PILOT: SELF-SUPERVISED VISUAL REPRESENTATION PRETRAINING FOR TEMPORAL CORRESPONDENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Self-distillation has emerged as a strong paradigm for self-supervised visual representation learning. However, extending this approach to learning video representations with strong temporal correspondence remains challenging: raw frames provide no explicit patch-to-patch alignment over time, while self-supervised pretraining must derive correspondence supervision without external optical flow, point tracks, or correspondence annotations. We address this bottleneck with **PILOT** (**P**rivileged **I**nformation **L**earning via **O**ptimal **T**ransport), a self-supervised representation pretraining framework based on explicit cross-frame target construction. PILOT uses asymmetric distillation: a privileged teacher combines information privilege with algorithmic privilege (Sinkhorn-Knopp optimal transport) to produce globally constrained cross-frame patch assignment predictions that serve as targets for a restricted student. These targets exhibit reduced assignment imbalance and many-to-one collisions, while the student predicts them from masked inputs using row-wise Softmax, encouraging the student encoder to internalize cross-frame correspondence. After pretraining, only a standard encoder is retained. Experiments show that PILOT achieves state-of-the-art overall performance among self-supervised pretrained encoders across three dense correspondence benchmarks and demonstrates strong transfer to action recognition and video retrieval.

1 INTRODUCTION

Self-distillation has emerged as a prominent paradigm for self-supervised visual representation learning. In the image domain, methods such as iBOT (Zhou et al., 2022) and DINOv2 (Oquab et al., 2023) train a student network to match targets produced by a momentum-updated teacher, enabling powerful representations to emerge without human annotations. In this setting, the teacher can generate targets that are anchored to visual content already present in the image, such as unmasked or augmented views.

However, this naturally available anchoring breaks down when extending self-distillation to dense temporal correspondence in videos. Given two video frames, there is no explicit ground-truth correspondence between their patches. Without external optical flow, tracking, or manual annotations, both the teacher and the student must infer temporal correspondence directly from the raw video itself. This makes target construction a central bottleneck: the model must learn from predicted correspondences, yet those correspondences are precisely what the model is expected to discover. This chicken-and-egg problem becomes especially severe under occlusion, where local appearance similarity often becomes unreliable.

To this end, we introduce **P**rivileged **I**nformation **L**earning via **O**ptimal **T**ransport (**PILOT**), a representation pretraining framework for learning stronger temporal correspondence through explicit cross-frame target construction. Inspired by Learning Using Privileged Information (Vapnik & Vashist, 2009) and prediction-as-supervision paradigms (Hsieh et al., 2023; Choudhury & Sodhi, 2025), we build on the idea that predictions can serve as effective learning targets when they are produced by a teacher that is better conditioned than the student. PILOT instantiates this principle by granting the teacher model two distinct advantages for constructing cross-frame patch assignment targets: an *information privilege* and an *algorithmic privilege*. Here, *privilege* refers to information or computation used to construct teacher targets but unavailable to the student encoder.

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.

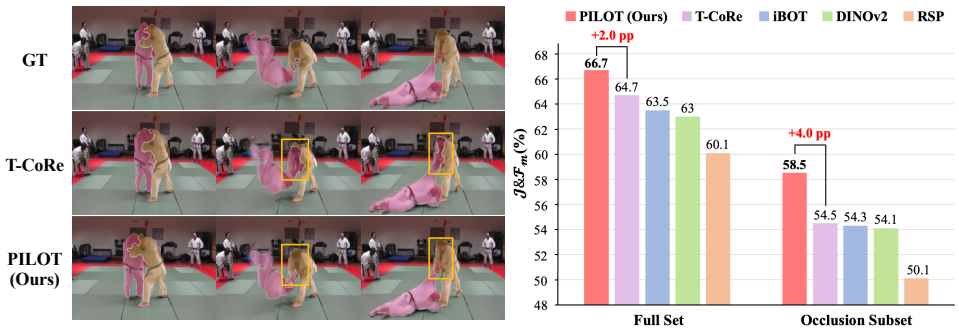


Figure 1: **Dense correspondence under occlusion. Left:** Compared with T-CoRe (Liu et al., 2025) (prior self-supervised SOTA), PILOT maintains more coherent object correspondences when the target undergoes occlusion. **Right:** PILOT outperforms state-of-the-art self-supervised methods on the DAVIS-2017 video object segmentation task (Pont-Tuset et al., 2018), with the larger gain on the occlusion subset indicating stronger robustness to occlusion.

First, the information privilege allows the teacher to operate on unmasked frames and perform cross-frame fusion to construct context-aware correspondence targets (Fig. 2), whereas the student must infer dense correspondence from heavily masked, partial observations.

Second, the algorithmic privilege equips the teacher to solve correspondence as an Optimal Transport (OT) problem. Dense temporal matching is inherently a global assignment task: the reliability of one match depends on competing matches, particularly under occlusion. While the independent row-wise Softmax matching used by the restricted student could assign many uncertain patches to the same visually similar target, the PILOT teacher applies Sinkhorn-Knopp (Cuturi, 2013), which iteratively normalizes *both* the rows and columns of the transport plan, producing an approximately doubly stochastic plan that discourages many-to-one correspondence collapse. As a result, patch assignments are optimized jointly across the full set of source and target patches, yielding globally balanced, routed targets for the student to approximate. Because global routing is used only to construct teacher predictions during pretraining, the student must learn representations with stronger dense temporal correspondence using only row-wise matching. As shown in Fig. 1, PILOT maintains more coherent object-level correspondences under occlusion. Consistently, Table 2 shows that Sinkhorn's global routing reduces reference-patch imbalance and many-to-one collisions, supporting its role in improving robustness when local appearance cues become unreliable.

When applying global assignment to dense ViT features, we further identify a shortcut: positional identity biases Sinkhorn-routed targets toward coordinate-to-coordinate matches. We mitigate this issue by applying latent perturbation to the teacher reference-frame features, weakening exact coordinate cues and encouraging targets to rely more on visual content and temporal context.

Our **contributions** are summarized as follows:

- We identify reliable target construction as a central bottleneck in self-supervised representation pretraining for dense temporal correspondence, where raw videos provide no explicit patch-level alignment, and formulate explicit cross-frame patch assignment as a pretraining target.
- We introduce PILOT, an asymmetric distillation framework where an unmasked teacher uses globally constrained Sinkhorn routing to form assignment targets for a masked row-wise student.
- We show that Sinkhorn routing introduces global competition that reduces reference-patch imbalance and many-to-one collisions, and identify a positional shortcut that biases routed targets toward coordinate-to-coordinate matches, which we mitigate through latent perturbation.
- PILOT achieves state-of-the-art performance among self-supervised pretrained encoders on dense correspondence benchmarks, with particularly strong robustness under occlusion.

2 Related Work

Self-supervised visual representation learning. Self-supervised visual pretraining learns transferable representations from unlabeled images and videos. Contrastive methods such as MoCo (He et al.,

2020; Chen et al., 2021) and SimCLR (Chen et al., 2020) learn invariance by pulling augmented views of the same image together while separating different images. Self-distillation methods, including the DINO family (Caron et al., 2021; Oquab et al., 2023; Siméoni et al., 2025) and iBOT (Zhou et al., 2022), replace explicit negative pairs with teacher–student prediction and learn strong dense features. Masked image modeling methods such as MAE (He et al., 2022) and BEiT (Bao et al., 2022) learn by reconstructing masked image content, and have been extended to videos through masked video modeling (Tong et al., 2022; Wang et al., 2023; Feichtenhofer et al., 2022). While video masked modeling can implicitly capture temporal dynamics through spatiotemporal reconstruction, its learning targets do not explicitly specify which local patches should correspond across frames. This creates a gap between general video self-supervision and dense correspondence-oriented pretraining.

Temporal correspondence learning. Learning temporal correspondence requires associating visual content across frames despite motion, deformation, and occlusion. Supervised methods learn cross-frame matching from optical flow or correspondence annotations (Ilg et al., 2017; Sun et al., 2018; Teed & Deng, 2020), while other correspondence-based approaches build on pretrained representations (Tumanyan et al., 2024; Nam et al., 2025; Liu et al., 2020; Qian et al., 2023). Our focus is instead on self-supervised pretraining from scratch to learn transferable, correspondence-aware encoders. Prior work in this setting exploits cycle consistency, temporal contrast, or cross-frame feature consistency (Wang et al., 2019; Jabri et al., 2020; Li et al., 2019; Xu & Wang, 2021). More recent methods learn through cross-frame reconstruction and prediction: SiamMAE (Gupta et al., 2023) uses masked reconstruction to encourage motion-aware features, RSP (Jang et al., 2024) uses temporal prediction, and T-CoRe (Liu et al., 2025) incorporates neighboring-frame information through cross-attention into masked-token prediction. In contrast to these reconstruction- and prediction-based approaches, PILOT directly supervises cross-frame correspondence learning with explicit patch-to-patch assignment targets, rather than relying solely on reconstruction or prediction to induce correspondence implicitly. PILOT directly distills globally routed patch-to-patch assignments from a privileged teacher into a frame-level encoder, retaining only the encoder for downstream use.

