ICLR 2025

International Conference On Learning Representations

Multiview Equivariance Improves 3D Correspondence Understanding with Minimal Feature Finetuning

Yang You[1], Yixin Li[1], Congyue Deng[1], Yue Wang[2], Leonidas Guibas[1]

[1]Stanford University [2]University of Southern California

Presenter: Yang You

SotA VLM is (pre)trained on 2D image

To what extent do ViT models possess an inherent awareness of 3D structures?

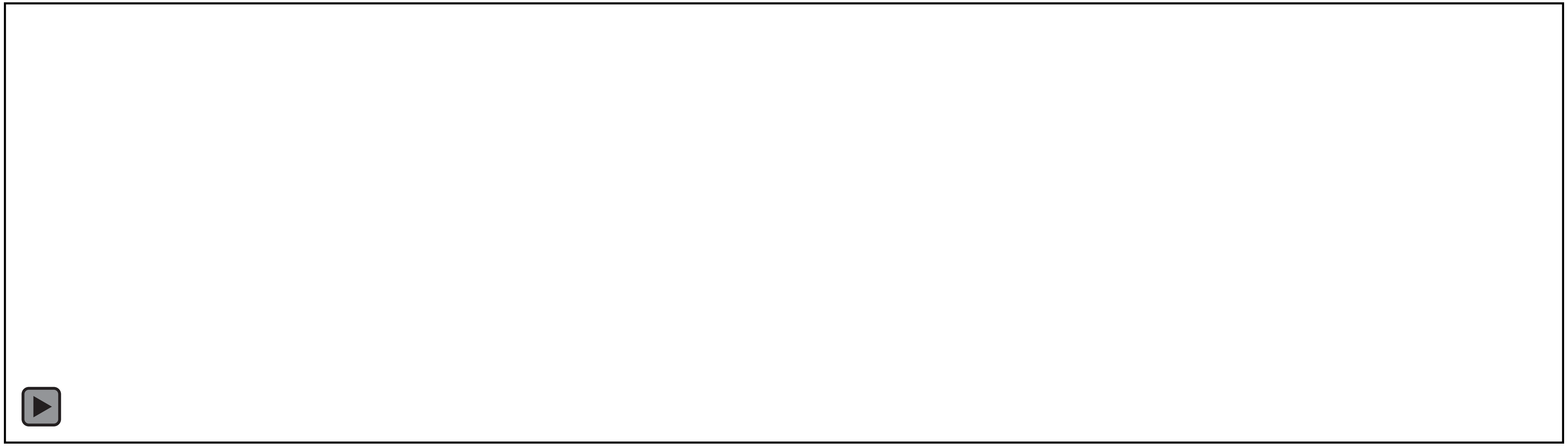
How does this awareness impact their performance on image-based 3D vision tasks?

Can we further enhance the 3D awareness of these vision foundation models?

To what extent do ViT models possess an inherent awareness of 3D structures?

How much does ViT know about 3D?

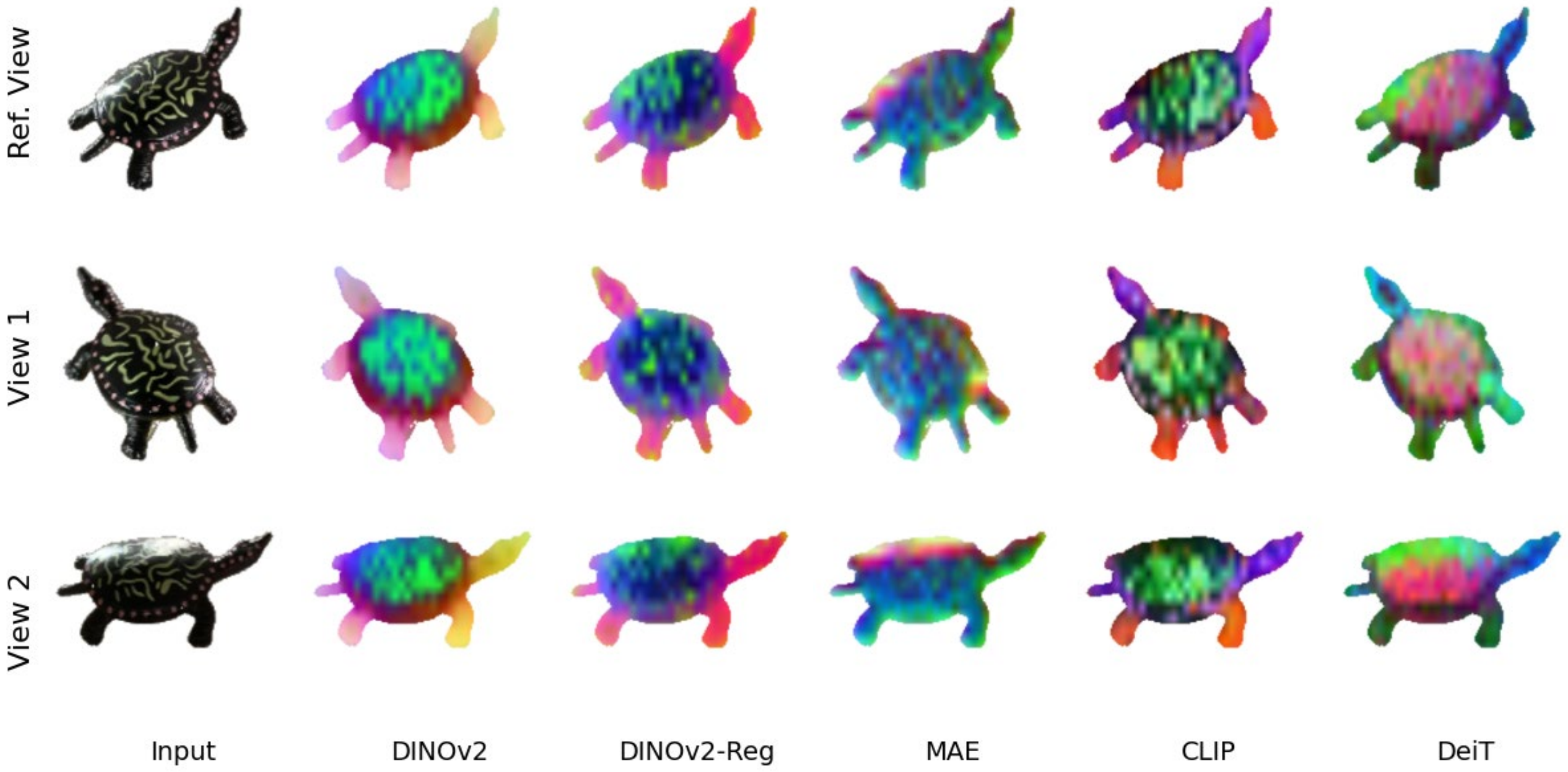
- 3D correspondence is one important aspect



