

Group meeting · ICLR submissions

Is It a Good Match?

2D Keypoint Matching with Feedback of a VLM

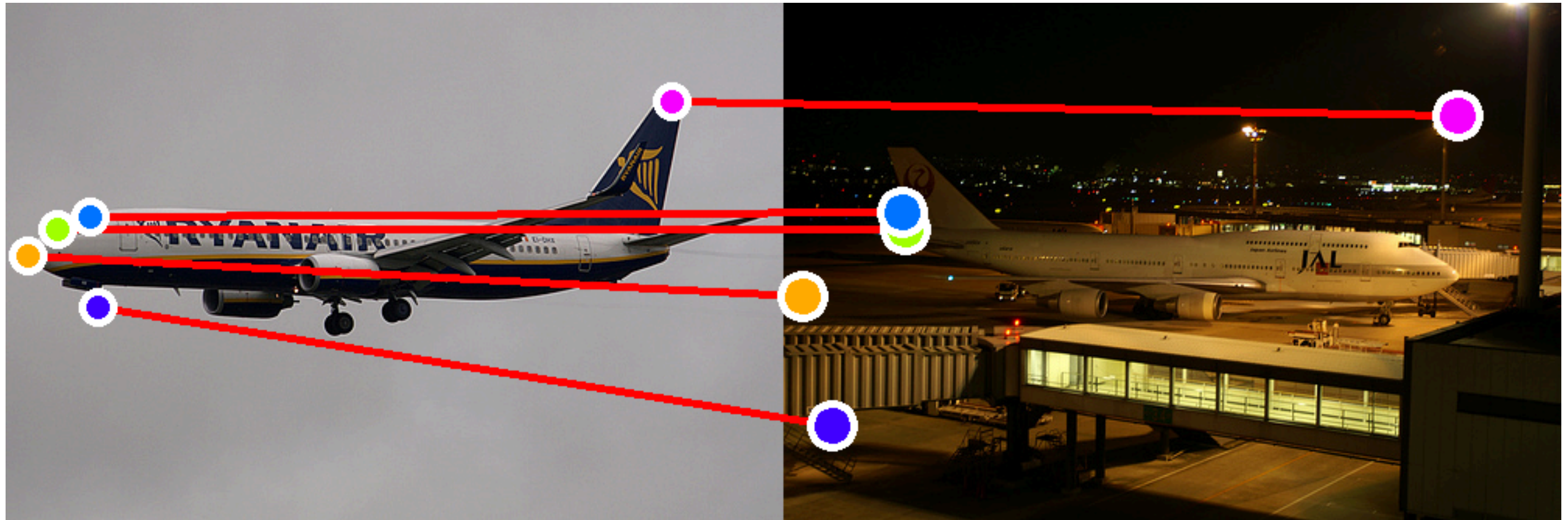
Bingchen Gong

with Viktoria Ehm, Emery Pierson and Maks Ovsjanikov

LIX, École Polytechnique

Match the same part across different objects

Given a source point, find its corresponding part in the target image.



A frozen DINOv3 example from SPair-71k. Red lines mark incorrect matches.

Appearance changes. The semantic part should stay the same.

Symmetry makes plausible matches wrong



One cat pair. Red: plain DINO pipeline. Blue: geometry-aware variant. Green: ground truth.

“The right object part” also requires the correct side and location.

