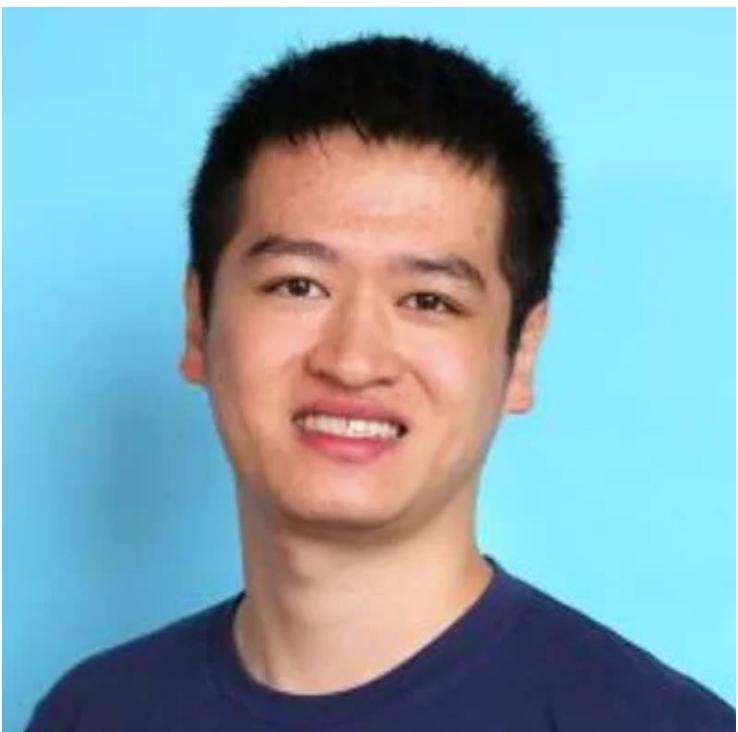


# Pose-Aware Self-Supervised Learning with Viewpoint Trajectory Regularization

Jiayun Wang



On job market!  
[peterw@caltech.edu](mailto:peterw@caltech.edu)

Yubei Chen



Stella X. Yu





NM2255 HD



- Car heading **towards** the camera
- **At danger!**



NM2255 HD



- Car heading **towards** the camera
- **At danger!**

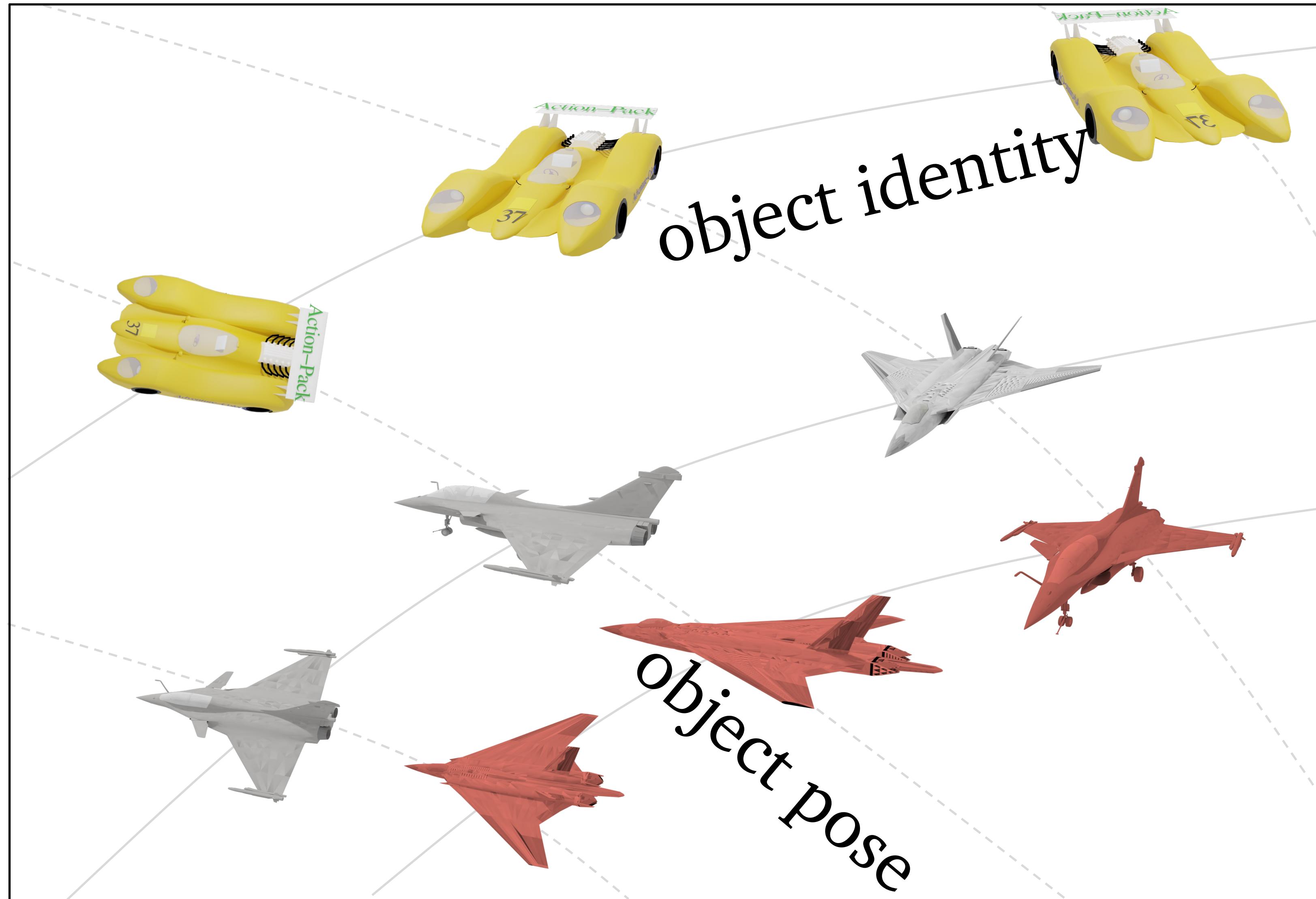


- Car heading **away** from the camera
- **No danger**

Recognition needs to understand both aspects:

- *What* is the object
- *How* is it presented

# Can we learn disentangled semantic-pose representation?



# Settings for disentangled representation learning?



**Scenario:** a robot moves around in the environment



A natural **data acquisition** scheme:

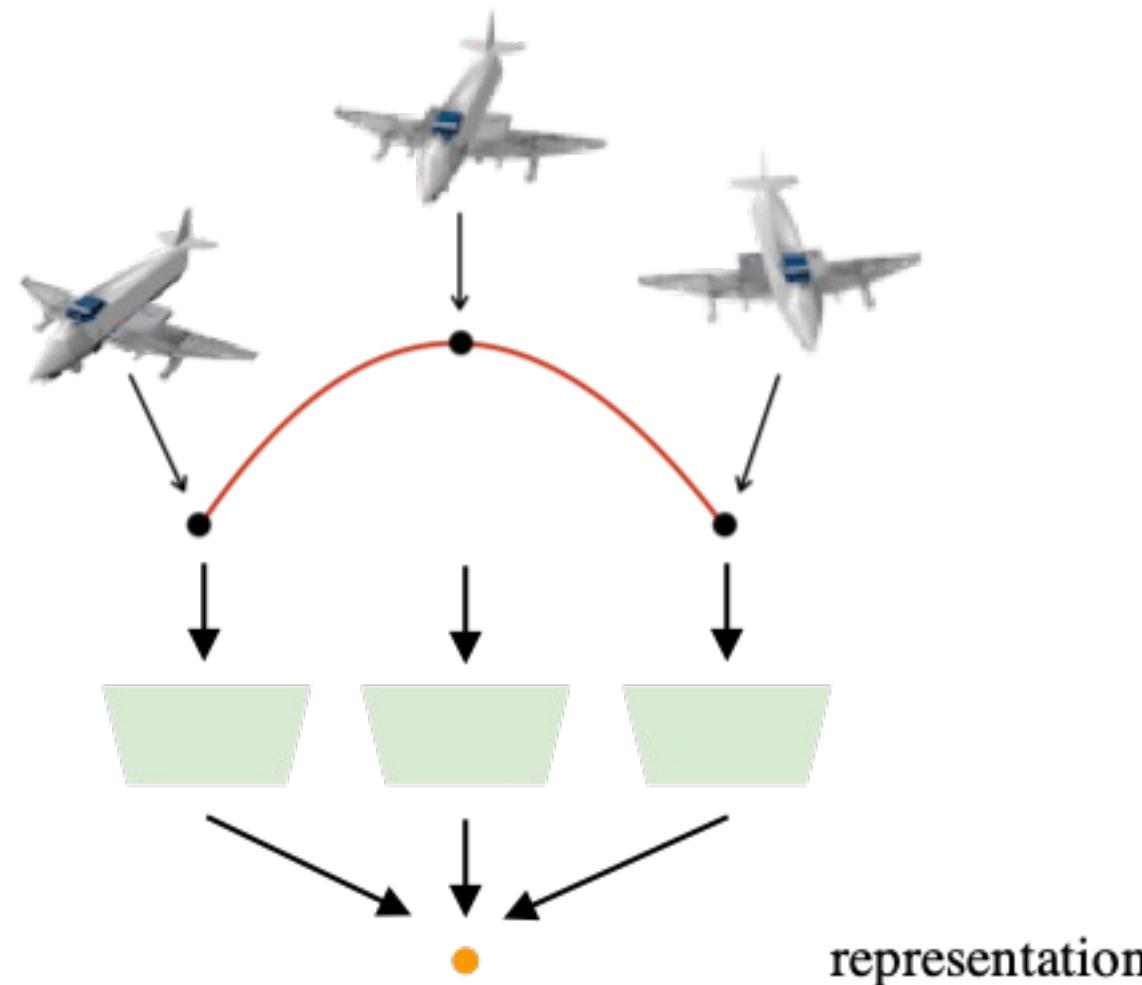
- No labels → Self-supervised learning (SSL)
- Adjacent images of the same object from a smooth **viewpoint trajectory**

# Existing SSL vs Ours

## Existing SSL

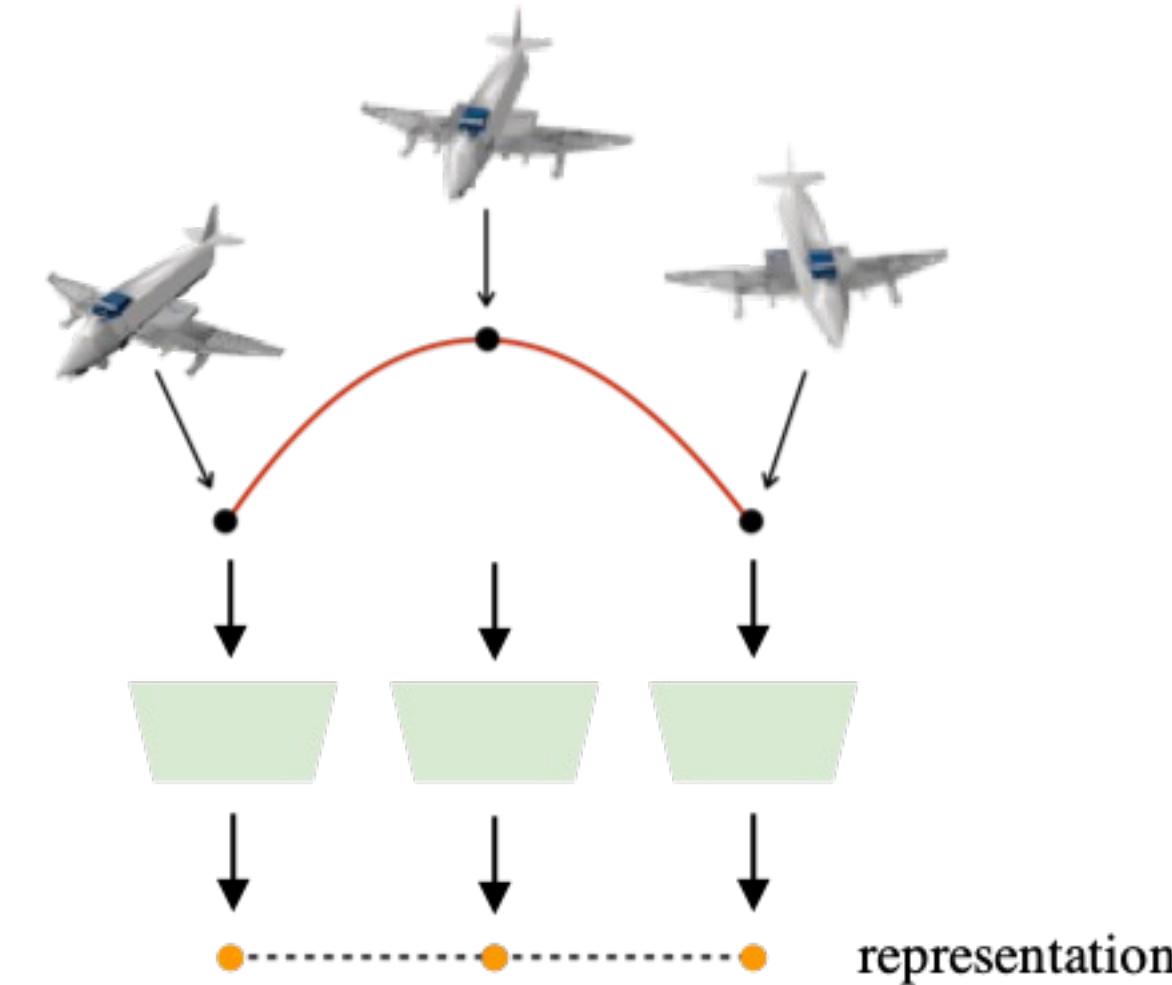
VICReg, SimCLR, SimSiam, MoCo,...

