On the use of Pre-trained Visual Representations in Visuo-Motor Robot Learning

Nikolaos Tsagkas[1] Andreas Sochopoulos[1] Duolikun Danier[1]
Sethu Vijayakumar[1] Chris Xiaoxuan Lu[2†] Oisin Mac Aodha[1†]
[1]University of Edinburgh, [2]UCL, [†]equal senior authorship
https://tsagkas.github.io/pvrobo/

Abstract

The use of pre-trained visual representations (PVRs) in visuo-motor robot learning offers an alternative to training encoders from scratch but we discover that it faces challenges such as temporal entanglement and poor generalisation to minor scene changes. These issues hinder performance in tasks requiring temporal awareness and scene robustness. We address these limitations by: (1) augmenting PVR features with temporal perception and task completion signals to disentangle them over time, and (2) introducing a module that selectively attends to task-relevant local features, improving robustness in out-of-distribution scenes. Our approach, particularly effective for PVRs trained with masking objectives, shows significant performance gains. This work summarises findings from Tsagkas et al. [32].

1. Temporal Entanglement

Problem Statement. Policies using frozen PVR features often violate the Markov assumption, as single-frame observations may lack sufficient information to determine the correct action. As shown in Fig. 1, PVR features from a pick-and-place trajectory exhibit temporal entanglement: (i) frames during static grasps cluster due to minimal pixel changes, and (ii) ascent/descent motions yield near-identical features, differing only slightly in regions affected by the cube's displacement. This ambiguity hampers learning a consistent mapping from observations to actions.

Proposed Solution. To resolve it, we map each timestep to a temporal encoding (TE) and append it to the corresponding observation, using $\gamma(t) = \left(\sin\left(\frac{2^k \pi t}{s^k}\right), \cos\left(\frac{2^k \pi t}{s^k}\right)\right)_{k=0}^{T-1}$ which we concatenate to the policy input. This simple augmentation injects temporal structure, helping disambiguate visually similar states and improve policy learning.

Results. Table 3 shows that feature disentanglement via TE significantly boosts performance, particularly for PVRs with temporal training objectives (*i.e.,* VIP, VFS, R3M). This suggests that potentially there is room for improvement for encoding temporal structure in PVRs.

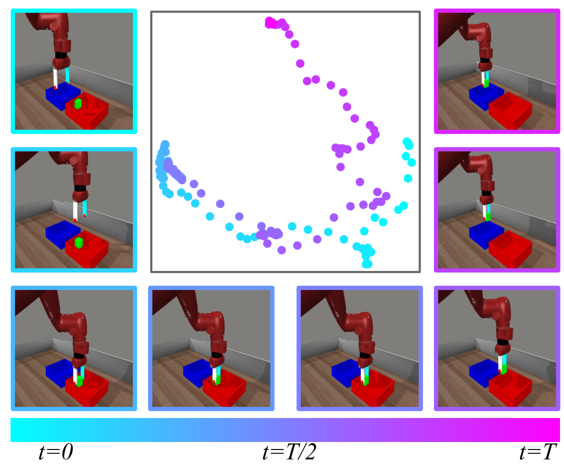

Figure 1. PCA of R3M [18] tokens from an expert demonstration in Bin Picking. Frame colours align with trajectory stages, suggesting feature entanglement during the gripper **descent** and **ascent**, and during the **gripper stop** phase.

Are Video-PVR Better Alternatives? A natural question is whether PVRs trained on video data inherently mitigate this issue. We evaluate three widely used video-PVRs on the same tasks and find that TE continues to improve success rates. Moreover, a negative correlation emerges between performance and the number of input frames (Table 1). This counter-intuitive result aligns with the findings of Chi et al. [9], regarding the observation horizon's length.

	ViT-B/16	**TimeSformer** [3]	**VideoMAE** [29]	**ViViT** [1]
Number of input frames	1	8	16	32
Average inference time	≈ 0.025s	≈ 0.145s	≈ 0.265s	≈ 0.550s
Video-PVR	–	56.9%	45.5%	18.8%
Video-PVR + TE	–	62.4%	44.8%	24.9%

Table 1. Policy success rate across 10 tasks for Video-PVRs.

	Peg Insert	**Bin Picking**	**Disassemble**	**Coffee Pull**	**Average**
CT	42%	80%	54%	96%	68.0%
CT + TE	62%	90%	93%	100%	86.3%

Table 2. Multitask results for Causal Transformer w/ and w/o TE.

