A Data-Centric Revisit of Pre-Trained Vision Models for Robot Learning

by ***Xin Wen***, *Bingchen Zhao, Yilun Chen, Jiangmiao Pang, and Xiaojuan Qi*

CVPR Nashville JUNE 11-15, 2025

香 港 大 學 THE UNIVERSITY OF HONG KONG

THE UNIVERSITY of EDINBURGH

上海人工智能实验室 Shanghai Artificial Intelligence Laboratory

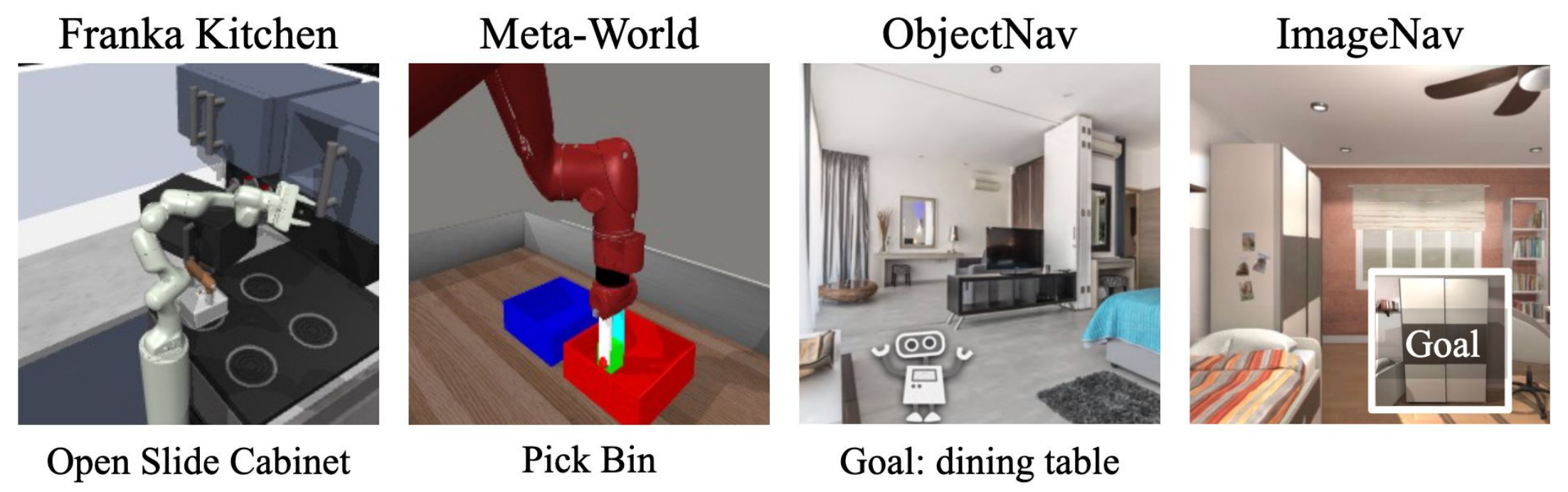
1. How Pre-Training Data Affect Vision Models on Robot Tasks?

PVM: 5 pre-training methods
BEiT, SplitMask, MAE, DINO, and iBOT

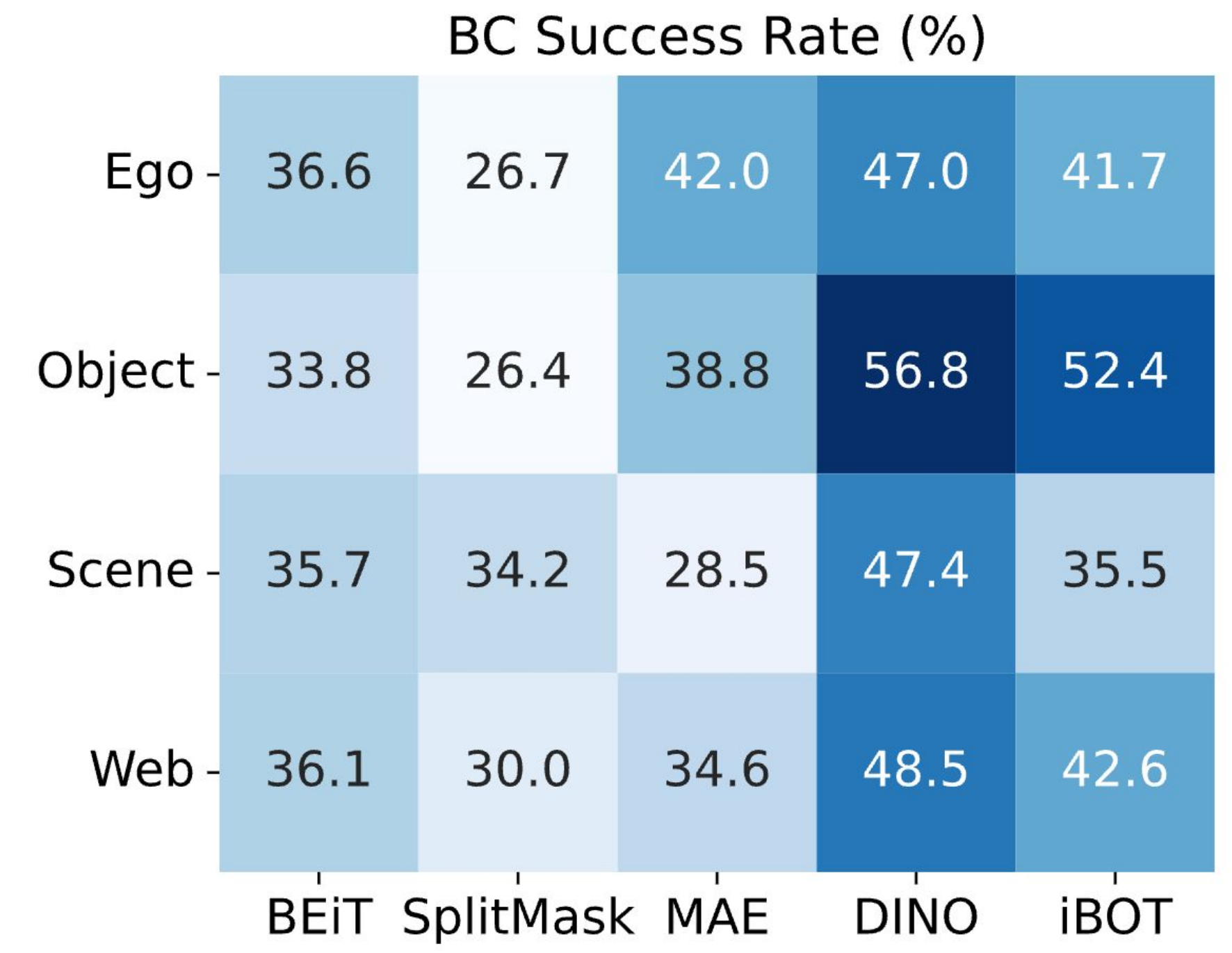
Data: 4 pre-training datasets
Ego-Centric, (Single-)Object-Centric, Scene-Centric, and Web-Crawled data

