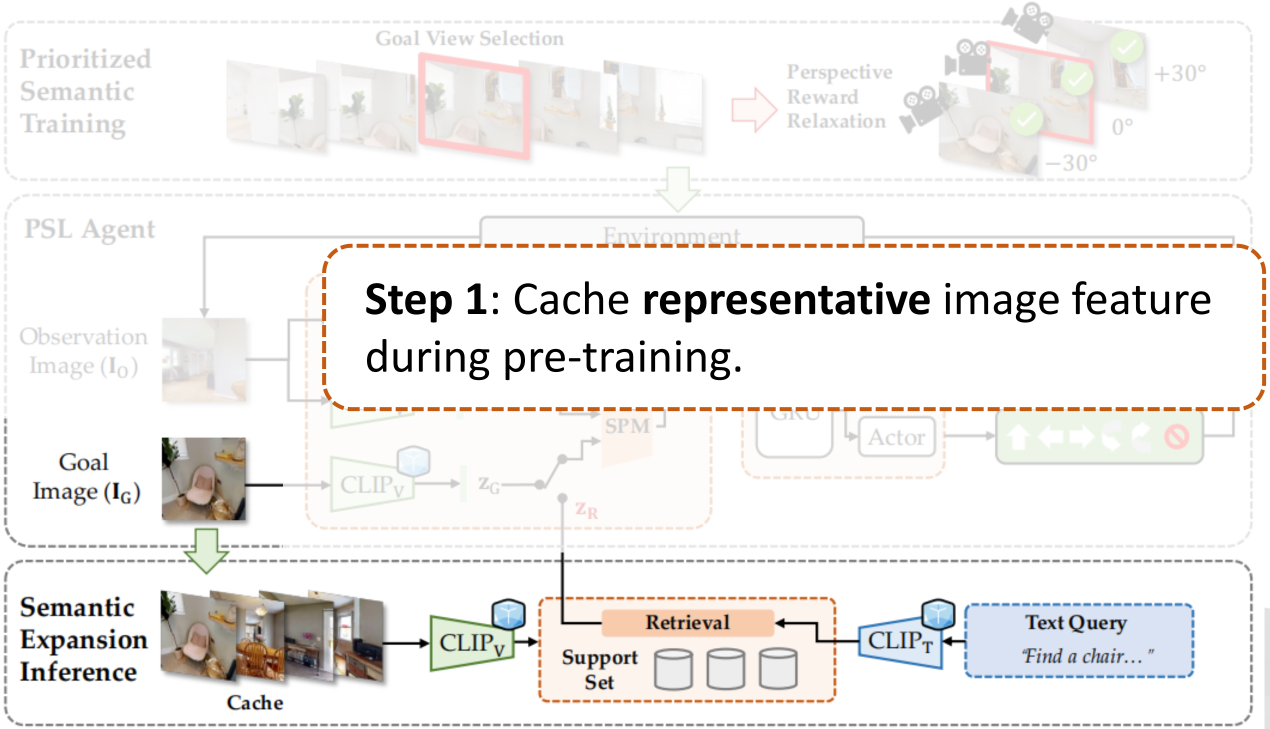
Step2: Relaxing the agent from **pitch heading.**