	DINOv2 [19]	**DINOv1** [6]	**MAE** [12]	**CLIP** [21]	**ViT** [10]	**iBot** [36]	**VC1** [17]	**MoCov2** [7]	**SWAV** [5]	**VIP** [16]	**DenseCL** [33]	**R3M** [18]	**VFS** [34]	**VICRegL** [2]
$-$	52.7%	51.2%	40.9%	48.6%	48.7%	52.6%	48.6%	65.3%	67.4%	50.6%	60.1%	67.5%	66.3%	65.1%
$\triangledown$	46.9%	57.0%	52.8%	48.9%	49.2%	**55.7%**	**52.9%**	65.3%	66.8%	54.7%	57.6%	71.4%	69.9%	67.7%
$\diamond$	**58.4%**	**61.9%**	**56.0%**	**56.4%**	**58.5%**	54.2%	52.8%	**71.2%**	**68.5%**	**65.1%**	**63.3%**	**75.3%**	**74.4%**	**70.9%**

Table 3. Average success rate across 10 tasks and 5 seeds. Results are reported without any temporal augmentation ($-$), with the FLARE method ($\triangledown$) and with TE of the timestep ($\diamond$). Colour indicates **first**, **second** and **third** best performing PVR with TE.

	DINOv2 [19]	**DINOv1** [6]	**MAE** [12]	**CLIP** [21]	**ViT** [10]	**iBot** [36]	**VC1** [17]	**MoCov2** [7]	**SWAV** [5]	**VIP** [16]	**DenseCL** [33]	**R3M** [18]	**VFS** [34]	**VICRegL** [2]
$\circ$	27.1%	18.6%	15.2%	**22.4%**	**17.8%**	17.6%	13.7%	20.4%	21.0%	10.6%	18.4%	4.6%	8.5%	22.6%
$*$	**41.2%**	**25.3%**	**39.6%**	20.2%	16.7%	**32.4%**	**41.4%**	**27.3%**	**30.5%**	**31.5%**	**28.8%**	**12.8%**	**17.6%**	**31.9%**

Table 4. Average success rate across 10 tasks, 5 seeds. Results are reported in visually perturbed scenes, for PVR+TE ($\circ$) and for PVR+TE+AFA ($*$). Colour indicates **first**, **second**, **third** and **fourth** best performing PVR with AFA.

Does TE Improve SoTA Methods? We also deploy TE along with SoTA approaches that implicitly model temporal structure. We use a Causal Transformer (CT) with context length and action chunking equal to 12. Note that we use rotary embeddings [28] in the CT input which encode the relevant position in the model's input, whereas our TE represent the position in the rollout. We studied a multitask learning scenario to emphasise that TE disentangles adjacent tokens, rather than encode absolute timesteps. Results in Table 2 validate the generality of TE.

2. Robustness Under Visual Perturbations

Problem Statement. Training policies using global features from PVRs (*i.e.,* the CLS token in ViTs or average pooled features in CNNs) can lead to overfitting to visually dominant but task-irrelevant scene attributes (*e.g.,* background textures). This dilutes the policy network's ability to focus on features critical for decision-making. Prior work suggests that only specific image regions contribute meaningfully to task success [8], and recent findings in PVR distillation [25] indicate that local information is especially valuable in robot learning, but this remains underexplored.

Proposed Solution. We introduce Attentive Feature Aggregation (AFA), a data-driven module built upon the attentive probing framework [8]. AFA appends a cross-attention layer with a trainable query token $\boldsymbol{q}$ to the frozen PVR, enabling selective aggregation of local features (*i.e.,* patch tokens in ViTs or channel-wise features in CNNs). The query attends to relevant regions via dot-product attention: $Attention(\boldsymbol{q}, F) = softmax\left(\frac{\boldsymbol{q}\cdot(F\cdot W_K)^\top}{\sqrt{d_k}}\right) F \cdot W_V$, where gradients update the query and projection weights (W_K, W_V), allowing the model to emphasize policy-relevant features while ignoring distractors. Multi-head attention enables focus across diverse feature subspaces.

Results. We train policies with and without AFA and evaluate them in scenes under visual perturbations, where we change either the tabletop texture with vibrant patterns, or change the position, brightness and intensity of the light source. Table 4 summarises these results and indicates that adding a module that learns to attend to task-relevant information increases robustness out-of-domain (OoD). The four top performing PVRs (VC-1 [17], DINOv1 [6], MAE [12] and iBOT [36]) have all been trained with Masked-Image Modelling (MIM), which reflects our original motivation from attentive probing, originally designed for evaluating MIM-trained models fairly.

The average in-domain (ID) performance increases slightly from 63.1% to 66.4% with AFA. This modest gain, compared to AFA's larger improvements in perturbed scenes, suggests it does not learn a new task-specific latent space. Instead, it refines the use of the existing one, by learning to leverage task-relevant information while discarding elements that are irrelevant to the policy. Real-world experiments are available in Tsagkas et al. [32] and project page.

Ablating the Pooling Mechanism. Pooling the feature input stream before the policy network is not novel in robot learning. Usually, however, it serves the role of compressing the input stream's length to increase the action inference speed. In this direction, TokenLearner [23] was used in RT-1 [4] and Spatial SoftMax in [11] and [9]. We compare AFA under visual perturbations against these methods (Table 5) and find that AFA outperforms them by more than 20%.