Eval: 13 behavior cloning tasks
Franka Kitchen and Meta-World

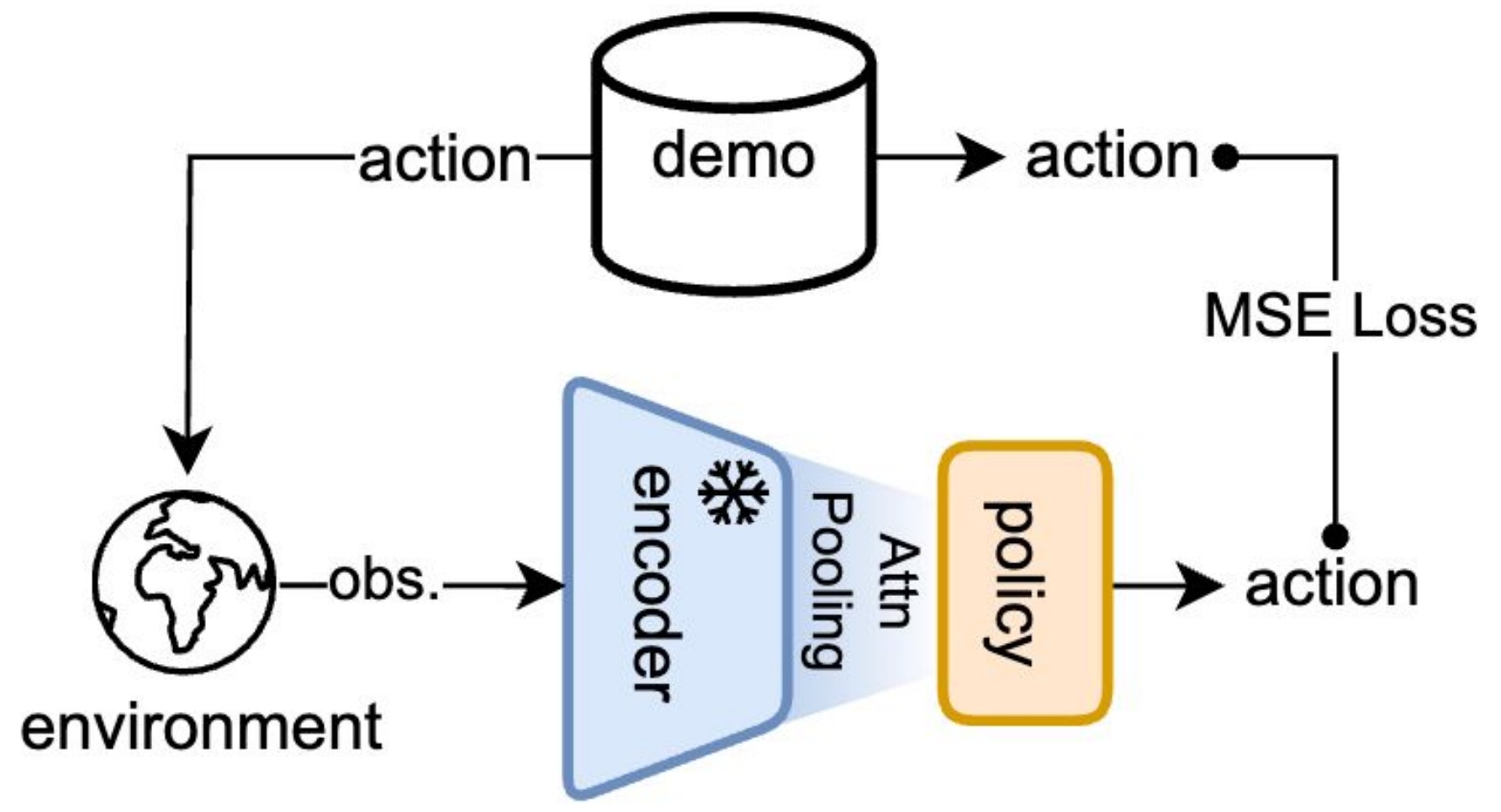
Pre-Training Data



Downstream Robot Learning Tasks



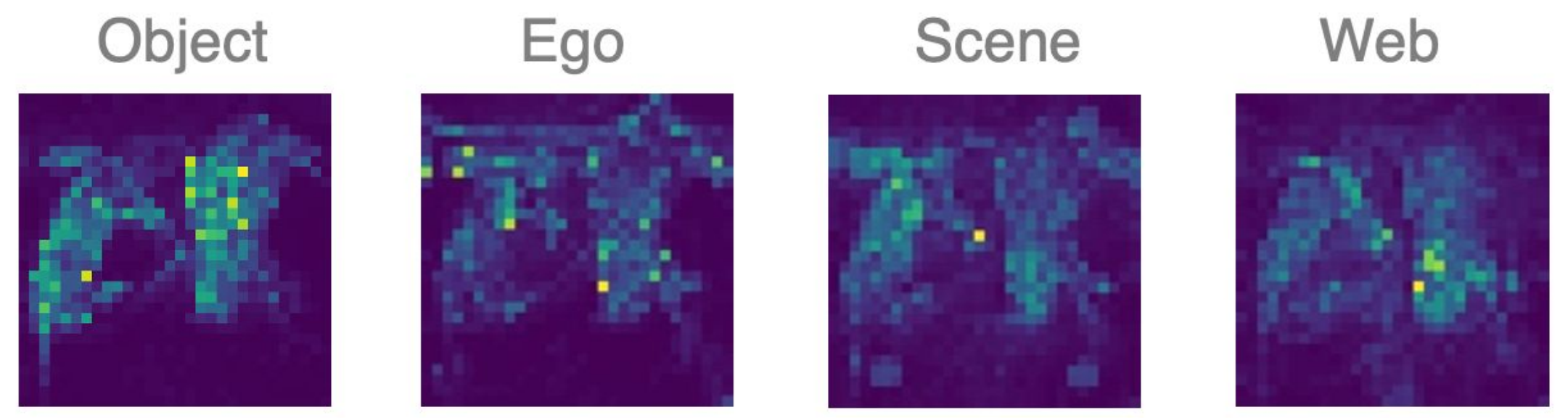