- Invariant representation
- Object identity only



## Ours

- Equivariant representation
- Object identity + pose



# Existing SSL vs Ours

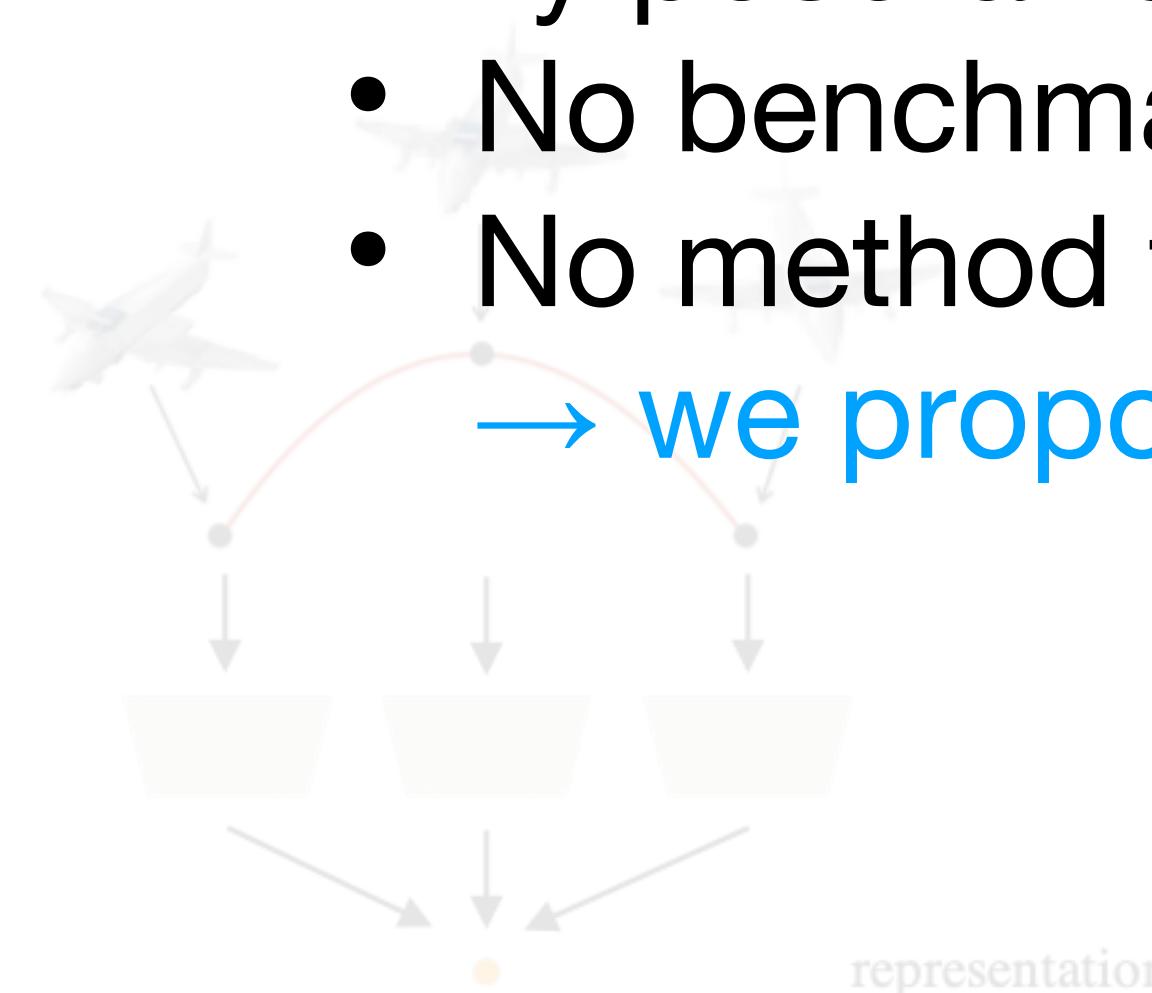
## Existing SSL

VICReg, SimCLR, SimSiam, MoCo,...

- Invariant representation
- Object identity only

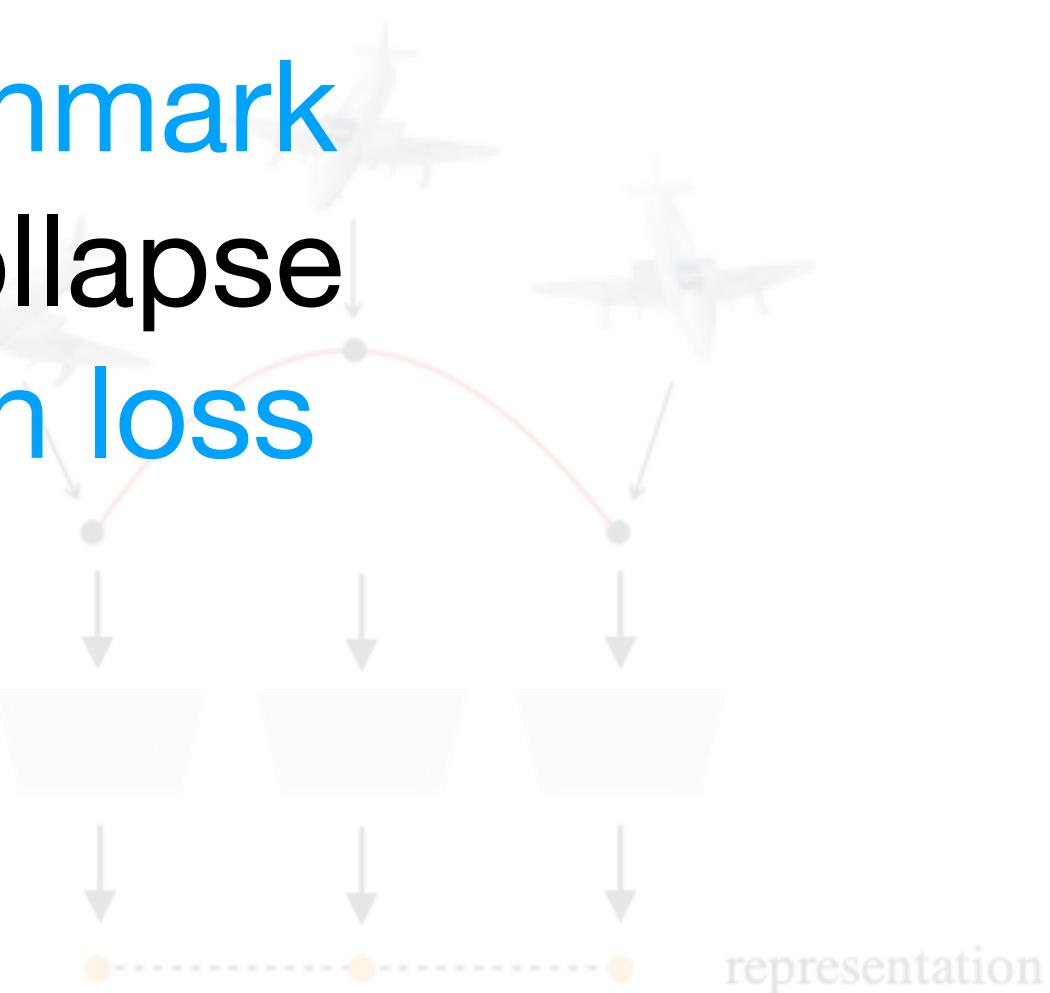
Why pose-aware SSL is hard?

- No benchmark → **we propose a benchmark**
- No method to avoid representation collapse  
→ **we propose trajectory regularization loss**



