2026 IEEE International Conference on Robotics and Automation (ICRA 2026)
June 1-5, 2026. Vienna, Austria

Order Matters: On Parameter-Efficient Image-to-Video Probing for Recognizing Nearly Symmetric Actions

Thinesh Thiyakesan Ponbagavathi[1,2] and Alina Roitberg[2]

***Abstract*— Fine-grained understanding of human actions is essential for safe and intuitive human–robot interaction. We study the challenge of recognizing *nearly symmetric actions* such as *picking up* vs. *placing down* a tool or *opening* vs. *closing* a drawer. These actions are common in close human-robot collaboration, yet they are rare and largely overlooked in mainstream vision frameworks. Pretrained vision foundation models (VFMs) are often adapted using *probing*, valued in robotics for its efficiency and low data needs, or *parameter-efficient fine-tuning* (PEFT), which adds temporal modeling through adapters or prompts. However, our analysis shows that probing is permutation-invariant and blind to frame order, while PEFT is prone to overfitting on smaller HRI datasets, and less practical in real-world robotics due to compute constraints.**

To address this, we introduce STEP (Self-attentive Temporal Embedding Probing), a lightweight extension to probing that models temporal order via frame-wise positional encodings, a global CLS token, and a simplified attention block. Compared to conventional probing, STEP improves accuracy by 4–10% on nearly symmetric actions and 6–15% overall across action recognition benchmarks in human-robot-interaction, industrial assembly, and driver assistance. Beyond probing, STEP surpasses heavier PEFT methods and even outperforms fully fine-tuned models on all three benchmarks, establishing a new state-of-the-art. Code and models will be made publicly available: https://github.com/th-nesh/STEP.

I. INTRODUCTION

Vision foundation models (VFMs) [1] are reshaping robotic perception [2], [3], [4], offering transferable visual representations that generalize across diverse tasks. In human–robot interaction (HRI), this is crucial: robots must reliably recognize human activities and anticipate intentions from subtle hand-object interactions [5], [6]. For example, in assembly, a robot may need to decide whether to hand over a workpiece or place it aside - decisions that hinge on accurate action recognition.

Unlike large-scale activity benchmarks such as Kinetics [7] or SSv2 [8], HRI datasets [9], [6] are small, domain-specific, and often contain *nearly symmetric actions* (e.g., *putting down* vs. *picking up* or *opening* vs. *closing*). These actions appear visually identical but differ in temporal order, and such confusion can compromise safe and effective robot collaboration. While prior studies explored order (e.g, arrow of time prediction [10], [11] and frame-order verification [12], [13]), they treated it mainly as an auxiliary signal. Consequently, evaluations that rely only on overall accuracy can obscure whether models truly capture temporal order. This limitation is particularly critical in HRI datasets, where

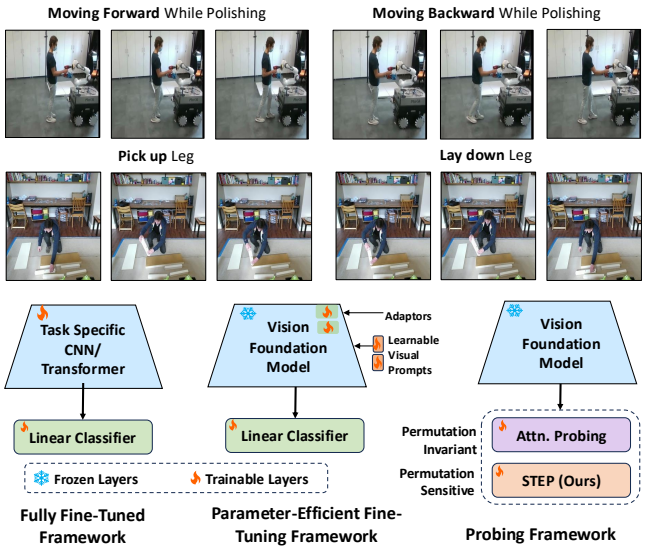

Fig. 1: **STEP: Probing Vision Foundation Models for HRI**. STEP makes probing (i.e., adding small task-specific heads on top of a frozen VFM backbone) sensitive to frame sequence, enabling accurate recognition of nearly symmetric actions (e.g., *pick up* vs. *lay down*) common in HRI scenarios. We focus on parameter-efficient VFM probing (right) and outperform parameter-heavier paradigms such as PEFT (middle) and task-specific models (left).

50–70% of the categories are symmetric and reliable recognition depends on explicitly modeling sequence information.

However, adapting VFMs to capture temporal nuances in HRI is non-trivial. Two strategies dominate: probing and parameter-efficient fine-tuning (PEFT). Probing [14], [15], [16] freezes the backbone and trains a lightweight classifier on top. It is efficient and data-friendly but inherently permutation-invariant, ignoring frame order. PEFT [17], [18] inserts small trainable modules (adapters or prompts) into the frozen backbone to learn temporal dynamics. While more expressive, PEFT is heavier, prone to overfitting on small HRI datasets, and costly in multi-task settings where robots must solve several perception tasks simultaneously. Although CNN-based models [9], [19], [20] remain attractive in HRI for their efficiency, VFMs with probing offer a more scalable alternative by supporting multiple tasks in a single backbone pass. This motivates us to revisit probing vision foundation models for action recognition in HRI, while directly addressing its central weakness: weak temporal modeling.

Motivated by these challenges, we propose **Self-attentive Temporal Embedding Probing (STEP)** – a lightweight extension of self-attention probing that introduces explicit

[1]Institute for Artificial Intelligence, University of Stuttgart, Germany.
[2]Intelligent Assistive Systems Lab, University of Hildesheim, Germany.

temporal modeling at the probing stage through three components: (1) a learnable frame-wise positional encoding reinforcing the video's sequential nature, (2) a global CLS token shared across frames for temporal coherence, and (3) a simplified attention block with no skip connections, layer norm, or FF layers. Unlike PEFT, which modifies the backbone, STEP keeps VFMs frozen and injects temporal order directly into the probing head, balancing efficiency and accuracy. Evaluated on HRI-30 [9] (human-robot collaboration), IKEA-ASM [21] (furniture assembly), and Drive&Act [22] (human-vehicle interaction), STEP achieves state-of-the-art overall performance, surpassing both probing and PEFT baselines, with improvements of 6–15% over probing baselines. Gains are especially pronounced on nearly symmetric actions, but STEP also delivers the best overall accuracy across all datasets, confirming that explicit temporal order modeling is essential.