Avg performance on manipulation tasks.
DINO/iBOT rival other methods a lot!

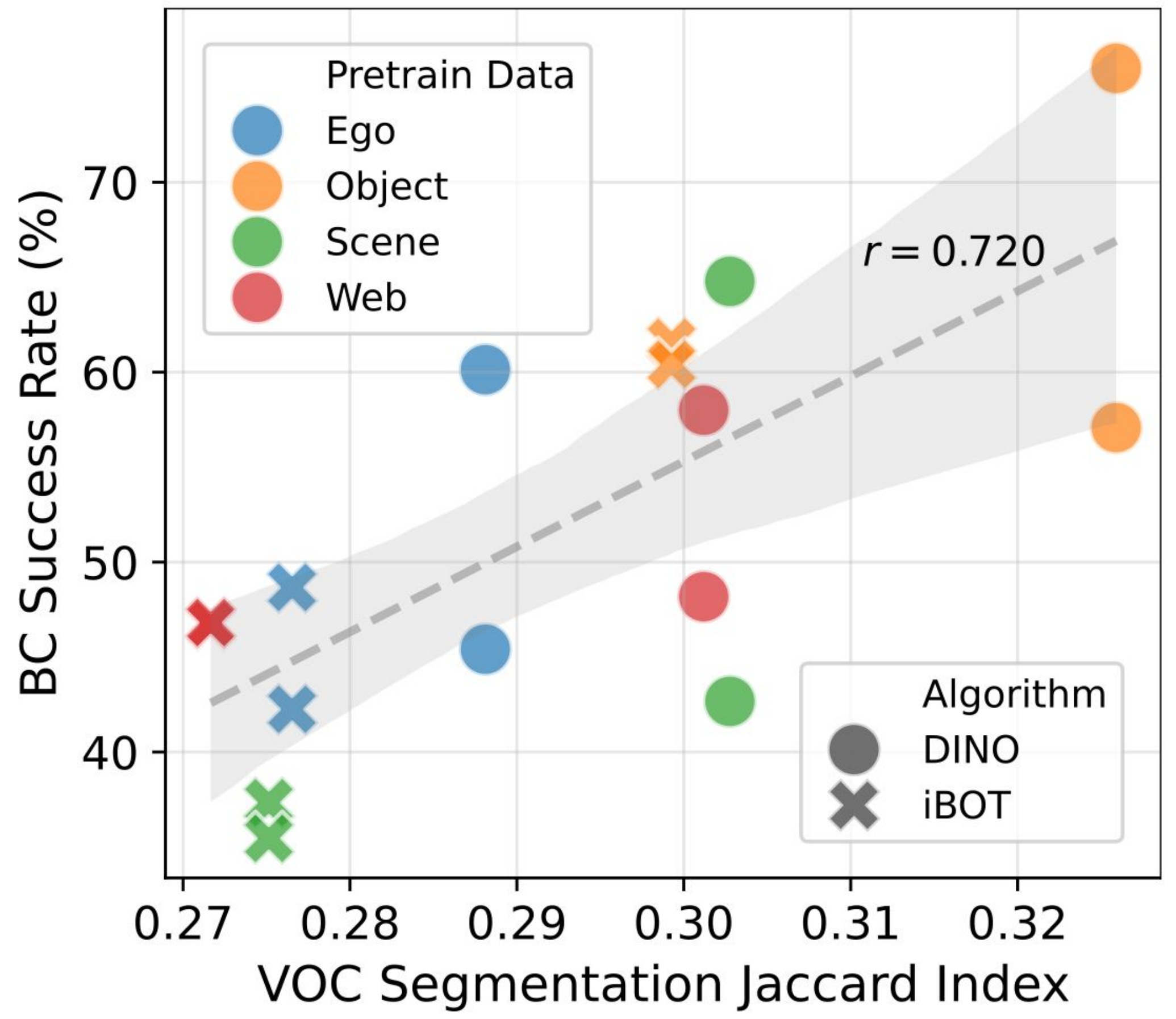


Evaluation Protocol: Behavior cloning with attentive probing on frozen PVMs.
Report success rate on multiple trials.

