DO FRAME-ORDER DIFFERENCES BETWEEN VIDEO PRETRAINING STAGES SURVIVE FINE-TUNING?

Anonymous authors
Paper under double-blind review

ABSTRACT

Linear probes on frozen features are a standard way to compare what different pretraining gives a video model. We ask whether the differences they reveal survive fine-tuning, for frame order, measured as the balanced-accuracy drop when a clip's frames are shuffled. Across the published pretraining stages of one architecture (VideoMAE), on two binary fall-detection datasets, one synthetic and one in the wild, frozen checkpoints differ markedly at the output, and supervised Kinetics training lowers the fall probe's order sensitivity even as it raises accuracy on the synthetic dataset. Fine-tuning collapses most of these differences. On the synthetic dataset the fraction of the gap between the self-supervised (MAE-only) and Kinetics checkpoints that fine-tuning closes is $\kappa = 1.08$ [0.84, 1.33], so the gap closes completely within error. It also closes under a standard fine-tuning schedule ($\kappa = 0.85$ [0.22, 1.36]), under re-drawn and tubelet-preserving shuffles and with nonlinear probes, and it does not survive on native-resolution input or with cross-fitted probes, where refitted probes even find the fine-tuned order reversed. On the in-the-wild dataset, where κ is uninformative, the spread between checkpoints shrinks by 69%. Between these two checkpoints, what closes is how the fall read-out uses frame order, not the linearly readable direction information the output carries: on a task defined by frame order their frozen outputs are not distinguishable, and both kept most of their direction information through fine-tuning for fall detection. In exploratory layer-wise probes, the frozen checkpoints differ mainly in the upper layers, and under the standard schedule they still differ at blocks 9 and 10 yet agree at the output. In TimeSformer, whose frozen Kinetics output carries no linearly readable direction information, the fall-task gap does not close under either recipe ($\kappa = 0.08$ and 0.43, both intervals containing 0), and most of the gap in direction information persists; the same holds in Hiera, where we predicted before fine-tuning that its fall-task gap would not close ($\kappa = 0.03$ [-0.54, 0.45]). On an order-defined task TimeSformer's gap closes ($\kappa = 0.86$ [0.69, 1.01]) and Hiera's, as the probe reads it, does not ($\kappa = 0.25$ [-0.24, 0.60]). Within VideoMAE, fall-task gaps closed even where differences in direction information remained. A frozen comparison of frame-order sensitivity on a task that does not require order thus describes, for VideoMAE, mostly how a probe reads the top of the network, and it cannot tell which differences fine-tuning will keep.

1 INTRODUCTION

Linear probes on frozen features are a standard instrument for asking what a pretraining objective puts into a video representation. A recent wave of studies probes video foundation models across depth for motion, intuitive physics and temporal structure (Joseph et al., 2026; Punzo et al., 2026; Musa et al., 2026), and a direct test of whether a model uses more than a bag of frames is its sensitivity to frame order (Yun et al., 2022). Such comparisons are often read as statements about the pretraining: this objective captures temporal structure, that one discards it.

Consider the published checkpoints of VideoMAE (Tong et al., 2022). The checkpoint supervised on Kinetics-400 is its self-supervised parent plus one supervised stage. Frozen, it is the more accurate of the two on both fall datasets (as point estimates), and it is also much less sensitive to frame order

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

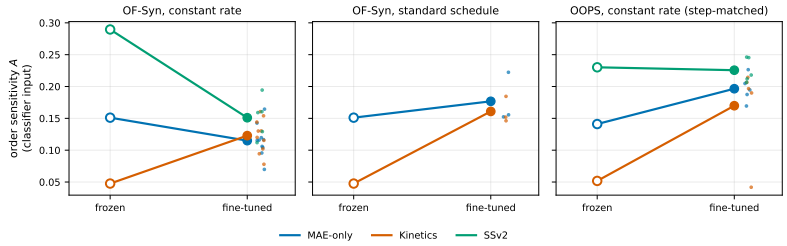


Figure 1: Fine-tuning brings the checkpoints together. Order sensitivity A (the drop in a linear probe's balanced accuracy when a clip's frames are shuffled) at the classifier's input before (open) and after (filled) fine-tuning, one line per VideoMAE checkpoint; small points are seeds. Left: OF-Syn, constant rate (10, 10 and 5 seeds). Centre: OF-Syn, the standard schedule (3 seeds). Right: OOPS, constant rate, step-matched budget (5 seeds each, including one Kinetics seed that failed to fit the task).

at its output (Section 4.1). A frozen comparison would rank it last for any task that needs temporal order, yet on a task defined by frame order it is not distinguishable from its parent (Section 4.2). Most applications, however, do not use frozen features; they fine-tune them. This paper asks whether the differences a frozen probe reveals between pretraining stages survive fine-tuning, for frame order.

We hold the architecture fixed and vary only the pretraining stage, using three published checkpoints of one network, and we measure order sensitivity with one instrument applied identically before and after fine-tuning: the drop in a linear probe's balanced accuracy when the frames of each test clip are shuffled (Section 3; Figure 3). We fine-tune on two binary fall-detection datasets, one synthetic (OF-Syn) and one in the wild (OOPS), under two recipes. We fixed the central test and its decision rule before their results existed; the reading that separates use from direction information was formed after the first results and then tested on further models, and the layer-wise analyses are exploratory.

Most of the difference does not survive (Figure 1). On OF-Syn, fine-tuning closes the frozen gap between the self-supervised and the Kinetics checkpoint completely within error (the fraction closed is $\kappa = 1.08$ [0.84, 1.33]), and the gap still closes under a standard schedule with VideoMAE's layer-wise decay, which trains markedly better models ($\kappa = 0.85$ [0.22, 1.36], three seeds). It closes under re-drawn and tubelet-preserving shuffles, with nonlinear probes and in the fine-tuned models' own heads, and it does not survive on native-resolution input or with a probe that never sees a clip the backbone was trained on, where the refitted probes (not the own heads) even find the fine-tuned order reversed. On OOPS, where κ is uninformative, the between-checkpoint spread at the output shrinks by 69%, against 90% on OF-Syn.

Between MAE-only and Kinetics, what closes is how the fall read-out uses order, not the linearly readable direction information (which way a clip runs) that the output carries, most of which both kept through fine-tuning for fall detection. The collapse is not universal (Section 4.2). In TimeSformer and Hiera, whose Kinetics outputs carry less direction information than their counterparts', the fall-task gap survives fine-tuning together with most of the direction gap; for Hiera we predicted the fall-task outcome from the frozen direction gap before fine-tuning ($\kappa = 0.03$ [-0.54, 0.45]). On the order-defined task TimeSformer's gap closes, while Hiera's does not, as the probe reads it, though the models' own heads close it. Within VideoMAE, fall fine-tuning closes most or all of each fall-task gap even where direction differences remain, so the frozen direction gap anticipated survival across architectures but not within VideoMAE (an observation made after the results, and confounded with architecture).

Exploratory layer-wise probing shows where the difference lies (Section 5): frozen, the checkpoints differ mainly in the upper layers, and Kinetics supervision lowers order sensitivity at the output but not its mid-network peak. Under the standard schedule the checkpoints keep differing at blocks 9 and 10, yet agree at the output, so the agreement does not require fine-tuning to overwrite the network. We also report two measurement cautions: a layer-wise probe's peak depth may not be identifiable at evaluation sizes like ours even when its drop is, and frozen values move with preprocessing, by checkpoint-specific amounts, and with the library version (Sections 3, 4.1 and 5).

The conclusion concerns what frozen comparisons measure: for VideoMAE, a frozen-probe ranking of pretraining stages by frame-order sensitivity describes the top of a pretrained network and is a poor guide to the fine-tuned model, and a frozen fall-task probe alone cannot tell which differences fine-tuning will keep. For accuracy the analogous observations are familiar: pretraining changes where training starts more than where it ends (He et al., 2019), and linear-probe and fine-tuned evaluations rank pretrained models differently (Zhai et al., 2019; Newell & Deng, 2020) (for MAE's variants, "linear probing and fine-tuning results are largely uncorrelated" (He et al., 2022)). We show the same for a probed temporal property of one architecture's pretraining stages, and locate it in the network. That the fall read-out's use of order changes while MAE-only's and Kinetics' direction information stays matches findings, in language models and synthetic tasks, that fine-tuning mostly re-uses mechanisms that pretraining built (Prakash et al., 2024; Jain et al., 2024).

2 Related work

Probing video models across depth. Yun et al. (2022) probed frame order at each block of supervised video transformers, found that it fades in later blocks, and added an order-prediction loss at the final block to keep it there. Dynamic information in Kinetics-trained video transformers peaks about halfway through the network (Kowal et al., 2022); layer-wise probes of video foundation models find intuitive physics (Joseph et al., 2026) and camera motion (Musa et al., 2026) most decodable mid-network, and intuitive physics most accessible at intermediate to late depth (Punzo et al., 2026), and a vision encoder's best embeddings lie in its intermediate layers (Bolya et al., 2025); Joseph et al. (2026) and Punzo et al. (2026) also use shuffled-video controls. These studies compare separately pretrained models, often differing in architecture, data and objective at once; for our main pair we vary only the supervision applied to one set of weights, read the ordered-minus-shuffled difference of a task probe at every block, and cross it with fine-tuning. Shuffling frames is an established way to remove temporal information (Sevilla-Lara et al., 2021), and the intermediate representations of most video models examined are biased toward static information (Kowal et al., 2022); we do not test methods that push representations toward motion by augmentation (Ding et al., 2022) or auxiliary temporal objectives (Bai et al., 2020), or architectures whose attention follows motion paths (Patrick et al., 2021). Bagad & Zisserman (2025) test time-sensitivity with pairs of temporally opposite actions that frozen embeddings often fail to separate linearly, and Ponbagavathi & Roitberg (2025) show that probes over per-frame image features are blind to the order of such nearly symmetric actions; our order-defined tasks follow that design. Intermediate-layer read-outs exist (Uselis & Oh, 2025; Ciernik et al., 2026); we propose none.

What fine-tuning does to representations. Probing, similarity analysis and layer ablations of BERT before and after fine-tuning show that it is conservative, changing the upper layers most and leaving most probed information decodable (Merchant et al., 2020); the sign of the change in a probed property varies across encoders and tasks, and strong pooling alone can match most apparent gains (Mosbach et al., 2020). Fine-tuning can distort pretrained features (Kumar et al., 2022), its change concentrates near the output in language models (Billa, 2026; Zhao et al., 2026), and models fine-tuned from the same weights stay in one basin of the loss landscape (Neyshabur et al., 2020). Closest to our question, Shekhar et al. (2023) find that image models trained with similar objectives learn similar features and, comparing a masked autoencoder with joint-embedding models layer by layer before and after fine-tuning, that fine-tuning brings the masked autoencoder's representation toward theirs, most in the layers that differed most. They compare separately trained objectives by representational similarity; none of these works compares a probed property *between the pretraining stages of one set of weights* before and after fine-tuning, which, with where in the network the difference lives, is our object of study.

