The Temporal Trap: Entanglement in Pre-Trained Visual Representations for Visuomotor Policy Learning

Nikolaos Tsagkas[1,α], Andreas Sochopoulos[1] Duolikun Danier[1],
Chris Xiaoxuan Lu[2,ε], Oisin Mac Aodha[1,ε]
[1]University of Edinburgh, [2]UCL

***Abstract*— The integration of pre-trained visual representations (PVRs) has significantly advanced visuomotor policy learning. However, effectively leveraging these models remains a challenge. We identify temporal entanglement as a critical, inherent issue when using these time-invariant models in sequential decision-making tasks. This entanglement arises because PVRs, optimised for static image understanding, struggle to represent the temporal dependencies crucial for visuomotor control. In this work, we quantify the impact of temporal entanglement, demonstrating a strong correlation between a policy's success rate and the ability of its latent space to capture task-progression cues. Based on these insights, we propose a simple, yet effective disentanglement baseline designed to mitigate temporal entanglement. Our empirical results show that traditional methods aimed at enriching features with temporal components are insufficient on their own, highlighting the necessity of explicitly addressing temporal disentanglement for robust visuomotor policy learning.**

I. INTRODUCTION

The integration of pre-trained visual representations (PVRs) into visuomotor robot learning has emerged as a promising alternative to training visual encoders from scratch. Despite the promising results of these models in downstream robotic applications, including affordance-based manipulation [21], semantically precise tasks [41], and language-guided approaches [40], [34], their deployment in policy learning is still nascent.

Overall, prior works have identified multiple issues revolving around the deployment of PVRs in robot learning. First, a consensus exists that we have not identified a single PVR that consistently leads to the best performance. This issue concerns not just the task at hand [28] but also the policy training paradigm [17], with performance varying greatly. Similarly, policy robustness might suffer when the scene undergoes visual perturbations, even though the key motivation behind using PVRs is their generalisation capabilities [46], [5], [42]. We add to this list of issues that hinder policy performance by identifying the problem of temporal entanglement.

In this work, we investigate a paradoxical phenomenon: models celebrated for their robustness on traditional computer vision benchmarks (e.g., classification) tend to underperform in robot learning tasks precisely because of this robustness. We show that the representations learned by such models exhibit temporal entanglement, not only between adjacent timesteps, but also over long-range temporal dependencies. Both entanglements correlate strongly with a model's inability to predict task progression, a metric we find to be a strong predictor of policy success. Building on this insight, we introduce a simple yet powerful baseline for evaluating temporal disentanglement. Using this baseline, we demonstrate that conventional approaches, including feature augmentation, architectural changes to the policy, and alternative pre-training objectives, consistently fall short in addressing the temporal entanglement challenge.

In summary, we make the following contributions:

1. Identifying temporal entanglement. We show that PVRs fall short in encoding temporal cues, even ones trained with temporal objective functions, which has a detrimental effect in visuomotor policy learning. We further demonstrate that these latent spaces lack a reliable notion of task progression.

2. Providing tools to quantify entanglement. We study both short and long-range temporal entanglement and reveal strong correlations with policy success. To this end, we propose a probing method that captures both types of entanglement and serves as a strong predictor of downstream performance.

3. Proposing a disentanglement baseline. We identify task-progression perception as a critical missing factor in current PVR-latent spaces and propose a simple yet effective approach to augment them with such a signal, yielding a significant boost in policy performance.

II. RELATED WORK

In PVR-based visuomotor policy learning, the incorporation of temporal information remains underexplored. Augmentation with temporal perception can happen either at feature level or during training time.

Feature Augmentation. While early fusion methods, such as stacking multiple frames before encoding [19], are common in training visual encoders from scratch, late fusion (*i.e.,* processing frames individually and stacking their representations [35]) has shown superior performance with fewer encoder parameters. Recent work [33] highlights that naive feature concatenation in latent space is insufficient;

*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) and ELIAI (Edinburgh Laboratory for Integrated Artificial Intelligence) - EPSRC (EP/W002876/1).

[ε] Indicates equal senior authorship.
[α] Corresponding author: N. Tsagkas – n.tsagkas@ed.ac.uk

