Is It a Good Match?
2D Keypoint Matching with Feedback of a VLM

Anonymous authors
Paper under double-blind review

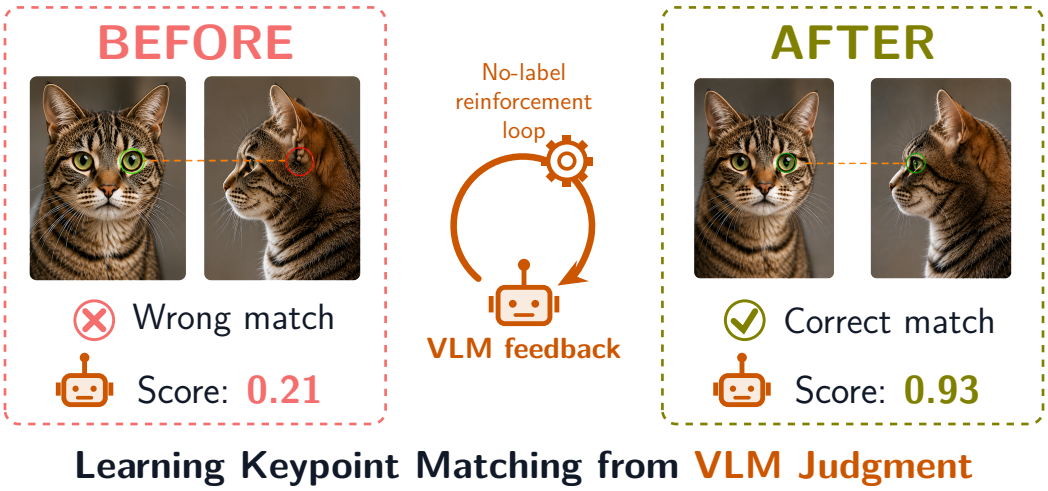


Figure 1: **Learning correspondence with VLM judgment.** A VLM scores predicted matches to guide vision-model fine-tuning via reinforcement learning, without ground-truth keypoint labels.

Abstract

Two-dimensional (2D) keypoint matching is a fundamental problem in computer vision, used in applications such as image registration, 3D reconstruction, and visual localization. Recent Vision Foundation Models, such as DINOv3, have demonstrated strong transferable representations across a range of vision tasks, including keypoint correspondence estimation. However, their performance remains limited by imperfect feature distinctiveness and weak geometric awareness. Their practical usage often relies on task-specific fine-tuning with carefully engineered losses. Vision-Language Models (VLMs) have shown strong capabilities in evaluating and reasoning on multiple vision tasks, including assessing the quality of vision model predictions. However, their high inference cost makes them impractical for direct use in matching pipelines. In this work, we leverage the evaluative strength of VLMs to guide the fine-tuning of vision foundation models for 2D keypoint matching. Specifically, we construct a feedback signal based on a VLM-based evaluation score (VLM-as-Judge) of predicted correspondences. We introduce a feedback-driven training framework using the VLM score to fine-tune a DINOv3 backbone for keypoint matching. By integrating this reward-based feedback loop, our method improves correspondence quality without requiring explicit ground-truth supervision or loss engineering. At test time, we rely only on the fine-tuned model, running 50–160$\times$ faster than a *single* VLM call. The same loop extends to text-queried keypoints: with Kimi K3 scoring eight numbered candidates per call, DINOv3 learns to point at named parts of KeypointNet objects without any keypoint label. The CLIP-query policy recovers about two-thirds of the gain of a ground-truth reward at 2,000 nominal queries, and its predictions lift to 3D by multi-view fusion.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Establishing reliable correspondences between images is a long-standing and fundamental problem in computer vision. In particular, 2D keypoint matching underpins numerous downstream applications, including image registration, structure-from-motion (Schönberger & Frahm, 2016; Schönberger et al., 2016), visual localisation, and augmented reality. Despite decades of progress, robust keypoint matching remains challenging in the presence of viewpoint changes, illumination variations, occlusions, and domain shifts. A central limitation of many existing approaches is their reliance on large-scale, high-quality supervision, which is often unavailable, noisy, or expensive to obtain in real-world scenarios.

Recent Vision Foundation Models, such as DINOv2 (Oquab et al., 2024) and DINOv3 (Siméoni et al., 2025), have demonstrated remarkable zero-shot generalisation capabilities across a wide range of visual tasks. Pretrained on large and diverse image corpora using self-supervised objectives, these models provide semantically rich and spatially consistent representations that are highly promising for correspondence estimation. However, adapting these generic features to the fine-grained demands of 2D keypoint matching without supervision remains an open challenge.

In parallel, Vision-Language Models (VLMs)—such as GPT-5, Claude, Gemini, Qwen, and Kimi—have exhibited impressive reasoning and evaluation capabilities for vision tasks. Beyond generation, these models are increasingly used as *judges* (Zheng et al., 2023) to assess the quality of predicted outputs across diverse domains. However, directly integrating VLMs into inference pipelines for vision tasks is computationally expensive and often impractical due to high latency and resource requirements.

We propose to bridge these two developments to fine-tune vision foundation models for keypoint matching. We construct a feedback signal for semantic correspondence using a VLM-based judge that outputs a score of matching prediction. We formulate our training as a reinforcement learning (RL) problem and propose a feedback-driven training framework based on the VLM score. This feedback signal guides the adaptation of a vision backbone via REINFORCE (Williams, 1992) policy gradients and a parameter-efficient Low-Rank Adaptation (LoRA) (Hu et al., 2022) module. By distilling the judgment capability of the VLM into the vision model at training time, we obtain a system that benefits from high-level evaluative reasoning while retaining the efficiency of a pure vision model at test time.

Our contributions are summarised as follows:

- We introduce a VLM-reward framework that adapts vision backbones through LoRA and REINFORCE without ground-truth correspondence labels.
- We evaluate multiple backbones and VLM judges on SPair-71k (Min et al., 2019) and PF-PASCAL (Ham et al., 2018), showing correspondence gains and 50–$160\times$ faster inference than a VLM call.
- We extend VLM-reward training to text-queried keypoints without labels, including reworded queries and multi-view 3D localisation.

Overall, our work highlights a new paradigm for geometric vision: using language-based evaluators as supervisory signals to improve foundation models in settings where traditional supervision is limited or unreliable.

2 RELATED WORK

2.1 VISION FOUNDATION MODELS FOR CORRESPONDENCE

Large, self-supervised Vision Transformers (Dosovitskiy et al., 2021) produce dense, expressive features that have become the dominant zero-shot signal for keypoint correspondence. They can be trained using different strategies such as self-supervised reconstruction (He et al., 2022), feature self-supervised training (Caron et al., 2021; Oquab et al., 2024; Siméoni et al., 2025), language supervision (Radford et al., 2021) or even extracted from pre-trained image generative diffusion models (Tang et al., 2023; Stracke et al., 2025). Also, fusing multiple features with a learnable aggregation network (Xu et al., 2023) improves open-vocabulary segmentation. Our method builds

directly on this line: we adopt a frozen DINOv3 backbone with a LoRA (Hu et al., 2022) adapter and, in the geo-aware variant, fuse it with cached Stable Diffusion features via an aggregation network with an identity-initialised DINO residual.

2.2 Few-shot Semantic Keypoint Matching

2D keypoint matching approaches explore a variety of formulations and strategies. Several methods cast semantic correspondence as an optimal transport problem (Liu et al., 2020; Lee et al., 2021), while others enhance supervision by leveraging sparse keypoint annotations (Huang et al., 2022) or probabilistic objectives for weakly supervised dense matching (Truong et al., 2022). On the architectural side, prior work investigates controlling token complexity through cost aggregation transformers (Cho et al., 2021; 2023), as well as consolidating multi-scale and multi-timestep feature maps into dense per-pixel descriptors (Luo et al., 2023). More recent approaches focus on feature design, combining representations from diffusion models and self-supervised vision encoders (Zhang et al., 2023), refining diffusion features without explicit noise modeling (Stracke et al., 2025), or distilling large models into compact vision backbones using LoRA adaptation (Fundel et al., 2025). The geometry-aware approach of Telling Left from Right (Zhang et al., 2024) fuses DINOv2 and Stable Diffusion features through a learned aggregation network and serves as the architectural baseline for our geo-aware variant.

2.3 Language-Grounded Dense Features and Named Keypoints

Open-vocabulary dense prediction aligns patch features with text using image–text supervision, either from CLIP's own features (Zhou et al., 2022; Wysoczańska et al., 2024) or by mapping text into DINO's space (Barsellotti et al., 2025; Jose et al., 2025). Zero-shot keypoint detectors instead prompt CLIP with visual markers (Shtedritski et al., 2023) or reason with large language models over rendered views (Gong et al., 2025), and Set-of-Mark prompting (Yang et al., 2023) lets VLMs refer to numbered image regions. Our text-queried extension trains the text–feature alignment from judge scores alone.

2.4 LLM-as-Judge and RL from AI Feedback

The LLM-as-judge paradigm has emerged as a scalable alternative to human evaluation (Zheng et al., 2023), and Reinforcement Learning from AI Feedback (RLAIF) (Bai et al., 2022; Lee et al., 2024) has been used to align language models without human preference data. Recent work generalises this beyond text: Eureka (Ma et al., 2024) uses LLMs to author dense reward functions for robotic control. Other methods use VLMs (Rocamonde et al., 2024; Baumli et al., 2023; Wang et al., 2024; Venuto et al., 2024; Huang et al., 2025) as a source of reward signal for tasks like robotic control, game environments and autonomous driving. To the best of our knowledge, ours is the first work to use a frozen VLM as the per-keypoint reward signal in a REINFORCE loop for dense semantic correspondence.

3 Method

We show an overview of our method in Figure 2.

3.1 Problem Formulation

Given two RGB images I_1 and I_2 with pixel domains Ω_1 and Ω_2, we predict a target location $q^i \in \Omega_2$ for each of N source keypoints:

$$\mathcal{K}_1 = \left\{ p^i = (u_i, v_i) \in \Omega_1 \right\}_{i=1}^{N}. \tag{1}$$

The state (I_1, I_2, p^i) conditions a categorical policy $\pi_\theta(a \mid I_1, I_2, p^i)$ over target pixels $a \in \Omega_2$, derived from the features of g_θ (Section 3.4). A frozen vision-language model (VLM) judges the sampled match and supplies its reward $r(a)$ (Section 3.5).

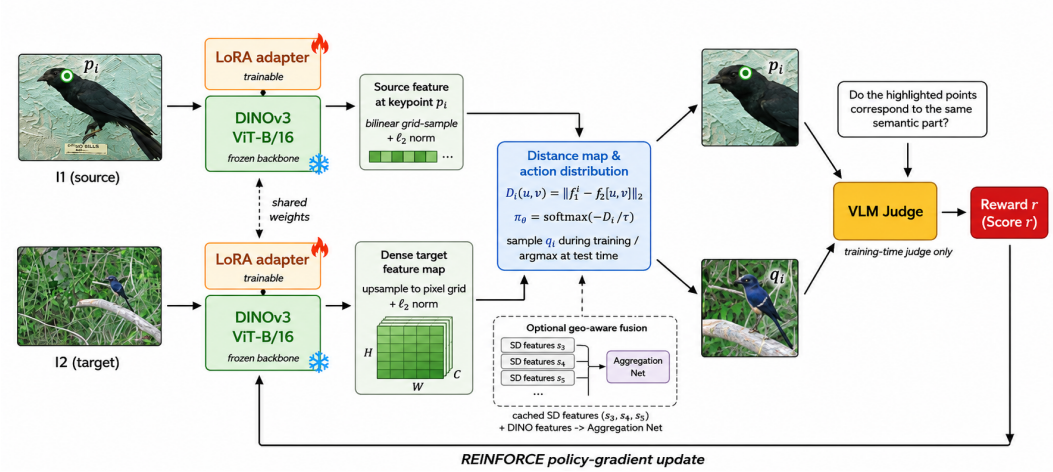


Figure 2: **Method pipeline.** LoRA-adapted DINO features define a target-pixel distribution for a source keypoint. A sampled match is marked in both images and scored by a VLM; REINFORCE uses this reward to update LoRA. The optional dashed path fuses cached Stable Diffusion features through a learnable Aggregation Network.

3.2 Feature Computation

We use a frozen Vision Transformer (ViT) (Dosovitskiy et al., 2021) with adaptation weights θ to compute dense d-dimensional features $f_1 = g_\theta(I_1)$ and $f_2 = g_\theta(I_2)$.

LoRA adaptation. We apply Low-Rank Adaptation (Hu et al., 2022) to the query-key-value projection matrices of the ViT. For each linear layer $W \in \mathbb{R}^{d\times d}$, the adapted weight is $W' = W + (\alpha/r)\, BA$ where $A \in \mathbb{R}^{r\times d}$, $B \in \mathbb{R}^{d\times r}$, where A and B are trainable, r is the rank, and α is the scaling factor.

3.3 Geo-Aware variant

In this *geo-aware* variant, we follow Zhang et al. (2024) and fuse the live DINO features with Stable Diffusion (SD) features (Rombach et al., 2022). Skip connections preserve input information through each bottleneck. The feature maps $f^1, ..., f^n$ are fused along the channel axis and passed through a learnable **Aggregation Network** (Xu et al., 2023), outputs a D-dimensional output:

$$F^{\text{agg}} = \text{AggNet}([f^1, ..., f^n]) \in \mathbb{R}^{1\times D\times P\times P} \tag{2}$$

Both the LoRA adapter and the AggregationNetwork parameters are optimised jointly.

Residual initialisation. We observed that setting weights randomly affects the performance of the vision model at the beginning of the training. We edit the aggregation network to present the vision model features at initialisation using a residual variant:

$$\hat{F}^{\text{agg}} = \phi(F^{\text{dino}}) + \alpha \cdot F^{\text{agg}}, \tag{3}$$

where ϕ is initialized at the identity. This simple residual strategy mitigates the accuracy gap between geo-aware and DINO variants of our approach (Table 1).

3.4 Keypoint Selection and Action Sampling

For a randomly selected keypoint $p^i \in \mathcal{K}_1$ with feature vector $f_1^i \in \mathbb{R}^d$, we compute the dense distance map

$$D_i(u,v) = \|f_1^i - f_2[u,v,:]\|_2, \quad (u,v) \in \Omega_2. \tag{4}$$

We convert distances to a probability distribution via temperature-scaled softmax ($\tau = 0.1$ by default):

$$\pi_\theta^i = \text{softmax}\left(\frac{-D_i}{\tau}\right). \tag{5}$$

During training, we sample $q^i \sim \text{Categorical}(\pi_\theta^i)$ for REINFORCE exploration (Sec. 3.6); higher τ encourages broader exploration. At test time, we select $q^i = \arg\max_{(u,v)\in\Omega_2} \pi_\theta^i[u, v]$.

3.5 VLM/LLM Reward Function

Given the source keypoint p^i in I_1 and its predicted counterpart q^i in I_2, we evaluate the quality of the match via a frozen VLM. Both keypoints are annotated in the image (see Table 5 in the supplementary material for the marker-shape ablation) and the annotated pair is sent to the VLM with an evaluation prompt:

> *Determine whether the two highlighted points correspond to the same semantic part of objects belonging to the same category (e.g., left eye of a dog to left eye of another dog). Only return a score in the format 'Score: X', where X is the score between 0 and 1.*

The response is parsed for a numeric `Score:` X value clamped to $[0, 1]$; when multiple responses are collected, the average is taken as the reward.

Mixed-backend mode. To explore and reduce bias from any single LLM, a *mixed* virtual backend is added, which randomly delegates each reward query to one of several backend instances (e.g., GPT-5, Claude Opus, Kimi K2). Backend instances can be weighted via repetition, and each call is independently rate-limited by provider name.

Supervised overfit reward (ablation). For ablation studies, we provide a deterministic reward based on PCK@0.1 against ground-truth keypoints:

$$r_{\text{overfit}} = \begin{cases} 1 & \text{if } \|q^i - g^i\|_2 \le 0.1 \cdot \max(h_{\text{bbox}}, w_{\text{bbox}}) \\ 0 & \text{otherwise,} \end{cases} \tag{6}$$

where g^i is the ground-truth target keypoint and $(h_{\text{bbox}}, w_{\text{bbox}})$ are the target bounding-box dimensions. This mode requires no LLM queries and enables rapid training-loop development.

3.6 REINFORCE Training

REINFORCE (Williams, 1992) updates the LoRA module and, in the geo-aware variant, the Aggregation Network. The exponential moving average baseline starts at $b_0 = 0$ and updates after every sample as $b_{t+1} = (1 - \alpha_b)b_t + \alpha_b r_t$, with $\alpha_b = 0.01$. The advantage $\delta_t = r_t - b_t$ gives the per-sample loss

$$\mathcal{L} = -\frac{1}{B} \log \pi_\theta^i[q^i] \cdot \delta_t, \tag{7}$$

where B is the gradient accumulation batch size (default $B = 10$); the factor $1/B$ normalizes the accumulated gradient. Table 1 compares against DPO (Rafailov et al., 2023) and Top-K rewards (selecting the Top-K rewards for each prompt).

3.7 Extension: Text-Queried Keypoints

The same loop applies when the query is a keypoint *name* instead of a source keypoint. Given an image I and a text t (e.g., "upper attachment point for the mug handle"), the *CLIP text embedding* (CLIP embd.) query representation uses a frozen CLIP text encoder (Radford et al., 2021) to embed t as $c(t)$. A learned linear head W maps it into the DINO feature space, and the policy is a softmax over pixels of the cosine similarity with the LoRA-adapted features:

$$\pi_\theta(a \mid I, t) = \text{softmax}_a\Big(\tfrac{1}{\tau} \cos\big(q(t),\, g_\theta(I)[a]\big)\Big), \qquad q(t) = W c(t) + \textstyle\sum_{w\in t} e_w, \tag{8}$$

where e_w are learned vectors for side words (left, right, front, ...), since CLIP barely separates mirrored names (median cosine 0.99). The *learned name embedding* (Learned embd.) control replaces $q(t)$ with an independently learned vector for each exact training name, implemented as an embedding lookup table without a text encoder. These vectors are randomly initialised and trained jointly with the LoRA adapters using the same reward. The *hybrid* $q(t) + r_t$ adds a zero-initialised learned residual r_t to the CLIP-based query for names seen in training (unseen or reworded names fall back to $q(t)$). All query representations use DINO image features; in VLM-reward runs, the judge receives the same natural-language name t.

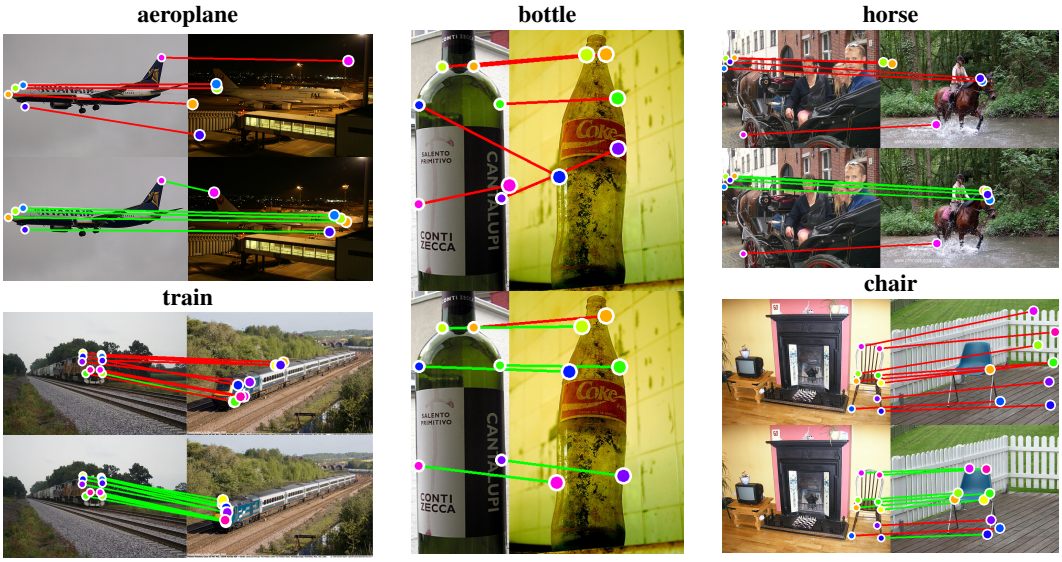


Figure 3: **SPair-71k test matches.** Per block, top: frozen DINOv3 ViT-B/16; bottom: DINOv3+LoRA after supervised REINFORCE. Green/red lines: correct/incorrect at α_{bbox}=0.1.

Multi-marker judge. For each query we sample K=8 pixels, draw them as numbered circles on one image (Set-of-Mark style (Yang et al., 2023)) and ask the VLM to score each marker against t in a single call (1: at the named part, 0.5: near it or on the right part but off, 0: elsewhere), which yields K rewards per call. Each sample uses the others as its baseline (Kool et al., 2019), $\delta_k = r_k - \frac{1}{K-1}\sum_{j\neq k} r_j$, in place of the moving average of Sec. 3.6. Queries follow a curriculum from object names ("the mug", a dense reward) to keypoint names, and images are cropped around the object, which covers only $\sim$3.6% of a rendered view.

4 Experiments

4.1 Experimental Setup

Vision backbone. We use DINOv2 ViT-B/14 or DINOv3 ViT-B/16 with LoRA adapters (rank r=16, scaling α=8).

Datasets We evaluate on SPair-71k (70k+ pairs, 18 categories, diverse viewpoint/scale variations) (Min et al., 2019) and PF-PASCAL (PASCAL-derived keypoint pairs) (Ham et al., 2018).

Unless noted, models train on SPair-71k `cat` and use its validation split; Table 1 gives the best where three seeds were run. The *PF-PASCAL* `cat` and *SPair* `cat` columns denote cat-only evaluation. *SPair all* covers all 18 categories; all-category training is in-distribution. The rows include different training budgets and evaluation splits (Appendix B.4).

Metric We report PCK@α_{bbox}, the percentage of predictions within $\alpha \max(h, w)$ of the ground-truth target keypoint, where h, w are the target bounding-box dimensions; $\alpha = 0.10$ unless stated.

4.2 Results

Table 1 summarises the results.

VLM-reward training. All three self-training methods outperform frozen DINOv3 (76.19) on SPair `cat` (val split): Top-K reaches 78.10, DPO 79.09, and Kimi-judged DINOv3+LoRA REINFORCE 85.65 (Table 1). In the separate matched-budget comparison, REINFORCE beats Top-K by $+0.88\pm0.15$ pt, while DPO matches REINFORCE within seed variance. On validation, the REINFORCE advantage over Top-K widens to $+1.31$ pt at α=0.05 (shrinking at 0.01); on test, it scales monotonically with difficulty: $+0.28$ on Easy to $+\mathbf{2.90}$ on Hard (Table 2). At α=0.05, the

Method	SPair all	SPair `cat`	PF-PASCAL `cat`
Zero-shot baselines (no training):			
DINOv2 ViT-B/14 (frozen patch features)	56.8	72.46	81.0
DINOv3 ViT-B/16 (frozen patch features)	54.2	76.19	88.56
Kimi K3 (VLM-as-direct-predictor; §4.4)	49.29	66.60	48.11
Gemini 3.8 Flash (VLM-as-direct-predictor)	16.82	24.29	48.11
Qwen3-VL-32B (VLM-as-direct-predictor)	22.41	42.84	29.60
Qwen3-VL-235B-A22B (VLM-as-direct-predictor)	24.37	44.30	45.75
GLM-4.6V (VLM-as-direct-predictor)	33.1	56.6	36.3
Alternative self-training:			
DINOv3 + LoRA, Top-K (K=5, Kimi)	54.28	78.10	87.74
DINOv3 + LoRA, DPO (β_{KL}=0.1, Kimi)	56.86	79.09	–
Ours — VLM-reward REINFORCE (no ground-truth labels):			
DINOv2 + LoRA (Kimi K3)	63.5	84.74	90.26
DINOv2 + LoRA (GPT-6 Astra)	58.10	85.39	90.92
DINOv2 + LoRA (Opus 5.5)	58.31	84.53	89.32
DINOv2 + LoRA (Gemini 3.8 Flash)	57.71	85.37	91.39
DINOv3 + LoRA (Kimi K3)	62.16	85.65	91.39
DINOv3 + LoRA (GPT-5.5)	57.19	86.07	92.35
DINOv3 + LoRA (GPT-6 Astra)	56.37	86.52	92.57
DINOv3 + LoRA (Opus 5.5)	58.82	86.40	92.67
DINOv3 + LoRA (Gemini 3 Pro)	67.83	–	–
DINOv3 + LoRA (Gemini 3.8 Flash)	63.46	86.81	92.99
Geo-aware + LoRA + resid. (Kimi K3)	61.75	85.89	93.95
Geo-aware + LoRA + resid. (GPT-5.5)	62.14	84.98	92.22
Geo-aware + LoRA + resid. (GPT-6 Astra)	61.37	86.84	94.94
Geo-aware + LoRA + resid. (Gemini 3 Pro)	**71.20**	–	–
Geo-aware + LoRA + resid. (Gemini 3.8 Flash)	61.03	**86.86**	**96.54**
Geo-aware + LoRA + resid. (Opus 5.5)	62.51	86.64	94.81
Supervised training			
DINOv3 + LoRA, GT-Reward (overfit)	56.8	82.56	91.63
Geo-aware, random AggNet init (overfit)	58.93	78.31	88.17
Geo-aware + DINO-resid. init (Eq. 3)	70.37	86.61	94.62
TLFR (Zhang et al., 2024) (DINOv3, 18-cat)	78.24	87.10	–

Table 1: **Keypoint matching on SPair-71k and PF-PASCAL** (PCK@α_{bbox}=0.10). VLM-reward fine-tuning is compared with frozen features, direct VLM prediction, alternative self-training and supervised baselines. Bold marks the best VLM-reward score per column.

Table 2: **REINFORCE vs Top-K by difficulty** (SPair-71k `cat` test; mean±SD, 3 seeds; Δ before rounding).

	$\alpha = 0.10$			$\alpha = 0.05$		
Bucket	REINFORCE	Top-K	Δ	REINFORCE	Top-K	Δ
Easy	77.92 ± 0.05	77.64 ± 0.05	$+0.28$	68.02 ± 0.07	67.91 ± 0.05	$+0.11$
Medium	69.04 ± 0.09	67.43 ± 0.07	$+1.62$	60.27 ± 0.08	59.15 ± 0.04	$+1.11$
Hard	63.73 ± 0.13	60.83 ± 0.14	$+\mathbf{2.90}$	54.78 ± 0.25	52.17 ± 0.08	$+\mathbf{2.61}$

test-set gain likewise grows from $+0.11$ on Easy to $+2.61$ on Hard, suggesting that reward-based updates are most useful for difficult matches. Figure 3 illustrates gains under supervised rewards.

Geo-aware advantage and supervised validation. With Kimi reward, the geo-aware stack reaches 85.89 vs plain DINOv3+LoRA's 85.65 on SPair `cat`, and 93.95 vs 91.39 on PF-PASCAL `cat` (Table 1). It also resolves left/right symmetry confusions (Fig. 4). GT-reward validation (Eq. 6) also favours the geo-aware stack (86.61 vs plain DINO's 82.56) and justifies the DINO-residual initialisation (Eq. 3): a random-init AggregationNetwork reaches only 78.31, below supervised plain DINO, whereas the residual variant starts at 81.6% (t=0), approaching the trained baseline before any updates. Upscaling plain DINO to 960×960 recovers ∼1 pt; the geo-aware stack also changes features and evaluation, so this is not an isolated resolution comparison.

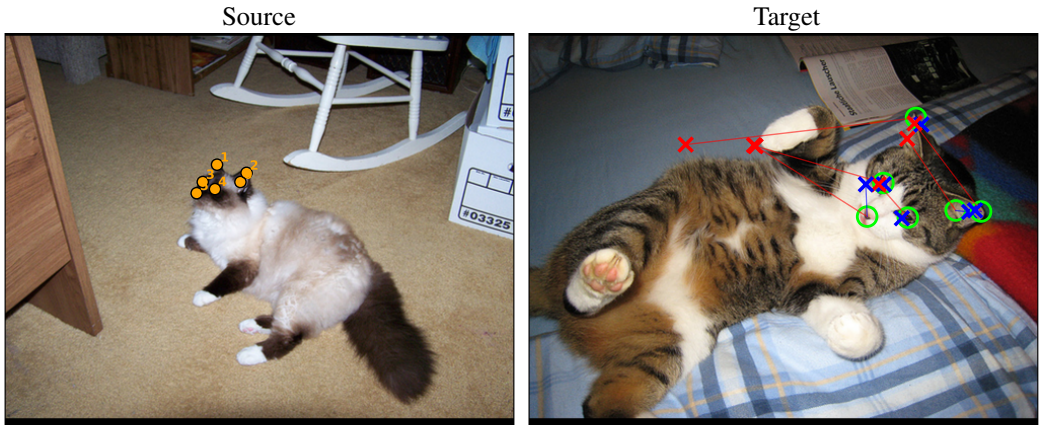


Figure 4: **Left/right correction** (SPair-71k `cat` test). Plain DINO (red) confuses 4/6 keypoints; geo-aware (blue) agrees with GT (green). Mean L_2: 233 vs. 21 px.

Efficiency and comparison with prior work. The plotted policies reach $\geq 98\%$ of final PCK within $\sim$22% of their VLM-call budget (Fig. 6). Inference runs at 26.3 FPS (plain DINO) and 8.5 FPS (geo-aware, cached SD features): 50–160$\times$ faster than a $\sim$6 s VLM call (Appendix B.3). On SPair-71k `cat` test, our geo-aware policy reaches 78.44, exceeding published TLFR zero-shot (DINOv2, 64.2) by 14.24 pt under different backbone/evaluation settings. The 4.98 pt gap to supervised TLFR-DINOv3 (83.42) includes a judge-alignment gap: our RL oracle (82.95) trails by 0.47 pt (Appendix B.1).

4.3 Analysis of the VLM-as-a-Judge Framework

Judge calibration. To verify that the VLM reward is informative, we measure how its score varies with prediction quality. On SPair-71k `cat` (N=126 keypoints), Qwen3-VL-32B-Thinking assigns a mean score of 0.83 when the predicted keypoint lands exactly on the ground-truth location (d=0 px) and only 0.11 when it is 160 px away (Spearman $\rho = -0.649$, $p \approx 2 \times 10^{-16}$). However, scores plateau once $d \leq 20$ px: the VLM cannot distinguish a 5 px error from a perfect match. The judge operates in a coarse manner, which is consistent with its patch representation for the vision part.

Multi-VLM robustness. To test robustness to judge choice, we sweep across VLMs from six vendors (Table 6 in the supplementary material). Cat-only DINOv3+LoRA scores range from Qwen3-VL-32B's 78.02 to Gemini 3.8 Flash's 86.81, above frozen DINOv3's 76.19; budgets include 1,000-epoch and longer runs. On the harder 18-category regime with DINOv3, **Gemini 3.1 Pro leads** (67.83 vs Kimi K2.5's 62.16). Kimi K2.5 without thinking outperforms its thinking counterpart (62.16 vs 58.63). Within Table 6, Claude Opus 5.5 leads all-category DINOv2 at 58.31; Gemini 3.8 Flash reaches 57.71 with DINOv2 and 63.46 with DINOv3.

Reward-hacking diagnostics. Judge–policy alignment and KL/entropy controls show no evidence of reward gaming; Appendix B.2 gives these diagnostics and the comparison with GT-reward training.

4.4 Ablation: VLM-as-Direct-Predictor

We test direct prediction by sending a marked source image and an unmarked target to the VLM and requesting fractional target coordinates in $[0, 1]^2$, with `thinking` disabled. The strongest direct predictor, Kimi K3, achieves 48.11 / 66.60 / 49.29 (PF-PASCAL `cat` / SPair `cat` / SPair all), compared with 91.39 / 85.65 / 62.16 for our Kimi-judged DINOv3+LoRA REINFORCE policy (Table 1). The direct-prediction scores use test data, whereas the cited SPair REINFORCE entries use validation data; these entries do not form a matched-split, matched-budget comparison. Per-model analysis and full PF-PASCAL results are in Appendix B.5.

Baseline EMA and temperature. Sweeps over the baseline EMA rate α_b and softmax temperature τ show flat PCK (spread < 1.3 pt); we retain α_b=0.01 and τ=0.10 (Fig. 5c). Visual-marker ablations are reported in Appendix B.2; cross-category and cross-dataset transfer are analysed in Appendix C.1.

Query representation	Reward	Query budget	PCK$_{.10}$	PCK$_{.05}$	Reworded	Retrieval	3D PCK$_{.10}$	Unseen
CLIP embd.	untrained	–	5.0	1.8	4.8	26.3	9.2	2.6
CLIP embd.	GT	2k	39.2	18.9	36.1	54.7	47.7	22.5
CLIP embd.	**Kimi K3**	2k	26.8	15.6	25.9	38.2	34.0	15.6
Learned embd.	GT	2k	44.9	18.8	–	56.4	56.3	–
Learned embd.	**Kimi K3**	2k	28.4	17.1	–	40.5	35.6	–
Hybrid	GT	20k	65.7	46.4	**49.6**	71.3	**76.1**	24.6
CLIP embd.	GT†	20k	61.8	41.8	47.4	64.4	70.7	**26.2**
Learned embd.	GT†	20k	**67.0**	**47.4**	–	**71.5**	–	–

Table 3: **Text-queried KeypointNet keypoints** (1 seed; test meshes from 12 training categories). Reworded reports PCK$_{.10}$ with paraphrased names; retrieval ranks visible names at a ground-truth point (chance 26.5%). The 3D metric fuses 14 views and normalises distances by the 3D ground-truth-keypoint bounding-box diagonal. Budgets count nominal queries; each VLM call scores 8 candidates, while GT rewards require no judge calls. "Unseen": PCK$_{.10}$ on 4 held-out categories; GT: ground-truth reward; †: repeated GT runs (protocol in Appendix D). Bold scores mark column maxima. Learned embeddings require exact training names; their repeated run's 3D evaluation is pending.

4.5 Text-Queried Keypoints

We evaluate named keypoints on KeypointNet (You et al., 2020), using PCK normalised by the visible-keypoint bounding-box diagonal. Appendix D provides the setup, baselines, judge calibration and expanded results.

VLM-judged training without labels. In Table 3, Kimi K3 training raises PCK@0.10 from 5.0 to 26.8–28.4 and retrieval from chance to ∼40%. At 2,000 nominal queries, the CLIP policy recovers ∼65% of the ground-truth reward's PCK gain but only 42% of its retrieval gain. The judge ranks the true keypoint above a different named keypoint in only 57% of calibration views, limiting name discrimination. The smaller recovery in retrieval than in localisation is consistent with this calibration: the judge provides useful spatial feedback but is less reliable at separating different named parts. On the shared 320-view subset, our Kimi-trained policies reach 29.8–30.7 PCK, compared with 17.7 for name-blind DINO saliency and 8.8 for CLIP-based Red Circle (Table 13).

Language at the query, part locations in the features. With ground-truth rewards, adapting DINO raises PCK from 54.7 to 62.3 (Table 14). A fresh text head reaches 63.4 on these frozen adapted features and 62.9 on features adapted without a text encoder: supervision teaches the part locations. Hybrid queries preserve learned embeddings' precision (65.7 vs. 67.3, equal retrieval) while enabling rewordings (49.6). Fusing 14 views improves every completed 3D evaluation (hybrid 76.1, learned embeddings 79.8). Held-out categories remain difficult (∼25% with ground-truth rewards at 20k queries); transfer follows geometric analogy, as with a skateboard's wheel, nose and tail. Reworded names and unseen categories therefore test different forms of transfer: recognising a paraphrase does not ensure that a familiar word identifies the correct geometry in a new category.

5 Conclusion

VLM feedback enables label-free reinforcement fine-tuning of vision backbones for keypoint matching, with efficient inference that needs no VLM call. The same loop learns text-queried keypoints: at 2,000 nominal queries, the Kimi K3-trained CLIP policy recovers about two-thirds of the ground-truth reward's PCK gain. Supervision teaches DINO part locations, while the query encoder supplies the language interface.

Limitation and promising directions. The reward plateaus below ∼20 px, limiting localization precision; higher-resolution judges or pairwise preferences may help. Kimi K3 distinguishes the true part from another named part in only 57% of calibration views, and unseen-category transfer follows geometric analogy rather than shared words. Our multi-view lift motivates extending feedback training to 3D keypoint and shape matching, with mesh-geodesic evaluation left to future work.

AI USE STATEMENT

VLMs serve as judges providing training rewards in the proposed method and as direct-prediction baselines in our experiments. Kimi K3 also generated keypoint-name paraphrases for the reworded-query evaluation (Appendix D). We additionally used AI assistance to edit the LaTeX manuscript, particularly tables, captions and layout. AI-assisted checks compared manuscript text with table entries and checked PDF compilation and layout. All AI-assisted content was manually reviewed and checked. The authors take responsibility for the final content, including text, claims and artifacts produced with AI assistance.

ETHICS STATEMENT

This work uses publicly available benchmarks (SPair-71k, PF-PASCAL, AP-10K) and renders of the public KeypointNet/ShapeNet meshes; it involves no human subjects and no personal data. The reward judges are commercial VLMs queried through their APIs; their scores may carry biases of those models, which our mixed-judge experiments (Sec. 4.3) partly probe.

REPRODUCIBILITY STATEMENT

Implementation details (backbones, LoRA configuration, optimisation, budgets) are given in Appendix A and the judge prompts in Sec. 3.5 and Appendix D. Data splits follow the public benchmarks; for KeypointNet we describe the renders, the object-centric crops and the held-out categories in Appendix D. Code, configuration files and the per-run logs behind every table will be released.

REFERENCES

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*, 2022.

Luca Barsellotti, Lorenzo Bianchi, Nicola Messina, Fabio Carrara, Marcella Cornia, Lorenzo Baraldi, Fabrizio Falchi, and Rita Cucchiara. Talking to DINO: Bridging self-supervised vision backbones with language for open-vocabulary segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22025–22035, 2025.

Kate Baumli, Satinder Baveja, Feryal Behbahani, Harris Chan, Gheorghe Comanici, Sebastian Flennerhag, Maxime Gazeau, Kristian Holsheimer, Dan Horgan, Michael Laskin, Clare Lyle, Hussain Masoom, Kay McKinney, Volodymyr Mnih, Alexander Neitz, Dmitry Nikulin, Fabio Pardo, Jack Parker-Holder, John Quan, Tim Rocktäschel, Himanshu Sahni, Tom Schaul, Yannick Schroecker, Stephen Spencer, Richie Steigerwald, Luyu Wang, and Lei Zhang. Vision-language models as a source of rewards. *arXiv preprint arXiv:2312.09187*, 2023.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021.

Seokju Cho, Sunghwan Hong, Sangryul Jeon, Yunsung Lee, Kwanghoon Sohn, and Seungryong Kim. CATs: Cost aggregation transformers for visual correspondence. In *Advances in Neural Information Processing Systems*, volume 34, pp. 9011–9023, 2021.

Seokju Cho, Sunghwan Hong, and Seungryong Kim. CATs++: Boosting cost aggregation with convolutions and transformers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(6):7174–7194, 2023.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Frank Fundel, Johannes Schusterbauer, Vincent Tao Hu, and Björn Ommer. Distillation of diffusion features for semantic correspondence. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pp. 6762–6774. IEEE, 2025.

Bingchen Gong, Diego Gomez, Abdullah Hamdi, Abdelrahman Eldesokey, Ahmed Abdelreheem, Peter Wonka, and Maks Ovsjanikov. ZeroKey: Point-level reasoning and zero-shot 3D keypoint detection from large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22089–22099, 2025. arXiv:2412.06292.

Bumsub Ham, Minsu Cho, Cordelia Schmid, and Jean Ponce. Proposal flow: Semantic correspondences from object proposals. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(7):1711–1725, 2018.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Shuaiyi Huang, Luyu Yang, Bo He, Songyang Zhang, Xuming He, and Abhinav Shrivastava. Learning semantic correspondence with sparse annotations. In *European Conference on Computer Vision*, pp. 267–284. Springer, 2022.

Zilin Huang, Zihao Sheng, Yansong Qu, Junwei You, and Sikai Chen. VLM-RL: A unified vision language models and reinforcement learning framework for safe autonomous driving. *Transportation Research Part C: Emerging Technologies*, 180:105321, 2025.

Cijo Jose, Théo Moutakanni, Dahyun Kang, Federico Baldassarre, Timothée Darcet, Hu Xu, Daniel Li, Marc Szafraniec, Michaël Ramamonjisoa, Maxime Oquab, Oriane Siméoni, Huy V. Vo, Patrick Labatut, and Piotr Bojanowski. DINOv2 meets text: A unified framework for image- and pixel-level vision-language alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24905–24916, 2025.

Wouter Kool, Herke van Hoof, and Max Welling. Buy 4 REINFORCE samples, get a baseline for free! In *ICLR Workshop on Deep Reinforcement Learning Meets Structured Prediction*, 2019.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Thomas Mesnard, Johan Ferret, Kellie Ren Lu, Colton Bishop, Ethan Hall, Victor Carbune, Abhinav Rastogi, and Sushant Prakash. RLAIF vs. RLHF: Scaling reinforcement learning from human feedback with AI feedback. In *International Conference on Machine Learning*, pp. 26874–26901, 2024.

Jae Yong Lee, Joseph DeGol, Victor Fragoso, and Sudipta N. Sinha. PatchMatch-based neighborhood consensus for semantic correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13153–13163, 2021.

Yanbin Liu, Linchao Zhu, Makoto Yamada, and Yi Yang. Semantic correspondence as an optimal transport problem. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4463–4472, 2020.

Grace Luo, Lisa Dunlap, Dong Huk Park, Aleksander Holynski, and Trevor Darrell. Diffusion hyperfeatures: Searching through time and space for semantic correspondence. In *Advances in Neural Information Processing Systems*, volume 36, pp. 47500–47510, 2023.

Yecheng Jason Ma, William Liang, Guanzhi Wang, De-An Huang, Osbert Bastani, Dinesh Jayaraman, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Eureka: Human-level reward design via coding large language models. In *International Conference on Learning Representations*, 2024.

Juhong Min, Jongmin Lee, Jean Ponce, and Minsu Cho. SPair-71k: A large-scale benchmark for semantic correspondence. *arXiv preprint arXiv:1908.10543*, 2019.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pp. 8748–8763, 2021.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, 2023.

Juan Rocamonde, Victoriano Montesinos, Elvis Nava, Ethan Perez, and David Lindner. Vision-language models are zero-shot reward models for reinforcement learning. In *International Conference on Learning Representations*, 2024.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Johannes Lutz Schönberger and Jan-Michael Frahm. Structure-from-motion revisited. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016.

Johannes Lutz Schönberger, Enliang Zheng, Jan-Michael Frahm, and Marc Pollefeys. Pixelwise view selection for unstructured multi-view stereo. In *European Conference on Computer Vision*, pp. 501–518, 2016.

Aleksandar Shtedritski, Christian Rupprecht, and Andrea Vedaldi. What does CLIP know about a red circle? Visual prompt engineering for VLMs. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025.

Nick Stracke, Stefan Andreas Baumann, Kolja Bauer, Frank Fundel, and Björn Ommer. CleanDIFT: Diffusion features without noise. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 117–127, 2025.

Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent correspondence from image diffusion. In *Advances in Neural Information Processing Systems*, 2023.

Prune Truong, Martin Danelljan, Fisher Yu, and Luc Van Gool. Probabilistic warp consistency for weakly-supervised semantic correspondences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8708–8718, 2022.

David Venuto, Mohammad Sami Nur Islam, Martin Klissarov, Doina Precup, Sherry Yang, and Ankit Anand. Code as reward: Empowering reinforcement learning with VLMs. In *International Conference on Machine Learning*, pp. 49368–49387, 2024.

Yufei Wang, Zhanyi Sun, Jesse Zhang, Zhou Xian, Erdem Biyik, David Held, and Zackory Erickson. RL-VLM-F: Reinforcement learning from vision language foundation model feedback. In *International Conference on Machine Learning*, pp. 51484–51501, 2024.

Ronald J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3–4):229–256, 1992.

Monika Wysoczańska, Oriane Siméoni, Michaël Ramamonjisoa, Andrei Bursuc, Tomasz Trzciński, and Patrick Pérez. CLIP-DINOiser: Teaching CLIP a few DINO tricks for open-vocabulary semantic segmentation. In *European Conference on Computer Vision*, 2024.

Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

Jianwei Yang, Hao Zhang, Feng Li, Xueyan Zou, Chunyuan Li, and Jianfeng Gao. Set-of-mark prompting unleashes extraordinary visual grounding in GPT-4V. *arXiv preprint arXiv:2310.11441*, 2023.

Yang You, Yujing Lou, Chengkun Li, Zhoujun Cheng, Liangwei Li, Lizhuang Ma, Cewu Lu, and Weiming Wang. KeypointNet: A large-scale 3D keypoint dataset aggregated from numerous human annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020.

Junyi Zhang, Charles Herrmann, Junhwa Hur, Luisa Polania Cabrera, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. A tale of two features: Stable diffusion complements DINO for zero-shot semantic correspondence. In *Advances in Neural Information Processing Systems*, volume 36, pp. 45533–45547, 2023.

Junyi Zhang, Charles Herrmann, Junhwa Hur, Eric Chen, Varun Jampani, Deqing Sun, and Ming-Hsuan Yang. Telling left from right: Identifying geometry-aware semantic correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, 2023.

Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from CLIP. In *European Conference on Computer Vision*, 2022.