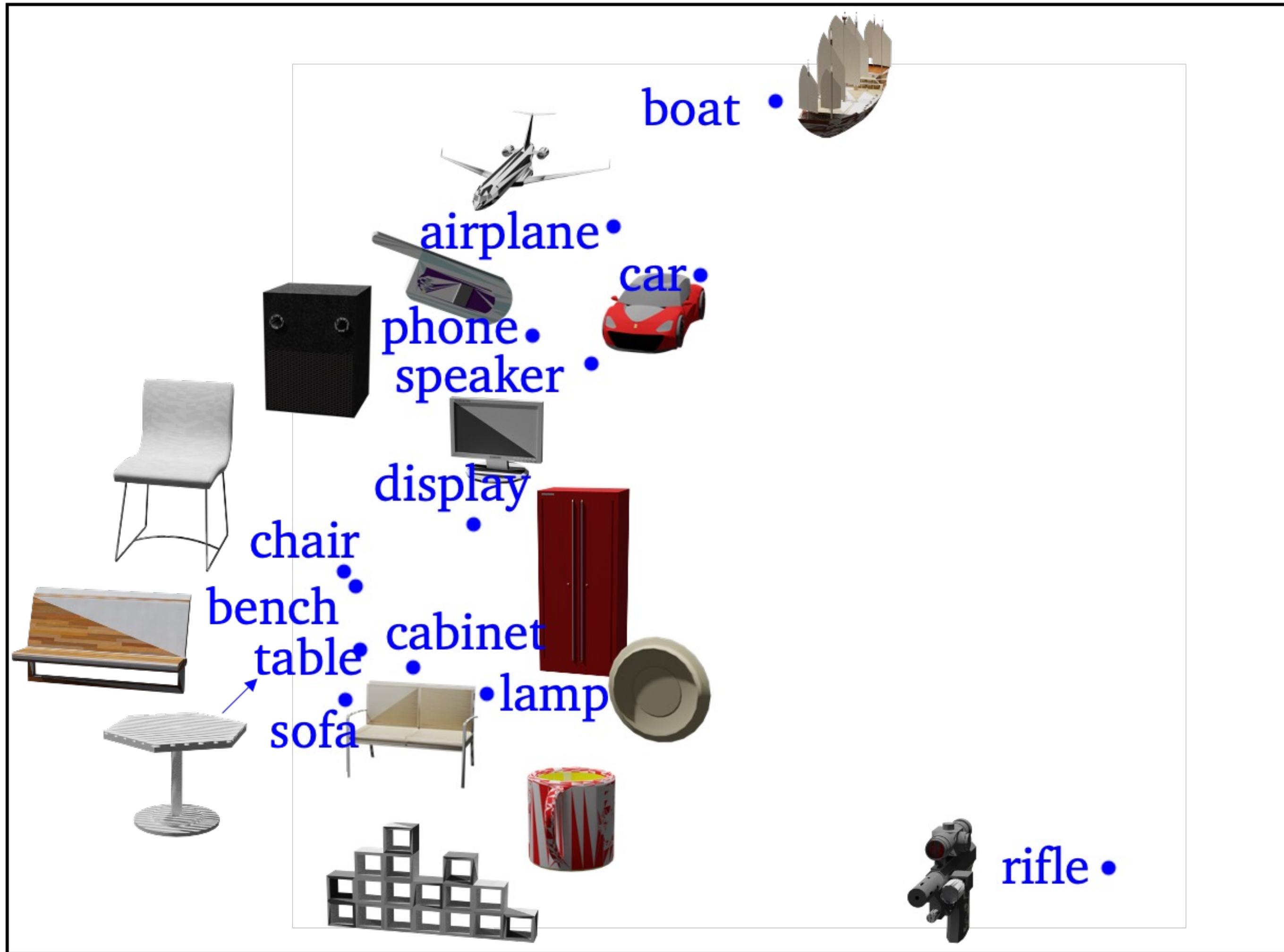
## Ours

- Equivariant representation
- Object identity + pose



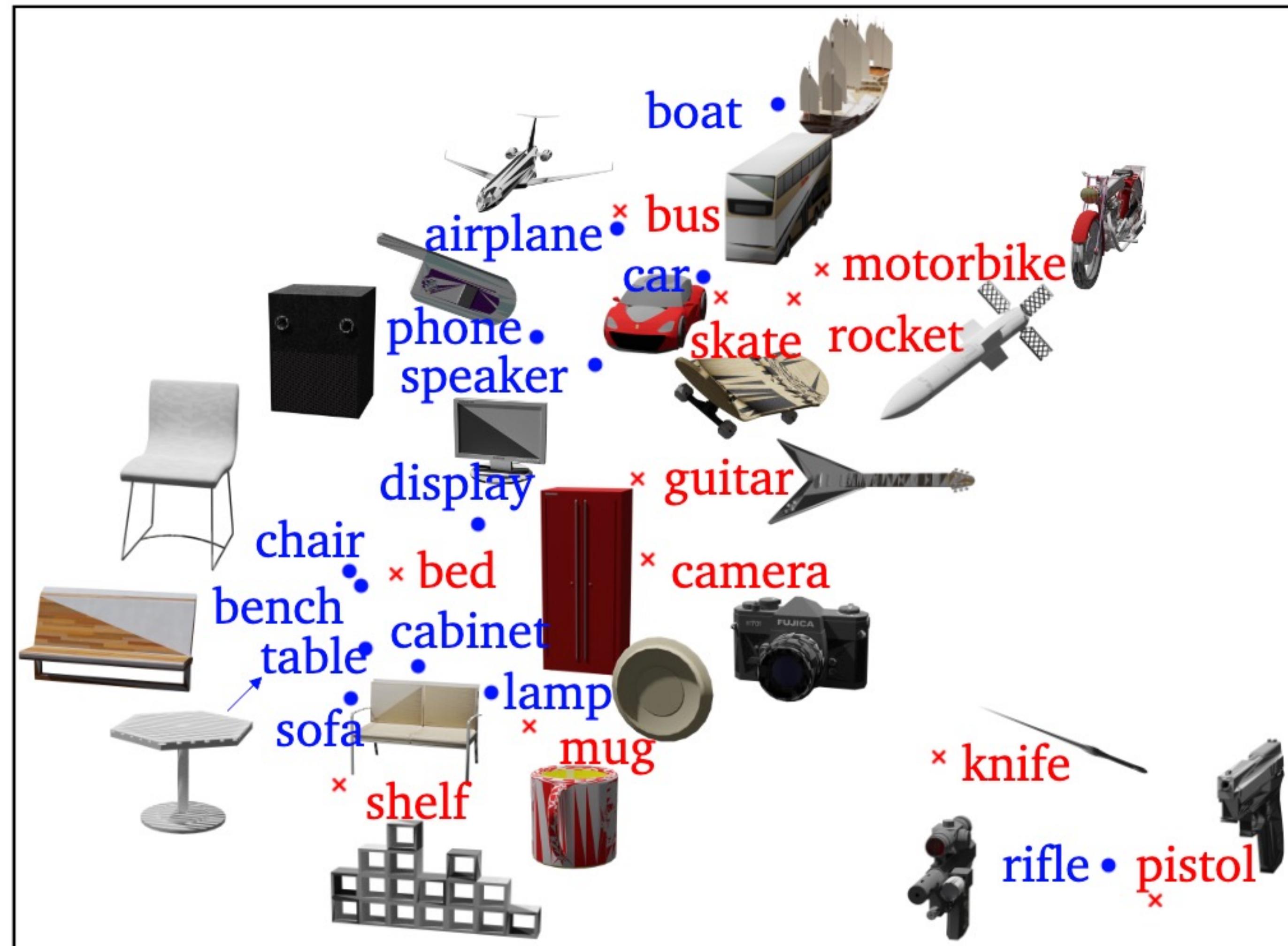
# Benchmark: Data

## 13 in-domain

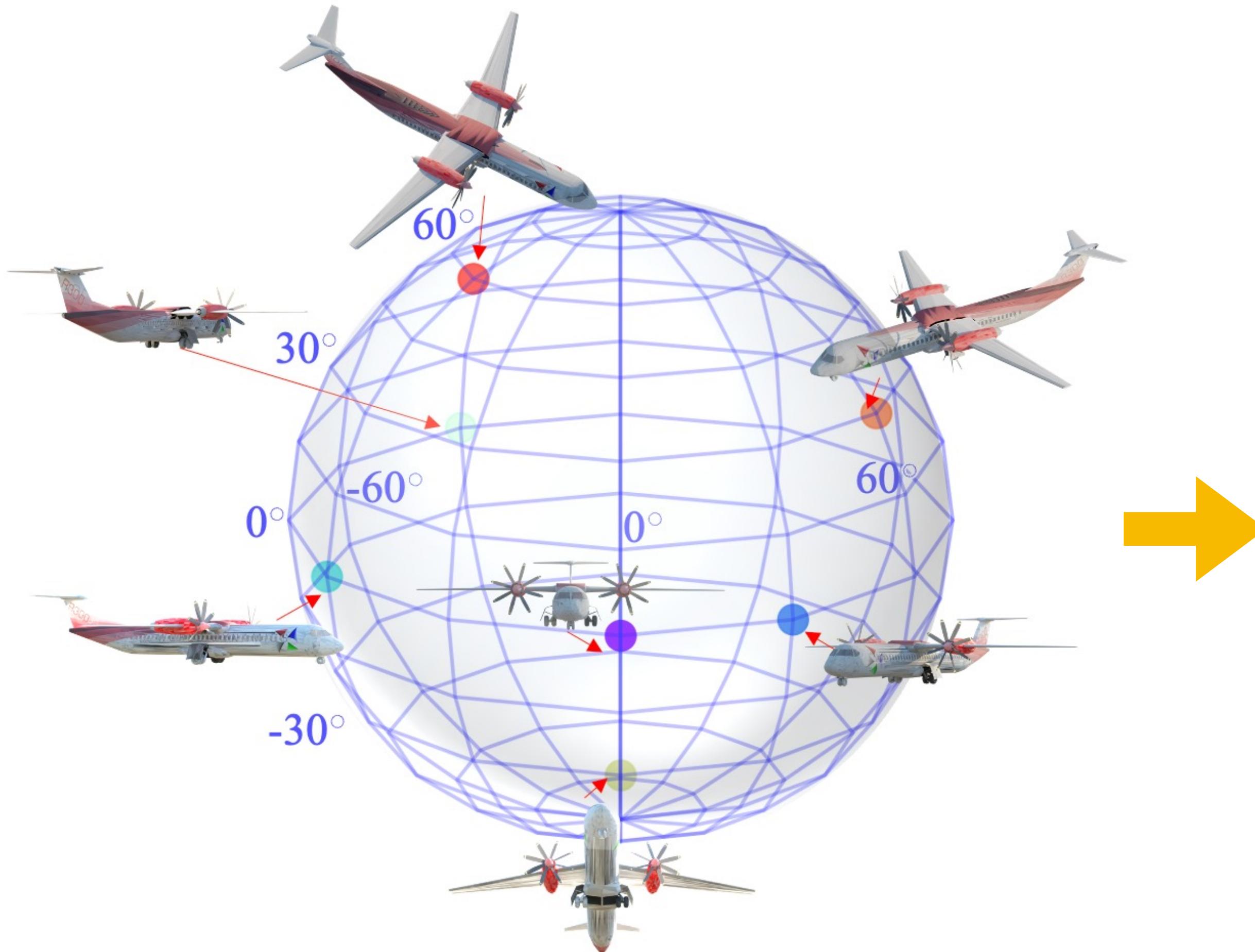


# Benchmark: Data

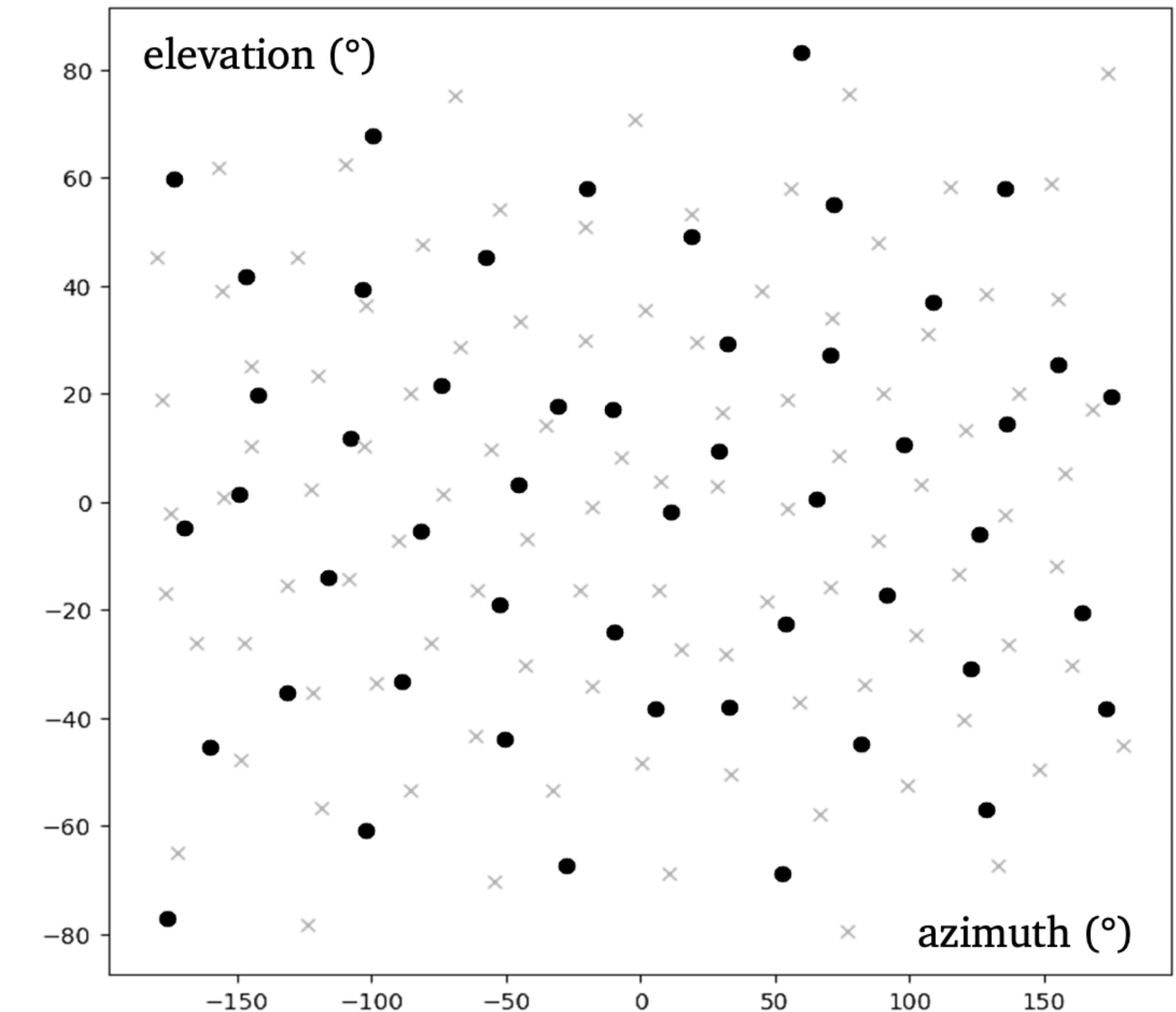
13 **in-domain** & 20 **out-of-domain** semantic categories