	Spatial SoftMax	**TokenLearner**	**AFA**
ID	**67.2%**	22.8%	59.2%
OoD	13.1%	19.4%	**41.5%**

Table 5. OoD comparison of Pooling Methods on VC-1.

3. Conclusion

PVRs have proven instrumental in downstream robotics tasks, ranging from encoding spatial-representations [24, 26, 30] to affordance learning [15] and precise zero-shot manipulation [31]. Nevertheless, their deployment in policy learning is still in its infancy [13, 14, 17, 18, 20, 22]. In this work, we discovered two important limitations in the way features from PVRs are utilised in imitation learning and proposed effective solutions to patch them. We conducted experiments both in simulation, in the MetaWorld [35] environment, and the real world. Our methods are agnostic to the policy architecture and can easily be deployed to popular models (*e.g.,* Chi et al. [9] and Sochopoulos et al. [27]).

Acknowledgments. This work was supported by the United Kingdom Research and Innovation (grant EP/S023208/1), EPSRC Centre for Doctoral Training in Robotics and Autonomous Systems (RAS). Funding for DD was provided by the Edinburgh Laboratory for Integrated Artificial Intelligence - EPSRC (EP/W002876/1).

References

[1] Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lučić, and Cordelia Schmid. Vivit: A video vision transformer. In *International Conference on Computer Vision (ICCV)*, 2021. 1

[2] Adrien Bardes, Jean Ponce, and Yann LeCun. Vicregl: Self-supervised learning of local visual features. *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 2

[3] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *International Conference on Machine Learning (ICML)*, 2021. 1

[4] Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Joseph Dabis, Chelsea Finn, Keerthana Gopalakrishnan, Karol Hausman, Alex Herzog, Jasmine Hsu, Julian Ibarz, Brian Ichter, Alex Irpan, Tomas Jackson, Sally Jesmonth, Nikhil J Joshi, Ryan Julian, Dmitry Kalashnikov, Yuheng Kuang, Isabel Leal, Kuang-Huei Lee, Sergey Levine, Yao Lu, Utsav Malla, Deeksha Manjunath, Igor Mordatch, Ofir Nachum, Carolina Parada, Jodilyn Peralta, Emily Perez, Karl Pertsch, Jornell Quiambao, Kanishka Rao, Michael Ryoo, Grecia Salazar, Pannag Sanketi, Kevin Sayed, Jaspiar Singh, Sumedh Sontakke, Austin Stone, Clayton Tan, Huong Tran, Vincent Vanhoucke, Steve Vega, Quan Vuong, Fei Xia, Ted Xiao, Peng Xu, Sichun Xu, Tianhe Yu, and Brianna Zitkovich. Rt-1: Robotics transformer for real-world control at scale. In *Robotics: Science and Systems (RSS)*, 2023. 2

[5] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. 2

[6] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *International Conference on Computer Vision (ICCV)*, 2021. 2

[7] Xinlei Chen, Haoqi Fan, Ross Girshick, and Kaiming He. Improved baselines with momentum contrastive learning. *arXiv preprint arXiv:2003.04297*, 2020. 2

[8] Xiaokang Chen, Mingyu Ding, Xiaodi Wang, Ying Xin, Shentong Mo, Yunhao Wang, Shumin Han, Ping Luo, Gang Zeng, and Jingdong Wang. Context autoencoder for self-supervised representation learning. *International Journal of Computer Vision (IJCV)*, 2024. 2

[9] Cheng Chi, Siyuan Feng, Yilun Du, Zhenjia Xu, Eric Cousineau, Benjamin Burchfiel, and Shuran Song. Diffusion policy: Visuomotor policy learning via action diffusion. In *Robotics: Science and Systems (RSS)*, 2023. 1, 2

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021. 2

[11] Chelsea Finn, Xin Yu Tan, Yan Duan, Trevor Darrell, Sergey Levine, and Pieter Abbeel. Deep spatial autoencoders for visuomotor learning. In *International Conference on Robotics and Automation (ICRA)*, 2016. 2

[12] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Doll'ar, and Ross B. Girshick. Masked autoencoders are scalable vision learners. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 2

[13] Yingdong Hu, Renhao Wang, Li Erran Li, and Yang Gao. For pre-trained vision models in motor control, not all policy learning methods are created equal. In *International Conference on Machine Learning (ICML)*, 2023. 2

[14] Ya Jing, Xuelin Zhu, Xingbin Liu, Qie Sima, Taozheng Yang, Yunhai Feng, and Tao Kong. Exploring visual pre-training for robot manipulation: Datasets, models and methods. In *International Conference on Intelligent Robots and Systems (IROS)*, 2023. 2