2. Objectness Matters

But is hard to obtain on non-object-centric (NOC) data

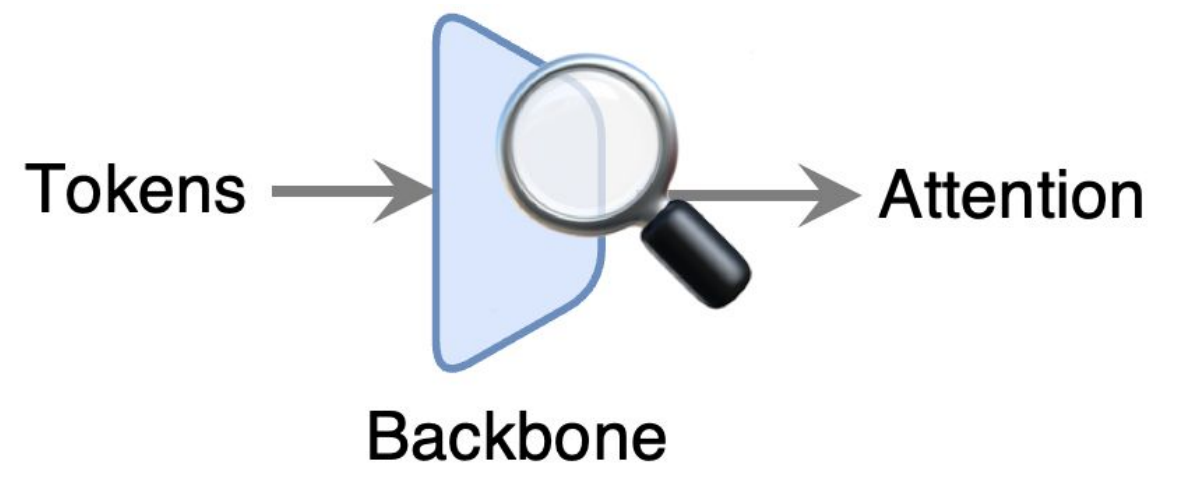




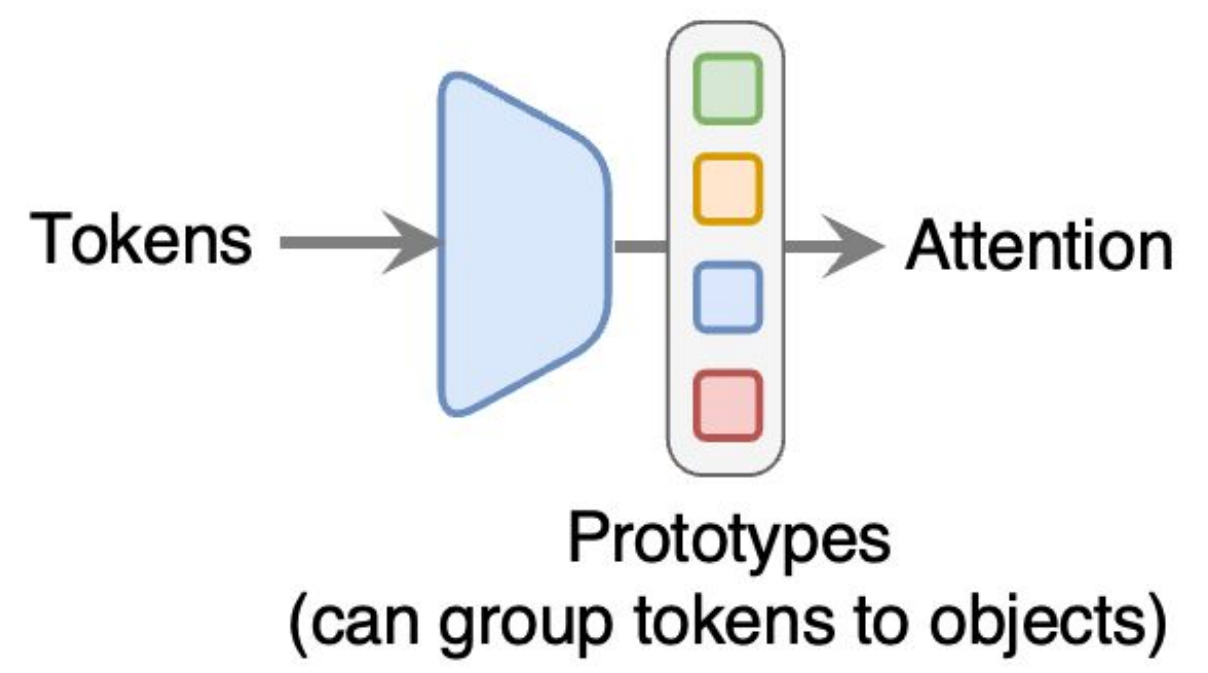
Top: Attention Masks of DINO
Poor objectness on non-object-centric data.
Right: corr. (objectness v.s. performance)

