Berkeley UNIVERSITY OF CALIFORNIA
UNIVERSITY OF MICHIGAN
Caltech
UCDAVIS

ECCV

Pose-Aware Self-Supervised Learning with Viewpoint Trajectory Regularization

Jiayun Wang

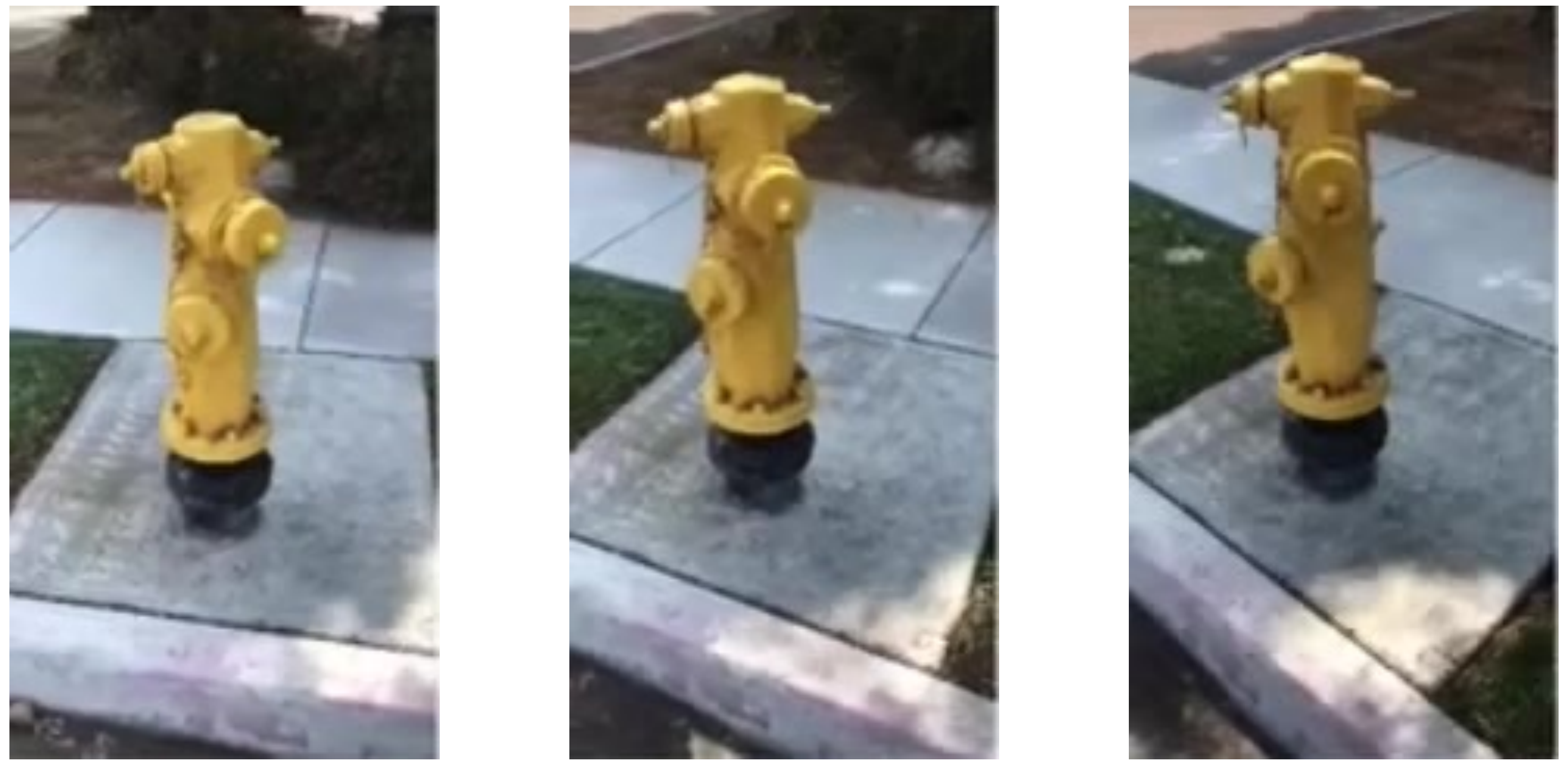
On job market!
peterw@caltech.edu

Yubei Chen

Stella X. Yu

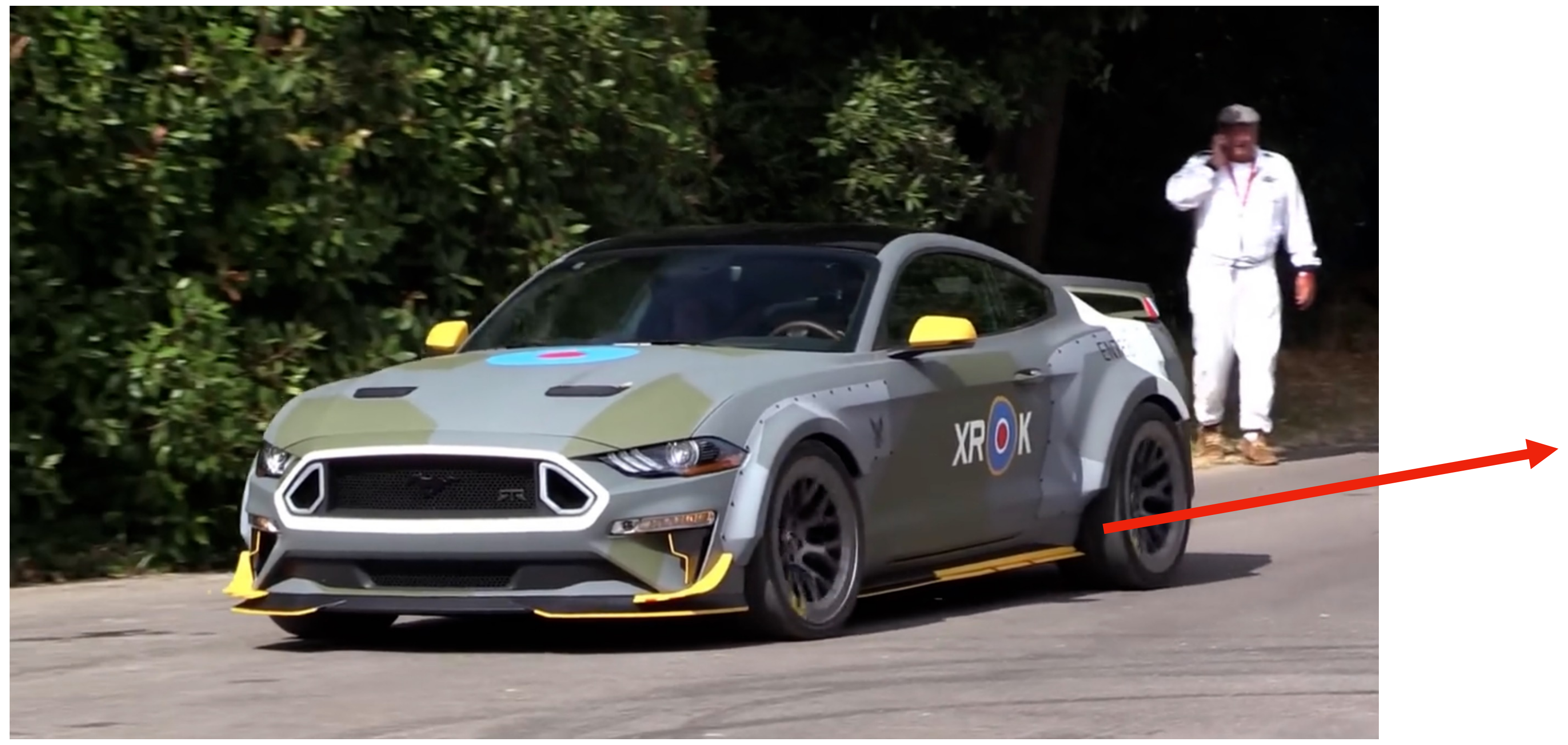
A robot moves around in the environment and encounters a new object

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



Video credit: Common objects in 3d. ICCV 2021

XR K
NM2255 HD



Existing SSL: car

Expected:

- Car heading towards the camera
- At danger!

Video credit: The BEST Car Donuts. Youtube

NM2255 HD

Existing SSL: car

Expected:

- Car heading towards the camera
- At danger!

Existing SSL: car

Expected

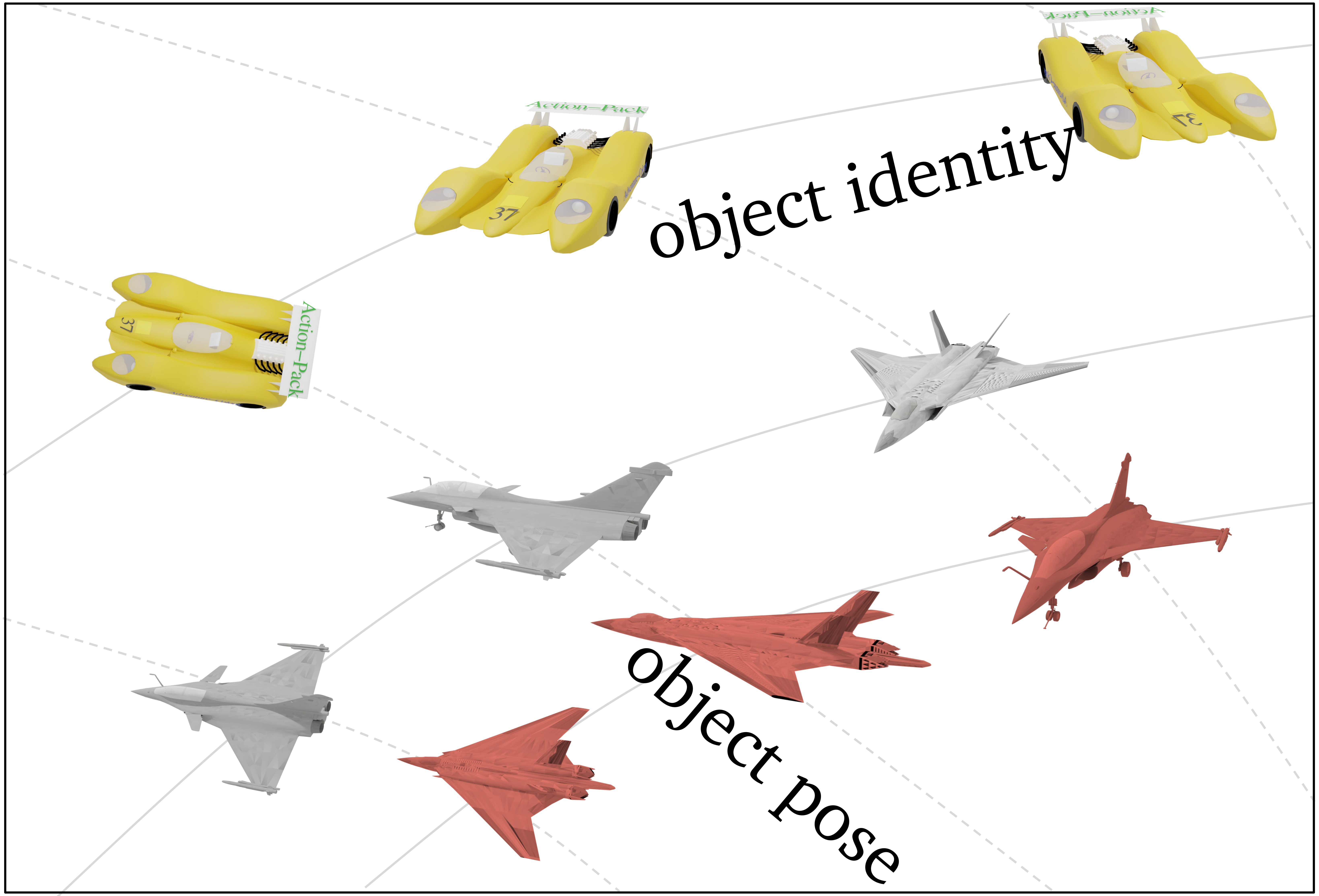
- Car heading away from the camera
- No danger

Recognition needs to understand both aspects:

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

Video credit: The BEST Car Donuts. Youtube

Learn **disentangled** semantic-pose representation with SSL?