To summarize our contributions: 1) We explore the concept of *nearly symmetric actions*: actions with visually similar frames but opposite temporal order, and provide dedicated evaluations on three HRI benchmarks. 2) We analyze the limitations of probing and PEFT in HRI scenarios comprising such nearly symmetric actions (often fine-grained object manipulation tasks), showing that probing is permutation-invariant to frame order, while PEFT overfits on smaller HRI datasets. 3) We present STEP – a simple yet effective attention-based probing mechanism, which integrates learnable frame-wise positional encodings, a frame-global CLS token, and a simplified attention block to better model temporal order. 4) Across benchmarks, STEP achieves state-of-the-art accuracy with fewer parameters than probing and PEFT approaches, while also delivering substantial gains on nearly symmetric actions. 5) We further show that STEP supports multi-task HRI in a single backbone pass, reducing computation up to 6× compared to PEFT.

II. Related Work

Action Recognition in HRI. Action recognition is central to HRI, enabling robots to understand human behavior and collaborate effectively. Earlier works relied on CNN-based pipelines, typically two-stream models combining RGB and optical flow [20], later extended with context-aware CNNs [23], human–object parsing [19], or CNN+LSTM backbones [24]. Digital twin [25] frameworks and multimodal fusion [5] have also been explored, but all rely on CNNs as feature extractors. While efficient on embedded hardware, these pipelines are inherently task-specific, and generalize poorly across domains. Vision Foundation Models [26], [27] (VFMs), by contrast, offer strong cross-task generalization but remain less popular in action recognition for HRI. Yet, despite their practical importance in HRI, *nearly symmetric actions* remain largely unexplored in current pipelines. Thus, we explicitly study this regime and investigate how VFMs can be adapted to disambiguate such order-sensitive actions.

Probing and PEFT for Adapting VFMs. There are two popular ways for adapting image VFMs for action recognition: probing and parameter-efficient fine-tuning (PEFT). Linear probes [15], [28] are simple but limited in expressiveness, while attentive probing with cross-attention [14], [16] improves feature aggregation. In video tasks, probing typically averages or concatenates frame embeddings [27], [29], [30], or applies attentive probing [14], [30], [29], but these remain permutation-invariant [16], failing to capture temporal order. Parameter-Efficient Fine-Tuning (PEFT) adapts the backbone more deeply than probing. In the image domain, techniques such as Visual Prompt Tuning (VPT) [31] introduce learnable prompts, while AdaptFormer [32] inserts lightweight adapters into transformer layers to enable efficient adaptation. For video, ST-Adapter [17] applies bottleneck layers for spatiotemporal transfer, Vita-CLIP [18] leverages multimodal prompts, and M2-CLIP [33] enhances temporal alignment with TED-Adapters. These methods achieve strong temporal modeling on large datasets, but risk overfitting in fine-grained HRI scenarios and scale poorly in multi-task robotics. This gap motivates our STEP framework, leveraging probing efficiency but also modeling temporal dynamics for improved action recognition in HRI.

Temporal Modeling for Image-to-Video Transfer. Temporal modeling in foundation models has evolved from frame-wise representations with learnable temporal encoders [34], [35] to spatiotemporal fusion via cross-attention [36]. Later works [37], [38], [39] refine temporal reasoning through auxiliary modules, bidirectional alignment, or temporal masks, while others embed temporal cues implicitly into input tokens via tube embeddings or spatiotemporal patches [40], [41], [42]. TaskAdapter++ [43] explored explicit order modeling, though mainly in text encoders. Beyond architecture design, prior work has also examined temporal order sensitivity itself: arrow-of-time prediction [10], [11], frame-order verification [12], [13], and Retro-Actions [44] showed that reversing or shuffling frames often causes little performance drop on large benchmarks, suggesting most models rely primarily on spatial cues. Overall, existing methods capture order only implicitly and are rarely tested on datasets with many symmetric actions. In contrast, we focus on explicit order modeling at the probing stage, making recognition sequence-sensitive while retaining efficiency.

III. Self-attentive Temporal Embedding Probing for Recognizing Nearly Symmetric Actions

Our goal is to advance temporal understanding in parameter-efficient image-to-video probing for human action recognition specifically for human-robot interaction (HRI) scenarios, where robots must distinguish subtle, order-dependent behaviors. A key challenge lies in *nearly symmetric actions*: visually similar activities that differ only in temporal order (e.g., manipulation actions like *picking* vs. *placing* an object). We first analyze the permutation invariance of self-attention blocks, then introduce STEP, which strengthens temporal reasoning through frame-wise embeddings and a global CLS token. To assess its impact, we curate symmetric action splits across three datasets focusing on HRI and human-vehicle interaction and report results on both symmetric subsets and full benchmarks (Sec. IV-A).

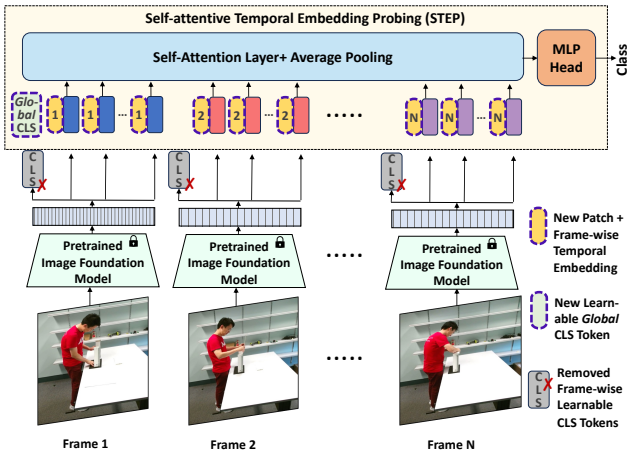

Fig. 2: **Overview of STEP.** Each video frame is independently processed by a frozen image model. We replace the frame-specific CLS token with learned patch-wise temporal encodings, while a newly added *frame-global* CLS token encourages temporal consistency, followed by a self-attention probing mechanism that tracks temporal order.

Self Attention and Permutation Invariance. Self-attention [45] computes pairwise token relations via dot-product attention, producing outputs $z_i = \sum_j \alpha_{ij} v_j$ with weights α_{ij} normalized over all tokens. Because these weights are agnostic to input order, self-attention is inherently permutation-invariant [16]. While effective when appearance dominates, it fails in order-dependent cases, such as *nearly symmetric actions*. Our results confirm this limitation: reversing frames yields almost no accuracy change (Table II).