Privileged teacher-student learning. Learning using privileged information studies how additional training information can benefit a predictor restricted to standard inputs (Vapnik & Vashist, 2009). Related reasoning and distillation methods similarly use predictions generated with richer information or structured computation as supervision for a more restricted student (Hsieh et al., 2023; Choudhury & Sodhi, 2025). These works suggest that privileged predictions can serve as training targets, allowing a restricted model to learn from better-conditioned predictions rather than ground-truth supervision. This perspective is particularly relevant to self-supervised temporal correspondence learning, where raw videos provide no explicit correspondence targets and supervision must therefore be constructed internally. Our work follows this line by using privileged correspondence predictions as training targets, with optimal transport (Cuturi, 2013) imposing global assignment constraints during dense correspondence target construction.

3 METHOD

3.1 PROBLEM SETUP AND METHOD OVERVIEW

Our goal is to pretrain a frame-level encoder for dense temporal correspondence without manual annotations, external optical flow, or point tracks. Given a video sequence $V = (v_1, \dots, v_T)$, we sample a current frame v_c together with a past frame v_p and a future frame v_f. We denote each reference frame as $v_r \in \{v_p, v_f\}$ and apply the same correspondence objective to both references. A ViT encoder (Dosovitskiy et al., 2020) maps each frame to patch embeddings $Z \in \mathbb{R}^{N \times D}$, where N is the number of patches and D is the feature dimension (the `[CLS]` token is omitted for dense correspondence). The task is to learn a patch-level correspondence distribution from the current frame v_c to each reference frame v_r. Since raw videos provide no ground-truth correspondence matrix, the central challenge is to construct reliable training targets from the video itself.

PILOT (Fig. 2) addresses this challenge with an asymmetric teacher-student framework. The Exponential-Moving-Average (EMA) teacher f_ξ receives unmasked frames and computes patch embeddings $Z_c^t = f_\xi(v_c)$ and $Z_r^t = f_\xi(v_r)$. For each v_r, the teacher fuses cross-frame context, perturbs the reference features, and applies the Sinkhorn-Knopp algorithm (Cuturi, 2013) to the current-reference similarity matrix, producing a globally routed correspondence target Q_r. The

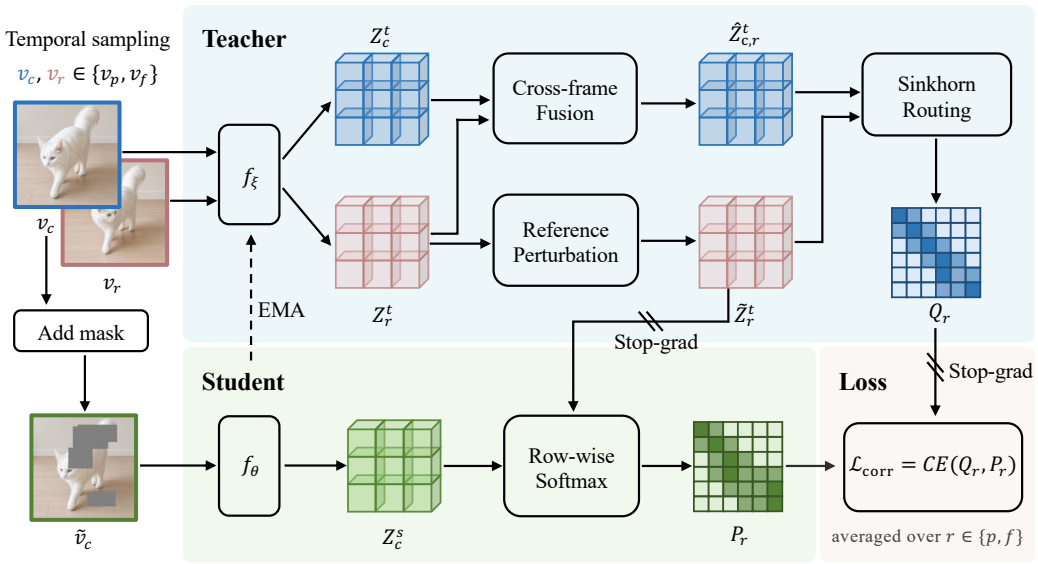


Figure 2: Overview of the proposed temporal correspondence objective. Given a current frame v_c and a temporal reference frame $v_r \in \{v_p, v_f\}$, the EMA teacher f_ξ processes unmasked frames and constructs a patch-level correspondence target used to supervise a masked student. Through cross-frame fusion, anti-shortcut reference perturbation, and Sinkhorn-Knopp routing, the teacher produces a globally constrained target Q_r. The student f_θ observes only the masked current frame $\tilde{v}_c$ and predicts a row-stochastic correspondence matrix P_r using stop-gradient teacher reference features. The correspondence loss distills Q_r into P_r and is averaged over the past and future references.

student f_θ receives the masked current frame $\tilde{v}_c$ and uses the perturbed teacher reference features as the matching candidates. It predicts a row-wise Softmax distribution P_r between current-frame patch embeddings and reference-frame patch embeddings. The correspondence loss distills Q_r into P_r and is averaged over the past and future references. The final objective also includes a Masked Image Modeling (MIM) loss (Zhou et al., 2022) to preserve intra-frame representation learning.

3.2 Privileged Teacher Target Construction

For each temporal reference frame $v_r \in \{v_p, v_f\}$, the teacher constructs a dense correspondence target from unmasked current and reference frames. This target construction uses two forms of privilege. First, the teacher has information privilege because it observes both frames without masking and fuses cross-frame context (Fig. 2). Second, it has algorithmic privilege because it uses Sinkhorn-Knopp routing to construct a globally constrained, approximately doubly stochastic assignment target, whereas the student is restricted to row-wise Softmax prediction.

Cross-frame fusion. The teacher computes patch embeddings $Z_c^t = f_\xi(v_c)$ and $Z_r^t = f_\xi(v_r)$. To condition the current-frame representation on the temporal reference, we use the following parameter-free cross-frame attention operation:

$$\hat{Z}_{c,r}^t = \mathbf{Norm}\left[Z_c^t + \mathbf{Softmax}\left[\frac{Z_c^t (Z_r^t)^\top}{\sqrt{D}}\right] Z_r^t\right], \tag{1}$$

where $\mathbf{Norm}[\cdot]$ and $\mathbf{Softmax}[\cdot]$ take a matrix and perform the $L2$-normalization and Softmax operations independently over the rows, respectively. This parameter-free fusion gives the teacher additional cross-frame context without changing the teacher architecture.

Anti-shortcut reference-frame feature perturbation. Vision Transformers inject spatial information through positional embeddings (Dosovitskiy et al., 2020). In dense temporal matching, this can induce a coordinate shortcut: patches at the same spatial index across frames may share a consistent positional component, biasing correspondence targets toward same-location matches rather than semantic correspondence. As shown in Fig. 3, without perturbation, the teacher target exhibits mostly coordinate-aligned correspondences and a near-diagonal correspondence pattern.

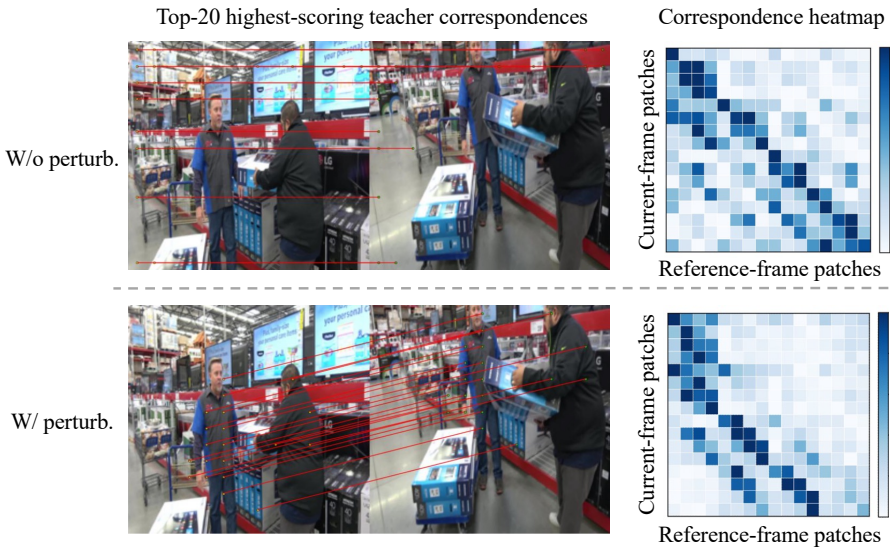


Figure 3: **Mitigating coordinate shortcuts through reference-frame feature perturbation.** Without perturbation, teacher targets exhibit a coordinate-aligned bias toward same-index matches; perturbing reference features reduces this bias, as reflected in both the visualized correspondences and the teacher correspondence heatmaps. For DAVIS-2017 frame pairs separated by 20 frames, perturbation reduces the same-index top-1 assignment rate from 25.7% to 13.2% (-12.5 pp).

This coordinate bias is a shortcut, as it reduces the training loss without capturing the desired correspondence structure (Geirhos et al., 2020). Motivated by noise-based regularization (Bishop, 1995) and denoising representation learning (Vincent et al., 2008), we apply latent perturbation to suppress this shortcut and encourage more robust correspondence targets. Specifically, we perturb only the temporal reference-frame features:

$$\tilde{Z}_r^t = Z_r^t + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \sigma_k^2 I), \tag{2}$$

where $r \in \{p, f\}$ and σ_k follows a cosine decay schedule over training iterations. The perturbed reference-frame features serve as shared reference-side matching candidates for both Sinkhorn-routed teacher target construction and row-wise Softmax student prediction. The perturbation weakens the shortcut-induced identity matching between the current frame and the past or future reference early in training, and gradually decays to allow fine-grained alignment.