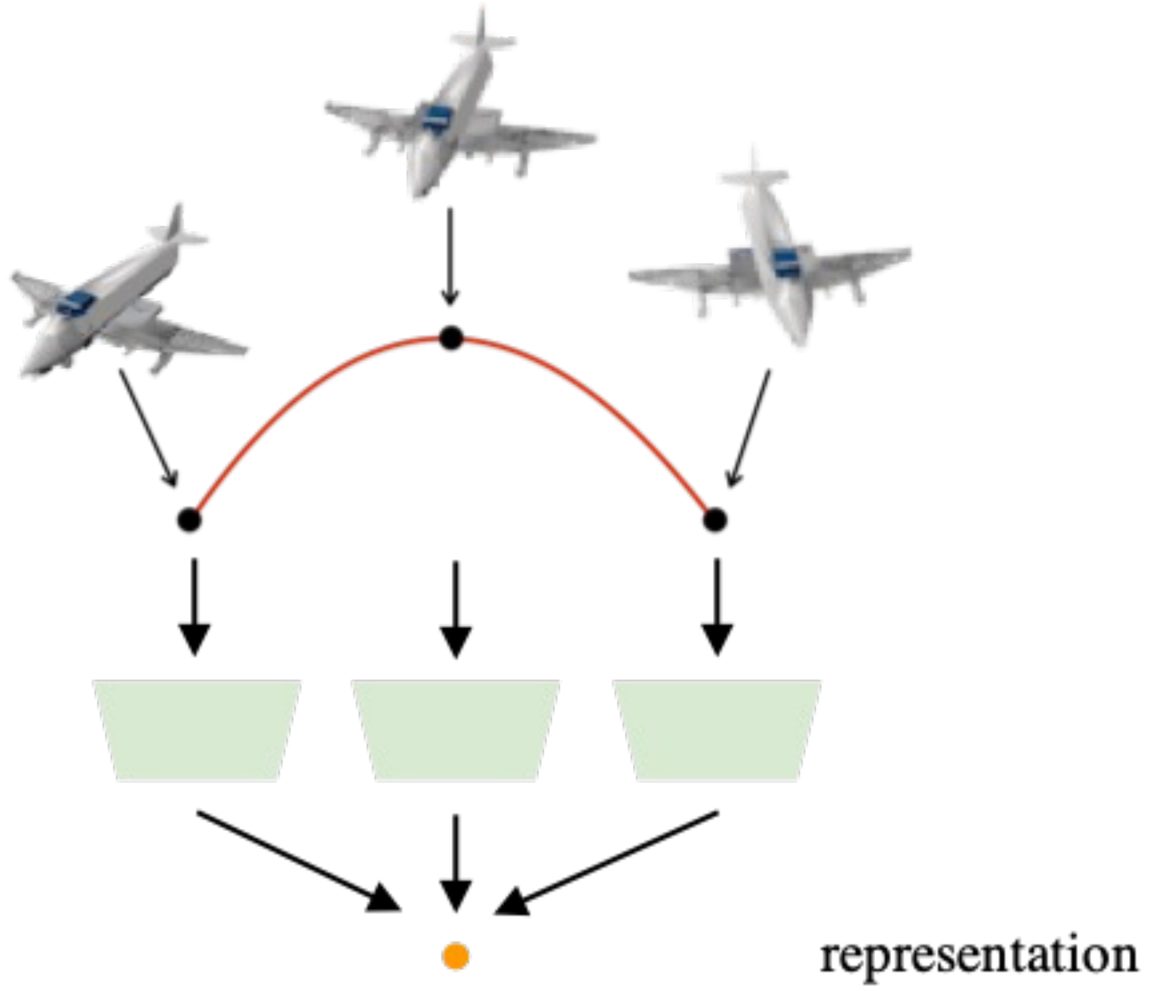


* Image embedding projected to 2D

Existing SSL

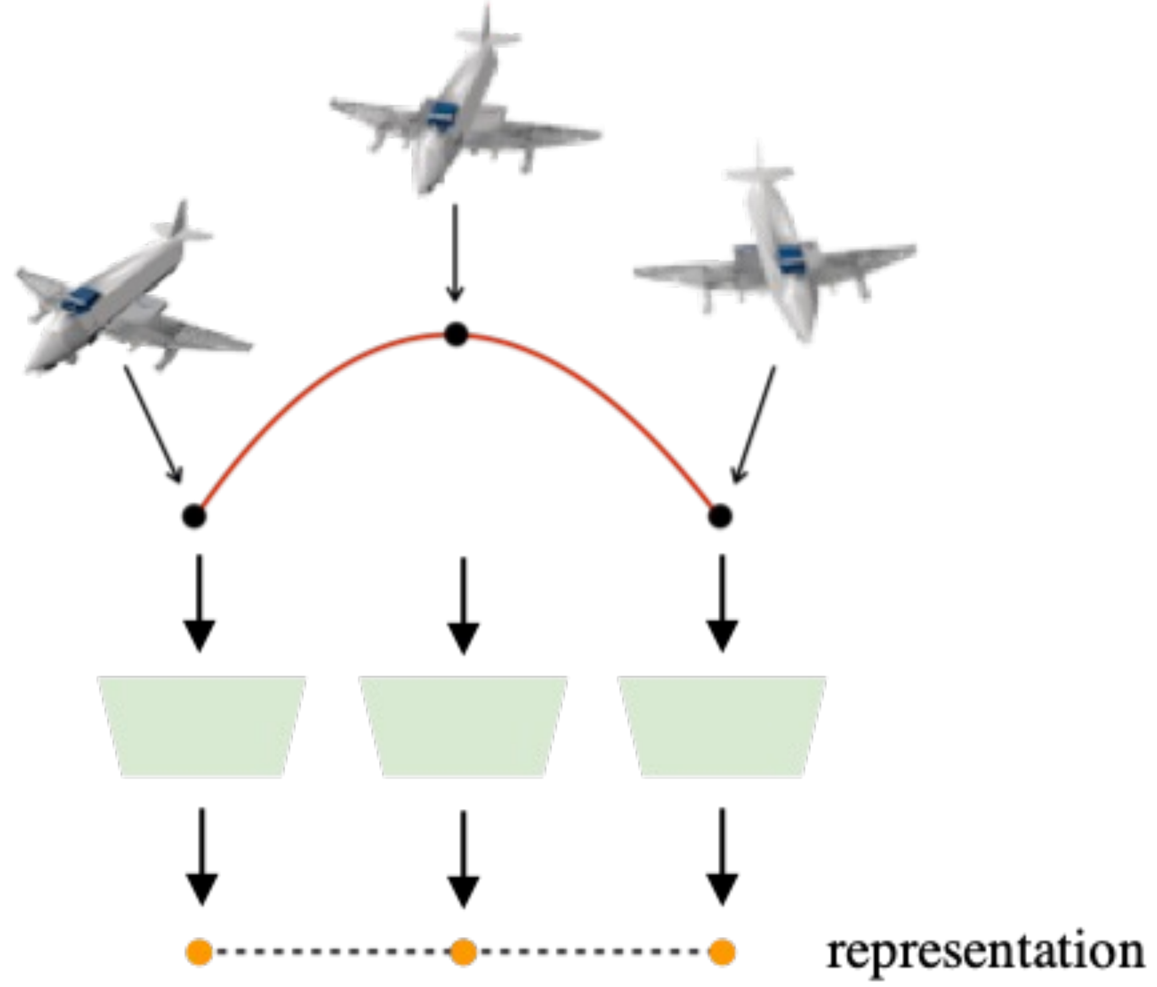
VICReg, SimCLR, SimSiam, MoCo,…

- Invariant representation
- Object identity only



Ours

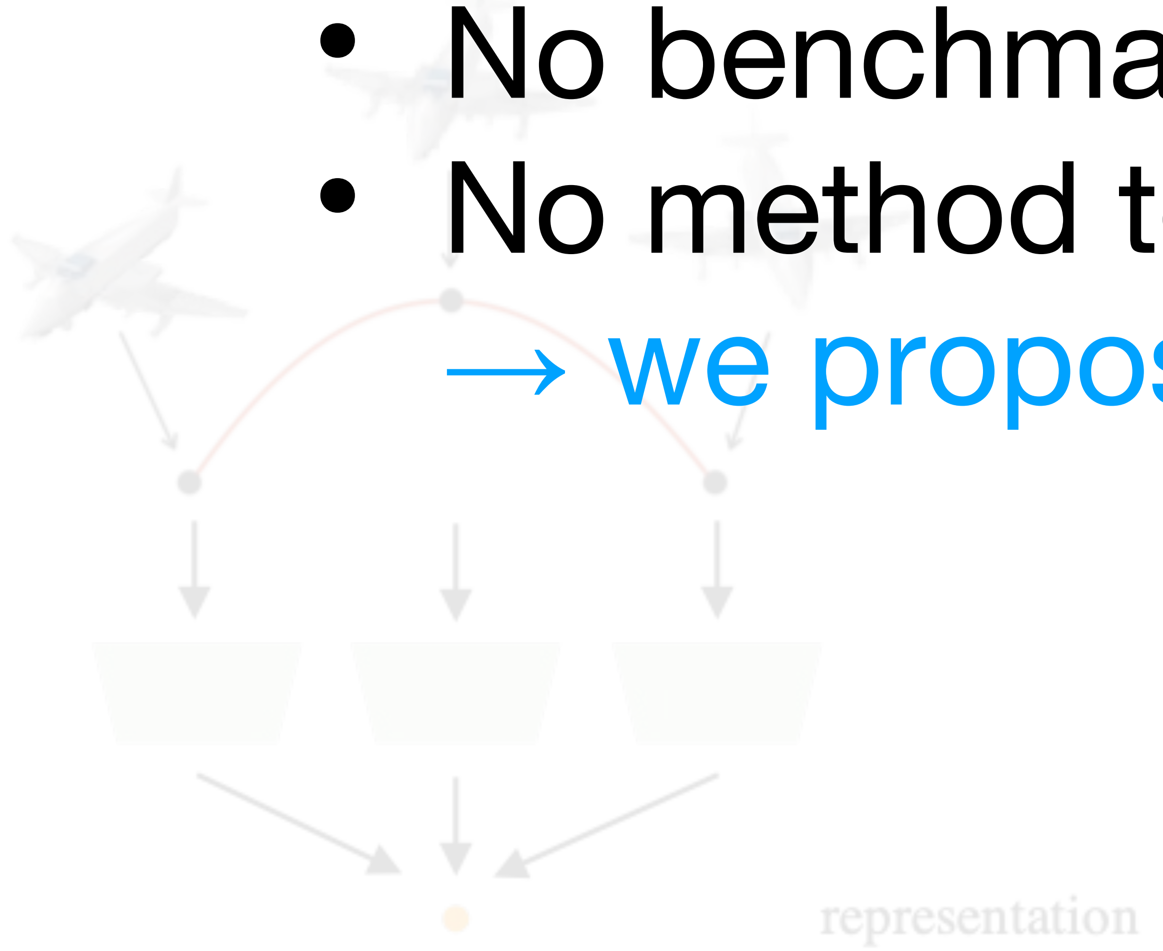
- Pose-aware representation
- Object identity + pose



Existing SSL

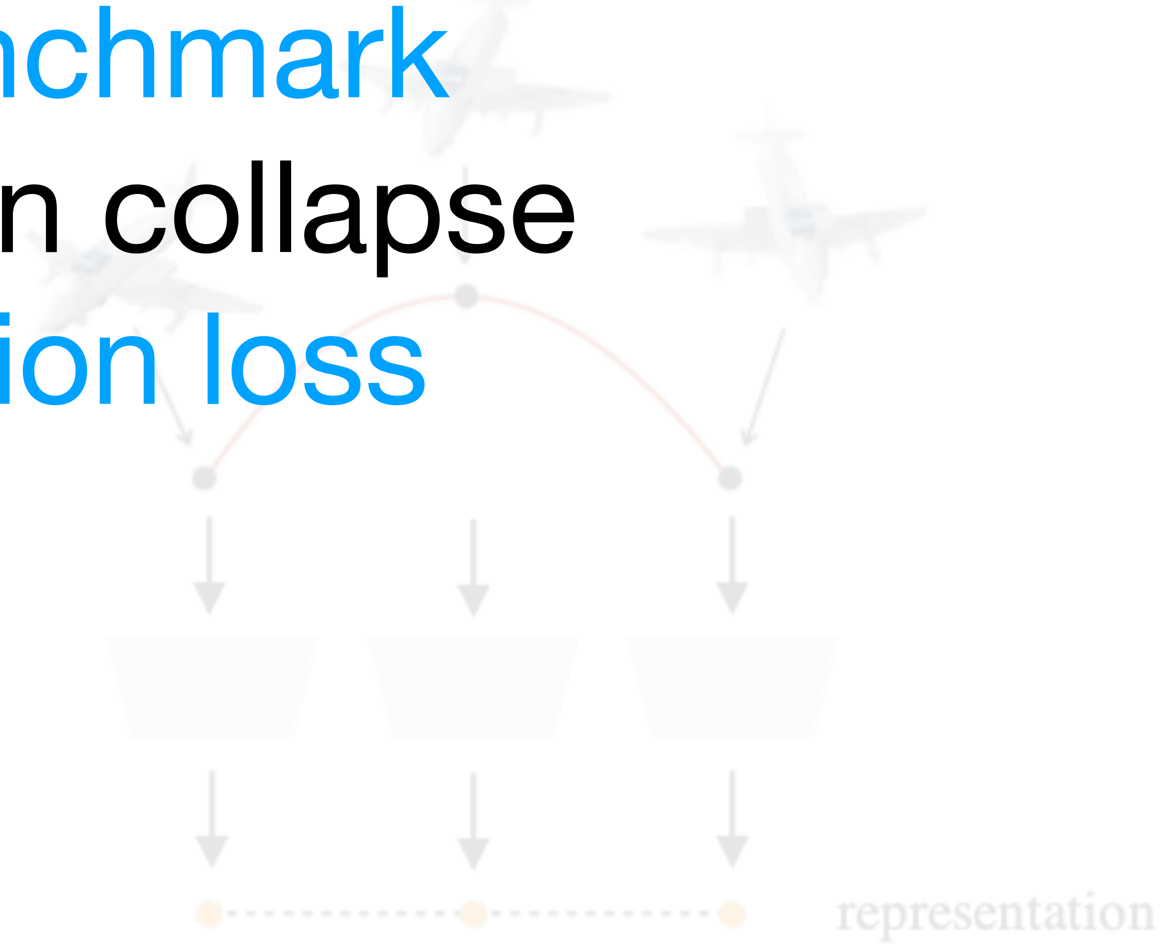
VICReg, SimCLR, SimSiam, MoCo,...

- Invariant representation
- Object identity only



Ours

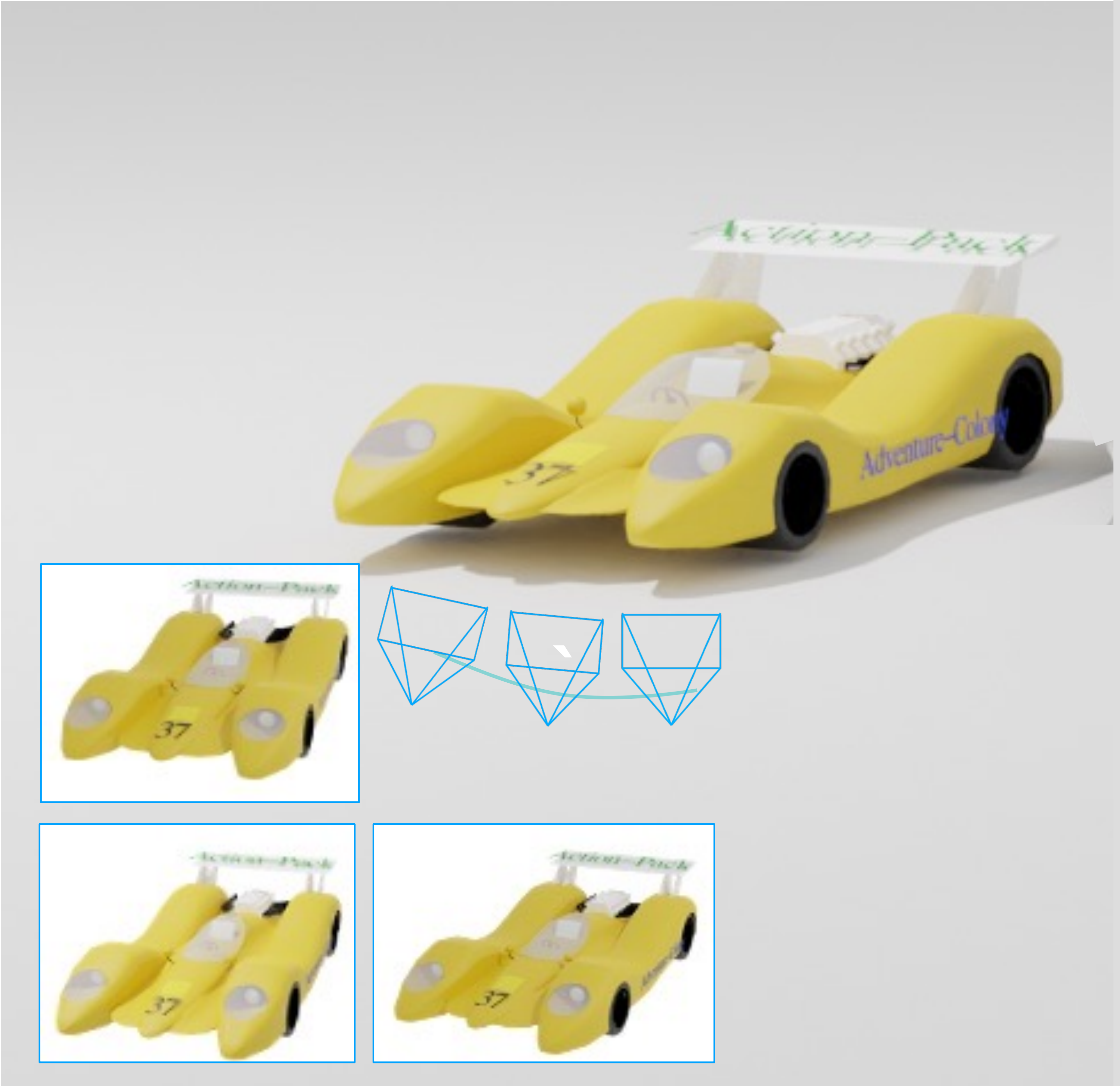
- Pose-aware representation
- Object identity + pose



Why pose-aware SSL is hard?