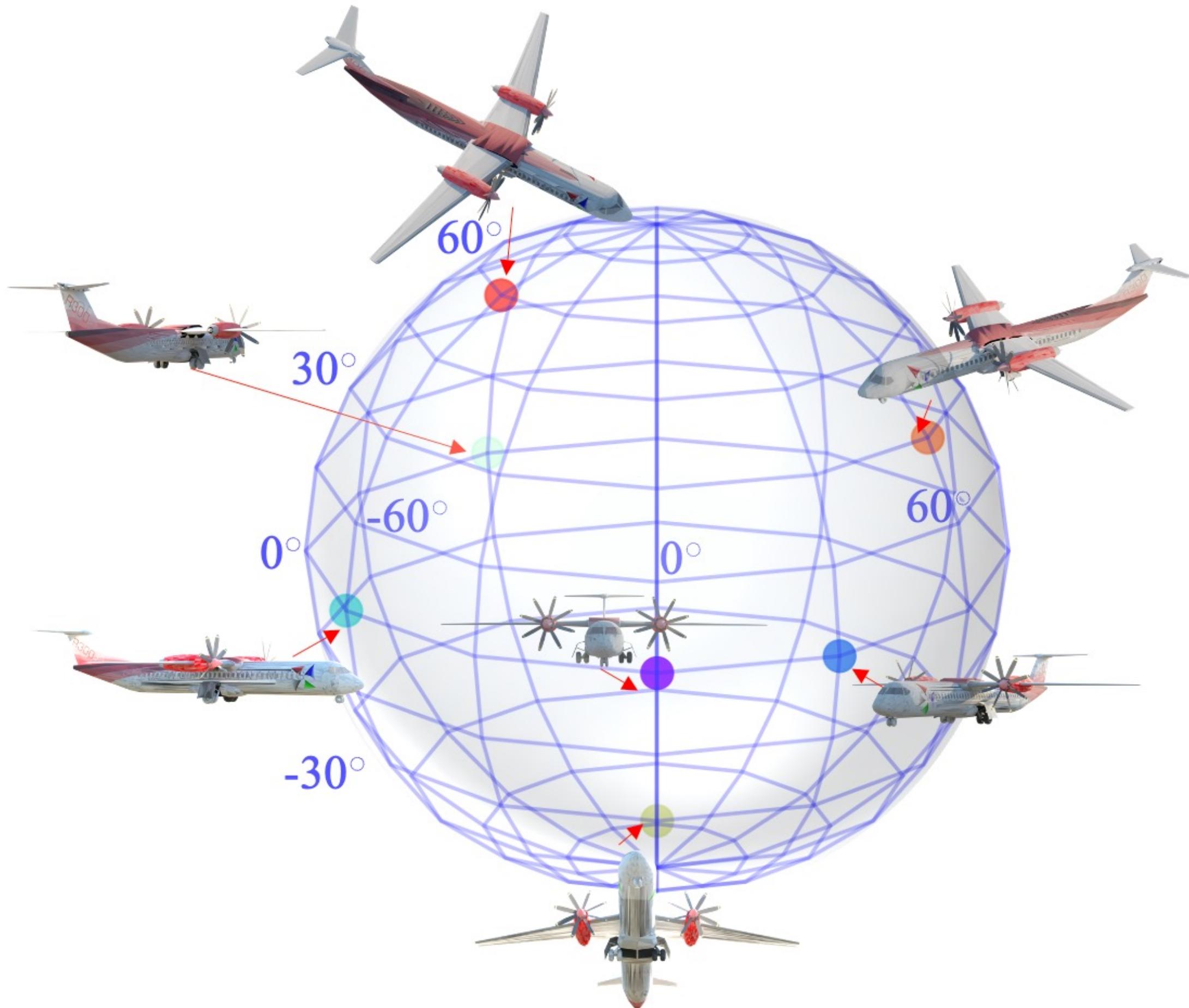
# Benchmark: Data



in-domain & out-of-domain pose

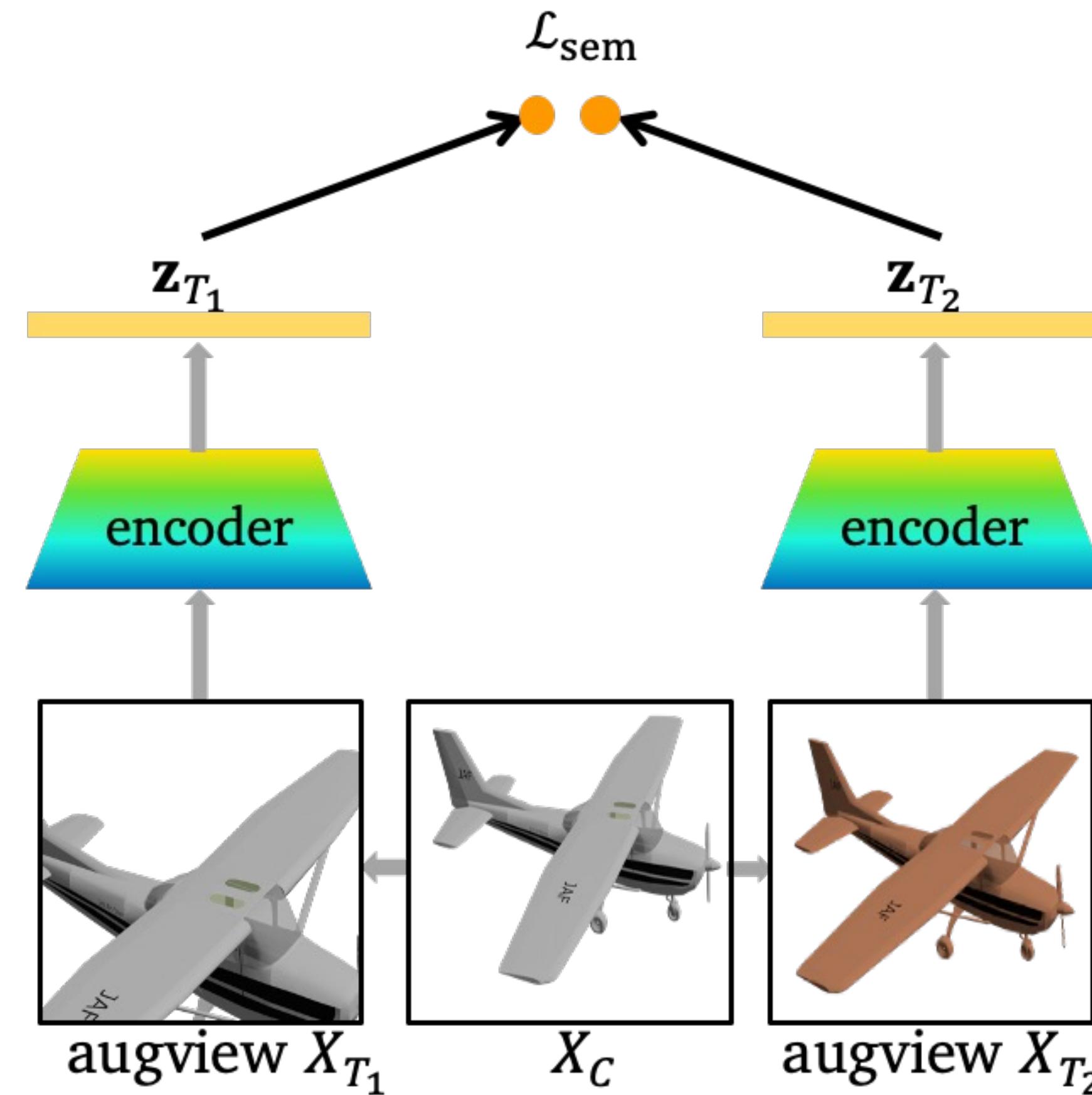


# Benchmark: Tasks



- Semantic classification
- Absolute pose → global pose
- Relative pose → generalizability
  - Category-specific pose free
  - Generalize to open categories

# Self-Supervised Representation Learning: Invariance

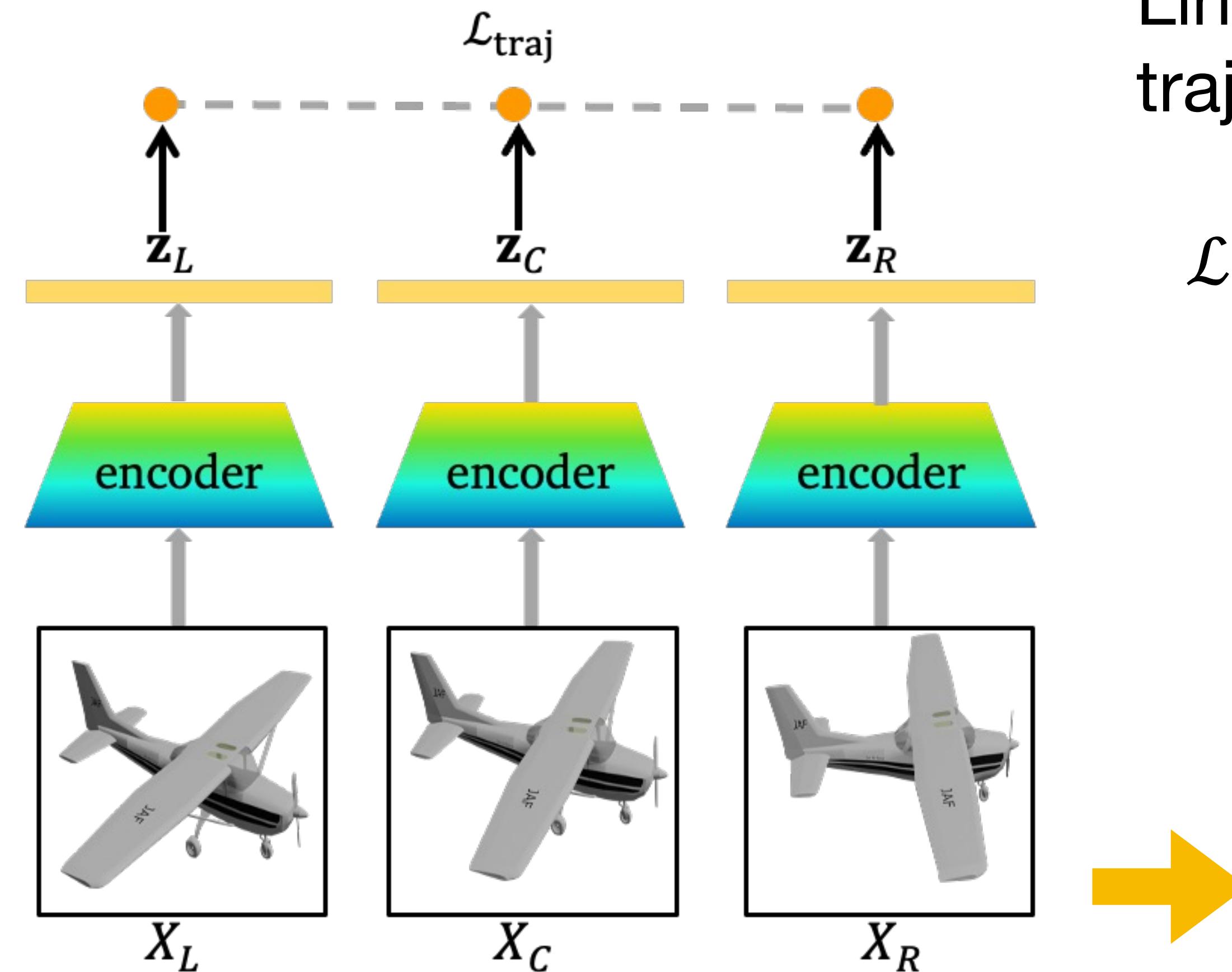


Example: VICReg

Augmentations:

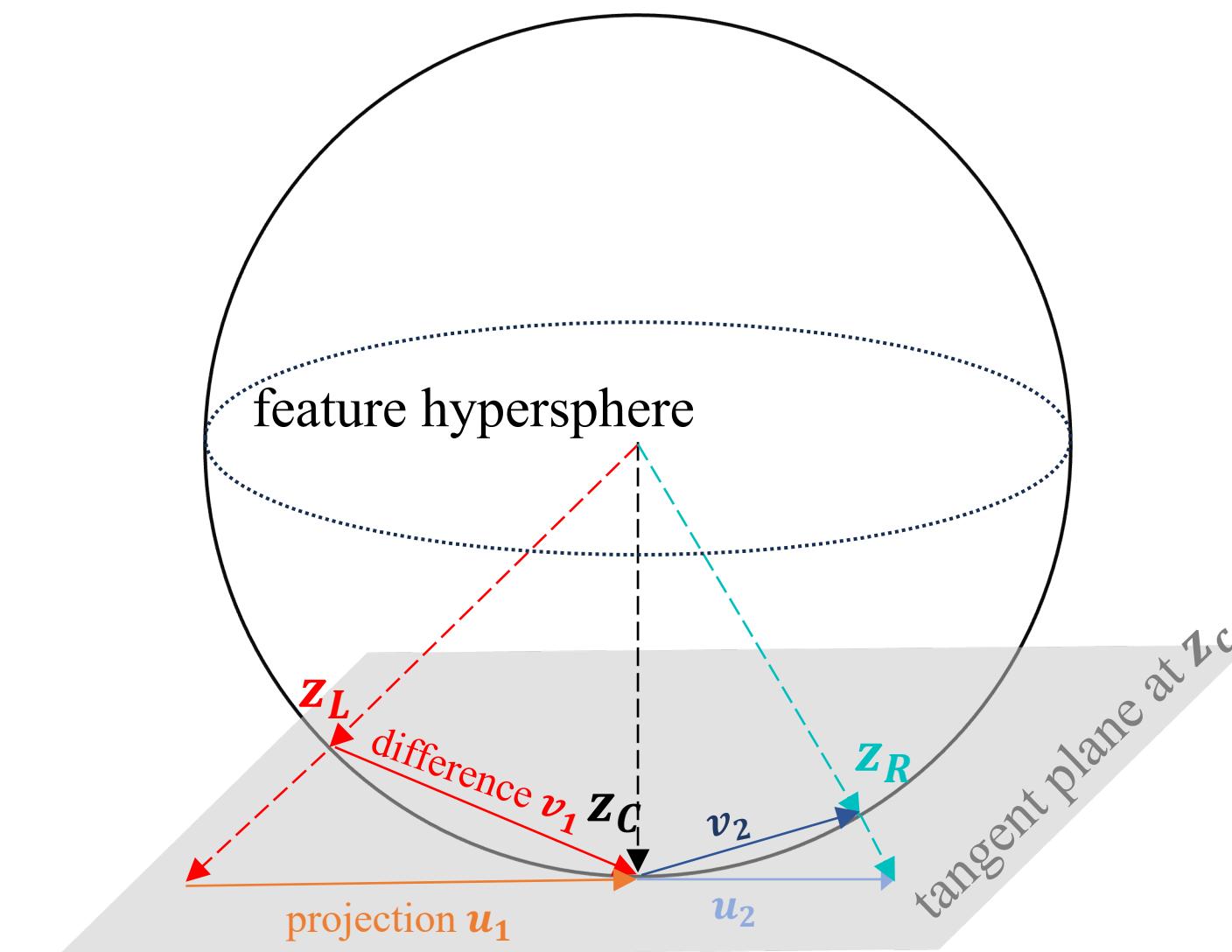
- Random crops
- Color jittering
- Gaussian Blur

# Self-Supervised Representation Learning: Trajectory Regularization



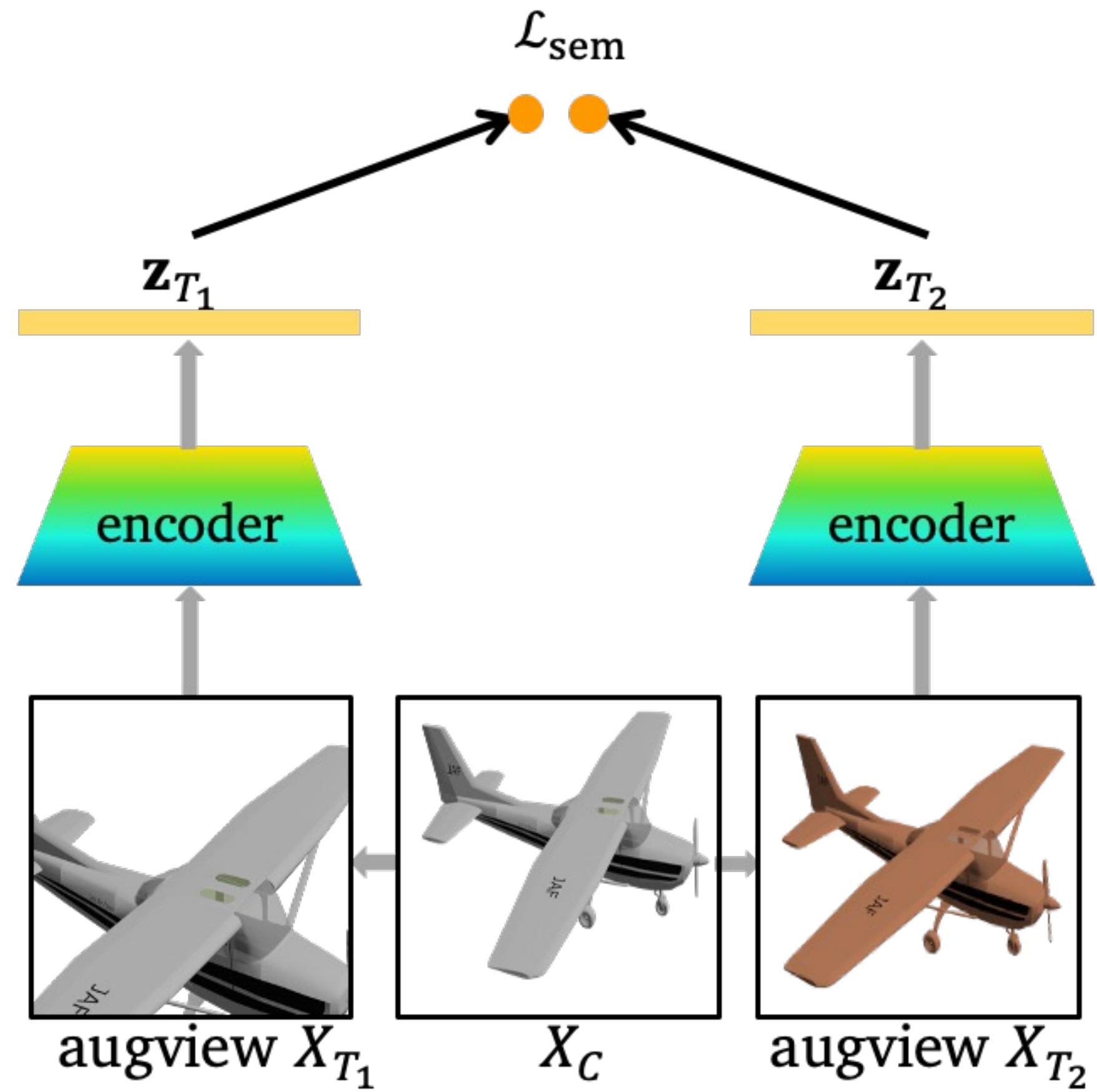
Line up 3 embeddings via trajectory regularization:

$$\mathcal{L}_{\text{traj}}(z_L, z_C, z_R) = -\frac{\mathbf{u}_1 \cdot \mathbf{u}_2}{\|\mathbf{u}_1\| \|\mathbf{u}_2\|}$$