3. Object-Centric Learning on Non-Object-Centric Data

1) Intuition: modeling objectness explicitly

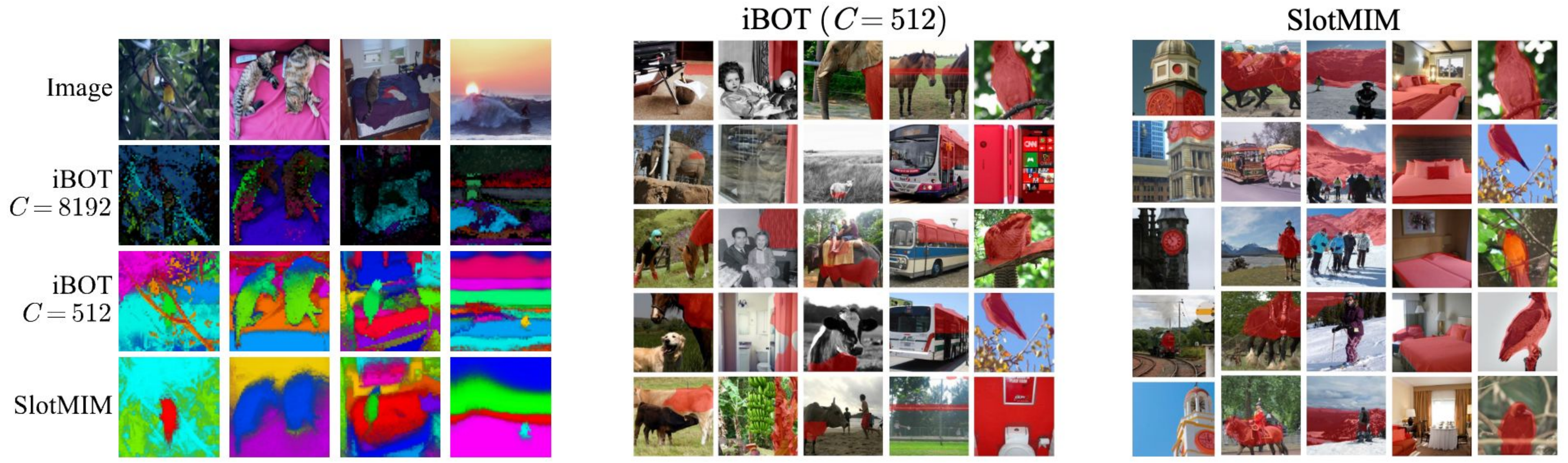


Objectness of DINO emerges internally, which relies on (Single-)OC data bias.



Our method (SlotMIM) receives explicit objectness supervision externally.

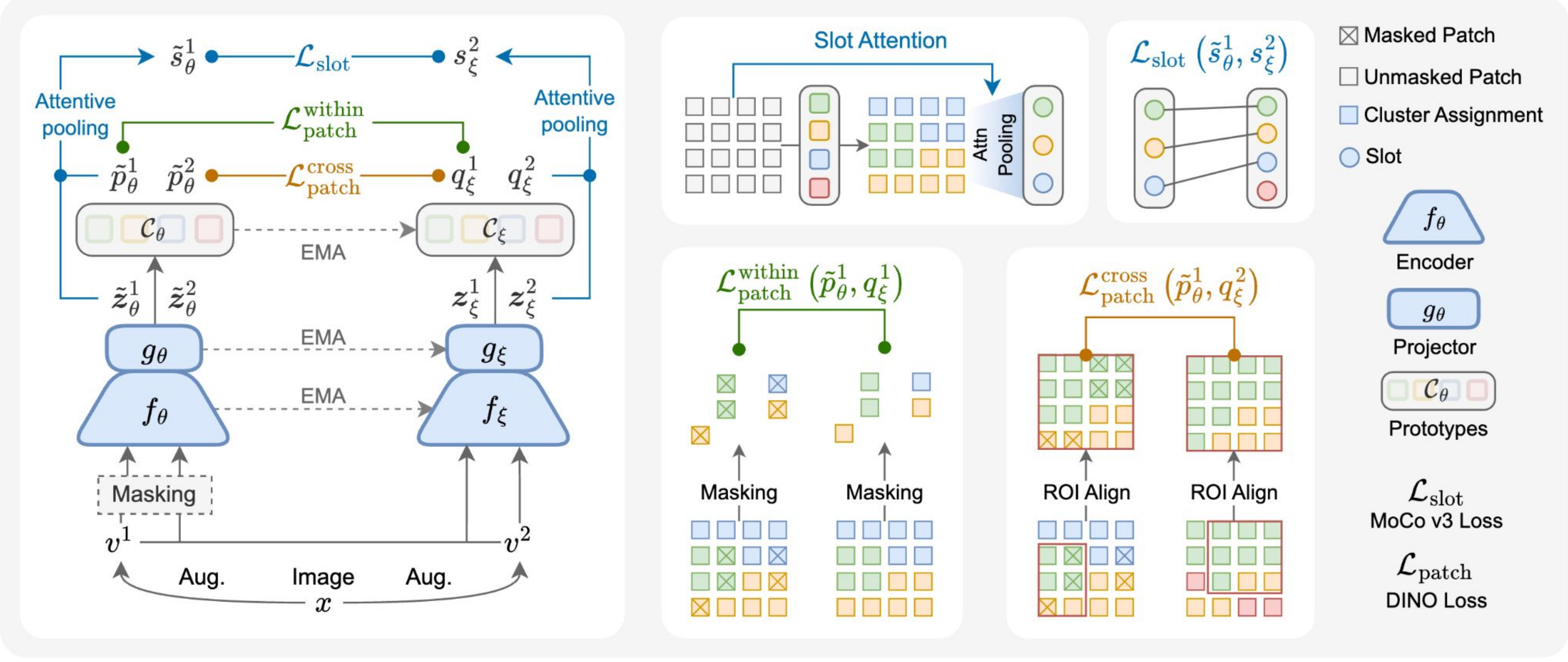
2) Observation: iBOT (DINO on patches) **discovers objects via prototypes even when trained on NOC data, despite semantically misaligned.**