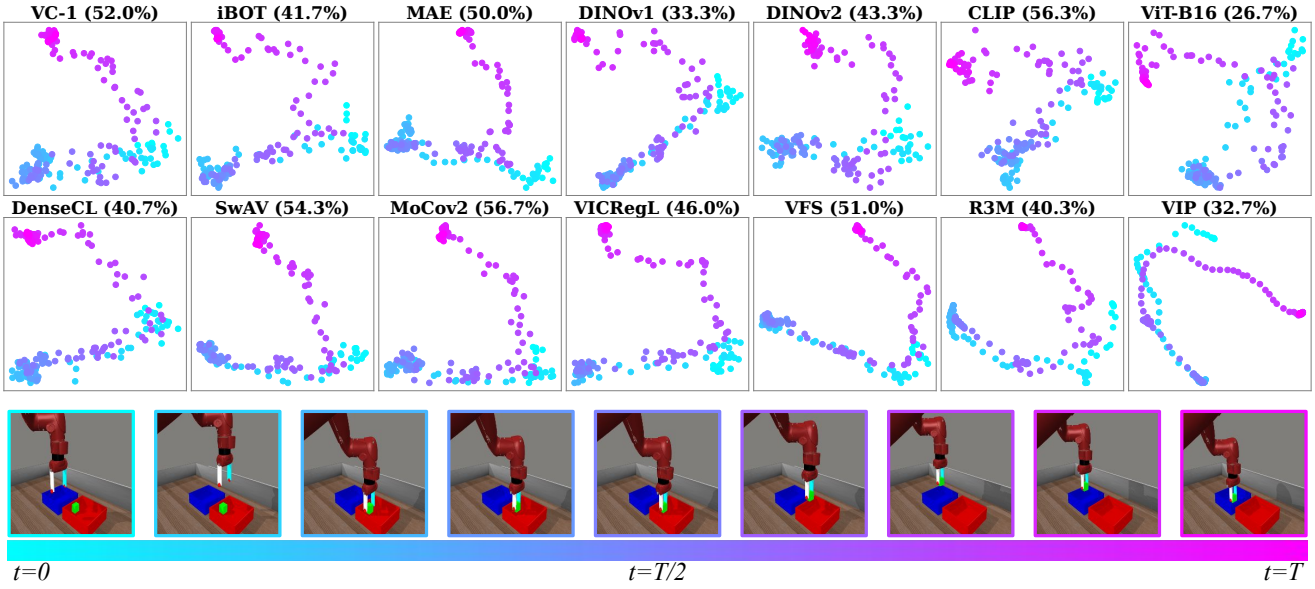

Fig. 1: PCA of features from an expert demonstration in Bin Picking across PVRs (Top row: ViT models; Bottom row: ResNet models). Frame colours align with trajectory stages, suggesting feature entanglement during the gripper descent and ascent, and during the gripper stop phase. Next to each PVR name we provide the success rate of the corresponding policy for the given task.

instead, approaches like FLARE [33] incorporate sequential embeddings and their differences, inspired by optical flow techniques. Nevertheless, concatenating sequential embeddings as input to policy networks has become standard in visuomotor policy learning [28] and SoTA generative policies [9], [36]. However, a gap remains in leveraging PVR features, which are primarily designed for vision tasks, within this temporal framework.

Loss Function Augmentation. A major limitation of many PVRs, in the context of visuomotor policy learning, is their inherit lack of temporal perception, as most are pre-trained on static 2D image datasets. Temporal perception can be added by employing loss functions that enforce temporal consistency during training (*e.g.,* R3M [26] and VIP [22]), when training with video data. However, there is no clear consensus on the superiority of this approach compared to alternatives like masked-image modelling (MIM) (*e.g.,* MVP [45], [17] and VC-1 [23]). This disparity suggests that existing temporal modelling strategies may be insufficient in isolation. In later experiments, we evaluate PVRs trained with temporal information and demonstrate that methods trained with a time-agnostic paradigm achieve comparable performance.

We hypothesise that this limitation arises from a lack of task-progression perception, which we address by incorporating positional encoding, a fundamental mechanism in many machine learning approaches. This straightforward operation has been instrumental in the success of Transformers [43], implicit spatial representations [25], and diffusion processes [16]. Our novelty lies in that we *do not* utilise this tool for encoding the position of an observation in the input stream, but rather in the global temporal stream of the rollout.

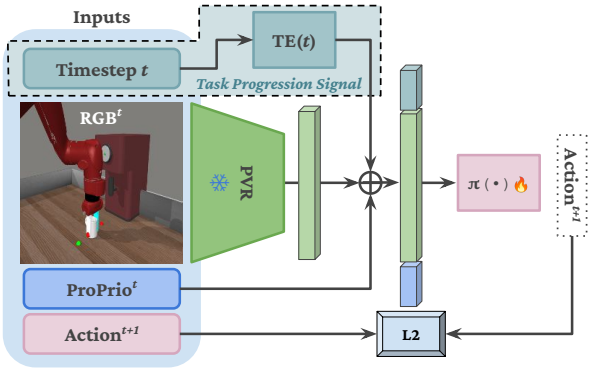

Fig. 2: Standard PVR-based behaviour cloning architecture, modified with out task-progression signal. We propose our disentangling baseline, which incorporates a task-progression signal, effectively disentangling features before the policy head.

III. PRELIMINARIES

A. Imitation Learning via Behaviour Cloning

We consider an expert policy $\pi^\star : \mathcal{P} \times \mathcal{O} \to \mathcal{A}$, which maps a robot's proprioceptive observation $p \in \mathcal{P}$ and visual observation $o \in \mathcal{O}$ to an action $a \in \mathcal{A}$. This policy generates a dataset $\mathcal{T}^{\text{e}} = \{(p_t^i, o_t^i, a_t^i)_{t=0}^{T}\}_{i=1}^{N}$ of N expert trajectories, where each trajectory contains T steps of observations and actions for a task.

We employ behaviour cloning to learn a policy π_θ, parameterised by θ, to imitate $\pi^\star$ by minimizing the action discrepancy over demonstrations:

$$\mathbb{E}_{(p_t^i, o_t^i, a_t^i) \sim \mathcal{T}^{\text{e}}} \| a_t^i - \pi_\theta(f_{\text{PVR}}(o_t^i), p_t^i) \|_2^2, \tag{1}$$

TABLE I: Overview of utilised PVRs. 🤖 indicates models specifically trained for robotic tasks, while ⏱ indicates a temporal training component. Dataset sizes are given in number of images, and † denotes number of frames from videos and * denotes number of videos. Abbreviations: **IN**: ImageNet[31], **LVD**: LVD-142M [27], **K**: Kinetics [20], **E4D**: Ego4D [12], **E4D+MNI** [23].