Oquab, M., Darcet, T., Moutakanni, T., Vo, H., Szafraniec, M., Khalidov, V., ... & Bojanowski, P. (2023). Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*.

A comparison on ViT features

- DINOv2 is good



A Comparison on ViT features

- Quantitative results
  - Again, DINOv2 is the best

Model	PCDP(%)			APE(%)↓
	0.05↑	0.1↑	0.2↑	
DINOv2	22.60	36.84	**58.88**	**19.12**
DINOv2-Reg	**23.05**	**37.24**	58.23	19.51
MAE	16.25	30.71	55.46	20.58
CLIP	17.05	33.00	57.17	20.11
DeiT	18.07	33.89	58.05	19.72

Results on Objaverse

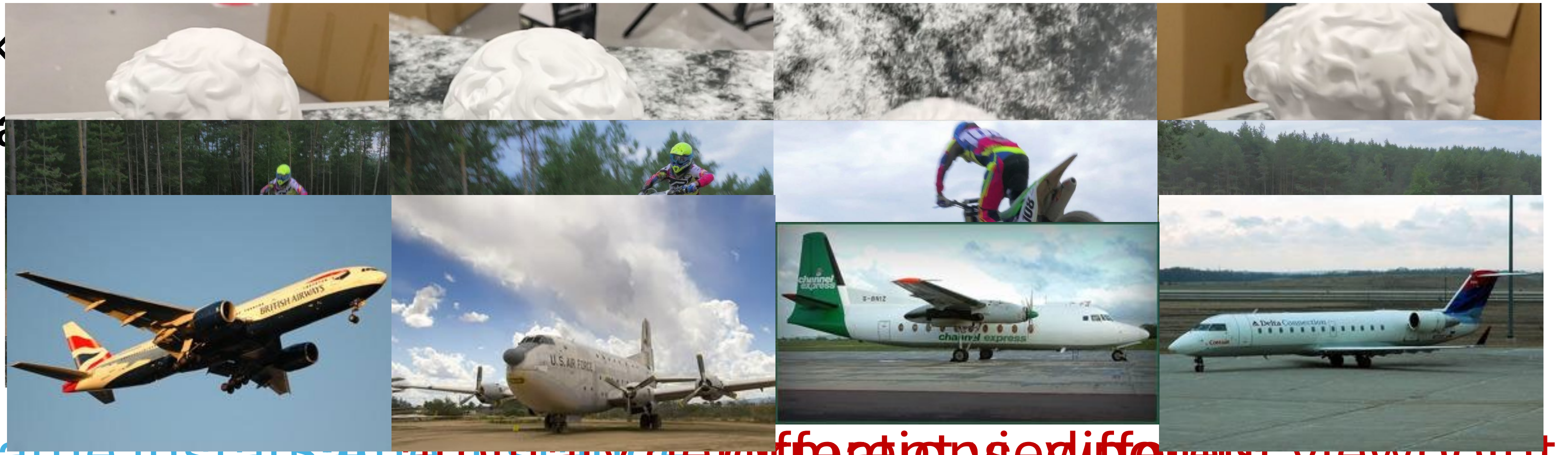
Model	PCDP(%)			APE(%)↓
	0.05↑	0.1↑	0.2↑	
DINOv2	62.09	77.94	**92.49**	6.24
DINOv2-Reg	**64.54**	**78.99**	92.25	**6.06**
MAE	59.10	75.82	91.42	6.73
CLIP	46.63	63.49	80.53	11.34
DeiT	54.63	72.36	87.64	8.34

Results on MVImgNet

How does this awareness impact their performance on image-based 3D vision tasks?