Method: *Semantic Perception Module.*



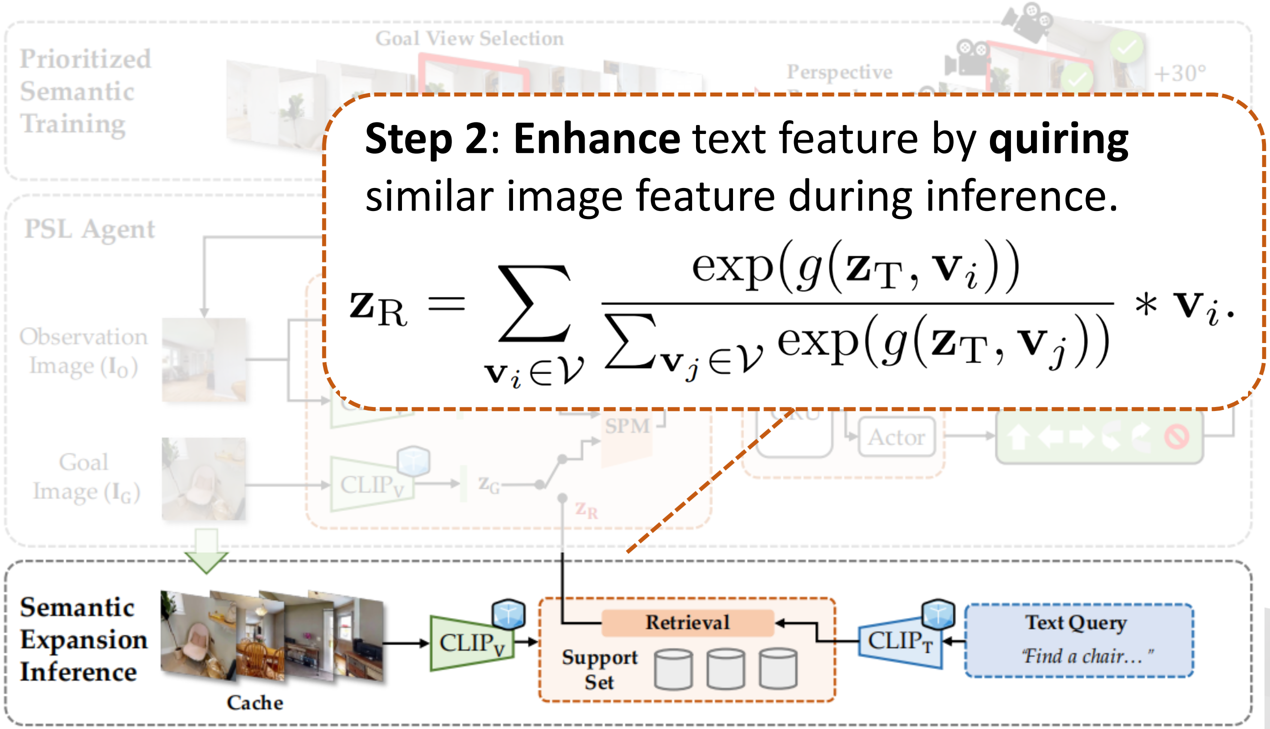
Capture **semantic similarity** between observation and goal images.

Method: *Semantic Expansion Inference Scheme*



Step 1: Cache **representative** image feature during pre-training.

Method: *Semantic Expansion Inference Scheme*



Step 2: **Enhance** text feature by **quiring** similar image feature during inference.

$$\mathbf{z}_{\mathrm{R}} = \sum_{\mathbf{v}_i \in \mathcal{V}} \frac{\exp(g(\mathbf{z}_{\mathrm{T}}, \mathbf{v}_i))}{\sum_{\mathbf{v}_j \in \mathcal{V}} \exp(g(\mathbf{z}_{\mathrm{T}}, \mathbf{v}_j))} * \mathbf{v}_i.$$

Experiments on ObjectNav

Table 1: Comparison with state-of-the-art methods on the ObjectNav task. Our PSL surpasses both LLM-based and Mapping-based methods in terms of Success Rate (SR).

Method	with Mapping	with LLM	LLM	Extra Sensors	SR	SPL
L3MVN [62]	✔	✔	GPT-2	Depth, GPS	35.2	16.5
PixelNav [6]	✘	✔	GPT-4	-	37.9	20.5
ESC [64]	✔	✔	GPT-3.5	Depth, GPS	39.2	22.3
CoW [11]	✔	✘	-	Depth, GPS	6.1	3.9
ProcTHOR [9]	✔	✘	-	Depth, GPS	13.2	7.7
ZSON [38]	✘	✘	-	-	25.5	12.6
PSL (Ours)	✘	✘	-	-	**42.4**	**19.2**

We achieve SOTA on the **ObjectNav** task, even surpassing **LLM-based methods** in SR!