Arch.	PVR	Training Objective	Dataset (Size)
ViT-B	MAE [14]	MIM	IN (1.2M)
	VC-1 [23] 🤖	MIM	E4D+MNI (5.6M†)
	DINOv1 [7]	Self-Distillation	IN (1.2M)
	iBOT [50]	MIM+Self-Distillation	IN (14M)
	DINOv2 [27]	MIM+Self-Distillation	LVD (142M)
	ViT [10]	Supervised	IN (14M)
	CLIP [30]	V-L Contrastive	LAION (2B)
ResNet-50	MoCov2 [8]	Contrastive	IN (1.2M)
	DenseCL [44]	Local Contrastive	IN (1.2M)
	SwAV [6]	Clustering	IN (1.2M)
	VICRegL [2]	VICReg (global+local)	IN (1.2M)
	VFS [47] ⏱	Self-Distillation (video)	K (240K*)
	VIP [22] 🤖,⏱	Value Function	E4D (5M†)
	R3M [26] 🤖,⏱	Time Contrastive+Language	E4D (5M†)

where f_{PVR} is a pre-trained visual representation (PVR) that extracts features from o_t^i. In visuomotor policy learning, it is common to assume the Markov property, whereby the current observation $x_t = (p_t, o_t)$ suffices for predicting the next state: $P(x_{t+1}|x_t) = P(x_{t+1}|x_t, x_{t-1}, \ldots, x_0)$. This allows tasks to be modelled as Markov decision processes, where each action depends only on the current state, enabling the use of behaviour cloning under this formulation.

As is common practice in similar work [28], [26], [17], we use a shallow (4-layer) MLP with ReLU activations and $\tanh$ before the output to predict the mean μ of a Gaussian with fixed standard deviation σ, modelling $\pi_\theta(a \mid o) = \mathcal{N}^{\mathcal{T}}(\mu, \sigma^2)$, where $\mathcal{N}^{\mathcal{T}}$ denotes a truncated Gaussian with support in $[-1, 1]^k$ (see Fig. 2).

B. Implementation Details

Environment. We conducted our experiments in simulations built on the MuJoCo [38] physics engine. Our core experiments were run on the widely used MetaWorld [48] environment, from which we selected ten tasks, based on their level of difficulty, as identified in prior work on PVR-based visuomotor control [24], [17], as well as from our empirical results. Using the provided heuristic policy, we generated for each task 25 expert demonstrations, with a maximum of 175 steps per rollout. Also, for the purpose of validating the generality of our proposed baseline, we verified its success on three robot arm tasks from the OpenAI Robotics Suite [4], [29].

PVRs. We conducted extensive evaluations across 14 PVRs, including the most popular vision encoders in the field of visuomotor policy learning, that have led to state-of-the-art performance, as summarised in Tab. I. We included PVRs from two main architecture families (Residual Networks (ResNets) [15] and Vision Transformers (ViT) [10]), maintaining a consistent backbone architecture for each group (ResNet-50 or ViT-B/16, with the exception of DINOv2 [27], which employs a smaller patch size of 14). In general, we selected PVRs with diverse characteristics, concerning the objective function, training dataset and balance between local and global perception. We also aimed to include PVRs that have a strong temporal component (*e.g.,* R3M) or simply trained with video data (*i.e.,* VC-1), to see if they deal better with temporal entanglement. Finally, we also briefly study whether VideoPVRs (*i.e.,* PVRs meant for video perception, and not singular image processing) could be a solution to the entanglement problem. For this, we selected three popular, state-of-the-art VideoPVRs: TimeSformer [3], ViViT [1] and VideoMAE [39].

Policy training. We deploy a similar behaviour cloning training scheme to [17], where we train each policy *(PVR, task)* pair 5 times, without updating the PVR weights, and report the interquartile mean (IQM) success rate. We train for 80K steps, using mini-batches of 128 samples and the Adam optimizer and learning rate equal to 10^{-4}.

IV. TEMPORAL ENTANGLEMENT

A. Intuition

We observe that the assumption of Markovian decision-making in policies using features from frozen PVRs is often invalid. This arises because, at each timestep, the available information may be insufficient for the policy to confidently map the current observation to the appropriate action.

Consider the example presented in Fig. 1, where PVR-features of the same pick-and-place trajectory are projected with PCA into 2D. Regardless of the PVR utilised, the extracted features seem to suffer from temporal entanglement. First, features extracted from the frames where the robot has stopped to pick up the box often form a tight cluster, since the only change is the movement of the gripper fingers, which corresponds to a very small percentage of pixels. Second, as the gripper moves down and subsequently ascends, the primary visual change is the cube's vertical displacement relative to the table. Consequently, the visual features extracted from the descent and ascent frames may differ only marginally, and only in dimensions affected by the small pixel region of the cube.

Training a policy network to map (p_t, o_t) to a_t becomes difficult under these conditions. When multiple observations are nearly indistinguishable, the mapping violates the functional requirement that each input must map to exactly one output.

B. Short-range Temporal Entanglement

To quantify the short-range temporal entanglement of features extracted from PVRs, we define the *average sequential cosine similarity* across N_d expert demonstrations and N_T tasks. Each demonstration trajectory consists of N frames, and each frame is mapped to a feature token $\mathbf{f}_n^{(i,t)} \in \mathbb{R}^d$ for the n-th frame of the i-th demonstration of task t. We compute the average feature token for each task:

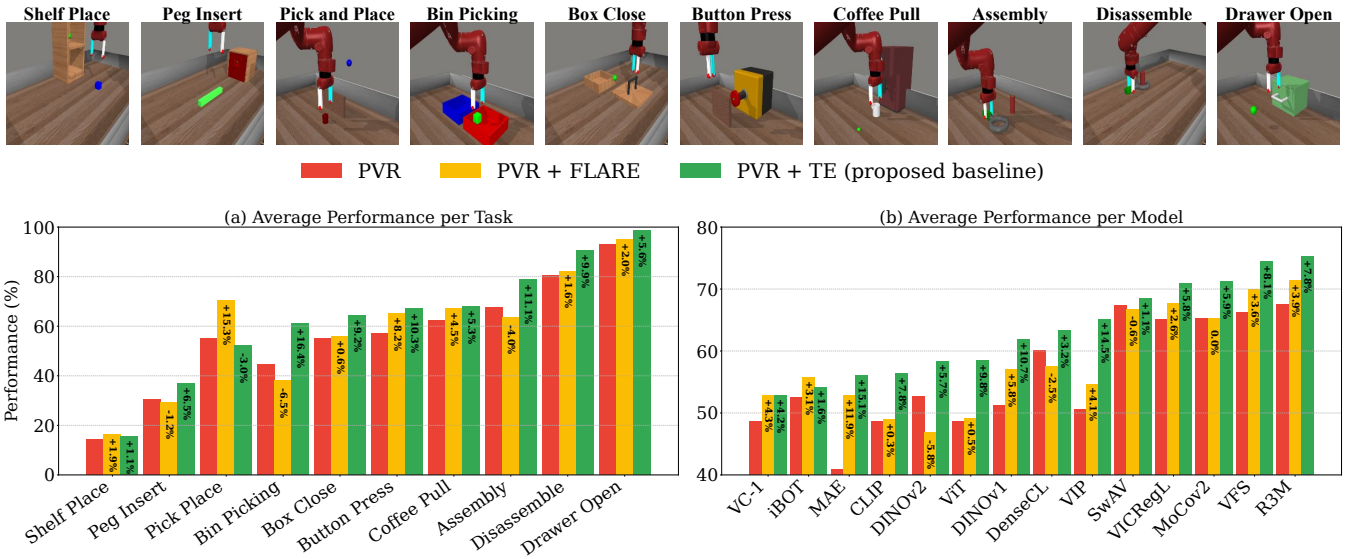

Fig. 3: Comparison of our Temporal Encoding (TE) against FLARE [33] and using no temporal augmentation on PVR features. Results (sorted by TE) show (a) per-task performance and (b) per-model performance. FLARE and TE bars indicate gains over no temporal information.

$$\bar{\mathbf{f}}^{(t)} = \frac{1}{N_d N} \sum_{i=1}^{N_d} \sum_{n=1}^{N} \mathbf{f}_n^{(i,t)}$$

and subtract it to reduce the influence of static visual cues. The *average sequential cosine similarity* is then defined as:

$$\frac{1}{N_d N_T (N-1)} \sum_{t=1}^{N_T} \sum_{i=1}^{N_d} \sum_{n=1}^{N-1} \cos\left(\tilde{\mathbf{f}}_n^{(i,t)}, \tilde{\mathbf{f}}_{n+1}^{(i,t)}\right)$$

where $\tilde{\mathbf{f}}_n^{(i,t)} = \mathbf{f}_n^{(i,t)} - \bar{\mathbf{f}}^{(t)}$ is the task-centred feature.

The left plot of Fig. 4 reveals a clear trend: within particular PVR subgroups, such as ResNets, short-range temporal entanglement is strongly and negatively correlated with the policy success rate. This finding substantiates our claim that PVRs producing more diverse sequential tokens during a rollout tend to achieve higher policy success rates.

C. Long-range Temporal Entanglement

We hypothesised that entanglement concerns not only sequential tokens, due to small disparities in the pixel domain, but also from the robot traversing similar areas during the task (*e.g.,* ascent/descent during pick-and-place). To measure the effect of long-range entanglement, we introduce another metric, the *average global cosine similarity*, where for each PVR we measure the average similarity of each token with all others in a rollout, for all expert demos and tasks.

$$\frac{1}{N_d N_T N} \sum_{t=1}^{N_T} \sum_{i=1}^{N_d} \sum_{\substack{n,m=1 \\ n \neq m}}^{N} \cos\left(\mathbf{f}_n^{(i,t)}, \mathbf{f}_m^{(i,t)}\right)$$

Similarly to the correlation results from section IV-C, a trend emerges where long-range temporal entanglement predicts the policy performance, which is visualised in the right plot of Fig. 4. This evidence supports our hypothesis that entanglement occurs throughout the trajectory and not only within a short-range of the current timestep. This finding is important since most methods, that work either in the latent space (*e.g.,* FLARE [33], concatenating prior tokens [28], etc.) or in the policy side (*e.g.,* LSTMs, Causal Transformers, etc.), aim to disentangle mostly locally, rather than globally.

D. General Temporal Entanglement

Each metric introduced in Sections IV-B, IV-C sheds light to part of the problem. We argue that a latent space overcomes both short and long-range entanglement if it is suitable for accurately perceiving the task-progression during deployment (*i.e.,* how close the policy is to solving the task).

For this, we probe the available PVRs with a shallow MLP, training it to predict the task-progression percentage, assuming that N_T corresponds roughly to the completion of the task. Then, using the test seeds utilized for evaluating the policies, we measure the *task-progression loss*, which is the mean-squared error between the predicted and ground-truth task-progression percentage.

First, we measure for each PVR the correlation between the task-progression loss and the product of the short and long-range similarity scores, to validate that our proposed probing methods is representative of both types of entanglement metrics. Indeed, as visualised in the left plot of Fig. 6, the entanglement scores for the different PVR subgroups are strongly correlated with the task-progression loss. This indicates that we can deploy this tool for measuring the general entanglement (*i.e.,* both local and global).