A. Self-attentive Temporal Embedding Probing

We now introduce **STEP** - **S**elf-attentive **T**emporal **E**mbedding **P**robing (overview in Figure 2), which incorporates several simple but effective modifications to attention-based probing tailored for sensitivity to subtle changes in temporal order, which often occur in HRI tasks. Consider a video sequence $x = \{x_1, x_2, \ldots, x_T\}$ processed frame-by-frame with a pre-trained and frozen image foundation model θ_{frame}, resulting in frame representations, e_i, $\theta_{\text{frame}} : \mathbf{x_i} \rightarrow e_i$. Each frame is spatially split into n patches, forming a set of patch tokens $e_i^{\text{patch}} = \{e_{i,1}, e_{i,2}, \ldots, e_{i,n}\}$ and an additional CLS token e_i^{CLS} that provides a unified representation of the frame.

$$e_i = \left\{ e_i^{\text{patch}}, e_i^{\text{CLS}} \right\} \tag{1}$$

The goal of our STEP probing method, denoted as f, is to link the frame-wise feature representations first into a sequence $e = \{e_1, e_2, \ldots, e_T\}$ then map this sequence to y – the target action label: $f : e \rightarrow y$.

Our method builds on the self-attention probing strategy for image-to-video transfer [41], [16] with two main modifications: (1) a global CLS token that aggregates information across frames instead of the frame-wise CLS tokens, and (2) frame-wise temporal embeddings to encode temporal order. These embeddings are then processed by a self-attention layer, average pooling, and a simplified classification head, as we consistently find that omitting certain common components of probing layers preserves action recognition accuracy while significantly reducing the parameter count.

Learnable Global CLS Token. STEP uses frame patch tokens and a **single learnable frame-global CLS** token, e_{global}^{CLS} to maintain a coherent global representation. This differs from the standard way [48], [27] of using a separate CLS token for each frame. We explicitly discard frame-specific CLS tokens and employ a single learnable global CLS token that attends to all the patch tokens across frames:

$$e = \left\{ e_{global}^{\text{CLS}}, e_1^{\text{patch}}, \ldots, e_T^{\text{patch}} \right\} \tag{2}$$

By attending to all patch tokens during self-attention, the global CLS token captures **global (sequence-level)** temporal dependencies. Meanwhile, **local (frame-level)** details are preserved by integrating frame-wise patch embeddings into the attention mechanism and classification layers, ensuring that STEP maintains fine-grained spatial information while modeling the overall temporal structure. This design enhances temporal consistency by reducing redundancy across frames and consolidating the sequence into a single coherent representation, allowing the model to focus on key temporal transitions critical for action recognition.

Injecting Temporal Embeddings in Self-attention Probing. To improve the temporal sensitivity of the existing probing mechanisms, we augment each frame representation e_i with a learned frame-specific temporal embedding, denoted by t_i. This results in a temporally enhanced frame embedding, $\tilde{e}_i$: $\tilde{e}_i = e_i + t_i$. However, since $\tilde{e}_i$ consists of multiple spatial patch tokens (Eq. 2), temporal embedding t_i is applied *to each patch token* within the frame to ensure temporal information across all regions. Thus, for each patch token e_i^{patch} within frame e_i, the temporal embedding is added as $\tilde{e}_i^{\text{patch}} = e_i^{\text{patch}} + t_i$. Consequently, self-attention probing is now permutation-invariant with respect to frame order, allowing the differentiation of actions with similar appearance but different temporal dynamics.

Feature Aggregation and Classification. The embeddings $\tilde{e}$ are passed through a Multihead Self-Attention (MHSA) layer followed by an average pooling operation: $p = \frac{1}{T}\sum_{t=1}^{T} \text{MHSA}(\tilde{e})$. Unlike prior works [14], [29], [30], [42], we use a pure MHSA block without layer normalization, residual connections, or Feedforward (FF) layers, reducing the parameter count by approximately 3× while maintaining or slightly improving performance. While average pooling itself is permutation-invariant [16], the MHSA embeddings fed into it explicitly encode temporal order through frame-wise positional encodings and interactions with the global CLS token. Thus, the temporal order information is effectively preserved despite pooling. Finally, the pooled representation p is passed to the linear classification layer.

IV. EXPERIMENTS

A. Evaluation setup

We evaluate STEP using image-pretrained CLIP [26] and DINOv2 [27] foundation models with ViT-B backbones

Method	Backbone	GFLOPs	Train Params (M)	Human-Robot Interaction						In-Vehicle Interaction		
				HRI-30			IKEA-ASM			Drive&Act		
				Sym Acc.	N-Sym Acc.	Ovr. Acc.	Sym Acc.	N-Sym Acc.	Ovr. Acc.	Sym Acc.	N-Sym Acc.	Ovr. Acc.
Comparison to published state-of-the-art (fully fine-tuned models)												
VideoSWINv2 [46]	IN-21K	282	88.1	-	-	-	-	-	72.60	-	-	-
SlowOnly [9]	K400	75	32	-	-	86.55	-	-	-	-	-	-
Uniformerv2 [47]	IN-21K	400	115	-	-	-	-	-	-	-	-	76.71
Image-to-Video PEFT Frameworks												
ST-Adapter [17]	CLIP	911	7.1	73.75	95.71	81.07	54.17	77.84	70.15	55.51	88.36	72.61
	DINOv2	1099		81.50	95.30	85.45	63.48	73.94	70.54	66.43	83.43	75.19
M2-CLIP [33]	CLIP	935	14.8	81.75	95.07	85.85	64.11	76.62	71.86	67.40	**85.70**	77.20
	DINOv2	1128		78.39	84.64	80.47	68.49	77.95	74.88	68.40	84.98	77.73
VitaCLIP [18]	CLIP	937	28.6	60.53	91.78	70.95	50.11	69.57	63.25	65.45	81.85	73.98
	DINOv2	1130		57.87	53.21	56.31	66.03	74.97	72.48	65.45	79.78	72.91
Probing Frameworks (main baseline)												
Linear Probing [28]	CLIP	269.6	**0.3**	19.46	35.35	24.76	25.29	36.39	32.70	28.77	66.27	48.28
	DINOv2	351.5		23.75	52.50	33.33	32.45	46.15	41.71	40.53	57.79	49.51
Attn Probing [16]	CLIP	273.6	7.3	30.71	66.07	42.50	60.77	64.43	61.94	49.41	78.69	64.64
	DINOv2	356.3		57.50	72.85	62.61	64.73	70.64	65.76	55.41	80.16	68.65
Self-Attn Probing [41]	CLIP	308.1	2.6	45.35	77.85	56.19	59.35	65.67	61.39	51.22	79.48	65.93
	DINOv2	413.5		74.28	80.00	76.19	56.30	68.01	60.78	59.57	80.85	71.16
STEP (Ours)	CLIP	307.8	2.6	56.25	87.50	66.66	64.65	70.44	64.83	57.22	80.37	69.27
	DINOv2	413.1		**82.14**	**96.78**	**87.02**	**69.46**	**80.59**	**76.80**	**69.98**	84.02	**78.40**

TABLE I: Comparison of symmetric and non-symmetric action recognition accuracy, model efficiency, and trainable parameters. STEP consistently outperforms fully fine-tuned, PEFT, and probing baselines.

across three diverse HRI-focused action recognition datasets. HRI-30 is a recent benchmark specifically designed for human-robot interaction [9], while IKEA-ASM [21] and Drive&Act [22] focus on fine-grained hand-object assembly and in-car driver behavior, respectively. Both assembly and driving represent critical HRI domains where robots must interpret subtle, order-dependent human actions to provide safe and effective assistance. We adhere to standard evaluation protocols of prior work for fair comparison.