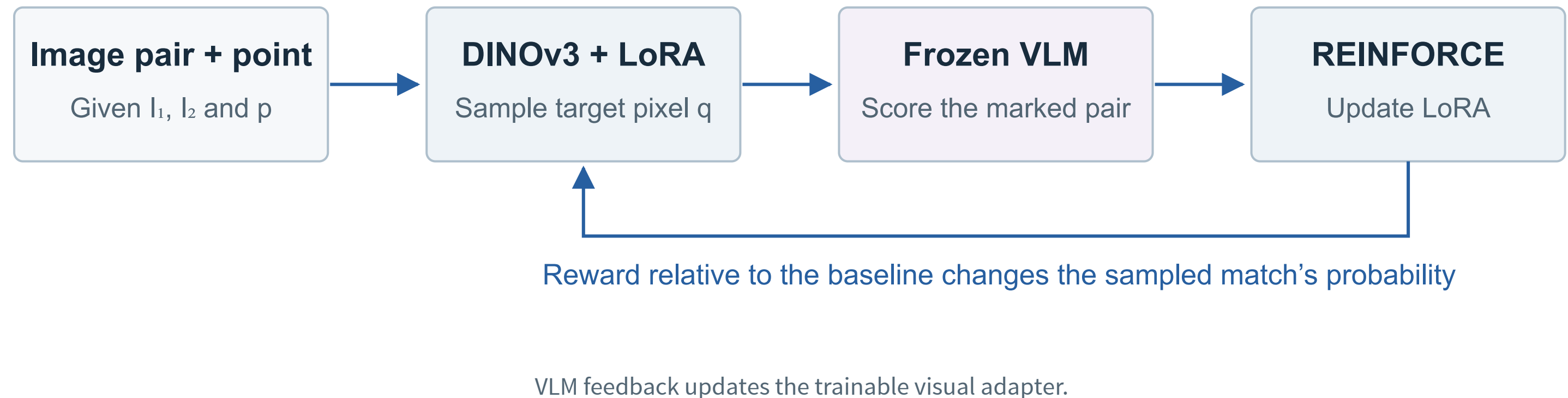
Three ways to obtain a match

Approach	What it offers	What is missing
Frozen visual features DINO / diffusion features	Fast, useful correspondences	Fine part and side distinctions
Supervised matching e.g. TLFR	Task-specific geometric alignment	Needs correspondence annotations
Direct VLM prediction	Semantic knowledge in the model	Slow calls and unreliable coordinates

We use semantic judgment to improve the fast visual matcher.

The VLM judges proposed matches

The vision model proposes a target point. The judge scores semantic agreement.



Increase the probability of matches that receive better feedback.

Source points are given. Ground-truth target coordinates do not enter the VLM reward.

The judge is needed only during training

Training

Propose a match

Score it with a VLM

Update the visual adapter

Optional DINO + diffusion features add geometric information.

Inference

Match visual features

Return the target point

26.3 frames per second
plain DINO pipeline

No VLM call at test time

Feedback improves two visual matching pipelines

SPair-71k cat test · 600 pairs · PCK@0.10 (%)

Feature pipeline	Before adaptation	Kimi K2.5 reward	Gain
Plain DINOv3	70.05	71.46 ± 0.03	+1.41
Geometry-aware	75.91	77.38 ± 0.25	+1.47

Judge feedback improves each pipeline over its own initialisation.

Final checkpoints after 1,000 training epochs.
Mean ± standard deviation over three training seeds.

A part name can replace the source point

“Cockpit of the airplane”

Encode the name with CLIP.

Propose locations using DINO features.

One judge call scores eight candidate points.

The VLM evaluates each candidate against the same name.



A KeypointNet render. Input illustration, not a predicted point.

Judge rewards teach named pointing

KeypointNet · Unseen test meshes of the 12 training categories · CLIP embd.

Reward	2D PCK@0.10
Untrained	5.0
Kimi K3	26.8
Ground truth	39.2

Same pretrained DINO and CLIP.
Untrained = initial query head and adapters.

≈64% of the gain from ground-truth rewards

Reworded-name PCK: 25.9

The learned policy responds to language.

Matched nominal budget up to 2,000 queries, development-selected checkpoints, one seed. PCK uses the visible-keypoint bounding box.

Predictions from multiple views give a 3D point



Rendered inputs from different cameras

Query each view.
Lift points with depth.
Fuse agreement in 3D.

34.0 3D PCK@0.10
CLIP embd. + Kimi K3

14 views per shape, using the same 2k model.

Euclidean 3D PCK, not the same metric as 2D PCK. No 3D reward training.

Fine localisation and transfer remain difficult

01 Coarse spatial judgment

Qwen3-VL rewards flatten below about 20 pixels in correspondence calibration.

02 Part-name discrimination

Kimi K3 ranks the true named part above another part in only 57% of calibration views.

03 Unseen categories

The 2k CLIP + Kimi model reaches 15.6 PCK on four categories excluded from training.

Better rewards must resolve fine spatial and semantic distinctions.

VLM feedback can train a fast spatial model

01 Semantic judgment provides a training signal

The VLM scores proposed correspondences.

02 The visual model retains the learned skill

Matching at test time needs no VLM call.

03 The loop also supports named keypoints

Text queries and multi-view fusion extend the task.

Questions