Second, we measure the correlation between the task-progression loss and the average policy success rate for each PVR. We discover a strong negative correlation between the two metrics, as visualised in the right plot of Fig. 6. We interpret this result as proof that if the PVR latent space is

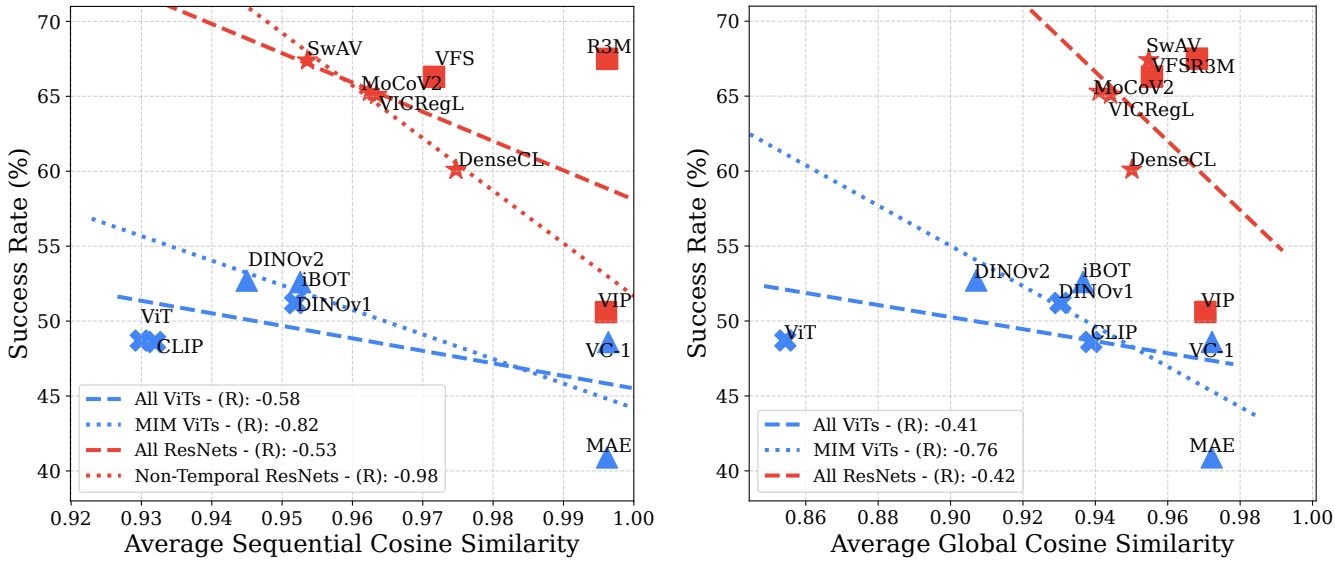

Fig. 4: Correlation plots between per-PVR average policy success rate and temporal entanglement. The left plot concerns short-range temporal entanglement, as described in Section IV-B and the right one concerns long-range temporal entanglement, as described in Section IV-C (we omit here the non-temporal ResNets sub-group, as no trend emerged).

suitable for providing a task-progression signal in the policy network, the success rate is more likely to be high.

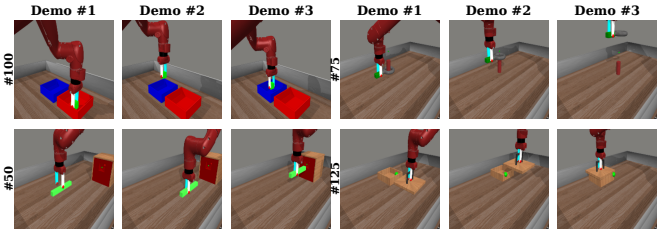

Fig. 5: Temporal variability in MetaWorld expert demonstrations: asynchronous task progression is evident in same-time-step frames from separate demonstrations.

V. TEMPORAL DISENTANGLEMENT

We leverage our findings from Section IV to better understand how we can effectively disentangle PVR-extracted features in visuomotor policy learning. First, in Section V-A we propose a baseline disentangling method, that injects a task-progression signal in the policy model's input. Then, in Sections V-B, V-C and V-D, we revisit traditional temporal disentanglement methods, comparing them against our proposed baseline.

A. Temporal Disentangling Baseline

Our experiments from Section IV-D indicated that the more suitable a PVR's latent space is for task-progression perception, the more likely it is to score high policy success rate. Motivated by this observation, we propose a simple yet very effective approach to augment each observation with a temporal component, by encoding the timestep index n of each frame as a high-dimensional vector, using:

$$\gamma(n) = \Big(\sin\left(\frac{2^0 \pi n}{s^0}\right), \cos\left(\frac{2^0 \pi n}{s^0}\right), \ldots, \sin\left(\frac{2^{N_T-1} \pi n}{s^{N_T-1}}\right), \cos\left(\frac{2^{N_T-1} \pi n}{s^{N_T-1}}\right) \Big) \quad (2)$$

and concatenating to the policy input (see Fig. 2). This augmentation can temporally disentangle similar (p_t, o_t) pairs, introducing a task progression signal into the robot state, which we argue can enhance policy performance. We empirically set the dimensionality of the temporal embeddings to be 64 and the scale parameter 100.

Using alternating sine and cosine functions at exponentially increasing frequencies 2^k, the lower-frequency terms capture coarse temporal trends, while the higher-frequency terms provide finer temporal resolution, enabling the policy to distinguish between temporally similar states in both short and long range. Note that traditionally such embeddings encode the relative position in a transformer's input [37], whereas here we encode the position of the embedding in the rollout.