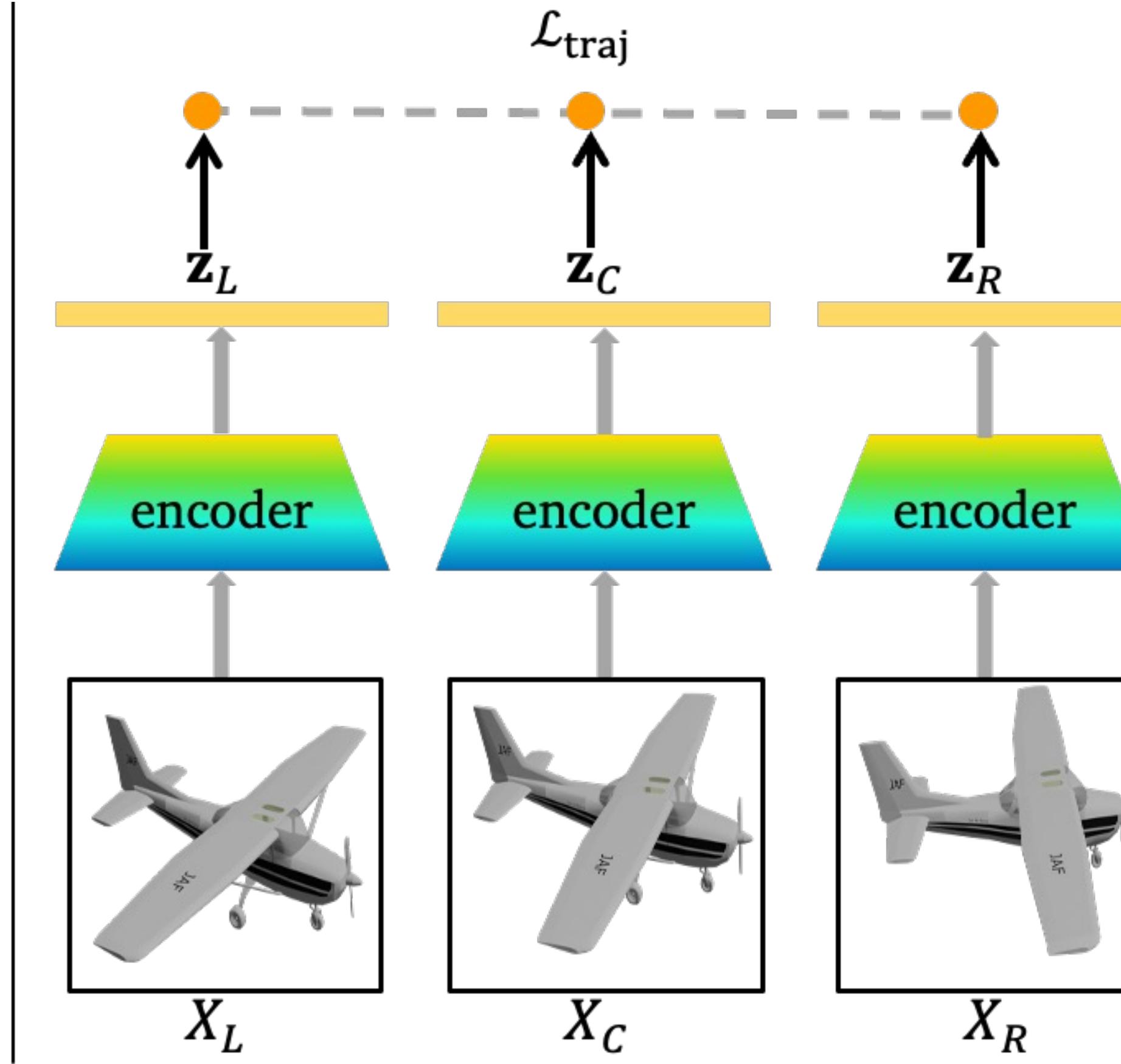


# Self-Supervised Representation Learning

## Invariant Learning

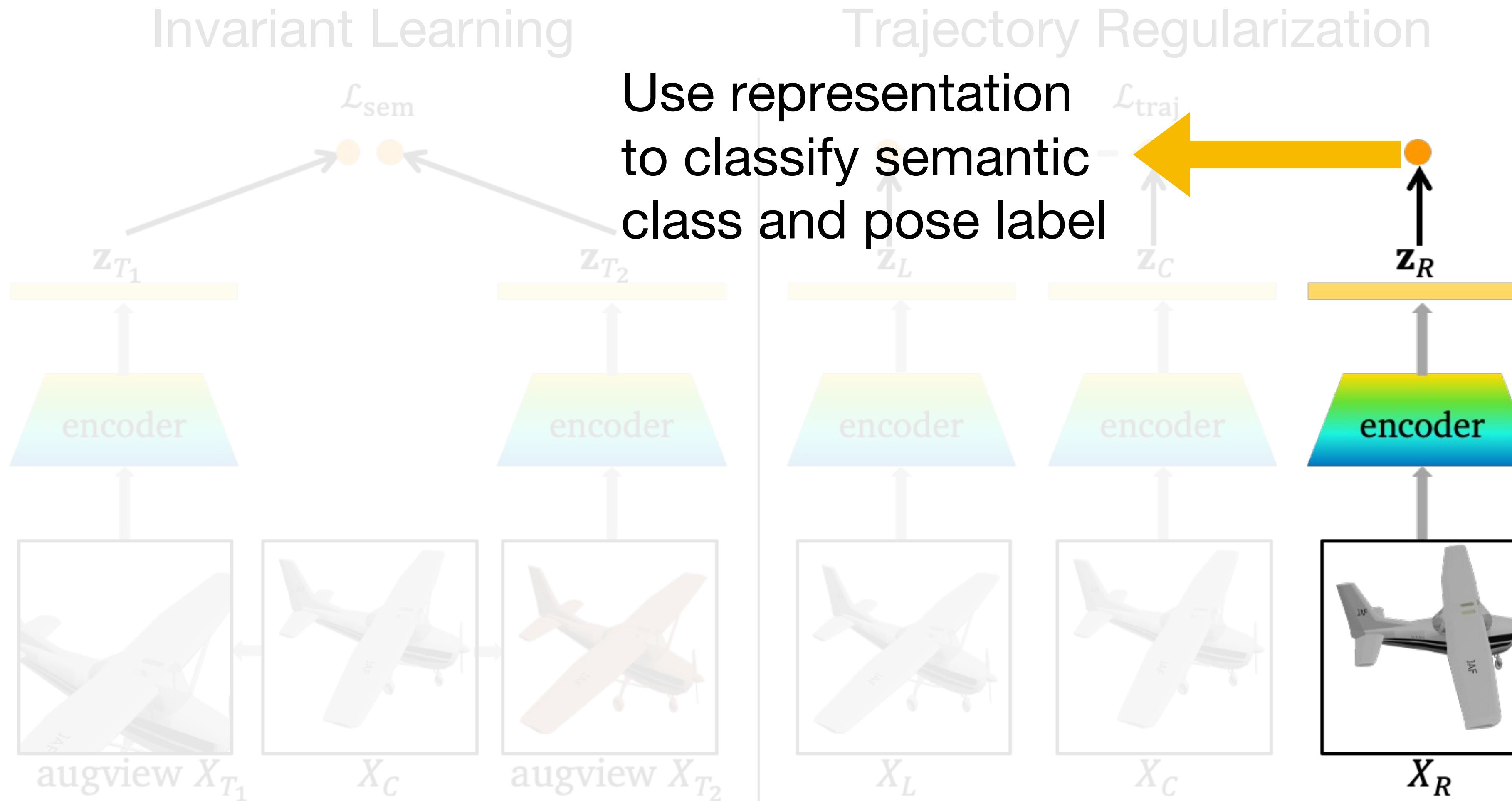


## Trajectory Regularization



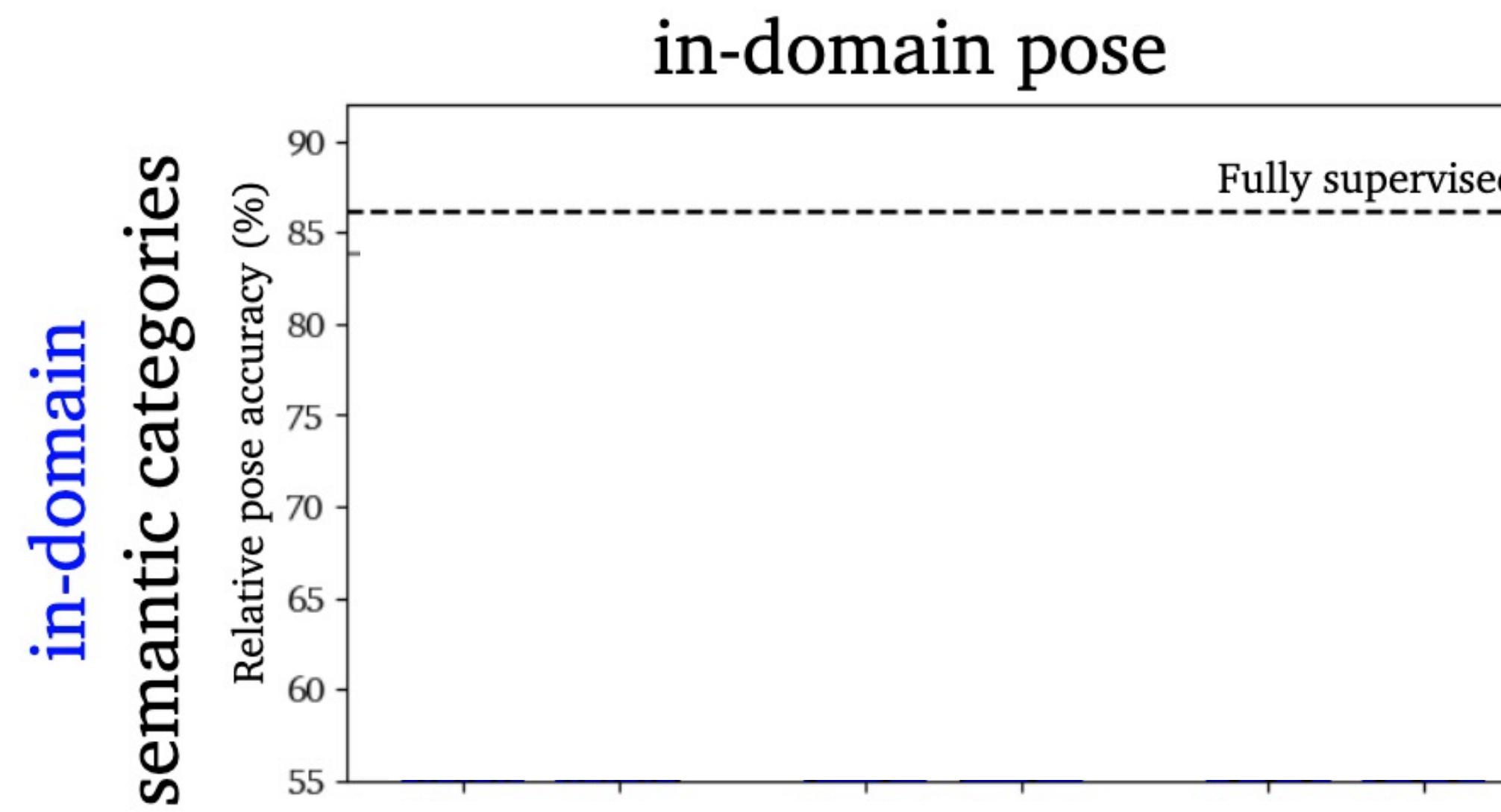
Final loss is a combination:  $\mathcal{L} = \mathcal{L}_{sem}(\mathbf{z}_{T_1}, \mathbf{z}_{T_2}) + \lambda \mathcal{L}_{traj}(\mathbf{z}_L, \mathbf{z}_C, \mathbf{z}_R)$

# After Representation Learning: Probing

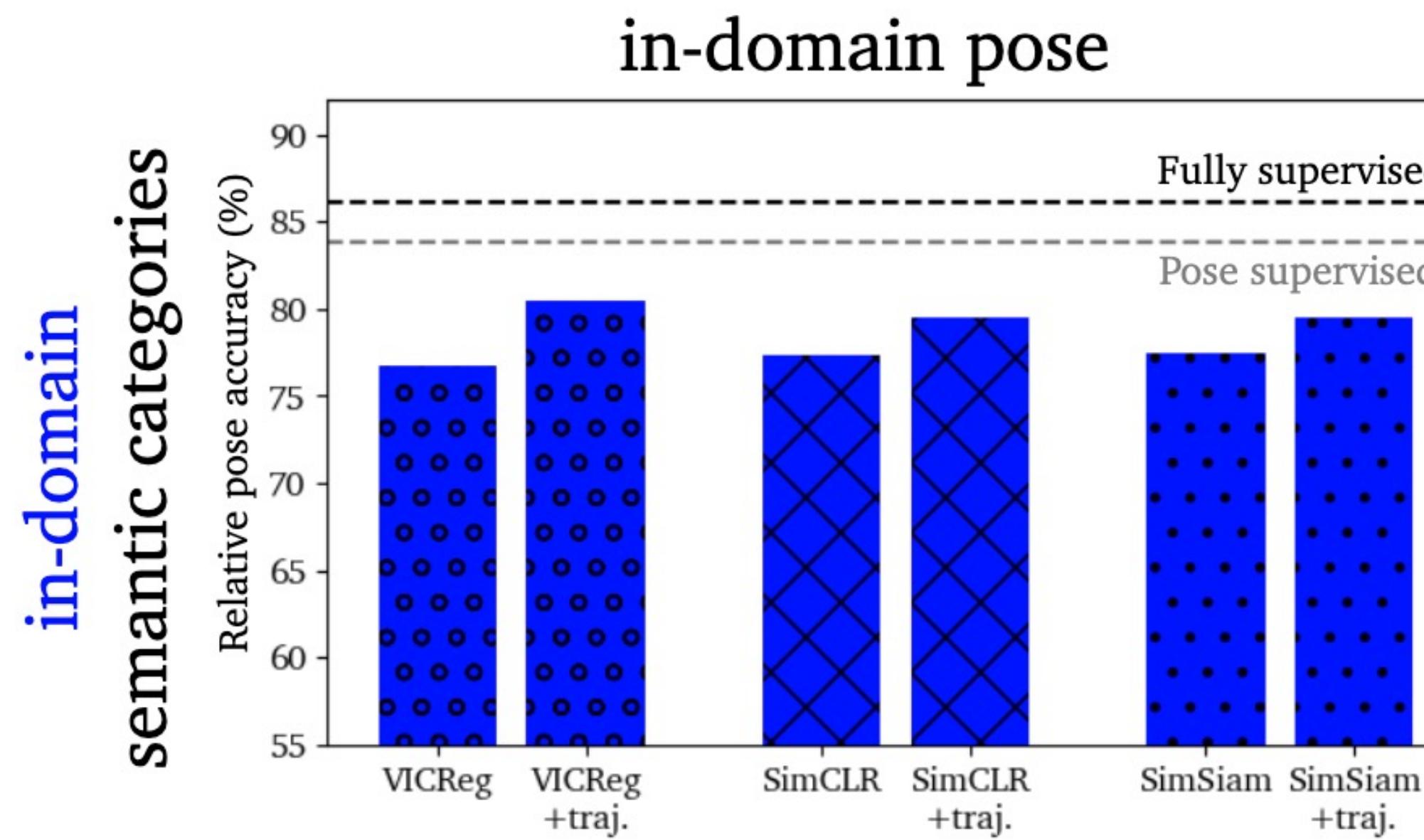


Final loss is a combination:  $\mathcal{L} = \mathcal{L}_{sem}(\mathbf{z}_{T_1}, \mathbf{z}_{T_2}) + \lambda \mathcal{L}_{traj}(\mathbf{z}_L, \mathbf{z}_C, \mathbf{z}_R)$

# Improved In-Domain and Out-of-Domain Pose Accuracy

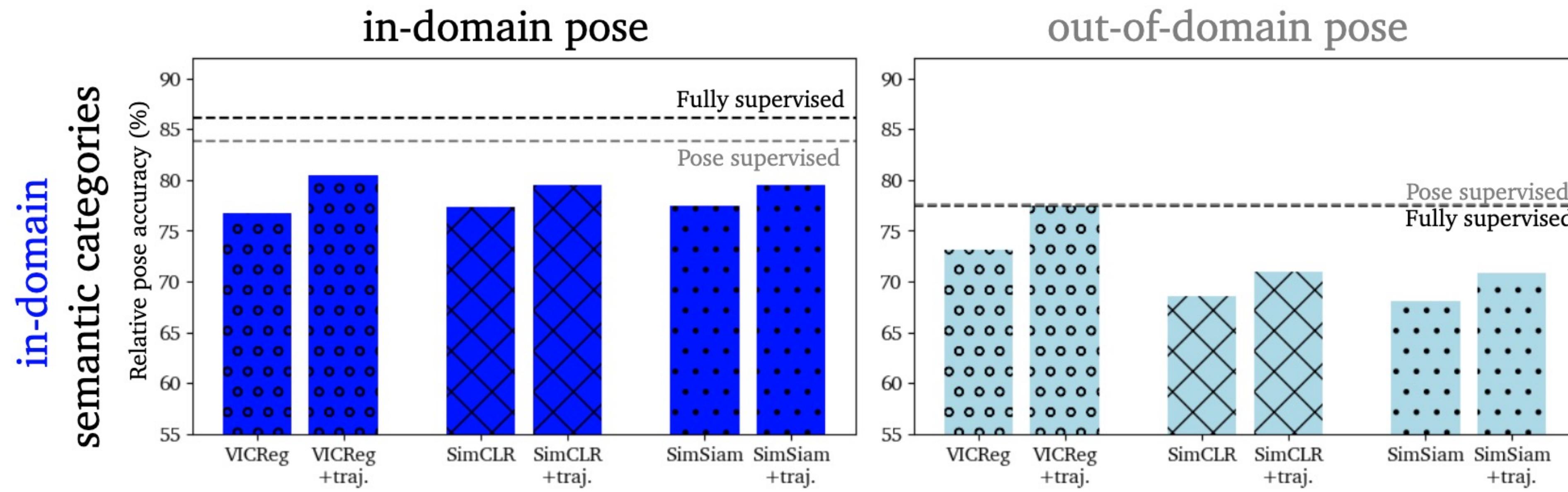


# Improved In-Domain and Out-of-Domain Pose Accuracy



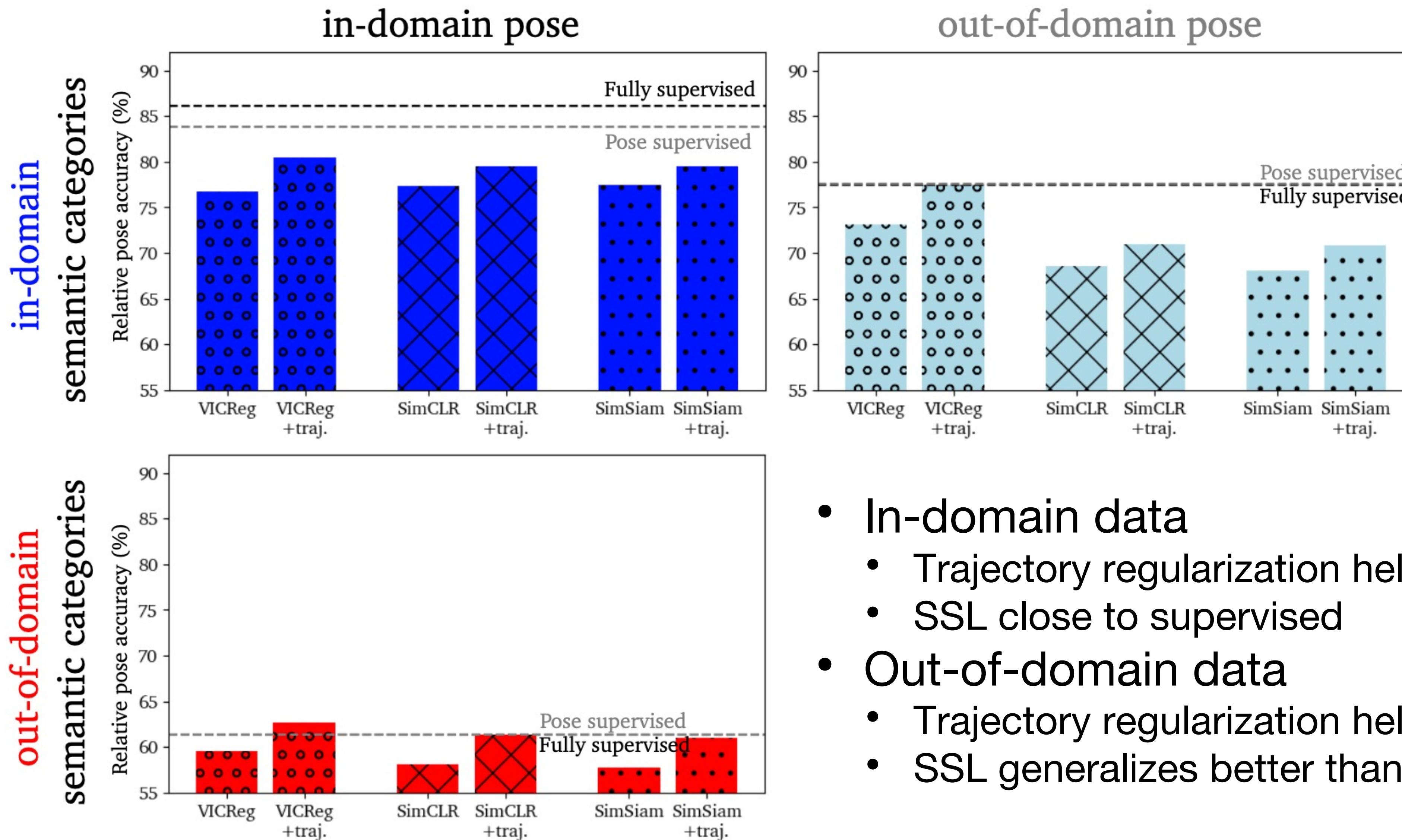
- In-domain data
  - Trajectory regularization helps
  - SSL close to supervised

# Improved In-Domain and Out-of-Domain Pose Accuracy



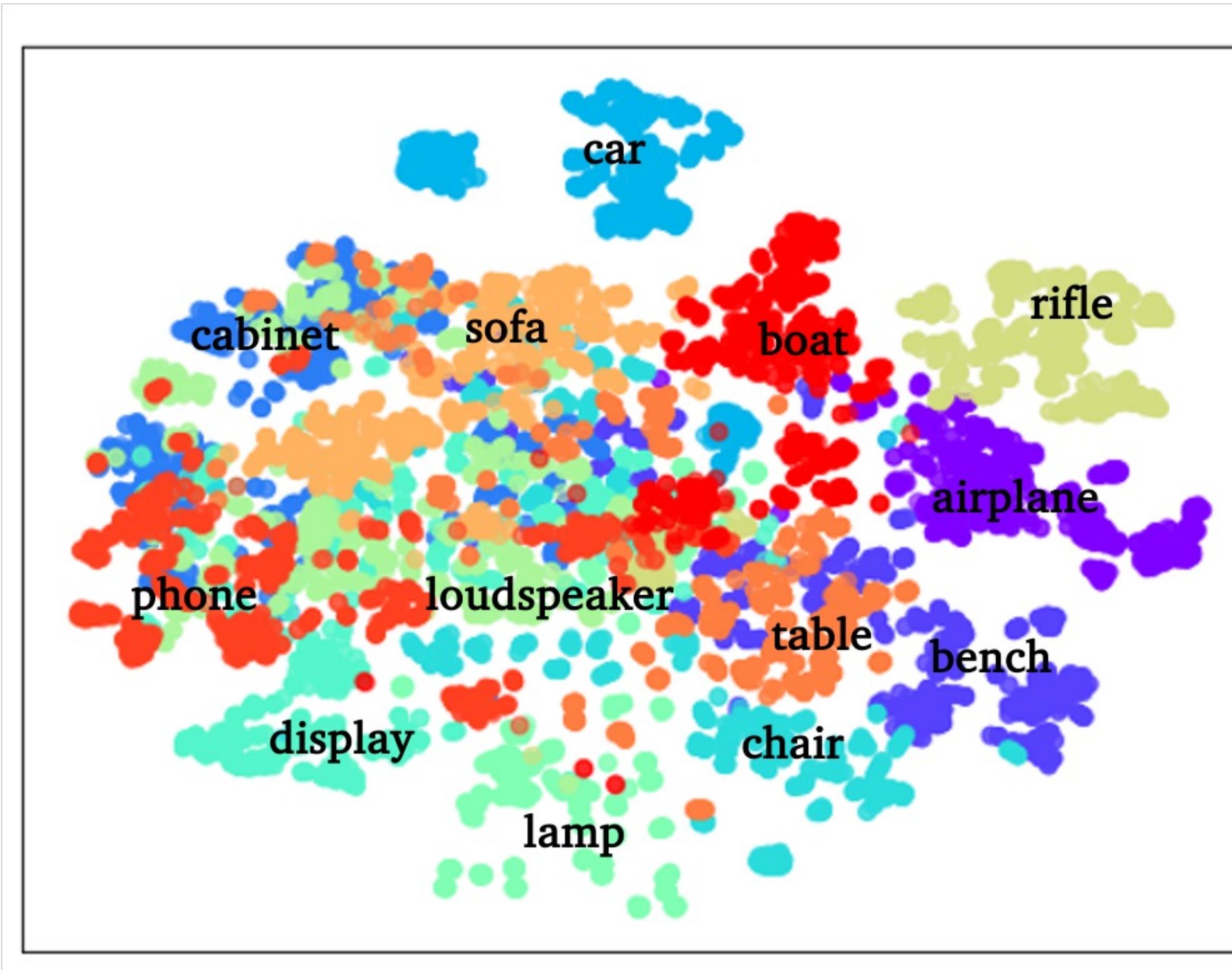
- In-domain data
  - Trajectory regularization helps
  - SSL close to supervised
- Out-of-domain data
  - Trajectory regularization helps
  - SSL generalizes better than supervised

# Improved In-Domain and Out-of-Domain Pose Accuracy



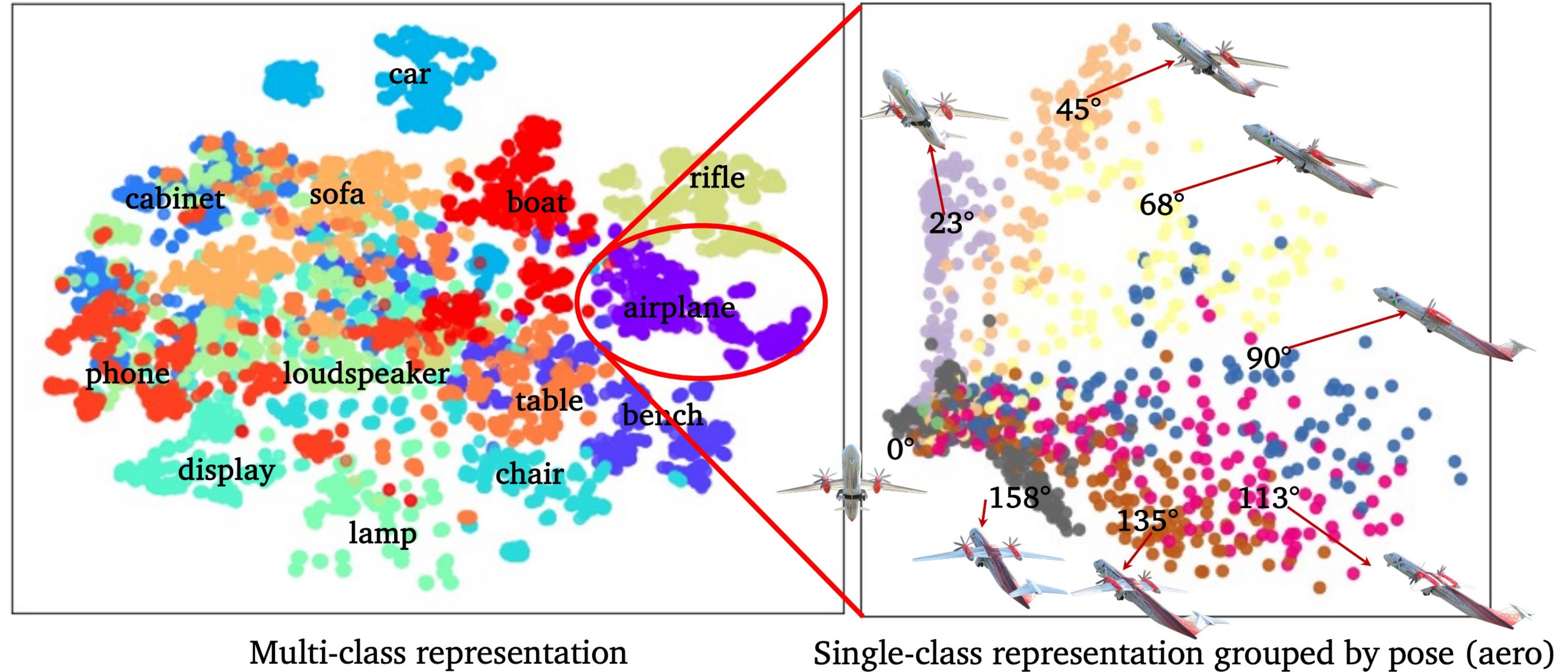
- In-domain data
  - Trajectory regularization helps
  - SSL close to supervised
- Out-of-domain data
  - Trajectory regularization helps
  - SSL generalizes better than supervised