InstanceNav vs. ObjectNav

Specific Destination

"Chair" vs. "The black leather chair."

Complex Instruction

Detailed description with attributes.

Intrinsic + **Extrinsic**

Experiments on InstanceNav

Table 2: Comparison with state-of-the-art methods in the text-goal track of the InstanceNav task. We report the baseline results based on the released code and models. †We perform evaluation with our proposed semantic expansion inference scheme.

Method	Backbone	with LLM	with Mapping	Extra Sensors	SR	SPL
CoW [11]	ViT-Base	✘	✔	Depth, GPS	1.8	1.1
GoW [64]	ViT-Base	✘	✔	Depth, GPS	7.2	4.2
ESC [64]	ViT-Base	✔	✔	Depth, GPS	6.5	3.7
OVRL† [58]	ResNet-50	✘	✘	-	3.7	1.8
ZSON [38]	ResNet-50	✘	✘	-	10.6	4.9
PSL (Ours)	ResNet-50	✘	✘	-	**16.5**	**7.5**

We achieve SOTA on the both tracks of **InstanceNav,** including **Text-goal track** and **Image-goal track.**

Experiments on InstanceNav

Table 3: Comparison with state-of-the-art methods in the image-goal track of the InstanceNav task. In this track, the "Supervised" mark means human labels on the objects are used. †We re-implement OVRL based on released pre-trained weight.

Method	Backbone	Supervised	Pre-training Data	SR	SPL
RL Agent [21]	ResNet-18	✔	-	8.3	3.5
OVRL-V2 [57]	ViT-Base	✔	Gibson	24.8	11.8
OVRL-V2 [57]	ViT-Base	✘	Gibson	0.6	0.2
OVRL† [58]	ResNet-50	✘	HM3D	8.0	4.2
FGPrompt [56]	ResNet-9	✘	HM3D	9.9	2.8
ZSON [38]	ResNet-50	✘	HM3D	14.6	7.3
PSL (Ours)	ResNet-50	✘	HM3D	**23.0**	**11.4**

We achieve SOTA on the both tracks of **InstanceNav,** including **Text-goal track** and **Image-goal track.**

Ablation Studies

Table 4: Ablation studies of different components in our Prioritized Semantic Learning (PSL) agent and Prioritized Semantic Training (PST) strategy under the Image-Goal setting of the `InstanceNav` task. The default entry is marked in gray. "SPM": Semantic Perception Module; "GVS": Goal View Selection; "PRR": Perspective Reward Selection.

	PSL	PST		ZSIN-image		ZSIN-text		ZSON	
	SPM	GVS	PRR	SR	SPL	SR	SPL	SR	SPL
ZSON	✘	✘	✘	12.7	6.5	10.6	6.5	25.5	12.6
PSL (Ours)	✔	✘	✘	19.5	7.9	13.0	5.6	33.7	15.8
	✘	✔	✘	14.8	7.7	11.8	6.1	30.4	14.7
	✔	✔	✘	16.5	6.5	12.3	5.7	35.0	18.1
	✔	✔	✔	**22.0**	**10.7**	**16.5**	**7.5**	**42.4**	**19.2**

All modules are crucial for our PSL agent.

Qualitative Results

Intrinsic Attributes:

The toilet in this image is white, and its seat appears to be yellowing.

Extrinsic Attributes:

In this image, there is a white toilet with a peeling lid and appears to be in a poor condition.

Intrinsic Attributes:

The bed in this image is white.

Extrinsic Attributes:

There are many paintings hanging on the wall around the bed.

Qualitative Results

Intrinsic Attributes:

The chair in this image is made of wood and has a brown color.

Extrinsic Attributes:

There are several chairs in the image, and one of them has a broken arm.

Intrinsic Attributes:

The chair is made of black wood and white leather.

Extrinsic Attributes:

The picture shows a rectangular black wooden dining table with white leather chairs.

THE HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY (GUANGZHOU)

Thank You!