While our approach currently has limitations that preclude it as a real-world solution (*e.g.,* policy cold-starts may be difficult), we believe it serves as a strong baseline for future disentanglement techniques.

B. Feature augmentation-based disentanglement

We first investigate the approach of temporally augmenting the PVR-features, by concatenating the three most recent past observations and their latent differences (*i.e.,* FLARE [33]). Even though this decreases the PVR short-range temporal entanglement from 0.9686 to 0.8904, the average success rate increases by only roughly 2%, from 56.11% to 58.34% (detailed results in Fig. 3). On the other hand, our baseline (PVR+TE) demonstrates a decrease in short-range temporal

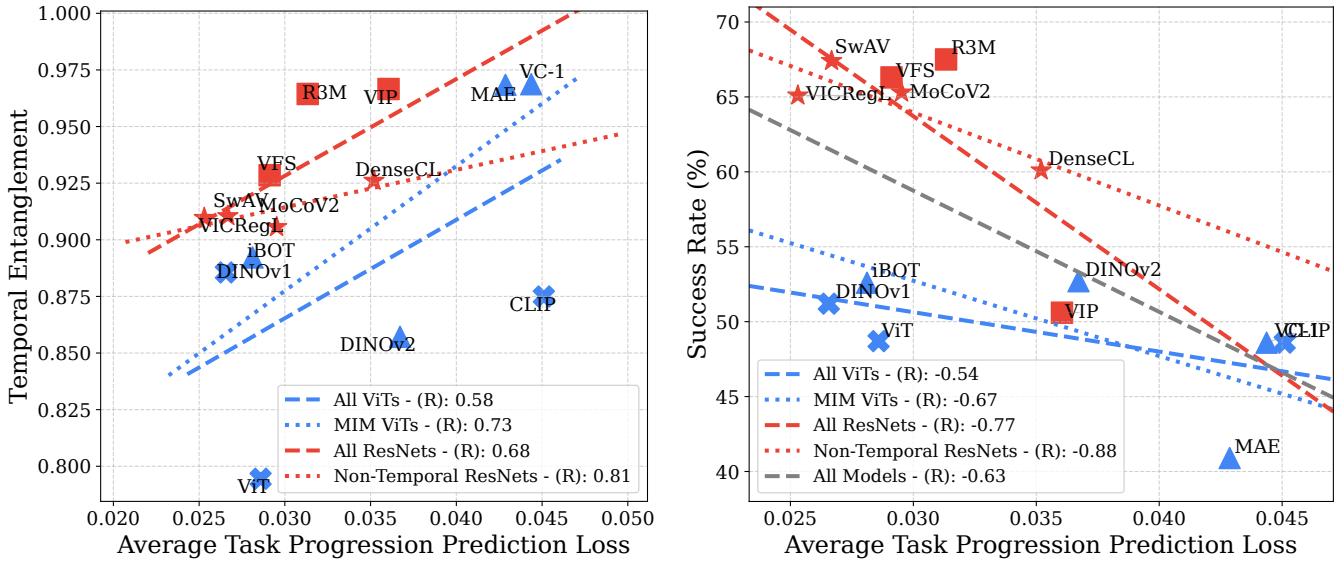

Fig. 6: Correlation plots for the per-PVR task-progression perception. On the left, we provide evidence that good task-progression perception correlates with reduced temporal entanglement. On the right, we showcase that task-progression perception is a good predictor of policy performance (gray line shows trend for all PVRs).

entanglement to only 0.9334, but the increase of the average PVR success rate is more than 7% (63.35%). This can be explained if we observe the long-range similarity of features. FLARE reduces it from 0.9426 to 0.9377, while TE scores 0.1520. Note, that as visualised in Fig. 5, MetaWorld expert demos exhibit temporal variability, and thus the increased performance should not be attributed to overfitting to *(time-index, action)* pairs, but rather to the injected progression signal.

These results further underscore that extracting a temporal signal from only nearby observations might not be enough, hinting that a more intricate solution is required that provides a sense of task-progression. An intriguing observation is that PVRs pre-trained with a temporal component in their objective function also benefit considerably from TE. For instance, R3M employs time-contrastive learning [32] to enforce similarity between representations of temporally adjacent frames-experience substantial gains from TE. In addition, VIP achieves an average performance boost of approximately 15%, making it one of the most positively impacted models. This finding suggests a potential reconsideration of how temporal perception is integrated into features designed for robot learning. It raises the possibility that existing approaches may not fully exploit the temporal structure necessary for optimal performance. Overall, while VC-1 and iBOT achieve slightly higher average scores with FLARE, all other PVRs benefit significantly from temporal augmentation, even when compared to FLARE-augmented results. Similarly, apart from the "Pick and Place" and "Shelf Place" tasks, TE significantly enhances the average task performance. Conducting a statistical analysis we confirmed that our TE baseline's gains over FLARE and no augmentation are significant. This is supported by both the Wilcoxon test ($p < 10^{-30}$) and paired t-test ($p < 10^{-26}$). Additionally, TE also outperforms FLARE, with statistically significant differences observed in both the Wilcoxon test ($p \approx 4.38 \times 10^{-5}$) and paired t-test ($p \approx 1.47 \times 10^{-4}$).

Finally, to make sure that our findings are general and not simulation-specific, we validate that PVR+TE improved the policy performance in tasks from a different simulation environment. We use the `Push`, `Pick and Place` and `Reach` tasks from the OpenAI Robotics Suite (visualised in Fig. 7) and 50 expert demonstrations per task from [13]. We present the average per-task success rates in Fig. 7, where TE more than doubles performance on the `Push` and `Pick and Place` tasks, and still delivers a notable, though smaller, improvement on `Reach`, likely due to the task's lower difficulty.