Nearly Symmetric Action Splits. To study temporal order sensitivity, we also define *nearly symmetric actions* as categories with visually similar frame appearances but opposite temporal order. We manually identify such categories in all three target datasets: HRI-30, IKEA-ASM, and Drive&Act, each of which contains a substantial proportion of such actions. Examples include *pick up* vs. *lay down* or *open bottle* vs. *close bottle*. Our splits result in 20/14/20 (10/7/10 pairs) symmetric and 14/19/10 non-symmetric actions in Drive&Act, IKEA-ASM, and HRI-30 respectively.

B. Comparison to Probing and PEFT Baselines

We first compare our model against two dominant adaptation paradigms for image-to-video transfer: probing and PEFT. Probing freezes the backbone with lightweight heads (linear [28], attentive [14], [30], self-attention [41]). These are efficient but largely permutation-invariant. PEFT methods (ST-Adapter [17], Vita-CLIP [18], M2-CLIP [33]) insert adapters or prompts for temporal modeling, but are heavier and remain underexplored on smaller domain-specific HRI datasets. Table I reports accuracy for nearly symmetric, non-symmetric, and all categories, alongside FLOPs and trainable parameters to quantify both the performance and efficiency of the models. The results reveal three consistent findings.

Large gains on symmetric actions. STEP consistently delivers the strongest results on *nearly symmetric actions*. On IKEA-ASM, it improves symmetric action recognition by 4.7% over attentive probing and still beats heavier PEFT methods (7–28M params). On HRI-30, STEP boosts symmetric performance to 82.1% (+7.8% over probing), surpassing PEFT baselines even with far lower computational cost. Similarly, on Drive&Act, STEP raises symmetric accuracy over probing by +10.2% to 69.98%, while PEFT methods fall short despite significantly higher compute budgets. These results confirm that explicit temporal modeling is essential for symmetric actions, where probing and PEFT fall short.

Strong performance on non-symmetric and overall accuracy. Importantly, STEP does not trade off symmetric accuracy for non-symmetric classes. On HRI-30, it achieves 96.8% non-symmetric and 87.02% overall accuracy, the best across all methods. On IKEA-ASM, it attains 80.6% non-symmetric and 76.8% overall, outperforming probing baselines, surpassing PEFT methods (e.g., M2-CLIP at 74.9%). On Drive&Act, PEFT closes the gap on non-symmetric actions (85.7% for M2-CLIP), yet STEP remains competitive with 84.0% non-symmetric and the strongest overall balance (78.4%). In summary, STEP is effective for both symmetric and non-symmetric actions, with the strongest gains observed for the symmetric ones, which are especially prominent in HRI tasks comprising fine-grained object manipulation.

Efficiency and robustness across backbones. A key strength of STEP is that it enables efficient VFM adaptation while delivering state-of-the-art performance. With only 2.6M trainable parameters and 410 GFLOPs, STEP is an order of magnitude smaller than typical PEFT frameworks (7-28M params and 900–1100 GFLOPs per pass). Although CNN-based task-specific models remain more lightweight, VFMs offer superior recognition quality and multi-task capabilities through shared frozen backbones (Table V). Compared to probing, STEP adds only 1-2% compute but delivers dramatic accuracy gains: +5 to +20 points on symmetric

and overall recognition. These improvements are consistent across both CLIP and DINOv2 backbones.

Comparison to fully fine-tuned. Previous approaches rely on fully fine-tuned heavy video backbones such as VideoSWIN [46], SlowOnly, and UniformerV2 [47], which represent the current SOTA on these benchmarks. These models achieve strong results (e.g., 86.6% overall on HRI-30 and 76.7% on Drive&Act) but at the cost of more trainable parameters, and they are highly task-specific. In contrast, STEP surpasses these models with minimal fine-tuning, showing that probing frozen VFMs can rival or exceed full fine-tuning while remaining more generalizable across tasks and domains.

C. Temporal Order Sensitivity Analysis

Impact of Test-time Frame Order Corruptions. Temporal order is generally an important element defining human activities, and a reliable action recognition model should show reduced performance if the event sequence is altered at test time. To evaluate this, we test probing and PEFT methods under correct and reverse frame order (Table II).

Dataset	Probing		PEFT	
	Attn.	STEP(Ours)	ST-Adapter	M2-CLIP
HRI-30	62.61	87.02	85.45	80.47
HRI-30_Reversed		42.26 (↓ 44.76)	39.76 (↓ 45.69)	37.38 (↓ 43.09)
IKEA-ASM	65.76	76.28	70.54	74.88
IKEA-ASM_Reversed		55.19(↓ 21.1)	67.28 (↓ 3.3)	68.91(↓ 5.9)
Drive&Act	68.65	78.40	75.19	77.73
Drive&Act_Reversed		59.83 (↓ 18.6)	61.16 (↓ 14.03)	59.44 (↓ 18.2)

TABLE II: Comparison of probing (Attentive probing vs. STEP) and PEFT methods with and without reversed frames at test time across HRI-30, Drive&Act, and IKEA-ASM.

Conventional probing methods remain unaffected in the reverse configuration, confirming their inherent permutation invariance and inability to model temporal order. In contrast, both PEFT methods and STEP show clear sensitivity to temporal corruption, with large drops in the reversed setting (e.g., –44.8% on HRI-30 for STEP). This demonstrates that STEP, despite operating only at the probing stage with 2.6M parameters, encodes order dependencies comparably to or better than heavy PEFT frameworks.

Class-wise symmetric action breakdown We further analyze the performance of STEP on individual *nearly symmetric actions* of the Drive&Act dataset. Table III presents their overall accuracy with the attentive probing baseline, the activity class most frequently confused with, and the corresponding confusion rate. Interestingly, we see that the model tends to reliably recognize *one* action in a symmetric pair, while the other is often mapped to its nearly symmetric counterpart. For example, *closing bottle* has a correct recognition rate of only 17% cases and is confused with *opening bottle* in 60% of cases. *Opening bottle*, on the other hand is correctly recognized 68% of times. With STEP, identifying *closing bottle* works 29% better, and the confusion with *opening bottle* falls by 20%. This pattern is repeated for most of the pairs.

True Activity Class	Top-1 Acc Attn.	Top-1 Acc Δ (↑)	Most Common Confusion Class	Confusion Attn.	Confusion Δ (↓)
closing_bottle	0.17	0.29	opening_bottle	0.60	-0.20
closing_door_inside	0.94	-0.12	opening_door_inside	0.06	0.12
closing_door_outside	0.73	0.27	opening_door_outside	0.27	-0.27
closing_laptop	0.22	0.06	Working on laptop	0.28	-0.06
entering_car	0.94	0.06	exiting_car	0.06	-0.06
exiting_car	1	0.0	-	0.0	0.0
fastening_seat_belt	0.89	0.05	unfastening_seat_belt	0.08	-0.06
fetching_an_object	0.51	0.32	placing_an_object	0.37	-0.29
opening_bottle	0.68	0.08	eating	0.11	0.05
opening_door_inside	0.19	0.56	closing_door_inside	0.56	-0.50
opening_door_outside	0.71	0.29	closing_door_outside	0.29	-0.29
opening_laptop	0.24	-0.12	working on laptop	0.47	-0.24
placing_an_object	0.62	0.18	fetching_an_object	0.22	-0.12
putting_laptop_into_bag	0.14	-0.14	placing_an_object	0.29	0.43
putting_on_jacket	0.33	0.05	taking_off_jacket	0.24	0.10
putting_on_sunglasses	0.84	-0.28	talking_on_phone	0.04	0.20
taking_laptop_from_bag	0.43	0.0	placing_an_object	0.43	-0.10
taking_off_jacket	0.67	-0.27	placing_an_object	0.20	0.27
taking_off_sunglasses	0.22	0.13	putting_on_sunglasses	0.22	-0.04
unfastening_seat_belt	0.5	0.25	fastening_seat_belt	0.25	-0.11

TABLE III: Analysis of the *nearly symmetric actions* of the Drive&Act dataset, including accuracy and most common confusions. Δ showcases comparison to STEP.

We also perform a similar analysis on the IKEA-ASM dataset, especially with the PEFT methods. As shown in Figure 3, STEP consistently outperforms PEFT methods across nearly all actions, with only a few cases where other methods perform comparably. Despite these exceptions, STEP maintains better performance overall, particularly in distinguishing nearly symmetric actions.

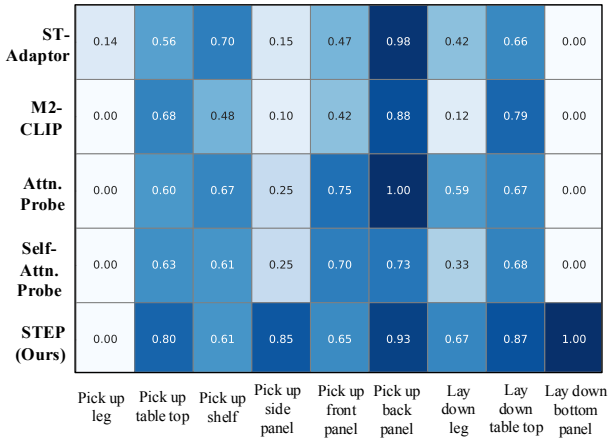

Fig. 3: Class-wise accuracy on nearly symmetric actions in IKEA-ASM. STEP outperforms PEFT and probing baselines.

D. Multi-Task Performance

Robotic systems must often perform several perception tasks at once, such as fine-grained activity recognition (FAR), atomic action recognition (AAR), and identifying the object under interaction (OUI), to provide intention-aware assistance. Existing PEFT methods [17], [31] train adapters or prompts separately for each task, forcing multiple backbone passes and scaling inference cost linearly with task count (Fig. 4), even under identical training budgets. Even when forced into a single-pass setup by reusing a single PEFT backbone across tasks, performance drops notably because the backbone overfits to the optimization task, hurting generalization on others (Table V). STEP avoids this by sharing a

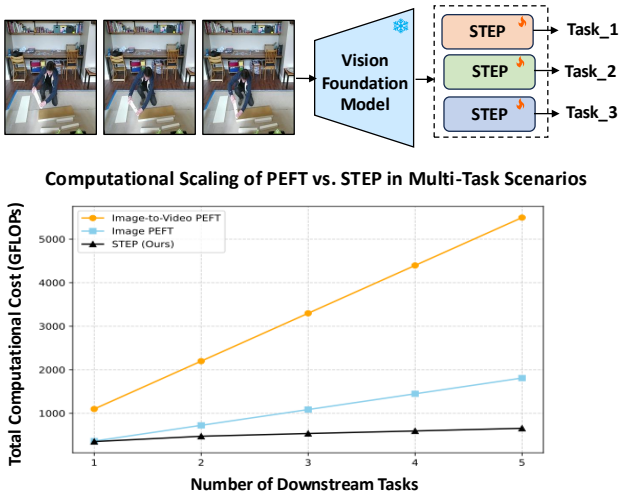

Fig. 4: Computational scaling of PEFT vs. STEP during inference in multi-task scenarios. PEFT cost grows linearly with tasks, while STEP remains constant.

frozen VFM backbone across tasks and attaching lightweight probes, enabling all objectives in a single pass and avoiding redundant backbone forward passes required by task-specific PEFT adapters. On IKEA-ASM, STEP reduces inference cost (GFLOPs) by up to 6× while improving accuracy across FAR, AAR, and OUI tasks. Moreover, adding new tasks only requires probes, not backbone re-optimization, making STEP highly scalable for real-time HRI.

E. STEP *as Temporal Extension of Image PEFT*

While image-to-video PEFTs [17], [33] perform strong spatio-temporal modeling, they are computationally heavy, requiring 7–28M parameters and high GFLOPs. In contrast, image-based PEFTs like Visual Prompt Tuning [31] (VPT) and AdaptFormer [32] are lightweight but focus mainly on spatial features, ignoring temporal reasoning. As shown in Table IV, adding STEP consistently boosts their performance across all datasets, with large gains on symmetric actions (+23.1% for VPT on HRI-30, +10.1% for AdaptFormer on Drive&Act) and up to +18.9% overall accuracy. Notably, STEP+image PEFT surpasses heavy video PEFTs on IKEA-ASM and Drive&Act, while approaching standalone STEP. Still, STEP alone delivers the best overall performance, establishing state-of-the-art results and highlighting its versatility as a temporal plug-in for lightweight Image-based PEFTs.

F. Qualitative Comparisons

To better understand how STEP improves recognition, we visualize attention maps (Fig. 5). Frozen DINOv2 largely attends to the background, overlooking fine-grained cues. PEFT shifts focus toward the human and furniture but still diffuses attention to irrelevant background objects. In contrast, DINOv2 + STEP sharply concentrates on the human–object interaction (e.g., the hand and manipulated object), indicating that our lightweight probe explicitly drives attention toward sequence-relevant regions.

We further analyze class separability through t-SNE embeddings of *nearly symmetric action* pairs such as *pick up drill* vs. *put down drill* (Figure 6). The results reinforce our earlier observations: attentive probing produces almost complete overlap between classes, confirming its permutation-invariant nature and inability to capture temporal order. PEFT methods achieve improved separation, demonstrating they perform temporal modeling, yet residual overlaps remain, reflecting their reduced effectiveness in symmetric actions. In contrast, STEP produces sharp, well-isolated clusters with minimal confusion, showing our lightweight probe effectively encodes sequence direction and disambiguates visually similar but temporally opposite actions.

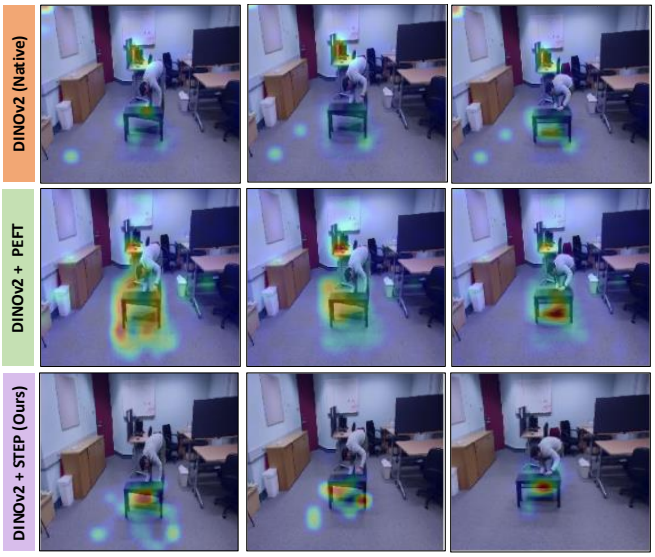

Fig. 5: Attention maps for action recognition. DINOv2 attends to background, PEFT spreads attention diffusely, while our DINOv2 + STEP focuses on the human–object interaction, aiding discrimination of symmetric actions.

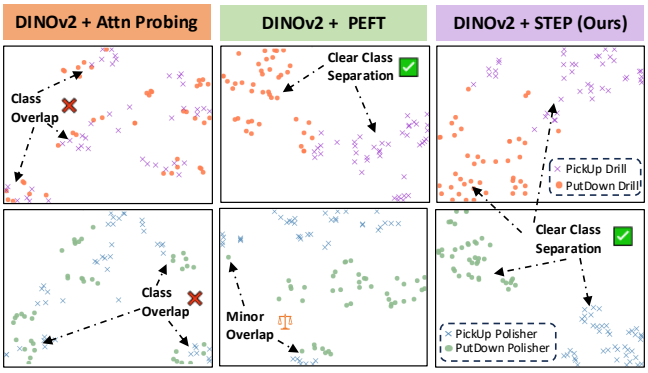

Fig. 6: t-SNE visualizations of HRI-30 dataset. Unlike probing (overlap) and PEFT (minor overlap), DINOv2 + STEP cleanly separates *nearly symmetric actions*, enabling safer and more reliable action understanding in HRI.

G. Ablation Studies

Impact of Simplified Attention block. We analyze the effect of simplifying the probing attention layer by progressively removing feedforward layers, normalization, and residual connections. These components mainly stabilize deep transformers but add unnecessary overhead in our

Method	Classifier	HRI-30			IKEA-ASM			Drive&Act		
		Sym	N-Sym	Ovr.	Sym	N-Sym	Ovr.	Sym	N-Sym	Ovr.
VPT[31]	Linear	51.07	85.53	62.50	60.62	73.82	69.54	56.04	79.98	68.49
	STEP	74.11	**96.07**	81.43	**70.88**	**78.30**	**75.89**	**72.72**	82.64	77.88
	Δ(↑)	+23.07	+10.54	+18.93	+10.26	+4.48	+6.35	+16.68	+2.66	+9.39
AdaptFormer[32]	Linear	64.28	76.79	68.45	65.16	73.13	70.54	60.43	83.99	73.73
	STEP	**75.53**	93.58	**81.55**	65.66	76.11	**72.56**	70.59	**85.99**	**78.61**
	Δ(↑)	+11.25	+16.78	+13.09	+0.50	+2.98	+2.02	+10.16	+2.00	+4.88

TABLE IV: Comparison of STEP with image PEFTs on HRI-30, IKEA-ASM, and Drive&Act. STEP consistently improves symmetric and overall accuracy.

Method	# Passes	GFLOPs	FAR Acc.	AAR Acc.	OUI Acc.
ST-Adapter [17]	1	1099	52.63	47.83	74.26
	3	3297	70.54	72.95	74.26
VPT [31]	1	**362.4**	56.05	54.03	66.28
	3	1087	69.54	65.82	66.28
STEP (Ours)	**1**	**536.5**	**76.80**	**75.28**	**76.51**

TABLE V: Multi-task: STEP achieves higher accuracy with less GFLOPs.

Method	D&A	IKEA
FF + LN + Skip (Block)	77.27	73.72
LN + Skip (No FF)	76.55	74.88
Ours (only Attn. layer)	**78.40**	**76.28**

TABLE VI: Comparison of attention block variants.

Method	D&A	IKEA
Only Global CLS	70.19	54.11
Only Patch Tokens	77.29	73.49
Combined (STEP)	**78.40**	**76.28**

TABLE VII: Ablation of CLS vs. Patch Tokens.

Method	D&A	IKEA
Fixed PE	76.76	69.53
Hybrid PE	77.78	74.88
Learnable PE (Ours)	**78.40**	**76.28**

TABLE VIII: Ablation of different PE types.

Method	D&A	IKEA
Token-wise PE	73.21	68.37
Frame-wise PE (STEP)	**78.40**	**76.28**

TABLE IX: Ablation of token vs. frame-wise PE.

Method	HRI-30	Drive&Act	IKEA-ASM
Global CLS Token			
Self-Attn. Probing	76.19	71.16	60.78
Self-Attn. Probing w Global CLS	78.88	72.00	68.76
Frame-wise Temporal PE			
Self-Attn. Probing w Temporal PE	84.74	77.07	71.56
STEP	**87.02**	**78.40**	**76.28**

TABLE X: Ablation of individual components and comparison to self-attention probing without our modifications.

single-layer setup. Table VI shows that our simplified attention block, retaining only multi-head attention, improves performance while reducing complexity, achieving 78.40% on Drive&Act (+1.13%), and 76.28% on IKEA-ASM (+1.93%). This demonstrates that a minimal design not only reduces complexity but also enhances accuracy in probing.

Impact of Global CLS Token and Frame-wise Temporal Positional Encoding. We analyze the contributions of STEP's components by evaluating the Global CLS Token and Frame-wise Temporal Positional Encoding (PE). Table X shows that adding the Global CLS token consistently improves accuracy, with the largest +8% gain on IKEA-ASM, where capturing global context is critical. The Frame-wise Temporal PE further boosts accuracy on both Drive&Act (+5.91%) and IKEA-ASM (+10.78%), reinforcing its importance for nearly symmetric actions. In Table VII, relying on CLS alone underperforms on fine-grained datasets where local detail is crucial, whereas patch embeddings complement CLS by capturing low-level cues. Combining both, as in STEP, yields the strongest results across all benchmarks.

Impact of Different Positional Encoding Schemes. Table VIII evaluates different positional encodings for DINOv2. Learnable PE, used in STEP, consistently outperforms Fixed and Hybrid variants, with the largest gains on IKEA-ASM. We further compare our frame-wise PE with conventional token-wise PE (Table IX) and find that, despite using fewer parameters, it achieves higher accuracy, highlighting its effectiveness in capturing temporal order.

V. Conclusion and Limitations

We studied parameter-efficient image-to-video probing, focusing on *nearly symmetric actions* – visually similar actions unfolding in reverse order. Existing approaches struggle with such actions as they ignore frame order due to permutation-invariant attention. To address this, we proposed STEP, introducing simple yet effective modifications to self-attention probing, improving temporal sensitivity. STEP outperforms probing, PEFT and fully fine-tuned methods across HRI-30, IKEA-ASM, and Drive&Act, achieving state-of-the-art accuracy with only a fraction of the trainable parameters. It also enables multi-task inference in a single backbone pass, reducing computation up to 6× compared to PEFT. Limitations remain: inference still requires a full pass through large transformer backbones; and while frozen probing excels in today's small-scale HRI benchmarks, its advantage may shrink if large symmetric-action datasets become available, where full fine-tuning and PEFT could dominate. Still, in realistic HRI scenarios, where data and compute are limited, STEP offers a practical balance of efficiency and accuracy.

Acknowledgments

The research published in this article is supported by the Deutsche Forschungsgemeinschaft (DFG) under Germany's Excellence Strategy – EXC 2120/1 –390831618. The authors also thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Thinesh Thiyakesan Ponbagavathi. The authors also gratefully acknowledge the computing time provided on the high-performance computer HoreKa by the National High-Performance Computing Center at KIT. HoreKa is partly funded by the German Research Foundation (DFG).

References

[1] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, *et al.*, "On the opportunities and risks of foundation models," *arXiv preprint arXiv:2108.07258*, 2021.

[2] J. Shang, K. Schmeckpeper, B. B. May, M. V. Minniti, T. Kelestemur, D. Watkins, and L. Herlant, "Theia: Distilling diverse vision foundation models for robot learning," 2024.

[3] X. Li, M. Liu, H. Zhang, C. Yu, J. Xu, H. Wu, C. Cheang, Y. Jing, W. Zhang, H. Liu, H. Li, and T. Kong, "Vision-language foundation models as effective robot imitators," in *The Twelfth International Conference on Learning Representations*, 2024.

[4] M. T. Khan and A. Waheed, "Foundation model driven robotics: A comprehensive review," 2025. [Online]. Available: https://arxiv.org/abs/2507.10087

[5] T. Wang, Z. Liu, L. Wang, M. Li, and X. V. Wang, "Data-efficient multimodal human action recognition for proactive human–robot collaborative assembly: A cross-domain few-shot learning approach," *Robotics and Computer-Integrated Manufacturing*, vol. 89, 2024.

[6] Z. Wang, P. Li, H. Liu, Z. Deng, C. Wang, J. Liu, J. Yuan, and M. Liu, "Recognizing actions from robotic view for natural human-robot interaction," 2025.

[7] W. Kay, J. Carreira, K. Simonyan, B. Zhang, C. Hillier, S. Vijayanarasimhan, F. Viola, T. Green, T. Back, P. Natsev, M. Suleyman, and A. Zisserman, "The kinetics human action video dataset," 2017. [Online]. Available: https://arxiv.org/abs/1705.06950

[8] R. Goyal, S. E. Kahou, V. Michalski, J. Materzyńska, S. Westphal, H. Kim, V. Haenel, I. Fruend, P. Yianilos, M. Mueller-Freitag, F. Hoppe, C. Thurau, I. Bax, and R. Memisevic, "The "something something" video database for learning and evaluating visual common sense," 2017. [Online]. Available: https://arxiv.org/abs/1706.04261

[9] F. Iodice, E. De Momi, and A. Ajoudani, "Hri30: An action recognition dataset for industrial human-robot interaction," in *2022 26th International Conference on Pattern Recognition (ICPR)*, 2022.

[10] D. Wei, J. J. Lim, A. Zisserman, and W. T. Freeman, "Learning and using the arrow of time," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[11] A. Ghodrati, E. Gavves, and C. G. Snoek, "Video time: Properties, encoders and evaluation," *arXiv preprint arXiv:1807.06980*, 2018.

[12] I. Misra, C. L. Zitnick, and M. Hebert, "Shuffle and learn: unsupervised learning using temporal order verification," in *European conference on computer vision*. Springer, 2016, pp. 527–544.

[13] B. Fernando, H. Bilen, E. Gavves, and S. Gould, "Self-supervised video representation learning with odd-one-out networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[14] J. Yu, Z. Wang, V. Vasudevan, L. Yeung, M. Seyedhosseini, and Y. Wu, "Coca: Contrastive captioners are image-text foundation models," 2022. [Online]. Available: https://arxiv.org/abs/2205.01917

[15] A. van den Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," 2019.

[16] J. Lee, Y. Lee, J. Kim, A. Kosiorek, S. Choi, and Y. W. Teh, "Set transformer: A framework for attention-based permutation-invariant neural networks," in *Proceedings of the 36th International Conference on Machine Learning*, 2019, pp. 3744–3753.

[17] J. Pan, Z. Lin, X. Zhu, J. Shao, and H. Li, "St-adapter: parameter-efficient image-to-video transfer learning," in *Proceedings of the 36th International Conference on Neural Information Processing Systems*, ser. NIPS '22. Red Hook, NY, USA: Curran Associates Inc., 2022.

[18] S. T. Wasim, M. Naseer, S. Khan, F. S. Khan, and M. Shah, "Vita-clip: Video and text adaptive clip via multimodal prompting," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 23 034–23 044.

[19] J. Zhang, P. Wang, and R. X. Gao, "Hybrid machine learning for human action recognition and prediction in assembly," *Robotics and Computer-Integrated Manufacturing*, vol. 72, p. 102184, 2021.