(a) **Clustering assignment of patch tokens.** Each patch is assigned to its nearest-neighbor prototype, with different colors indicating different prototypes.

(b) **Top-5 segments retrieved by the prototypes (by column).** A segment consists of patches assigned to the same prototype within an image. Each column shows the top-5 segments with the highest cosine similarity to the prototype corresponding to the column.

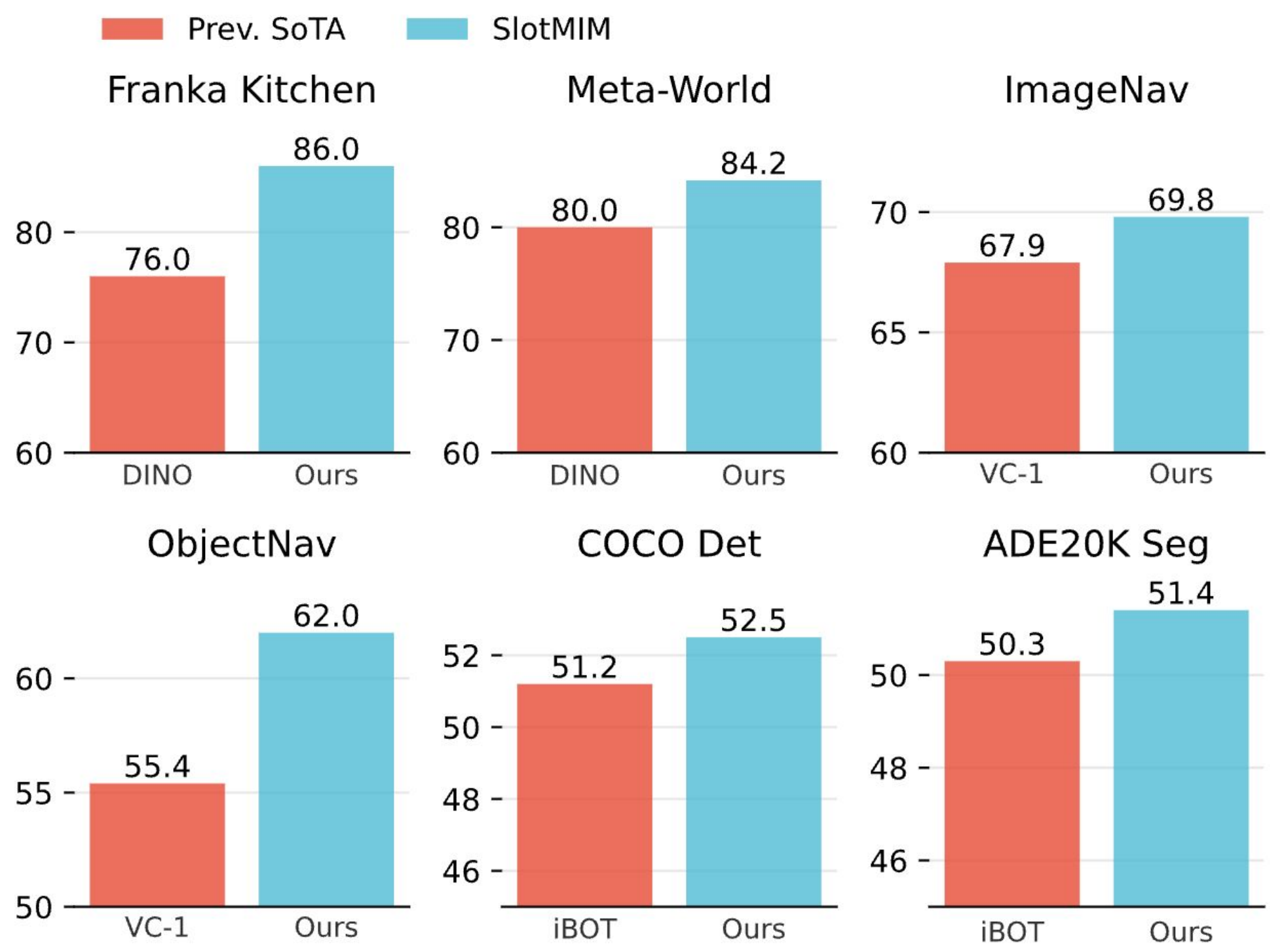
3) Method: DINO loss on image patches within and cross views, and contrastive learning between (grouped) object-level features.