Sinkhorn-routed correspondence targets. Given the fused current features and perturbed reference features, the teacher computes current-reference similarities in a shared projected feature space. Specifically, a lightweight MLP head $g(\cdot)$ maps patch embeddings before the temperature-scaled similarity matrix is computed:

$$S_r^t = \frac{g(\hat{Z}_{c,r}^t) g(\tilde{Z}_r^t)^\top}{\tau}, \tag{3}$$

where τ controls the sharpness of the similarity logits before Sinkhorn routing. The teacher then applies Sinkhorn-Knopp iterations (Cuturi, 2013) to obtain a globally routed correspondence target:

$$Q_r = \mathrm{Sinkhorn}(S_r^t). \tag{4}$$

Sinkhorn routing improves teacher target construction by introducing **global competition** over reference patches. For the Sinkhorn target Q_r, the column mass received by reference patch j is

$$m_j = \sum_i Q_r(i, j). \tag{5}$$

Under independent row-wise Softmax, multiple ambiguous current patches can assign high probability to the same reference patch, causing excessive column mass to concentrate on that location. Sinkhorn additionally normalizes each column by its total assignment mass, alternating row-wise and column-wise normalization over multiple iterations to balance column masses across reference patches and create competition among current patches for reference locations. This reference-side competition reduces many-to-one concentration, as illustrated in Fig. 4.

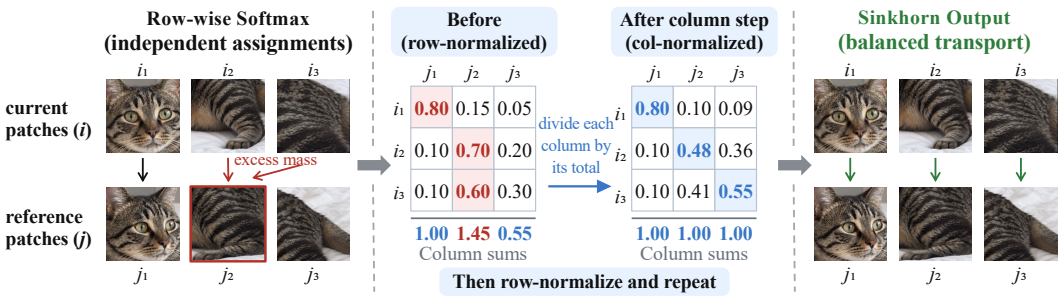


Figure 4: **Global competition in teacher routing.** Row-wise Softmax can concentrate ambiguous matches on one reference patch, whereas Sinkhorn introduces reference-side competition that discourages many-to-one collapse. The illustrative matrices show one column-normalization step within Sinkhorn. A more detailed analysis is provided in Table 2 and Appendix B.

3.3 Restricted Student Prediction and Correspondence Distillation

The student is trained to predict the teacher's globally routed correspondence target from restricted information. For each reference frame $v_r \in \{v_p, v_f\}$, the student receives the masked current frame $\tilde{v}_c$ and produces patch embeddings $Z_c^s = f_\theta(\tilde{v}_c)$.

The teacher–student asymmetry lies in both the available current-frame information and the assignment procedure: the student matches its masked current-frame features to the same perturbed teacher reference features, but converts similarities into correspondences using independent row-wise Softmax rather than globally constrained Sinkhorn routing. Using the same projected similarity computation as in Eq. (3), with Z_c^s replacing $\hat{Z}_{c,r}^t$, the student obtains the logits S_r^s that represent the similarity matrix. In contrast to the teacher, the student predicts the transport plan P_r via row-wise Softmax. That is, $P_r = \mathbf{Softmax}(S_r^s)$. The final correspondence loss $\mathcal{L}_{\text{corr}}$ distills the teacher's Sinkhorn-routed assignments into the student's row-wise predictions. For each current patch i, the inner sum computes the cross-entropy over all reference patches, and the loss then averages these terms over all current patches and both past and future references:

$$\mathcal{L}_{\text{corr}} = -\frac{1}{2N} \sum_{r\in\{p,f\}} \sum_{i=1}^{N} \left[\sum_{j=1}^{N} [Q_r \odot \log P_r]_{ij} \right]. \tag{6}$$

Here, $\odot$ denotes element-wise multiplication and $[\cdot]_{ij}$ indexes the (i, j) entry.

3.4 Overall Training Objective

The final pretraining objective combines the proposed correspondence loss with the standard iBOT-style masked image modeling loss (Zhou et al., 2022):

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{MIM}} + \lambda \mathcal{L}_{\text{corr}}, \tag{7}$$

where λ controls the weight of the correspondence loss. Here, $\mathcal{L}_{\text{MIM}}$ preserves intra-frame representation learning, while $\mathcal{L}_{\text{corr}}$ encourages temporally consistent patch-level correspondence.

4 Experiments

4.1 Experimental Setup

Datasets and evaluation protocols. We evaluate PILOT on three tracking-oriented dense correspondence benchmarks: video object segmentation on DAVIS-2017 (Pont-Tuset et al., 2018), video human-part segmentation on VIP (Zhou et al., 2018), and human pose tracking on JHMDB (Jhuang et al., 2013). Following prior work (Gupta et al., 2023; Eymaël et al., 2024; Jang et al., 2024), we use a semi-supervised label propagation protocol. The pretrained encoder is given the ground-truth annotation of the first frame and propagates it to subsequent frames without fine-tuning. The propagation is performed by patch-level k-nearest-neighbor (k-NN) matching using cosine similarity

Table 1: **Comparison with prior work on dense correspondence benchmarks.** We evaluate video object segmentation on DAVIS-2017 (Pont-Tuset et al., 2018), video human-part segmentation on VIP (Zhou et al., 2018), and human pose tracking on JHMDB (Jhuang et al., 2013) under the semi-supervised label propagation protocol (Sec. 4). Results marked with † are taken from prior work, and ‡ denotes results reproduced using official code. We include a supervised ResNet-50 reference for context. Best self-supervised results are in **bold**.

Method	Backbone	Pretrain data	Epochs	DAVIS-2017			VIP	JHMDB	
				$\mathcal{J}\&\mathcal{F}_m$	$\mathcal{J}_m$	$\mathcal{F}_m$	mIoU	PCK@0.1	PCK@0.2
Supervised† (He et al., 2016)	ResNet-50	ImageNet	100	66.0	63.7	68.4	39.5	59.2	78.3
VideoMAE† (Tong et al., 2022)	ViT-S/16	Kinetics	1600	39.3	39.7	38.9	22.3	41.0	67.9
V-JEPA 2‡ (Assran et al., 2025)	ViT-S/16	Kinetics	400	47.8	46.2	49.4	28.4	33.7	60.1
MAE† (He et al., 2022)	ViT-S/16	Kinetics	400	53.5	50.4	56.7	32.5	43.0	71.3
iBOT† (Zhou et al., 2022)	ViT-S/16	Kinetics	400	63.5	61.7	65.4	37.4	44.5	74.2
DINOv2† (Oquab et al., 2023)	ViT-S/16	Kinetics	400	63.0	62.0	64.1	37.0	46.6	74.8
SiamMAE† (Gupta et al., 2023)	ViT-S/16	Kinetics	400	57.9	56.0	60.0	33.2	46.1	74.0
CropMAE† (Eymaël et al., 2024)	ViT-S/16	Kinetics	400	58.6	55.8	61.4	33.7	42.9	71.1
RSP† (Jang et al., 2024)	ViT-S/16	Kinetics	400	60.1	57.4	62.8	33.8	44.6	73.4
T-CoRe† (Liu et al., 2025)	ViT-S/16	Kinetics	400	64.7	63.5	66.0	37.8	**47.0**	75.2
PILOT (Ours)	ViT-S/16	Kinetics	400	**66.7**	**64.8**	**68.6**	**40.1**	46.7	**75.7**

Table 2: **Global competition in teacher targets.** We compare Softmax and Sinkhorn routing for target construction in terms of reference-patch assignment balance and top-1 collisions, and report DAVIS-2017 performance.

Metric	Softmax	Sinkhorn	Δ
Column dev. ↓	42.2%	**8.8%**	−33.4 pp
Column std ↓	0.6	**0.1**	−0.5
Top-1 collision ↓	40.2%	**24.2%**	−16.0 pp
DAVIS ↑	64.3	**66.7**	+2.4 pp
Occluded DAVIS ↑	54.3	**58.5**	+4.2 pp

Table 3: **Transfer to video-level tasks.** We evaluate action recognition by linear probing and report top-1 accuracy. For video retrieval, we use cosine similarity between video representations and report top-1 recall.

Method	Action Recognition			Video Retrieval	
	UCF101	HMDB51	SSv2	UCF101	HMDB51
RSP	59.3	36.0	17.9	37.6	18.9
iBOT	74.3	**49.4**	26.0	55.3	29.9
DINOv2	75.3	46.9	**26.7**	56.0	27.5
T-CoRe	75.4	44.5	25.1	55.6	24.9
PILOT	**78.3**	49.0	26.1	**62.0**	**30.5**

between pretrained features. As a complementary evaluation of transfer beyond dense correspondence, we evaluate video action recognition via linear probing on UCF101 (Soomro et al., 2012), HMDB51 (Kuehne et al., 2011), and Something-Something V2 (SSv2) (Goyal et al., 2017), as well as video retrieval via cosine similarity on UCF101 and HMDB51. Additional implementation and evaluation details for all tasks are provided in Appendix A.