[15] Gen Li, Nikolaos Tsagkas, Jifei Song, Ruaridh Mon-Williams, Sethu Vijayakumar, Kun Shao, and Laura Sevilla-Lara. Learning precise affordances from egocentric videos for robotic manipulation. *arxiv preprint arXiv:2408.10123*, 2024. 2

[16] Yecheng Jason Ma, Shagun Sodhani, Dinesh Jayaraman, Osbert Bastani, Vikash Kumar, and Amy Zhang. VIP: Towards universal visual reward and representation via value-implicit pre-training. In *International Conference on Learning Representations (ICLR)*, 2023. 2

[17] Arjun Majumdar, Karmesh Yadav, Sergio Arnaud, Jason Ma, Claire Chen, Sneha Silwal, Aryan Jain, Vincent-Pierre Berges, Tingfan Wu, Jay Vakil, Pieter Abbeel, Jitendra Malik, Dhruv Batra, Yixin Lin, Oleksandr Maksymets, Aravind Rajeswaran, and Franziska Meier. Where are we in the search for an artificial visual cortex for embodied intelligence? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 2

[18] Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3m: A universal visual representation for robot manipulation. In *Conference on Robot Learning (CoRL)*, 2022. 1, 2

[19] Maxime Oquab, Timothée Darcet, Theo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Russell Howes, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research (TMLR)*, 2024. 2

[20] Simone Parisi, Aravind Rajeswaran, Senthil Purushwalkam, and Abhinav Gupta. The (un)surprising effectiveness of pre-trained vision models for control. In *International Conference on Machine Learning (ICML)*, 2022. 2

[21] 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 (ICML)*, 2021. 2

[22] Ilija Radosavovic, Tete Xiao, Stephen James, Pieter Abbeel, Jitendra Malik, and Trevor Darrell. Real-world robot learning with masked visual pre-training. *CoRL*, 2022. 2

[23] Michael S Ryoo, AJ Piergiovanni, Anurag Arnab, Mostafa Dehghani, and Anelia Angelova. Tokenlearner: Adaptive space-time tokenization for videos. In *Advances in Neural Information Processing Systems*, 2021. 2

[24] N. M. M. Shafiullah, C. Paxton, L. Pinto, S. Chintala, and A. Szlam. Clip-fields: Weakly supervised semantic fields for robotic memory. In *Robotics: Science and Systems (RSS)*, 2023. 2

[25] Jinghuan Shang, Karl Schmeckpeper, Brandon B. May, Maria Vittoria Minniti, Tarik Kelestemur, David Watkins, and Laura Herlant. Theia: Distilling diverse vision foundation models for robot learning. In *Conference on Robot Learning (CoRL)*, 2024. 2

[26] William Shen, Ge Yang, Alan Yu, Jansen Wong, Leslie Pack Kaelbling, and Phillip Isola. Distilled feature fields enable few-shot language-guided manipulation. In *Conference on Robot Learning (CoRL)*, 2023. 2

[27] Andreas Sochopoulos, Nikolay Malkin, Nikolaos Tsagkas, João Moura, Michael Gienger, and Sethu Vijayakumar. Fast flow-based visuomotor policies via conditional optimal transport couplings. *arXiv preprint arXiv:2505.01179*, 2025. 2

[28] Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 2024. 2

[29] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. 1

[30] Nikolaos Tsagkas, Oisin Mac Aodha, and Chris Xiaoxuan Lu. Vl-fields: Towards language-grounded neural implicit spatial representations. In *International Conference on Robotics and Automation Workshops (ICRA)*, 2023. 2

[31] Nikolaos Tsagkas, Jack Rome, Subramanian Ramamoorthy, Oisin Mac Aodha, and Chris Xiaoxuan Lu. Click to grasp: Zero-shot precise manipulation via visual diffusion descriptors. In *International Conference on Intelligent Robots and Systems (IROS)*, 2024. 2

[32] Nikolaos Tsagkas, Andreas Sochopoulos, Duolikun Danier, Chris Xiaoxuan Lu, and Oisin Mac Aodha. When pre-trained visual representations fall short: Limitations in visuo-motor robot learning. *arXiv preprint arXiv:2502.03270*, 2025. 1, 2

[33] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[34] Jiarui Xu and Xiaolong Wang. Rethinking self-supervised correspondence learning: A video frame-level similarity perspective. In *International Conference on Computer Vision (ICCV)*, 2021. 2

[35] Tianhe Yu, Deirdre Quillen, Zhanpeng He, Ryan Julian, Karol Hausman, Chelsea Finn, and Sergey Levine. Meta-world: A benchmark and evaluation for multi-task and meta reinforcement learning. In *Conference on Robot Learning (CoRL)*, 2020. 2

[36] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. In *International Conference on Learning Representations (ICLR)*, 2022. 2