Frozen evaluation and fine-tuning. For accuracy (see also Section 1), regularisers that slightly raise ImageNet accuracy make frozen features transfer worse (Kornblith et al., 2019), and rankings of video self-supervised methods change with the downstream domain, data and task (Thoker et al., 2022; 2025), and video foundation models rank differently with frozen backbones than after end-to-end fine-tuning (Yuan et al., 2024). We ask the same question of a representational property. On method we follow the probing literature's cautions: probe accuracy does not show that a property is used (Ravichander et al., 2021; Belinkov, 2022), and probe results need controls (Hewitt & Liang, 2019); ours include a structural null on which A must be exactly zero (Section 3).

3 MEASURING ORDER SENSITIVITY BEFORE AND AFTER FINE-TUNING

The quantity. Given a feature extractor and a labelled video task, we fit a linear probe (logistic regression with $C = 1$ on standardised features) on the training clips and measure its balanced accuracy on held-out clips twice: as presented, $\mathrm{bAcc}_{\mathrm{ord}}$, and with each clip's frames permuted, $\mathrm{bAcc}_{\mathrm{shuf}}$. Order sensitivity is

$$A \;=\; \mathrm{bAcc}_{\mathrm{ord}} - \mathrm{bAcc}_{\mathrm{shuf}}. \tag{1}$$

Every test clip has its own random permutation of its 16 frames, drawn once and reused for every model, and the permutation is applied to the pixels before the backbone, so the two conditions present exactly the same set of frames. A read-out that is a symmetric function of the frames therefore has $A = 0$ by construction. We check the construction for every model by encoding each frame separately and mean-pooling, which gives exactly zero in every case, so a nonzero A must come from an encoder that is sensitive to order. Because VideoMAE embeds frames in pairs, as tubelets, shuffling also changes the tubelets; a shuffle of whole frame pairs, which leaves every tubelet intact, keeps the checkpoints' ordering and the fine-tuning result (Appendix B). A is signed, since shuffling can help, and is never read as a fraction of the ordered accuracy. A larger fitting set (the clips the probe is fitted on) would likely give larger values, so we compare values only at one fitting size and within a dataset.

Where it is read. We write A for the value at the classifier's input and $A(\ell)$ for the mean-pooled output of block ℓ with no normalisation, the bare read-out; "the output" is always the last block, so $A(12)$ and A are two read-outs of the same block. Comparisons across all three checkpoints (Table 1, the spreads of Table 2) and all layer-wise results use the bare read-out, defined identically for every checkpoint. The κ test and Figure 1 use the classifier's input (the pooled state after the final normalisation, freshly initialised for MAE-only; Appendix A), which the fine-tuned classifier reads, measured in the same job for the frozen model (A_{pre}) and each fine-tuned seed (A_{ft}). The two are close for frozen models (MAE-only/Kinetics gap 0.103 at the classifier's input, 0.094 bare) but not for fine-tuned ones, whose normalisation is trained, so we never mix them. The probe on a fine-tuned backbone is fitted on the clips it was fine-tuned on; a probe cross-fitted on the evaluation split agrees (Section 4.2).

Checkpoints. We use the three published VideoMAE ViT-B checkpoints (Tong et al., 2022). *MAE-only* is self-supervised on Kinetics-400 for 1600 epochs; *Kinetics* is that model after supervised training on Kinetics-400 (Kay et al., 2017); *SSv2* comes from a separate 2400-epoch self-supervised run on Something-Something v2 itself, followed by supervised training on it (Goyal et al., 2017). MAE-only and Kinetics are the one pair that differs by a single supervised stage (apart from the normalisation before the classifier; Appendix A); SSv2 differs in its self-supervised run as well, and we use it as a third point rather than as a step on the same path. For scale we add the VideoMAE-large pair (MAE-only and Kinetics, 24 blocks), frozen and in one fine-tuning arm. To test whether the result extends beyond VideoMAE, we add published checkpoint pairs of three further architectures (Appendix C): R(2+1)D-34 (Tran et al., 2018), weakly supervised on Instagram videos and then on Kinetics (Ghadiyaram et al., 2019); TimeSformer-B (Bertasius et al., 2021), supervised on Kinetics or on SSv2 from ImageNet weights; and Hiera-B (Ryali et al., 2023), a masked autoencoder and its Kinetics child.

Data. OF-Syn is the synthetic domain of OmniFall (Schneider et al., 2026): 15,344 training and 1,928 evaluation clips of 16 frames, labelled fall or no fall, with 1,133 source groups in the evaluation split. OOPS is our clip set from the test split of OmniFall's in-the-wild component, built on the fail videos of Epstein et al. (2020): 2,804 clips from 572 videos, most with clips of both classes, so the label depends largely on where in a video a clip lies; since OmniFall uses this component only for testing, we split it ourselves by source video into 1,962 training and 842 evaluation clips. Both are binary tasks, and every interval resamples the evaluation groups (source groups on OF-Syn, videos on OOPS). Neither strictly needs frame order, so we also relabel OF-Syn into an order-defined task, *up/down*: clips of a person standing up against clips of a person falling, sitting down or lying down, so that a reversed clip of either class resembles a clip of the other (5,141 training and 653 evaluation clips from the same videos and splits; Appendix C).

Table 1: Frozen order sensitivity at the output (bare read-out of the last block), $A_{\mathrm{pre}}(12)$ with its 95% group-bootstrap interval, and the balanced accuracy of the same probe on ordered clips. *Shared*: the 112-pixel pipeline of every fine-tuning and depth result except the native-input arm; *native*: each checkpoint's own processor at full resolution.

	OF-Syn, shared		OF-Syn, native		OOPS, shared	
checkpoint	$A_{\mathrm{pre}}(12)$	bAcc	$A_{\mathrm{pre}}(12)$	bAcc	$A_{\mathrm{pre}}(12)$	bAcc
MAE-only	0.137 [0.112, 0.162]	0.792	0.180 [0.155, 0.209]	0.797	0.128 [0.095, 0.162]	0.702
Kinetics	0.043 [0.020, 0.069]	0.838	0.116 [0.090, 0.141]	0.848	0.057 [0.024, 0.095]	0.720
SSv2	0.279 [0.255, 0.306]	0.814	0.305 [0.279, 0.329]	0.827	0.227 [0.193, 0.262]	0.747

Fine-tuning. Every model is fine-tuned end to end with cross-entropy and a linear head on the pooled, normalised features. The *constant-rate* recipe uses AdamW with learning rate 10^{-4}, batch size 8 and weight decay 10^{-4}, for 10 epochs on OF-Syn; on OOPS, whose training set is eight times smaller, it runs 80 epochs so that the number of optimisation steps matches (the step-matched budget, our primary one because probe-measured change tracks updates rather than epochs (Mehrafarin et al., 2022); the epoch-matched 10-epoch budget is secondary). The *standard* schedule adds VideoMAE's schedule to this recipe: layer-wise learning-rate decay 0.75, one epoch of linear warmup and cosine decay (VideoMAE's own recipe also uses weight decay 0.05, a batch-scaled learning rate and heavy augmentation, which we do not). We use 5 seeds per arm for the κ test and 10 for its repeat, 5 for SSv2 and for OOPS, and 3 for the standard schedule, the layer-wise probes and the further arms of Section 4.2.

Statistics. Intervals are 95% percentile bootstrap intervals: those on κ and fine-tuned values resample evaluation groups and then seeds, those on spreads and layer-wise quantities resample groups only, and one group resample is shared by every model scored on the same clips. A frozen checkpoint is uninformative if its probe reaches less than 0.65 balanced accuracy on ordered clips, and a fine-tuned arm if its models do on average.

Loading the weights one thinks one loaded. All VideoMAE and TimeSformer models are loaded with `transformers` 5.0.0, and every job asserts that the loaded attention biases match the checkpoint file: releases 5.6.0 to 5.17.0 silently load VideoMAE with its query and value biases at zero, which inflates the Kinetics checkpoint's frozen A 2.0- to 4.3-fold (Appendix F).

4 In VideoMAE, fine-tuning collapses most of the frozen difference at the output

4.1 Frozen, the pretraining stages differ

Table 1 reports order sensitivity at the output for the three frozen checkpoints. On OF-Syn the ordering is SSv2 $>$ MAE-only $>$ Kinetics, with every adjacent pair's intervals disjoint, on the shared and on native input; native input raises the Kinetics value 2.7-fold, MAE-only's 1.3-fold and SSv2's 1.1-fold, so we report no ratio between checkpoints as a finding. On OOPS the ordering is the same; the marginal intervals of MAE-only and Kinetics meet (0.095 against 0.095), but their paired difference, 0.071 [0.024, 0.116], excludes zero.

In the informative step, from MAE-only to Kinetics, the single supervised stage lowers order sensitivity at the output, by 69% on the shared pipeline (0.137 to 0.043) and by 36% on native input (0.180 to 0.116), while the frozen features become *more* accurate on ordered clips in all three settings (point estimates; Table 1). A frozen probe therefore ranks the more accurate checkpoint as the less order-sensitive one, as one would expect of supervision on videos VideoMAE's authors describe as "mostly stationary and scene-related" (Tong et al., 2022): the kind of evidence from which a frozen comparison would conclude that Kinetics supervision discards temporal information. On a task defined by frame order the frozen Kinetics output is not distinguishable from MAE-only's (Section 4.2), so the drop does not show that the information is gone.

4.2 AFTER FINE-TUNING, MOST OF THE DIFFERENCE IS GONE

The test. Before any of their results existed we fixed a two-arm test on OF-Syn: fine-tune MAE-only and Kinetics with the same recipe and seeds, and estimate the fraction of their frozen gap that fine-tuning closes,

$$\kappa \;=\; 1 - \frac{A_{\mathrm{ft}}^{\mathrm{MAE}} - A_{\mathrm{ft}}^{\mathrm{Kin}}}{A_{\mathrm{pre}}^{\mathrm{MAE}} - A_{\mathrm{pre}}^{\mathrm{Kin}}}, \tag{2}$$

where $\kappa = 0$ means the gap is carried through fine-tuning unchanged and $\kappa = 1$ that it closes completely. We fixed in advance how its interval would be read: excluding 0 and containing 1, the gap closes completely within error; inside (0, 1), partly; containing 0 but not 1, it does not close; containing both, the test is uninformative; lying above 0, the gap does not survive. An interval above 1 therefore means that the fine-tuned order is the reverse of the frozen one, a reading we added after two such results. Frozen, the two checkpoints differ by 0.103 [0.071, 0.135] at the classifier's input. Fine-tuned, MAE-only reads 0.115 and Kinetics 0.123 (Figure 1, left): MAE-only falls by 0.036 [0.003, 0.070] and Kinetics rises by 0.075 [0.044, 0.107]. The five-seed test gives $\kappa = 1.07$ [0.76, 1.39], and its ten-seed repeat on a single GPU type (Appendix A) gives $\kappa = 1.08$ [0.84, 1.33]. Both intervals exclude 0 and contain 1, so by our rule the frozen gap closes completely within error; by the ten-seed interval, at least 84% of it closes. The bare read-out agrees: over all ten seeds the same ratio is 1.06 [0.79, 1.35] (Appendix B).

