ONE-SHOT RELATIONAL ADAPTATION OF VISION FOUNDATION MODELS FOR 3D CORRESPONDENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision foundation models (VFMs) have demonstrated generalization across diverse 2D vision tasks, yet adapting them for 3D correspondence tasks, such as multi-view matching and pose estimation, remains challenging due to the lack of large-scale 3D correspondence annotations. Existing adaptation strategies typically rely on extensive training data or task-specific supervision, limiting their applicability in data-scarce scenarios. In this work, we propose a one-shot adaptation framework that enhances the 3D correspondence capability of VFMs using only a single overlapping image pair with cross-view correspondences. Our key insight is that a correspondence-annotated image pair provides not only point-level supervision but also rich relational structures among matched points. To exploit this property, we introduce a two-stage adaptation framework. First, an attention-gated convolutional adapter is designed to improve local cross-view feature consistency while introducing minimal modification to the pretrained VFM. Second, a relational adaptation module explicitly models structural relationships among corresponding points and enforces cross-view relational consistency, enabling the adapted representation to preserve relational structures across viewpoints. Experiments on multiple 3D correspondence benchmarks show that our approach generally improves pretrained VFMs and outperforms recent VFM adaptation methods in many settings, while requiring only a single image pair for adaptation. These results highlight the effectiveness of exploiting intrinsic correspondence and relational information for efficient 3D-aware VFM adaptation.

1 INTRODUCTION

Vision foundation models (VFMs), particularly those based on the Vision Transformer (ViT) architecture (Dosovitskiy et al., 2020), have emerged as powerful approaches for learning generalizable visual representations. By leveraging large-scale 2D image data, these models capture rich semantic and structural knowledge, enabling strong performance across a wide range of downstream tasks, including image understanding, dense prediction, and vision-based geometric reasoning. Such transferable representations have established VFMs as a fundamental paradigm for modern computer vision systems.

Recent efforts have focused on adapting VFMs to geometry-sensitive downstream tasks while preserving their generalization ability. FiT3D (Yue et al., 2024) and DVT (Yang et al., 2024a) improve VFM features for dense prediction through geometry-aware adaptation and artifact reduction. 3DCorrEnhance (You et al., 2025) shows that limited 3D correspondence supervision can improve feature-level equivariance in VFMs. Shape-of-You (Im et al., 2026) further leverages complementary geometric cues to improve correspondence robustness across viewpoint changes.

Despite these advances, existing methods still face several practical limitations. Many require complex fine-tuning pipelines (Yang et al., 2024a), introduce substantial parameter overhead (Yue et al., 2024), or rely on multiple foundation models and auxiliary representations (Im et al., 2026).

These requirements increase the cost of adaptation and limit applicability in data-scarce settings. More importantly, how to effectively adapt VFMs for 3D correspondence under extremely limited supervision remains underexplored.

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.

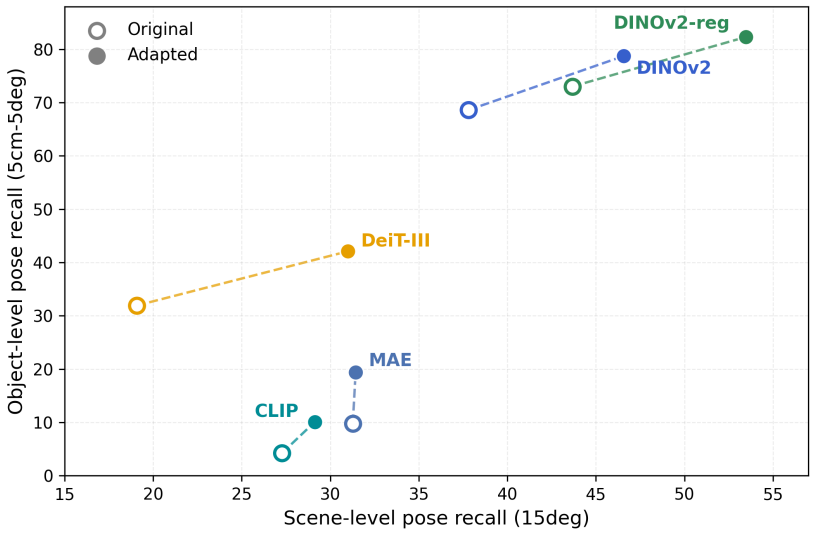


Figure 1: Our adapter consistently improves pose estimation across diverse pretrained models. Each line shows the performance change from the pretrained model (hollow) to the adapted model (solid) for scene-level and object-level pose estimation, demonstrating the applicability of our approach.

To address these challenges, we propose a one-shot adaptation approach for VFMs that requires only a single image pair with 3D spatial overlap and cross-view correspondences. Our lightweight adapter integrates seamlessly into VFM architectures, adding only 6.6% parameters via two lightweight adaptation stages and supporting dynamic enabling or disabling for flexible deployment. As illustrated in Figure 1, our approach consistently improves pose estimation performance across both scene-level and object-level settings.

Our approach takes a single overlapping image pair with 3D correspondences as supervision for one-shot adaptation. The lightweight adapter first refines native VFM features with an attention-gated convolutional module, where correspondence-aware contrastive learning improves local cross-view consistency. It then aligns pairwise feature distances across the two views through a relational consistency objective, preserving structural relationships beyond individual correspondences. Together, the two stages efficiently refine pretrained VFM features from a single image pair. Extensive experiments demonstrate consistent improvements on multi-view matching and pose estimation, while also showing transferability to video tracking and semantic correspondence.

Our contributions are summarized as follows:

- We propose a one-shot adaptation method that enhances the 3D correspondence understanding of VFMs using only a single overlapping image pair with 3D correspondences, enabling effective adaptation under extremely limited correspondence supervision.
- We design a lightweight two-stage VFM adapter that jointly improves local correspondence features and cross-view relational structure through attention-guided refinement and relational metric adaptation.
- Extensive experiments show that our approach improves diverse VFMs across multi-view correspondence and pose estimation, and achieves competitive performance against existing VFM fine-tuning methods with minimal parameter overhead.

2 Related Work

2.1 Vision Foundation Models and 3D Extensions

The emergence of VFMs has marked a paradigm shift in computer vision. Built largely on ViT architectures (Dosovitskiy et al., 2020), these models learn highly transferable representations from

large-scale data. Multimodal models such as CLIP (Radford et al., 2021) align visual and textual concepts for zero-shot recognition and semantic search, while vision models including DINO (Caron et al., 2021), MAE (He et al., 2022a), DINOv2 (Oquab et al., 2023), and DeiT-III (Touvron et al., 2022) provide dense visual features transferable to classification, segmentation, depth estimation, and correspondence.

Specialized foundation models further extend these capabilities. SAM 2 (Ravi et al., 2025) targets image and video segmentation, while MiDaS (Ranftl et al., 2020) and Depth Anything V2 (Yang et al., 2024b) support robust monocular depth estimation. In 3D vision, DUSt3R (Wang et al., 2024), VGGT (Wang et al., 2025), and FoundationStereo (Wen et al., 2025) extend large-scale representation learning to geometric reconstruction, stereo understanding, and multi-view correspondence, highlighting the potential of foundation models for spatial and structural reasoning.

2.2 Probing 3D Awareness in Foundation Models

A growing body of research investigates the latent 3D awareness of VFMs beyond conventional 2D perceptual capabilities. El Banani et al. (2024) evaluate capabilities in monocular depth and multi-view correspondence. SpatialVLM (Chen et al., 2024) investigates spatial reasoning abilities of vision-language models. Zhan et al. (2024) evaluate physical relationship understanding via classification tasks, and Feat2GS (Chen et al., 2025) extends probing to texture consistency through 3D Gaussian splatting (Kerbl et al., 2023). Together, these studies indicate that VFMs encode geometric priors. However, such priors are primarily inherited from large-scale 2D pretraining and are not explicitly optimized for dense cross-view geometric consistency, limiting their ability to produce reliable correspondences under viewpoint changes. This motivates adaptation techniques that explicitly enhance 3D correspondence.

2.3 Adapting Foundation Models for 3D Tasks

Beyond probing intrinsic properties, recent works adapt VFMs for geometric reasoning. FiT3D (Yue et al., 2024) lifts 2D features into 3D representations through 3D-aware fine-tuning, while DVT (Yang et al., 2024a) reduces position-dependent artifacts via denoising-based adaptation. You et al. (2025) demonstrate that correspondence supervision improves cross-view feature equivariance through feature fine-tuning, and Shape-of-You (Im et al., 2026) exploits structural relationships for semantic correspondence under challenging conditions. However, how to effectively exploit the limited supervision provided by a small number of correspondence pairs remains underexplored.

In contrast, our method adapts VFMs from a single correspondence-annotated image pair by exploiting both local matching constraints and cross-view relational consistency.

3 Method

Our approach adapts VFMs to enhance 3D correspondence using a single overlapping image pair with given correspondences. A lightweight two-stage adapter first refines local features through attention-gated convolution and contrastive supervision, then aligns cross-view relational structure through a shared projection. Figure 2 provides an overview of the approach. Additional implementation details are provided in Appendix A.

3.1 Data Preprocessing

To adapt VFMs for 3D correspondence reasoning, we consider a one-shot setting where a single overlapping image pair and its cross-view geometric correspondences are available during adaptation. Given images $I_0, I_1 \in \mathbb{R}^{H\times W\times C}$, the VFM produces token features $F_0, F_1 \in \mathbb{R}^{\frac{H}{P}\times\frac{W}{P}\times d}$, where P is the patch size and d is the feature dimension. A lightweight adapter then refines these pretrained features toward improved cross-view correspondence consistency. In our experiments, cross-view correspondences are derived from the available geometric annotations of each dataset.

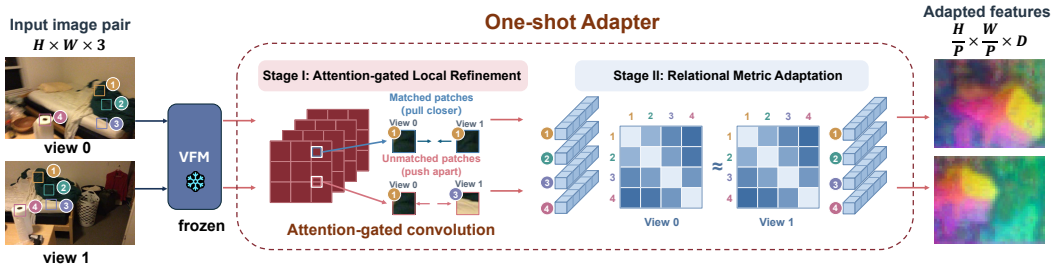


Figure 2: Overview of our pipeline. Given an overlapping image pair and its cross-view correspondences, the proposed two-stage adapter is attached to the frozen VFM. Stage I performs attention-gated local refinement with contrastive supervision to improve patch-level correspondence consistency, while Stage II learns a shared projection that preserves pairwise relational structure across views. Only the adapter is optimized, while the VFM backbone remains frozen.

3.2 Adapter Training

Adapting a VFM using only a single image pair provides limited supervision. We therefore exploit the available cross-view correspondences at progressively richer levels. In Stage I, matched patches are contrasted with multiple non-matching patches, pulling corresponding features closer while pushing mismatched ones apart. Stage II exploits pairwise relationships among correspondence features, providing additional relational constraints to preserve feature-space structure across viewpoints.

Given an overlapping image pair with cross-view correspondences $\mathcal{C} = (\mathbf{p}_i^0, \mathbf{p}_i^1)_{i=1}^N$, we design a lightweight two-stage adapter. During training, only the adapter parameters are optimized sequentially, while the pretrained VFM backbone remains frozen.

Stage I: Attention-gated local refinement. Stage I aims to improve cross-view patch consistency under limited one-pair supervision. While convolution provides effective local aggregation, its kernel weights are shared across spatial locations and do not account for content-dependent variations in local neighborhoods. We therefore introduce an attention gate that adaptively modulates the contributions of different convolutional offsets based on local feature context, enabling content-aware local refinement.

Let $F^v \in \mathbb{R}^{H' \times W' \times d}$ denote the frozen VFM feature map of view $v \in \{0, 1\}$. The refined feature map is obtained as

$$\widetilde{F}^v = \Phi_\theta(F^v), \tag{1}$$

where Φ_θ computes local attention within each 3×3 neighborhood and uses the resulting weights to adaptively modulate the contributions of different convolutional offsets. This enables content-dependent local refinement while retaining the spatial inductive bias of convolution.

The refined features are optimized with a bidirectional contrastive objective. For each correspondence i, the matched cross-view features form a positive pair, while M non-matching patches are sampled as negatives. For the direction from view 0 to view 1, the loss is

$$\mathcal{L}_{\text{con}}^{0\to1} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp\left(s_i^+/\tau\right)}{\exp\left(s_i^+/\tau\right) + \sum_{m=1}^{M} \exp\left(s_{im}^-/\tau\right)}, \tag{2}$$

where N is the number of correspondences, s_i^+ denotes the cosine similarity between the matched cross-view features of correspondence i, s_{im}^- denotes the cosine similarity between the anchor feature from view 0 and the m-th non-matching feature from view 1. The reverse direction is computed symmetrically, and the final Stage-I objective is

$$\mathcal{L}_{\text{con}} = \mathcal{L}_{\text{con}}^{0\to1} + \mathcal{L}_{\text{con}}^{1\to0}. \tag{3}$$

By contrasting each matched feature against multiple non-matching features, the correspondence supervision is converted into multiple cross-view discriminative constraints. This encourages corresponding patches to remain similar across viewpoints while pushing non-corresponding patches

apart. During this stage, the attention-gated convolution module is optimized for 200 iterations, while the pretrained VFM backbone remains frozen.

Stage II: Relational metric adaptation. Although Stage I improves patch-to-patch correspondence consistency, it does not explicitly preserve the pairwise relations among corresponding features across views. Stage II therefore learns a shared projection applied to the Stage-I features to align these relational structures, extending adaptation from local patch alignment to cross-view relational consistency.

Let $\mathbf{f}_i^v$ denote the Stage-I feature of the i-th matched patch in view v. We learn a shared projection matrix $L \in \mathbb{R}^{d\times d}$, which is optimized so that the distance between two patches within one view remains consistent across viewpoint changes. To reduce parameter overhead while keeping the transformation nonsingular, L is parameterized as a lower-triangular matrix with a positive diagonal. The projection is initialized to the identity matrix, allowing Stage II to start from the Stage-I representation. The projected features are then ℓ_2-normalized before relational comparison.

For a set of N matched patches, we characterize the relational structure within each view using the complete pairwise Euclidean distance matrix:

$$D_{ij}^v = \left|\mathbf{z}_i^v - \mathbf{z}_j^v\right|_2, \qquad D^v \in \mathbb{R}^{N\times N}, \tag{4}$$

where $\mathbf{z}_i^v$ denotes the normalized feature obtained by projecting $\mathbf{f}_i^v$ with $L^\top$. Since the overall pairwise distance scale may vary across views, we normalize each distance matrix by its mean off-diagonal value before relational comparison. This removes global scale differences while preserving the relative relational structure within each view. We denote the normalized distance matrix as $\widehat{D}^v$ and enforce consistency between the pairwise relations in the two views:

$$\mathcal{L}_{\text{rel}} = \frac{1}{N(N-1)} \sum_{i\neq j} \left(\widehat{D}_{ij}^0 - \widehat{D}_{ij}^1\right)^2. \tag{5}$$

By comparing all within-view patch pairs, this objective encourages the relative organization of the transformed features to remain consistent across viewpoints, extending the patch-level supervision of Stage I to relational structure consistency.

Since Stage II is adapted from only a single image pair, the learned projection may overfit the observed relational structure. We therefore apply LogDet regularization (Davis et al., 2007) to constrain excessive changes in the learned projection:

$$\mathcal{L}_{\text{logdet}} = \text{tr}(LL^\top) - \log\det(LL^\top) - d. \tag{6}$$

This regularization discourages unnecessary distortion of the Stage-I feature space while allowing L to adapt under the relational supervision.

The final Stage-II objective is

$$\mathcal{L}_{\text{stage2}} = \mathcal{L}_{\text{rel}} + \lambda_{\text{logdet}}\mathcal{L}_{\text{logdet}}. \tag{7}$$

4 Experiments

We first introduce the 3D geometric tasks considered in our evaluation and detailed experimental setup. We then organize the results into three parts: an evaluation of diverse raw VFMs, a comparison with recent VFM adaptation methods under a unified backbone, and ablation studies.

4.1 Task Selection

Following prior work (You et al., 2025; El Banani et al., 2024), we evaluate 3D correspondence understanding across multiple downstream tasks and datasets:

1. **Multi-view correspondence estimation** (El Banani et al., 2024) is evaluated on the scene-level ScanNet (Dai et al., 2017) and object-level NAVI (Jampani et al., 2023) datasets. Dense feature maps are extracted from image pairs, and correspondences are established via cosine similarity following keypoint-free pipelines (El Banani & Johnson, 2021; Sun et al., 2021). We report correspondence recall under pixel- and meter-based error thresholds for scenes and objects, respectively.

Table 1: Comparison across different vision foundation models. All results are reported as percentages (%). Higher values indicate better performance.

Model	Multi-view Correspondence						Pose Estimation					
	Scene			Object			Scene			Object		
	R@5px	R@10px	R@20px	R@1cm	R@3cm	R@5cm	5deg	10deg	15deg	1cm-1deg	3cm-3deg	5cm-5deg
CLIP	0.67	2.42	9.13	7.26	32.01	57.20	2.00	11.87	27.27	0.15	1.55	4.22
Ours	**0.74**	**2.56**	**9.25**	**9.55**	**34.90**	**59.85**	**2.08**	**12.73**	**29.13**	**0.20**	**3.89**	**10.10**
Gain(%)	+10.45	+5.79	+1.31	+31.54	+9.03	+4.63	+4.00	+7.25	+6.82	+33.33	+150.97	+139.34
MAE	1.97	4.18	11.08	14.36	43.45	67.57	2.13	14.00	31.27	0.35	4.58	9.80
Ours	**2.26**	**4.43**	**11.26**	**16.16**	**44.65**	**68.30**	**2.35**	**14.33**	**31.43**	**0.67**	**7.87**	**19.37**
Gain(%)	+14.72	+5.98	+1.62	+12.53	+2.76	+1.08	+10.33	+2.36	+0.51	+91.43	+71.83	+97.65
DeiT-III	6.37	12.21	20.54	14.86	43.67	65.92	3.87	11.20	19.07	0.19	16.04	31.92
Ours	**8.25**	**12.51**	**20.69**	**18.70**	**45.62**	**66.26**	**4.00**	**14.93**	**31.00**	**2.26**	**23.31**	**42.10**
Gain(%)	+29.51	+2.46	+0.73	+25.84	+4.47	+0.52	+3.36	+33.30	+62.56	+1089.47	+45.32	+31.89
DINOv2	17.42	26.64	38.31	35.63	67.00	82.45	11.04	23.98	37.80	6.98	48.15	68.60
Ours	**31.57**	**40.54**	**50.72**	**40.79**	**70.49**	**85.12**	**17.73**	**32.93**	**46.56**	**10.87**	**55.79**	**79.22**
Gain(%)	+81.23	+52.18	+32.39	+14.48	+5.21	+3.24	+60.60	+37.32	+23.17	+55.73	+15.87	+15.48
DINOv2-reg	20.94	31.14	43.09	33.71	65.81	81.74	14.53	29.27	43.67	9.50	49.87	73.02
Ours	**36.24**	**45.85**	**55.95**	**37.20**	**67.99**	**83.43**	**23.67**	**41.80**	**53.47**	**9.85**	**60.19**	**82.31**
Gain(%)	+73.07	+47.24	+29.84	+10.35	+3.31	+2.07	+62.90	+42.81	+22.44	+3.68	+20.69	+12.72

2. **Pose estimation** is evaluated on ScanNet and OnePose-LowTex (Sun et al., 2022; He et al., 2022b) for scene- and object-level pose estimation, respectively. Feature correspondences are passed to the corresponding robust pose-estimation pipelines. We report angular-threshold recall on ScanNet and pose recall at 1cm-1deg, 3cm-3deg, and 5cm-5deg on OnePose-LowTex.

3. **Video tracking** is conducted on TAP-Vid-DAVIS (Doersch et al., 2022), which evaluates the capability of VFMs to trace moving objects in videos. Following Tumanyan et al. (2024), tracking is performed by nearest-neighbor search in the feature space using cosine similarity. We adopt Average Jaccard Index (AJ) and Occlusion Accuracy (OA).

4. **Semantic correspondence** is evaluated on PF-PASCAL (Ham et al., 2017). Given a keypoint in a source image, its correspondence in another instance of the same semantic category is predicted by maximizing cosine similarity in the feature space. Following prior work (You et al., 2025), performance is measured using Percentage of Correct Keypoints (PCK) under normalized distance thresholds.

The evaluated tasks serve two complementary purposes. Multi-view correspondence and pose estimation constitute our geometric evaluation, directly measuring one-shot adaptation under multi-view geometry. Video tracking and semantic correspondence, in contrast, serve as transfer evaluations, testing whether the adapted representation generalizes beyond the adaptation domain to temporal and cross-instance correspondence without additional target-task adaptation.

4.2 Implementation Details

For Stage I, we set the number of negative tokens $M = 2$, the contrastive temperature $\tau = 0.1$, and weight decay to 10^{-4}, while keeping the pretrained VFM backbone frozen. For Stage II, the metric transformation is initialized to the identity matrix and optimized with AdamW using a learning rate of 3×10^{-4} and $\lambda_{\text{logdet}} = 10^{-3}$.

For each geometric task, we construct candidate overlapping image pairs with relative rotations of 0.4–$0.6\times$ the horizontal camera field of view. Adaptation pairs are strictly separated from evaluation samples. We randomly sample 10 pairs, perform one-shot adaptation independently on each pair, and report the average results. For transfer tasks, the NAVI-adapted model is directly evaluated on TAP-Vid-DAVIS and PF-PASCAL without additional adaptation.

4.3 Comparison with Pretrained VFMs

We compare models adapted with our method against CLIP (Radford et al., 2021), MAE (He et al., 2022a), DeiT-III (Touvron et al., 2022), DINOv2 (Oquab et al., 2023), and DINOv2-reg (Darcet et al., 2024). Tables 1 and 2 report results across tasks.

Table 2: Cross-task transfer evaluation on video tracking and semantic correspondence. The adapted models are directly evaluated on TAP-Vid-DAVIS and PF-PASCAL without additional target-dataset adaptation. All results are reported as percentages (%). Higher values indicate better performance.

	Video Tracking		Semantic Correspondence		
	OA	AJ	PCK@0.05	PCK@0.10	PCK@0.15
CLIP	**79.60**	**6.37**	2.28	5.64	11.93
Ours	70.04	5.95	**2.99**	**8.02**	**13.49**
Gain(%)	-12.01	-6.59	+31.14	+42.20	+13.08
MAE	**80.92**	26.34	**13.37**	25.45	**37.83**
Ours	80.23	**27.45**	13.01	**25.52**	36.66
Gain(%)	-0.85	+4.21	-2.69	+0.28	-3.09
DeiT-III	81.37	21.60	6.47	13.98	20.53
Ours	**82.34**	**27.89**	**8.05**	**16.87**	**26.19**
Gain(%)	+1.19	+29.12	+24.42	+20.67	+27.57
DINOv2	80.93	39.54	53.86	72.14	79.93
Ours	**84.08**	**41.74**	**56.81**	**75.26**	**83.26**
Gain(%)	+3.89	+5.56	+5.48	+4.32	+4.17
DINOv2-reg	80.41	33.97	43.55	60.62	69.63
Ours	**81.04**	**38.18**	**48.78**	**67.08**	**77.45**
Gain(%)	+0.78	+12.39	+12.01	+10.66	+11.23

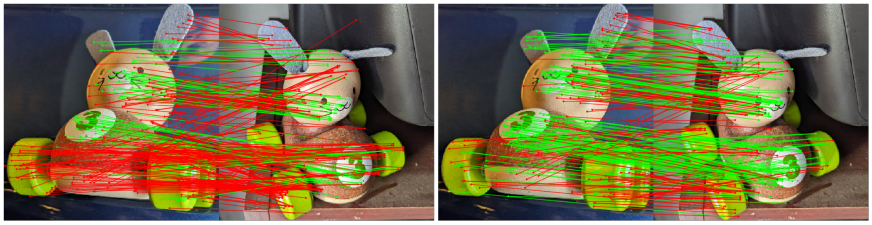


Figure 3: 3D correspondence improvement. Left: correspondences estimated by DINOv2. Right: adapted DINOv2. Green and red lines indicate correct and incorrect matches, respectively.

Our approach consistently improves pretrained VFMs on 3D geometric tasks, including multi-view correspondence and pose estimation. The gains are particularly notable for stronger visual backbones such as DINOv2 and DINOv2-reg, suggesting that the one-shot adapter complements powerful pretrained representations while enhancing geometric correspondence. These improvements require only lightweight adaptation from a single image pair, with the VFM backbone frozen. Figures 3 and 4 further illustrate these gains: the adapted features produce more accurate cross-view matches with fewer errors and yield clearer object-part distinctions with more spatially consistent responses in both single-view and multi-view settings.

For video tracking and semantic correspondence, we evaluate whether the adapted representations transfer beyond the static multi-view geometry used during adaptation. Video tracking involves temporal dynamics and object motion, introducing challenges not explicitly considered in the adaptation setting and therefore leading to varying improvements across VFMs. In contrast, semantic correspondence between instances of the same category is more closely related to correspondence matching and generally benefits more from the adapted features. These results suggest that the proposed one-shot adaptation can transfer beyond the original geometric adaptation domain.

We further show that one-shot adaptation using an external CO3Dv2 (Reizenstein et al., 2021) image pair improves the pretrained DINOv2-B across the downstream evaluation suite, without any target-dataset adaptation in Appendix B.

4.4 Comparison with VFM-Tuning Methods

To ensure fairness, we benchmark against representative VFM-tuning frameworks under a common DINOv2 backbone, as these methods have been fine-tuned on either DINOv2-Small or DINOv2-

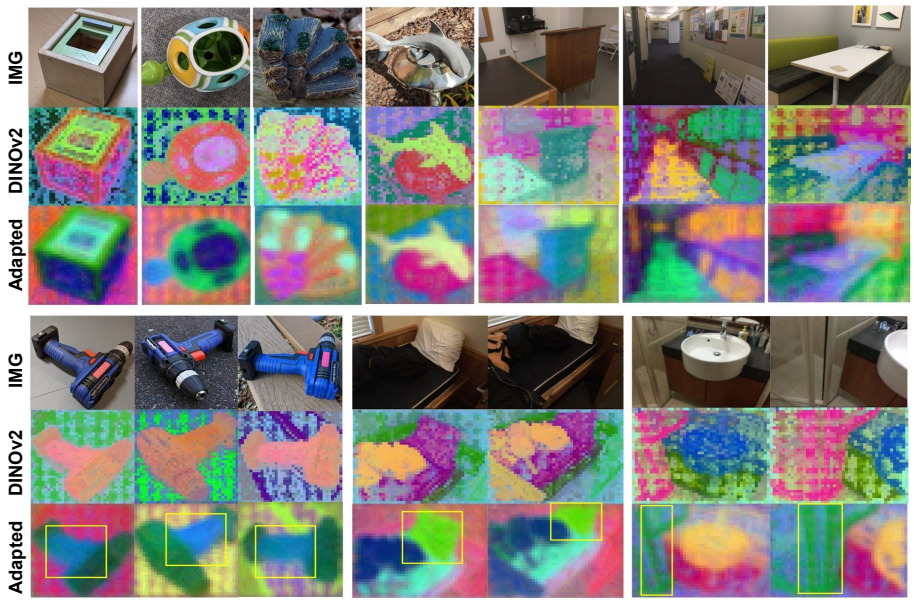


Figure 4: Qualitative results. The first group shows single-view feature maps, and the second group shows multi-view feature maps. Each group contains three rows: the first row shows the input image, the second row shows DINOv2 feature maps, and the third row shows feature maps from our adaptation model. Our method exhibits fewer artifacts and clearer distinctions between object parts, thereby benefiting 3D correspondence tasks.

Table 3: Comparison with representative VFM fine-tuning methods using matched vision foundation model backbones. S denotes a scene dataset and O an object dataset. Best results are in **bold**, and second-best results are underlined.

	MV Corr. (S)		MV Corr. (O)		Pose (S)		Pose (O)		Tracking		Semantic Corr.	
Model	R@5px	R@20px	R@1cm	R@5cm	5deg	15deg	1cm-1deg	5cm-5deg	OA	AJ	PCK@.05	PCK@.15
DINOv2-S	13.35	33.41	30.80	80.71	7.53	34.53	7.12	67.98	80.63	37.62	53.59	79.97
FiT3D	9.32	26.85	29.37	78.87	5.27	31.67	9.02	66.89	78.92	34.24	**57.39**	**84.54**
DVT	16.68	39.62	29.55	79.27	11.67	34.60	5.59	61.46	80.28	36.63	37.96	67.09
3DCorrEnhance	**24.68**	41.69	**40.62**	83.03	13.00	40.13	8.73	**85.25**	**86.42**	**44.12**	55.61	82.04
Ours (S)	23.94	**42.75**	35.31	**83.38**	**13.07**	**40.53**	**9.68**	76.23	82.20	39.66	51.69	81.19
DINOv2-B	17.42	38.31	35.63	82.45	11.04	37.80	6.98	68.60	80.93	39.54	53.86	79.93
FiT3D	11.12	29.28	33.61	81.57	6.20	33.80	6.78	79.02	81.55	40.41	59.63	**87.78**
DVT	17.09	37.32	35.14	81.37	9.20	34.53	6.27	70.22	81.13	39.56	50.45	75.97
3DCorrEnhance	31.25	49.94	**43.48**	85.02	16.87	45.67	9.12	**88.51**	**86.89**	**44.56**	**62.56**	87.18
Ours (B)	**31.57**	**50.72**	40.79	**85.12**	**17.73**	**46.56**	**10.87**	79.22	84.08	41.74	56.81	83.26

Base. The comparison includes FiT3D (Yue et al., 2024), DVT (Yang et al., 2024a), and 3DCorrEnhance (You et al., 2025). The results are shown in Table 3.

On multi-view 3D tasks, our method consistently outperforms FiT3D (Yue et al., 2024) and DVT (Yang et al., 2024a). Under the same evaluation protocols and backbones, our one-shot adaptation achieves stronger scene- and object-level correspondence and pose estimation, demonstrating that lightweight adaptation can enhance pretrained representations with limited supervision.

Compared with 3DCorrEnhance (You et al., 2025), our main comparison follows its reported dataset-scale adaptation setting, whereas our method performs adaptation using a single correspondence-annotated image pair. 3DCorrEnhance performs strongly on object-centric tasks, whereas our method achieves comparable or better results on scene-level correspondence and pose estimation. A controlled one-pair comparison in Appendix E further shows that our method consistently outperforms one-shot 3DCorrEnhance, demonstrating the effectiveness of our one-shot adapter under limited correspondence supervision.

For video tracking and semantic correspondence, we evaluate the transferability of the adapted representations without additional target-task adaptation. Our method remains competitive with exist-

Table 4: Comparison of different VFM adaptation methods in terms of parameter count, inference speed, and adaptation cost.

Model	Params	Time/img	Adapt.
DINOv2	86.6M	31.6ms	–
FiT3D	173.1M	59.4ms	8h
DVT	94.7M	32.6ms	2h
3DCorrEnhance	91.9M	32.0ms	2h
Ours	92.3M	31.9ms	2min

Table 5: Ablation study of the proposed two-stage adapter on multi-view correspondence.

Setting	R@5px	R@20px	R@1cm	R@5cm
Baseline	17.42	38.31	35.63	82.45
Stage I	29.73	47.81	38.86	84.14
Stage II	22.60	41.95	37.31	83.09
Both	**31.57**	**50.72**	**40.79**	**85.12**

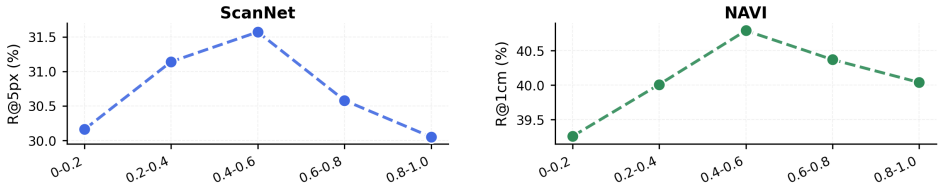


Figure 5: Effect of adaptation-pair selection across relative rotation-to-FOV ratio intervals. Left: R@5px on ScanNet. Right: R@1cm on NAVI.

ing adaptation approaches while requiring substantially less adaptation data. Overall, these results highlight the complementary advantages of our lightweight one-shot strategy: it enables efficient geometric adaptation while maintaining competitive performance across diverse correspondence-related tasks. A further comparison with Shape-of-You (Im et al., 2026), which uses a different adaptation setting from ours, is provided in Appendix D.

Table 4 compares computational efficiency under a unified evaluation setup. All runtime measurements are obtained on a single NVIDIA A100 GPU using the same input resolution and inference settings. Our adapter introduces only a small parameter overhead with negligible impact on inference speed. Under the same hardware setting, one-shot adaptation completes within minutes, substantially reducing the adaptation cost compared with dataset-scale VFM tuning methods.

4.5 Ablation Studies

We conduct ablation studies on ScanNet and NAVI multi-view correspondence to assess each component (Table 5). Stage I yields the largest individual gain through attention-gated local refinement and patch-level contrastive supervision, while Stage II also improves the baseline by promoting cross-view consistency of relational structure. Combining both stages performs best, showing that local refinement and relational consistency provide complementary benefits.

We further study adaptation-pair selection, which trades off correspondence reliability and viewpoint diversity: small rotations provide more reliable matches but limited geometric variation, whereas large rotations reduce valid overlap. As shown in Figure 5, performance first improves and then declines as the rotation-to-FOV ratio increases. We therefore use pairs with relative rotations of 0.4–0.6 times the horizontal field of view. Additional ablations are provided in Appendix C.

5 Conclusion

We introduce a lightweight one-shot adapter for 3D correspondence that uses a single overlapping image pair while keeping the VFM backbone frozen. The method combines local feature refinement with relational metric adaptation, improving multi-view correspondence and pose estimation while transferring to video tracking and semantic correspondence. Ablations validate the complementary roles of both stages and the importance of pair selection.

Future work will explore broader low-resource settings, including few-shot adaptation and scenarios without explicit geometric annotations, while incorporating richer semantic cues for more diverse correspondence tasks.

AI USE STATEMENT

Generative AI tools were used solely to improve the language, clarity, and organization of the manuscript. The authors reviewed and revised all AI-assisted text and take full responsibility for the content of the paper.

REFERENCES

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *2021 IEEE/CVF international conference on computer vision (ICCV)*, pp. 9630–9640. IEEE, 2021.

Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14455–14465, 2024.

Yue Chen, Xingyu Chen, Anpei Chen, Gerard Pons-Moll, and Yuliang Xiu. Feat2gs: Probing visual foundation models with gaussian splatting. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6348–6361. IEEE, 2025.

Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *2017 IEEE conference on computer vision and pattern recognition (CVPR)*, pp. 2432–2443. IEEE, 2017.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *International conference on learning representations*, volume 2024, pp. 2632–2652, 2024.

Jason V Davis, Brian Kulis, Prateek Jain, Suvrit Sra, and Inderjit S Dhillon. Information-theoretic metric learning. In *Proceedings of the 24th international conference on Machine learning*, pp. 209–216, 2007.

Carl Doersch, Ankush Gupta, Larisa Markeeva, Adria Recasens, Lucas Smaira, Yusuf Aytar, Joao Carreira, Andrew Zisserman, and Yi Yang. Tap-vid: A benchmark for tracking any point in a video. *Advances in Neural Information Processing Systems*, 35:13610–13626, 2022.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Mohamed El Banani and Justin Johnson. Bootstrap your own correspondences. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6413–6422. IEEE, 2021.

Mohamed El Banani, Amit Raj, Kevis-Kokitsi Maninis, Abhishek Kar, Yuanzhen Li, Michael Rubinstein, Deqing Sun, Leonidas Guibas, Justin Johnson, and Varun Jampani. Probing the 3d awareness of visual foundation models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21795–21806. IEEE, 2024.

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, 2017.

Ali Hassani, Steven Walton, Jiachen Li, Shen Li, and Humphrey Shi. Neighborhood attention transformer. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6185–6194. IEEE, 2023.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 15979–15988. IEEE, 2022a.

Xingyi He, Jiaming Sun, Yuang Wang, Di Huang, Hujun Bao, and Xiaowei Zhou. Onepose++: Keypoint-free one-shot object pose estimation without cad models. *Advances in Neural Information Processing Systems*, 35:35103–35115, 2022b.

Jiin Im, Sisung Liu, and Je Hyeong Hong. Shape-of-you: Fused gromov-wasserstein optimal transport for semantic correspondence in-the-wild. *arXiv preprint arXiv:2603.11618*, 2026.

Varun Jampani, Kevis-Kokitsi Maninis, Andreas Engelhardt, Arjun Karpur, Karen Truong, Kyle Sargent, Stefan Popov, André Araujo, Ricardo Martin Brualla, Kaushal Patel, et al. Navi: Category-agnostic image collections with high-quality 3d shape and pose annotations. *Advances in Neural Information Processing Systems*, 36:76061–76084, 2023.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, George Drettakis, et al. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *2023 IEEE/CVF international conference on computer vision (ICCV)*, pp. 3992–4003. IEEE, 2023.

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, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE transactions on pattern analysis and machine intelligence*, 44(3):1623–1637, 2020.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. In *International Conference on Learning Representations*, volume 2025, pp. 28085–28128, 2025.

Jeremy Reizenstein, Roman Shapovalov, Philipp Henzler, Luca Sbordone, Patrick Labatut, and David Novotny. Common objects in 3d: Large-scale learning and evaluation of real-life 3d category reconstruction. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10881–10891. IEEE, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pp. 10674–10685. ieee, 2022.

Jiaming Sun, Zehong Shen, Yuang Wang, Hujun Bao, and Xiaowei Zhou. Loftr: Detector-free local feature matching with transformers. *arXiv preprint arXiv:2104.00680*, 2021.

Jiaming Sun, Zihao Wang, Siyu Zhang, Xingyi He, Hongcheng Zhao, Guofeng Zhang, and Xiaowei Zhou. Onepose: One-shot object pose estimation without cad models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6815–6824. IEEE, 2022.

Hugo Touvron, Matthieu Cord, and Hervé Jégou. Deit iii: Revenge of the vit. In *European conference on computer vision*, pp. 516–533. Springer, 2022.

Narek Tumanyan, Assaf Singer, Shai Bagon, and Tali Dekel. Dino-tracker: Taming dino for self-supervised point tracking in a single video. In *European Conference on Computer Vision*, pp. 367–385. Springer, 2024.

Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggt: Visual geometry grounded transformer. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5294–5306. IEEE, 2025.

Shuzhe Wang, Vincent Leroy, Yohann Cabon, Boris Chidlovskii, and Jerome Revaud. Dust3r: Geometric 3d vision made easy. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20697–20709. IEEE, 2024.

Bowen Wen, Matthew Trepte, Joseph Aribido, Jan Kautz, Orazio Gallo, and Stan Birchfield. Foundationstereo: Zero-shot stereo matching. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5249–5260. IEEE, 2025.

Jiawei Yang, Katie Z Luo, Jiefeng Li, Congyue Deng, Leonidas Guibas, Dilip Krishnan, Kilian Q Weinberger, Yonglong Tian, and Yue Wang. Denoising vision transformers. In *European Conference on Computer Vision*, pp. 453–469. Springer, 2024a.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in neural information processing systems*, 37:21875–21911, 2024b.

Yang You, Yixin Li, Congyue Deng, Yue Wang, and Leonidas Guibas. Multiview equivariance improves 3d correspondence understanding with minimal feature finetuning. In *International Conference on Learning Representations*, volume 2025, pp. 58081–58094, 2025.

Yuanwen Yue, Anurag Das, Francis Engelmann, Siyu Tang, and Jan Eric Lenssen. Improving 2d feature representations by 3d-aware fine-tuning. In *European Conference on Computer Vision*, pp. 57–74. Springer, 2024.

Guanqi Zhan, Chuanxia Zheng, Weidi Xie, and Andrew Zisserman. A general protocol to probe large vision models for 3d physical understanding. *Advances in Neural Information Processing Systems*, 37:43468–43498, 2024.

A Implementation Details

A.1 Attention-Gated Local Refinement

We provide additional implementation details for the attention-gated convolution used in Stage I. The same refinement module is shared between the two views. For clarity, we omit the view superscript and denote the frozen VFM feature at spatial location $\mathbf{p}$ by $\mathbf{f}_{\mathbf{p}} \in \mathbb{R}^d$.

Local attention. We first apply channel-wise LayerNorm and obtain query and key embeddings using two bias-free 1×1 projections:

$$\mathbf{q}_{\mathbf{p}} = W_q \,\mathrm{LN}(\mathbf{f}_{\mathbf{p}}), \qquad \mathbf{k}_{\mathbf{p}} = W_k \,\mathrm{LN}(\mathbf{f}_{\mathbf{p}}). \tag{8}$$

The query/key dimension is $D_a = \frac{D}{8}$ and is divided into $H_a = 8$ heads, each with dimension $d_a = \frac{D_a}{H_a}$.

Let $\mathcal{N} = \{-1, 0, 1\} \times \{-1, 0, 1\}$ denote the offsets of a 3×3 neighborhood. For head h, the attention score for offset $\boldsymbol{\delta} \in \mathcal{N}$ is

$$e^h_{\mathbf{p},\boldsymbol{\delta}} = \frac{(\mathbf{q}^h_{\mathbf{p}})^\top \mathbf{k}^h_{\mathbf{p}+\boldsymbol{\delta}}}{\sqrt{d_a}} + b^h_{\boldsymbol{\delta}}, \tag{9}$$

where $b^h_{\boldsymbol{\delta}}$ is a learnable relative-position bias. Invalid neighbors at image boundaries are masked before normalization:

$$a^h_{\mathbf{p},\boldsymbol{\delta}} = \frac{\exp(e^h_{\mathbf{p},\boldsymbol{\delta}})}{\sum_{\boldsymbol{\delta}' \in \mathcal{N}_{\mathbf{p}}} \exp(e^h_{\mathbf{p},\boldsymbol{\delta}'})}, \tag{10}$$

where $\mathcal{N}_{\mathbf{p}}$ contains the valid offsets at location $\mathbf{p}$.

Attention-modulated convolution. Rather than replacing convolution with attention, we use the attention weights to modulate the spatial contributions of a standard convolution. Specifically, the gate for head h is defined as

$$g^h_{\mathbf{p},\boldsymbol{\delta}} = 1 + \beta_h \left(9 a^h_{\mathbf{p},\boldsymbol{\delta}} - 1\right), \tag{11}$$

where β_h is a learnable scalar.

For each offset, let $W_{\boldsymbol{\delta}} \in \mathbb{R}^{D \times D}$ denote the corresponding slice of a bias-free 3×3 convolution kernel, and let S_h select the D/H_a output channels associated with head h. The refined feature is

$$\widetilde{\mathbf{f}}_{\mathbf{p}} = \mathrm{Concat}_{h=1}^{H_a} \left[\sum_{\boldsymbol{\delta} \in \mathcal{N}_{\mathbf{p}}} g^h_{\mathbf{p},\boldsymbol{\delta}} S_h \left(W_{\boldsymbol{\delta}} \mathbf{f}_{\mathbf{p}+\boldsymbol{\delta}}\right) \right]. \tag{12}$$

The convolution mixes all input channels before the output is partitioned into heads. The module is a full convolution with head-wise spatial gates rather than a depthwise or grouped convolution. LayerNorm is used only in the query and key branch, while the convolution operates directly on the backbone features.

The convolution kernel is initialized with Kaiming uniform initialization, while the relative-position biases and β_h are initialized to zero. Therefore, $g^h_{\mathbf{p},\boldsymbol{\delta}} = 1$ at initialization, and the module initially behaves as a standard bias-free 3×3 convolution. The convolution branch uses zero padding, while invalid neighbors are excluded from the attention softmax.

A.2 Correspondence Feature Sampling

The supplied correspondence coordinates are expressed in the preprocessed image coordinate system and do not necessarily coincide with token centers. We therefore obtain correspondence descriptors by bilinearly interpolating the refined feature maps:

$$\widetilde{\mathbf{f}}^v_i = \mathrm{Interp}\left(\widetilde{F}^v, \mathbf{p}^v_i\right). \tag{13}$$

The interpolated descriptors are ℓ_2-normalized before computing cosine similarities in the Stage-I contrastive objective. The same interpolation procedure is used for positive and negative locations, avoiding quantization to the nearest token.

A.3 Sequential Adapter Optimization

The two adaptation stages are optimized sequentially. During Stage I, the VFM backbone remains frozen and only the attention-gated refinement module is updated using the bidirectional contrastive objective. The same refinement parameters are shared by both views.

After Stage I, the refinement module is frozen and its sampled correspondence descriptors are detached from the computational graph. Stage II then optimizes only the shared projection using the relational and LogDet objectives defined in the main paper.

At inference time, the learned Stage-I refinement and Stage-II projection are applied sequentially to the frozen VFM features. No correspondence supervision or additional adapter optimization is required during inference.

Training Configuration. Stage I is optimized using Adam with a learning rate of 1×10^{-4} and weight decay of 1×10^{-4} for 200 iterations, with the learning rate kept fixed. The pretrained VFM backbone is frozen, and only the parameters of the refinement module are updated.

Stage II is optimized using AdamW with a learning rate of 3×10^{-4} and zero weight decay. The optimization runs for at most 300 iterations with a fixed learning rate. Only the parameters of the metric transformation are optimized in this stage, while the backbone and Stage-I refinement module remain frozen.

B Cross-dataset One-shot Adaptation

To examine whether the proposed adaptation relies on target-dataset information, we perform one-shot adaptation using image pairs sampled exclusively from CO3Dv2 (Reizenstein et al., 2021). The adapted DINOv2-B model is directly evaluated on downstream benchmarks without any additional target-dataset adaptation.

As shown in Table 6, external one-shot adaptation consistently improves the pretrained DINOv2-B baseline across geometric and correspondence tasks. In particular, ScanNet correspondence improves substantially, with R@5px increasing from 17.42 to 29.14 and R@20px from 38.31 to 47.55. The improvements also transfer to object-level pose estimation, video tracking, and semantic correspondence, providing evidence that the learned correspondence refinement can generalize beyond the adaptation source and target-specific settings.

Table 6: Cross-dataset one-shot adaptation using CO3Dv2 as the external adaptation source. The adapted DINOv2-B model is evaluated directly on downstream benchmarks without target-dataset adaptation.

Model	MV Corr. (S)		MV Corr. (O)		Pose (S)		Pose (O)		Tracking		Semantic Corr.	
	R@5px	R@20px	R@1cm	R@5cm	5deg	15deg	1cm-1deg	5cm-5deg	OA	AJ	PCK@.05	PCK@.15
DINOv2-B	17.42	38.31	35.63	82.45	11.04	37.80	6.98	68.60	80.93	39.54	53.86	79.93
CO3Dv2 Adapt.	**29.14**	**47.55**	**39.96**	**84.28**	**15.80**	**43.74**	**7.72**	**81.04**	**84.32**	**41.72**	**57.14**	**84.23**
Gain(%)	+67.28	+24.12	+12.15	+2.22	+43.12	+15.71	+10.60	+18.13	+4.19	+5.51	+6.09	+5.38

C Architecture Ablation

Since our adapter is trained from only a single image pair, the available supervision is inherently limited. We therefore restrict the adapter to lightweight architectures to reduce the risk of overfitting under one-shot adaptation. Within this lightweight design space, we compare alternative architectures for both stages while keeping the backbone, correspondence supervision, and training protocol unchanged.

For Stage I, we replace the proposed attention-gated convolution with three alternatives: a residual MLP, a standard 3×3 convolution, and a Neighborhood Attention Transformer (NAT) (Hassani et al., 2023). The residual MLP performs token-wise nonlinear adaptation without explicit spatial interaction, while the 3×3 convolution introduces fixed local aggregation. NAT instead models content-dependent interactions within local neighborhoods through neighborhood attention. These

Table 7: Architecture ablation under one-shot adaptation using DINOv2-B. For Stage I ablations, Stage II uses the default linear projection with LogDet regularization. For Stage II ablations, Stage I uses the proposed attention-gated convolution.

Stage I	Stage II	ScanNet		NAVI	
		R@5px	R@20px	R@1cm	R@5cm
Residual MLP	Linear Projection	30.36	47.10	39.14	82.98
3×3 Conv	Linear Projection	31.23	50.47	40.24	84.08
NAT	Linear Projection	30.72	48.68	39.45	84.61
Attention-Gated Conv	Linear Projection w/o LogDet	31.39	50.28	40.60	84.77
Attention-Gated Conv	Residual MLP	31.04	48.21	39.96	84.39
Attention-Gated Conv	**Linear Projection**	**31.57**	**50.72**	**40.79**	**85.12**

comparisons allow us to examine the effect of different local refinement mechanisms under one-pair supervision.

For Stage II, we evaluate two variants of the relational metric adaptation. The first removes the LogDet regularization while keeping the linear projection unchanged. The second replaces the linear projection with a residual MLP and optimizes it using only the relational objective. These comparisons disentangle the effects of metric regularization and transformation complexity under one-shot supervision.

As shown in Table 7, the proposed attention-gated convolution achieves the strongest overall Stage I performance. The standard 3×3 convolution is the closest alternative, indicating that local spatial aggregation is important, while the attention gate provides additional benefit by making this aggregation content-dependent. The residual MLP performs worse across all four metrics, consistent with the lack of explicit local spatial interaction. The NAT also does not improve over the proposed module, suggesting that a more flexible local interaction mechanism is not necessarily advantageous under one-pair supervision.

For Stage II, we separately analyze the effects of metric regularization and transformation complexity. Removing the LogDet regularization while keeping the linear projection unchanged causes a slight performance drop, from 31.57 to 31.39 on ScanNet R@5px and from 40.79 to 40.60 on NAVI R@1cm, indicating that the regularization provides additional stability under one-shot adaptation. Replacing the linear projection with a residual MLP further reduces performance, from 31.57 to 31.04 on ScanNet R@5px and from 40.79 to 39.96 on NAVI R@1cm, suggesting that a more expressive nonlinear transformation is not necessarily beneficial under extremely limited supervision. Together, these results support the use of a lightweight linear metric adaptation with LogDet regularization, which provides sufficient flexibility while avoiding unnecessary complexity in the one-shot setting.

D Comparison with Shape-of-You

We additionally compare our method with Shape-of-You (Im et al., 2026), a recent approach for improving semantic and geometric correspondence. We emphasize that this comparison is not conducted under a matched adaptation setting. Shape-of-You relies on multiple external foundation models and auxiliary representations, including DINOv2 and Stable Diffusion features (Rombach et al., 2022), SAM (Kirillov et al., 2023) masks, and VGGT-based 3D geometry, and its released model is trained on SPair-71k (Min et al., 2019). In contrast, our method adapts a single pretrained VFM using only one overlapping image pair and its cross-view correspondences, without additional foundation models during inference. The results are shown in Table 8.

Despite Shape-of-You leveraging large-scale training data, multiple external foundation models, and auxiliary geometric representations, our method remains competitive across the evaluated tasks while requiring only a single image pair for adaptation. Notably, our approach achieves stronger performance on NAVI correspondence, OnePose pose estimation, and video tracking, demonstrating the effectiveness of one-shot adaptation under substantially more limited supervision.

Table 8: Comparison with Shape-of-You across the downstream evaluation suite. All results are reported as percentages (%). Higher values indicate better performance.

	MV Corr. (S)		MV Corr. (O)		Pose (S)		Pose (O)		Tracking		Semantic Corr.	
Model	R@5px	R@20px	R@1cm	R@5cm	5deg	15deg	1cm-1deg	5cm-5deg	OA	AJ	PCK@.05	PCK@.15
DINOv2-B	17.42	38.31	35.63	82.45	11.04	37.80	6.98	68.60	80.93	39.54	53.86	79.93
Shape-of-You	**35.35**	**57.70**	32.39	80.32	**25.87**	**57.00**	8.41	72.22	80.13	41.49	**72.75**	**92.81**
Ours	31.57	50.72	**40.79**	**85.12**	17.73	46.56	**10.87**	**79.22**	**84.08**	**41.74**	56.81	83.26

Table 9: Comparison with 3DCorrEnhance under the same one-pair adaptation setting using DINOv2-B. For each geometric dataset, both methods are independently adapted using the same 10 sampled image pairs, with each adaptation using only a single image pair. Results are averaged across the 10 adaptations. Their NAVI-adapted models are then evaluated on TAP-Vid-DAVIS and PF-PASCAL without additional adaptation. All results are reported as percentages (%). Higher values indicate better performance.

	MV Corr. (S)		MV Corr. (O)		Pose (S)		Pose (O)		Tracking		Semantic Corr.	
Model	R@5px	R@20px	R@1cm	R@5cm	5deg	15deg	1cm-1deg	5cm-5deg	OA	AJ	PCK@.05	PCK@.15
DINOv2-B	17.42	38.31	35.63	82.45	11.04	37.80	6.98	68.60	80.93	39.54	53.86	79.93
3DCorrEnhance (one-shot)	25.08	42.03	37.02	83.93	15.00	41.53	7.68	73.38	81.39	40.03	54.26	80.82
Ours	**31.57**	**50.72**	**40.79**	**85.12**	**17.73**	**46.56**	**10.87**	**79.22**	**84.08**	**41.74**	**56.81**	**83.26**

E Comparison with 3DCorrEnhance under One-Shot Adaptation

To provide a controlled comparison under the same one-pair supervision budget, we further compare our method with 3DCorrEnhance (You et al., 2025) under a one-shot adaptation setting. Both methods use DINOv2-B as the backbone and are independently adapted using the same 10 sampled image pairs, with each adaptation using only a single image pair. No additional target-dataset pairs are used.

As shown in Table 9, both methods improve over the pretrained DINOv2-B baseline, while our approach consistently outperforms one-shot 3DCorrEnhance across all evaluated tasks and metrics. The gains are observed on both scene- and object-level geometric tasks, as well as video tracking and semantic correspondence. This controlled comparison removes differences in adaptation data scale and shows that our method is more effective under the same one-pair supervision.