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

Benchmark: Training Data



* Synthetic data from ShapeNet

Benchmark: Training Data

Action-Pack
Adventure-Colony
37

Benchmark: Training Data

Action-Pack
Adventure-Colony
37

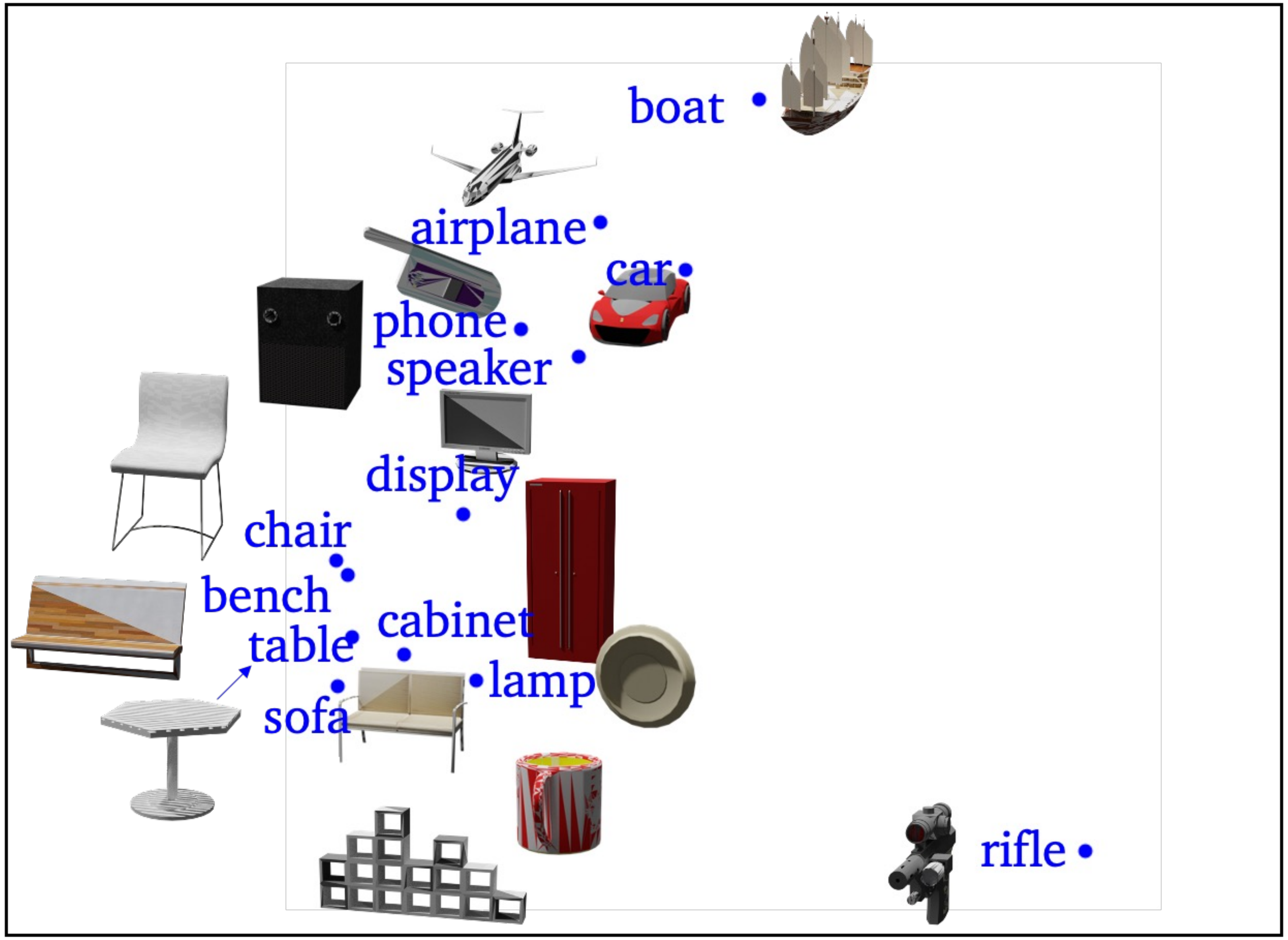
Benchmark: Training Data

Action-Pack
Adventure-Colony
37

Training data: **Image triplets** with small pose changes
No semantic or pose **labels.**