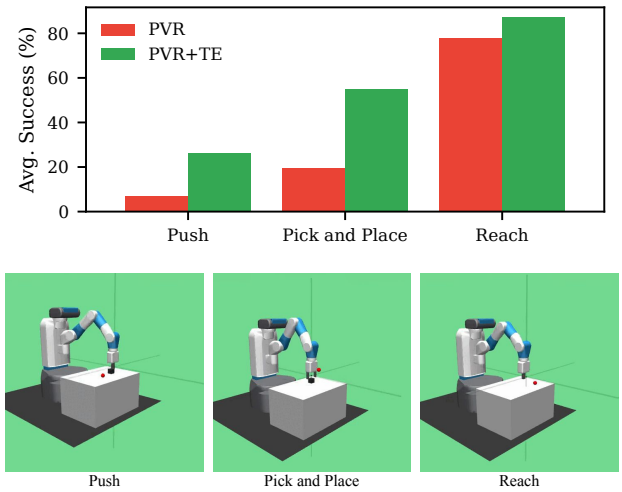

Fig. 7: Validation results of the TE baseline on three additional tasks from the OpenAI Robotics Suite.

TABLE II: Multi-task success rate (Peg Insert, Bin Picking, Disassemble, Coffee Pull) of a Causal Transformer trained with and without TE.

	Task 1	Task 2	Task 3	Task 4	Average
CT	42%	80%	54%	96%	68.0%
CT+TE	62%	90%	93%	100%	86.3%

C. Policy-based disentanglement

We revisit the idea of disentangling features by relying on a policy architecture that captures temporal relationships. We deploy a 1-layer Causal Transformer (CT) [49], [11], [18], which encodes temporal relationships both in the input and in the output, by utilising context and action chunking of length 12, and train it with the features of the strong-performing R3M for 0.8M steps on four MetaWorld tasks. We choose to test our CT in a multi-task setting to further validate TE's ability to temporally disentangle similar features, rather than encode absolute timesteps.

Indeed, Table II reveals that augmenting the input feature space with a task-progression signal benefits greatly the policy in successfully completing a task, leading to an average task success rate boost of +20%.

D. PVR-based disentanglement

Finally, we challenge the starting point of this work, which was to rely on PVRs meant for static image processing, even if some of them were trained with frames extracted from video datasets. We explore the idea that potentially PVRs inherently struggle to deal with temporal entanglement due the nature of their training data and explore the idea of relying on Video-PVRs (*i.e.,* pre-trained models that were meant for video applications, and thus potentially encode implicitly temporal structures). For this purpose, we deploy three powerful video encoders (TimeSformer [3], ViViT [1] and VideoMAE [39]) that have all been trained with the Kinetics-400 dataset [20] and utilise the ViT-B/16 backbone.

In these experiments we preserve our methodology, apart from the way the frame input stream is processed. In the case of video encoders, for both training and evaluating, a frame buffer is created of length N_f, which has the most recent frame, followed by the $N_f - 1$ previous ones. Until the number of available frames become equal to N_f, the buffer is padded, by repeating the oldest frame. We set N_f to match the frame history length of each encoder.

TABLE III: Comparison of inference time and input frames across three video encoders versus the vanilla ViT-B/16, which processes a single frame and is the base for all our ViT-based PVRs except DINOv2 (patch size 14).

	ViT-B/16	TimeSformer	VideoMAE	ViViT
N_f	1	8	16	32
T_p	≈ 0.025s	≈ 0.145s	≈ 0.265s	≈ 0.550s
VPVR	–	56.9%	45.5%	18.8%
VPVR+TE	–	62.4%	44.8%	24.9%

Table III summarises three important aspects of pre-trained video encoders. First, Video-PVRs seem to struggle to outperform even the average PVR performance. Second, we measure the average inference time of each encoder and compare it against the time it takes to process a single frame for the same backbone (*i.e.,* ViT-B/16), which is the one utilised by other PVRs in our experiments. It is not a surprise that the larger N_f is, the slower inference gets [1]. Finally, an interesting find concerns the average success rate itself, which seems to be negatively correlated with the number of frames in the buffer. This counter-intuitive result aligns with the findings of [9], regarding the length of the observation horizon, where the performance would decline as the length increased.

VI. CONCLUSIONS

Our findings suggest that temporal entanglement and a lack of task-progression perception are imperative limitations that accompany the deployment of PVRs in visuomotor policy learning. Our proposed baseline provides evidence that current techniques that aim to enhance the policy input with a temporal component fall short. At the same time, PVRs trained with objective functions and dataset that contain a temporal component also seem to show room for improvement. We believe that our findings can help pave the way towards a temporally aware PVR that will mitigate the investigated issues.

REFERENCES

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

[2] Adrien Bardes, Jean Ponce, and Yann LeCun. Vicregl: Self-supervised learning of local visual features. *NeurIPS*, 2022.

[3] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, 2021.

[4] Greg Brockman, Vicki Cheung, Ludwig Pettersson, Jonas Schneider, John Schulman, Jie Tang, and Wojciech Zaremba. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

[5] Kaylee Burns, Zach Witzel, Jubayer Ibn Hamid, Tianhe Yu, Chelsea Finn, and Karol Hausman. What makes pre-trained visual representations successful for robust manipulation? In *CoRL*, 2024.

[6] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised learning of visual features by contrasting cluster assignments. *NeurIPS*, 2020.

[7] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021.