A third checkpoint and a second dataset. At the classifier's input, SSv2, the most order-sensitive checkpoint, falls from 0.290 to 0.151 under the same recipe, most of the way to the other two. On OOPS, at the primary, step-matched budget, MAE-only rises from 0.141 to 0.197 and Kinetics from 0.052 to 0.170, while SSv2 stays where it was (0.230 to 0.226), so the fine-tuned values keep their frozen order. On OOPS κ is not informative: 0.70 [-0.41, 1.34], covering both 0 and 1, as the Kinetics arm's seeds vary nearly three times as much as on OF-Syn (standard deviation 0.072), all of it from the seed that failed to fit (Appendix D). Over all three checkpoints, the between-checkpoint spread at the bare read-out falls by 90% on OF-Syn and by 69% on OOPS (Table 2; the OOPS Kinetics arm keeps its failed seed, as our rule requires, and without it the reduction is 90%).

A standard fine-tuning schedule. The recipe above, a constant rate with no warmup or layer-wise decay, updates the lower blocks more strongly than a layer-decayed schedule, so the collapse could be the recipe overwriting the backbone. We re-ran the pair with the standard schedule (three seeds each), with $\kappa \geq 0.5$ on the point estimate as the test. It holds: $\kappa = 0.85$, with a wide interval, [0.22, 1.36], that excludes 0. The pair's gap falls from 0.103 to 0.016 at the classifier's input (Figure 1, centre) and from 0.094 to -0.006 at the bare read-out, where the same ratio is 1.06. The standard schedule also trains much better models: end-to-end balanced accuracy 0.898 and 0.916, against 0.794 and 0.817 under the constant rate, which is no better than a linear probe on the frozen features (Table 1), so the collapse is not a symptom of a poorly fitted network. The two checkpoints meet near 0.17 at the classifier's input (near 0.12 under the constant rate), so there is no single fine-tuned value for them to converge to.

Checks of the instrument and of the scope. On the shared 112-pixel pipeline the frozen MAE-only/Kinetics gap at the bare read-out is larger than on native input (0.094 against 0.064), so we re-ran the test on native-resolution input, three seeds per arm (on those clips the frozen gap at the classifier's input is again 0.094; Table 3): the gap does not survive, and the refitted probe finds it reversed, $\kappa = 1.51$ [1.16, 2.07], with fine-tuned Kinetics above MAE-only (0.151 against 0.103), though the models' own heads do not show the reversal. On the ten-seed pair (Appendix B), four fresh permutations give κ between 1.03 and 1.11; a tubelet-preserving shuffle of whole frame pairs gives 0.76 [0.18, 1.13]; an MLP and an attentive probe give 1.08 and 1.06; and a linear probe cross-fitted on held-out evaluation groups, which never sees a clip the backbone was trained on, gives 1.94 [1.30, 4.64], again above 1. Read from the fine-tuned models' own heads instead of a refitted probe, κ is 1.00 [0.72, 1.31].

Other architectures and an order-defined task. We ran the same test for three further architectures and on the order-defined up/down task (Appendix C). In TimeSformer-B the frozen gap between the Kinetics and SSv2 checkpoints survives fine-tuning on OF-Syn, $\kappa = 0.08$ [-0.75,

Table 2: Fine-tuning collapses most of the between-checkpoint difference at the output. κ (Eq. 2) at the classifier's input, interval over groups and seeds. Spread: maximum minus minimum of $A(12)$ over the checkpoints (bare read-out, seeds 0 to 2); the interval on its reduction resamples groups only.

	OF-Syn		OOPS
	constant rate	standard schedule	constant rate
checkpoints in the spread	3	2	3
κ	1.08 [0.84, 1.33][a]	0.85 [0.22, 1.36]	0.70 [−0.41, 1.34]
spread, frozen	0.237	0.094	0.170
spread, fine-tuned	0.023	0.006	0.052
reduction	0.214 [0.172, 0.248]	0.088 [0.045, 0.122]	0.118 [0.063, 0.171]

[a] Ten seeds per arm; the five-seed test gives 1.07 [0.76, 1.39].

0.60], so at most 60% of it closes, leaving a gap of 0.045 [0.021, 0.072]. Under the standard schedule, although both arms' models beat their frozen probes, it does not close either: $\kappa = 0.43$ [−0.03, 0.73]. On up/down the same pair starts further apart (0.231 [0.193, 0.270]) and fine-tuning closes the gap completely within error, $\kappa = 0.86$ [0.69, 1.01]; there both fine-tuned models end near chance on shuffled clips, so A largely tracks accuracy (Appendix C). In R(2+1)D-34 the frozen checkpoints are not distinguishable on either task, so the test is not defined. Hiera-B's pair, shaped like VideoMAE's, has a frozen gap of 0.140 [0.101, 0.180]; its models did not fit the task under the constant rate, but under the standard schedule they do (end-to-end balanced accuracy 0.91 and 0.92), and the gap does not close: $\kappa = 0.03$ [−0.54, 0.45]. Unlike TimeSformer's, Hiera's gap also survives fine-tuning on up/down ($\kappa = 0.25$ [−0.24, 0.60]), although read from the models' own heads it closes (0.62 [0.12, 1.04]).

Use and direction information. Neither fall dataset strictly needs frame order, so A on them measures how much a fall read-out uses order, not how much direction information the output carries; on up/down, where a shuffled clip has no direction, a probe's gain over its shuffled accuracy measures the direction information it can read. Frozen, the MAE-only and Kinetics outputs are not distinguishable there ($A = 0.325$ and 0.303 at the classifier's input, gap 0.022 [−0.033, 0.078]), whereas on the fall task, at the same read-out, Kinetics is far less order-sensitive (0.048 against 0.151, gap 0.103 [0.071, 0.135]), so the fall-task drop does not show that Kinetics supervision removes direction information. After the TimeSformer results we tested whether fall fine-tuning changes use rather than direction information by probing the fall-fine-tuned backbones on up/down (Table 5). VideoMAE's keep most of it: MAE-only and Kinetics keep 0.23 to 0.29 under both recipes, while their fall-task values converge. In TimeSformer the frozen Kinetics output carries none ($A = -0.001$); after fine-tuning for fall detection it carries a little under the constant rate ($A = 0.056$) and none under the standard schedule, while SSv2 keeps its own (0.217 and 0.263): most of the gap in direction information persists (0.162 and 0.263). What fine-tuning removes between MAE-only and Kinetics is thus how the fall read-out uses order; what persists in TimeSformer comes with a surviving gap in the direction information itself, and the models' own heads also keep the fall-task gap ($\kappa = 0.08$ and 0.01; Appendix B). We then used the reading to predict Hiera's result before any of its standard-schedule models existed. Frozen on up/down, Hiera's Kinetics output carries less direction information than its parent's (0.253 against 0.403, a gap of 0.150 [0.101, 0.198], whose lower bound clears our threshold of 0.10 by 0.001), so the reading predicted that the fall-task gap would not close, and it did not ($\kappa = 0.03$; from the models' own heads 0.09 [−0.34, 0.38]); as in TimeSformer, the direction gap itself survives fine-tuning for fall detection (0.133). Within VideoMAE, though, the frozen direction gap does not predict what happens: its SSv2 checkpoint differs in direction information from Kinetics by 0.119 [0.063, 0.172] and from MAE-only by 0.097 [0.037, 0.157] (cases our rule left open), yet both fall-task gaps close at least mostly ($\kappa = 0.90$ [0.73, 1.08] and 0.77 [0.56, 0.96]) while direction gaps of 0.045 [0.001, 0.092] and 0.060 [0.014, 0.107] remain; under the standard schedule MAE-only and Kinetics even end 0.064 [0.007, 0.123] apart in direction information while their fall-task gap closes. In TimeSformer and Hiera, then, a surviving direction difference accompanied a surviving fall-task gap, whereas VideoMAE's fine-tuning closed its fall-task gaps either way (an observation made after the results, and confounded with architecture).

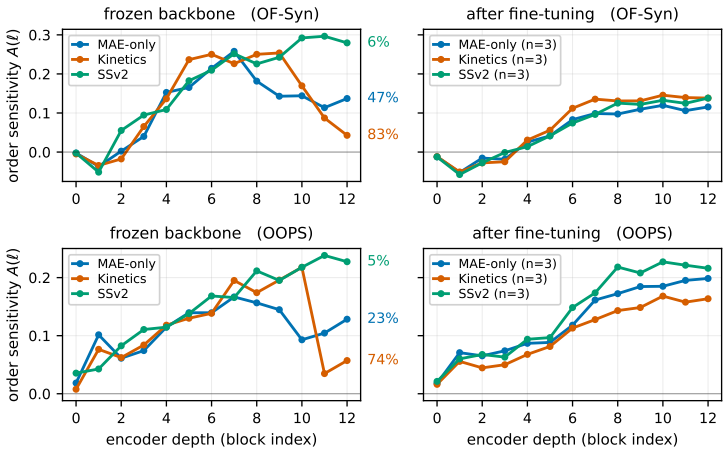


Figure 2: Order sensitivity $A(\ell)$ at every block ℓ (bare mean pool; block 0 is the tubelet embedding; a negative value, as at block 1, means shuffling helps the probe), frozen (left) and after constant-rate fine-tuning (right; mean of seeds 0 to 2), on OF-Syn (top) and OOPS (bottom); percentages give each frozen checkpoint's drop from its maximum to the last block (the bare read-out, not Figure 1's classifier input).

5 Where the difference lives: the upper layers

We probe every block (Figure 2) and ask only what separates one network's pretraining stages and what fine-tuning does to that separation; that probed information peaks inside a network is established (Section 2). These analyses are exploratory.

Where the frozen checkpoints differ. Frozen, the three checkpoints are close through the lower two thirds and separate in the upper third: their spread at blocks 11 and 12 lies above its upper bound at every block from 0 to 8 on OF-Syn and from 0 to 9 on OOPS (Table 6). After constant-rate fine-tuning no block's spread has an upper bound above 0.060 (OF-Syn) or 0.102 (OOPS). An intervention does not confirm these contrasts: transplanting the top three blocks between the frozen checkpoints moves 42% of the gap one way and 207% the other.

Kinetics removes order sensitivity from the output, not at its peak. For the MAE-only and Kinetics pair the ordering depends on depth. At blocks 8 and 9 of OF-Syn, Kinetics is *more* order-sensitive than MAE-only (0.254 [0.224, 0.281] against 0.143 [0.116, 0.170] at block 9); at the last block the ordering is reversed (0.043 against 0.137), both with disjoint intervals. OOPS shows the same inversion between blocks 10 and 11. Kinetics supervision therefore leaves the network's peak order sensitivity about where its parent's is (0.254 against 0.258), while removing most of that sensitivity from the output, as the fall probe reads it. We measure this by the drop from each checkpoint's maximum $A(\ell)$ to its last block, as a percentage of the maximum, rather than by the peak's depth, which studies often report (Joseph et al., 2026; Musa et al., 2026) but which is unstable here: the full-sample peak block is the maximum in as few as 40% of resamples for OF-Syn Kinetics, whose drop, re-selecting the maximum per resample, is still tightly estimated (Appendix E). On OF-Syn the drop is 83% [74, 93] for Kinetics, against 47% [37, 56] for MAE-only and 6% [0, 14] for SSv2; on OOPS it is 74% [57, 90] against 23% [4, 45] and 5% [0, 19]; for the VideoMAE-large pair it is 68% [58, 77] against 45% [32, 57] (Figure 4). Kinetics loses more than MAE-only with disjoint intervals on both datasets and at both scales, while its maximum is never below MAE-only's beyond noise (Figure 4). This fits what each supervision rewards: Something-Something's classes describe what is done to an unspecified object (Goyal et al., 2017), and its checkpoint (whose self-supervised run also differs) stays order-sensitive at the output, while Kinetics is largely recognisable from appearance (Tong et al., 2022; Bertasius et al., 2021), and its checkpoint does not.

Agreement is made at the top. A deflationary reading is that fine-tuning simply overwrites the backbone, which would make the collapse trivial and predict agreement at every depth. Under the

constant-rate recipe on OF-Syn the whole profile does move: mid-network order sensitivity (mean over blocks 5 to 9) falls by 0.106 for MAE-only and 0.130 for Kinetics, and the spread between checkpoints stays below 0.04 at every depth, while on OOPS the same recipe moves the mid-network level by at most 0.044. The standard schedule separates the moving profile from the agreement at the output. It lowers mid-network order sensitivity by less than half as much, or raises it (block 5 to 9 averages change by $+0.059$ and -0.014); every standard-schedule model keeps a peak inside the network, losing 13% to 39% of it by the last block; and the two checkpoints still differ at blocks 9 and 10. At block 9 under this schedule, MAE-only exceeds Kinetics in every seed, by 0.106 [0.081, 0.133] on average and in the reverse of their frozen ordering at that depth, and at block 10 they differ by 0.066 [0.042, 0.090]; they agree only in the last two blocks. So the output difference can be removed without making the middle agree: the checkpoints differed mainly in their upper layers when frozen, and fine-tuning brings them together in the last two, compatible with Merchant et al. (2020)'s finding that fine-tuning leaves most probed information decodable, though under the constant rate it changes the whole profile.

6 What a frozen ranking measures

A frozen probe ranks VideoMAE's pretraining stages by frame-order sensitivity, and the ranking is real: it holds on both datasets and on native input. But it describes the top of a pretrained network, where the MAE-only/Kinetics ordering inverts with depth, its magnitudes change with preprocessing, and fine-tuning removes 69% to 94% of it. It also describes use more than information (a reading formed after the results, then tested; Section 4.2): between MAE-only and Kinetics, fine-tuning removed the fall read-out's use of order while both checkpoints kept most of their direction information. This echoes the distinction between what a probe can decode and what a model uses (Ravichander et al., 2021; Elazar et al., 2021), here across fine-tuning, with the caveat that our "use" is how a fall-task linear read-out depends on order, which the fine-tuned models' own heads confirm (Appendix B). A frozen fall-task probe therefore cannot tell whether a difference will survive: it closes in VideoMAE, survives in TimeSformer and Hiera, and in R(2+1)D-34 fine-tuning separates checkpoints that were not distinguishable frozen (a descriptive, three-seed contrast; Appendix C). For VideoMAE's checkpoints, a frozen comparison of pretraining objectives is evidence about frozen use and little evidence about what the objectives contribute once the network is fine-tuned, as such checkpoints usually are.

Four practices follow. Report a probed property across depth, as the drop from each model's own maximum, not only at the output or as a peak location. Check a ranking between checkpoints in the regime in which it will be used: a frozen ranking licenses a choice for frozen use. Where the probing task does not need the property, also probe one that does (Section 4.2). And record the preprocessing and library version, and assert a checkpoint-file statistic at load: both move frozen values, preprocessing by checkpoint-specific amounts.

Limitations. The collapse is established for VideoMAE on the fall task and for TimeSformer-B on one order-defined task: TimeSformer's and Hiera's fall-task gaps survive (Hiera's also on the order-defined task, as the probe reads it), R(2+1)D-34's frozen checkpoints are not distinguishable, and our VideoMAE-large fine-tuning arm and an arm on time-reversed Something-Something action pairs are uninformative. TimeSformer also differs from VideoMAE in its checkpoints, read-out and frames (Appendix C), and why VideoMAE's fine-tuning closes fall-task gaps that TimeSformer's and Hiera's keep we do not identify; architecture is confounded with that outcome. The use/direction reading was formed after the first results and tested out of sample on one architecture; "direction information" means what a linear probe reads at the classifier input on one relabelled task. Both main tasks are binary fall classification, and κ is informative on one fall dataset and two order-defined pairs. We fine-tune every weight at one clip length and stride; regularised (Li et al., 2018) or parameter-efficient fine-tuning, which move the weights less, may keep more of a frozen difference. A comes from a linear probe at a fixed fitting size; other probes change its values, though not the conclusion that VideoMAE's frozen gap does not survive. Intervals of three-seed arms resample only three seeds, and we make no correction for multiple tests. The SSv2 checkpoint comes from a separate self-supervised run, all analyses but the native-input test share a 112-pixel input pipeline, and we make no claim about fine-tuning from random initialisation.

Reproducibility statement

All checkpoints and datasets are public (Section 3). Appendix A gives the input pipelines, probe, bootstrap, fine-tuning recipes, splits, library version, load check and compute. The results rest on 164 fine-tuned models, trained and probed in 159 GPU jobs (110 GPU-hours, about half of them on NVIDIA H200s). Code and per-clip predictions will be released with the paper.

Ethics statement

We analyse public research datasets and collect no new data. OF-Syn is synthetic; OOPS consists of publicly posted videos of unintentional failures (Epstein et al., 2020), which we use through OmniFall's annotations (Schneider et al., 2026) without attempting to identify anyone in them. We train no model for deployment.

AI use statement

A large language model (Claude, Anthropic) was used to help propose and refine hypotheses, design experiments, implement the analysis code, dataset splits and cluster job scripts, interpret results, find related work, and aid and polish the writing; it was not used to generate data, to formulate or prove mathematical claims, or for translation. All AI-assisted work was checked by means that do not depend on a language model: every result number is traced to a stored result and the cluster job that produced it, and every reference was taken from the publisher's or arXiv's own record. The authors take full responsibility for the content of this submission.

References

Piyush Bagad and Andrew Zisserman. Chirality in action: Time-aware video representation learning by latent straightening. In *Advances in Neural Information Processing Systems*, 2025.

Yutong Bai, Haoqi Fan, Ishan Misra, Ganesh Venkatesh, Yongyi Lu, Yuyin Zhou, Qihang Yu, Vikas Chandra, and Alan Yuille. Can temporal information help with contrastive self-supervised learning? *arXiv preprint arXiv:2011.13046*, 2020.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219, 2022. doi: 10.1162/coli_a_00422.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pp. 813–824, 2021.

Jayadev Billa. Decomposing the depth profile of fine-tuning. arXiv preprint arXiv:2604.17177, 2026.

Daniel Bolya, Po-Yao Huang, Peize Sun, Jang Hyun Cho, Andrea Madotto, Chen Wei, et al. Perception Encoder: The best visual embeddings are not at the output of the network. In *Advances in Neural Information Processing Systems*, volume 38, pp. 60884–60937, 2025.

Laure Ciernik, Marco Morik, Lukas Thede, Luca Eyring, Shinichi Nakajima, Zeynep Akata, and Lukas Muttenthaler. Attentive multilayer fusion for vision transformers. arXiv preprint arXiv:2601.09322, 2026.

Shuangrui Ding, Maomao Li, Tianyu Yang, Rui Qian, Haohang Xu, Qingyi Chen, Jue Wang, and Hongkai Xiong. Motion-aware contrastive video representation learning via foreground-background merging. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9716–9726, 2022.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 2021. doi: 10.1162/tacl_a_00359.

Dave Epstein, Boyuan Chen, and Carl Vondrick. Oops! Predicting unintentional action in video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 916–926, 2020. doi: 10.1109/CVPR42600.2020.00100.

Deepti Ghadiyaram, Du Tran, and Dhruv Mahajan. Large-scale weakly-supervised pre-training for video action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12038–12047, 2019.

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

Kaiming He, Ross Girshick, and Piotr Dollár. Rethinking ImageNet pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4918–4927, 2019.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 16000–16009, 2022.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2733–2743, 2019. doi: 10.18653/v1/D19-1275.

Samyak Jain, Robert Kirk, Ekdeep Singh Lubana, Robert P. Dick, Hidenori Tanaka, Edward Grefenstette, Tim Rocktäschel, and David Scott Krueger. Mechanistically analyzing the effects of fine-tuning on procedurally defined tasks. In *International Conference on Learning Representations*, 2024.

Sonia Joseph, Quentin Garrido, Randall Balestriero, Matthew Kowal, Thomas Fel, Shahab Bakhtiari, Blake Richards, and Mike Rabbat. Interpreting physics in video world models. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, 2026.

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. arXiv preprint arXiv:1705.06950, 2017.

Simon Kornblith, Jonathon Shlens, and Quoc V. Le. Do better ImageNet models transfer better? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019.

Matthew Kowal, Mennatullah Siam, Md Amirul Islam, Neil D. B. Bruce, Richard P. Wildes, and Konstantinos G. Derpanis. A deeper dive into what deep spatiotemporal networks encode: Quantifying static vs. dynamic information. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13999–14009, 2022.

Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. In *International Conference on Learning Representations*, 2022.

Xuhong Li, Yves Grandvalet, and Franck Davoine. Explicit inductive bias for transfer learning with convolutional networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2825–2834, 2018.

Houman Mehrafarin, Sara Rajaee, and Mohammad Taher Pilehvar. On the importance of data size in probing fine-tuned models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 228–238, 2022.

Amil Merchant, Elahe Rahimtoroghi, Ellie Pavlick, and Ian Tenney. What happens to BERT embeddings during fine-tuning? In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pp. 33–44, 2020. doi: 10.18653/v1/2020.blackboxnlp-1.4.

Marius Mosbach, Anna Khokhlova, Michael A. Hedderich, and Dietrich Klakow. On the interplay between fine-tuning and sentence-level probing for linguistic knowledge in pre-trained transformers. In *Proceedings of the Third BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pp. 68–82, 2020. doi: 10.18653/v1/2020.blackboxnlp-1.7.

Sharon S. Musa, Fereshteh Forghani, Harrish Thasarathan, Sonia Joseph, Matthew Kowal, and Konstantinos G. Derpanis. What, where, and how: Probing spatiotemporal representations in video foundation models. arXiv preprint arXiv:2609.01551, 2026.

Alejandro Newell and Jia Deng. How useful is self-supervised pretraining for visual tasks? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

Behnam Neyshabur, Hanie Sedghi, and Chiyuan Zhang. What is being transferred in transfer learning? In *Advances in Neural Information Processing Systems*, volume 33, pp. 512–523. Curran Associates, Inc., 2020.

Ambroise Odonnat, Vasilii Feofanov, Laetitia Chapel, Romain Tavenard, and Ievgen Redko. Layer by layer, module by module: Choose both for optimal OOD probing of ViT. arXiv preprint arXiv:2603.05280, 2026.

Mandela Patrick, Dylan Campbell, Yuki M. Asano, Ishan Misra, Florian Metze, Christoph Feichtenhofer, Andrea Vedaldi, and João F. Henriques. Keeping your eye on the ball: Trajectory attention in video transformers. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

Thinesh Thiyakesan Ponbagavathi and Alina Roitberg. Order matters: On parameter-efficient image-to-video probing for recognizing nearly symmetric actions. *arXiv preprint arXiv:2503.24298*, 2025. Accepted to ICRA 2026.

Nikhil Prakash, Tamar Rott Shaham, Tal Haklay, Yonatan Belinkov, and David Bau. Fine-tuning enhances existing mechanisms: A case study on entity tracking. In *International Conference on Learning Representations*, 2024.

Samuele Punzo, Niccolò Caselli, Ippokratis Pantelidis, Francesco Massafra, Salvatore Lo Sardo, and Mohammadreza Salehi. How do video foundation models encode intuitive physics? Probing across pretraining paradigms. arXiv preprint arXiv:2606.09646, 2026.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. Probing the probing paradigm: Does probing accuracy entail task relevance? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 3363–3377, 2021. doi: 10.18653/v1/2021.eacl-main.295.

Chaitanya Ryali, Yuan-Ting Hu, Daniel Bolya, Chen Wei, Haoqi Fan, Po-Yao Huang, Vaibhav Aggarwal, Arkabandhu Chowdhury, Omid Poursaeed, Judy Hoffman, Jitendra Malik, Yanghao Li, and Christoph Feichtenhofer. Hiera: A hierarchical vision transformer without the bells-and-whistles. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 29441–29454, 2023.

David Schneider, Zdravko Marinov, Moritz Mistol, Zeyun Zhong, Alexander Jaus, Rodi Düger, Rafael Baur, M. Saquib Sarfraz, and Rainer Stiefelhagen. OmniFall: From staged through synthetic to wild, a unified multi-domain dataset for robust fall detection. In *Computer Vision – ECCV 2026*, Lecture Notes in Computer Science, pp. 112–129. Springer, 2026. doi: 10.1007/978-3-032-37023-5_7.

Laura Sevilla-Lara, Shengxin Zha, Zhicheng Yan, Vedanuj Goswami, Matt Feiszli, and Lorenzo Torresani. Only time can tell: Discovering temporal data for temporal modeling. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 535–544, 2021.

Shashank Shekhar, Florian Bordes, Pascal Vincent, and Ari Morcos. Objectives matter: Understanding the impact of self-supervised objectives on vision transformer representations. arXiv preprint arXiv:2304.13089, 2023.

Fida Mohammad Thoker, Hazel Doughty, Piyush Bagad, and Cees Snoek. How severe is benchmark-sensitivity in video self-supervised learning? In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2022.

Fida Mohammad Thoker, Letian Jiang, Chen Zhao, Piyush Bagad, Hazel Doughty, Bernard Ghanem, and Cees G. M. Snoek. SEVERE++: Evaluating benchmark sensitivity in generalization of video representation learning. arXiv preprint arXiv:2504.05706, 2025.

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*, volume 35, pp. 10078–10093. Curran Associates, Inc., 2022. doi: 10.52202/068431-0732.

Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6450–6459, 2018.

Arnas Uselis and Seong Joon Oh. Intermediate layer classifiers for OOD generalization. In *International Conference on Learning Representations*, pp. 46059–46088, 2025.

Liangzhe Yuan, Nitesh Bharadwaj Gundavarapu, Long Zhao, Hao Zhou, Yin Cui, Lu Jiang, Xuan Yang, Menglin Jia, Tobias Weyand, Luke Friedman, Mikhail Sirotenko, Huisheng Wang, Florian Schroff, Hartwig Adam, Ming-Hsuan Yang, Ting Liu, and Boqing Gong. VideoGLUE: Video general understanding evaluation of foundation models. *Transactions on Machine Learning Research*, 2024.

Sukmin Yun, Jaehyung Kim, Dongyoon Han, Hwanjun Song, Jung-Woo Ha, and Jinwoo Shin. Time is MattEr: Temporal self-supervision for video transformers. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 25804–25816. PMLR, 2022.

Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, Andre Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, Lucas Beyer, Olivier Bachem, Michael Tschannen, Marcin Michalski, Olivier Bousquet, Sylvain Gelly, and Neil Houlsby. A large-scale study of representation learning with the visual task adaptation benchmark. arXiv preprint arXiv:1910.04867, 2019.

Qinghua Zhao, Xueling Gong, Xinyu Chen, Zhongfeng Kang, and Xinlu Li. A layer-wise analysis of supervised fine-tuning. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9981–9992. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.453.

A Experimental details

Figure 3 summarises the setup.

Checkpoints. From the `MCG-NJU` organisation on the Hugging Face Hub: `videomae-base`, `videomae-base-finetuned-kinetics`, `videomae-base-finetuned-ssv2`, `videomae-large` and `videomae-large-finetuned-kinetics`, loaded with `transformers` 5.0.0 and checked at load time (Appendix F). The SSv2 checkpoint is distributed only as a PyTorch pickle; we converted it to `safetensors` and verified the conversion tensor by tensor. For Appendix C: `facebook/timesformer-base-finetuned-k400` and `-ssv2`, converted the same way (`transformers` 5.0.0 refuses their pickles under our PyTorch) and checked by the summed magnitude of every tensor against the file; and the 32-frame R(2+1)D-34 files of `ig65m-pytorch` release 1.0.0 (`r2plus1d_34_clip32_ig65m_from_scratch` and `r2plus1d_34_clip32_ft_kinetics_from_ig65m`), loaded strictly with their hashes checked and fed our 16-frame clips; and, for Hiera-B, the 16-frame 224-pixel files of its official release (`mae_hiera_base_16x224` and `hiera_base_16x224`), loaded into its reference implementation. Every VideoMAE model uses the architecture of VideoMAE's fine-tuned checkpoints, which mean-pools the last block and normalises the pooled vector before the classifier. The

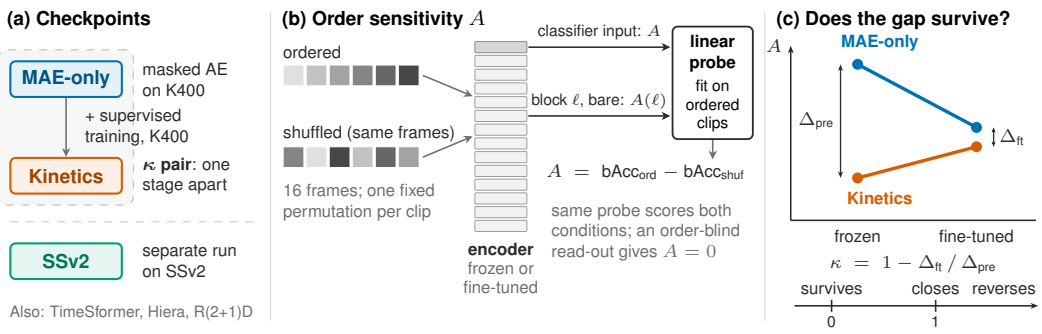


Figure 3: Setup. (a) Three published VideoMAE ViT-B checkpoints: Kinetics is MAE-only after one supervised stage, the pair behind κ; SSv2 comes from a separate run on Something-Something v2. The same test is applied to checkpoint pairs of three further architectures (Section 4.2). (b) Order sensitivity A: a linear probe fitted on ordered training clips is scored on each test clip as presented and with its 16 frames shuffled (one fixed permutation per clip, applied to the pixels); A is read at the classifier's input or, as $A(\ell)$, at the mean-pooled output of block ℓ without normalisation. (c) κ: the fraction of the frozen MAE-only/Kinetics gap that fine-tuning with the same recipe and seeds closes (values drawn are schematic).

MAE-only checkpoint was never trained with a classifier: its encoder's final normalisation of the tokens has no place in that architecture and is not loaded, and the normalisation of the pooled vector is freshly initialised. With weight 1 and bias 0 it only standardises each pooled vector, and that is why the MAE-only classifier-input A_{pre} (0.151) is higher than its bare mean-pool value (0.137, Table 1). The bare read-out has no such layer, and the same ratio computed on it agrees with κ (Section 4.2).

Input. Clips are 16 frames. The shared pipeline stores frames at 112×112, the input size of the 3D-convolutional backbones for which the clips were first extracted, normalises them with Kinetics-400 channel statistics and resizes them bilinearly to 224×224 before the backbone; it is used for every fine-tuning and depth result except the native-input arm. The native pipeline decodes the raw video and applies each checkpoint's own VideoMAE processor; it is used for the native column of Table 1. The native-input fine-tuning arm re-encodes the same clips from the raw video with the processor's geometry (shortest edge 224, centre crop) and normalisation statistics, so no 112-pixel stage is involved.

Probe and shuffle. Logistic regression (scikit-learn, L-BFGS, $C = 1$, at most 1000 iterations) on standardised features, fitted on the training split's ordered clips and evaluated on the evaluation split's ordered and shuffled clips. Each evaluation clip receives one independent permutation of its 16 frames (fixed seed), stored and reused for every model. Because every clip has its own permutation, resampling evaluation groups also resamples the permutations, so their randomness enters every interval through the clips they belong to. Over the persisted permutation and four fresh draws, a single model's A has a standard deviation of up to 0.016, more than the 0.010 we had predicted, while a cell's mean over seeds has one of at most 0.007, and κ stays between 1.03 and 1.11 (Appendix B).

Intervals. Percentile bootstrap intervals. Layer-wise values, spreads and drops use 2,000 resamples of evaluation groups, with the maximum re-selected inside each resample for the drop. Fine-tuned values and κ use 4,000 two-level resamples (groups, then seeds). One group resample is shared by every model scored on the same clips.

GPU types. Re-extracting features on a different GPU type moves A by about 0.004, so every contrast was measured on one GPU type (NVIDIA H200, or H100 for the TimeSformer and Hiera pairs of Appendix C and for the up/down probes of fall-fine-tuned backbones), with two exceptions: the two arms of the five-seed κ test were measured on an H200 and an A100, which κ tolerates because each arm's change is measured within one job (its denominator can shift by about 4%); and

in one descriptive pair with no κ (R(2+1)D-34 on the temporal mirrors) the two arms were measured on different types.

Fine-tuning. Cross-entropy, a linear head on the normalised mean-pooled features, AdamW, batch size 8, weight decay 10^{-4}, no augmentation. Constant rate: learning rate 10^{-4} for 10 epochs on OF-Syn and 80 on OOPS, about 19,200 and 19,700 optimisation steps; the epoch-matched OOPS budget is 10 epochs. Standard schedule: peak learning rate 10^{-4} for the head, each block's rate 0.75 times that of the block above it, one epoch of linear warmup and cosine decay to zero, 10 epochs. The native-input and VideoMAE-large arms use the constant rate for 10 epochs on OF-Syn; the temporal-mirror arm uses it for 8 epochs, which matches OF-Syn's number of optimisation steps.

Compute. The results rest on 164 fine-tuned models: on OF-Syn under the constant rate, MAE-only and Kinetics with 10 seeds each and SSv2 with 5; under the standard schedule, MAE-only and Kinetics with 3 each; on OOPS, all three checkpoints with 5 seeds each at the step-matched budget and MAE-only and Kinetics with 5 each at the epoch-matched one; for the further arms of Appendix B, MAE-only and Kinetics with 3 seeds each on native input, at the large scale and on the temporal-mirror task; for Appendix C, 21 arms of 3 seeds each; two 20-epoch trajectory runs; and 25 models for random-initialisation and 3D-convolutional comparisons not reported here. Every model was probed at its output, and the layer-wise sweeps probe every block of each base checkpoint, frozen and after fine-tuning (seeds 0 to 2), and of the frozen large pair; the checks of the instrument, the up/down probes of fall-fine-tuned backbones, the temporal-attention knockout and the block transplants re-probe existing models and train none. The runs for this paper comprised 159 GPU jobs on a university cluster, 109.6 GPU-hours in total: 55.2 on NVIDIA H200s (73 jobs), 26.3 on H100s (52 jobs), 26.1 on A100s (27 jobs), 1.9 on V100s (2 jobs) and 0.1 on one faulty H200 NVL (5 jobs that failed or were cancelled at start). Every job's outcome was confirmed from its return code.

Splits. OF-Syn uses OmniFall's split. OOPS is split by source video with 30% of videos held out (fixed seed; 399 and 173 videos), verified group-disjoint, with both classes on both sides. The temporal-mirror task uses Something-Something v2's own training and validation splits (19,084 and 2,990 videos), each video its own group.

B Checks of the instrument and of the scope

The checks below all ran on NVIDIA H200s; the last paragraph adds a re-reading of stored predictions. The checks of the instrument re-probe the three frozen checkpoints and the fine-tuned models of Table 2 (OF-Syn under both recipes, OOPS at the step-matched budget) under new test conditions, from one fp32 forward pass per clip; the first three groups of Table 3 use the ten-seed pair. This re-extraction reproduces every model's committed classifier-input A within 0.002 and the bare values of Table 1 within 0.004, so values here can differ from the body's in the last digit. Training clips are always ordered. The perturbations are four fresh per-clip permutations; a pair shuffle, which permutes the eight consecutive frame pairs of a clip and keeps the order inside each pair, so that every tubelet is one of the ordered clip's own tubelets in a new position; frame reversal; and pair reversal. The read-outs are the classifier input with the linear probe; the bare mean pool with the linear probe and with an MLP probe; the last block's tokens averaged over space at each of the eight tubelet time steps, read by an attentive probe (one learned query, eight heads, AdamW with learning rate 10^{-3} for 30 epochs) with no position embedding on the probe side, so that like the mean pool it is a symmetric function of its inputs and a nonzero A must still come from the encoder; and a linear probe cross-fitted on the evaluation split, fitted on the ordered clips of half the evaluation groups and scored on the other half, in both directions, so that no probe sees a clip its backbone was trained on. κ uses the estimator of Section 4.2, and is computed only when the interval on the frozen gap excludes 0 and both frozen probes reach 0.65.

On every re-drawn permutation, under the pair shuffle, and with every read-out, the interval on κ excludes 0. It also excludes 0 under frame reversal, and it covers 0 and 1 under pair reversal, the perturbation under which the frozen gap is smallest (0.041 [0.010, 0.072]). Under the pair shuffle every frozen checkpoint's A falls, on both datasets, and the frozen ordering SSv2 $>$ MAE-only $>$ Kinetics holds on both; on OF-Syn the bare values fall from 0.134 to 0.065 for MAE-only, from

Table 3: κ under the checks. Frozen gap: the denominator of Eq. 2, $A^{\mathrm{MAE}}_{\mathrm{pre}} - A^{\mathrm{Kin}}_{\mathrm{pre}}$, on each row's read-out. The first three groups re-probe the ten-seed OF-Syn pair; the last group is new fine-tuning, three seeds per arm. Intervals are 95% and resample groups and then seeds.

check	frozen gap	κ
Persisted permutation, re-extracted	0.102 [0.069, 0.134]	1.08 [0.85, 1.33]
Re-drawn permutation 1	0.106 [0.072, 0.138]	1.03 [0.81, 1.24]
Re-drawn permutation 2	0.079 [0.046, 0.113]	1.08 [0.78, 1.39]
Re-drawn permutation 3	0.090 [0.057, 0.123]	1.10 [0.82, 1.43]
Re-drawn permutation 4	0.094 [0.061, 0.127]	1.11 [0.89, 1.34]
Pair shuffle (tubelets kept)	0.047 [0.021, 0.073]	0.76 [0.18, 1.13]
Frame reversal	0.114 [0.083, 0.144]	0.97 [0.78, 1.16]
Pair reversal	0.041 [0.010, 0.072]	0.56 [−0.82, 1.03]
Bare mean pool	0.091 [0.057, 0.125]	1.06 [0.79, 1.35]
MLP probe, bare mean pool	0.051 [0.016, 0.088]	1.08 [0.51, 1.73]
Attentive probe, 8 time tokens	0.114 [0.074, 0.156]	1.06 [0.84, 1.27]
Probe cross-fitted on the evaluation split	0.049 [0.010, 0.088]	1.94 [1.30, 4.64]
Native-resolution input (3 seeds)	0.094 [0.056, 0.132]	1.51 [1.16, 2.07]
VideoMAE-large (3 seeds)	0.052 [0.020, 0.084]	1.10 [−1.83, 3.58]
VideoMAE-large, without the seed below 0.65	0.052 [0.020, 0.084]	0.38 [−2.21, 2.06]
Something-Something temporal mirrors (3 seeds)	0.054 [0.029, 0.081]	not computed: a frozen probe at 0.59

0.043 to 0.019 for Kinetics and from 0.277 to 0.204 for SSv2. Shuffling barely affects the tubelet embedding itself (block 0: $|A(0)| \leq 0.005$ on OF-Syn and at most 0.035 on OOPS), and since the pair shuffle keeps the ordering and the fine-tuning result, unfamiliar tubelets are not needed to produce the effect. A value above 1 means that fine-tuning reverses the frozen order: with the cross-fitted probe, fine-tuned Kinetics exceeds MAE-only (0.111 against 0.065), as it does on native input (0.151 against 0.103). The cross-fitted probe is fitted on half the evaluation clips, so its values are comparable only with each other. Under the pair shuffle the between-checkpoint spread at the bare read-out falls from 0.185 to 0.004 on OF-Syn, a reduction of 0.181 [0.138, 0.207], and from 0.141 to 0.042 on OOPS; under each of the eight test conditions, on both datasets, the interval on the reduction excludes 0. The bare-read-out row, over all ten seeds, contains the three-seed value of the layer-wise sweeps (1.24) within its interval.

Two descriptive predictions failed. We predicted that each model's A would have a standard deviation of at most 0.010 over the five permutation draws; single models reach 0.016, although the mean over a cell's seeds has a standard deviation of at most 0.007 and κ stays between 1.03 and 1.11. And we predicted that reversal would lower frozen A less than shuffling; that holds for SSv2, but MAE-only and Kinetics have *higher* A under reversal (0.152 against 0.134, and 0.047 against 0.043, bare mean pool).

Of the new arms, native input is informative: on its clips (Appendix A) the frozen probes reach 0.80 and 0.84 balanced accuracy, and every fine-tuned model at least 0.76 end to end. VideoMAE-large is not: one MAE-only seed does not fit the task (end-to-end balanced accuracy 0.61, below 0.65), the arm's mean clears our floor, and the three-seed intervals cover 0 and 1 with and without that seed; fine-tuning lowers A in both large arms. The temporal-mirror task uses five Something-Something templates whose members are each other's time reversal (opening and closing something, folding and unfolding it, covering and uncovering it, putting something into something and taking it out, and moving something up and down); the label is which member of its pair a video shows, so the task cannot be solved without frame order. The SSv2 checkpoint is left out, since it was trained on these labels. The frozen Kinetics probe reaches only 0.59, so by our rule κ is not computed. Fine-tuning raises A in both arms (0.150 to 0.230 for MAE-only, 0.096 to 0.162 for Kinetics); the arms start 0.054 apart and end 0.067 apart, from which, by the same rule, we draw no conclusion.

The fine-tuned models' own heads. Order sensitivity after fine-tuning is read by a linear probe refitted on each fine-tuned backbone, so that it is estimated the same way as the frozen value. Each fine-tuned model's own classification head was also scored in the same job, on the same ordered and shuffled clips. Using it in place of the refitted probe in the estimator of Section 4.2, with the frozen

side unchanged (a frozen MAE-only checkpoint has no head of its own), gives $\kappa = 1.00$ [0.72, 1.31] for the ten-seed pair, 1.02 [0.43, 1.51] under the standard schedule and 1.16 [0.82, 1.55] on native input: in VideoMAE the gap still closes completely within error, although on native input the fine-tuned order no longer reverses (fine-tuned MAE-only minus Kinetics -0.015 [-0.045, 0.015]). For TimeSformer on OF-Syn it gives 0.08 [-0.92, 0.74] under the constant rate and 0.01 [-0.65, 0.40] under the standard schedule, where the models' own heads keep a gap of 0.049 [0.031, 0.067], the size of the frozen one; on up/down it gives 0.79 [0.63, 0.94], an interval between 0 and 1, so read this way that gap closes partly rather than completely within error. For Hiera it gives 0.09 [-0.34, 0.38] on OF-Syn, agreeing with the probe, but 0.62 [0.12, 1.04] on up/down, against the probe's 0.25 [-0.24, 0.60]: there the models' own heads leave a gap of 0.057 [-0.005, 0.119], and read this way the gap closes completely within error, the one cell in which the two read-outs disagree on whether the gap closes at all.

C Other architectures and an order-defined task

The body's main results use one architecture, and neither fall dataset strictly needs frame order. We ran these tests with the constant-rate recipe of Section 3 unless noted, three seeds per arm, the read-out and decision rule of Section 4.2, and both arms of every pair measured on one GPU type (NVIDIA H200 unless noted), apart from the R(2+1)D-34 pair on the temporal mirrors, whose Kinetics arm was measured on an H100.

Further pairs of published checkpoints. Each pair is two published checkpoints of one architecture. R(2+1)D-34 (Tran et al., 2018) is a convolutional network whose blocks factorise each 3D convolution into spatial and temporal parts; its checkpoints are the model weakly supervised on 65 million Instagram videos and that model after supervised training on Kinetics-400 (Ghadiyaram et al., 2019), the same parent-plus-one-supervised-stage shape as the MAE-only/Kinetics pair, with no masked autoencoding anywhere, although the parent's hashtag labels were themselves seeded from Kinetics-400's class names. TimeSformer-B (Bertasius et al., 2021) is a video transformer with divided space-time attention that embeds each frame separately, so it has no tubelets; its Kinetics and SSv2 checkpoints are supervised siblings, each trained from ImageNet-pretrained weights (the released configurations name the same ViT-B, while the original paper reports ImageNet-21k pretraining for its Kinetics results and ImageNet-1k for its SSv2 ones), so they may differ in initialisation as well as in the video dataset. It reads eight frames, every other frame of each clip, under its own persisted permutation. Hiera-B (Ryali et al., 2023) is a hierarchical video transformer; its checkpoints are a masked autoencoder pretrained on Kinetics-400 and that model after supervised training on Kinetics-400, the parent-plus-child shape of VideoMAE's pair in a second architecture, read at the normalised mean of its tokens on all sixteen frames, with the MAE-only model's normalisation freshly initialised. Every model reads the shared 112-pixel pipeline, and κ is read at each model's classifier input (for TimeSformer, the class token after its final normalisation).

An order-defined task: up/down. From OF-Syn's own labels we built a task in which each class is the time reversal of the other: clips of a person standing up against clips of a person falling, sitting down or lying down, the chiral-pair design of Bagad & Zisserman (2025) in our domain. A symmetric read-out of the frames can separate the classes only by scene, which the shuffled condition measures. The task uses the videos, splits and frame sampling of the fall task (5,141 training and 653 evaluation clips, 1,335 and 180 of them standing up), with a step-matched budget of 30 epochs. We ran it for the VideoMAE pair, with SSv2 as a third point, and for the R(2+1)D-34 and TimeSformer pairs, and, after its fall result, for the Hiera pair under the standard schedule; the R(2+1)D-34 pair also ran on the Something-Something temporal mirrors of Appendix B, which neither of its checkpoints was trained on.

On OF-Syn the gap survives in TimeSformer. In TimeSformer the frozen checkpoints differ as the test needs: SSv2 reads 0.047 [0.030, 0.064] at the classifier input and Kinetics -0.002 [-0.005, 0.000], a gap of 0.049 [0.031, 0.067], and every probe and arm clears 0.65. Fine-tuning does not close it: $\kappa = 0.08$ [-0.75, 0.60], an interval that contains 0 and excludes 1, so by our rule the gap does not close; at most 60% of it closes, and the fine-tuned gap is 0.045 [0.021, 0.072]. κ is near 0 because fine-tuning moves both checkpoints by about the same amount ($+0.013$ [-0.018, 0.045] for SSv2 and $+0.016$ [0.006, 0.027] for Kinetics), whereas under the same recipe it moves

Table 4: The tests on further architectures and on order-defined tasks. A at the classifier's input, frozen and fine-tuned (mean over three seeds), for the two checkpoints of each pair; the frozen gap, first-listed minus second-listed checkpoint (up to sign the denominator of Eq. 2); and κ. Intervals are 95% and resample evaluation groups and then seeds. κ is not computed when a frozen probe or an arm is below 0.65 balanced accuracy, or when the gap's interval covers 0. "std.": the standard schedule.

architecture, task	checkpoints	frozen A	fine-tuned A	frozen gap	κ
VideoMAE-B, up/down	MAE-only / Kinetics	0.325 / 0.303	0.410 / 0.388	0.022 [−0.033, 0.078]	not computed (gap covers 0)
R(2+1)D-34, fall	IG-65M / Kinetics	0.081 / 0.085	0.141 / 0.178	−0.004 [−0.035, 0.026]	not computed (gap covers 0)
R(2+1)D-34, up/down	IG-65M / Kinetics	0.064 / 0.100	0.416 / 0.390	−0.036 [−0.091, 0.020]	not computed (gap covers 0)
R(2+1)D-34, SSv2 mirrors	IG-65M / Kinetics	0.046 / 0.055	0.196 / 0.217	−0.009 [−0.034, 0.016]	not computed (frozen IG-65M probe 0.54)
TimeSformer-B, fall	SSv2 / Kinetics	0.047 / −0.002	0.059 / 0.014	0.049 [0.031, 0.067]	0.08 [−0.75, 0.60]
TimeSformer-B, up/down	SSv2 / Kinetics	0.230 / −0.001	0.378 / 0.345	0.231 [0.193, 0.270]	0.86 [0.69, 1.01]
TimeSformer-B, fall, std.	SSv2 / Kinetics	0.047 / −0.002	0.027 / −0.001	0.049 [0.031, 0.067]	0.43 [−0.03, 0.73]
Hiera-B, fall	MAE-only / Kinetics	0.214 / 0.074	−0.062 / −0.044	0.140 [0.101, 0.180]	not computed (both arms below 0.65)
Hiera-B, fall, std.	MAE-only / Kinetics	0.214 / 0.074	0.244 / 0.108	0.140 [0.101, 0.180]	0.03 [−0.54, 0.45]
Hiera-B, up/down, std.	MAE-only / Kinetics	0.403 / 0.253	0.402 / 0.290	0.150 [0.102, 0.198]	0.25 [−0.24, 0.60]

the VideoMAE pair toward each other, by -0.036 and $+0.075$. A check with VideoMAE's kind of read-out, the mean of the patch tokens instead of the class token, agrees: the frozen gap is 0.030 [0.013, 0.048] and $\kappa = -0.58$ [−2.84, 0.46]. Four differences from the VideoMAE arms qualify the comparison. The recipe is the one we use for VideoMAE, not TimeSformer's own (which uses SGD), and its models vary more across seeds than VideoMAE's (end-to-end balanced accuracy 0.82 and 0.76, standard deviations 0.05 and 0.06, against 0.04 and 0.03 over VideoMAE's ten seeds); two of its three Kinetics models, at 0.73, end below the frozen probe's 0.83, though under this recipe seven of VideoMAE's ten Kinetics models (all five in the five-seed test) end below theirs too, so we cannot exclude a poorly fitted network. It reads eight frames through the shared pipeline, spaced differently from the eight frames 32 apart its Kinetics model was trained on, and with Kinetics statistics rather than its own, and Section 4.1 shows that preprocessing moves frozen values by checkpoint-specific amounts. Its checkpoints are siblings, not a parent and its supervised child. And its measurement ran on NVIDIA H100s for both arms, since the freed H200s went to other jobs; its training ran on H200s. We therefore ran the same pair under the standard schedule, with a manipulation check that both arms' fine-tuned models beat their frozen probes by at least 0.03. They do, by $+0.084$ (SSv2) and $+0.055$ (Kinetics), at end-to-end balanced accuracy 0.891 and 0.885, and the gap still does not close: $\kappa = 0.43$ [−0.03, 0.73], an interval that contains 0 and excludes 1. The gap narrows here because SSv2 falls (to 0.027) while Kinetics stays at zero (-0.001), with training and measurement both on H100s. On OF-Syn, then, the collapse of the body does not extend to TimeSformer under either recipe; on the order-defined task below, it does.

R(2+1)D-34 has no frozen gap to close. Its two checkpoints are indistinguishable at the classifier input (0.081 and 0.085), so the gap's interval covers 0 and κ is not computed. The Kinetics stage that removes most of VideoMAE's output order sensitivity on OF-Syn removes no detectable amount of it after IG-65M pretraining, whose labels already derive from Kinetics' classes. Fine-tuning raises both and leaves them slightly apart (0.141 and 0.178; Kinetics minus IG-65M 0.036 [0.009, 0.065]): here the frozen comparison fails to predict the fine-tuned one in the other direction, indistinguishable frozen and different fine-tuned, in a descriptive contrast of three seeds per arm.

Hiera keeps its gap, as predicted. Hiera's frozen checkpoints repeat VideoMAE's pattern: MAE-only reads 0.214 [0.183, 0.242] at the classifier input and Kinetics 0.074 [0.049, 0.103], a gap of 0.140 [0.101, 0.180], while the Kinetics probe is the more accurate (0.838 against 0.822). Under our constant-rate recipe every fine-tuned Hiera model predicts no fall for every test clip (end-to-end balanced accuracy 0.50, all six), and probes on their backbones reach only 0.59 and 0.56, so both arms are below our floor and κ is not computed. We then ran the pair under the standard schedule, which adds warmup and layer-wise decay, and, before any of its models existed, made a prediction

from the use/direction reading: if the frozen up/down gap's interval covers 0, $\kappa \geq 0.5$; if it lies above 0.10, $\kappa < 0.5$. Frozen on up/down, MAE-only reads 0.403 and Kinetics 0.253 (probes 0.90 and 0.80), a gap of 0.150 [0.101, 0.198] whose lower bound clears 0.10 by 0.001, so the prediction was $\kappa < 0.5$. Under the standard schedule every model fits the task (end-to-end balanced accuracy 0.909 to 0.930), and the gap does not close: $\kappa = 0.03$ [-0.54, 0.45], an interval that contains 0 and excludes 1, so the prediction holds. Both checkpoints move up by about the same amount ($+0.030$ and $+0.034$), every MAE-only model stays above every Kinetics model (0.17 to 0.28 against 0.09 to 0.14), and the models' own heads agree ($\kappa = 0.09$ [-0.34, 0.38]). Training and measurement ran on H100s.

The up/down task. The task needs frame order as designed: with the frames shuffled, every fine-tuned probe of the first three pairs falls to between 0.51 and 0.58 balanced accuracy and every fine-tuned model to between 0.50 and 0.59, against at least 0.89 on ordered clips. So for these pairs A is close to ordered accuracy minus chance, and it mostly tracks accuracy. Hiera's fine-tuned models are the exception (below).

The TimeSformer pair is the one that differs frozen: SSv2 reads 0.230 [0.196, 0.269] at the classifier input and Kinetics -0.001 [-0.003, 0.000], a gap of 0.231 [0.193, 0.270], with frozen probes at 0.83 and 0.70. Fine-tuning closes it: $\kappa = 0.86$ [0.69, 1.01], an interval that excludes 0 and contains 1, so by our rule the gap closes completely within error. Unlike on OF-Syn, fine-tuning moves the two checkpoints by different amounts, $+0.148$ [0.094, 0.203] for SSv2 and $+0.346$ [0.303, 0.391] for Kinetics, leaving a fine-tuned gap of 0.033 [-0.004, 0.069], and its models reach 0.93 and 0.92 balanced accuracy end to end. The frozen Kinetics checkpoint's classifier input is insensitive to frame order on this task as on OF-Syn (its probe reaches 0.699 on ordered clips and 0.700 on shuffled ones), consistent with Kinetics being largely solvable from appearance (Bertasius et al., 2021). Both arms were measured on H100s; their seeds 0 and 1 trained on H200s and seed 2 on H100s.

Hiera's pair differs frozen by the gap that fixed the prediction above, and fine-tuning on up/down under the standard schedule does not close it: $\kappa = 0.25$ [-0.24, 0.60], an interval that contains 0 and excludes 1, and unlike TimeSformer's, Hiera's gap survives on the order-defined task as on OF-Syn. Every model fits the task (end-to-end balanced accuracy 0.982 to 0.991), but fine-tuning leaves MAE-only where it started (0.403 to 0.402) while Kinetics rises (0.253 to 0.290), leaving a fine-tuned gap of 0.112 [0.064, 0.159]. Here A does not track accuracy: on shuffled clips the Kinetics models' probes still reach 0.67 to 0.71 balanced accuracy (MAE-only's 0.58 to 0.59), so they separate part of the task without frame order. Read from the models' own heads, the gap closes ($\kappa = 0.62$ [0.12, 1.04], an interval that excludes 0 and contains 1; Appendix B). Training and measurement ran on H100s.

Frozen, the VideoMAE and R(2+1)D-34 pairs are not distinguishable: the VideoMAE checkpoints read 0.325 and 0.303 (gap 0.022 [-0.033, 0.078]) and the R(2+1)D-34 checkpoints 0.064 and 0.100 (gap -0.036 [-0.091, 0.020]), so κ is not computed for them. (On the Something-Something mirrors the VideoMAE pair did differ, by 0.054 [0.029, 0.081], but there the frozen Kinetics probe fell below our floor; Appendix B.) Three observations follow without a test. First, on up/down the frozen Kinetics probe is about as order-dependent as MAE-only's (0.836 on ordered clips and 0.533 on shuffled ones, against 0.866 and 0.542), so the fall-task drop of Section 4.1 does not show that Kinetics supervision discards temporal information. Second, the SSv2 checkpoint starts higher (0.421 [0.376, 0.466], an interval disjoint from Kinetics'), and fine-tuning brings MAE-only and Kinetics to it (0.410 and 0.388, against SSv2's own 0.415): the three VideoMAE checkpoints span 0.303 to 0.421 frozen and 0.388 to 0.415 fine-tuned (for the SSv2/Kinetics pair, $\kappa = 0.77$ [0.34, 1.09]). Third, fine-tuning raises A in every arm of the three pairs, and it brings their six checkpoints from between -0.001 and 0.325 frozen to between 0.345 and 0.416 fine-tuned; that comparison is across architectures, and it is largely a convergence of accuracy, so we report it only as a description. On the Something-Something mirrors both frozen R(2+1)D-34 probes fall below our floor (0.54 and 0.56), so κ is not computed there; fine-tuning raises A in both arms (from 0.046 to 0.196 for IG-65M and from 0.055 to 0.217 for Kinetics, whose arm was measured on an H100).

Direction information after fall fine-tuning. The body reads A on up/down as the direction information a probe can read at the classifier input, and A on the fall task as how much a fall read-out uses order (Section 4.2). After the TimeSformer results we tested the consequence that fall fine-

tuning leaves the direction information roughly where pretraining put it: the up/down probe, fitted and evaluated as above, applied to the fall-fine-tuned backbones of both recipes, seeds 0 to 2, with no further training and the frozen values re-measured in the same job (they reproduce the values above exactly). Table 5 gives the results. Every frozen and fine-tuned probe clears 0.65 (lowest 0.699 and 0.712, both TimeSformer Kinetics). Every VideoMAE arm keeps at least 0.20 (one Kinetics seed under the standard schedule reads 0.199), and TimeSformer's gap stays above 0.10; after constant-rate fall fine-tuning TimeSformer Kinetics reads 0.056 (seeds 0.044, 0.061 and 0.061), and under the standard schedule 0.000. After the Hiera result we ran the same test for Hiera's standard-schedule fall models and for VideoMAE's SSv2 checkpoint under the constant rate. Hiera keeps its direction gap (0.150 frozen, 0.133 after fine-tuning for fall detection), with both arms above 0.20 (0.369 and 0.237). VideoMAE's fall fine-tuning instead narrows SSv2's advantage over Kinetics from 0.119 to 0.045, although SSv2 itself stays at 0.332; the gaps that remain, 0.045 [0.001, 0.092] to Kinetics and 0.060 [0.014, 0.107] to MAE-only, still exclude 0 (Section 4.2).

Table 5: Direction information after fall fine-tuning: A on up/down at the classifier input, frozen and on the fall-fine-tuned backbones (mean over seeds 0 to 2). Intervals on the gaps after fine-tuning are in Section 4.2.

architecture	checkpoint	frozen	constant rate	standard schedule
VideoMAE-B	MAE-only	0.325	0.272	0.290
	Kinetics	0.303	0.287	0.226
	SSv2	0.421	0.332	–
	SSv2 minus Kinetics	0.119	0.045	–
	SSv2 minus MAE-only	0.097	0.060	–
	MAE-only minus Kinetics	0.022	−0.015	0.064
TimeSformer-B	SSv2	0.230	0.217	0.263
	Kinetics	−0.001	0.056	0.000
	SSv2 minus Kinetics	0.231	0.162	0.263
Hiera-B	MAE-only	0.403	–	0.369
	Kinetics	0.253	–	0.237
	MAE-only minus Kinetics	0.150	–	0.133

Knocking out temporal attention. We also set the output of the temporal-attention branch to zero in every block of each fine-tuned TimeSformer, so that nothing moves between frames through it, and scored each model's own head on its own task's ordered test clips. The up/down models fall to near chance (balanced accuracy 0.50 to 0.57; drops of 0.410 for Kinetics and 0.395 for SSv2), so the knockout can matter. It costs the Kinetics fall models 0.010 under the standard schedule and 0.077 under the constant rate, and the SSv2 fall models 0.077 and 0.074. Setting a branch to zero also moves later activations off their usual range, so part of any drop can come from that alone; and since the Kinetics fall models' order sensitivity stays near zero, what they lose need not be order information.

D THE FAILED OOPS KINETICS SEED

One of the five OOPS Kinetics seeds did not fit the task: its fine-tuned probe reaches a balanced accuracy of 0.592 on ordered clips and the fine-tuned model 0.575. Our exclusion rule is an arm-mean floor, which this arm clears (0.714), so the seed stays in every primary number. Without it, OOPS Kinetics reaches $A_{\mathrm{ft}} = 0.202$ rather than 0.170 at the classifier's input; the spread between checkpoints at the bare read-out falls from 0.170 to 0.018 rather than 0.052, a reduction of 0.153 [0.094, 0.201] rather than 0.118 [0.063, 0.171], or 90% rather than 69%; and the largest fine-tuned spread at any depth has an upper bound of 0.073 rather than 0.102. The seed's own drop from peak to output is 4%, so it does not affect the drop-from-peak results.

E SPREAD BY DEPTH

The spread is the maximum minus the minimum of $A(\ell)$ over the three checkpoints at the bare read-out; fine-tuned values are constant-rate means over seeds 0 to 2, including the failed OOPS

Table 6: Between-checkpoint spread of $A(\ell)$ by block, with 95% group-bootstrap intervals.

	OF-Syn		OOPS	
block	frozen	fine-tuned	frozen	fine-tuned
0	0.002 [0.001, 0.010]	0.001 [0.000, 0.004]	0.028 [0.011, 0.047]	0.005 [0.001, 0.018]
1	0.017 [0.004, 0.045]	0.007 [0.002, 0.023]	0.059 [0.023, 0.098]	0.015 [0.005, 0.036]
2	0.073 [0.044, 0.103]	0.012 [0.004, 0.031]	0.021 [0.006, 0.064]	0.023 [0.008, 0.049]
3	0.055 [0.025, 0.087]	0.024 [0.009, 0.041]	0.036 [0.009, 0.081]	0.024 [0.008, 0.047]
4	0.044 [0.016, 0.077]	0.017 [0.004, 0.036]	0.003 [0.005, 0.059]	0.026 [0.007, 0.056]
5	0.071 [0.043, 0.101]	0.015 [0.005, 0.036]	0.010 [0.005, 0.062]	0.015 [0.004, 0.042]
6	0.041 [0.020, 0.073]	0.038 [0.022, 0.057]	0.030 [0.008, 0.075]	0.035 [0.014, 0.066]
7	0.032 [0.009, 0.067]	0.039 [0.023, 0.060]	0.029 [0.008, 0.077]	0.046 [0.023, 0.074]
8	0.069 [0.037, 0.101]	0.033 [0.018, 0.053]	0.055 [0.020, 0.100]	0.075 [0.049, 0.102]
9	0.111 [0.082, 0.146]	0.021 [0.006, 0.041]	0.051 [0.019, 0.104]	0.059 [0.036, 0.083]
10	0.148 [0.114, 0.185]	0.026 [0.008, 0.046]	0.125 [0.089, 0.179]	0.059 [0.036, 0.085]
11	0.209 [0.176, 0.243]	0.033 [0.015, 0.054]	0.204 [0.160, 0.245]	0.064 [0.037, 0.091]
12	0.237 [0.202, 0.270]	0.023 [0.010, 0.044]	0.170 [0.122, 0.220]	0.052 [0.026, 0.081]

Kinetics seed. Seed variation is not in these intervals, and a maximum minus minimum over seed means overstates a fine-tuned spread, so the fine-tuned columns do not test whether a spread is zero. A maximum minus minimum is also biased upward under resampling, so where the checkpoints nearly coincide an interval can lie above its point estimate (OOPS, frozen, block 4).

The location of the peak. Because we re-select the maximum inside every bootstrap resample, the drop's interval carries the uncertainty in where the peak is. The peak block itself is unstable: for OF-Syn Kinetics the full-sample peak is the maximum in only 40% of resamples, although that cell's drop is tightly estimated, and across the frozen checkpoints the share ranges from 40% to over 99%. Layer-wise studies commonly report the depth at which a property peaks (Joseph et al., 2026; Musa et al., 2026); at evaluation sizes like ours that location may not be identifiable, so we report drops, not peak depths. We also do not interpret where the peak sits: a mid-network optimum can arise from distribution shift between pretraining and probing data (Odonnat et al., 2026), and both our datasets are far from Kinetics-400 and Something-Something v2. The contrasts of Section 5 are between checkpoints at a fixed depth, which that account does not address.

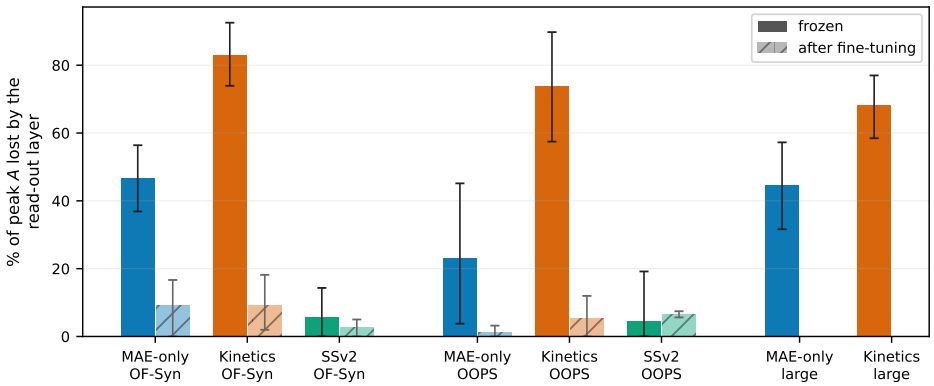


Figure 4: Drop from each checkpoint's maximum order sensitivity to its last block, frozen (solid, 95% interval) and after constant-rate fine-tuning (hatched, mean of per-seed drops, range over seeds), including the VideoMAE-large pair (frozen only). Kinetics' maximum is never below MAE-only's beyond noise: 0.254 against 0.258 on OF-Syn, 0.218 against 0.167 on OOPS, and 0.251 against 0.228 for the large pair.

F THE ATTENTION-BIAS LOADING REGRESSION

VideoMAE checkpoints store the attention query and value biases under keys of their own, which earlier releases of `transformers` applied by hand. Pull request 45139, first released in 5.6.0, moved these biases into the linear projections without mapping the checkpoint keys. From 5.6.0 through 5.17.0, the latest release at the time of submission, the stored biases are therefore dropped as unexpected keys, the new bias parameters are left at zero, and the model loads, trains and evaluates normally. A fix has since been merged to the main branch but not released. A third party reported the same defect, on the MAE-only checkpoint, in the pull request that fixes it.

On 16 OF-Syn clips a model loaded this way produces features whose cosine similarity to the correctly loaded model's averages 0.823 (relative L_2 difference 0.564). Measured in pairs on the Kinetics checkpoint, same clips and permutation, it inflates frozen A 4.3-fold on the fine-tuning pipeline's extraction path (0.203 against 0.047, at the classifier's input) and 4.0-fold on a second 112-pixel extraction path (0.171 against 0.043). On native input, measured three months apart with the same code and data, it inflates A 2.0-fold (0.237 against 0.116). In all three cases the correctly loaded model is also slightly *more* accurate on ordered clips. So the defect makes frozen VideoMAE features look more order-sensitive than they are. We do not know of a published result that it affects, and cannot check, because papers rarely report the library version. Every job in this paper asserts, before computing anything, that the summed magnitude of the loaded query biases matches the checkpoint file (2816.998 for the Kinetics checkpoint).