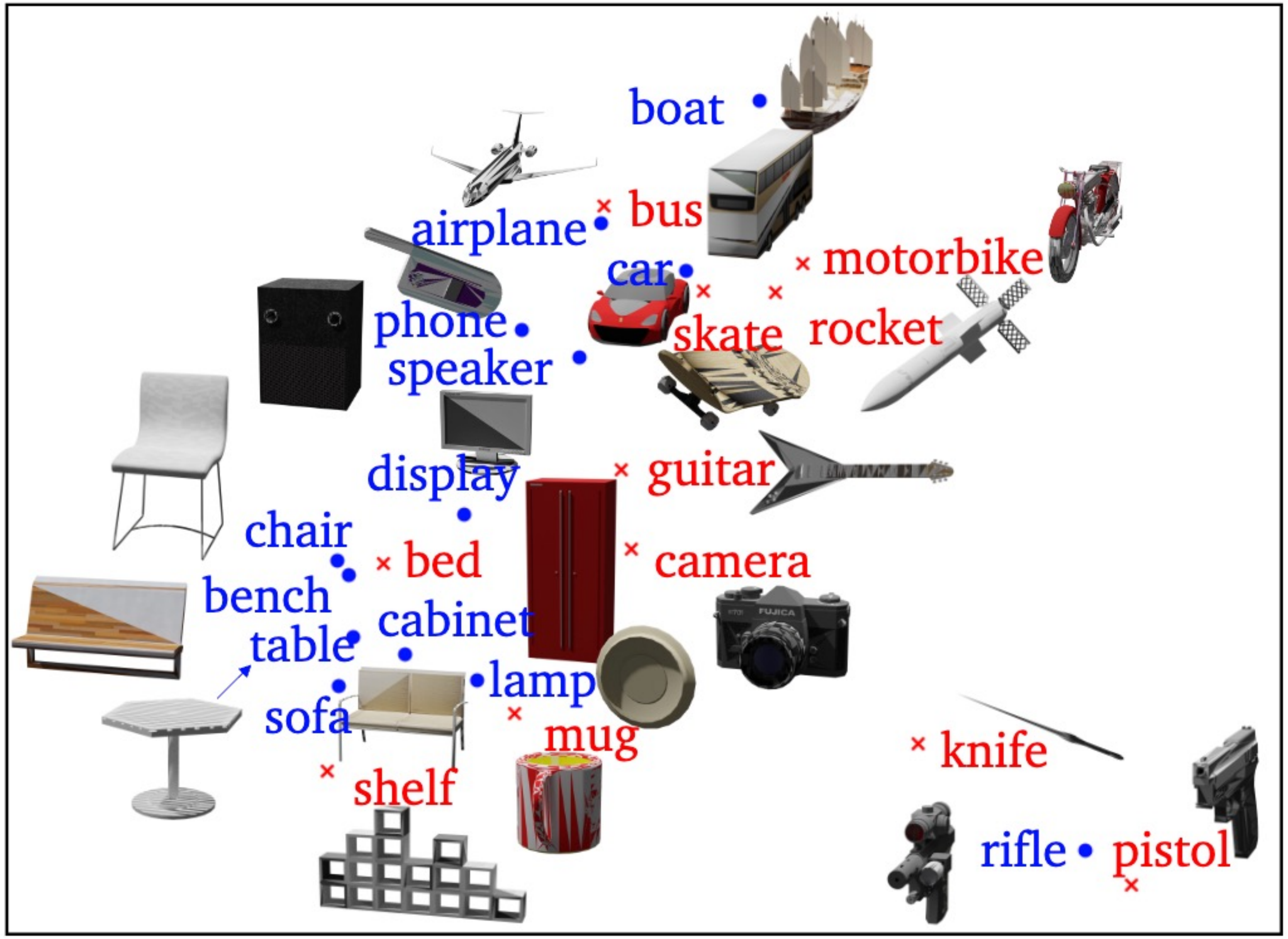
Benchmark: Data Configuration

13 in-domain

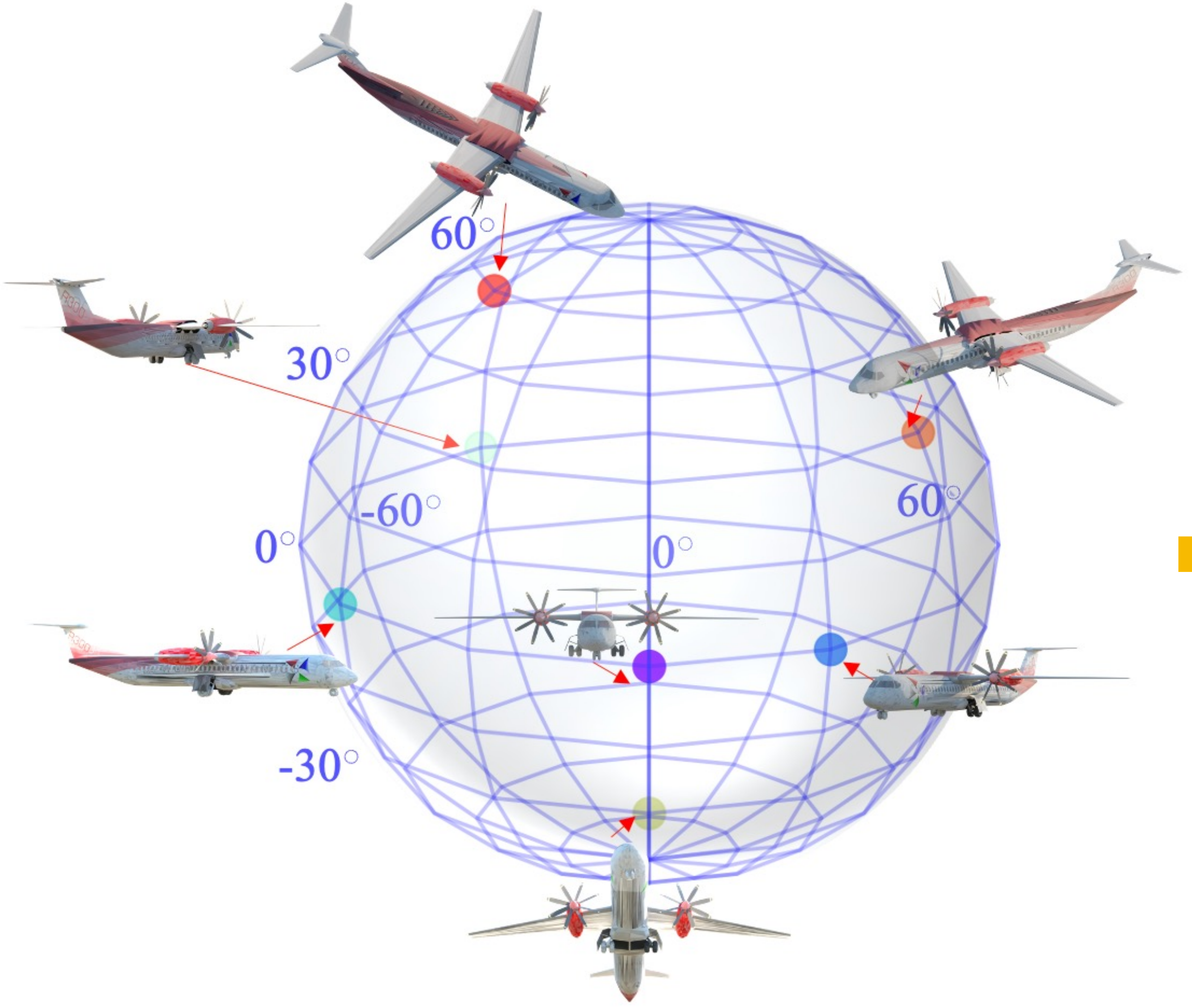


Benchmark: Data Configuration

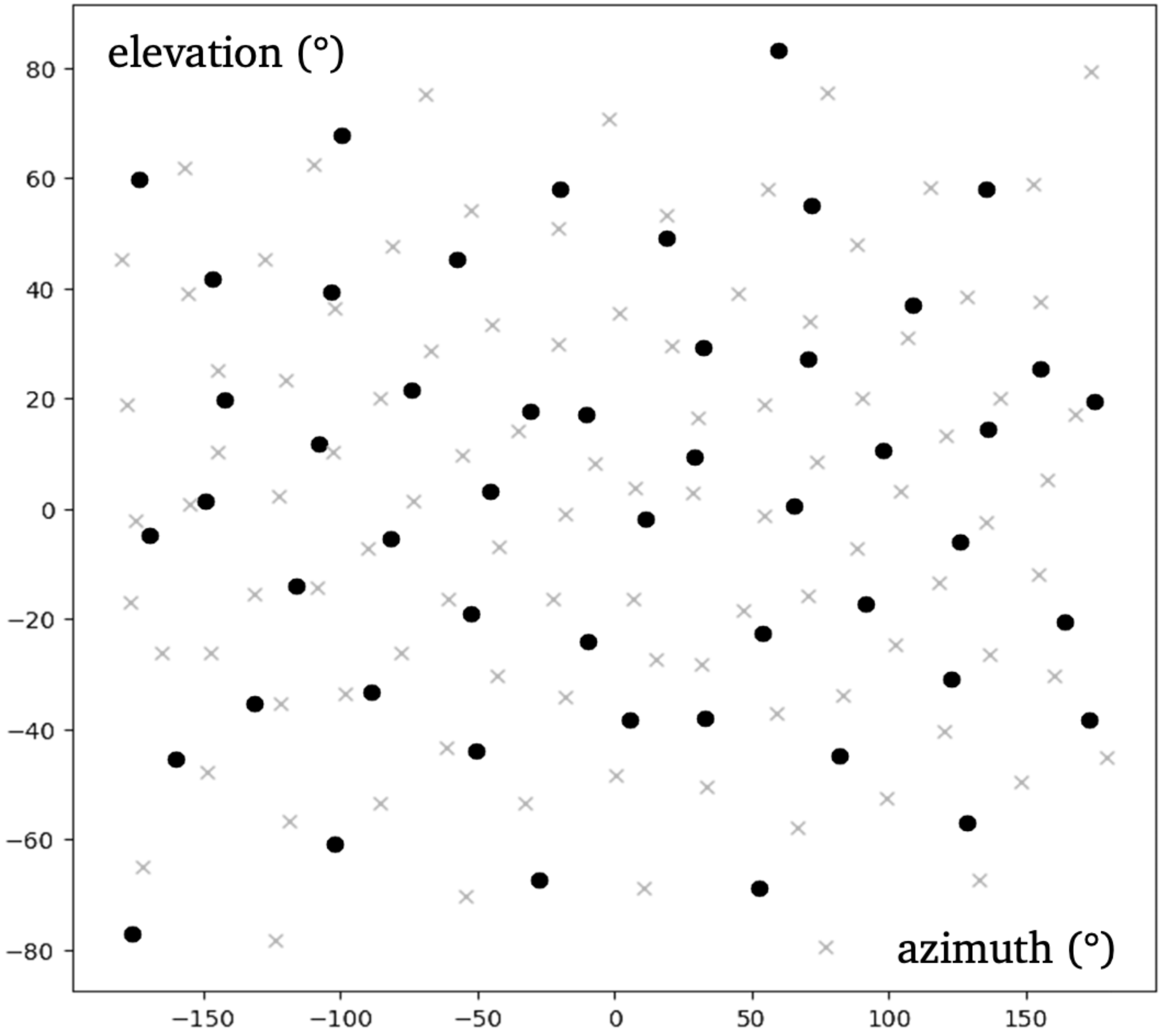
13 in-domain & 20 out-of-domain semantic categories



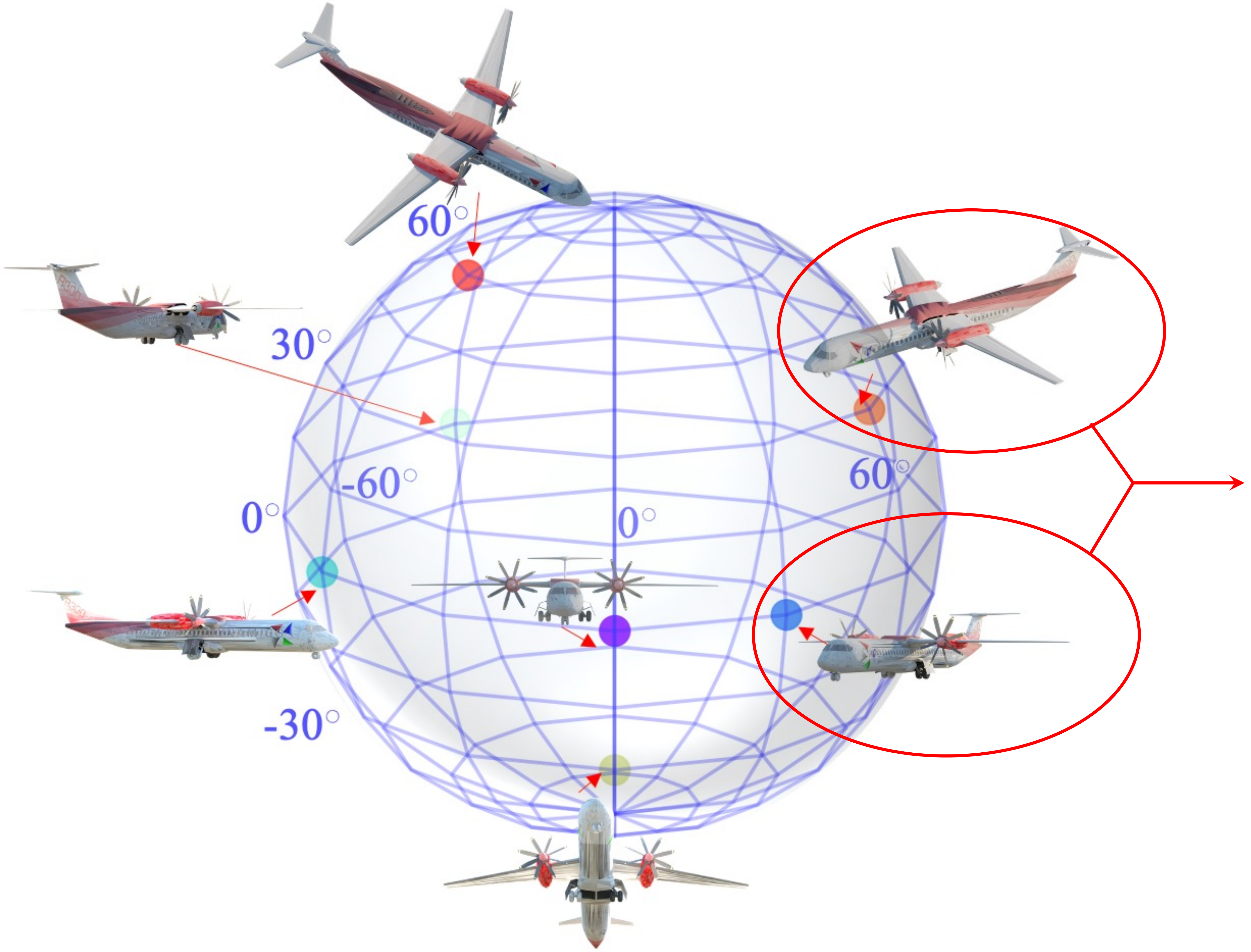
Benchmark: Data Configuration



in-domain & out-of-domain pose

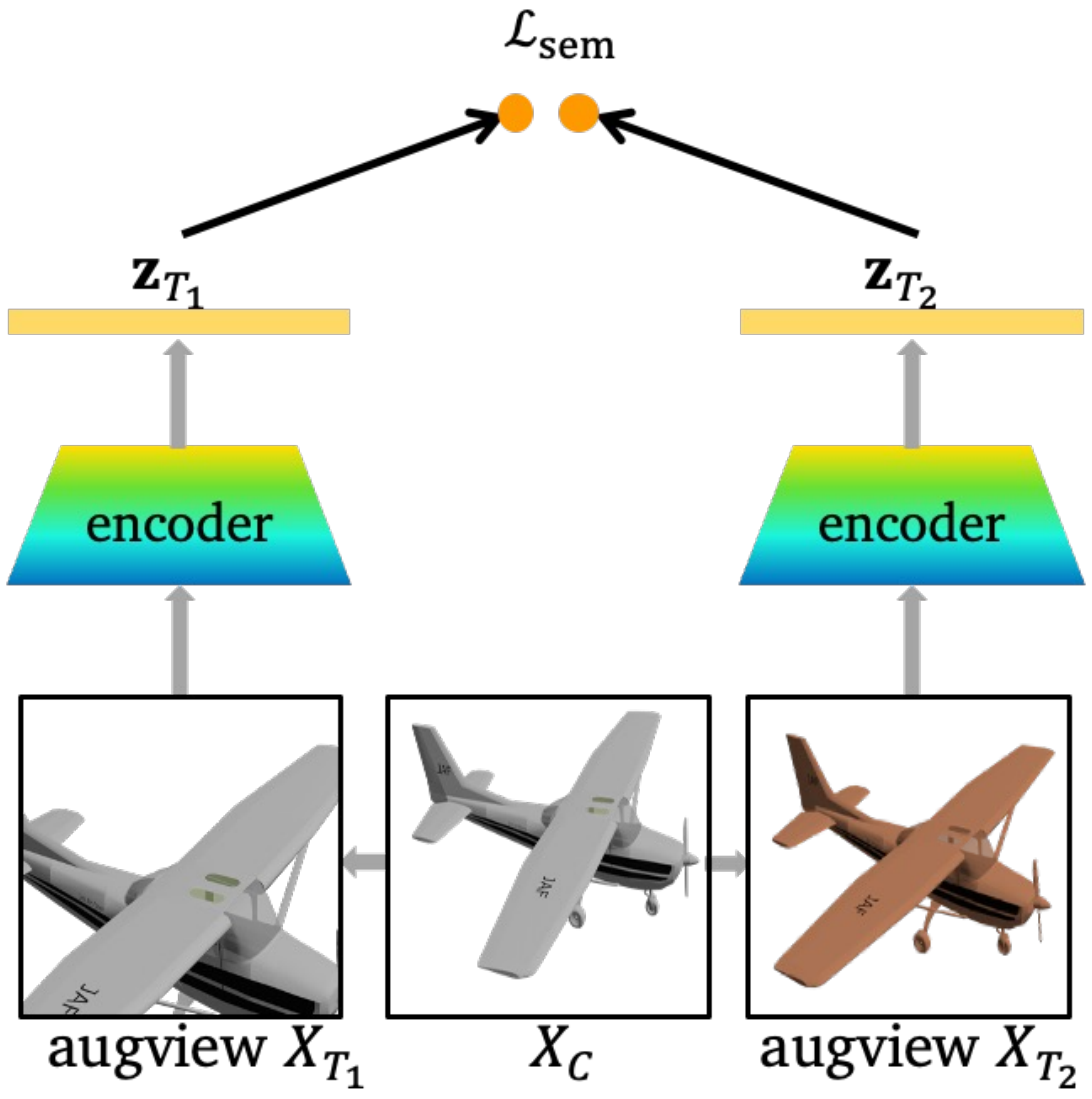


Benchmark: Tasks



- Semantic classification
- Absolute pose → how SSL learns global pose from local adjacent pose
- Relative pose → how SSL generalizes to OOD class/pose
  - Category-specific pose free
  - Generalize to open categories

SSL Training: Invariance



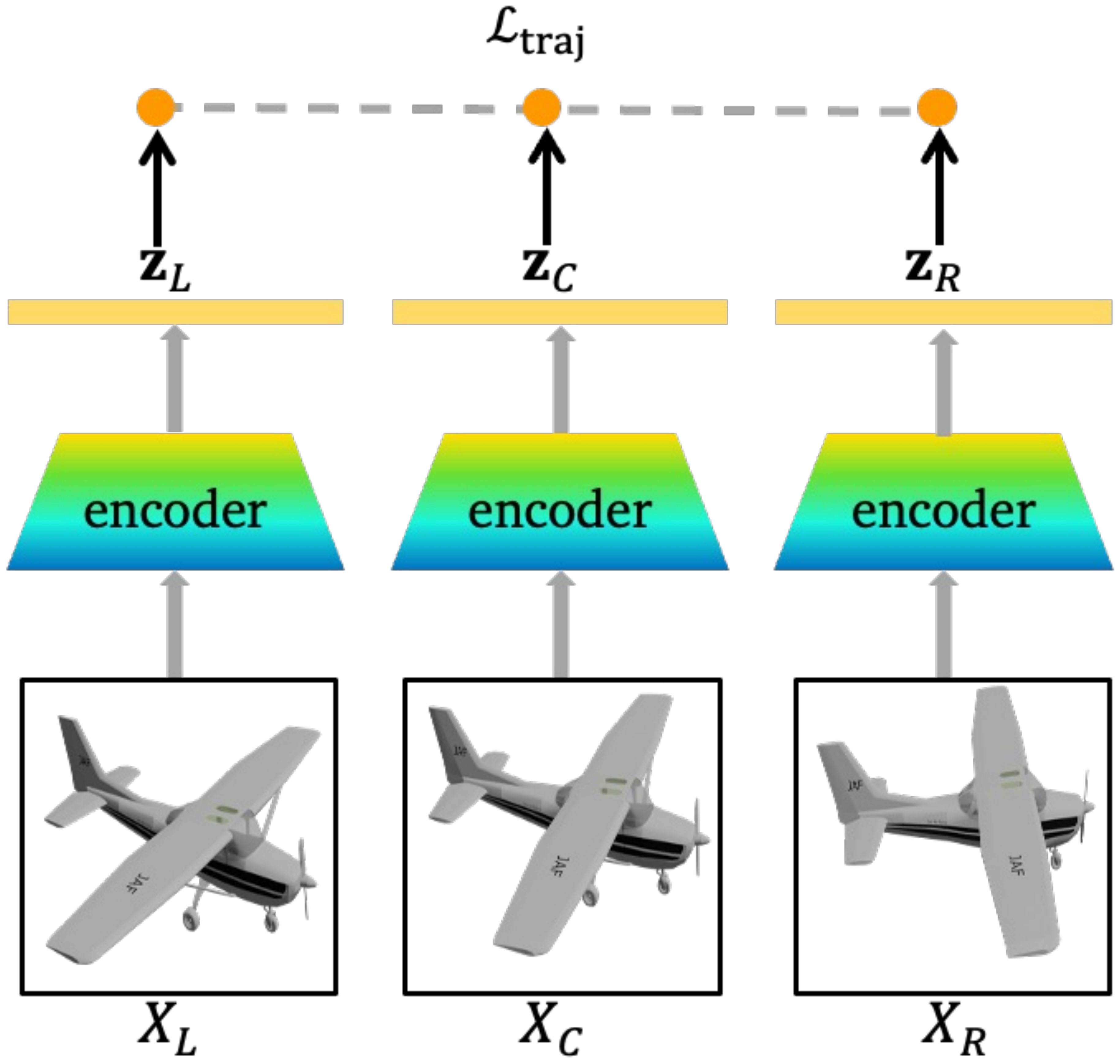
Example: VICReg

Augmentations:

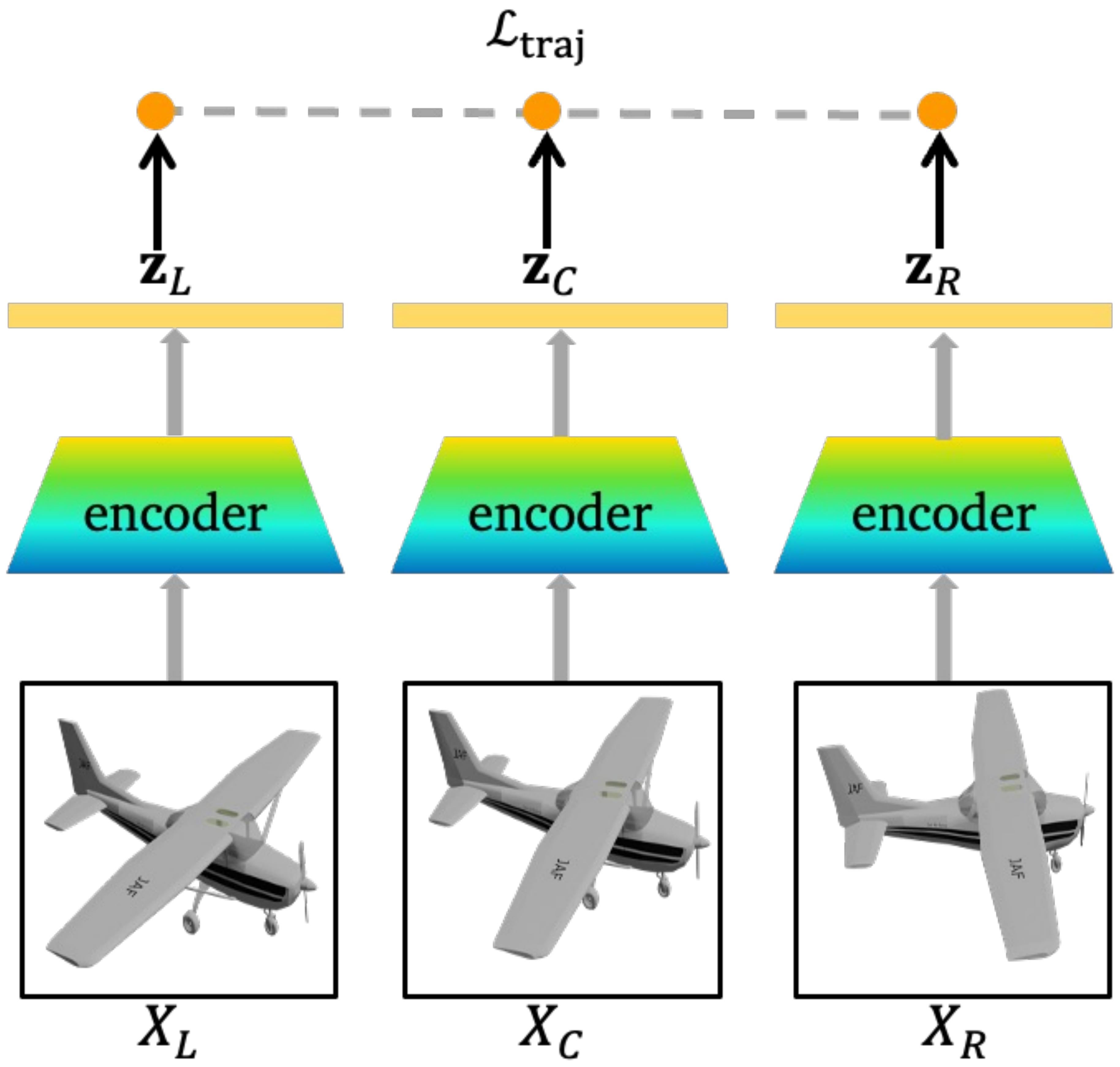
- Random crops
- Color jittering
- Gaussian Blur

VICReg: Variance-invariance-covariance regularization for self-supervised learning. ICLR 2022

SSL Training: Trajectory Regularization

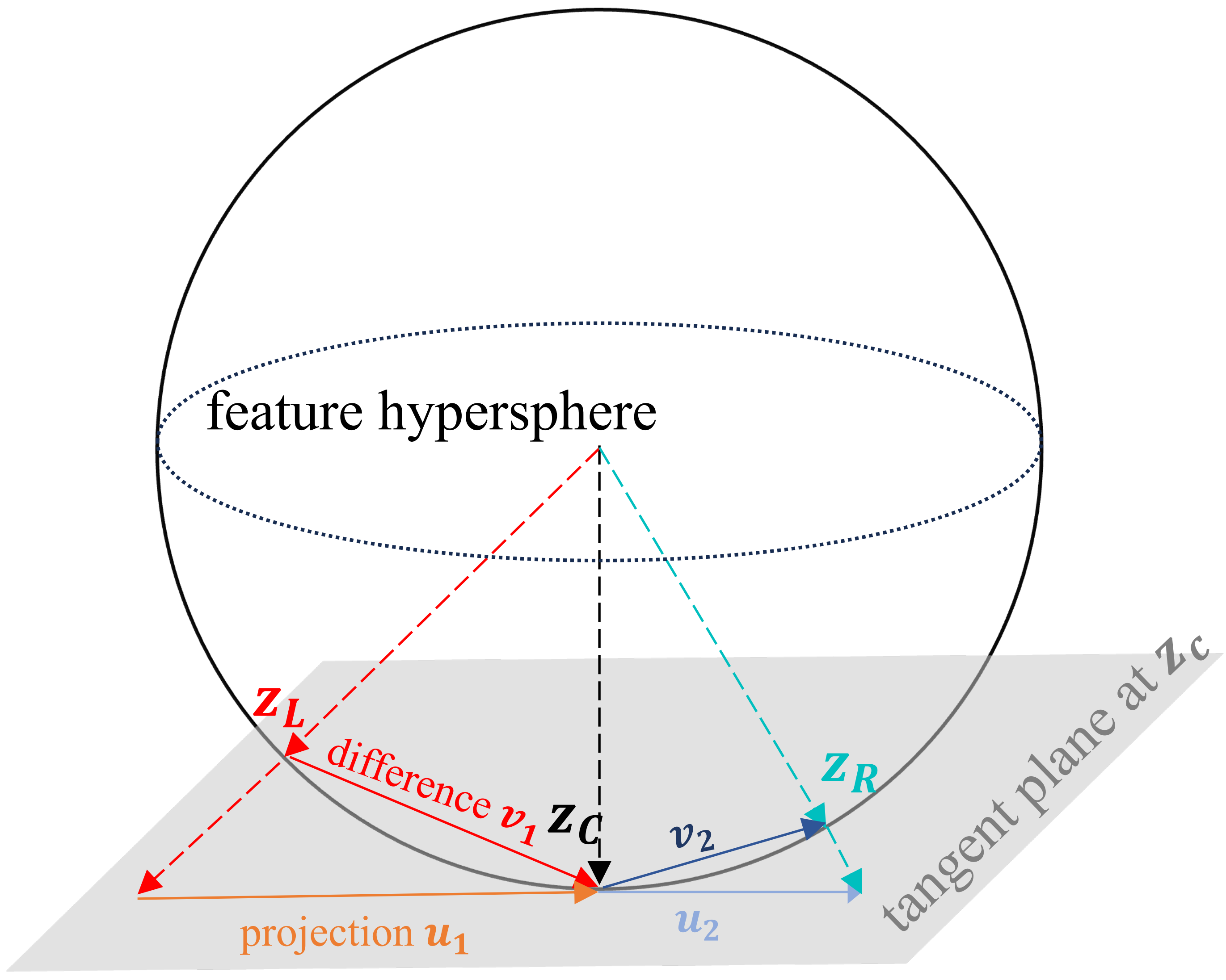


SSL Training: Trajectory Regularization

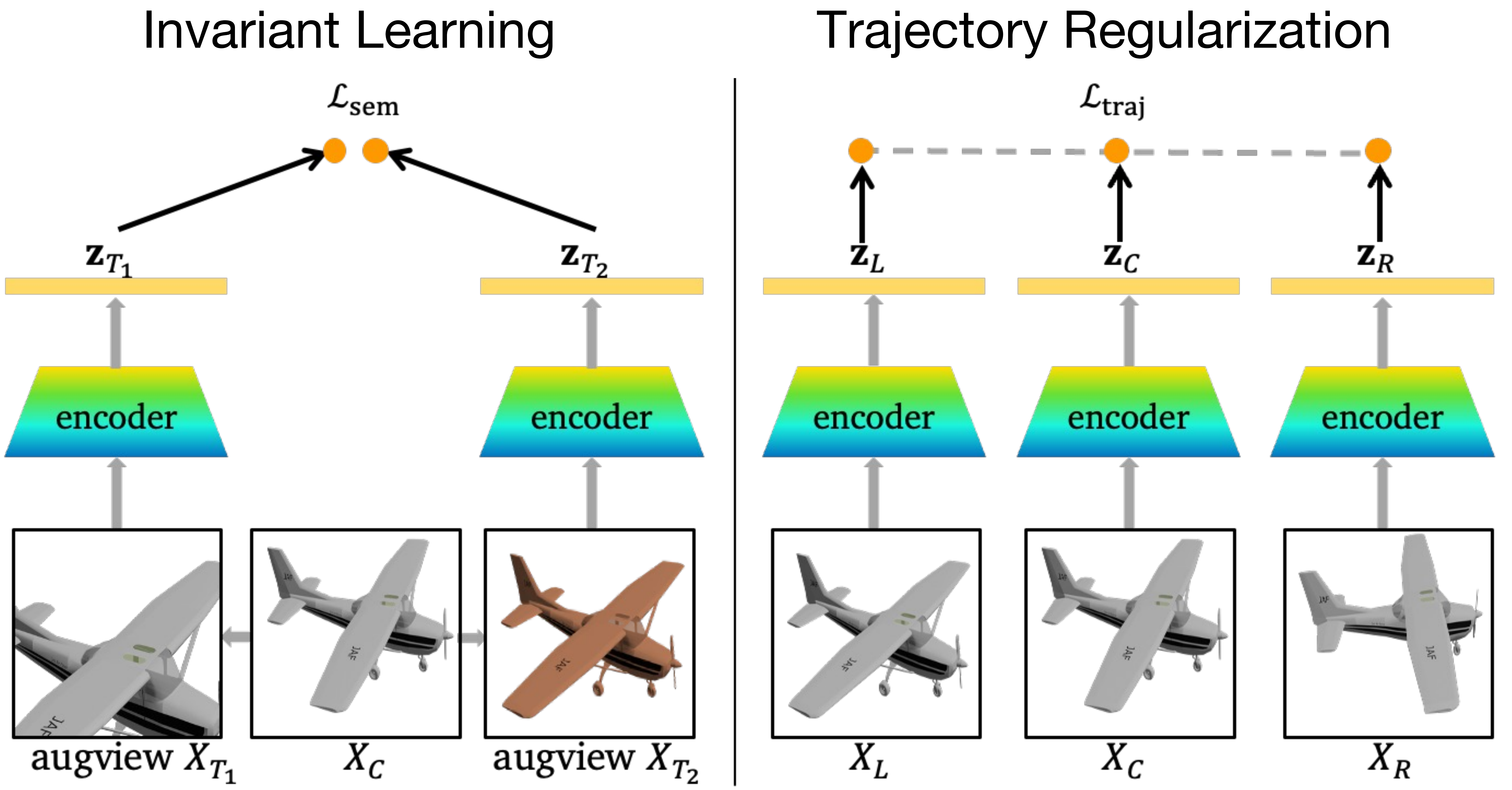


Line up 3 embeddings via trajectory regularization:

$$\mathcal{L}_{traj}(\mathbf{z}_L,\mathbf{z}_C,\mathbf{z}_R) = -\frac{\mathbf{u}_1\cdot\mathbf{u}_2}{\|\mathbf{u}_1\|\|\mathbf{u}_2\|}$$



SSL Training

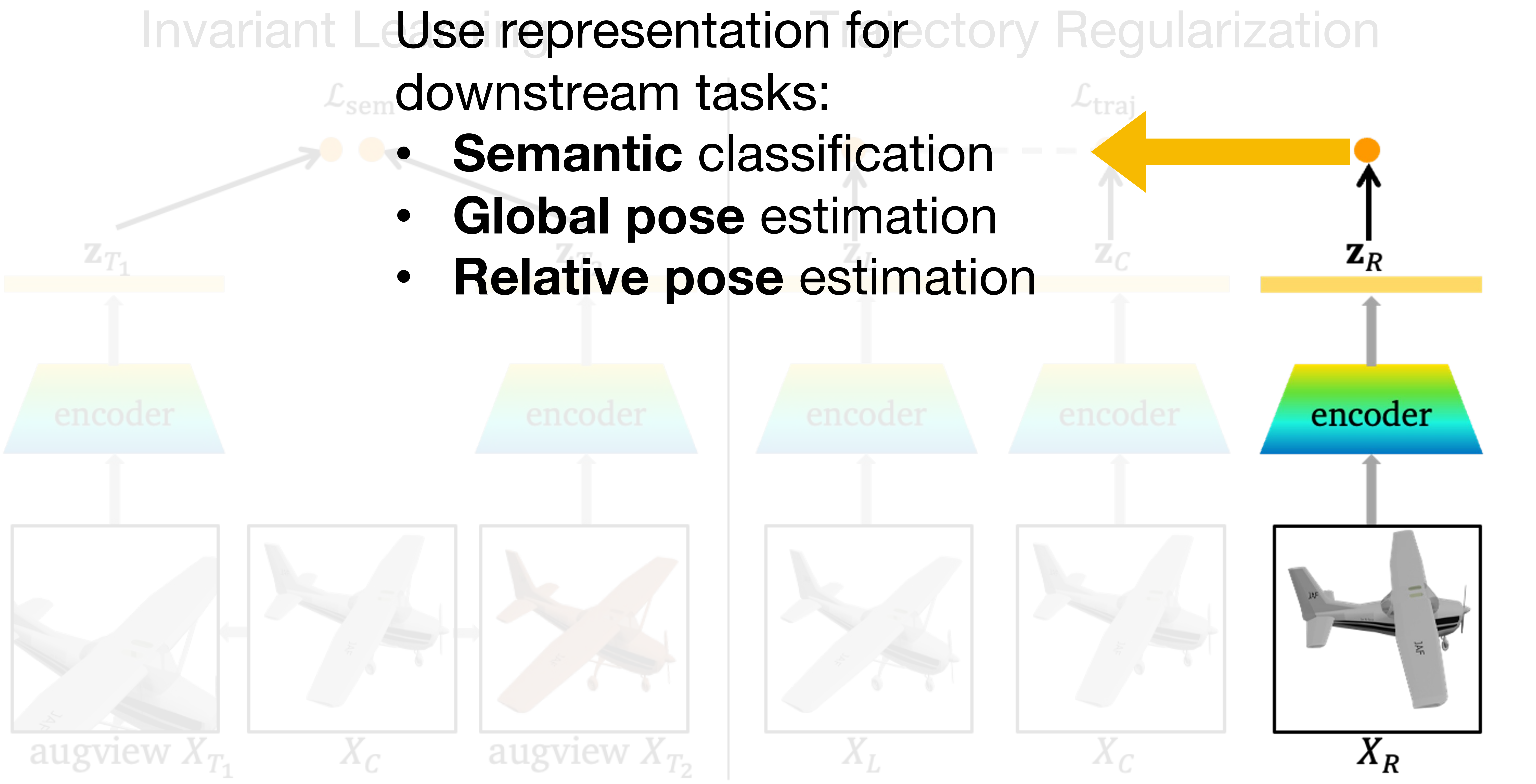


Final loss is a combination: $\mathcal{L} = \mathcal{L}_{sem}(\mathbf{z}_{T_1}, \mathbf{z}_{T_2}) + \lambda\mathcal{L}_{traj}(\mathbf{z}_L, \mathbf{z}_C, \mathbf{z}_R)$