Pretraining details. To ensure fair comparison with prior self-supervised dense correspondence methods, we use ViT-S/16 (Dosovitskiy et al., 2020) as the default encoder. PILOT is pretrained on Kinetics-400 (Carreira & Zisserman, 2017) for 400 epochs on 4 NVIDIA H100 GPUs, using AdamW with a base learning rate of 5×10^{-4} and a total batch size of 800. Following Liu et al. (2025), the past and future reference frames are sampled at 15%–25% temporal offsets from the current frame. For the correspondence objective, we set $\lambda = 0.6$, use a similarity temperature of $\tau = 0.1$, and perform three Sinkhorn iterations. For latent perturbation, we add Gaussian noise to teacher reference features, with σ cosine-decayed from 0.05 to 0.005.

4.2 Results and Analysis

We compare self-supervised pretrained encoders under a common frozen encoder protocol to isolate representation quality. We include both frame-level self-supervised encoders and clip-based video encoders. A fully supervised ResNet-50 (He et al., 2016) is included as a reference.

Main results. As shown in Table 1, PILOT achieves the strongest overall performance among self-supervised methods with the same ViT-S/16 backbone. Compared with the strongest prior

Table 4: **DAVIS-2017 performance on occlusion and non-occlusion subsets.** Following the official DAVIS-2017 sequence attributes, we split the validation set into occlusion and non-occlusion subsets and report label propagation performance on each subset.

Method	Occlusion Set			Non-occlusion Set		
	$\mathcal{J}\&\mathcal{F}_m$	$\mathcal{J}_m$	$\mathcal{F}_m$	$\mathcal{J}\&\mathcal{F}_m$	$\mathcal{J}_m$	$\mathcal{F}_m$
RSP	50.1	47.1	53.1	62.5	61.6	63.4
iBOT	54.3	51.3	57.3	67.0	65.4	68.6
DINOv2	54.1	51.1	57.2	65.8	64.8	66.9
T-CoRe	54.5	51.4	57.5	67.8	66.8	68.7
PILOT	**58.5** (+4.0)	**54.7** (+3.3)	**62.3** (+4.8)	**69.9** (+2.1)	**68.7** (+1.9)	**71.1** (+2.4)

Table 5: **Robustness under synthetic occlusions.** DAVIS label propagation under varying occlusion ratios.

Method	Occlusion Ratio		
	0.1	0.2	0.3
RSP	52.4	44.7	41.3
iBOT	54.0	47.2	43.0
DINOv2	53.2	46.1	40.8
T-CoRe	55.1	47.4	43.5
PILOT	**55.8**	**48.9**	**44.9**

self-supervised baseline, T-CoRe, PILOT improves $\mathcal{J}\&\mathcal{F}_m$ by $+\mathbf{2.0}$ pp on DAVIS-2017 and mIoU by $+\mathbf{2.3}$ pp on VIP. PILOT also surpasses the supervised ResNet-50 reference on VIP. On JHMDB, PILOT achieves the best self-supervised performance on PCK@0.2 and comparable performance on PCK@0.1. In this protocol, clip-based video encoders are less competitive for dense correspondence, suggesting that clip-level spatiotemporal objectives may not directly preserve patch-level correspondences across frames. In contrast, PILOT explicitly constructs and distills dense temporal alignment targets, leading to stronger transfer across dense prediction tasks without downstream fine-tuning. The qualitative propagation results are provided in Appendix C.1.

Analysis of global competition in teacher target construction. To directly test the global-competition mechanism described in Sec. 3.2, we compute cosine-similarity matrices from fixed DAVIS-2017 teacher features and apply row-wise Softmax or Sinkhorn to the same matrices, changing only the routing operator. Column-marginal deviation and standard deviation quantify imbalance in the assignment mass received by reference patches, while the top-1 collision rate measures how often multiple current patches select the same reference patch. As shown in Table 2, Sinkhorn reduces the column-marginal deviation from 42.2% to 8.8%, the standard deviation from 0.6 to 0.1, and the collision rate from 40.2% to 24.2%. The corresponding Sinkhorn-teacher variant also improves $\mathcal{J}\&\mathcal{F}_m$ by 2.4 pp on DAVIS-2017 and by 4.2 pp on its natural-occlusion subset. These results demonstrate that Sinkhorn routing produces more balanced target usage, fewer many-to-one collisions, and stronger downstream correspondence performance.

Transfer to video-level tasks. PILOT performs best in three of the five action recognition and video retrieval evaluations (Table 3), indicating that the learned representation transfers beyond dense temporal correspondence to broader video-level understanding tasks.

Performance under natural occlusions. To examine whether PILOT's gains are particularly pronounced under occlusion, we group the DAVIS-2017 validation set into occlusion and non-occlusion subsets according to the official attribute for naturally occurring occlusion. As shown in Table 4, PILOT improves over the strongest prior baseline by $+\mathbf{4.0}$ pp in $\mathcal{J}\&\mathcal{F}_m$ on the occlusion subset, compared with $+\mathbf{2.1}$ pp on the non-occlusion subset. The larger gain under occlusion further supports Sinkhorn's global-competition mechanism, where target-side competition becomes more valuable as reliable reference locations become scarce.

Robustness under synthetic occlusions. We further stress-test robustness by applying block masks to target object regions in DAVIS-2017 with occlusion ratios of 0.1, 0.2, and 0.3, where the target object remains partially visible. As shown in Table 5, all methods degrade as occlusion becomes stronger, while PILOT maintains the best performance across all tested ratios. At 30% occlusion, PILOT reaches 44.9 $\mathcal{J}\&\mathcal{F}_m$, outperforming the strongest prior baseline by $+\mathbf{1.4}$ pp.

4.3 Ablation Studies

To understand the contribution of each component, we conduct ablation studies using the ViT-S/16 backbone. We first evaluate the core components across three dense correspondence benchmarks to assess their consistency across tasks. For subsequent design analyses, we report DAVIS-2017 results to highlight performance differences under a controlled setting. Additional analyses beyond the main ablations, including backbone scaling, hyperparameter sensitivity, further design studies, and pretraining tradeoffs, are provided in Appendix C.

Table 6: **Ablation of core components.** We evaluate the contributions of information privilege (unmasked cross-frame fusion) and algorithmic privilege (Sinkhorn routing with the default latent perturbation) under the correspondence objective $\mathcal{L}_{\text{corr}}$, using $\mathcal{L}_{\text{MIM}}$ as the base objective. Results are reported on DAVIS-2017, VIP, and JHMDB. Default PILOT settings are highlighted in blue.

$\mathcal{L}_{\text{MIM}}$	Privileged $\mathcal{L}_{\text{corr}}$		DAVIS-2017			VIP	JHMDB	
	+ Information	+ Algorithmic	$\mathcal{J}\&\mathcal{F}_m$	$\mathcal{J}_m$	$\mathcal{F}_m$	mIoU	PCK@0.1	PCK@0.2
✓			63.5	61.7	65.4	37.4	44.5	74.2
✓	✓		64.1	62.3	65.9	37.9	44.6	74.8
✓		✓	65.4	63.1	67.7	39.5	46.0	75.3
✓	✓	✓	**66.7**	**64.8**	**68.6**	**40.1**	**46.7**	**75.7**

Table 7: **Design analysis of privileged target construction.** We analyze the routing–perturbation interaction, teacher-student routing asymmetry, and information privilege. All results are reported on DAVIS-2017.

(a) Routing and perturbation

Teacher	Perturb.	$\mathcal{J}\&\mathcal{F}_m$	$\mathcal{J}_m$	$\mathcal{F}_m$
Softmax	×	64.1	62.3	65.9
Softmax	✓	64.3	62.6	66.0
Sinkhorn	×	62.8	61.2	64.3
Sinkhorn	✓	**66.7**	**64.8**	**68.6**

(b) Teacher-student routing

Teacher	Student	$\mathcal{J}\&\mathcal{F}_m$	$\mathcal{J}_m$	$\mathcal{F}_m$
Softmax	Softmax	64.3	62.6	66.0
Sinkhorn	Softmax	**66.7**	**64.8**	**68.6**
Sinkhorn	Sinkhorn	65.1	63.0	67.2

(c) Information privilege

Teacher input	Fusion	$\mathcal{J}\&\mathcal{F}_m$	$\mathcal{J}_m$	$\mathcal{F}_m$
Masked	×	65.4	63.1	67.7
Unmasked	×	65.9	63.8	68.0
Unmasked	✓	**66.7**	**64.8**	**68.6**

Core component analysis. Table 6 evaluates the contribution of each component. Starting from the MIM baseline, adding information privilege through temporal fusion consistently improves performance across all benchmarks, while algorithmic privilege yields a larger gain, improving DAVIS-2017 $\mathcal{J}\&\mathcal{F}_m$ from 63.5 to 65.4. Combining both privileges achieves the best results on all metrics, reaching 66.7 $\mathcal{J}\&\mathcal{F}_m$ on DAVIS-2017, 40.1 mIoU on VIP, and 46.7 PCK@0.1 on JHMDB. These results indicate that temporal fusion and global routing provide complementary benefits for constructing dense correspondence targets.

Design analysis. Table 7 analyzes the key design choices in privileged target construction. **(a) Routing and perturbation.** Without perturbation, Sinkhorn routing underperforms the Softmax teacher, suggesting that global constraints can amplify coordinate bias in the similarity scores. Latent perturbation reverses this trend: with perturbation, Sinkhorn achieves the best result, improving from 62.8 to 66.7 and outperforming the perturbed Softmax teacher by $+\mathbf{2.4}$ pp. This interaction, consistent with the shortcut diagnostics in Fig. 3, suggests that mitigating coordinate bias is important for effective global routing. **(b) Teacher-student routing.** The Sinkhorn-teacher and Softmax-student design outperforms both Softmax-Softmax and Sinkhorn-Sinkhorn variants. This supports our asymmetric design, where global routing constructs capacity-constrained teacher targets, while the student remains a restricted row-wise Softmax predictor. **(c) Information privilege.** An unmasked teacher improves over the masked-teacher variant, and cross-frame fusion further improves performance. This indicates that privileged visual evidence and reference-conditioned fusion both contribute to constructing reliable correspondence targets.

5 Conclusion

We addressed the target-construction bottleneck in self-supervised representation pretraining for dense temporal correspondence, where raw videos lack explicit cross-frame patch alignment. To this end, we introduced PILOT, which uses privileged teacher–student distillation to construct explicit cross-frame correspondence targets and learn representations with stronger temporal correspondence. By equipping the teacher with information privilege from unmasked cross-frame fusion and algorithmic privilege from Sinkhorn-Knopp optimal transport, PILOT constructs globally routed assignment targets to guide a restricted masked student. Coupled with latent perturbation to suppress positional shortcuts, PILOT achieves state-of-the-art performance among self-supervised pretrained encoders across dense correspondence benchmarks and transfers strongly to action recognition and video retrieval. Together, these results support using privileged cross-frame predictions as explicit supervision for learning correspondence-aware patch representations. Future work will extend this framework to long-form videos with broader temporal context.

AI USE STATEMENT

Generative AI tools were used to assist with language editing and polishing, literature search and discovery, and code editing and organization. All AI-assisted outputs were reviewed and verified by the authors.

REFERENCES

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-JEPA 2: Self-Supervised Video Models Enable Understanding, Prediction and Planning. *arXiv*, 6 2025.

Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. BEiT: BERT Pre-Training of Image Transformers. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=p-BhZSz59o4.

Chris M Bishop. Training with noise is equivalent to Tikhonov regularization. *Neural Comput.*, 7(1):108–116, 1 1995. ISSN 0899-7667. doi: 10.1162/neco.1995.7.1.108. URL https://doi.org/10.1162/neco.1995.7.1.108.

Mathilde Caron, Hugo Touvron, Ishan Misra, Herve Jegou, 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, 10 2021. ISBN 978-1-6654-2812-5. doi: 10.1109/ICCV48922.2021.00951.

Joao Carreira and Andrew Zisserman. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4724–4733. IEEE, 7 2017. ISBN 978-1-5386-0457-1. doi: 10.1109/CVPR.2017.502.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org, 2020.

Xinlei Chen, Saining Xie, and Kaiming He. An Empirical Study of Training Self-Supervised Vision Transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9620–9629. IEEE, 10 2021. ISBN 978-1-6654-2812-5. doi: 10.1109/ICCV48922.2021.00950.

Sanjiban Choudhury and Paloma Sodhi. Better than Your Teacher: LLM Agents that learn from Privileged AI Feedback. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=st7XqFgbAH.

Marco Cuturi. Sinkhorn Distances: Lightspeed Computation of Optimal Transport. In C J Burges, L Bottou, M Welling, Z Ghahramani, and K Q Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc., 2013. URL https://proceedings.neurips.cc/paper_files/paper/2013/file/af21d0c97db2e27e13572cbf59eb343d-Paper.pdf.

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. *arXiv*, abs/2010.11929, 2020. URL https://api.semanticscholar.org/CorpusID:225039882.

Alexandre Eymaël, Renaud Vandeghen, Anthony Cioppa, Silvio Giancola, Bernard Ghanem, and Marc Van Droogenbroeck. Efficient Image Pre-training with Siamese Cropped Masked Autoencoders. In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part XXIII*, pp. 348–366, Berlin, Heidelberg, 2024. Springer-Verlag. ISBN 978-3-031-73336-9. doi: 10.1007/978-3-031-73337-6_20. URL https://doi.org/10.1007/978-3-031-73337-6_20.

Christoph Feichtenhofer, Haoqi Fan, Yanghao Li, and Kaiming He. Masked Autoencoders As Spatiotemporal Learners. In S Koyejo, S Mohamed, A Agarwal, D Belgrave, K Cho, and A Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 35946–35958. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/e97d1081481a4017df96b51be31001d3-Paper-Conference.pdf.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11): 665–673, 2020. ISSN 2522-5839. doi: 10.1038/s42256-020-00257-z. URL https://doi.org/10.1038/s42256-020-00257-z.

Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. The "something something" video database for learning and evaluating visual common sense. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 5843–5851. IEEE, 10 2017. ISBN 978-1-5386-1032-9. doi: 10.1109/ICCV.2017.622.

Agrim Gupta, Jiajun Wu, Jia Deng, and Li Fei-Fei. Siamese masked autoencoders. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778. IEEE, 6 2016. ISBN 978-1-4673-8851-1. doi: 10.1109/CVPR.2016.90.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum Contrast for Unsupervised Visual Representation Learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9726–9735. IEEE, 6 2020. ISBN 978-1-7281-7168-5. doi: 10.1109/CVPR42600.2020.00975.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollar, 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, 6 2022. ISBN 978-1-6654-6946-3. doi: 10.1109/CVPR52688.2022.01553.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling Step-by-Step! Outperforming Larger Language Models with Less Training Data and Smaller Model Sizes. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki (eds.), *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 8003–8017, Toronto, Canada, 7 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.507. URL https://aclanthology.org/2023.findings-acl.507/.

Eddy Ilg, Nikolaus Mayer, Tonmoy Saikia, Margret Keuper, Alexey Dosovitskiy, and Thomas Brox. FlowNet 2.0: Evolution of Optical Flow Estimation with Deep Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1647–1655. IEEE, 7 2017. ISBN 978-1-5386-0457-1. doi: 10.1109/CVPR.2017.179.

Allan A Jabri, Andrew Owens, and Alexei A Efros. Space-time correspondence as a contrastive random walk. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Huiwon Jang, Dongyoung Kim, Junsu Kim, Jinwoo Shin, Pieter Abbeel, and Younggyo Seo. Visual representation learning with stochastic frame prediction. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Hueihan Jhuang, Juergen Gall, Silvia Zuffi, Cordelia Schmid, and Michael J. Black. Towards Understanding Action Recognition. In *2013 IEEE International Conference on Computer Vision*, pp. 3192–3199. IEEE, 12 2013. ISBN 978-1-4799-2840-8. doi: 10.1109/ICCV.2013.396.

Hilde Kuehne, Hueihan Jhuang, Estibaliz Garrote, Tomaso Poggio, and Thomas Serre. HMDB: A large video database for human motion recognition. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 2556–2563, 2011. doi: 10.1109/ICCV.2011.6126543.

Xueting Li, Sifei Liu, Shalini De Mello, Xiaolong Wang, Jan Kautz, and Ming-Hsuan Yang. Joint-task self-supervised learning for temporal correspondence. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Curran Associates Inc., Red Hook, NY, USA, 2019.

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 (CVPR)*, pp. 4463–4472, 6 2020. doi: 10.1109/CVPR42600.2020.00452.

Yang Liu, Qianqian Xu, Peisong Wen, Siran Dai, and Qingming Huang. When the Future Becomes the Past: Taming Temporal Correspondence for Self-supervised Video Representation Learning. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24033–24044. IEEE, 6 2025. ISBN 979-8-3315-4364-8. doi: 10.1109/CVPR52734.2025.02238.

Jisu Nam, Soowon Son, Dahyun Chung, Jiyoung Kim, Siyoon Jin, Junhwa Hur, and Seungryong Kim. Emergent Temporal Correspondences from Video Diffusion Transformers. In *Advances in Neural Information Processing Systems 38*, 2025. doi: 10.52202/085713-4122. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/b2fea79b1137d917e8b7cce9434ab5fa-Abstract-Conference.html.

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 Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning Robust Visual Features without Supervision. *arXiv*, 4 2023.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: an imperative style, high-performance deep learning library. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Curran Associates Inc., Red Hook, NY, USA, 2019.

Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alex Sorkine-Hornung, and Luc Van Gool. The 2017 DAVIS Challenge on Video Object Segmentation. *arXiv*, 3 2018.

Rui Qian, Shuangrui Ding, Xian Liu, and Dahua Lin. Semantics Meets Temporal Correspondence: Self-supervised Object-centric Learning in Videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16675–16687, 10 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Qian_Semantics_Meets_Temporal_Correspondence_Self-supervised_Object-centric_Learning_in_Videos_ICCV_2023_paper.html.

Paul-Edouard Sarlin, Daniel DeTone, Tomasz Malisiewicz, and Andrew Rabinovich. SuperGlue: Learning Feature Matching With Graph Neural Networks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4937–4946. IEEE, 6 2020. ISBN 978-1-7281-7168-5. doi: 10.1109/CVPR42600.2020.00499.

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*, 8 2025.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

Deqing Sun, Xiaodong Yang, Ming-Yu Liu, and Jan Kautz. PWC-Net: CNNs for Optical Flow Using Pyramid, Warping, and Cost Volume. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8934–8943. IEEE, 6 2018. ISBN 978-1-5386-6420-9. doi: 10.1109/CVPR.2018.00931.

Zachary Teed and Jia Deng. RAFT: Recurrent All-Pairs Field Transforms for Optical Flow. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (eds.), *Computer Vision – ECCV 2020*, pp. 402–419, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58536-5.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA, 2022. Curran Associates Inc. ISBN 9781713871088.

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 (ECCV)*, pp. 367–385, 2024. doi: 10.1007/978-3-031-73347-5_21. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/html/3799_ECCV_2024_paper.php.

Vladimir Vapnik and Akshay Vashist. A new learning paradigm: Learning using privileged information. *Neural Networks*, 22(5-6):544–557, 7 2009. ISSN 08936080. doi: 10.1016/j.neunet.2009.06.042.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, pp. 1096–1103, New York, NY, USA, 2008. Association for Computing Machinery. ISBN 9781605582054. doi: 10.1145/1390156.1390294. URL https://doi.org/10.1145/1390156.1390294.

Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. VideoMAE V2: Scaling Video Masked Autoencoders with Dual Masking. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14549–14560. IEEE, 6 2023. ISBN 979-8-3503-0129-8. doi: 10.1109/CVPR52729.2023.01398.

Xiaolong Wang, Allan Jabri, and Alexei A. Efros. Learning Correspondence From the Cycle-Consistency of Time. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2561–2571. IEEE, 6 2019. ISBN 978-1-7281-3293-8. doi: 10.1109/CVPR.2019.00267.

Qiangqiang Wu, Tianyu Yang, Ziquan Liu, Baoyuan Wu, Ying Shan, and Antoni B. Chan. DropMAE: Masked Autoencoders with Spatial-Attention Dropout for Tracking Tasks. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14561–14571. IEEE, 6 2023. ISBN 979-8-3503-0129-8. doi: 10.1109/CVPR52729.2023.01399.

Jiarui Xu and Xiaolong Wang. Rethinking Self-supervised Correspondence Learning: A Video Frame-level Similarity Perspective. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 10055–10065. IEEE, 10 2021. ISBN 978-1-6654-2812-5. doi: 10.1109/ICCV48922.2021.00992.

Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. iBOT: Image BERT Pre-Training with Online Tokenizer. *International Conference on Learning Representations (ICLR)*, 2022.

Qixian Zhou, Xiaodan Liang, Ke Gong, and Liang Lin. Adaptive Temporal Encoding Network for Video Instance-level Human Parsing. In *Proceedings of the 26th ACM International Conference on Multimedia*, MM '18, pp. 1527–1535, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450356657. doi: 10.1145/3240508.3240660. URL https://doi.org/10.1145/3240508.3240660.

Appendix

A Experimental details

This section provides additional evaluation and implementation details, including dense label propagation, global-competition diagnostics, video-level transfer, occlusion protocols, and pretraining settings.

A.1 Downstream tasks and evaluation metrics

We evaluate dense temporal correspondence using the standard semi-supervised label propagation protocol on three downstream tasks: video object segmentation on DAVIS-2017 (Pont-Tuset et al., 2018), semantic part propagation on VIP (Zhou et al., 2018), and human pose propagation on JHMDB (Jhuang et al., 2013). In all tasks, the pretrained encoder is kept frozen and no downstream fine-tuning is performed. Given the ground-truth annotation of the first frame, labels are propagated through the video sequentially, with each frame receiving labels from annotated and previously propagated reference frames using patch-level nearest-neighbor matching in the pretrained feature space.

Video object segmentation. DAVIS-2017 evaluates whether object masks can be propagated over time. We report Jaccard region similarity $\mathcal{J}_m$, boundary F-measure $\mathcal{F}_m$, and their average $\mathcal{J}\&\mathcal{F}_m$. The Jaccard region similarity measures mask overlap using intersection over union:

$$\mathcal{J}_m = \frac{1}{N}\sum_{i=1}^{N}\frac{|P_i \cap G_i|}{|P_i \cup G_i|}, \tag{8}$$

where P_i and G_i denote the predicted and ground-truth masks for the i-th evaluated object-frame pair. The boundary F-measure $\mathcal{F}_m$ measures contour quality by comparing the predicted and ground-truth object boundaries. For each evaluated object-frame pair, boundary precision Prec_i is the fraction of predicted boundary pixels that match a ground-truth boundary pixel within a tolerance distance, while boundary recall Rec_i is the fraction of ground-truth boundary pixels that are matched by a predicted boundary pixel. The boundary F-measure is then computed as the harmonic mean of boundary precision and recall:

$$\mathcal{F}_m = \frac{1}{N}\sum_{i=1}^{N}\frac{2\,\mathrm{Prec}_i\,\mathrm{Rec}_i}{\mathrm{Prec}_i + \mathrm{Rec}_i}. \tag{9}$$

The final DAVIS-2017 score is computed as

$$\mathcal{J}\&\mathcal{F}_m = \frac{\mathcal{J}_m + \mathcal{F}_m}{2}. \tag{10}$$

Semantic part propagation. VIP evaluates fine-grained human part propagation. We report mean intersection-over-union (mIoU) over semantic part classes:

$$\mathrm{mIoU} = \frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}}\frac{\sum_{i=1}^{N}|P_{i,c}\cap G_{i,c}|}{\sum_{i=1}^{N}|P_{i,c}\cup G_{i,c}|}, \tag{11}$$

where $\mathcal{C}$ denotes the set of semantic part classes, and $P_{i,c}$ and $G_{i,c}$ denote the predicted and ground-truth regions of class c in frame i.

Human pose propagation. JHMDB evaluates whether keypoint annotations can be propagated across frames. We report Percentage of Correct Keypoints (PCK) at thresholds 0.1 and 0.2. A predicted keypoint is counted as correct if its distance to the ground truth is below a fixed fraction of the human body scale:

$$\mathrm{PCK@}k = \frac{1}{|\mathcal{K}|}\sum_{(i,j)\in\mathcal{K}}\mathbf{1}\left[D(\hat{p}_{i,j}, p_{i,j}) < k\cdot d_i\right], \tag{12}$$

where $\mathcal{K}$ denotes the set of evaluated keypoints, d_i denotes the human body scale for frame i, and $D(\cdot,\cdot)$ denotes Euclidean distance.

Table 8: Hyperparameters used in our semi-supervised label propagation protocol, following the T-CoRe evaluation setting.

Dataset	Top-K	Queue Length	Neighborhood Size
DAVIS-2017	7	20	20
VIP	10	20	20
JHMDB	7	20	20

A.2 Label propagation protocol

Following prior dense correspondence evaluation protocols (Gupta et al., 2023; Eymaël et al., 2024; Liu et al., 2025; Jang et al., 2024), we perform label propagation using frozen patch features from the pretrained encoder. For each target frame, we extract patch-level features and match them to features from annotated or previously propagated reference frames using cosine similarity. For segmentation-style tasks, the labels of matched reference patches are transferred to target patches and resized to the original image resolution. For keypoint propagation, each keypoint is transferred according to the matched patch correspondence and evaluated against the ground-truth keypoint location. All evaluation hyperparameters, including the number of neighbors, queue length, and local neighborhood size, are fixed across methods for fair comparison.

A.3 Evaluation hyperparameters

We use the same label-propagation hyperparameters as SiamMAE, CropMAE, RSP, and T-CoRe (Gupta et al., 2023; Eymaël et al., 2024; Liu et al., 2025; Jang et al., 2024). The number of nearest neighbors, queue length, and local neighborhood size are fixed for each benchmark and are not tuned separately for PILOT. The detailed evaluation hyperparameters are summarized in Table 8.

A.4 Global-competition diagnostic protocol

For the routing-only diagnostics in Table 2, we keep the teacher features and cosine-similarity matrices fixed and apply either row-wise Softmax or three-iteration Sinkhorn to each matrix. Let $A \in \mathbb{R}^{N\times N}$ denote a row-normalized assignment matrix and $m_j = \sum_i A_{ij}$ the assignment mass received by reference patch j; because every row sums to one, the mean column mass is one. We define column-marginal deviation as $\frac{100}{N}\sum_j |m_j - 1|$ and column-marginal standard deviation as $\sqrt{\frac{1}{N}\sum_j (m_j-1)^2}$. For $\pi_i = \arg\max_j A_{ij}$, the top-1 collision rate is $100\big(1 - |\{\pi_i\}_{i=1}^N|/N\big)$, which measures the fraction of assignments beyond the unique reference locations selected by the current-frame patches. We compute these quantities for each evaluated DAVIS-2017 teacher similarity matrix and average them across frame pairs. The DAVIS columns in Table 2 instead report the frozen-encoder label-propagation performance of the separately trained Softmax-target and Sinkhorn-target variants.

A.5 Video-level transfer protocol

We evaluate action recognition on UCF101 (Soomro et al., 2012), HMDB51 (Kuehne et al., 2011), and Something-Something V2 (Goyal et al., 2017), and video retrieval on UCF101 and HMDB51. For every method, the ViT-S/16 encoder remains frozen. We uniformly sample eight frames from each video, average the patch-token features within each frame, and concatenate the eight frame representations in temporal order. For action recognition, we train only a linear classifier on these frozen video representations and report top-1 accuracy. The classifier is optimized for 100 epochs with Adam and a learning rate of 0.01. For video retrieval, each test-set video retrieves its nearest neighbor from the training set using cosine similarity between video representations, and we report top-1 recall.

A.6 Occlusion evaluation protocols

For the natural-occlusion analysis, we split the DAVIS-2017 validation sequences according to the official sequence-level occlusion attribute and apply the same frozen label-propagation evaluation to the occlusion and non-occlusion subsets. For the synthetic stress test, we apply block masks to target object regions in the DAVIS-2017 validation set at occlusion ratios of 0.1, 0.2, and 0.3, while keeping the target object partially visible. The label-propagation hyperparameters are unchanged, so only the input occlusion severity varies across the three conditions.

A.7 Pretraining implementation details

We use Kinetics-400 as the pretraining dataset and ViT-S/16 (Dosovitskiy et al., 2020) as the default backbone unless otherwise specified. PILOT is pretrained for 400 epochs in total. We use AdamW with a base learning rate of 5×10^{-4}, a total batch size of 800, and a 20-epoch learning-rate warmup at the beginning of training. All experiments are conducted with PyTorch 2.5.1 (Paszke et al., 2019) on 4 NVIDIA H100 GPUs.

For each video, we sample a current frame together with past and future reference frames following the temporal sampling strategy described in the main paper. Specifically, following Liu et al. (2025), the past and future reference frames are sampled at temporal offsets ranging from 15% to 25% of the video duration relative to the current frame. The same correspondence objective is applied to both past and future references. The student receives a masked current-frame view, while the EMA teacher processes unmasked frames to construct privileged correspondence targets. For the correspondence objective, we set the loss weight to $\lambda = 0.6$, use a similarity temperature of $\tau = 0.1$, and perform three Sinkhorn iterations. For latent perturbation, we add Gaussian noise to the teacher reference-frame features with standard deviation σ cosine-decayed from 0.05 to 0.005 over training. Unless otherwise stated, we use the no-dustbin variant as the default target construction design.

We follow the standard iBOT-style pretraining recipe to apply block-wise masking with mask ratios randomly sampled from $[0.1, 0.5]$ to 50% of the global crops. The masked student views are used for both correspondence prediction and the masked image modeling loss. We use the same masked student features for the correspondence objective, while the teacher constructs targets from unmasked frames. The backbone is first trained with the masked image modeling objective for 350 epochs to obtain stable intra-frame patch representations. We then continue training for 50 epochs with the full PILOT objective. This avoids constructing dense correspondence targets from weak early-stage features, where teacher assignments can be noisy and overly affected by low-level appearance or positional shortcuts. Unless otherwise specified, ViT-S/16 results use the final checkpoint of this 400-epoch pretraining schedule. Additional ablations on the training schedule, correspondence loss weight, latent perturbation magnitude, and the other design choices are provided in Appendix C.

Code and model availability. Code and pretrained model checkpoints will be publicly released upon acceptance.

B Structural analysis of Sinkhorn global competition

We provide an additional analysis of why Sinkhorn routing gives the teacher an algorithmic privilege over independent row-wise matching.

Let $S \in \mathbb{R}^{N \times N}$ denote the teacher similarity matrix between current-frame patches and reference-frame patches. A row-wise Softmax constructs the correspondence distribution independently for each current patch:

$$P_{ij} = \frac{\exp(S_{ij})}{\sum_{k=1}^{N} \exp(S_{ik})}. \tag{13}$$

This normalization guarantees a valid distribution for each current patch,

$$\sum_{j=1}^{N} P_{ij} = 1, \tag{14}$$



Figure 5: Qualitative results on dense correspondence tasks. We show object-mask propagation on DAVIS-2017, semantic-part propagation on VIP, and human-pose propagation on JHMDB. The results illustrate temporally consistent propagation across object-level, part-level, and keypoint-level targets.

but it does not constrain the amount of assignment mass received by each reference patch. For a reference patch j, the column marginal is

$$m_j = \sum_{i=1}^{N} P_{ij}. \tag{15}$$

Since row-wise Softmax imposes no constraint on m_j, the resulting column marginals can be highly imbalanced:

$$m_j \gg 1 \quad \text{for some columns, while} \quad m_{j'} \approx 0 \quad \text{for others.} \tag{16}$$

In dense temporal correspondence, large column marginals correspond to many-to-one assignment collapse: ambiguous or visually similar current patches can be assigned to the same high-scoring reference patch without any global competition. Very small column marginals indicate the complementary failure mode, where some reference patches receive almost no assignment mass. Together, these imbalanced marginals weaken the dense correspondence structure of the teacher target.

Sinkhorn routing addresses this issue by imposing both row and column constraints. This routing can be viewed as constructing an entropy-regularized transport plan over the similarity scores:

$$Q = \arg\max_{Q \in \mathcal{B}} \langle Q, S \rangle + \varepsilon H(Q), \tag{17}$$

where $\varepsilon > 0$ controls the strength of entropy regularization, $H(Q) = -\sum_{i,j} Q_{ij} \log Q_{ij}$ is the entropy term, and $\mathcal{B}$ denotes the set of doubly stochastic matrices:

$$\mathcal{B} = \left\{ Q \in \mathbb{R}_{+}^{N \times N} \;\middle|\; \sum_{j=1}^{N} Q_{ij} = 1, \; \sum_{i=1}^{N} Q_{ij} = 1 \right\}. \tag{18}$$

Table 9: Comparison with prior dense correspondence methods using larger backbones pretrained on Kinetics-400 (Carreira & Zisserman, 2017). All methods are evaluated under the frozen-backbone semi-supervised label propagation protocol. Results marked with † are taken from prior work. Best results are in **bold**.

Method	Backbone	Epochs	DAVIS-2017			VIP	JHMDB	
			$\mathcal{J}\&\mathcal{F}_m$	$\mathcal{J}_m$	$\mathcal{F}_m$	mIoU	PCK@0.1	PCK@0.2
VideoMAE† (Tong et al., 2022)	ViT-L/16	1600	45.0	43.6	46.5	25.6	43.3	70.5
MAE-ST† (Feichtenhofer et al., 2022)	ViT-L/16	1600	54.6	55.5	53.6	33.2	44.4	72.5
DropMAE† (Wu et al., 2023)	ViT-B/16	1600	53.4	51.8	55.0	31.1	42.3	69.2
CropMAE† (Eymaël et al., 2024)	ViT-B/16	400	57.8	56.9	58.7	33.0	45.3	73.3
RSP† (Jang et al., 2024)	ViT-B/16	400	60.5	57.8	63.2	34.0	46.0	74.6
DINOv2† (Oquab et al., 2023)	ViT-B/16	200	64.9	63.3	66.6	38.2	46.8	**75.8**
T-CoRe† (Liu et al., 2025)	ViT-B/16	200	66.4	64.6	68.2	38.9	**47.1**	**75.8**
PILOT (Ours)	ViT-B/16	200	**67.1**	**64.8**	**69.4**	**40.5**	46.8	**75.8**

The row constraint keeps each current patch assigned as a valid distribution over reference patches. The column constraint limits the total assignment mass received by each reference patch:

$$\sum_{i=1}^{N} Q_{ij} = 1. \tag{19}$$

Therefore, no single reference patch can absorb the probability mass of many current patches, and the target avoids highly degenerate column usage. Instead, assignments must be distributed under a shared column capacity, encouraging ambiguous current patches to redistribute mass toward alternative plausible matches rather than collapsing onto the same dominant reference patch.

The column constraint is a training-time regularizer rather than an assumption that real patch correspondences are exactly one-to-one. Occlusion, disocclusion, and repeated appearance can violate one-to-one matching in real videos. Its role is to prevent the evolving teacher from concentrating supervision on a small set of reference locations.

This distinction explains the algorithmic privilege of the teacher. The student predicts correspondences using independent row-wise Softmax, which is efficient and naturally defines one-way current-to-reference predictions. The teacher, however, constructs its target using a globally constrained routing procedure. Thus, the supervision signal encodes not only local pairwise similarity, but also global assignment competition across reference patches. This produces a better-conditioned correspondence target for distillation than independent row-wise normalization.

C Additional Experiments and Analyses

We provide additional experiments and analyses of PILOT, including training schedule and initialization, qualitative dense propagation, backbone scaling, hyperparameter sensitivity, design choices in privileged target construction, and pretraining tradeoffs. Unless otherwise specified, ablation experiments use the ViT-S/16 backbone and report DAVIS-2017 metrics under the same label propagation protocol as the main paper.

C.1 Qualitative Dense Propagation Results

Fig. 5 visualizes propagation results on object masks, semantic parts, and human poses. The predictions show stable temporal transfer across diverse targets, including objects, body parts, and keypoint structures.

C.2 Scaling to a Larger Backbone

To examine whether the proposed privileged correspondence learning remains effective with a larger encoder, we additionally train PILOT with a ViT-B/16 backbone and compare it with prior dense correspondence methods using larger-scale pretraining settings. As shown in Table 9, PILOT with ViT-B/16 achieves the best performance on DAVIS-2017 and VIP, matches the best PCK@0.2 on

Table 10: Epoch-matched training schedule analysis on DAVIS-2017.

Schedule	$\mathcal{J}\&\mathcal{F}_m$	$\mathcal{J}_m$	$\mathcal{F}_m$
400 MIM	63.5	61.7	65.4
350 MIM + 50 MIM	63.3	61.6	64.9
400 PILOT from initialization	65.0	62.9	67.0
200 MIM + 200 PILOT	66.4	64.3	68.6
350 MIM + 50 PILOT	**66.7**	**64.8**	**68.6**

Table 11: Cycle-consistency quality at different initializations. Lower mean return distance is better.

Initialization	Cycle consistency ↑	Mean return distance ↓
Random initialization	34.2%	4.0
200-epoch MIM	64.5%	1.6
350-epoch MIM	65.1%	1.6

Table 12: Run-to-run variation of PILOT. Each run uses a 350-epoch MIM initialization followed by 50 epochs of PILOT pretraining.

	DAVIS-2017			VIP	JHMDB	
Run	$\mathcal{J}\&\mathcal{F}_m$	$\mathcal{J}_m$	$\mathcal{F}_m$	mIoU	PCK@0.1	PCK@0.2
1	66.7	64.8	68.6	40.1	46.7	75.7
2	66.6	64.4	68.7	40.4	46.6	75.7
3	66.2	63.8	68.6	40.3	46.5	75.6
Mean ± std	66.5 ± 0.3	64.3 ± 0.5	68.6 ± 0.1	40.3 ± 0.2	46.6 ± 0.1	75.7 ± 0.1

JHMDB, and remains competitive on PCK@0.1. These results indicate that the gains of PILOT are not limited to the ViT-S/16 setting used in our ablations, but also transfer to a larger-capacity encoder.

C.3 Training schedule and initialization analysis

Table 10 compares epoch-matched 400-epoch schedules. PILOT trained from random initialization improves over 400 epochs of MIM by 1.5 pp, showing that PILOT does not require a 350-epoch MIM initialization. The 350+50 MIM control does not improve over uninterrupted MIM, ruling out stage splitting or an optimization restart as the source of the gain. Switching after either 200 or 350 MIM epochs performs strongly, indicating only weak sensitivity to the switching point.

To characterize the initialization, we measure nearest-neighbor cycle consistency and mean return distance before PILOT training (Table 11). MIM rapidly establishes stable intra-frame patch features: cycle consistency rises from 34.2% at random initialization to 64.5% after 200 MIM epochs and changes only slightly by epoch 350. This explains the default switch point as a practical choice for constructing stable early targets, not a requirement for PILOT to work.

C.4 Run-to-run variation

To assess run-to-run stability, we repeat PILOT pretraining three times. As shown in Table 12, the sample standard deviation is at most 0.5 points across all six metrics, demonstrating consistent performance across independent runs.

C.5 Effect of perturbation scale

We further analyze the perturbation schedule used to mitigate coordinate shortcuts in Sinkhorn-routed teacher target construction. As shown in Table 13, removing perturbation substantially degrades performance, confirming that global routing alone can be sensitive to coordinate-biased similarities. A cosine schedule with maximum perturbation $\sigma = 0.05$ performs best. Both smaller and larger

Table 13: Effect of perturbation scale and schedule in Sinkhorn-routed teacher target construction on DAVIS-2017.

Schedule	Max σ	$\mathcal{J}\&\mathcal{F}_m$	$\mathcal{J}_m$	$\mathcal{F}_m$
None	0.00	62.8	61.2	64.3
Cosine	0.02	65.7	63.8	67.6
Cosine (default)	**0.05**	**66.7**	**64.8**	**68.6**
Constant	0.05	66.2	63.9	68.5
Cosine	0.10	65.3	63.5	67.1

Table 14: Effect of the correspondence loss weight λ on DAVIS-2017.

λ	$\mathcal{J}\&\mathcal{F}_m$	$\mathcal{J}_m$	$\mathcal{F}_m$
0.2	64.9	62.8	67.0
0.4	65.8	63.8	67.8
0.6 (default)	**66.7**	**64.8**	**68.6**
0.8	65.5	63.6	67.4
1.0	65.8	64.0	67.6

Table 15: Effect of the similarity temperature τ on DAVIS-2017.

τ	$\mathcal{J}\&\mathcal{F}_m$	$\mathcal{J}_m$	$\mathcal{F}_m$
0.05	65.7	63.6	67.8
0.10 (default)	**66.7**	**64.8**	**68.6**
0.20	64.8	62.8	66.8

perturbation magnitudes reduce downstream performance relative to the default, suggesting a tradeoff between regularization strength and the preservation of useful matching cues. The cosine schedule also outperforms constant perturbation with the same maximum scale, suggesting that decaying the noise over training helps refine fine-grained correspondences after the shortcut has been weakened.

C.6 Effect of the correspondence loss weight

We vary the correspondence loss weight λ while keeping all other settings fixed. As shown in Table 14, a small λ underuses the privileged correspondence targets, whereas a large λ can dominate the masked image modeling objective and weaken intra-frame representation learning. The default value ($\lambda = 0.6$) balances temporal correspondence learning and general visual representation quality.

C.7 Effect of similarity temperature

The temperature τ controls the sharpness of the teacher and student correspondence distributions (Table 15). Lower temperatures produce sharper assignments but can reduce early training stability, while higher temperatures yield smoother targets with weaker correspondence supervision. The default temperature ($\tau = 0.10$) gives the highest overall performance.

C.8 Effect of Sinkhorn iterations

We ablate the number of Sinkhorn-Knopp iterations used by the teacher (Table 16). More iterations impose stronger global assignment constraints but increase training-time computation. We observe that 3 iterations are sufficient to obtain most of the performance gain, suggesting that approximate global routing is adequate for constructing privileged targets.

Table 16: Effect of the number of Sinkhorn-Knopp iterations on DAVIS-2017.

Iterations	$\mathcal{J}\&\mathcal{F}_m$	$\mathcal{J}_m$	$\mathcal{F}_m$
1	64.2	62.2	66.2
3 (default)	**66.7**	**64.8**	68.6
10	66.6	64.4	**68.8**

Table 17: Effect of dustbin modeling in Sinkhorn-routed teacher target construction on DAVIS-2017. We compare fixed-score and learnable-score one-sided dustbin variants against the no-dustbin default.

Dustbin	Dustbin Score	$\mathcal{J}\&\mathcal{F}_m$	$\mathcal{J}_m$	$\mathcal{F}_m$
Fixed score	0.1	65.8	63.7	67.9
Fixed score	0.6	65.3	63.6	67.0
Learnable score	learned scalar	64.6	62.8	66.4
Without dustbin (default)	–	**66.7**	**64.8**	**68.6**

Table 18: Pretraining cost and performance tradeoff. We compare pretraining time, GPU memory, and downstream dense correspondence performance using the ViT-S/16 backbone under the same hardware setting.

Method	Epochs	Pretraining Time	GPU Mem.	DAVIS-2017 $\mathcal{J}\&\mathcal{F}_m$
iBOT (Baseline)	400	∼23 h	4×62.6 GB	63.5
PILOT (Ours)	400	∼26 h	4×73.5 GB	**66.7**

C.9 Effect of dustbin modeling

We evaluate whether adding a dustbin state improves correspondence target construction under unmatched or ambiguous regions. In dense temporal correspondence, some current-frame patches may not have reliable matches in the reference frame due to temporal changes, newly appearing content, or ambiguous local appearance. A dustbin state is one possible way to handle such cases by adding an explicit unmatched option alongside the reference patches. We implement the dustbin as an additional reference column in the assignment logits, allowing each current-frame patch to assign probability mass either to one of the reference patches or to the dustbin. We consider two one-sided dustbin variants, consistent with our current-to-reference correspondence formulation. The first uses a fixed dustbin score, where the dustbin column is appended with a constant logit before Sinkhorn normalization. The second uses a learnable dustbin score, where the fixed dustbin logit is replaced by a learnable scalar, following prior partial matching formulations (Sarlin et al., 2020).

As shown in Table 17, neither fixed-score nor learnable-score dustbin modeling improves over the standard Sinkhorn-routed target. The learnable dustbin variant also performs worse than both fixed-score variants, suggesting that learning an explicit rejection option does not benefit self-supervised correspondence target construction in our setting. One possible explanation is that explicit dustbin modeling introduces an additional shortcut: uncertain patches can be assigned to the dustbin instead of learning useful visual correspondences from the available context. Although the no-dustbin Sinkhorn target cannot explicitly represent non-correspondence, it can still express uncertainty through soft transport assignments.

In particular, the column constraint limits the total assignment mass that each reference patch can receive. Reference patches with reliable counterparts in the current frame are therefore preferentially claimed by current patches with stronger similarities. Ambiguous or newly appearing current patches may still assign some mass to these reference patches under an independent row-wise assignment, but under Sinkhorn they must compete with more confident matches for the same column capacity.

As a result, they are less likely to collapse onto a dominant reference patch that is already better explained by another current patch, and their mass is instead redistributed over multiple weaker candidates. This allows uncertainty to be represented through distributed soft assignments within

the transport plan rather than absorbed by an explicit dustbin state. We therefore use the no-dustbin variant as the default design.

C.10 Pretraining cost and performance tradeoff

PILOT introduces additional computation during pretraining because the teacher constructs Sinkhorn-routed correspondence targets over dense patch features. We report training time, GPU memory, and downstream DAVIS-2017 performance for the iBOT baseline and PILOT under the same backbone (ViT-S/16). As shown in Table 18, the privileged routing objective increases pretraining time by approximately 13% and GPU memory by approximately 17%, while improving DAVIS-2017 $\mathcal{J}\&\mathcal{F}_m$ by +3.2 pp. To measure the cost of Sinkhorn routing itself, we apply row-wise Softmax and three Sinkhorn iterations to the same input similarity matrix. Compared with Softmax, three Sinkhorn iterations add approximately 0.7 ms to each routing operation and require approximately 7.4 MB of additional temporary memory. These values measure only the routing step, whereas Table 18 reports the total end-to-end pretraining cost. Sinkhorn is used only when the teacher constructs correspondence targets during pretraining. After pretraining, both the teacher and Sinkhorn routing are discarded, leaving only the standard encoder for downstream evaluation.