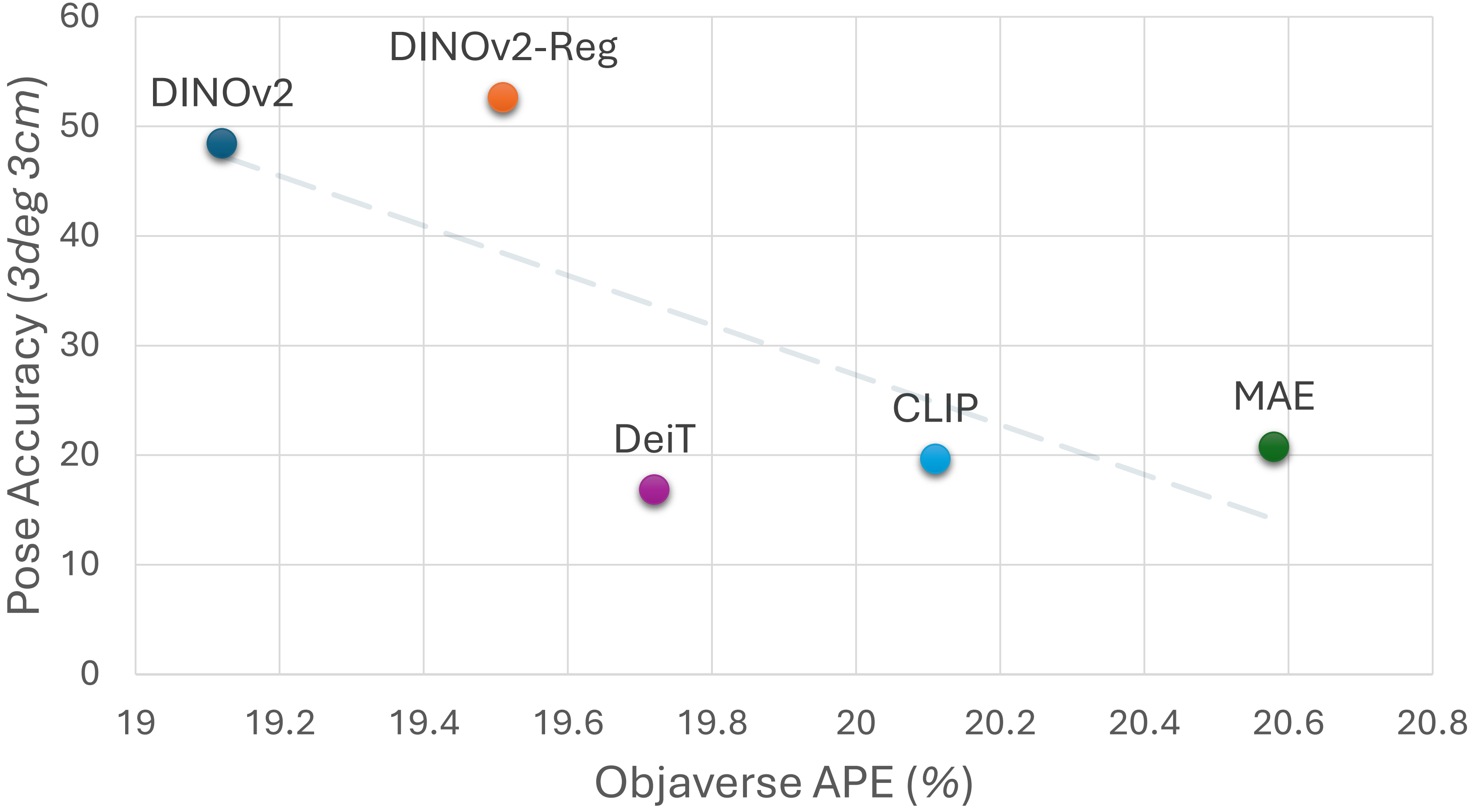
Why do we study 3D correspondence?

- 3D correspondence itself is not interesting but there are interesting tasks that leverage 3D correspondence
  - 3D pose estimation
  - Track
  - Sema

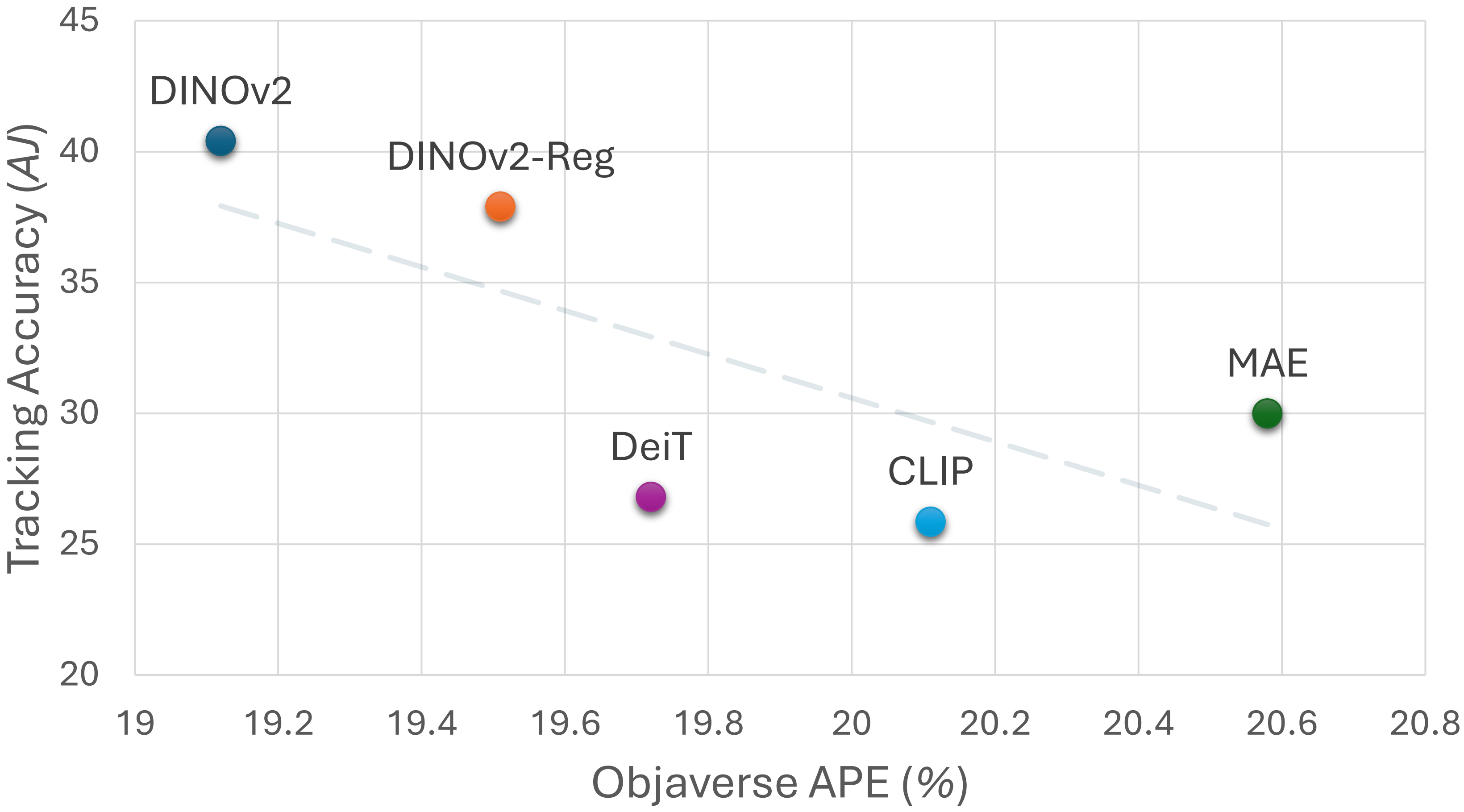


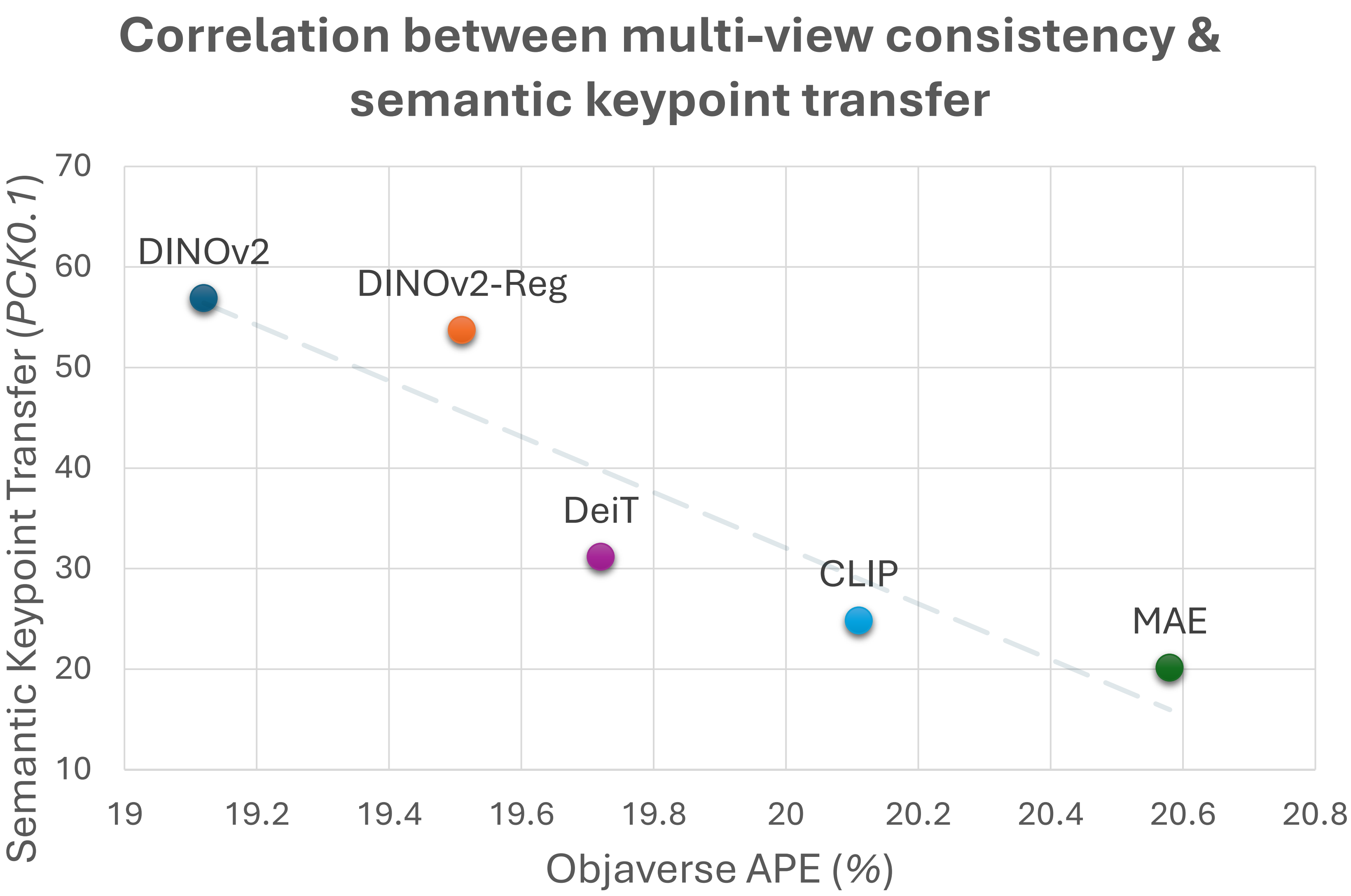
Different instances, different viewpoints

Correlation between multi-view consistency & pose estimation



Correlation between multi-view consistency & tracking




Correlation between multi-view consistency & semantic keypoint transfer
Semantic Keypoint Transfer (PCK0.1)
Objaverse APE (%)
70
60
50
40
30
20
10
19
19.2
19.4
19.6
19.8
20
20.2
20.4
20.6
20.8
DINOv2
DINOv2-Reg
DeiT
CLIP
MAE

There is an obvious correlation between multi-view consistency and downstream tasks.

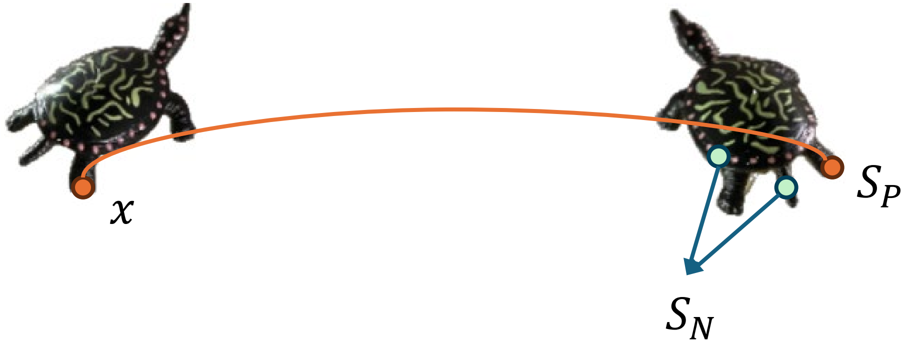
Can we further enhance the 3D awareness of these vision foundation models?

And thus, benefit downstream 3D applications

A simple but effective method

- LoRA finetuning with SmoothAP -- A loss function optimized for retrieval

$$SmoothAP = \frac{1}{S_P} \sum_{i \in S_P} \frac{1 + \sum_{j \in S_P} \sigma(D_{ij})}{1 + \sum_{j \in S_P} \sigma(D_{ij}) + \sum_{j \in S_N} \sigma(D_{ij})} \quad \text{where} \quad D_{ij} = f_j \cdot f_{x_1} - f_i \cdot f_{x_1}$$

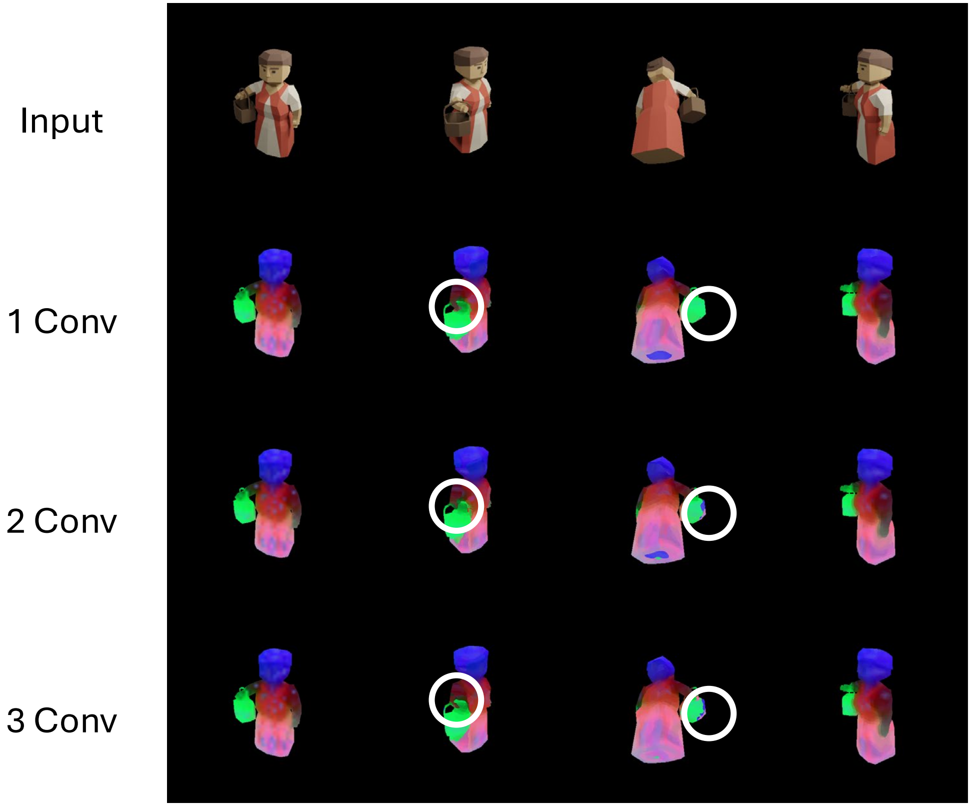


View 1 View 2

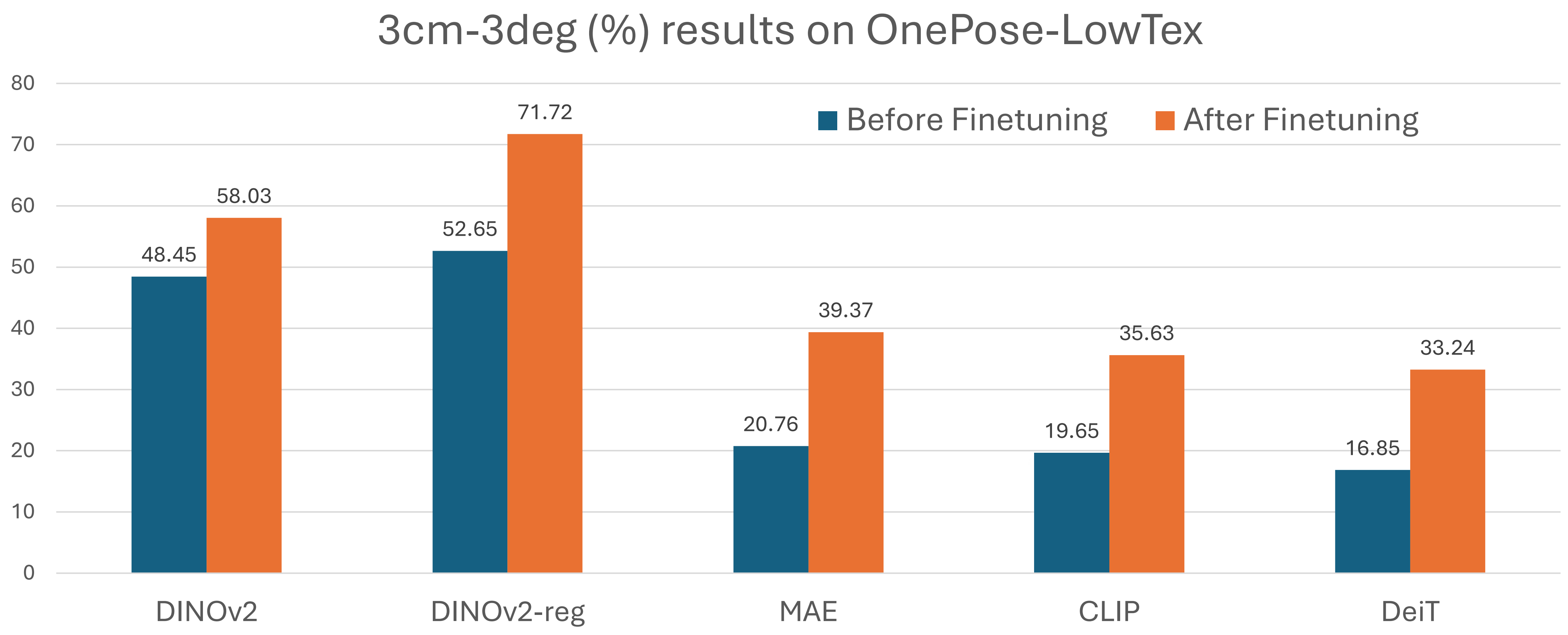
Brown, Andrew, Weidi Xie, Vicky Kalogeiton, and Andrew Zisserman. "Smooth-ap: Smoothing the path towards large-scale image retrieval." In European conference on computer vision, pp. 677-694. Cham: Springer International Publishing, 2020.

A simple but effective method

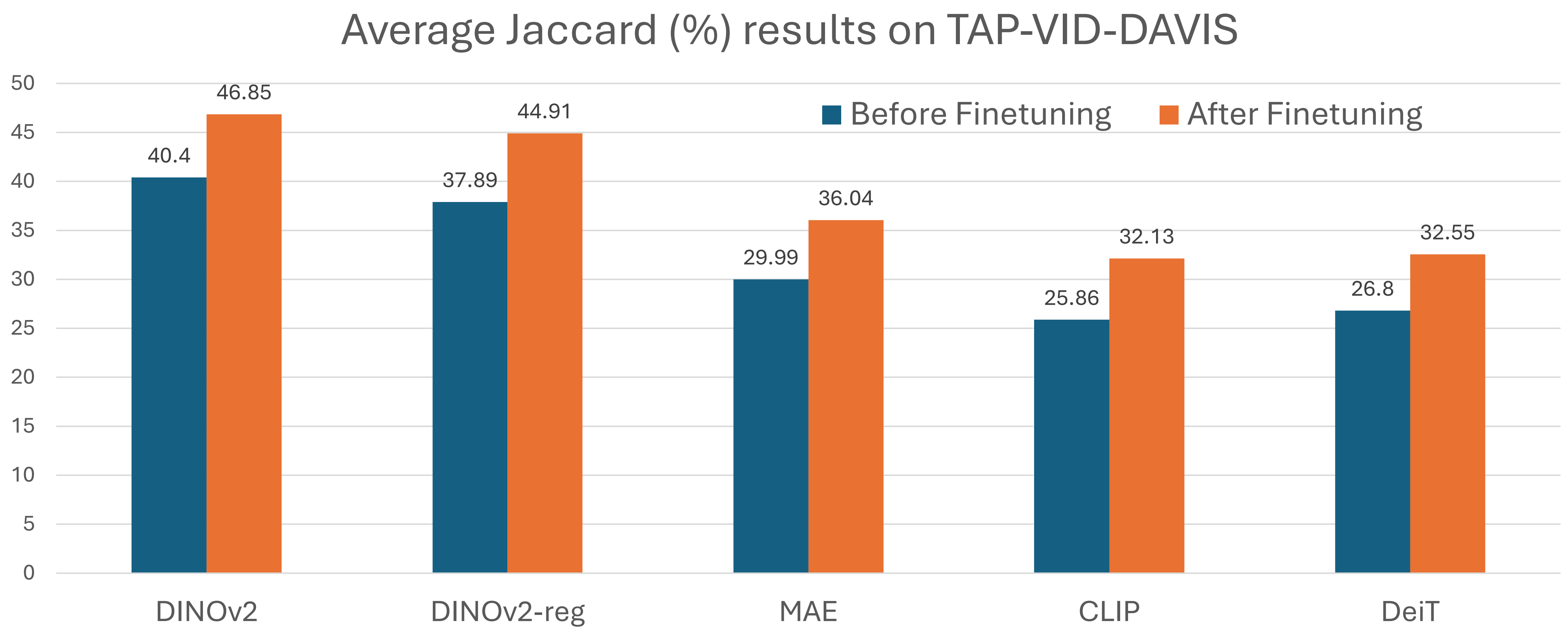
- Adding one convolution layer is helpful, but not more



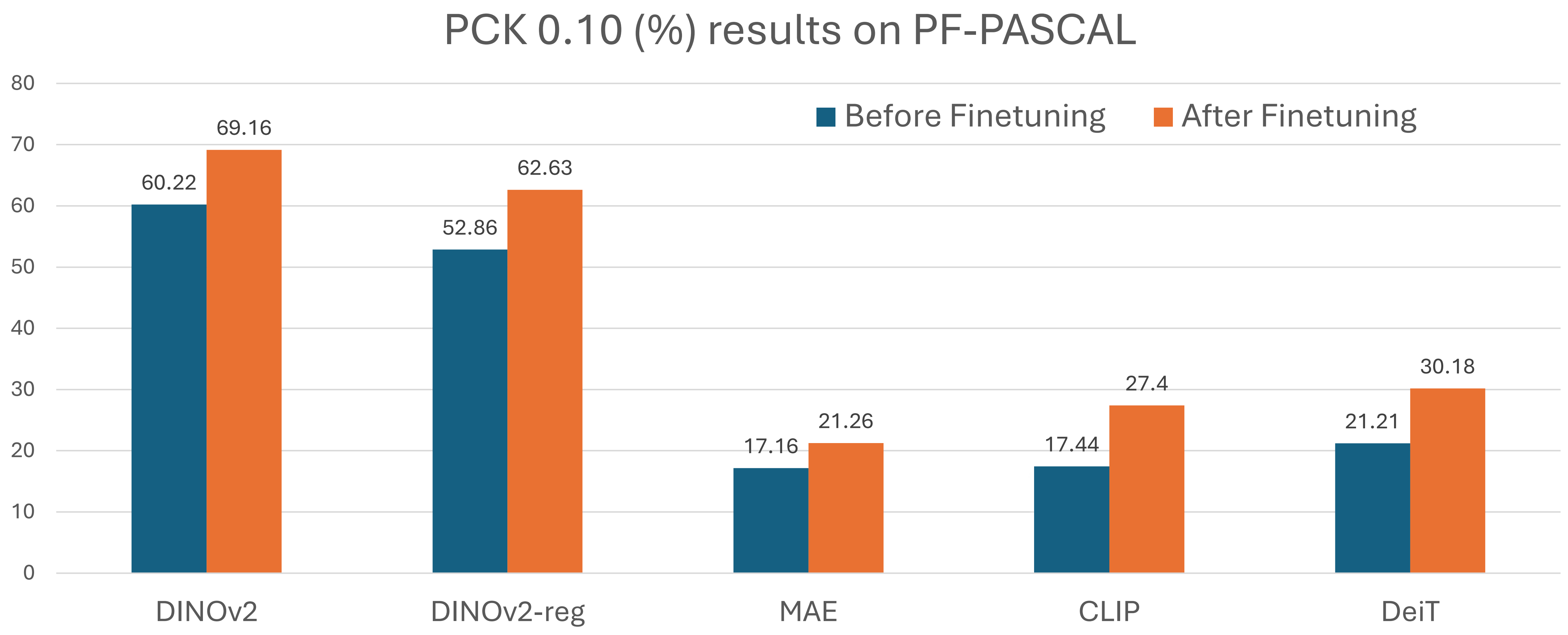
Finetuning improving 3D pose estimation



Finetuning improving tracking

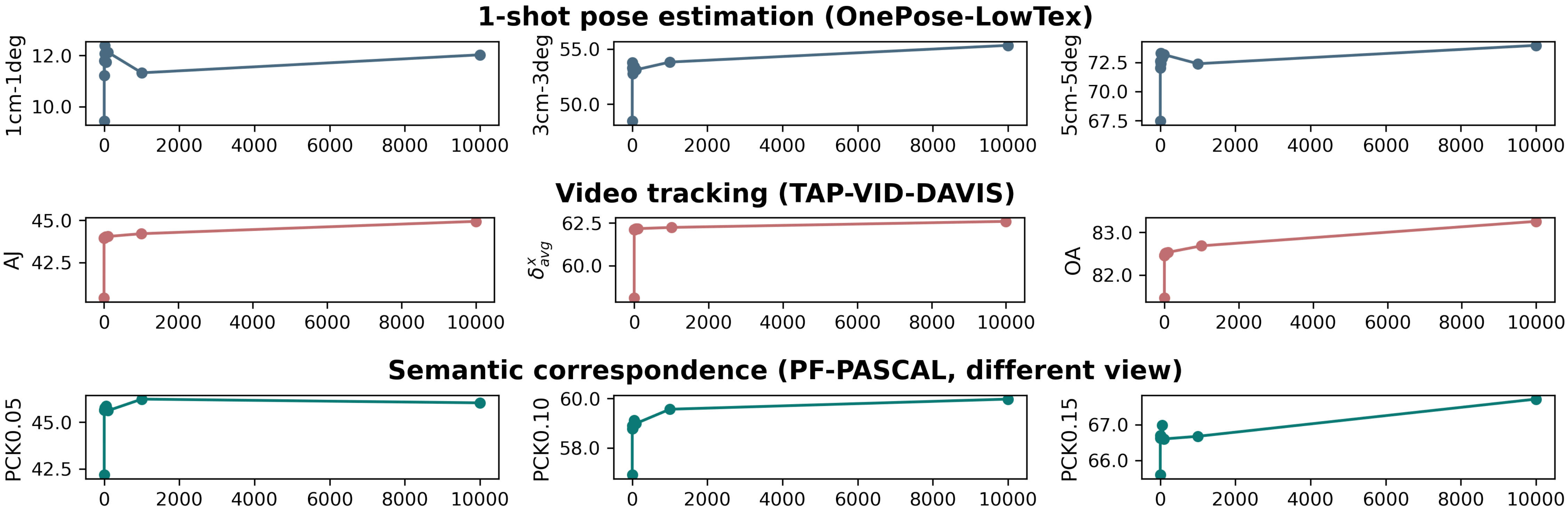


Finetuning improving semantic transfer



One object, one iteration is enough

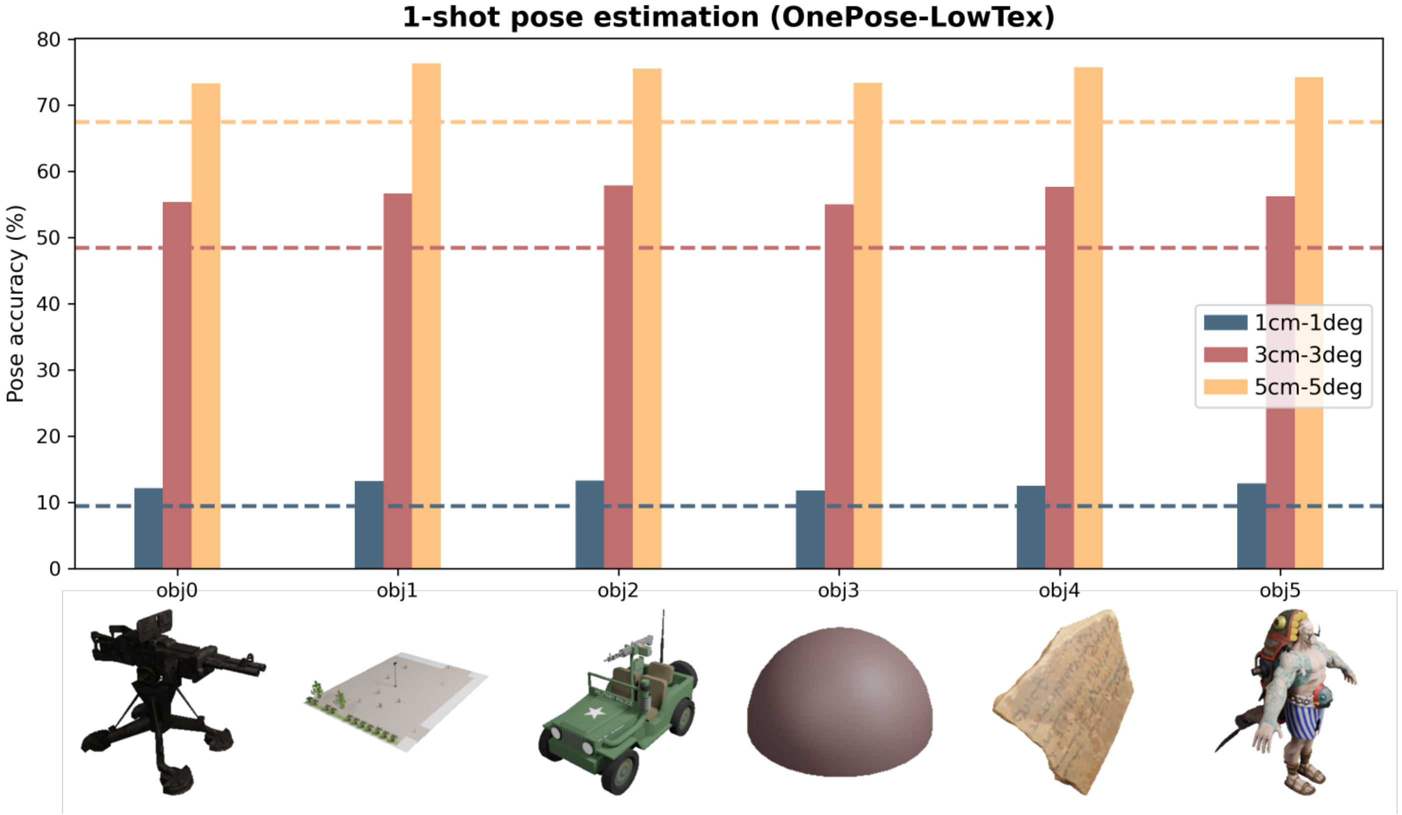
- Interestingly, only one iteration on one object can boost a lot



Performance v.s. # training iterations, on one object

Agnostic to the specific object choice

- Finetuning on different objects gives similar performance



Thanks for listening!

Try our demo below!