# Visualizing Representation



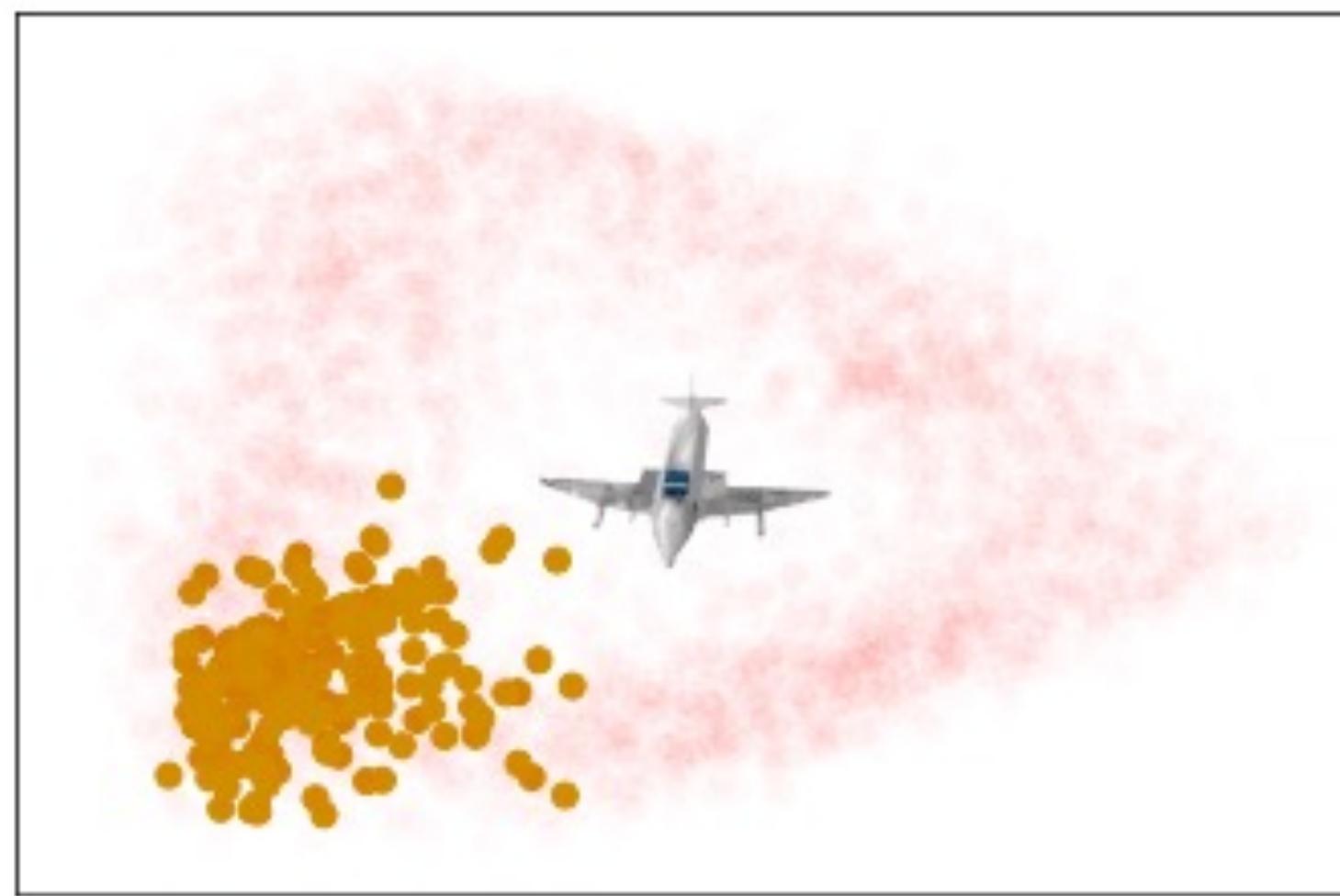
Multi-class representation

# Visualizing Representation

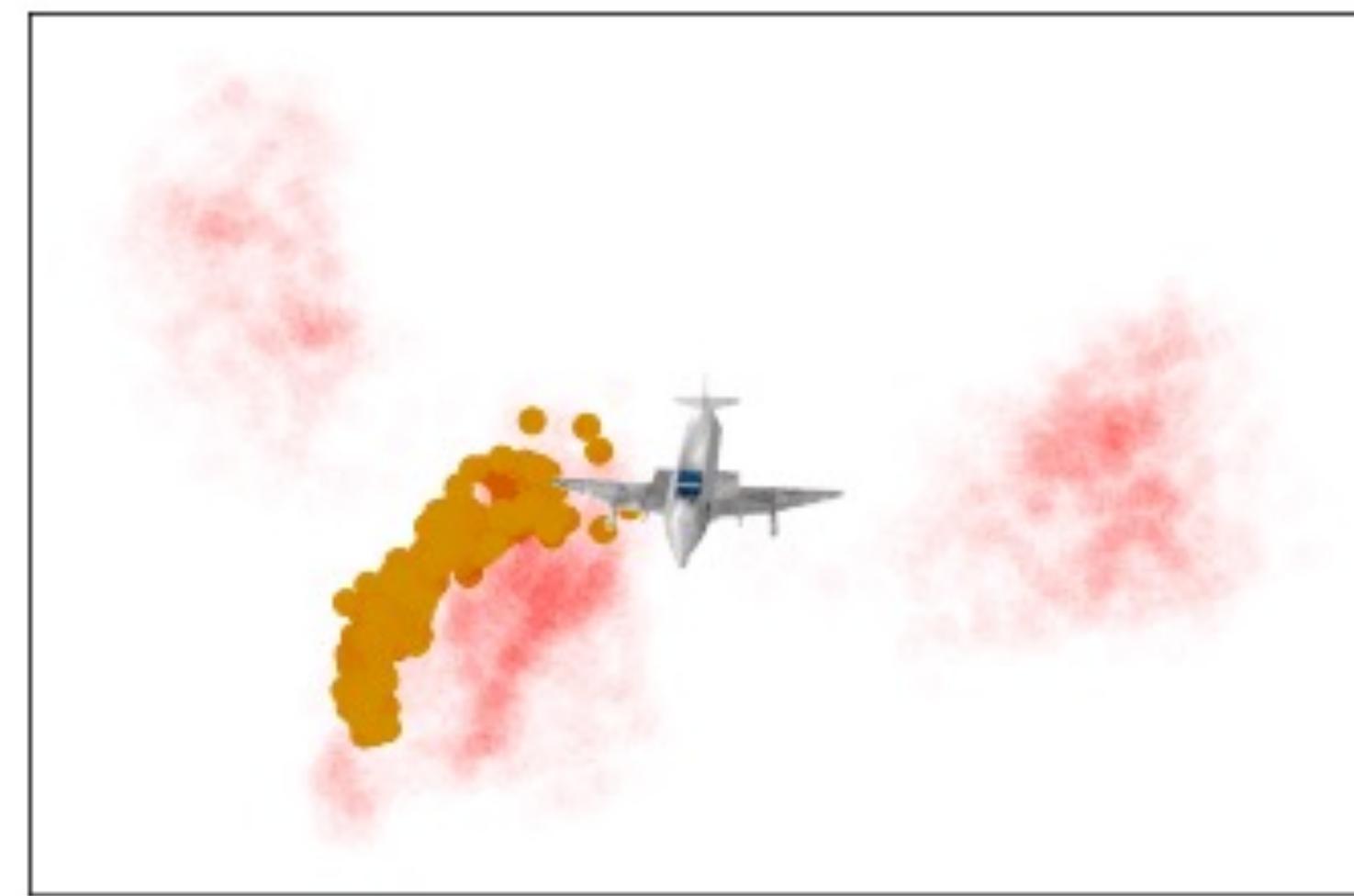


Emergent pose-semantic representation without labels!

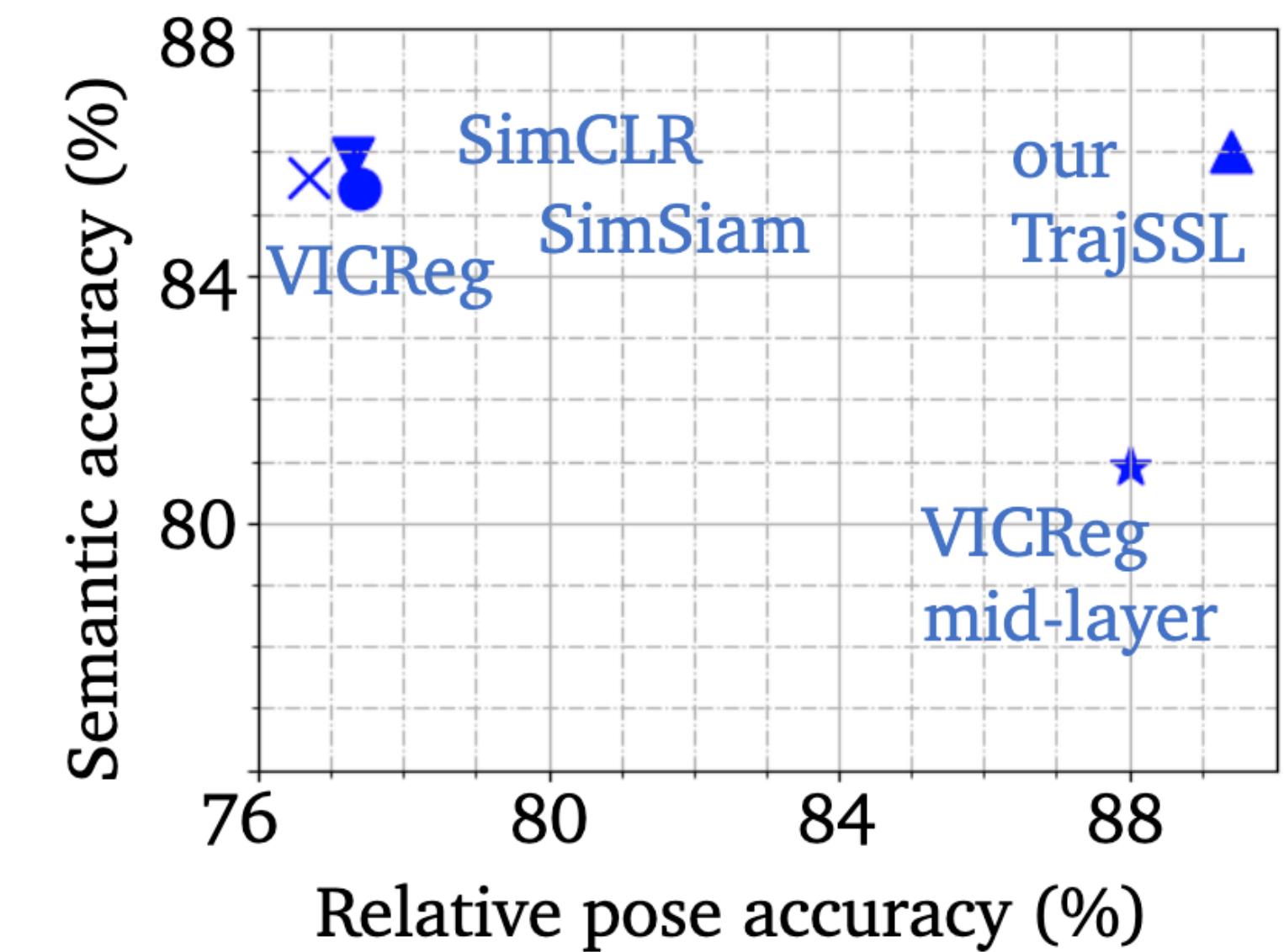
# Compared to Baseline SSL



VICReg  
+trajectory regularization



VICReg



**Thank you!**

Please come to our poster: #256  
Paper/code/data:



Contact: [peterw@caltech.edu](mailto:peterw@caltech.edu)