After SSL Training: Probing

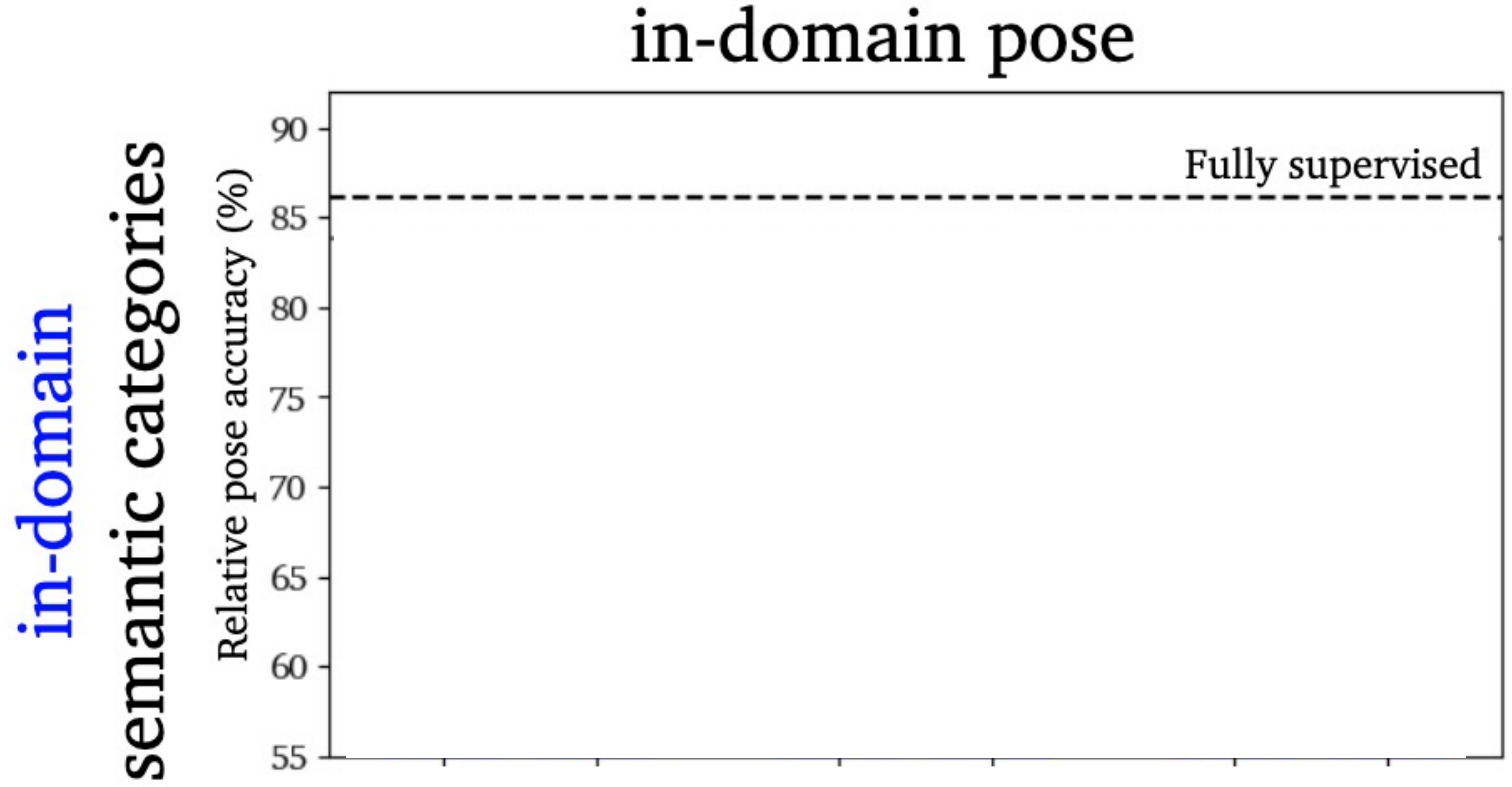
Use representation for downstream tasks:

- **Semantic** classification
- **Global pose** estimation
- **Relative pose** estimation

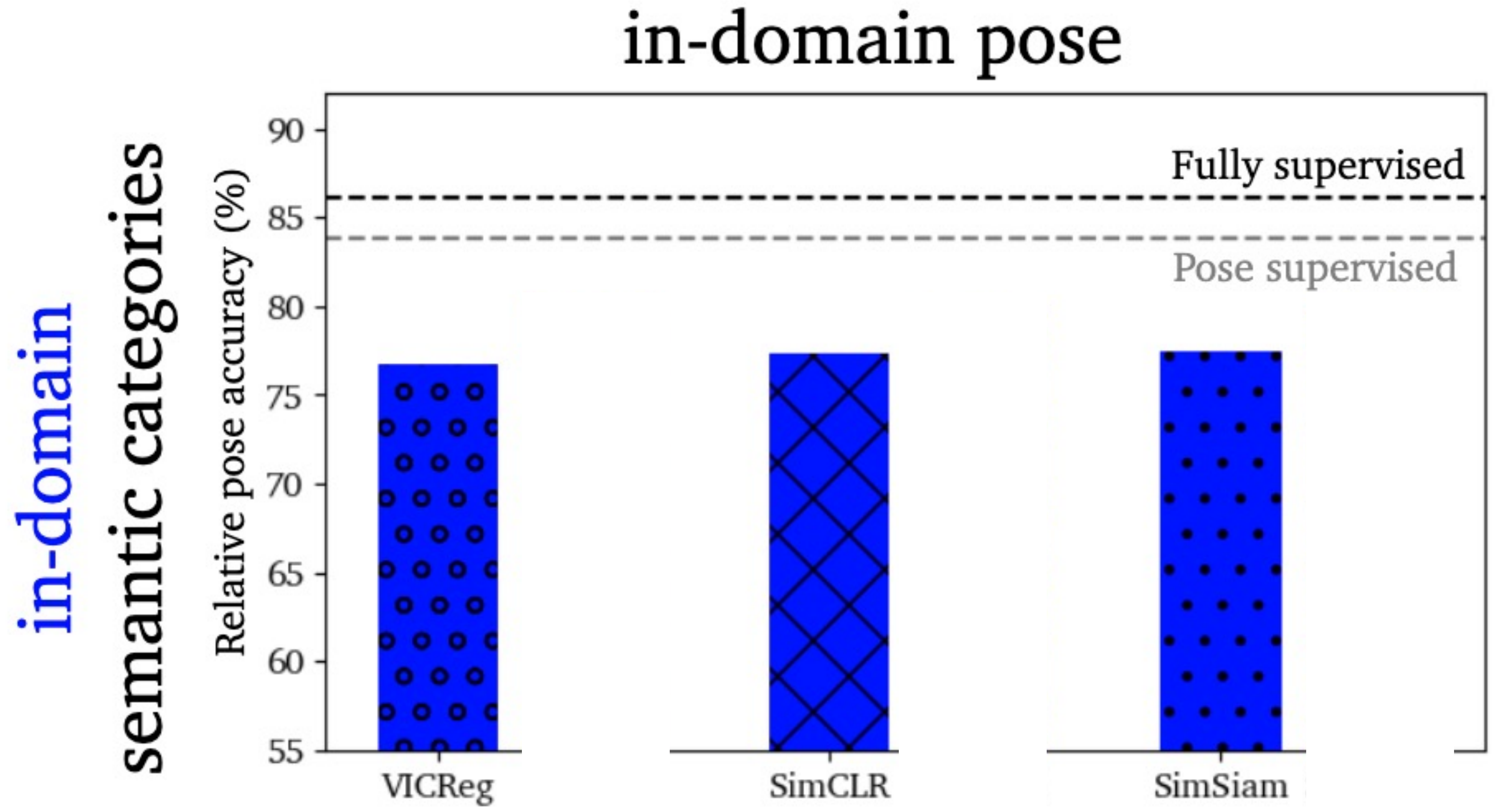


Final loss is a combination: $\mathcal{L} = \mathcal{L}_{sem}(\mathbf{z}_{T_1}, \mathbf{z}_{T_2}) + \lambda \mathcal{L}_{traj}(\mathbf{z}_L, \mathbf{z}_C, \mathbf{z}_R)$

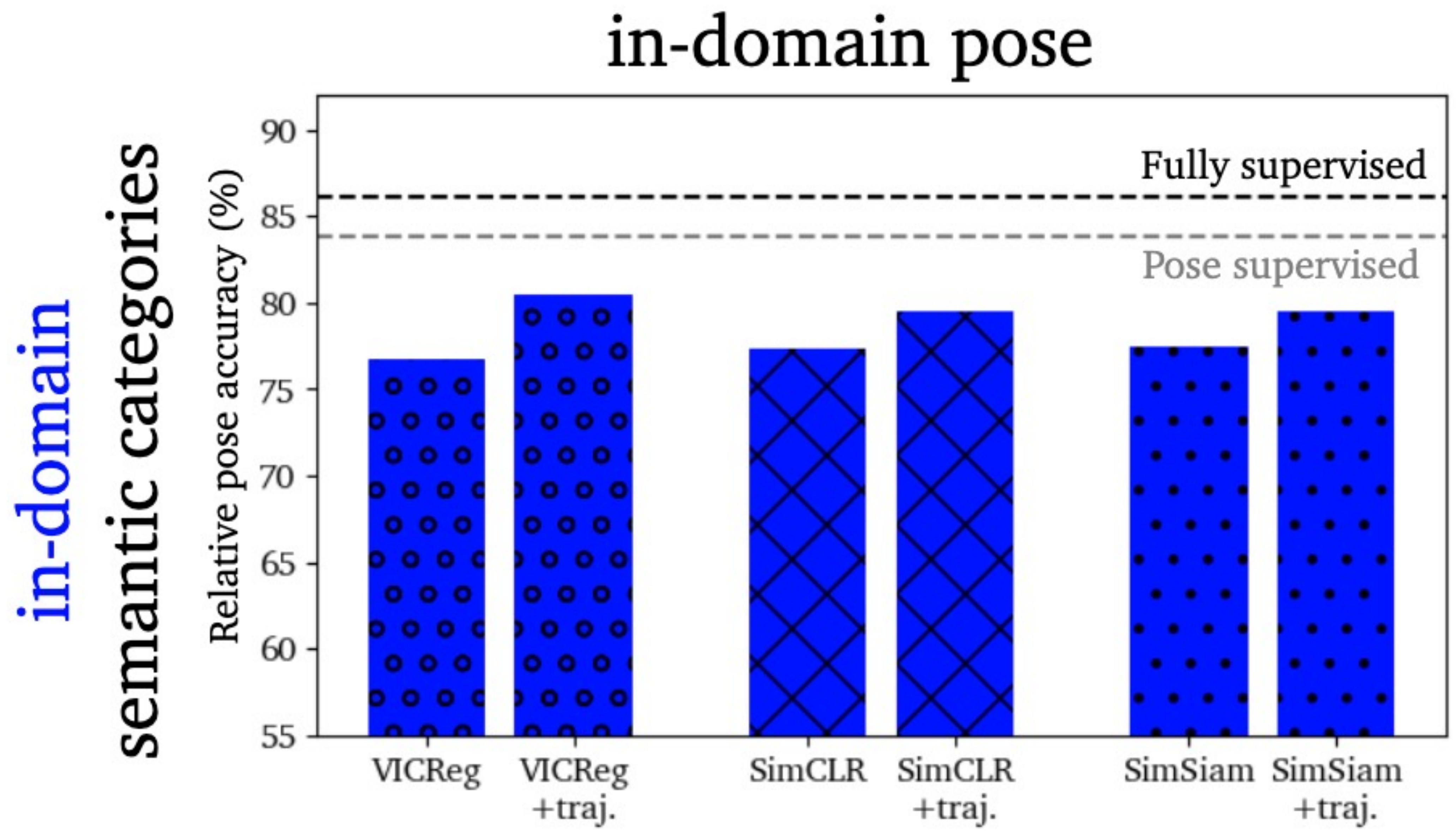
In-Domain and Out-of-Domain Pose Accuracy