$\tilde{s}^1_\theta$
$\mathcal{L}_{\text{slot}}$
s^2_ξ
Attentive pooling
$\mathcal{L}^{\text{within}}_{\text{patch}}$
Attentive pooling
$\tilde{p}^1_\theta$
$\tilde{p}^2_\theta$
$\mathcal{L}^{\text{cross}}_{\text{patch}}$
q^1_ξ
q^2_ξ
$\mathcal{C}_\theta$
$\mathcal{C}_\xi$
EMA
$\tilde{z}^1_\theta$
$\tilde{z}^2_\theta$
z^1_ξ
z^2_ξ
g_θ
g_ξ
EMA
f_θ
f_ξ
EMA
Masking
v^1
v^2
Aug.
Image
x
Aug.
Slot Attention
Attn Pooling
$\mathcal{L}_{\text{slot}}\left(\tilde{s}^1_\theta, s^2_\xi\right)$
$\mathcal{L}^{\text{within}}_{\text{patch}}\left(\tilde{p}^1_\theta, q^1_\xi\right)$
Masking
Masking
$\mathcal{L}^{\text{cross}}_{\text{patch}}\left(\tilde{p}^1_\theta, q^2_\xi\right)$
ROI Align
ROI Align
☒ Masked Patch
□ Unmasked Patch
□ Cluster Assignment
○ Slot
f_θ Encoder
g_θ Projector
$\mathcal{C}_\theta$ Prototypes
$\mathcal{L}_{\text{slot}}$ MoCo v3 Loss
$\mathcal{L}_{\text{patch}}$ DINO Loss

4. Experiments: SoTA on Manip., Nav., Det., and Seg.

Code & Models

Home Page



Benchmark Suite	RGB	Proprio.	Physics	Action	Goal	Learning
Franka Kitchen [26]	✓	✗	✓	Continu.	–	IL
Meta-World [76]	✓	✗	✓	Continu.	–	IL
ObjectNav [6]	✓	✓	✗	Discrete	Class	IL
ImageNav [81]	✓	✗	✗	Discrete	Image	RL

Tasks: control and perception

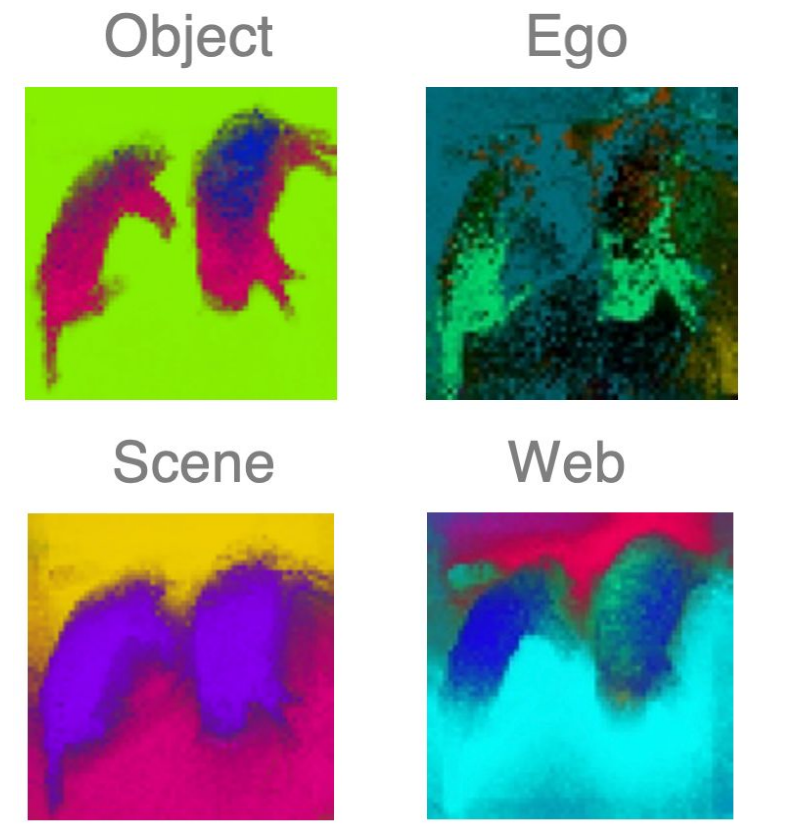
Prev. SoTAs: DINO, iBOT, VC-1

Detailed comparisons available in the paper.

Ablation Study

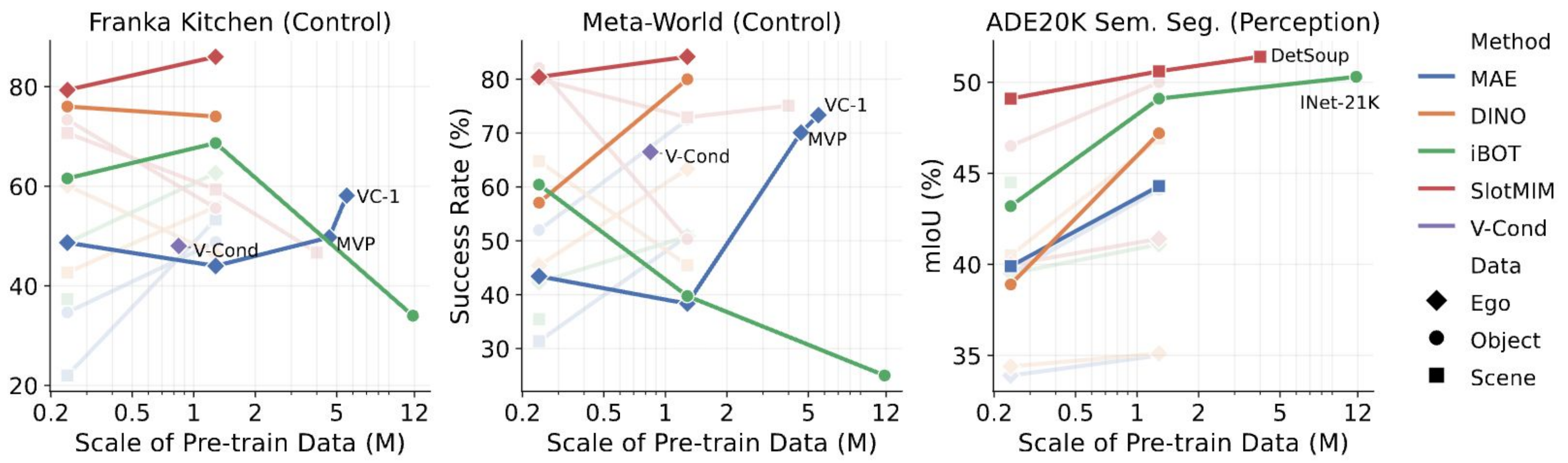
	mask	$\mathcal{L}^{\text{cross}}_{\text{patch}}$	$\mathcal{L}^{\text{within}}_{\text{patch}}$	$\mathcal{L}_{\text{slot}}$	k-NN	ADE	Jacc	$\overline{K}$
1	✗	✓	✗	✗	45.1	47.4	42.5	8.3
2	✓	✓	✗	✗	44.9	48.6	42.3	10.3
3	✓	✗	✓	✗	27.7	45.7	39.3	20.7
4	✓	✓	✗	✓	45.3	47.5	42.9	8.4
5	✓	✓	✓	✓	**46.2**	**49.1**	**43.9**	9.4

Qualitative Results

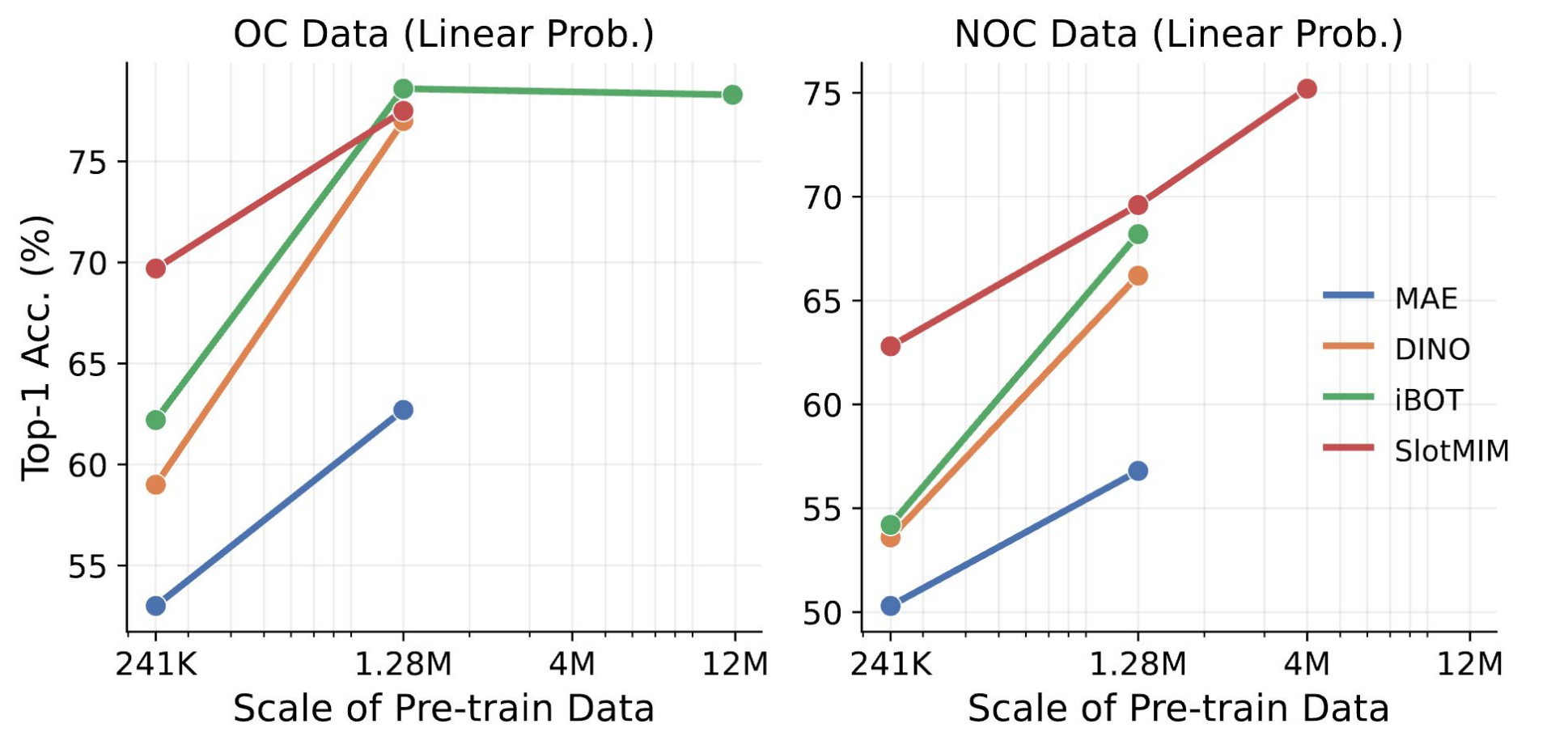


SlotMIM learns objective-ness adaptively.

Scaling Curves (Ctrl. & Percept.) highlighting best (model, data)



Scaling Curves (ImageNet L.P.)



Better scaling on NOC data means 1) less dependence on data curation, 2) better scalability, and 3) better data efficiency.

Future investigations on scalability could be valuable.