[20] Q. Xiong, J. Zhang, P. Wang, D. Liu, and R. X. Gao, "Transferable two-stream convolutional neural network for human action recognition," *Journal of Manufacturing Systems*, vol. 56, pp. 605–614, 2020.

[21] Y. Ben-Shabat, X. Yu, F. Saleh, D. Campbell, C. Rodriguez-Opazo, H. Li, and S. Gould, "The ikea asm dataset: Understanding people assembling furniture through actions, objects and pose," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2021, pp. 847–859.

[22] M. Martin, A. Roitberg, M. Haurilet, M. Horne, S. Reiß, M. Voit, and R. Stiefelhagen, "Drive&act: A multi-modal dataset for fine-grained driver behavior recognition in autonomous vehicles," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019.

[23] R. X. Gao, L. Wang, P. Wang, J. Zhang, H. Liu, X. V. Wang, J. Váncza, and Z. Kemény, *Human Motion Recognition and Prediction for Robot Control", bookTitle="Advanced Human-Robot Collaboration in Manufacturing*. Springer International Publishing, 2021.

[24] D. Moutinho, L. F. Rocha, C. M. Costa, L. F. Teixeira, and G. Veiga, "Deep learning-based human action recognition to leverage context awareness in collaborative assembly," *Robotics and Computer-Integrated Manufacturing*, vol. 80, p. 102449, 2023.

[25] J. Fan, P. Zheng, and C. K. M. Lee, "A vision-based human digital twin modeling approach for adaptive human–robot collaboration," *Journal of Manufacturing Science and Engineering*, vol. 145, no. 12, p. 121002, 07 2023.

[26] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," 2021.

[27] M. Oquab, T. Darcet, T. Moutakanni, H. Vo, M. Szafraniec, V. Khalidov, P. Fernandez, D. Haziza, F. Massa, A. El-Nouby, M. Assran, N. Ballas, W. Galuba, R. Howes, P.-Y. Huang, S.-W. Li, I. Misra, M. Rabbat, V. Sharma, G. Synnaeve, H. Xu, H. Jegou, J. Mairal, P. Labatut, A. Joulin, and P. Bojanowski, "Dinov2: Learning robust visual features without supervision," 2024.

[28] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," 2020.

[29] L. Yuan, N. B. Gundavarapu, L. Zhao, H. Zhou, Y. Cui, L. Jiang, X. Yang, M. Jia, T. Weyand, L. Friedman, M. Sirotenko, H. Wang, F. Schroff, H. Adam, M.-H. Yang, T. Liu, and B. Gong, "Videoglue: Video general understanding evaluation of foundation models," 2024.

[30] A. Bardes, Q. Garrido, J. Ponce, M. Rabbat, Y. LeCun, M. Assran, and N. Ballas, "Revisiting feature prediction for learning visual representations from video," *arXiv:2404.08471*, 2024.

[31] M. Jia, L. Tang, B.-C. Chen, C. Cardie, S. Belongie, B. Hariharan, and S.-N. Lim, "Visual prompt tuning," in *European Conference on Computer Vision (ECCV)*, 2022.

[32] S. Chen, C. Ge, Z. Tong, J. Wang, Y. Song, J. Wang, and P. Luo, "Adaptformer: Adapting vision transformers for scalable visual recognition," *arXiv preprint arXiv:2205.13535*, 2022.

[33] M. Wang, J. Xing, B. Jiang, J. Chen, J. Mei, X. Zuo, G. Dai, J. Wang, and Y. Liu, "M2-clip: A multimodal, multi-task adapting framework for video action recognition," *arXiv preprint arXiv:2401.11649*, 2024.

[34] Y. Ma, G. Xu, X. Sun, M. Yan, J. Zhang, and R. Ji, "X-clip: End-to-end multi-grained contrastive learning for video-text retrieval," in *ACMMM*, 2022.

[35] M. Wang, J. Xing, and Y. Liu, "Actionclip: A new paradigm for video action recognition," 2021.

[36] Z. Lin, S. Geng, R. Zhang, P. Gao, G. de Melo, X. Wang, J. Dai, Y. Qiao, and H. Li, "Frozen clip models are efficient video learners," 2022. [Online]. Available: https://arxiv.org/abs/2208.03550

[37] R. Liu, J. Huang, G. Li, J. Feng, X. Wu, and T. H. Li, "Revisiting temporal modeling for clip-based image-to-video knowledge transferring," 2023.

[38] W. Wu, X. Wang, H. Luo, J. Wang, Y. Yang, and W. Ouyang, "Bidirectional cross-modal knowledge exploration for video recognition with pre-trained vision-language models," 2023.

[39] S. Tu, Q. Dai, Z. Wu, Z.-Q. Cheng, H. Hu, and Y.-G. Jiang, "Implicit temporal modeling with learnable alignment for video recognition," 2023. [Online]. Available: https://arxiv.org/abs/2304.10465

[40] Z. Tong, Y. Song, J. Wang, and L. Wang, "Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training," *NeurIPS*, 2022.

[41] H. Luo, L. Ji, M. Zhong, Y. Chen, W. Lei, N. Duan, and T. Li, "Clip4clip: An empirical study of clip for end to end video clip retrieval," 2021. [Online]. Available: https://arxiv.org/abs/2104.08860

[42] Y. Wang, K. Li, X. Li, J. Yu, Y. He, C. Wang, G. Chen, B. Pei, Z. Yan, R. Zheng, J. Xu, Z. Wang, Y. Shi, T. Jiang, S. Li, H. Zhang, Y. Huang, Y. Qiao, Y. Wang, and L. Wang, "Internvideo2: Scaling foundation models for multimodal video understanding," 2024. [Online]. Available: https://arxiv.org/abs/2403.15377

[43] C. Cao, P. Han, Y. zhang, Y. Yu, Q. Lv, L. Min, and Y. zhang, "Task-adapter++: Task-specific adaptation with order-aware alignment for few-shot action recognition," 2025.

[44] W. Price and D. Damen, "Retro-actions: Learning'close'by time-reversing'open'videos," in *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019, pp. 0–0.

[45] A. Vaswani *et al.*, "Attention is all you need," in *NeurIPS*, 2017.

[46] Z. Liu, J. Ning, Y. Cao, Y. Wei, Z. Zhang, S. Lin, and H. Hu, "Video swin transformer," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 3202–3211.

[47] K. Li, Y. Wang, Y. He, Y. Li, Y. Wang, L. Wang, and Y. Qiao, "Uniformerv2: Spatiotemporal learning by arming image vits with video uniformer," 2022.

[48] A. Dosovitskiy *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," in *ICLR*, 2021.