In-Domain and Out-of-Domain Pose Accuracy

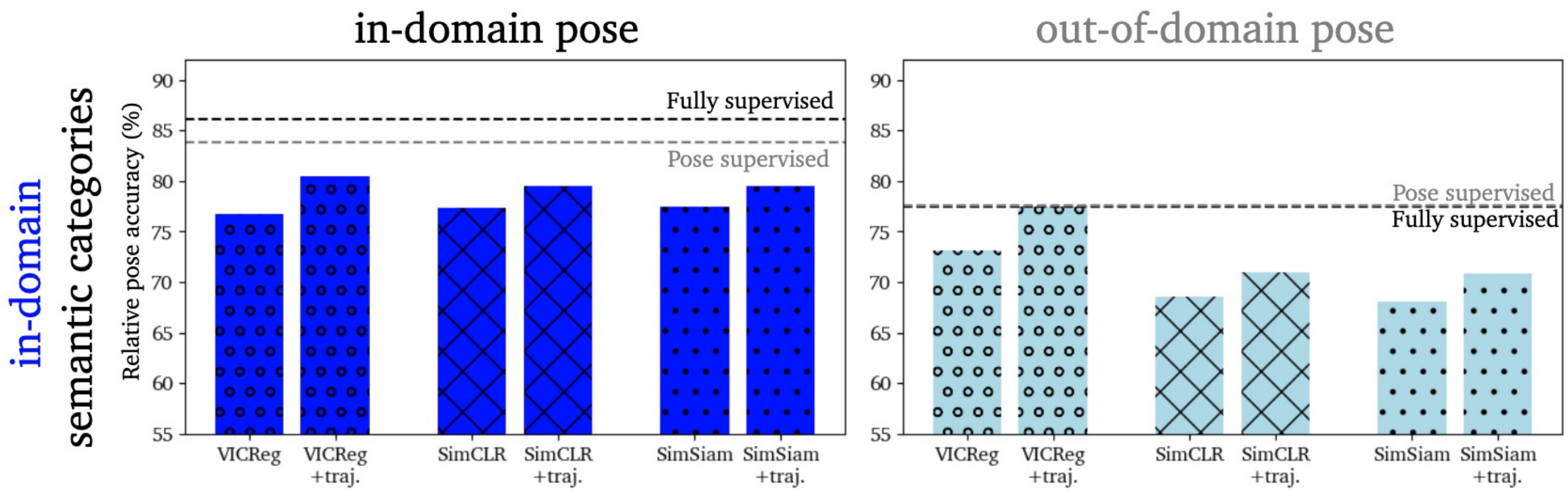


In-Domain and Out-of-Domain Pose Accuracy



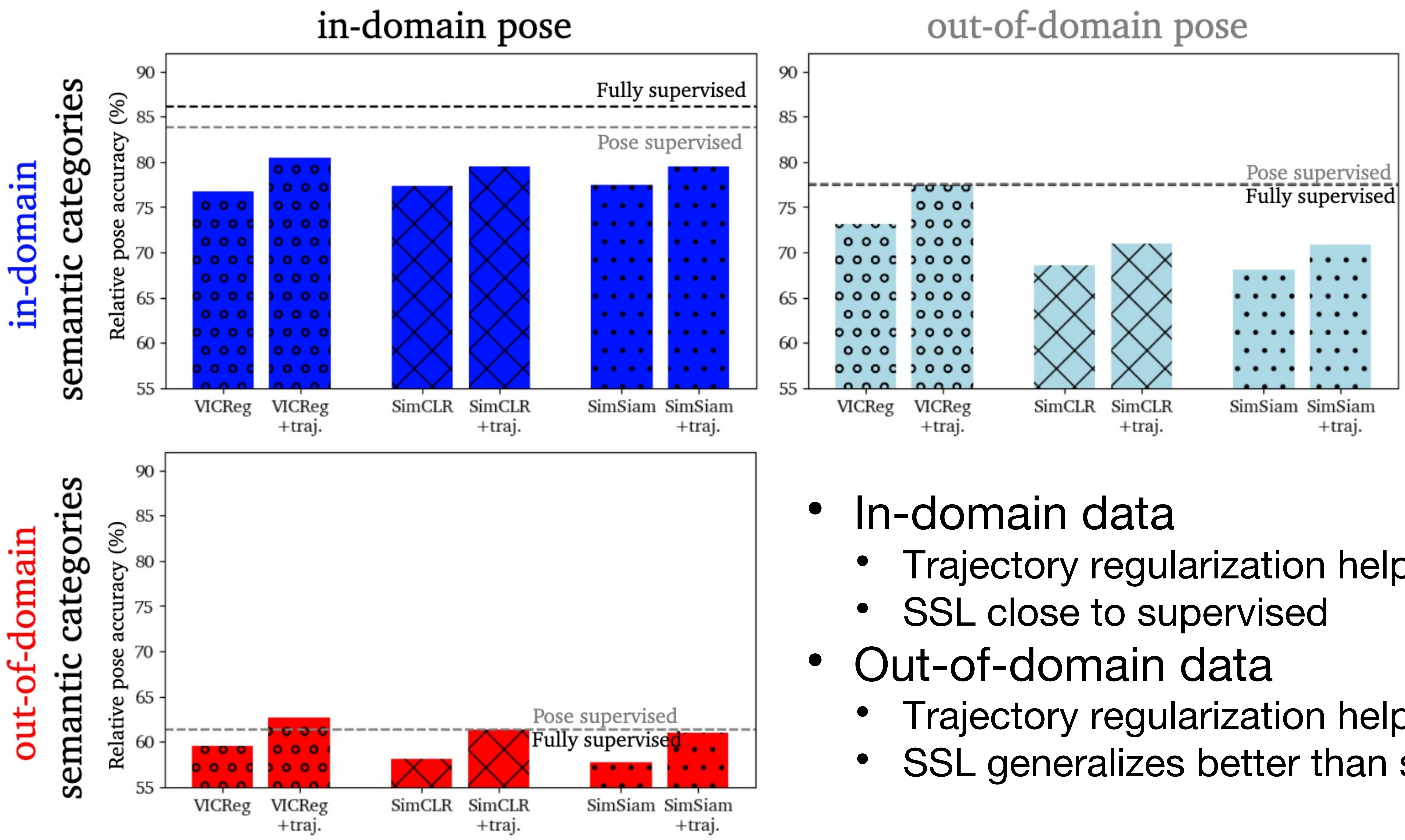
- In-domain data
  - Trajectory regularization helps
  - SSL close to supervised

In-Domain and Out-of-Domain Pose Accuracy



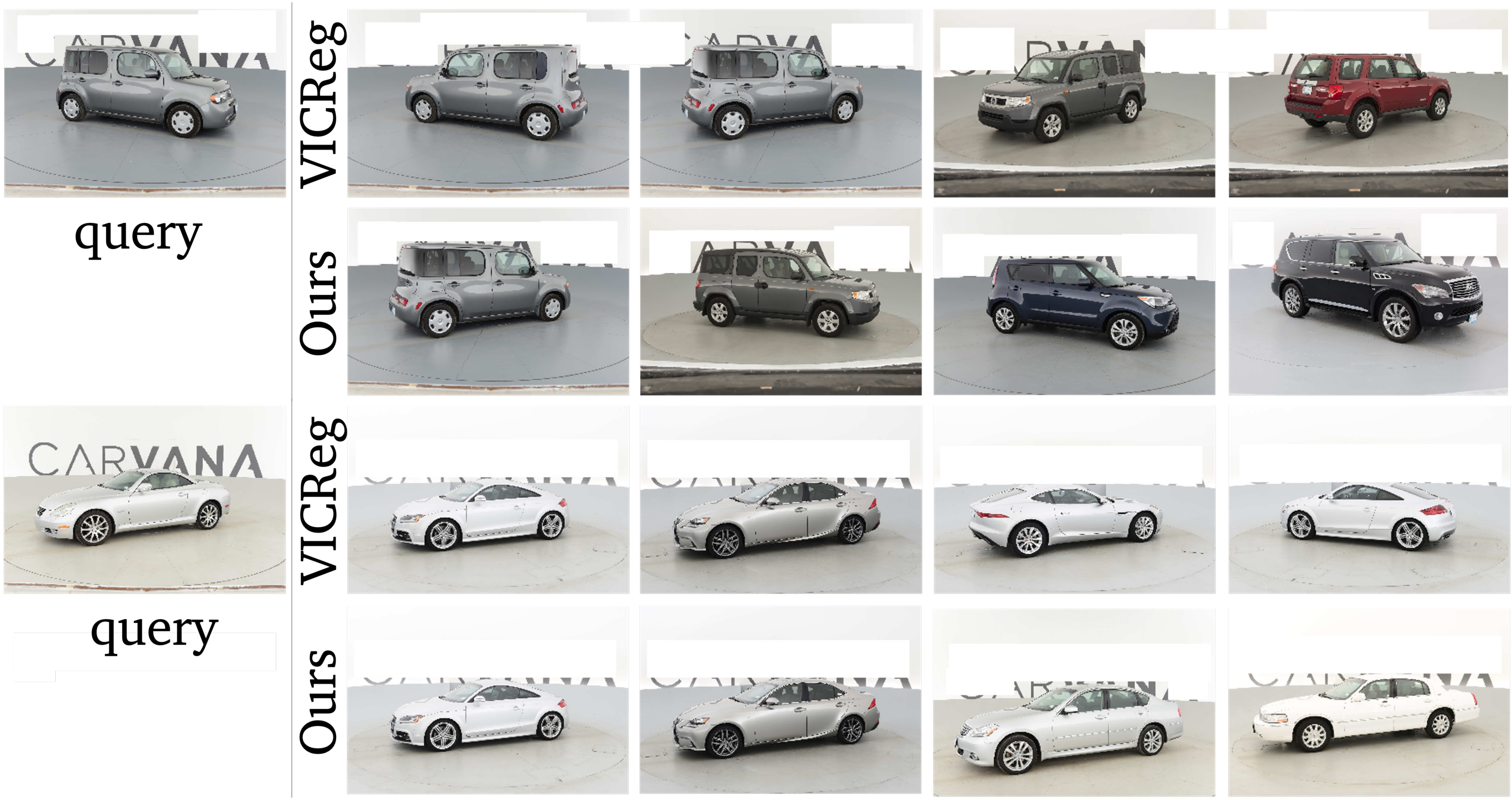
- In-domain data
  - Trajectory regularization helps
  - SSL close to supervised
- Out-of-domain data
  - Trajectory regularization helps
  - SSL generalizes better than supervised

In-Domain and Out-of-Domain Pose Accuracy



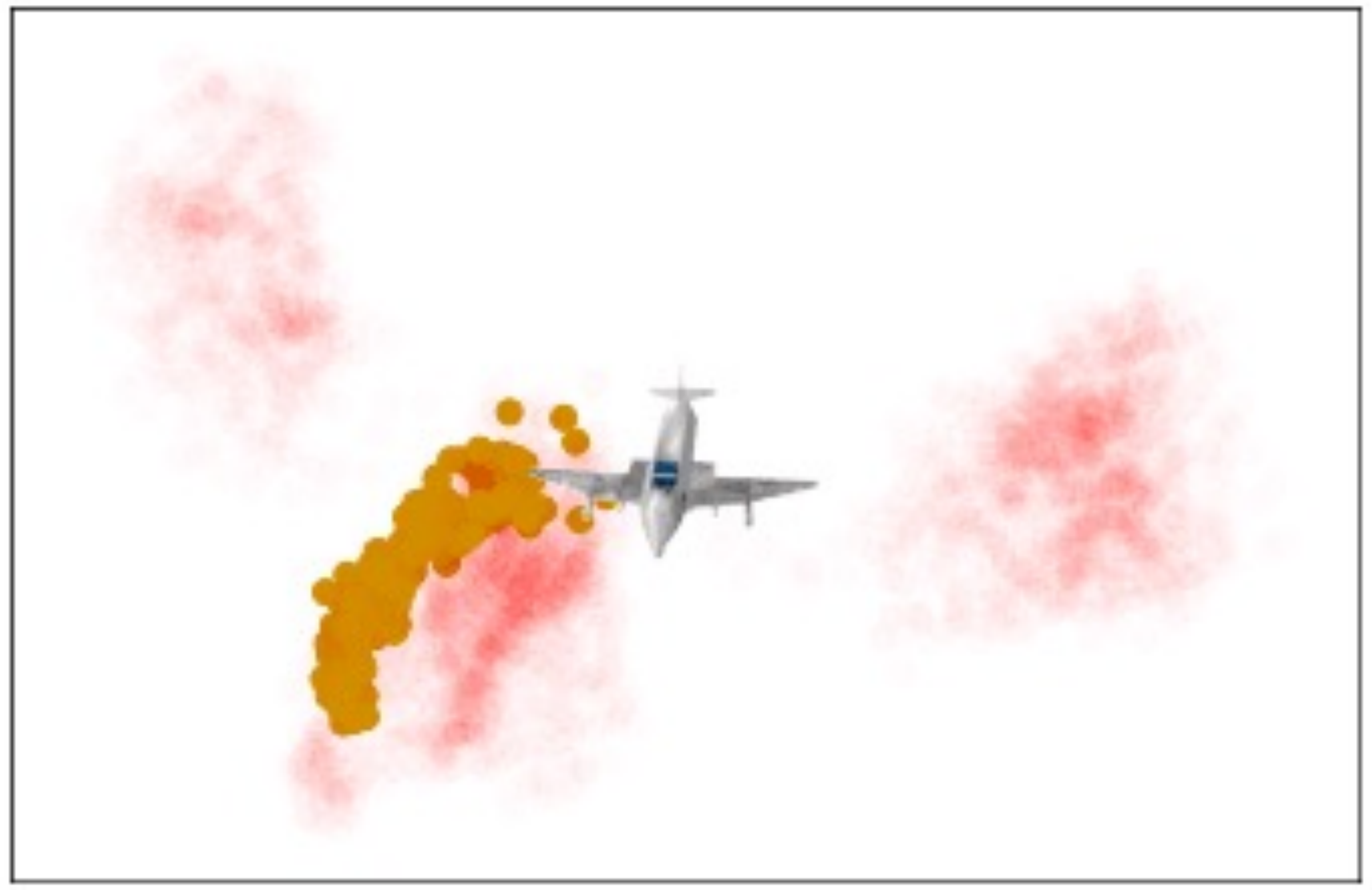
- In-domain data
  - Trajectory regularization helps
  - SSL close to supervised
- Out-of-domain data
  - Trajectory regularization helps
  - SSL generalizes better than supervised