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

[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 *RSS*, 2023.

[10] Alexey Dosovitskiy, Lucas Beyer, and et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2021.

[11] Letian Fu, Huang Huang, Gaurav Datta, Lawrence Yunliang Chen, William Chung-Ho Panitch, Fangchen Liu, Hui Li, and Ken Goldberg. In-context imitation learning via next-token prediction. In *ICRA*, 2025.

[1]Note that all preprocessing modules and model inference times were measured using code from `huggingface.co/docs/transformers` and tested on a NVIDIA GeForce RTX 4090 GPU with 24GB VRAM, using batches of size 25.

[12] Kristen Grauman, Andrew Westbury, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *CVPR*, pages 18995–19012, 2022.

[13] Siddhant Haldar, Vaibhav Mathur, Denis Yarats, and Lerrel Pinto. Watch and match: Supercharging imitation with regularized optimal transport. *CoRL*, 2022.

[14] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Doll'ar, and Ross B. Girshick. Masked autoencoders are scalable vision learners. *CVPR*, 2022.

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[16] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020.

[17] 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 *ICML*, 2023.

[18] Huang Huang, Fangchen Liu, Letian Fu, Tingfan Wu, Mustafa Mukadam, Jitendra Malik, Ken Goldberg, and Pieter Abbeel. Otter: A vision-language-action model with text-aware feature extraciton. *arXiv*, 2025.

[19] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *CVPR*, 2014.

[20] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, Mustafa Suleyman, and Andrew Zisserman. The kinetics human action video dataset, 2017.

[21] 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.

[22] 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 *ICLR*, 2023.

[23] Arjun Majumdar, Karmesh Yadav, Sergio Arnaud, and et al. Where are we in the search for an artificial visual cortex for embodied intelligence? In *NeurIPS*, 2023.

[24] Atharva Mete, Haotian Xue, Albert Wilcox, Yongxin Chen, and Animesh Garg. Quest: Self-supervised skill abstractions for learning continuous control. In *NeurIPS*, 2024.

[25] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision (ECCV)*, 2020.

[26] Suraj Nair, Aravind Rajeswaran, Vikash Kumar, Chelsea Finn, and Abhinav Gupta. R3m: A universal visual representation for robot manipulation. In *CoRL*, 2022.

[27] 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.

[28] Simone Parisi, Aravind Rajeswaran, Senthil Purushwalkam, and Abhinav Gupta. The (un)surprising effectiveness of pre-trained vision models for control. In *ICML*, 2022.

[29] Matthias Plappert, Marcin Andrychowicz, Alex Ray, Bob McGrew, Bowen Baker, Glenn Powell, Jonas Schneider, Josh Tobin, Maciek Chociej, Peter Welinder, Vikash Kumar, and Wojciech Zaremba. Multi-goal reinforcement learning: Challenging robotics environments and request for research. *arXiv preprint arXiv:1802.09464*, 2018.

[30] 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 *ICML*, 2021.

[31] Tal Ridnik, Emanuel Ben-Baruch, Asaf Noy, and Lihi Zelnik-Manor. Imagenet-21k pretraining for the masses. In *NeurIPS Datasets and Benchmarks*, 2021.

[32] Pierre Sermanet, Kelvin Xu, and Sergey Levine. Unsupervised perceptual rewards for imitation learning. In *RSS*, 2017.

[33] Wenling Shang, Xiaofei Wang, Aravind Srinivas, Aravind Rajeswaran, Yang Gao, Pieter Abbeel, and Misha Laskin. Reinforcement learning with latent flow. In *NeurIPS*, 2021.

[34] William Shen, Ge Yang, Alan Yu, Jansen Wong, Leslie Pack Kaelbling, and Phillip Isola. Distilled feature fields enable few-shot language-guided manipulation. In *CoRL*, 2023.

[35] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. In *NeurIPS*, 2014.

[36] Andreas Sochopoulos, Nikolay Malkin, Nikolaos Tsagkas, João Moura, Michael Gienger, and Sethu Vijayakumar. Fast flow-based visuomotor policies via conditional optimal transport couplings. *CoRL*, 2025.

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

[38] Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: A physics engine for model-based control. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2012.

[39] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In *NeurIPS*, 2022.

[40] 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.

[41] 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.

[42] Nikolaos Tsagkas, Andreas Sochopoulos, Duolikun Danier, Sethu Vijayakumar, Chris Xiaoxuan Lu, and Oisin Mac Aodha. On the use of pre-trained visual representations in visuo-motor robot learning. In *6th Embodied AI Workshop CVPR*, 2025.

[43] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

[44] Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *Conference on CVPR*, 2021.

[45] Tete Xiao, Ilija Radosavovic, Trevor Darrell, and Jitendra Malik. Masked visual pre-training for motor control. *arXiv preprint arXiv:2203.06173*, 2022.

[46] Annie Xie, Lisa Lee, Ted Xiao, and Chelsea Finn. Decomposing the generalization gap in imitation learning for visual robotic manipulation. In *ICRA*, 2024.

[47] Jiarui Xu and Xiaolong Wang. Rethinking self-supervised correspondence learning: A video frame-level similarity perspective. In *ICCV*, 2021.

[48] 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 *CoRL*, 2020.

[49] Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. Dino-wm: World models on pre-trained visual features enable zero-shot planning. *arXiv preprint arXiv:2411.04983*, 2025.

[50] Jinghao Zhou, Chen Wei, Huiyu Wang, Wei Shen, Cihang Xie, Alan Yuille, and Tao Kong. ibot: Image bert pre-training with online tokenizer. In *ICLR*, 2022.