

What supervision does the method use?

Setting	Available during training	Reward
Image-pair correspondence	Image pairs and source points	VLM semantic score
Named keypoints	Rendered images, part names, visibility-based query selection	VLM scores of eight markers
GT-reward control	Ground-truth target coordinates	Ground-truth correctness

Pretrained visual/text encoders carry prior supervision.

“Without labels” refers to the target-coordinate reward. Ground truth still supports evaluation and checkpoint selection.

Is REINFORCE the only way to use feedback?

SPair-71k cat test · 600 pairs · Plain DINOv3 · Kimi K2.5 judge

Training objective	PCK@0.10 (%)
Frozen control	70.05
Top-K filtering	70.14 ± 0.07
Direct preference optimisation	71.49 ± 0.37
REINFORCE	71.46 ± 0.03

DPO and REINFORCE give similar gains in this comparison.

Same held-out test protocol. Mean ± standard deviation over three seeds.

CLIP embd. and Learned embd. answer different queries

Query representation	Exact training name	Reworded / unseen name
CLIP embd.	Frozen text encoder + learned head	Can form a query
Learned embd.	One trainable vector per name	No vector available
Hybrid	CLIP query + per-name residual	Falls back to CLIP query

All three query DINO image features. The judge receives the natural-language name.

Language enables new wording. It does not by itself guarantee transfer to new categories.