Direct Evaluation on Real Data

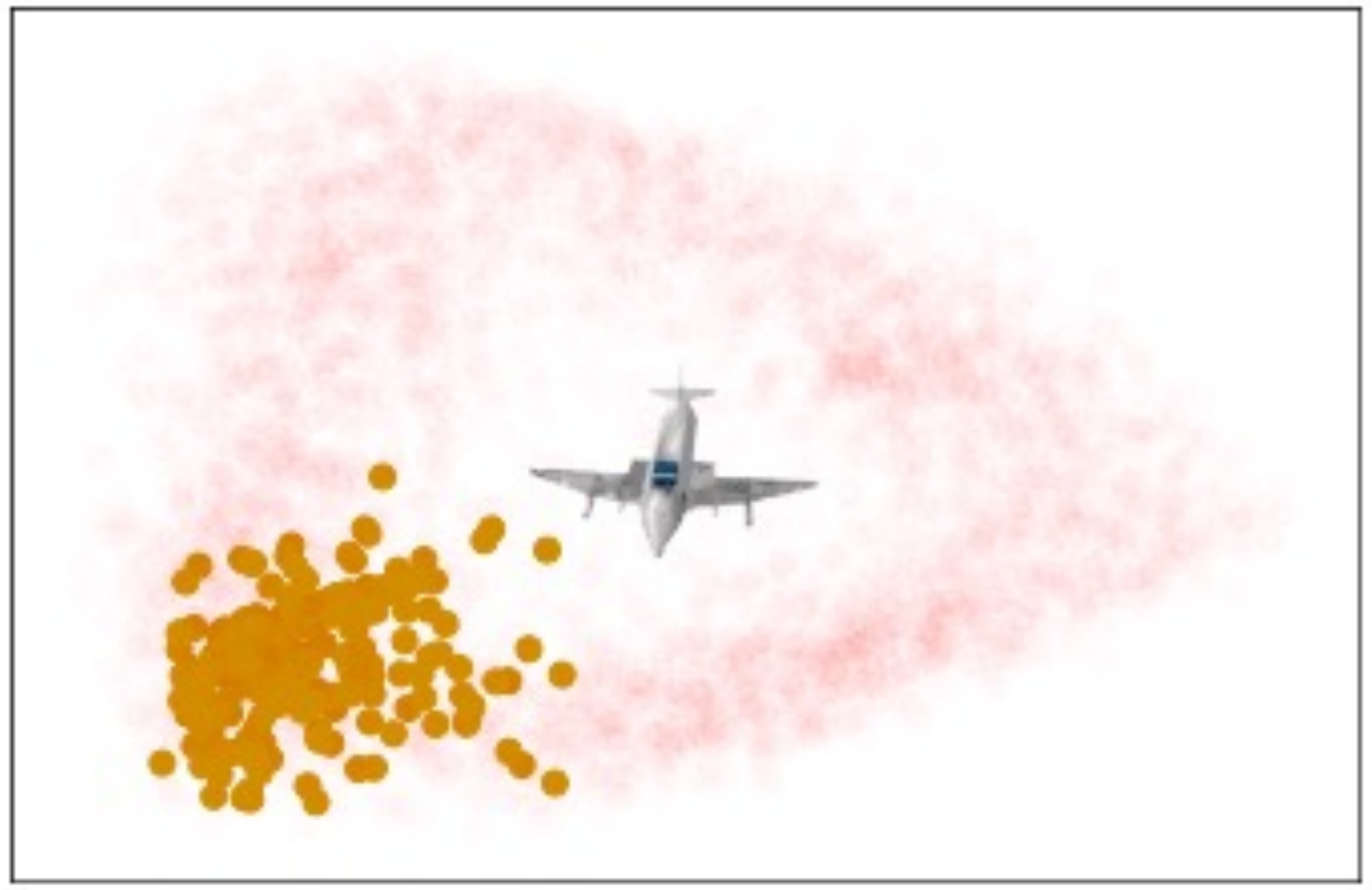


Rotating car dataset CARVANA

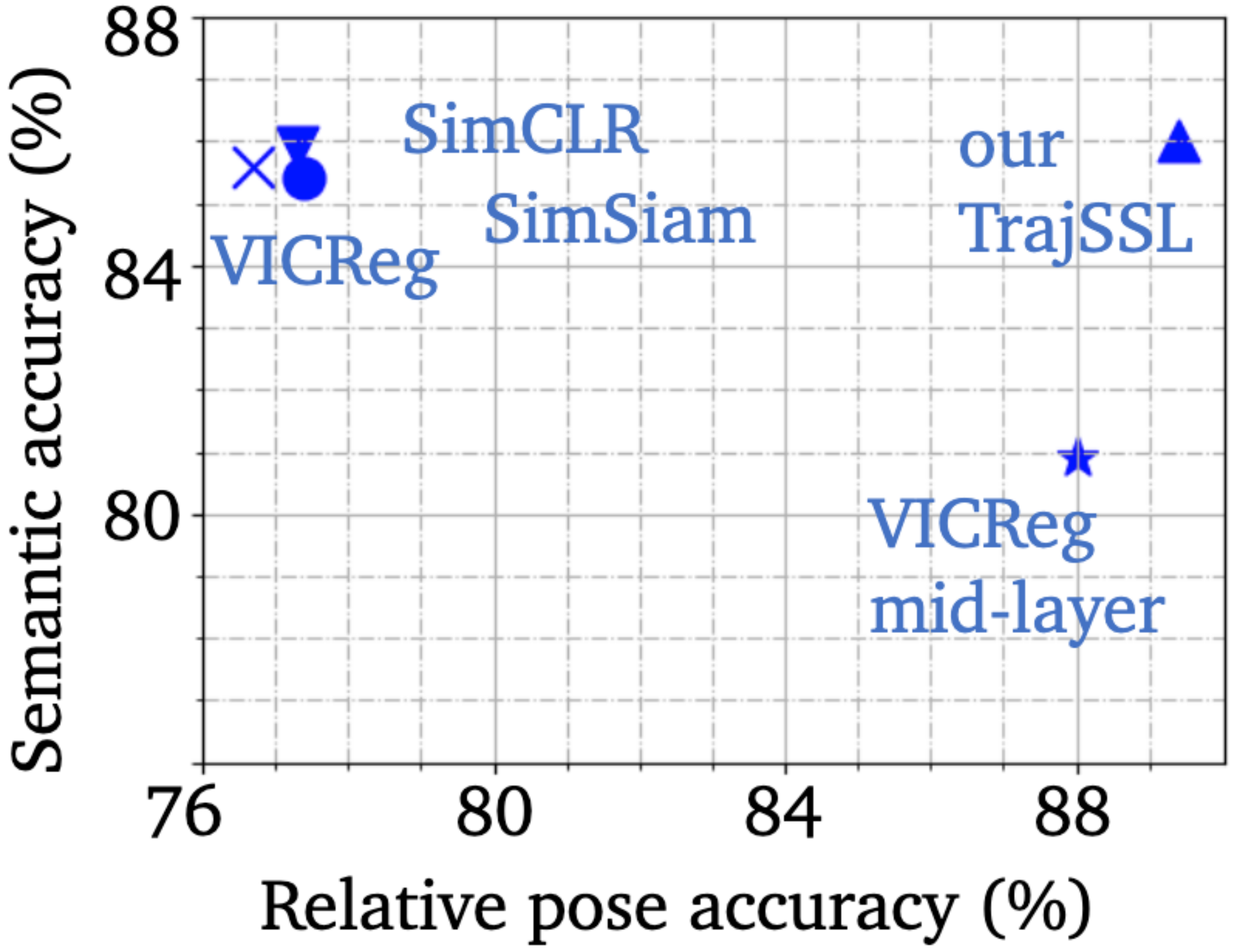
Visualizing Representation



VICReg

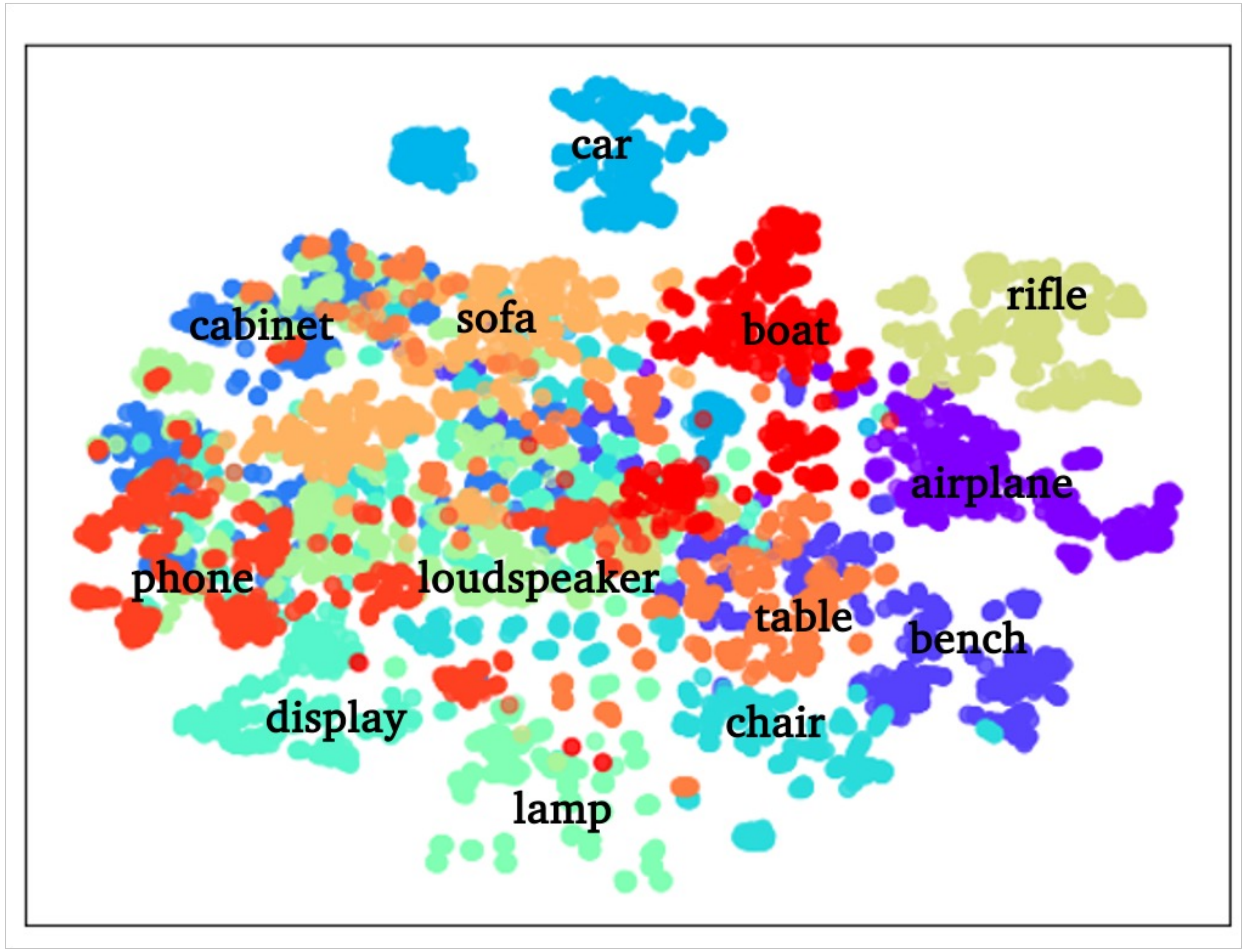


VICReg
+trajectory regularization



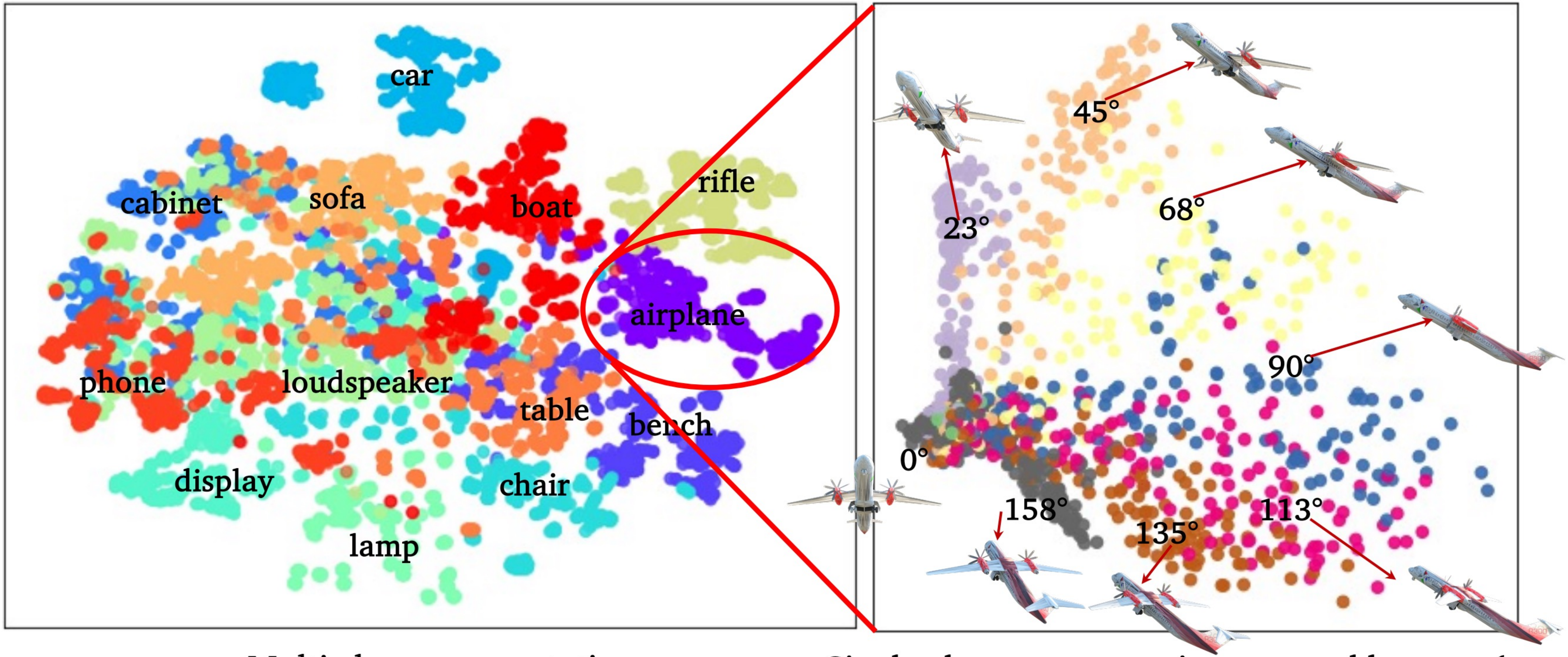
VICReg: Variance-invariance-covariance regularization for self-supervised learning. ICLR 2022

Visualizing Representation



Multi-class representation

Visualizing Representation



Multi-class representation

Single-class representation grouped by pose (aero)

Emergent pose-semantic representation without labels!

Pose-Aware Self-Supervised Learning with Viewpoint Trajectory Regularization

Thank you!

Please come to our poster: #256

Paper/code/data: