IMAGE ROBUSTNESS DOES NOT IMPLY VIDEO ROBUSTNESS: THE LIMITS OF SINGLE-FRAME EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Single-image corruption benchmarks such as ImageNet-C are a cheap way to choose a robust backbone for a video system, but this shortcut works only if image and video corruptions rank the models the same way. We test how well a model's robustness to single-image corruptions actually predicts its robustness to video corruptions. We evaluate models on image and video corruptions under identical conditions and check whether the single-image robustness ranking predicts the video robustness ranking, using a rank-correlation threshold for reliable prediction. Our central finding is that image robustness transfers to video only under certain corruption conditions. Specifically, we observed that the single-image ranking transfers to video to the extent that a single frame shows the corruption. When a corruption realization is shared across a video clip, the single-image ranking is preserved almost perfectly across seven settings. By contrast, when a corruption disrupts the temporal structure of a clip, the single-image ranking does not reliably predict the video ranking in any setting. These temporal-structure corruptions include frame drop, repetition, and reordering, which damage the frame sequence even though every delivered frame stays clean. Corruptions that a single frame shows only in part give mixed results, for example when the corruption strength drifts over time or the camera shakes. Single-image robustness had essentially no rank correlation with video robustness to temporal-structure corruptions among 37 video-pretrained 3D backbones. The same pattern was observed in codec and transport pipelines: Single-image robustness predicts robustness to codec compression well, but poorly predicts robustness to packet loss or variable-frame-rate judder. Our theoretical analysis proves that two corruptions can look identical to a single-image benchmark yet reverse the ranking of two video models, implying that image-only testing cannot guarantee the right model choice for video. We recommend adding a separate targeted temporal evaluation whenever temporal structure can be disrupted, and selecting video models by single-image robustness only when a single frame shows the expected corruptions.

1 INTRODUCTION

Single-image benchmarks such as ImageNet-C, the corrupted Common Objects in Context benchmark, and Cityscapes-C provide a standard way to evaluate robustness to noise, blur, weather, and compression artifacts (Hendrycks & Dietterich, 2019; Michaelis et al., 2019; Kamann & Rother, 2021). These image-robustness scores can also be used to choose a backbone for a video system. Using image-robustness scores for backbone selection assumes that image robustness transfers to video robustness (Fig. 1).

However, video can be corrupted in ways that a single frame cannot represent. Some corruptions alter the appearance of individual frames: Fog or compression may persist across a video clip, but the haze or compression artifacts are visible even in a single frame. Other corruptions disrupt the frame sequence: Frames can be dropped, repeated, reordered, or delivered at an irregular rate. Video pipelines deployed in the real world routinely produce these faults. When packets are lost during video transmission over a network, video decoders often fill in missing content using earlier frames, making old visual information reappear (Wang & Zhu, 1998). The Real-time Transport Protocol

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.

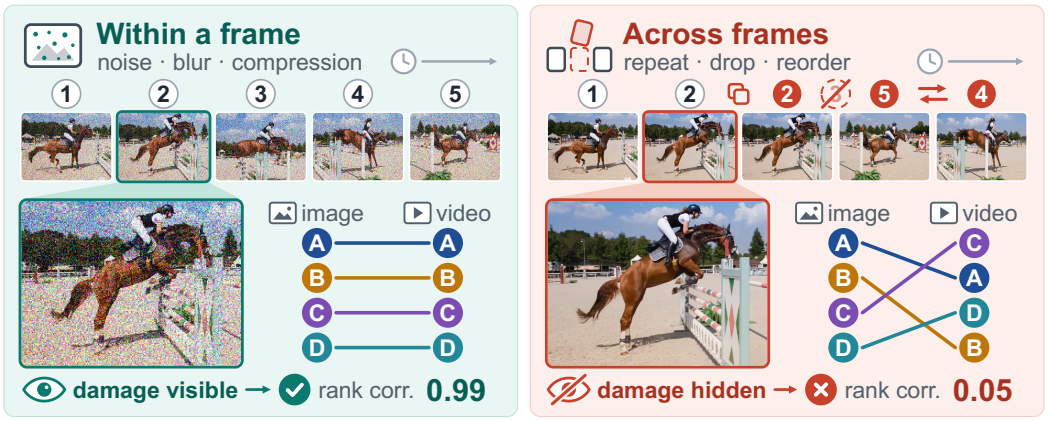


Figure 1: Where the single-image proxy holds and fails. An image-robustness benchmark ranks candidate video models one frame at a time, shown here with schematic rankings for models A to D; each side enlarges one such frame. Corruptions that a single frame can represent, such as noise, blur, and compression, are visible in every delivered frame, and the image-robustness ranking transfers almost perfectly to video evaluation (left). Temporal-structure corruptions, such as frame repetition, drop, and reordering, leave each delivered frame clean, so no single frame reveals the fault, and the same image-robustness ranking has almost no correlation with the video-robustness ranking (right). Frames come from a video clip in the Densely Annotated VIdeo Segmentation (DAVIS) dataset (Pont-Tuset et al., 2017).

(RTP), used for streaming video over networks, guarantees neither packet delivery nor arrival order, and live analytics systems skip frames to fit bandwidth and compute budgets (Schulzrinne et al., 2003; Li et al., 2020b). Evaluating robustness from one frame alone cannot reveal whether frames in the video were dropped, repeated, reordered, or delivered at irregular intervals. A video model can amplify the resulting prediction errors by combining information across frames and can also carry these errors into later predictions by retaining information from earlier frames. These properties of video raise our research question of when single-image robustness predicts video robustness and when temporal structure breaks that prediction. Indeed, Theorem 1 shows that the image-robustness ranking can transfer for one corruption yet fail for another that looks identical frame by frame. The success or failure of this transfer therefore requires empirical validation, which Secs. 3 and 4 provide.

We address this question with matched evaluations that keep model weights fixed and change only the input and corruption protocol. Two axes organize the study: how much of a corruption one frame can show, and whether the task can be solved from appearance or requires temporal information.

Our experiments reveal a consistent pattern: Image robustness predicts video robustness best when a single frame shows the corruption (Table 1). When the same corruption pattern persists throughout a video clip, the ranking of models by robustness to single-image corruptions is preserved across seven settings. Frame drops, repetitions, and reordering change which models are more robust, even though each delivered frame remains visually intact. Corruptions that one frame only partly reveals, such as blur that changes over time or camera shake, give less consistent rank transfer. For temporal-structure robustness, the video ranking is almost unrelated to the image-robustness ranking among video-pretrained three-dimensional (3D) models on Diving48. Selecting between two models by image robustness is then close to random choice. We also find that the temporal-structure robustness ranking transfers across datasets, showing that temporal robustness is a distinct model property that image evaluation misses. Codec and transport pipelines show the same pattern: Single-image robustness predicts robustness to compression but is a much weaker guide to packet loss and variable-frame-rate (VFR) judder. Fig. 4 in App. F.1 illustrates why.

Prior corruption benchmarks have characterized image or video robustness within benchmark-specific protocols (Hendrycks & Dietterich, 2019; Yi et al., 2021; Schiappa et al., 2023; Zeng et al., 2024; Zhang et al., 2026). We instead test whether the model ranking from the image protocol remains valid under the video protocol. This ranking-level view connects corruption robustness to benchmark and construct validity (Jacobs, 2021; Dehghani et al., 2021; Gulrajani & Lopez-Paz, 2021) and adopts

the model-zoo rank agreement by which generalization measures are validated (Jiang et al., 2020); App. B reviews the broader literature.

Our contributions are as follows:

- We test whether the image-robustness ranking predicts the video-robustness ranking, keeping each model's weights fixed across image and video corruptions. We propose a matched-pair audit protocol that implements this test for video models built from pretrained backbones.
- Our experiments show that the transfer of robustness rankings follows how much of the corruption one frame can represent, with evidence across appearance-dominated and motion-dependent tasks, video-pretrained 3D models, public checkpoints, video object segmentation (VOS), and codec and transport pipelines.
- Our theoretical analysis proves that two video corruptions can look identical under single-frame evaluation yet reverse model rankings, establishing an unavoidable limit on image-only model selection. We derive a lower bound on the resulting worst-case selection regret, identify when frame-level information is sufficient, and bound the effect of temporal interactions on model risk.
- We propose a model-selection method that improves prediction of video-robustness rankings under packet loss and VFR judder. The method combines the image-robustness ranking with a temporal probe that uses only eight corruption conditions at one severity.

2 Problem Formulation and Experimental Setup

We test whether robustness to corruptions applied to individual frames can guide the choice of a video model. The key question is whether that measurement ranks candidate models in the same order as evaluation on corrupted videos.

A matched-pair design. For the matched classification experiments, we initialize each backbone from a pretrained checkpoint and fine-tune the backbone and classification head together on clean training clips from the target dataset. We evaluate the static reference and every video corruption family with the same fine-tuned weights. Diving48 (Li et al., 2018) and Something-Something V2 (SSv2) (Goyal et al., 2017) provide motion-dependent tasks, where recognition relies on changes across frames; the University of Central Florida 101-class dataset (UCF101) and the 51-class Human Motion Database (HMDB51) provide the appearance-dominated controls.

Matched robustness scores. Fix a model set $\mathcal{M}$, an evaluation protocol p, and a nonnegative, higher-is-better metric P. The protocol may be the $T = 1$ image protocol or a multi-frame video protocol. The metric is top-1 accuracy for classification and the stated task metric in later evaluations. For model m, let the performance on uncorrupted input under protocol p be $P_{m,p,\text{clean}} > 0$, and let the model's performance under corruption c at severity q be $P_{m,p,c,q} \geq 0$. We use retention as our primary robustness score: corrupted-input performance divided by the same model's clean performance,

$$g(m,p,c,q) := \frac{P_{m,p,c,q}}{P_{m,p,\text{clean}}} \in [0,\infty), \qquad \Delta_{\text{rel}}(m,p,c,q) := 1 - g(m,p,c,q) \in (-\infty, 1]. \tag{1}$$

A retention of 0.8 means that corrupted-input performance is 80% of the model's clean performance. Each score is relative to the clean baseline of the corresponding evaluation protocol; the image and video protocols can have different clean performances. We use a group-balanced average so that groups with more corruption types, or types with more severity levels, do not receive more weight within family f:

$$s^{(f)}_{m,p} := \frac{1}{|\mathcal{G}_f|} \sum_{r \in \mathcal{G}_f} \frac{1}{|\mathcal{C}_{f,r}|} \sum_{c \in \mathcal{C}_{f,r}} \frac{1}{|\mathcal{Q}_c|} \sum_{q \in \mathcal{Q}_c} g(m,p,c,q), \tag{2}$$

where $\mathcal{G}_f$ are the spatial groups within family f and $\mathcal{C}_{f,r}$ the corruption types in group r. The protocol p is held fixed throughout the sum. We test each family separately; the rows labeled Overall in this paper's rank-correlation tables use video scores averaged across corruption families and do not determine rejection decisions.

The scores we compare. The video score $s_m^{\mathrm{vid}(f)} := s_{m,\mathrm{vid}}^{(f)}$ is the model's average retention under corruption family f with multi-frame input, using the weighting in Eq. 2. We compare the video score with a static reference that measures robustness to spatial corruptions, such as noise, blur, and compression, applied independently to individual frames. We call this measurement the image-style static reference because the spatial corruptions are applied frame by frame, as in an image benchmark, even when the video model receives several frames.

On appearance-dominated tasks, the reference is the single-frame score $s_m^{\mathrm{stat}} := s_m^{\mathrm{img}} := s_{m,\mathrm{img}}^{(\text{frame-independent})}$, measured with $T = 1$. On Diving48 and SSv2, both evaluations use the same multi-frame clips because one frame does not show how the action unfolds. For the static reference, we apply the same spatial corruptions independently to each input frame: $s_m^{\mathrm{stat}} := s_{m,\mathrm{vid}}^{(\text{frame-independent})}$. This reference preserves the video clip's original sampling and order. On the appearance tasks, the single-frame and multi-frame references can both be evaluated and produce almost identical model rankings, with Spearman $\rho = 0.98$. The ranking given by s^{stat} is the static ranking, called the image-robustness ranking in Sec. 1; the ranking given by $s^{\mathrm{vid}(f)}$ is the video ranking for family f.

For comparisons on Diving48 and SSv2, the static reference and video scores use the same network weights, video clips, temporal head, input length, and aggregation rule (Eq. 2). Only the corruption family changes, so the comparison directly tests whether frame-by-frame spatial-corruption robustness predicts the video ranking.

Measuring agreement between the rankings. The primary endpoint is Spearman's rank correlation (Spearman, 1987) between the static and video scores over $\mathcal{M}$,

$$\widehat{\rho}^{(f)} := \mathrm{Spearman}\big((s_m^{\mathrm{stat}})_{m\in\mathcal{M}}, (s_m^{\mathrm{vid}(f)})_{m\in\mathcal{M}}\big). \tag{3}$$

We write $\rho^{(f)}$ for this observed coefficient over the evaluated model set and report Kendall's tie-corrected τ_b (KENDALL, 1938; 1945) as a second rank statistic. For the model-population test, let $\rho_*^{(f)}$ denote the corresponding population rank association. We also report concordance: the fraction of model pairs for which the two scores prefer the same model, excluding pairs tied in either score. Concordance is one minus the reversal fraction and has a chance value of $1/2$. A high rank correlation means that models judged more robust by the static reference also tend to be more robust under the video corruption. The absolute retention values cannot be compared directly across protocols with different clean baselines and attainable score ranges.

Testing whether the static reference is reliable. Finite samples, tied scores, and measurement noise can prevent perfect rank agreement even for a useful proxy. We set $\rho_0 = 0.9$ in advance as the required level of rank agreement (App. E.1), rather than treating every departure from $\rho = 1$ as a failure. For each family, we test

$$\mathcal{H}_0^{(f)} : \rho_*^{(f)} \geq \rho_0 \quad \text{against} \quad \mathcal{H}_A^{(f)} : \rho_*^{(f)} < \rho_0. \tag{4}$$

We use the pre-specified approximate Fisher-z rule detailed in App. E.2 and report model-set and paired clip-level bootstrap confidence intervals (CIs), described in App. E.3.

If video corruptions hurt some models much more than other models, the static reference should rank the more robust models higher. A weak correlation between the static and video rankings means the static reference fails to do so reliably. If all models retain almost the same performance, a different ranking has little practical effect. We distinguish these cases using the score spread: the sample standard deviation of the models' retention scores within each family.

Representability hypothesis H1. We expect the static ranking to predict the video ranking best when a single frame shows the corruption. When corruption strength stays constant over time, every frame shows the damage, so we expect the two rankings to agree with $\rho_*^{(f)} \geq \rho_0$, whether the random pattern changes between frames or is shared throughout the clip. By contrast, temporal-structure corruptions change which frames are delivered or the frame order while leaving each delivered frame visually intact. We expect rank agreement to fall below ρ_0 when these corruptions cause clear differences in retention across models. Drifting and motion-coupled corruptions are partly visible in one frame, so we expect these two families to break the rank-agreement criterion less consistently.

A representability limit for marginal-only audits. For a video corruption channel $C : X \mapsto \widetilde{X}$, define the channel's coordinate kernel by $K_{C,t}(A \mid x) \coloneqq \Pr_C(\widetilde{X}_t \in A \mid X = x)$. Call an audit marginal-only if the audit's information about C is unchanged whenever all coordinate kernels $K_{C,t}$ are unchanged; such an audit can know the clean distribution and the candidate models, but not the corruption's cross-frame coupling.

Theorem 1 (Temporal coupling is not identifiable from frame marginals)**.** *For every flip rate $q \in (1/4, 1/2]$, there exist a binary two-frame classification distribution, three corruption channels $C_{\mathrm{coh}}, C_{\mathrm{anti}}, C_{\mathrm{ind}}$ with identical coordinate kernels, and two deterministic video classifiers h_+, h_- such that, for zero-one risk $R_C(h) \coloneqq \Pr_C(h(\widetilde{X}) \neq Y)$,*

	h_+	h_-
C_{coh}	0	1
C_{anti}	$2q$	$1-2q$
C_{ind}	$2q(1-q)$	$1-2q(1-q)$.

(5)

Corollary 1 (Minimax selection limit for marginal-only decisions)**.** *In the setting of Theorem 1, every possibly randomized marginal-only selector A taking values in $\{h_+, h_-\}$ satisfies*

$$\max_{C \in \{C_{\mathrm{coh}}, C_{\mathrm{anti}}\}} \mathbb{E}[R_C(A) - \min_{h \in \{h_+, h_-\}} R_C(h)] \geq \frac{4q-1}{4q}, \tag{6}$$

where the expectation includes the selector's data and randomization. At $q = 1/2$, the randomized selection lower bound is $1/2$, and every deterministic selector has regret 1 on one of the two channels.

These are existence and worst-case limits, not a claim that every dynamic corruption must break every static proxy. App. C.2 states the rank-reversal consequence, App. C.3 identifies when frame-level information is sufficient, and App. D.3 gives the proofs.

A temporal corruption taxonomy. We group corruptions into five families according to how corruption patterns change across frames. Frame-independent corruptions sample random corruption parameters separately for each frame, such as the pixel-wise noise values or the motion-blur angle, while keeping severity fixed. Temporally coherent corruptions reuse those parameters for every frame at the same severity. Joint Photographic Experts Group (JPEG) compression and other deterministic operators use no random corruption parameters. For these deterministic operators, the frame-independent and temporally coherent modes produce the same output for a given input frame at a fixed severity. Temporally drifting corruptions keep the same random parameters but gradually change the strength from severity 1 to the selected severity over the clip. Temporal-structure corruptions change frame sampling or order through frame drop, repeat, shuffle, VFR, and temporal jitter. Motion-coupled corruptions apply camera shake or time-varying blur and brightness. The static reference uses the frame-independent family; we evaluate severities $\{1, 3, 5\}$ and aggregate with Eq. 2.

3 Motion Regime: Where the Proxy Transfers and Fails

We ask whether selecting a model by static robustness also selects a model that is more robust to video corruption. We compare the appearance regime of UCF101, where a single frame often identifies the action, with the motion regime of Diving48 and SSv2, where recognition depends on information across frames. On UCF101, we use a head that averages the frame predictions without using frame order, a competitive choice for clean action-recognition accuracy when individual frames already identify the action. On the motion-dependent tasks, we use a temporal-Transformer (tx) head that can use frame order to recognize the action. On UCF101, models are nearly equally robust to dynamic corruption, so choosing a different model makes little difference to robustness (App. P.1). On Diving48 and SSv2, dynamic-corruption robustness differs across models, so an inaccurate static ranking can lead to a less robust choice.

We therefore focus on the motion regime. On Diving48, spatial corruptions reduce accuracy more than on UCF101, and dynamic corruptions now cause substantial losses. These two changes let us test whether the static reference identifies the models that best withstand corrupted video.

Spatial corruptions cause more damage, but the static ranking still transfers. Mean static retention is 0.763, compared with 0.868 on UCF101. Li et al. (2018) designed Diving48 to reduce reliance on static scene and object cues, leaving fewer useful static cues to support recognition when frames are degraded. We observed that the static ranking still transfers essentially perfectly, at coherent $\rho = 1.00$ in Table 3, across a wide cross-model spread of 0.093. The coherent condition is the positive control, which tests whether reusing a corruption pattern across frames is enough to change which models are more robust (App. C.3). In these experiments, the ranking is preserved even with a head that uses frame order on a task that needs information from multiple frames. Even when a spatial corruption persists across frames, the static score still identifies the more robust video models in this setting.

Dynamic corruptions now cause large losses that differ across models. Dynamic corruptions produce both larger accuracy losses and measurable differences in how much accuracy models retain. Choosing between models matters for robustness under these corruptions. Across UCF101 and HMDB51, the same frame-repeat operator has mean retention of at least 0.83 across models even at severity 5; on Diving48, that operator and the other dynamic corruptions cause large, consistent accuracy losses. At severity 5, frame-repeat retains only 0.36, frame-drop 0.68, and focus-drift 0.71 of clean accuracy. The worst dynamic operator retains 0.36. A single-image benchmark cannot fully measure robustness to these failures. The benchmark corrupts one frame at a time, so a gradual change in corruption strength appears only as one fixed severity, and missing, repeated, or reordered frames do not appear at all. The image-style static reference thus leaves these temporal failure modes unmeasured even when the resulting accuracy losses are large.

Static robustness only partly predicts dynamic robustness. We now examine whether models that tolerate spatial corruption also retain more accuracy under these temporal disruptions. Table 3 shows that the correlations fall from near-perfect static transfer to moderate-to-strong values below the required $\rho_0 = 0.9$: drift $\rho = 0.77$, temporal structure $\rho = 0.73$ with video clip CI $[0.49, 0.79]$, and motion $\rho = 0.81$. Between 0.20 and 0.23 of model pairs reverse order. Permutation tests find positive association for every dynamic family, with $p < 0.001$. These positive correlations show that the static score provides useful information about dynamic robustness, but a model's spatial-corruption robustness does not fully determine the same model's temporal-corruption robustness. Among video-pretrained 3D models, the static score no longer provides this information for temporal structure, with $\rho = 0.05$ (Sec. 4.2).

The comparison with the coherent control identifies what static evaluation misses. With the same models, task, and head, reusing a spatial corruption pattern preserves the ranking, whereas drift, temporal structure, and motion coupling lower the correlation between the static and video rankings below our required $\rho_0 = 0.9$ (App. H). These results distinguish two aspects of video robustness: resistance to spatial damage within frames, which the static score measures, and resistance to disruptions across frames, such as repetition, reordering, or camera shake, which the static score does not measure directly.

Robustness to temporal structure also transfers across datasets. Temporal-structure robustness is a model's ability to retain accuracy when frames are dropped, repeated, reordered, or retimed. We evaluate the identical 37-model 3D set on Diving48 and SSv2 to test whether this property is reproducible across datasets. Temporal-structure robustness measured on Diving48 predicts the model ranking for the same corruption family on SSv2 at $\rho = 0.67$. The image-style static reference on Diving48 predicts that same SSv2 ranking at only 0.27, with Williams $p = 0.0180$ for the difference between the two correlations. Temporal-structure robustness also transfers more strongly across the datasets than clean accuracy, whose correlation is $\rho = 0.43$. These results show reproducible differences in temporal robustness that the static score only partly captures, supporting an explanation beyond measurement noise for the weak static→structure correlation.

4 Validating and Improving Video Model Selection

The previous section shows that static robustness can mis-rank models under temporal corruption. We now test whether this problem persists under codec and transport faults and across model architectures

and tasks, showing how this mismatch can lead to a less robust model choice. We then add a small set of temporal tests to improve that choice.

4.1 Rank prediction under codec and transport faults

Compression: a fault visible in one frame. We first test compression, whose artifacts are visible in an individual frame. We rank models by the fraction of clean accuracy each model retains after JPEG compression of individual frames, then compare this JPEG ranking with the ranking under H.264 and H.265 video compression. We re-encode every video clip through low-bitrate H.264 and H.265 video codecs (Wiegand et al., 2003; Sullivan et al., 2012). The ranking under frame-by-frame JPEG compression tracks the video-codec ranking at $\rho = 0.89$ on UCF101 and 0.93 on Diving48, and the image-style static reference predicts codec robustness at 0.92 on Diving48 (Table 8). Spatial-corruption tests can therefore help identify models that tolerate video codec compression. We next ask whether the static reference also identifies the more robust models when a pipeline disrupts frame timing.

Temporal faults: dropped, repeated, or retimed frames. The central failure in Sec. 3 occurs under temporal-structure corruption: Models ranked higher by the static reference need not retain more accuracy when frames are dropped, repeated, reordered, or retimed. Our five tests of frame order and timing simulate these faults directly in the input clips: The tests drop frames and hold the last available frame, repeat or shuffle frames, or change frame sampling to simulate VFR and temporal jitter. We check whether robustness to these synthetic faults predicts robustness to two faults that deployed streams exhibit: transport packet loss and container-level VFR. For each model, we compare the synthetic temporal-structure retention score and the image-style static reference as predictors of robustness under packet loss and container-level VFR (Table 9).

Tests of frame order and timing predict pipeline robustness better. Among the 37 video-pretrained 3D models, the synthetic temporal-structure score predicts the packet-loss and VFR ranking at $\rho = 0.78$, with CI $[0.57, 0.90]$ and pairwise concordance 0.79. The static reference predicts the same ranking much more weakly, at $\rho = 0.34$ with CI $[0.02, 0.60]$. Because both correlations use the same pipeline ranking as the prediction target, we compare the two correlations with Williams's T_2 test (Williams, 1959; Steiger, 1980), which gives one-sided $p = 0.0036$. We also control for the codec's own compression damage by dividing each pipeline retention score by a matched compression-only baseline. After this adjustment, the synthetic temporal-structure score predicts the pipeline ranking at $\rho = 0.84$, while the static reference reaches only 0.25; App. K.2 details the paired comparison and the compression control. The same pattern appears in the 27-model Diving48 tx set: The static reference predicts codec robustness at 0.92 but compression-controlled packet-loss and VFR robustness at only 0.57. Testing frame order and timing therefore gives a better guide to which models tolerate these pipeline faults. Better rank prediction alone does not show that these tests can replace pipeline evaluation.

4.2 Model choice across architectures and tasks

Video-pretrained 3D backbones. To test whether the poor agreement between static and temporal-robustness rankings is specific to image encoders with a lightweight tx head, we fine-tune 37 video-pretrained 3D models drawn from three independent frameworks and the three pretraining datasets Kinetics-400 (K400), the motion-centric SSv2 (Goyal et al., 2017), and Kinetics-600 (K600) (Carreira et al., 2018). The set spans 3D convolutional neural networks (CNNs) and 3D Transformers and includes video masked autoencoder (VideoMAE) models ranging from 22M to 632M parameters, with clean top-1 from 46.5% to 84.1% (Table 18). For this 37-model set, the pre-specified Fisher-z critical cutoff is 0.83. Table 1 compares this set with the other settings in the study.

The static ranking works best for spatial corruption at a constant strength. For temporally coherent corruption, such as the same noise pattern or blur applied throughout a clip at a fixed strength, the static and video rankings among these 3D models agree almost perfectly: $\rho = 0.99$ with CI $[0.96, 1.00]$ wholly above ρ_0. When corruption strength gradually increases across the clip (drift), each frame still contains a spatial distortion, and the static and video rankings remain strongly correlated at $\rho = 0.88$ with CI $[0.76, 0.94]$. The motion-coupled family tests camera shake, blur that

Table 1: Static→video rank correlation ρ in the seven settings where drift, temporal-structure, and motion-coupled corruptions produce differences in robustness between models. Bold marks rejection of $\rho_0 = 0.9$ under the pre-specified rule.

Damage one frame shows		all	part		none
Setting	$\lvert\mathcal{M}\rvert$	coherent	drifting	motion-coupled	structure
Diving48, tx head	27	1.00	**0.77**	**0.81**	**0.73**
SSv2, tx head	17	1.00	**0.40**	**0.24**	**0.72**
Diving48, 3D models	37	0.99	0.88	**0.77**	**0.05**
SSv2, 3D models	37	0.99	**0.60**	0.85	**0.22**
K400, public checkpoints	27	1.00	0.93	0.85	**0.45**
SSv2, public checkpoints	16	0.99	0.90	**0.76**	**−0.47**
DAVIS VOS, frozen backbones	28	0.99	0.83	**0.66**	**0.00**
Median ρ		0.99	0.83	0.77	0.22
Settings rejecting ρ_0		0 of 7	3 of 7	5 of 7	7 of 7

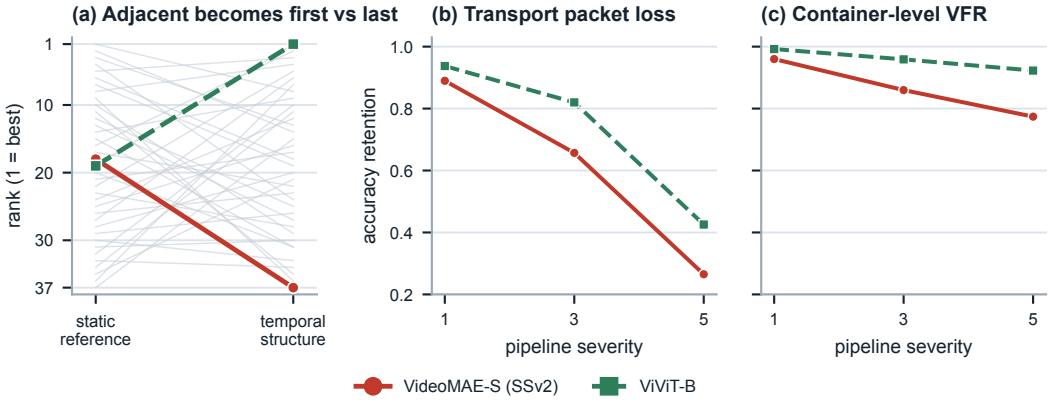


Figure 2: Two models with almost equal static scores reverse order under temporal evaluation. Among all pairs of the 37 video-pretrained 3D models with a static-score gap at most 0.005, a rule that does not use model identity selects the rank-reversing pair with the largest temporal-structure gap. (a) VideoMAE-S (SSv2) and the Base variant of the Video Vision Transformer (ViViT-B) are adjacent under the static reference, at 0.8408 and 0.8399, yet become last and first under temporal structure, at 0.7236 and 0.9204; thin lines show how the rankings of all models change. (b,c) The same pairwise ranking appears under pipeline-produced packet loss and VFR, with severity-5 retentions of 0.265 and 0.426 under packet loss and of 0.774 and 0.923 under VFR. This pair illustrates the largest reversal meeting the static-gap cutoff.

changes over time, and brightness that changes over time. Camera shake displaces the view between frames, so one frame alone cannot reveal that displacement. The correlation between the static and video rankings for this family is lower, at $\rho = 0.77 < 0.83$ with CI $[0.59, 0.88]$ entirely below $\rho_0 = 0.9$. The correlation between the static and video rankings is lowest for temporal structure: Frames are dropped, repeated, reordered, or retimed without altering the pixels of the frames that are delivered. Here, $\rho = 0.05$ with model-set CI $[-0.27, 0.34]$ and a one-sided pre-specified Fisher-z p-value below 10^{-3}. Temporal-structure retention ranges from 0.724 to 0.920 across this model set, yet the static ranking does not reliably distinguish which models retain more accuracy under these faults.

Choosing by the static score can select the less robust model. The tests reject the $\rho_0 = 0.9$ criterion for 2 of the three dynamic families, as expected from how much of each corruption a single frame can show. For temporal structure, rejection persists under every tested clean-accuracy floor and top-K restriction (Table 16).

Temporal-structure selection cost. On the full model set, selection by static robustness gives pairwise concordance 0.52, near chance (Table 4). The static-best model ranks 8th of 37 on frame-timing robustness, with selection regret 0.059. Among discordant pairs, the retention shortfall reaches 0.197; 0.48 of pairs are mis-ranked, with expected per-pair loss 0.026. This threshold-independent cost occurs on a family that a single frame cannot represent. Fig. 2 shows the largest temporal-structure reversal among pairs with almost equal static scores: VideoMAE-S (SSv2) and ViViT-B. The statically preferred model also performs worse under pipeline-produced packet loss and VFR.

The same failure appears in public checkpoints and segmentation. We evaluate two sets of public checkpoints using each checkpoint's released weights, head, and evaluation input protocol. The correlation between static and temporal-structure robustness drops to $\rho = 0.45$ over 27 K400 checkpoints and -0.47 over 16 SSv2 checkpoints, each rejecting $\rho_0 = 0.9$. On DAVIS VOS, where object masks are propagated across frames, the same correlation is 0.00 over 28 frozen backbones. Static scores thus also mis-rank public models. Across the seven settings of Table 1, the coherent control always passes, with the correlation's model-set CI above ρ_0, and temporal structure always fails the test. The two partly represented families, drift and motion coupling, fail in some settings and pass in other settings.

4.3 Improving selection with eight temporal tests

To select more robust video models, we add a temporal probe of eight tests to the static evaluation. The probe measures robustness to changes in frame order, timing, and motion that the static reference does not directly test. The probe contains the five temporal-structure and three motion-coupled operators, all evaluated at severity $q = 3$ under the video protocol. For model m, let s_m^{stat} be the static reference and s_m^{probe} the probe's retention score, computed by averaging the two family scores with equal weight using Eq. 2 within each family. Because the static and probe scores have different retention ranges, we combine the static and probe rankings,

$$s_m^{\mathrm{proxy}}(\lambda) := (1-\lambda)\,\mathrm{rank}(s_m^{\mathrm{stat}}) + \lambda\,\mathrm{rank}(s_m^{\mathrm{probe}}), \qquad \lambda \in [0,1], \tag{7}$$

where $\lambda = 0$ uses only the static ranking. We report $\lambda = 1/2$, giving the two rankings equal weight without tuning the weight to maximize pipeline rank correlation.

The probe improves prediction of packet-loss and VFR robustness. Combining the static and temporal-probe rankings with equal weight improves prediction over the static ranking alone in 4 of the 4 model-set evaluations of pipeline-level temporal corruption, with 3 gains significant at one-sided $p < 0.05$. For the 37 Diving48 models, this combination raises the rank correlation from $\rho = 0.34$ to 0.74, with one-sided p-value $<$0.001. The pooled rank correlation rises from 0.63 to 0.78. For model sets evaluated under the video protocol, the eight-condition probe alone also predicts the pipeline ranking better than the full static-corruption sweep while using roughly $1/5$ as many forward passes (Fig. 6).

5 Discussion and Conclusion

We found that image robustness predicts video robustness best when a single frame shows the corruption, but is unreliable for selecting models when frame order or timing is disrupted. The static ranking transfers almost perfectly under temporally coherent corruption, with ρ from 0.99 to 1.00, across all seven settings (Table 1). The static ranking fails the pre-specified $\rho_0 = 0.9$ criterion under temporal-structure corruption in every setting. Drift and motion-coupled corruption, which a single frame shows only in part, fail the criterion in three and five of the seven settings, respectively.

Our theory proves that single-frame evaluation alone cannot guarantee the correct video ranking: Two video corruptions can look identical to every single-frame measurement yet reverse which model is more robust (Theorem 1). Corollary 1 gives a lower bound on the resulting worst-case loss from choosing the wrong model, so collecting more single-frame measurements cannot remove this limit.

Guidelines for selecting robust video models. Select models by image robustness when the expected faults are visible in individual frames and the task is appearance-dominated. Add our proposed temporal probe (Sec. 4.3) whenever frame drop, repetition, reordering, timing changes, or packet loss is expected, or the task depends on frame order.

AI USE STATEMENT

Generative AI tools were used as assistive tools for writing and editing, mathematical proofreading and consistency checking, and code development and debugging. Any AI-assisted outputs incorporated into this work were reviewed and validated by the authors as appropriate. The authors take full responsibility for the final content of this work, including any text, claims, or artifacts produced with the assistance of generative AI.

ETHICS STATEMENT

This study collects no new data and adds no annotations of people; all datasets and pretrained checkpoints are publicly released. Those checkpoints may inherit pretraining biases, and our evaluation of nuisance corruptions makes no claim about demographic fairness or deployment safety.

REPRODUCIBILITY STATEMENT

Sec. 2 specifies the matched-pair scores and hypotheses. The substantive threshold $\rho_0 = 0.9$ and one-sided test were fixed before the rank correlations were computed; App. E.1 documents these analysis choices, and App. E.2 gives the test calibration. Theory is developed in Sec. 2 and App. C and proved in App. D.

Setup details. The appendix collects the notation (App. A), the rationale for the rank statistic and the bootstrap procedures (App. E), corruption operators (App. F.1), dataset, training, and evaluation settings (App. F.2), and model-set descriptions and complete checkpoint lists (App. G). Listing 1 provides an implementation excerpt of the temporal-mode dispatch for the photometric corruption families.

Regime details. Fig. 3 illustrates the two task regimes. App. P documents the HMDB51 replication and the overlap among the three static measurements. App. F.1 specifies the corruption operators, and Table 17 gives the individual Diving48 model scores. App. J provides the full SSv2 replication and cross-dataset comparisons in three tables.

Validation and selection details. Codec protocols and model lists are in Apps. K.1 and G. App. I compares the tx and 3D results and gives model-set sensitivity. Selection costs are reported in Sec. 4.2 for temporal structure and App. P for motion coupling. Public-checkpoint and segmentation protocols are in Apps. M and N; the full probe results and weight sweep are in App. L.

App. E.3 details the model and video clip bootstraps, and App. K.2 specifies the pipeline-level temporal-corruption protocols and controls.

REFERENCES

Anurag Arnab, Mostafa Dehghani, Georg Heigold, Chen Sun, Mario Lucic, and Cordelia Schmid. ViViT: A Video Vision Transformer. In *ICCV*, pp. 6816–6826, 2021.

Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Mojtaba Komeili, Matthew J. Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, Sergio Arnaud, Abha Gejji, Ada Martin, Francois Robert Hogan, Daniel Dugas, Piotr Bojanowski, Vasil Khalidov, Patrick Labatut, Francisco Massa, Marc Szafraniec, Kapil Krishnakumar, Yong Li, Xiaodong Ma, Sarath Chandar, Franziska Meier, Yann LeCun, Michael Rabbat, and Nicolas Ballas. V-JEPA 2: Self-Supervised Video Models Enable Understanding, Prediction and Planning. *CoRR*, abs/2506.09985, 2025.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Mach. Learn.*, 79(1-2):151–175, 2010.

Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is Space-Time Attention All You Need for Video Understanding? In *ICML*, volume 139, pp. 813–824, 2021.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging Properties in Self-Supervised Vision Transformers. In *ICCV*, pp. 9630–9640, 2021.

João Carreira and Andrew Zisserman. Quo Vadis, Action Recognition? A New Model and the Kinetics Dataset. In *CVPR*, pp. 4724–4733, 2017.

João Carreira, Eric Noland, Andras Banki-Horvath, Chloe Hillier, and Andrew Zisserman. A Short Note about Kinetics-600. *CoRR*, abs/1808.01340, 2018.

MMAction2 Contributors. OpenMMLab's Next Generation Video Understanding Toolbox and Benchmark. https://github.com/open-mmlab/mmaction2, 2020.

Alexander D'Amour, Katherine A. Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D. Hoffman, Farhad Hormozdiari, Neil Houlsby, Shaobo Hou, Ghassen Jerfel, Alan Karthikesalingam, Mario Lucic, Yi-An Ma, Cory Y. McLean, Diana Mincu, Akinori Mitani, Andrea Montanari, Zachary Nado, Vivek Natarajan, Christopher Nielson, Thomas F. Osborne, Rajiv Raman, Kim Ramasamy, Rory Sayres, Jessica Schrouff, Martin Seneviratne, Shannon Sequeira, Harini Suresh, Victor Veitch, Max Vladymyrov, Xuezhi Wang, Kellie Webster, Steve Yadlowsky, Taedong Yun, Xiaohua Zhai, and D. Sculley. Underspecification Presents Challenges for Credibility in Modern Machine Learning. *J. Mach. Learn. Res.*, 23:226:1–226:61, 2022.

Mostafa Dehghani, Yi Tay, Alexey A. Gritsenko, Zhe Zhao, Neil Houlsby, Fernando Diaz, Donald Metzler, and Oriol Vinyals. The Benchmark Lottery. *CoRR*, abs/2107.07002, 2021.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*, 2021.

Bradley Efron and Robert Tibshirani. *An Introduction to the Bootstrap*. 1993.

Haoqi Fan, Tullie Murrell, Heng Wang, Kalyan Vasudev Alwala, Yanghao Li, Yilei Li, Bo Xiong, Nikhila Ravi, Meng Li, Haichuan Yang, Jitendra Malik, Ross B. Girshick, Matt Feiszli, Aaron Adcock, Wan-Yen Lo, and Christoph Feichtenhofer. PyTorchVideo: A Deep Learning Library for Video Understanding. In *ACM Multimedia*, pp. 3783–3786, 2021a.

Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale Vision Transformers. In *ICCV*, pp. 6804–6815, 2021b.

Christoph Feichtenhofer. X3D: Expanding Architectures for Efficient Video Recognition. In *CVPR*, pp. 200–210, 2020.

Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. SlowFast Networks for Video Recognition. In *ICCV*, pp. 6201–6210, 2019.

Joseph Fioresi, Ishan Rajendrakumar Dave, and Mubarak Shah. ALBAR: Adversarial Learning approach to mitigate Biases in Action Recognition. In *ICLR*, 2025.

Ronald A Fisher. On the" probable error" of a coefficient of correlation deduced from a small sample. *Metron*, 1:3–32, 1921.

Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In *ICLR*, 2019.

Phillip Good. *Permutation, parametric and bootstrap tests of hypotheses*. 2005.

Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fründ, 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 *ICCV*, pp. 5843–5851, 2017.

Ishaan Gulrajani and David Lopez-Paz. In Search of Lost Domain Generalization. In *ICLR*, 2021.

Dan Hendrycks and Thomas G. Dietterich. Benchmarking Neural Network Robustness to Common Corruptions and Perturbations. In *ICLR*, 2019.

Dan Hendrycks, Norman Mu, Ekin Dogus Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. AugMix: A Simple Data Processing Method to Improve Robustness and Uncertainty. In *ICLR*, 2020.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The Many Faces of Robustness: A Critical Analysis of Out-of-Distribution Generalization. In *ICCV*, pp. 8320–8329, 2021.

Wassily Hoeffding. A Class of Statistics with Asymptotically Normal Distribution. *The Annals of Mathematical Statistics*, 19(3):293–325, 1948. ISSN 0003-4851. doi: 10.1214/aoms/1177730196. URL http://dx.doi.org/10.1214/aoms/1177730196.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, pp. 65–70, 1979.

De-An Huang, Vignesh Ramanathan, Dhruv Mahajan, Lorenzo Torresani, Manohar Paluri, Li Fei-Fei, and Juan Carlos Niebles. What Makes a Video a Video: Analyzing Temporal Information in Video Understanding Models and Datasets. In *CVPR*, pp. 7366–7375, 2018.

Allan Jabri, Andrew Owens, and Alexei A. Efros. Space-Time Correspondence as a Contrastive Random Walk. In *NeurIPS*, 2020.

Abigail Z. Jacobs. Measurement and Fairness. In *FAccT*, pp. 375–385, 2021.

Junchen Jiang, Ganesh Ananthanarayanan, Peter Bodík, Siddhartha Sen, and Ion Stoica. Chameleon: scalable adaptation of video analytics. In *SIGCOMM*, pp. 253–266, 2018.

Yiding Jiang, Behnam Neyshabur, Hossein Mobahi, Dilip Krishnan, and Samy Bengio. Fantastic Generalization Measures and Where to Find Them. In *ICLR*, 2020.

Christoph Kamann and Carsten Rother. Benchmarking the Robustness of Semantic Segmentation Models with Respect to Common Corruptions. *Int. J. Comput. Vis.*, 129(2):462–483, 2021.

Oguzhan Fatih Kar, Teresa Yeo, Andrei Atanov, and Amir Zamir. 3D Common Corruptions and Data Augmentation. In *CVPR*, pp. 18941–18952, 2022.

Will Kay, João 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. *CoRR*, abs/1705.06950, 2017.

M. G. KENDALL. A NEW MEASURE OF RANK CORRELATION. *Biometrika*, 30(1-2):81–93, 1938. ISSN 1464-3510. doi: 10.1093/biomet/30.1-2.81. URL http://dx.doi.org/10.1093/biomet/30.1-2.81.

M. G. KENDALL. THE TREATMENT OF TIES IN RANKING PROBLEMS. *Biometrika*, 33(3): 239–251, 1945. ISSN 1464-3510. doi: 10.1093/biomet/33.3.239. URL http://dx.doi.org/10.1093/biomet/33.3.239.

Simon Kornblith, Jonathon Shlens, and Quoc V. Le. Do Better ImageNet Models Transfer Better? In *CVPR*, pp. 2661–2671, 2019.

Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso A. Poggio, and Thomas Serre. HMDB: A large video database for human motion recognition. In *ICCV*, pp. 2556–2563, 2011.

Ananya Kumar, Aditi Raghunathan, Robbie Matthew Jones, Tengyu Ma, and Percy Liang. Fine-Tuning can Distort Pretrained Features and Underperform Out-of-Distribution. In *ICLR*, 2022.

Mengtian Li, Yu-Xiong Wang, and Deva Ramanan. Towards Streaming Perception. In *ECCV (2)*, volume 12347, pp. 473–488, 2020a.

Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. MViTv2: Improved Multiscale Vision Transformers for Classification and Detection. In *CVPR*, pp. 4794–4804, 2022.

Yingwei Li, Yi Li, and Nuno Vasconcelos. RESOUND: Towards Action Recognition Without Representation Bias. In *ECCV (6)*, volume 11210, pp. 520–535, 2018.

Yuanqi Li, Arthi Padmanabhan, Pengzhan Zhao, Yufei Wang, Guoqing Harry Xu, and Ravi Netravali. Reducto: On-Camera Filtering for Resource-Efficient Real-Time Video Analytics. In *SIGCOMM*, pp. 359–376, 2020b.

Thomas Liao, Rohan Taori, Inioluwa Deborah Raji, and Ludwig Schmidt. Are We Learning Yet? A Meta Review of Evaluation Failures Across Machine Learning. In *NeurIPS Datasets and Benchmarks*, 2021.

Ji Lin, Chuang Gan, and Song Han. TSM: Temporal Shift Module for Efficient Video Understanding. In *ICCV*, pp. 7082–7092, 2019.

Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video Swin Transformer. In *CVPR*, pp. 3192–3201, 2022.

Claudio Michaelis, Benjamin Mitzkus, Robert Geirhos, Evgenia Rusak, Oliver Bringmann, Alexander S. Ecker, Matthias Bethge, and Wieland Brendel. Benchmarking Robustness in Object Detection: Autonomous Driving when Winter is Coming. *CoRR*, abs/1907.07484, 2019.

John Miller, Rohan Taori, Aditi Raghunathan, Shiori Sagawa, Pang Wei Koh, Vaishaal Shankar, Percy Liang, Yair Carmon, and Ludwig Schmidt. Accuracy on the Line: on the Strong Correlation Between Out-of-Distribution and In-Distribution Generalization. In *ICML*, volume 139, pp. 7721–7735, 2021.

Eric Mintun, Alexander Kirillov, and Saining Xie. On Interaction Between Augmentations and Corruptions in Natural Corruption Robustness. In *NeurIPS*, pp. 3571–3583, 2021.

Federico Perazzi, Jordi Pont-Tuset, Brian McWilliams, Luc Van Gool, Markus H. Gross, and Alexander Sorkine-Hornung. A Benchmark Dataset and Evaluation Methodology for Video Object Segmentation. In *CVPR*, pp. 724–732, 2016.

Jordi Pont-Tuset, Federico Perazzi, Sergi Caelles, Pablo Arbeláez, Alexander Sorkine-Hornung, and Luc Van Gool. The 2017 DAVIS Challenge on Video Object Segmentation. *CoRR*, abs/1704.00675, 2017.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*, volume 139, pp. 8748–8763, 2021.

Inioluwa Deborah Raji, Emily Denton, Emily M. Bender, Alex Hanna, and Amandalynne Paullada. AI and the Everything in the Whole Wide World Benchmark. In *NeurIPS Datasets and Benchmarks*, 2021.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet Classifiers Generalize to ImageNet? In *ICML*, volume 97, pp. 5389–5400, 2019.

Madeline Chantry Schiappa, Naman Biyani, Prudvi Kamtam, Shruti Vyas, Hamid Palangi, Vibhav Vineet, and Yogesh S. Rawat. A Large-Scale Robustness Analysis of Video Action Recognition Models. In *CVPR*, pp. 14698–14708, 2023.

Henning Schulzrinne, Stephen L. Casner, Ron Frederick, and Van Jacobson. RTP: A Transport Protocol for Real-Time Applications. *RFC*, 3550:1–104, 2003.

Kalpana Seshadrinathan, Rajiv Soundararajan, Alan Conrad Bovik, and Lawrence K. Cormack. Study of Subjective and Objective Quality Assessment of Video. *IEEE Trans. Image Process.*, 19(6): 1427–1441, 2010.

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 *WACV*, pp. 535–544, 2021.

Ziyao Shangguan, Chuhan Li, Yuxuan Ding, Yanan Zheng, Yilun Zhao, Tesca Fitzgerald, and Arman Cohan. TOMATO: Assessing Visual Temporal Reasoning Capabilities in Multimodal Foundation Models. In *ICLR*, 2025.

Vaishaal Shankar, Achal Dave, Rebecca Roelofs, Deva Ramanan, Benjamin Recht, and Ludwig Schmidt. Do Image Classifiers Generalize Across Time? In *ICCV*, pp. 9641–9649, 2021.

Karen Simonyan and Andrew Zisserman. Two-Stream Convolutional Networks for Action Recognition in Videos. In *NIPS*, pp. 568–576, 2014.

M Sklar. Fonctions de répartition à n dimensions et leurs marges. In *Annales de l'ISUP*, volume 8, pp. 229–231, 1959.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A Dataset of 101 Human Actions Classes From Videos in The Wild. *CoRR*, abs/1212.0402, 2012.

C. Spearman. The Proof and Measurement of Association between Two Things. *The American Journal of Psychology*, 100(3/4):441, 1987. ISSN 0002-9556. doi: 10.2307/1422689. URL http://dx.doi.org/10.2307/1422689.

James H. Steiger. Tests for comparing elements of a correlation matrix. *Psychological Bulletin*, 87(2):245–251, March 1980. ISSN 0033-2909. doi: 10.1037/0033-2909.87.2.245. URL http://dx.doi.org/10.1037/0033-2909.87.2.245.

Thomas Stockhammer, Miska M. Hannuksela, and Thomas Wiegand. H.264/AVC in wireless environments. *IEEE Trans. Circuits Syst. Video Technol.*, 13(7):657–673, 2003.

Gary J. Sullivan, Jens-Rainer Ohm, Woojin Han, and Thomas Wiegand. Overview of the High Efficiency Video Coding (HEVC) Standard. *IEEE Trans. Circuits Syst. Video Technol.*, 22(12): 1649–1668, 2012.

Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring Robustness to Natural Distribution Shifts in Image Classification. In *NeurIPS*, 2020.

Damien Teney, Yong Lin, Seong Joon Oh, and Ehsan Abbasnejad. ID and OOD Performance Are Sometimes Inversely Correlated on Real-world Datasets. In *NeurIPS*, 2023.

Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training. In *NeurIPS*, 2022.

Antonio Torralba and Alexei A. Efros. Unbiased look at dataset bias. In *CVPR*, pp. 1521–1528, 2011.

Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A Closer Look at Spatiotemporal Convolutions for Action Recognition. In *CVPR*, pp. 6450–6459, 2018.

Du Tran, Heng Wang, Matt Feiszli, and Lorenzo Torresani. Video Classification With Channel-Separated Convolutional Networks. In *ICCV*, pp. 5551–5560, 2019.

Cédric Villani. *Optimal Transport*. 2009. ISBN 9783540710509. doi: 10.1007/978-3-540-71050-9. URL http://dx.doi.org/10.1007/978-3-540-71050-9.

Carl Vondrick, Abhinav Shrivastava, Alireza Fathi, Sergio Guadarrama, and Kevin Murphy. Tracking Emerges by Colorizing Videos. In *ECCV (13)*, volume 11217, pp. 402–419, 2018.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno A. Olshausen, and Trevor Darrell. Tent: Fully Test-Time Adaptation by Entropy Minimization. In *ICLR*, 2021.

Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal Segment Networks: Towards Good Practices for Deep Action Recognition. In *ECCV (8)*, volume 9912, pp. 20–36, 2016.

Yao Wang and Qin-Fan Zhu. Error control and concealment for video communication: a review. *Proc. IEEE*, 86(5):974–997, 1998.

Thomas Wiegand, Gary J. Sullivan, Gisle Bjøntegaard, and Ajay Luthra. Overview of the H.264/AVC video coding standard. *IEEE Trans. Circuits Syst. Video Technol.*, 13(7):560–576, 2003.

Ross Wightman. PyTorch Image Models. https://github.com/rwightman/pytorch-image-models, 2019.

E. J. Williams. The Comparison of Regression Variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 21(2):396–399, 1959. ISSN 1467-9868. doi: 10.1111/j.2517-6161.1959.tb00346.x. URL http://dx.doi.org/10.1111/j.2517-6161.1959.tb00346.x.

Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking Spatiotemporal Feature Learning: Speed-Accuracy Trade-offs in Video Classification. In *ECCV (15)*, volume 11219, pp. 318–335, 2018.

Taojiannan Yang, Yi Zhu, Yusheng Xie, Aston Zhang, Chen Chen, and Mu Li. AIM: Adapting Image Models for Efficient Video Action Recognition. In *ICLR*, 2023.

Chenyu Yi, Siyuan Yang, Haoliang Li, Yap-Peng Tan, and Alex C. Kot. Benchmarking the Robustness of Spatial-Temporal Models Against Corruptions. In *NeurIPS Datasets and Benchmarks*, 2021.

Chenyu Yi, Siyuan Yang, Yufei Wang, Haoliang Li, Yap-Peng Tan, and Alex C. Kot. Temporal Coherent Test Time Optimization for Robust Video Classification. In *ICLR*, 2023.

Dong Yin, Raphael Gontijo Lopes, Jonathon Shlens, Ekin Dogus Cubuk, and Justin Gilmer. A Fourier Perspective on Model Robustness in Computer Vision. In *NeurIPS*, pp. 13255–13265, 2019.

Runhao Zeng, Xiaoyong Chen, Jiaming Liang, Huisi Wu, Guangzhong Cao, and Yong Guo. Benchmarking the Robustness of Temporal Action Detection Models Against Temporal Corruptions. In *CVPR*, pp. 18263–18274, 2024.

Bingqing Zhang, Zhuo Cao, Heming Du, Yang Li, Xue Li, Jiajun Liu, and Sen Wang. Robust Test-time Video-Text Retrieval: Benchmarking and Adapting for Query Shifts. *CoRR*, abs/2604.20851, 2026.

Bolei Zhou, Alex Andonian, Aude Oliva, and Antonio Torralba. Temporal Relational Reasoning in Videos. In *ECCV (1)*, volume 11205, pp. 831–846, 2018.

APPENDIX TABLE OF CONTENTS

A List of notation **17**

B Related work **17**

C Mechanism: when temporal context helps or hurts **20**

D Proofs of the theoretical results **23**

E Inference protocol and uncertainty **26**

F Implementation details **28**

G Evaluated model sets **31**

H Robustness of the motion-regime rank divergence **32**

I Video-pretrained 3D backbones: extended analysis **33**

J SSv2 replication and cross-dataset consistency **34**

K Cross-validation against pipeline-produced corruptions **36**

L Prescription details **40**

M Off-the-shelf operating-point replications **42**

N Dense online regime: VOS label-propagation protocol **44**

O Scope and limitations **44**

P Appearance-regime results, collinearity, and motion-coupling cost **46**

A LIST OF NOTATION

Table 2 collects the symbols used throughout the paper. One naming convention runs throughout: We call s^{stat} the image-style static reference across regimes and reserve single-frame for a literal $T = 1$ measurement, since the image-style static reference describes the operator family and need not be a one-frame protocol.

What the paired scores measure. The static reference score measures robustness to frame-independent spatial corruption, exactly the quantity an ImageNet-C-style benchmark reports, and the video scores measure robustness under each family of a temporal corruption taxonomy. One consequence of that definition governs how the pair is read. Where a single frame carries the class, the static reference is the literal $T = 1$ image score. Where one frame does not, as on a motion task, the static reference is the frame-independent score read under the video protocol, which applies the identical frame-by-frame spatial operator to the identical video clips at the identical severities and so hands the proxy the model, the video clips, and the head of the video score. In the appearance regime, where both the single-frame and frame-independent video scores are available, the two scores are collinear across models at $\rho = 0.98$, so the substitution carries the image measurement into the motion regime rather than replacing that measurement (Sec. 2). Because the paired evaluations of each model use identical weights, the divergence cannot be attributed to swapping models or weights between measurements; the divergence reflects how the fixed model zoo interacts with the changed input and corruption protocol. The design combines a pre-specified threshold, explicit rejection calibration, dual uncertainty over model sets and video clips, rank-scale dependent-correlation comparisons, and a representability rule fixed in advance. That rule holds that single-image robustness predicts robustness to a corruption to the extent one frame can represent that corruption, so the rule names the cells where the proxy lacks direct measurement before those cells are measured.

B RELATED WORK

Single-image corruption robustness. Hendrycks & Dietterich (2019) introduced ImageNet-C and ImageNet-P and standardized corruption type, severity, and the relative-drop metric; the paradigm was extended to detection (Michaelis et al., 2019) and segmentation (Kamann & Rother, 2021). The benchmark also became the yardstick for interventions. Geirhos et al. (2019) showed that moving an ImageNet model from texture toward shape cues raises the model's corruption robustness, and AugMix (Hendrycks et al., 2020) became the standard augmentation for the ImageNet-C score; both interventions are defined and scored one image at a time. Two parallel lines of work refine the spatial operators. Kar et al. (2022) argue that image-space corruptions ignore the scene's 3D structure and rebuild those corruptions to respect scene geometry, a correction that stays inside the single-image regime and leaves the temporal axis untouched. Yin et al. (2019) characterize corruption sensitivity in the frequency domain and show that robustness gained on one band is often paid for on another, and Mintun et al. (2021) report that ImageNet-C robustness drops sharply on a perceptually dissimilar corruption set. Both concerns are about which spatial operators a benchmark covers and are orthogonal to the axis we test. These benchmarks evaluate each image, or equivalently each frame, and do not model how errors aggregate, persist, or recover over time. Our study takes the single-image evaluation paradigm as the object of validation.

Benchmark validity and the external validity of robustness measures. Whether a cheap benchmark's score predicts the deployment behavior that benchmark stands in for is a question of external validity, close to the construct validity that measurement modeling makes explicit (Jacobs, 2021), and a growing measurement literature has pressed that question on machine learning. Torralba & Efros (2011) exposed dataset bias as a threat to cross-dataset generalization; Recht et al. (2019) showed that ImageNet accuracy drops under a faithful test-set replication, evidence that headline numbers can overfit the measurement instrument; and Raji et al. (2021) and Liao et al. (2021) argue that many benchmarks lack the construct validity to support the general capability claims made for those benchmarks. Closest to our framing, Dehghani et al. (2021) show that swapping the benchmark alone can change the ranking of a field's methods, and Gulrajani & Lopez-Paz (2021) show that in domain generalization the choice of model-selection criterion alone can change which algorithm leads, so a ranking is a property of the evaluation protocol as much as of the models ranked; that ranking-level failure mode is the one we test along the temporal axis.

Table 2: List of notation. Symbols are grouped by the setup that introduces each symbol: evaluation protocol and data, robustness scores, the external-validity rank statistic and the associated test calibration, corruption channels and risk, and the path-transport and temporal-interaction quantities of App. C. A few symbols are reused with an explicit local scope: The symbol q is the corruption severity in Sec. 2 and the per-frame flip rate in Theorem 1, and A is the marginal-only selector of Corollary 1, which is unrelated to the linearized transition matrix A_t of App. D.1.

Symbol	Meaning
Models, data, and evaluation protocol	
$\mathcal{M}$, n	Evaluated model set and model count $n := \lvert\mathcal{M}\rvert$
m	A model with fixed weights across the paired protocols
p	Evaluation protocol: $T = 1$ image, or multi-frame video
$X := (x_1, \ldots, x_T)$, Y	Clean T-frame video clip and corresponding label
$\widetilde{X}$	Corrupted video clip produced by a corruption channel
f	Corruption family; one of the five taxonomy families
c, q	Corruption type and severity level
P	Task metric: top-1 accuracy, or DAVIS region similarity $\mathcal{J}$
Robustness scores	
$P_{m,p,\text{clean}}$, $P_{m,p,c,q}$	Clean and corrupted performance of m under protocol p
$g(m,p,c,q)$	Retention $P_{m,p,c,q}/P_{m,p,\text{clean}}$; the robustness score
$s^{(f)}_{m,p}$	Group-balanced family score, Eq. 2
s^{img}	Literal single-frame ($T = 1$) image score
s^{stat}	Image-style static reference: the frame-independent score
$s^{\text{vid}(f)}$	Video score of family f; the prediction target
s^{probe}	One-severity battery over the non-representable families
$s^{\text{proxy}}(\lambda)$	Rank blend of s^{stat} and s^{probe}, Eq. 7
External-validity statistic and test calibration	
$\rho^{(f)}$	Observed Spearman correlation of s^{stat} and $s^{\text{vid}(f)}$
$\rho^{(f)}_*$	Population rank association that the test targets
$\rho_0 = 0.9$	Pre-specified proxy-fidelity threshold
ρ_{crit}	Critical observed-correlation cutoff of the Fisher-z rule
$\mathcal{H}^{(f)}_0$, $\mathcal{H}^{(f)}_A$	Null $\rho^{(f)}_* \geq \rho_0$ against alternative $\rho^{(f)}_* < \rho_0$
$\Delta\rho$, T_2	Dependent-correlation gap and Williams test statistic
Corruption channels and risk	
C, C'	Video corruption channels, mapping X to $\widetilde{X}$
$R_C(m)$	Risk of model m under channel C
$K_{C,t}$	Per-frame coordinate kernel of C at time t
C_{coh}, C_{anti}, C_{ind}	Coherent, anti-coherent, and independent flip channels
h_+, h_-, A	Two classifiers of Theorem 1; a marginal-only selector
Path transport and temporal interaction	
z_t, $z_{1:T}$	Corrupted frame at t and path, in frame space $(\mathcal{Z}, d)$
$h_t := F_t(h_{t-1}, z_t)$	Model state at time t and state-update map
α_t, β_t	State- and input-Lipschitz constants of F_t
ℓ_t, L_t, q_t	Per-time loss, loss Lipschitz constant, and loss weight
$\mathcal{L}_{m,x}$	Video clip loss $\sum_t q_t \ell_t(h_t, y)$
$\omega_{m,s}$	Influence of frame s on the video clip loss, Eq. 9
$\Pi(P,Q)$, P_x, Q_x	Couplings of P, Q; path laws of C, C' given x
W_m, $B_m(C,C')$	Path-transport discrepancy and average over clean inputs
$\ell_m(z,y)$	Bounded video clip loss of model m on corrupted path z
$\Gamma_m(y)$	Residual oscillation of ℓ_m after removing additivity
d_{TV}, $\delta(x)$	Total-variation distance and the channel gap at x
ε_m	Interaction transfer bound $\mathbb{E}[\Gamma_m(Y)\delta(X)]$

Transport of a cheap benchmark's model ranking. Most directly, Taori et al. (2020) tested whether synthetic corruption-robustness interventions predict robustness to natural distribution shift, the image-domain analogue of our inquiry, and found that those interventions largely do not. Hendrycks et al. (2021) qualify that verdict: Artificial-benchmark gains do transfer to some natural shifts, yet no method that study evaluates helps across every shift type. Transfer is therefore

conditional on the shift, the shape of the answer we obtain along the temporal axis. Closest to our setting along the temporal axis, Shankar et al. (2021) showed that ImageNet classifiers lose substantial accuracy on nearby, perceptually similar video frames, establishing video as a source of natural perturbation, and D'Amour et al. (2022) supply a complementary mechanism, since models a benchmark cannot distinguish can behave very differently under shift. Pulling the other way, Miller et al. (2021) document a strong, often near-linear, relation between in-distribution and out-of-distribution accuracy across many image shifts and model families, so a cheap benchmark's model ranking frequently does transport, whereas Teney et al. (2023) exhibit datasets on which the two accuracies are instead inversely correlated and attribute earlier positive reports partly to a biased selection of models. Outside robustness entirely, Kornblith et al. (2019) correlate ImageNet accuracy with downstream accuracy over a fixed model zoo, an association that is strong in aggregate but weakens as the target task departs from the source benchmark. In the same spirit, Jiang et al. (2020) validate more than forty candidate generalization measures by the measures' rank correlation with the generalization gap those measures are meant to predict, over thousands of trained networks. That model-zoo rank correlation between a cheap score and an expensive target is precisely the statistic we adopt, computed across the temporal axis rather than across downstream datasets or training configurations. We also add what these image-domain audits do not have: a matched-pair design that holds weights fixed and thereby rules out changing weights as the source of the rank change, a pre-specified rejection threshold, and an a priori representability rule that identifies corruptions the proxy does not directly measure and whose video rankings the proxy may fail to predict. Our design holds model weights fixed while measuring how a model ranking transfers from image to video evaluation, separating static and temporal transfer. We observed that temporally coherent corruption preserves the static ranking, a positive result that the video-specific character of coherent corruption would not predict, while the static score does not rank video-pretrained 3D backbones by frame-timing robustness at all. Both statements are quantities that a design varying model identity between the two measurements cannot produce.

Video corruption robustness. Yi et al. (2021) proposed Mini Kinetics-C and Mini SSV2-C and compared the spatial and temporal corruption robustness of video classifiers; Yi et al. (2023) improved corrupted-video classification with temporal-coherence test-time optimization, carrying fully test-time adaptation (Wang et al., 2021) from images to video clips; Zeng et al. (2024) showed temporal action detection is fragile under temporal corruptions, with a larger performance drop in localization than in classification; and Zhang et al. (2026) extended corruption benchmarking to video–text retrieval with twelve video perturbations at five severities, arguing that image-focused robustness methods miss the spatio-temporal dynamics of those shifts. Closest in scale, Schiappa et al. (2023) evaluated six action-recognition models against 90 perturbations on four corrupted benchmarks and found temporal perturbations largely harmless except on SSv2, a task dependence that our contrast between appearance- and motion-dominated regimes turns into a controlled variable. These works characterize video-model robustness or compare defenses and architectures. Our inquiry and design are different in kind: We hold the weights fixed and test transfer of the single-image ranking using rank correlation between matched static and temporal scores as the primary endpoint, across both appearance- and motion-dominated tasks. The matched-pair, external-validity framing, together with the explicit separation of an absent-signal null from a consequential vulnerability the image-style static reference cannot represent, distinguishes this study from corrupted-video benchmarking.

Temporal tasks and video architectures. Motion-dependent recognition benchmarks such as SSv2 (Goyal et al., 2017) and Diving48 (Li et al., 2018) were built so that temporal order, not static appearance, carries the label; we use Diving48 precisely because that benchmark makes temporally dynamic corruptions bite. Huang et al. (2018) isolated how much of a video model's accuracy depends on temporal information rather than static appearance, and Sevilla-Lara et al. (2021) quantified how rarely the classes of standard video datasets require temporal information at all. The same diagnosis has since been pressed from two directions. Fioresi et al. (2025) suppress static bias in action recognition by training adversarially against a static video clip in which every frame is the same frame, the limiting case of our frame-repeat operator. Shangguan et al. (2025) report that many questions in the video benchmarks used for multimodal foundation models can be answered from a single frame, a few frames, or frames out of order, and propose multi-frame gain and frame-order sensitivity as the metrics that expose that shortcut. That task-level temporal dependence is the axis along which our two regimes are separated. Perazzi et al. (2016) established VOS benchmarks with temporal coherence metrics beyond frame-wise scores, which inform our temporal diagnostics.



Figure 3: The two regimes, on eight frames spanning each video clip. (a) A UCF101 video clip from appearance-dominated recognition, where a single frame already signals the class, which is why the temporally dynamic cell there is low-signal and non-diagnostic rather than a proxy failure (App. P.1). (b,c) The identical eight frames of a Diving48 test video clip, differing only in the frames' order, forward and reversed. No single frame of either row fixes the action, and only one of the two orders is physically possible, so the label is carried by the arrangement rather than by any frame. That pair is the empirical shape of Theorem 1, where two channels that agree on every per-frame marginal still differ in what the video clip encodes. The digit under each frame in (c) gives the position that frame occupies in (b), so (c) reads as a permutation of (b).

On the architecture side, two-stream networks with an explicit optical-flow branch (Simonyan & Zisserman, 2014), Temporal Segment Networks (TSNs) (Wang et al., 2016), space-time attention (Bertasius et al., 2021), SlowFast (Feichtenhofer et al., 2019), video self-supervised pretraining (Tong et al., 2022), and adapters that add temporal attention to a frozen image backbone (Yang et al., 2023) provide temporal mechanisms such as motion estimation, averaging, and memory. We treat those mechanisms not as performance levers but as sources of an image-to-video robustness gap: Each temporal mechanism provides a route by which a perturbation that a single frame would localize is instead aggregated, propagated, or absorbed across the video clip.

Rationale for a temporally engaged task. On appearance-dominated recognition, a single frame already signals the class (Sevilla-Lara et al., 2021), so temporally dynamic corruptions such as drift, frame drop or repeat, and motion coupling barely hurt and barely discriminate models; a low image-to-video correlation there is an absence of signal rather than a proxy failure, so we treat the cell as low-signal and non-diagnostic (App. P.1). A model-discriminating test requires a task that uses the temporal stream. We therefore evaluate Diving48 (Li et al., 2018), a fine-grained dive-recognition benchmark explicitly built to suppress static scene and object bias, so that strong performance requires aggregating information across many frames, including takeoff, somersault and twist counts, flight, and entry. There, dynamic corruptions become consequential, often severe vulnerabilities, and the proxy comparison becomes model-discriminating (Sec. 3). To show this result is not a property of one dataset, we replicate the evaluation on a second, more strongly order-dependent benchmark, SSv2 (Goyal et al., 2017), whose classes are defined by the direction and order of motion and where an order-invariant baseline performs poorly (App. J). Fig. 3 shows one video clip from each of the two regimes.

C Mechanism: When Temporal Context Helps or Hurts

This appendix complements Theorem 1 and Corollary 1 in Sec. 2. Together, the results bound how pathwise coupling changes propagate through a stateful model, identify what no marginal-only proxy can recover without assumptions on cross-frame coupling, and characterize exactly how loss interactions bound that coupling effect and when marginals are exactly sufficient on finite alphabets. The results supply a formal boundary for the empirical representability principle.

An impossibility boundary behind H1. The failure that H1 in Sec. 2 expects under temporal structure is not only empirical, and the boundary governs how the empirical result should be read. Call an audit marginal-only when the audit's information about a corruption channel is unchanged whenever every per-frame kernel is unchanged, which is exactly the information a single-image benchmark collects. Theorem 1 constructs channels that agree on every per-frame kernel yet reverse which of two video models has the lower risk, so the optimal ranking is not identifiable from frame marginals, and Corollary 1 turns that into a minimax lower bound of $(4q-1)/(4q)$ on the selection regret of every marginal-only selector, reaching $1/2$ for randomized and 1 for deterministic selectors at $q = 1/2$. Enlarging a single-image benchmark cannot repair this failure at any measurement budget, because the missing quantity is the cross-frame coupling that no per-frame measurement varies. The boundary is two-sided rather than only negative: Theorem 4 shows that frame marginals are universally sufficient exactly for frame-additive video clip losses, and Theorem 3 bounds the risk gap between same-marginal channels by the loss's non-additive temporal interaction, which is the positive side that the coherent families realize. Because same-marginal channels can reverse a ranking, the coherent transfer observed in all seven settings of Table 1 is an empirical finding and not a logical necessity, whereas a temporal-structure fault delivers only clean frames and so leaves the input of a single-image benchmark unchanged. These results are existence and worst-case statements, and Secs. 3 and 4 measure how often the failure occurs in the model zoos we evaluate and what the failure costs.

C.1 Memory-based error propagation

For the linear memory $m_t := Am_{t-1} + Bz_t$, perturbations obey $e_t = Ae_{t-1} + B\delta z_t$, and more generally a local time-varying linearization around a clean trajectory has $e_t = A_t e_{t-1} + \eta_t$. App. D.1 solves this recursion through the recursion's transition products, records the resulting norm bound on $\|e_t\|$, and states what contractivity, the spectral radius, and nonnormal transient amplification do and do not imply about the persistence of an impulse.

The linear calculation concerns one realized perturbation. The next theorem upgrades this pathwise calculation to a risk-level comparison between two full corruption channels, allowing nonlinear time-varying state updates and either a terminal video clip loss or an online average. Let $\Pi(P, Q)$ denote the set of couplings of probability laws P and Q, in the sense of optimal transport (Villani, 2009). For this result, let $(\mathcal{Z}, d)$ be a metric frame space and, for a fixed model m, let $h_0 := h_0(x)$ and $h_t := F_t(h_{t-1}, z_t)$. Suppose that, for constants $\alpha_t, \beta_t \geq 0$,

$$\|F_t(h, z) - F_t(h', z')\| \leq \alpha_t \|h - h'\| + \beta_t d(z, z') \qquad (t = 1, \ldots, T). \tag{8}$$

The factors α_t need not be below one; the result uses only the Lipschitz inequality of Eq. 8, not a contraction assumption. For a fixed clean video clip x, let the video clip loss be $\mathcal{L}_{m,x}(z_{1:T}, y) := \sum_{t=1}^{T} q_t \ell_t(h_t, y)$, where $q_t \geq 0$, $\sum_t q_t = 1$, and each $\ell_t(\cdot, y)$ is L_t-Lipschitz uniformly in y. Define the influence of frame s by

$$\omega_{m,s} := \beta_s \sum_{t=s}^{T} q_t L_t \prod_{u=s+1}^{t} \alpha_u, \tag{9}$$

with an empty product equal to one.

For two corruption channels C, C' under this stateful-model setup, conditioned only on the clean video clip X and acting under a common clean law for (X, Y), write the conditional path laws of the two channels as $P_x := C(\cdot \mid x)$ and $Q_x := C'(\cdot \mid x)$ and the risks as $R_C(m) := \mathbb{E}[\mathcal{L}_{m,X}(\widetilde{X}, Y)]$. Define the weighted path-transport discrepancy and the average discrepancy over clean inputs by

$$\mathsf{W}_m(P_x, Q_x) := \inf_{\pi \in \Pi(P_x, Q_x)} \mathbb{E}_\pi \left[\sum_{s=1}^{T} \omega_{m,s} d(Z_s, Z'_s) \right], \qquad B_m(C, C') := \mathbb{E}_X \mathsf{W}_m(P_X, Q_X). \tag{10}$$

Theorem 2 (Path-transport certificate for a stateful video model)**.** *Whenever the preceding expectations are finite, the channel risks satisfy*

$$|R_C(m) - R_{C'}(m)| \leq B_m(C, C'). \tag{11}$$

Theorem 2 combines the deterministic propagation calculation with the channel comparison. The bound takes the shape familiar from domain-adaptation theory, where a risk difference between two

distributions is controlled by a divergence between those distributions (Ben-David et al., 2010); what is specific here is that the divergence is a path-level transport cost whose per-frame weights $\omega_{m,s}$ are supplied by the model's own memory. For two channels with the same frame marginals, the theorem isolates the missing object as a transport discrepancy between the two full paths; that discrepancy cannot in general be read from the marginals. The constants are global upper bounds and can be conservative for a deep network, so the result is a certificate rather than an estimator of the empirical risk gap.

A memory can introduce transient-response and recovery dynamics absent in a stateless per-frame map: As a function of the transition products and the direction and timing of the impulses, a memory can attenuate an impulse or carry and amplify that impulse. Contractive post-corruption products can yield recovery under the condition made precise in App. D.1.

C.2 Rank reversal under matched frame marginals

Theorem 1 and Corollary 1 in Sec. 2 establish the marginal-only audit's identification and selection limits. Fixing the coordinate kernels defined there fixes only the frame marginals and leaves the coupling free, which is the classical fact that a joint law is pinned down by the coordinate margins only together with a copula (Sklar, 1959). The following corollary makes the resulting rank reversal explicit.

Corollary 2 (Rank reversal under matched frame marginals)**.** *In the construction of Theorem 1, for $q < 1/2$, the independent per-frame benchmark ranks h_+ above h_-, whereas the per-frame-matched channel C_{anti} ranks h_- above h_+. At $q = 1/2$, C_{coh} and C_{anti} reverse the two risks from $(0, 1)$ to $(1, 0)$. The construction extends to every $T \geq 2$ by adding frames ignored by the classifiers.*

Theorem 1 and Corollaries 2 and 1 give identifiability limits for ranking and selection, not a claim that every dynamic corruption must break every static proxy. Additional structure can couple marginal and temporal robustness; indeed, the coherent families empirically preserve the static ranking. What these results rule out is a universal guarantee based only on per-frame matching. The empirical rank tests in Secs. 3 and 4 measure the effect of the missing temporal coupling on rankings in realistic model zoos, while the pipeline-level cross-validation evaluates recovery by a targeted video-aware measurement.

C.3 Conditions for frame-marginal sufficiency

The impossibility theorem establishes that temporal coupling can matter, but not what property of a particular model makes that coupling matter or how large the resulting error can be. The next result supplies that positive, quantitative boundary, and the following theorem isolates the exact finite-alphabet case. For a bounded function f, write $\operatorname{osc}(f) \coloneqq \sup f - \inf f$, and let the total-variation distance d_{TV} use the convention $d_{\mathrm{TV}}(P, Q) \coloneqq \sup_A |P(A) - Q(A)|$. For a model m with bounded video clip loss $\ell_m(z, y)$, define the loss's label-conditional temporal-interaction magnitude

$$\Gamma_m(y) \coloneqq \inf_{a_0, a_1, \ldots, a_T} \operatorname{osc}_z \left(\ell_m(z, y) - a_0 - \sum_{t=1}^{T} a_t(z_t) \right), \tag{12}$$

where $a_0 \in \mathbb{R}$ and the infimum is over bounded measurable coordinate functions $a_t : \mathcal{Z}_t \to \mathbb{R}$ for $t \geq 1$. The constant and coordinate functions may be chosen separately for each y. Thus Γ_m is the infimal residual oscillation after subtracting a frame-additive function and measures only the cross-frame interaction left behind, the interaction remainder of a functional analysis-of-variance decomposition into a constant, per-coordinate main effects, and everything else (Hoeffding, 1948).

For two video-corruption channels C and C' acting under a common clean law for (X, Y) and having identical coordinate kernels $K_{C,t}(\cdot \mid x) = K_{C',t}(\cdot \mid x)$ for every x, t, write $\delta(x) \coloneqq d_{\mathrm{TV}}(C(\cdot \mid x), C'(\cdot \mid x))$ and $R_C(m) \coloneqq \mathbb{E}[\ell_m(\widetilde{X}, Y)]$. Define

$$\varepsilon_m \coloneqq \mathbb{E}[\Gamma_m(Y)\delta(X)]. \tag{13}$$

Theorem 3 (Temporal-interaction transfer bound)**.** *Under the preceding setup,*

$$|R_C(m) - R_{C'}(m)| \leq \varepsilon_m. \tag{14}$$

Theorem 4 (Exact marginal sufficiency on finite frame alphabets)**.** *Fix a label y and finite coordinate frame alphabets. The function $z \mapsto \ell_m(z, y)$ has the same expectation under every pair of joint laws with the same coordinate marginals if and only if that function is frame-additive:*

$$\ell_m(z, y) = a_0(y) + \sum_{t=1}^{T} a_t(z_t, y). \tag{15}$$

Equivalently, label by label, frame marginals are universally sufficient exactly for additive losses; if a finite-alphabet video clip loss is non-additive at a label, some two same-marginal temporal couplings change the loss's expectation at that label.

Theorem 3 refines representability from a binary property of the corruption into an interaction between channel and model, while Theorem 4 identifies the exact finite-alphabet boundary. A large joint-channel mismatch is harmless to a frame-additive loss, while even a channel pair with identical frame marginals can matter for an order-sensitive or stateful loss with a large non-additive component. For higher-is-better robustness, one applies the same result to $-\ell_m$.

The coherent control. The frame-independent and coherent conditions apply the same spatial operator at the same severities and differ only in whether a new random corruption pattern is sampled for each frame or one pattern is reused throughout the clip. For any given frame, the two conditions therefore have the same distribution of corrupted images; what changes is how the corruption patterns are shared across frames. In theory, this per-frame match alone cannot guarantee the same video-model ranking (Theorem 1). Theorem 3 bounds the possible change in risk using the difference between the two video corruption distributions and the part of the model's loss that depends on interactions between frames.

D Proofs of the theoretical results

D.1 Linear error propagation in a memory

The symbol e_t is local to this subsection and is redefined in the proof of Theorem 2. For the linear memory $m_t := Am_{t-1} + Bz_t$, perturbations obey $e_t = Ae_{t-1} + B\delta z_t$. More generally, a local time-varying linearization around a clean trajectory has $e_t = A_t e_{t-1} + \eta_t$. With the transition products $\Phi(t, s) := A_t A_{t-1} \cdots A_{s+1}$ and $\Phi(s, s) := I$, that recursion gives

$$e_t = \Phi(t, 0)e_0 + \sum_{j=1}^{t} \Phi(t, j)\eta_j, \qquad \|e_t\| \le \|\Phi(t, 0)\| \cdot \|e_0\| + \sum_{j=1}^{t} \|\Phi(t, j)\| \cdot \|\eta_j\|,$$

where the norm bound holds for any compatible vector norm and the corresponding induced matrix norm.

Contractive transition products guarantee attenuation of the corresponding earlier-impulse contribution, not necessarily of the total error when new impulses continue to arrive. The converse does not follow from this upper bound: The condition $\|\Phi(t, j)\| \ge 1$ only permits persistence or amplification in some direction, whereas the realized impulse persists only if $\|\Phi(t, j)\eta_j\|$ remains large. For a fixed finite-dimensional A, the homogeneous dynamics with $\eta_t = 0$ are asymptotically stable exactly when the spectral radius $r_{\mathrm{sp}}(A) := \max\{|\lambda| : \lambda \text{ is an eigenvalue of } A\}$ is below 1, while nonnormal matrices can still have finite-horizon transient amplification.

Contractive post-corruption products can yield recovery. Precisely, if $\eta_j = 0$ for all $j > J$ and $\|\Phi(t, j)\| \to 0$ as $t \to \infty$ for every $j \in \{0, \dots, J\}$, then the finite-sum representation above gives $e_t \to 0$.

D.2 Path transport through a stateful model

Proof of Theorem 2. Fix a clean input x, a label y, and two paths $z_{1:T}, z'_{1:T}$. Run the two state trajectories from the same initial state $h_0(x)$, and write $e_t := \|h_t - h'_t\|$ and $d_t := d(z_t, z'_t)$. Because $e_0 = 0$, Eq. 8 and induction give

$$e_t \le \sum_{s=1}^{t} \beta_s \left(\prod_{u=s+1}^{t} \alpha_u \right) d_s. \tag{16}$$

Indeed, the formula holds at $t = 1$. If the formula holds at $t - 1$, then

$$e_t \le \alpha_t e_{t-1} + \beta_t d_t \le \sum_{s=1}^{t-1} \beta_s \left(\prod_{u=s+1}^{t} \alpha_u \right) d_s + \beta_t d_t,$$

which is Eq. 16 at time t.

The Lipschitz property of the per-time losses, followed by Eq. 16 and an interchange of two finite sums, yields the pathwise bound

$$\begin{aligned} |\mathcal{L}_{m,x}(z, y) - \mathcal{L}_{m,x}(z', y)| &\le \sum_{t=1}^{T} q_t L_t e_t \le \sum_{t=1}^{T} q_t L_t \sum_{s=1}^{t} \beta_s \left(\prod_{u=s+1}^{t} \alpha_u \right) d_s \\ &= \sum_{s=1}^{T} \beta_s \left(\sum_{t=s}^{T} q_t L_t \prod_{u=s+1}^{t} \alpha_u \right) d_s = \sum_{s=1}^{T} \omega_{m,s} d_s. \end{aligned} \tag{17}$$

Now fix any coupling $\pi \in \Pi(P_x, Q_x)$. On one draw $(Z, Z') \sim \pi$, Eq. 17 applies. Since π has marginals P_x and Q_x,

$$\begin{aligned} &|\mathbb{E}_{P_x} \mathcal{L}_{m,x}(Z, y) - \mathbb{E}_{Q_x} \mathcal{L}_{m,x}(Z', y)| \\ &\quad = |\mathbb{E}_\pi[\mathcal{L}_{m,x}(Z, y) - \mathcal{L}_{m,x}(Z', y)]| \le \mathbb{E}_\pi \sum_{s=1}^{T} \omega_{m,s} d(Z_s, Z'_s). \end{aligned}$$

This bound holds for every coupling, so taking the infimum over π shows that, for every (x, y), the absolute conditional loss difference is at most $\mathsf{W}_m(P_x, Q_x)$. Therefore, under the common clean law of the preceding setup,

$$\begin{aligned} |R_C(m) - R_{C'}(m)| &\le \mathbb{E}_{X,Y}[|\mathbb{E}_{P_X} \mathcal{L}_{m,X}(Z, Y) - \mathbb{E}_{Q_X} \mathcal{L}_{m,X}(Z', Y)|] \\ &\le \mathbb{E}_X \mathsf{W}_m(P_X, Q_X) = B_m(C, C'), \end{aligned}$$

which proves Eq. 11. □

D.3 Marginal non-identifiability and model-selection regret

Proof of Theorem 1. Let Y and U be independent Rademacher variables, each uniform on $\{-1, +1\}$, and set $X_1 := U$ and $X_2 := YU$. Thus $X_1 X_2 = Y$, although either frame alone is independent of the label. Represent a sign flip by a bit $B \in \{0, 1\}$ through $(-1)^B$. Define the channels as follows. Under C_{coh}, draw $B \sim \text{Bernoulli}(q)$ and use the same bit in both frames, $(\widetilde{X}_1, \widetilde{X}_2) := ((-1)^B X_1, (-1)^B X_2)$. Under C_{anti}, draw (B_1, B_2) with

$$\Pr(B_1, B_2) := \begin{cases} q, & (B_1, B_2) \in \{(1, 0), (0, 1)\}, \\ 1 - 2q, & (B_1, B_2) = (0, 0), \\ 0, & (B_1, B_2) = (1, 1), \end{cases}$$

which is a valid distribution because $q \le 1/2$, and set $\widetilde{X}_t := (-1)^{B_t} X_t$. Under C_{ind}, draw B_1 and B_2 independently from $\text{Bernoulli}(q)$ and use the same formula. In all three channels, for every clean video clip x and coordinate t,

$$\Pr(\widetilde{X}_t = -x_t \mid X = x) = q, \qquad \Pr(\widetilde{X}_t = x_t \mid X = x) = 1 - q.$$

The three channels' coordinate kernels are therefore identical; only the joint coupling of the two flip bits differs.

Define $h_+(z_1, z_2) := z_1 z_2$ and $h_-(z_1, z_2) := -z_1 z_2$. For any pair of flip bits,

$$h_+(\widetilde{X}_1, \widetilde{X}_2) = (-1)^{B_1 + B_2} X_1 X_2 = (-1)^{B_1 + B_2} Y.$$

Thus h_+ errs exactly when $B_1 + B_2$ is odd, and h_- errs exactly when that sum is even. The parity is always even under C_{coh}, is odd with probability $2q$ under C_{anti}, and is odd with probability $q(1 - q) + (1 - q)q = 2q(1 - q)$ under C_{ind}. This parity computation proves the risk matrix in Eq. 5. □

Proof of Corollary 2. Since $q > 1/4$, $2q > 1 - 2q$, so C_{anti} prefers h_-. For $q < 1/2$, $2q(1-q) < 1/2$, so C_{ind} prefers h_+. At $q = 1/2$, the risk matrix in Eq. 5 gives the complete reversal. Adding any number of frames with identical channel behavior and letting h_+, h_- ignore those frames proves the extension to $T > 2$. □

Proof of Corollary 1. By the definition of a marginal-only procedure, the available channel information, and hence the procedure's output law after accounting for any common data and internal randomization, is the same under C_{coh} and C_{anti}. Let p be the resulting probability that selector A chooses h_+. The selector's expected selection regret is $1 - p$ under C_{coh}. Under C_{anti}, the risk gap between h_+ and the optimal h_- is $2q - (1 - 2q) = 4q - 1$, so the selector's expected regret is $p(4q - 1)$. Put $d := 4q - 1 > 0$. For every $p \in [0, 1]$, $\max\{1 - p, pd\} \geq d/(1 + d)$. Indeed, if both terms were smaller than $d/(1 + d)$, the first inequality would imply $p > 1/(1 + d)$ and the second inequality $p < 1/(1 + d)$, a contradiction. Substituting $d = 4q - 1$ proves Eq. 6. At $q = 1/2$, the risk matrix in Eq. 5 gives the complete reversal, Eq. 6 equals $1/2$, and a deterministic choice incurs unit regret under one of the two channels. □

D.4 Temporal interaction, coupling transfer, and marginal sufficiency

Proof of Theorem 3. Fix (x, y) and let $P := C(\cdot \mid x)$ and $Q := C'(\cdot \mid x)$. Consider any bounded frame-additive function $a(z) := a_0 + \sum_{t=1}^{T} a_t(z_t)$. Equality of the coordinate kernels gives

$$\begin{aligned}\mathbb{E}_P[a(Z)] &= a_0 + \sum_{t=1}^{T} \int a_t(z_t) K_{C,t}(dz_t \mid x) \\ &= a_0 + \sum_{t=1}^{T} \int a_t(z_t) K_{C',t}(dz_t \mid x) = \mathbb{E}_Q[a(Z)]. \end{aligned} \tag{18}$$

Put $r(z) := \ell_m(z, y) - a(z)$. We first establish the elementary oscillation–total-variation inequality

$$|\mathbb{E}_P r - \mathbb{E}_Q r| \leq \operatorname{osc}(r) d_{\text{TV}}(P, Q). \tag{19}$$

If $\operatorname{osc}(r) = 0$, both expectations are the same. Otherwise let $r_- := \inf_z r(z)$, $r_+ := \sup_z r(z)$, and $h := (r - r_-)/(r_+ - r_-)$, so $0 \leq h \leq 1$. The layer-cake identity and Tonelli's theorem yield

$$\mathbb{E}_P h - \mathbb{E}_Q h = \int_0^1 (P(h > u) - Q(h > u)) du.$$

Taking absolute values and using the definition of total variation bounds $|\mathbb{E}_P h - \mathbb{E}_Q h|$ by $d_{\text{TV}}(P, Q)$. Multiplication by $r_+ - r_-$ proves Eq. 19.

By Eq. 18, subtracting a does not change the difference of the two conditional expected losses. Applying Eq. 19 and then taking the infimum over all additive a therefore gives

$$|\mathbb{E}_C[\ell_m(\widetilde{X}, y) \mid X = x] - \mathbb{E}_{C'}[\ell_m(\widetilde{X}, y) \mid X = x]| \leq \Gamma_m(y)\delta(x).$$

Taking expectation over (X, Y) and using $|\mathbb{E}Z| \leq \mathbb{E}|Z|$, together with the definition in Eq. 13, proves Eq. 14. □

Proof of Theorem 4. Fix y and write $\Omega := \prod_{t=1}^{T} \mathcal{Z}_t$, with every $\mathcal{Z}_t$ finite. Identify functions and signed measures on Ω with vectors in $V := \mathbb{R}^{\Omega}$ under the standard inner product. Define the linear marginalization map

$$M : V \longrightarrow \bigoplus_{t=1}^{T} \mathbb{R}^{\mathcal{Z}_t}, \qquad (M\nu)_{t,u} := \sum_{z \in \Omega : z_t = u} \nu(z).$$

Thus $\ker M$ is exactly the space of signed measures whose every coordinate marginal is zero. The adjoint satisfies $(M^* b)(z) = \sum_{t=1}^{T} b_t(z_t)$, so $\operatorname{range}(M^*)$ is precisely the space of frame-additive functions. A constant is included by taking any one b_t constant. Finite-dimensional linear algebra gives $(\ker M)^{\perp} = \operatorname{range}(M^*)$.

If $\ell_m(\cdot, y)$ is additive, equality of coordinate marginals makes the expected losses equal by Eq. 18. Conversely, suppose the expected loss is equal under every pair of probability laws with the same coordinate marginals. Let $\nu \in \ker M$. The total mass of ν is zero, because summing any zero coordinate marginal of ν gives $\sum_{z\in\Omega} \nu(z) = 0$. Let π be the uniform probability vector on Ω. For sufficiently small $c > 0$, both $P := \pi + c\nu$ and $Q := \pi - c\nu$ are nonnegative probability vectors; moreover, $MP = MQ$ because $M\nu = 0$. The assumed invariance then gives $0 = \langle \ell_m, P - Q\rangle = 2c\langle \ell_m, \nu\rangle$. Hence $\ell_m \perp \ker M$, so $\ell_m \in (\ker M)^\perp = \mathrm{range}(M^*)$ and is additive. Taking the contrapositive shows that every non-additive finite-alphabet loss admits a same-marginal pair with different expectations, completing the proof. □

E Inference protocol and uncertainty

Why the endpoint is a ranking. Four reasons make the induced ranking, rather than the level of either score, the quantity to test. The practitioner's action is a choice among candidates, so only the ranking a cheap score induces over the zoo can change that action, and the selection regret of Corollary 1 is defined on that ranking. The two protocols also have different clean performances and different attainable score ranges, so absolute retention values are not commensurable across a protocol change while ranks are. A rank statistic is moreover invariant to any monotone rescaling of the score, so the conclusion does not depend on our choosing retention over some other robustness parameterization. Finally, the instrument matches the failure mode, because a proxy can rank models wrongly while every model retains acceptably, which a rank-correlation test detects and a retention threshold does not (App. I). Because a rank correlation over a curated zoo depends on that zoo's composition, we report the model-set bootstrap beside every coefficient and the composition sensitivity of the central result in Table 16.

E.1 Pre-specified proxy threshold

The substantive proxy threshold $\rho_0 = 0.9$ and the one-sided test of Eq. 4 were fixed before computing any rank correlation, together with the rule that a low $\rho^{(f)}$ counts as a proxy failure only when the family is model-discriminating and has a non-negligible cross-model score spread. We refer to these analysis choices as pre-specified and document them here. The threshold expresses the high rank agreement required for using the static ranking to select video models. The critical correlation follows from ρ_0, α, and the model count alone; the cutoff is not a power-based minimum detectable effect.

E.2 Calibration of the rejection rule

With $|\mathcal{M}| = n > 3$ models, the Fisher-z rule we specified before computing any rank correlations applies the usual Pearson Fisher-z normal approximation (Fisher, 1921) to the observed Spearman coefficient and rejects $\mathcal{H}_0^{(f)}$ at nominal level α when

$$\widehat{\rho}^{(f)} < \tanh\left(\operatorname{arctanh}\rho_0 - \frac{z_{1-\alpha}}{\sqrt{n-3}}\right) =: \rho_{\mathrm{crit}}.$$

This Fisher-z rule provides an approximate calibration for Spearman correlation; the discussion below explains the calibration assumptions and distinguishes the threshold test from the permutation p-values reported in the tables.

The pre-specified Fisher-z rule of Sec. 2 rejects $\mathcal{H}_0^{(f)}$ at nominal level α when the observed coefficient falls below

$$\rho_{\mathrm{crit}} := \tanh\left(\operatorname{arctanh}\rho_0 - \frac{z_{1-\alpha}}{\sqrt{n-3}}\right).$$

This threshold is a critical observed-correlation cutoff, not a minimum detectable population correlation: A power calculation at target power $1-\beta$ would additionally replace $z_{1-\alpha}$ by $z_{1-\alpha} + z_{1-\beta}$. The Fisher transform with standard error $1/\sqrt{n-3}$ is a large-sample approximation for Pearson correlation under a bivariate-normal model and is less directly justified for Spearman correlation. This limitation is especially relevant in the presence of ties, so we report the cutoff as a transparent sample-size calibration and interpret that calibration together with the reported bootstrap intervals.

The rule also counts checkpoints as independent model units; shared architecture or pretraining lineages can reduce the effective information, so the calibration is approximate rather than exact. Every use of the name pre-specified Fisher-z rule in the body inherits the approximations recorded here.

Rank-correlation table conventions. The column definitions of Table 3 carry over to every rank-correlation table in this paper. The predictor is the image-style static reference of Sec. 2, which is the frame-independent video score wherever one frame cannot carry the task and the literal single-frame image score in the appearance regime. The correlations are Spearman ρ and Kendall's tie-corrected τ_b, the two CI columns propagate model-set resampling and a paired clip-level bootstrap, and the reversals column is the fraction of model pairs that swap order between the static and video rankings. Rows are banded by the corruption's temporal stationarity, with the static-magnitude families above the gap and the temporally dynamic families below, and the rule above Overall sets off the descriptive aggregate row, which never enters a rejection decision.

Score and rank-summary conventions. Three conventions govern the scores defined in Sec. 2. We do not clip the retention ratio: A value $g > 1$ records the occasional case in which a corruption improves the measured performance relative to the clean run. We suppress the protocol index in $s^{(f)}_{m,p}$ only when the family–protocol pairing is unambiguous. Pairwise reversal and concordance summaries omit pairs tied in either score; on the remaining comparable pairs, concordance is one minus the reversal fraction, and the concordance chance value is $1/2$ under independent continuous scores. Retention normalizes corrupted performance by clean accuracy, and that normalization does not make the two quantities statistically independent.

Permutation test and the scope of a reject statement. Tables also carry a permutation p-value obtained by randomly permuting the model labels of one of the two scores (Good, 2005); that test targets the existence of any positive association rather than the proxy criterion of Eq. 4. Accordingly, unless a different interval or test is named, every reject statement in the body refers to the pre-specified working rule above rather than to an exact finite-sample Spearman test.

Descriptive aggregates and score spread. Two reported quantities are descriptive and do not enter the rejection rule. First, for the descriptive overall rows only, with $\mathcal{F}_{\text{vid}}$ denoting the five evaluated video families, we use

$$s_m^{\text{vid,overall}} := \frac{1}{|\mathcal{F}_{\text{vid}}|} \sum_{f \in \mathcal{F}_{\text{vid}}} s^{(f)}_{m,\text{vid}};$$

family-specific hypotheses and rejection decisions never use this aggregate. Second, score spread means the sample standard deviation of the relevant protocol-specific scores $s^{(f)}_{m,p}$ over the evaluated model set. The spread uses divisor $|\mathcal{M}| - 1$ and serves only to separate a diagnostic proxy failure with low ρ and high spread from a low-signal cell with low ρ and near-zero spread.

E.3 Uncertainty propagation

We report two bootstrap CIs for each $\rho^{(f)}$ (Efron & Tibshirani, 1993). The model-set interval resamples models with replacement and recomputes the correlation. The paired clip-level interval resamples evaluation video clips with replacement, using the same sampled clips for every model and both scores, and recomputes the scores and the correlation between the scores. These two intervals show how sensitive the rank correlation is to the choice of candidate models and evaluation video clips, respectively.

The model-set CI resamples models with replacement and recomputes $\rho^{(f)}$. That CI is an empirical model-zoo sensitivity interval: Because the zoo is curated rather than a probability sample from a defined universe of architectures, the interval is not a design-based coverage guarantee for all possible models.

For classification, we record whether each model predicts each evaluation clip correctly under each input protocol, both on clean input and under each corruption condition. These records let us recompute accuracy and retention for a resampled set of clips. The clip-level CI is a paired bootstrap:



Figure 4: Why representability matters. (a) Clean input. (b,c) Independent and temporally coherent color shift is applied to the same Diving48 video clip; each affected frame reveals the corruption, and the image-robustness ranking transfers to video. (d) Intact frames are repeated, leaving no visible sign of repetition in any individual frame; single-image robustness has little correlation with temporal-structure robustness. Digits identify source-frame indices, with repeated positions tinted.

We resample the evaluation video clips with replacement using the same indices for every model, recompute each model's group-balanced static and video scores from these per-clip records, and recompute $\rho^{(f)}$. The nested equal-weight aggregation of the corrupted per-condition accuracies in Eq. 2 is a fixed linear functional, computed by a matrix multiply in each resample; the clean-normalized retention is then recomputed by elementwise division by that resample's corresponding protocol-specific clean accuracy. Thus, whenever the resampled protocol-specific clean accuracy is nonzero, the implementation exactly recomputes the stated ratio score rather than treating the full score as linear; a 10^{-6} numerical floor guards the otherwise undefined zero-denominator case. We report both intervals; the video clip interval isolates finite-test-set noise from model-set composition.

F Implementation Details

F.1 Corruption Operators

The frame-independent family draws a fresh realization per frame; the temporally coherent family freezes one realization across the video clip; and the drifting family ramps the magnitude from severity 1 to the nominal severity over the video clip. Temporal-structure operators act on temporal sampling or order: frame drop with hold-last, repeat, shuffle, VFR via index resampling, and temporal jitter. Motion-coupled operators act on geometry or time-varying magnitude: camera shake as a smooth affine random walk and time-varying blur or brightness with cosine-modulated magnitude.

As a control, we apply the same spatial corruption at the same severity and vary whether the corruption's random pattern changes between frames. For example, frame-independent Gaussian noise adds a newly sampled noise pattern to each frame, whereas temporally coherent Gaussian noise reuses one pattern throughout the video clip. This comparison tests whether persistence across frames alone changes the model ranking.

Fig. 4 illustrates the frame-independent, coherent, and frame-repeat conditions.

Deployment origins of the temporal-structure operators. The five temporal-structure operators abstract faults that deployed capture, transport, and analytics stacks produce. Frame drop with hold-last and frame repeat are what a fixed-rate consumer receives when pictures arrive late or never arrive. The standard transport protocol for interactive media, RTP, neither guarantees delivery nor

prevents out-of-order delivery, and RTP sequence numbers exist so that a receiver can detect loss and restore packet order (Schulzrinne et al., 2003); a decoder then conceals the missing data by reusing previously decoded pictures, and the mismatch propagates through inter prediction until an intact refresh (Wang & Zhu, 1998; Stockhammer et al., 2003). VFR and temporal jitter are what irregular frame timing becomes once a consumer resamples the stream onto a fixed grid. Capture systems lower the frame rate under limited bandwidth (Yi et al., 2021), live analytics systems treat the frame rate as a tunable knob or filter frames on the camera to fit compute and bandwidth budgets (Jiang et al., 2018; Li et al., 2020b), and a streaming perception stack under a latency budget ignores the frames that arrive while an earlier frame is still being processed (Li et al., 2020a). Frame shuffle has the weakest deployment counterpart, because a receiver restores packet order before decoding; we keep shuffle as the control that tests whether a model uses temporal order at all. Earlier video-corruption benchmarks already include members of this family, namely packet loss, bit error, and frame-rate conversion (Yi et al., 2021) together with temporal sampling, reversal, jumbling, and freezing (Schiappa et al., 2023). This family captures faults that occur in deployed video pipelines. The index-level operators isolate the timeline component of these faults, and App. K.2 reproduces two of the faults with a codec and transport stack, where concealment artifacts and timeline shifts occur together. Neither pipeline reorders frames, so the weak transfer of the static ranking to the pipeline-level ranking does not depend on the shuffle operator.

The matched-pair manipulation in code. The three photometric families of frame-independent, temporally coherent, and temporally drifting corruption share one spatial operator, one severity grid, and one video clip, and differ only in how the spatial operator's random realization and magnitude are scheduled along the video clip. Listing 1 is that dispatch and is where the spatial operator is held fixed across the matched pair of Sec. 2. Every operator receives the random corruption parameters as an explicit argument instead of reseeding a generator internally, as the reference ImageNet-C implementation does. Passing the random corruption parameters explicitly makes one spatial corruption replayable under a different temporal structure. `Realization.sample` returns whatever the operator's realization consists of: the additive field for the noise operators, the kernel angle for motion blur, the airlight for fog, or the per-channel shift direction for color shift. The return is empty for the operators that are deterministic at a fixed magnitude, such as the blurs and JPEG.

Two lines carry the whole contrast. The `shared` assignment draws the realization once and reuses that realization for every frame unless the mode is independent, in which case each frame draws a new realization; the `drift` branch ramps the magnitude from the severity-1 value to the nominal one across the video clip, while the two static-magnitude families hold the magnitude at the nominal value throughout. The operator, the severity grid, the severity index, and the video clip are identical across the three families, so a rank change between those families cannot be attributed to a change of spatial operator. The `flicker` schedule is not a sixth family: The motion-coupled operators `time_varying_blur` and `time_varying_brightness` are the corresponding spatial operators driven by that cosine-modulated magnitude. Camera shake and the five temporal-structure operators act on geometry and on sampling order rather than on a per-pixel magnitude, and are separate functions; the frame-independent family is also the only one evaluated under both the image and the video protocols, which is where the matched pair of Sec. 2 is formed.

```python
# vrg/corruptions/temporal.py.  OPS[name] is the spatial operator,
# SEVERITY[name] is the five-entry magnitude grid, and Realization.sample
# the operator's random draw (noise field, blur angle, fog airlight, ...).

def _magnitude_schedule(name: str, severity: int, T: int,
                        mode: str) -> np.ndarray:
    nominal = SEVERITY[name][severity - 1]
    if mode in ("independent", "coherent"):
        return np.full(T, nominal, np.float32)
    grid = SEVERITY[name]
    if mode == "drift":  # ramp from severity 1 up to the nominal severity
        lo = grid[0]
        return np.linspace(lo, nominal, T).astype(np.float32)
    if mode == "flicker":  # oscillate between mild and nominal
        lo = grid[0]
        osc = 0.5 * (1 - np.cos(np.linspace(0, 3 * np.pi, T)))
        return (lo + (nominal - lo) * osc).astype(np.float32)
    raise ValueError(mode)


def apply_photometric(clip: np.ndarray, name: str, severity: int,
                      mode: str, rng: np.random.Generator) -> np.ndarray:
    T, H, W, C = clip.shape    # (T, H, W, 3) float32 in [0, 1]
    op = OPS[name]
    mags = _magnitude_schedule(name, severity, T, mode)
    out = np.empty_like(clip)
    shared = (Realization.sample(name, (H, W, C), rng)
              if mode != "independent" else None)
    for t in range(T):
        param = (shared if shared is not None
                 else Realization.sample(name, (H, W, C), rng))
        out[t] = op(clip[t], mags[t], param)
    return out
```

Listing 1: The temporal-mode dispatch behind the frame-independent, temporally coherent, and temporally drifting families, shown as an implementation excerpt. The two lines that separate the families are the `shared` assignment and the `drift` branch of the schedule.

F.2 Training and evaluation details

Datasets. We compare two task settings: appearance-dominated and motion-dependent action recognition. For the appearance-dominated setting, we use UCF101 (Soomro et al., 2012), which has 101 classes and 9,537 train and 3,783 test video clips in official split 1. HMDB51 (Kuehne et al., 2011), with 51 classes in split 1, provides the cross-dataset replication in App. P.1. The primary motion-dependent task is Diving48 (Li et al., 2018), which has 48 fine-grained dive classes and 15,027 train and 1,970 test video clips. The dataset's creators aimed to reduce static appearance biases so that recognition depends more on temporal information. SSv2 provides a second motion-dependent task, with classes that distinguish the direction and order of motion. Each video is decoded to 16 uniformly sampled frames at 256×256; training uses random-resized-crop and flip to 224, evaluation a center crop. Unless stated otherwise, evaluation uses a fixed random test subset of 1,500 video clips shared across models, so each model's static and video scores are paired for each video clip.

A common operating point. We use operating point for the configuration at which a checkpoint is scored, comprising the checkpoint's fine-tuning recipe, head, and evaluation input protocol. The recipe belongs in that definition because fine-tuning changes out-of-distribution behavior: Full fine-tuning can distort pretrained features and underperform linear probing under shift (Kumar et al., 2022). Holding one recipe fixed across a zoo keeps that effect from being read as a difference between backbones. Every model fine-tuned by us is trained with adaptive moment estimation with decoupled weight decay (AdamW), a cosine schedule, 16-bit brain floating-point (bf16) arithmetic, and label

smoothing 0.1, so robustness is measured at a competitive operating point rather than a weak one. The off-the-shelf zoos of Sec. 4 instead keep each checkpoint's published operating point, which we never restate as a property of a task or a zoo. App. G lists every checkpoint, and the recipe below gives the full settings, including the evaluation input length each head receives.

Appearance-regime backbones are initialized from pretrained PyTorch Image Models (Wightman, 2019) weights and fine-tuned on the training split for 25 epochs, or 30 for HMDB51, with AdamW at weight decay 0.05, a cosine schedule with 3 warmup epochs, base learning rate 1.5×10^{-4}, and a $10\times$ rate for the randomly initialized head. We use bf16 autocast, gradient clipping at 5.0, label smoothing 0.1, batch size 32, and $T = 8$ sparsely sampled frames with random-resized-crop at scale $[0.7, 1.0]$ and horizontal flip.

Motion-regime Diving48 temporal-head models use the same recipe with the order-sensitive tx head: 2 layers, learned temporal positional embeddings, and a classification (CLS) readout. Those models are fully fine-tuned for 40 epochs at batch 64; the order-invariant averaging baselines use the same budget. The model-family names below abbreviate Vision Transformer (ViT), Data-efficient Image Transformer (DeiT), Twins Spatially Separable ViT (Twins-SVT), Cross-Covariance Image Transformer (XCiT), Global Context ViT (GC-ViT), and Residual Network (ResNet). The pretraining labels denote Contrastive Language-Image Pre-training (CLIP), its OpenAI checkpoint (CLIP-OA), and ImageNet-21k (IN21k). DeiT3-M, ViT-B/16 (CLIP-OA), and ViT-B/16 (IN21k) use base learning rate 5×10^{-5}, head multiplier 4, and 5 warmup epochs. ViT-Ti/16, Twins-SVT-S, XCiT-S12, and GC-ViT-T use learning rate 1×10^{-4}; ResNet-34, ResNeXt-50, RegNetY-016, RegNetY-080, and EfficientNet-B1 use learning rate 1.2×10^{-4}. Both groups use head multiplier 5 and 5 warmup epochs. ResNet-101 uses learning rate 8×10^{-5}, head multiplier 5, 5 warmup epochs, and a 45-epoch budget. All other models use base learning rate 1.5×10^{-4}, head multiplier 10, and 3 warmup epochs.

The input length follows each Hugging Face 3D model's configuration: VideoMAE uses $T = 16$ and the Time-Space Transformer (TimeSformer) uses $T = 8$. Torchvision and PyTorchVideo models use the matched 16-frame cache at 224^2. Native-32 wrappers for ViViT-B and the Base variant of the Multiscale ViT (MViT), MViT-B 32×3, nearest-index upsample the same 16-frame content to 32 rather than sampling 32 unique frames. Evaluation uses center crops and uniform temporal sampling over the 16-frame cache. We record results for every combination of dataset, model, protocol, family, corruption type, and severity.

G Evaluated model sets

Models. The UCF101 comparison uses 17 backbones, spanning CNNs and Transformers across scales plus ViT-B variants initialized from CLIP (Radford et al., 2021) and IN21k checkpoints. Each backbone encodes every frame with shared weights, and a temporal-average (avg) head, also known as the TSN (Wang et al., 2016) consensus head, averages the frame predictions without using temporal frame order. The Diving48 comparison uses 27 architecturally diverse backbones with an order-sensitive tx head that combines information across frames. Clean top-1 reaches 78.6%, with a mean of 63.3%, a strong single-stream Diving48 result well above chance and single-frame baselines. To test whether the findings extend beyond image backbones with temporal heads, we also fine-tune and evaluate 37 video-pretrained 3D backbones on the target classification datasets (Sec. 4). We also evaluate public checkpoints without further fine-tuning and frozen backbones on DAVIS segmentation (Sec. 4), with evaluation details in Apps. M and N.

Appearance-regime backbones. In the appearance regime, our model set spans 17 backbones, each wrapped as a shared frame encoder with a linear avg, or TSN (Wang et al., 2016), consensus head. The set includes CNNs: ResNet-18, ResNet-50, ResNet-101, ConvNeXt-T, ConvNeXt-S, EfficientNet-B0, EfficientNet-B3, and RegNetY-040. The Transformers are ViT-Ti, ViT-S, ViT-B (Dosovitskiy et al., 2021), DeiT3-S, DeiT3-B, Shifted Window (Swin) variants Swin-T and Swin-S, plus ViT-B variants initialized from CLIP (Radford et al., 2021) and IN21k checkpoints. The avg head is parameter-light, order-invariant, and ideal for the appearance regime, where a single frame already signals the class.

Motion-regime backbones. In the motion regime, where strong performance requires multi-frame aggregation, the matched set uses an architecturally diverse pool of 27 backbones. Across scales, the

pool covers the CNN families ResNet, ResNeXt, RegNetY, EfficientNet, and ConvNeXt, together with the Transformer families ViT, DeiT3, Swin, Twins, XCiT, and GC-ViT. The pool also includes three CLIP and IN21k pretrainings: CLIP trained on the Large-scale Artificial Intelligence Open Network's two-billion-pair dataset, CLIP-OA, and supervised IN21k. The models each use a tx head, a CLS token attending over T frame tokens with learned temporal positional embeddings, so each model is an image backbone given a light temporal module, as in image-to-video adaptation (Yang et al., 2023). Every model is fully fine-tuned rather than frozen. Clean top-1 reaches 78.6%, with a mean of 63.3%. This clean accuracy marks a strong single-stream Diving48 operating point well above chance and single-frame baselines. The 27-model pool supports model-set uncertainty estimation for the dynamic-family correlations (App. H). Both regimes are fine-tuned with the recipe of App. F.2, so robustness is measured at a competitive operating point rather than a weak one.

Video-pretrained 3D backbones. The 37 video-pretrained 3D models are drawn from three independent model frameworks and the three pretraining datasets K400 (Kay et al., 2017), the motion-centric SSv2 (Goyal et al., 2017), and K600 (Carreira et al., 2018), spanning both 3D CNNs and 3D Transformers and including a VideoMAE scale sweep from 22M to 632M parameters. From Hugging Face: VideoMAE-S, VideoMAE-B, VideoMAE-L, and VideoMAE-H (Tong et al., 2022), together with TimeSformer-B (Bertasius et al., 2021), all pretrained on K400; short-schedule VideoMAE-B variants on K400 and SSv2, plus VideoMAE-S, VideoMAE-B, and TimeSformer-B fine-tuned on SSv2; TimeSformer-B pretrained on K600, a third large-scale supervised-video corpus; and the factorized space-time ViViT-B (Arnab et al., 2021) pretrained on K400.

Torchvision and PyTorchVideo checkpoints. The checkpoint names below distinguish residual 3D (R3D), mixed convolution (MC3), residual spatial–temporal factorization (R(2+1)D), and separable 3D (S3D) architectures. From Torchvision, nine K400 backbones: R3D-18, MC3-18, and R(2+1)D-18 (Tran et al., 2018); S3D (Xie et al., 2018); Video-Swin-T, Video-Swin-S, and Video-Swin-B (Liu et al., 2022); and MViTv1-B (Fan et al., 2021b) and MViTv2-S (Li et al., 2022). From PyTorchVideo (Fan et al., 2021a), a third source: SlowFast-R50, SlowFast-R101, and a deeper SlowFast-16×8-R101 (Feichtenhofer et al., 2019); Expand 3D (X3D) variants X3D-M, X3D-L, X3D-S, and X3D-XS, as well as the mobile EfficientX3D-S and EfficientX3D-XS (Feichtenhofer, 2020); a PyTorchVideo R(2+1)D-R50; the MViT-B 3D Transformer at both 16×4 and the longer 32×3 (Fan et al., 2021b); and the 50-layer Inflated 3D model (I3D-R50) (Carreira & Zisserman, 2017), the 101-layer Channel-Separated Network (CSN-R101) (Tran et al., 2019), the two-dimensional (2D) convolutional baseline (C2D-R50), and the Slow-R50 3D-ResNet, all on Diving48. Per-model clean top-1 and per-family retention are in Table 18.

H Robustness of the motion-regime rank divergence

The motion-regime result separates a representability fact from an empirical claim. By construction, a single-frame benchmark cannot encode dropped, repeated, or reordered frames, and encodes a temporal drift only as one magnitude at a time. The empirical questions are whether those operators cause substantial performance loss and whether the image-style static reference preserves the resulting video ranking. On Diving48, we observed that frame-repeat retention falls to 0.36 at severity 5 and that the dynamic-family rank correlations fall below ρ_0.

The static ranking fails the $\rho_0 = 0.9$ test for every dynamic family. On the 17-model UCF101 set, rankings differed under dynamic corruption, but the models' retention scores were nearly equal. On the 27-model Diving48 set, retention differences are substantial, which allows us to test whether the static ranking reliably identifies the more robust video models. The pre-specified Fisher-z rule gives a critical observed correlation of approximately 0.81. All three dynamic correlations fall below that cutoff before rounding, so we reject the hypothesis $\mathcal{H}_0^{(f)} : \rho_*^{(f)} \geq \rho_0$ for each dynamic family. All three tests remain significant after Holm correction, as detailed below. Temporal structure, which no single frame reveals, has the lowest rank correlation ρ; motion coupling, which one frame partly reveals, has the correlation closest to the cutoff. The model-set bootstrap intervals exclude zero, so the static score still provides information, but the static score does not predict the video ranking accurately enough to meet our $\rho_0 = 0.9$ criterion.

Table 3: Static→video rank correlation by corruption family for 27 Diving48 tx models. Model-set and clip-level bootstrap CIs accompany each estimate. Every dynamic-family point estimate is below $\rho_0 = 0.9$ and the critical cutoff; all three one-sided tests remain significant after Holm correction, with permutation $p < 0.001$. Cross-model score spreads are 0.093 for the static reference and 0.036, 0.016, and 0.024 for the drifting, temporal-structure, and motion-coupled families.

Video score	ρ	model CI	clip CI	τ_b	reversals
Temporally coherent	1.000	[1.00, 1.00]	[1.00, 1.00]	1.000	0.00
Temporally drifting	0.767	[0.54, 0.88]	[0.71, 0.83]	0.550	0.23
Temporal structure	0.728	[0.41, 0.91]	[0.49, 0.79]	0.571	0.21
Motion-coupled	0.808	[0.60, 0.90]	[0.69, 0.86]	0.601	0.20
Overall	0.981	[0.93, 0.99]	[0.96, 0.99]	0.886	0.06

Table 4: Static-to-video rank transfer among 37 video-pretrained 3D models on Diving48. We report Spearman ρ, model-set bootstrap CIs, the decision against the pre-specified $\rho_0 = 0.9$ criterion, and pairwise concordance; ✔ denotes rejection.

Family	ρ	model-set CI	reject $\rho_0 = 0.9$	concord.
Temporally coherent	0.99	[0.96, 1.00]	✘	0.97
Temporally drifting	0.88	[0.76, 0.94]	✘	0.85
Temporal structure	0.05	[−0.27, 0.34]	✔	0.52
Motion-coupled	0.77	[0.59, 0.88]	✔	0.79

Uncertainty over models and video clips. Because external validity concerns a population of models, the model-set bootstrap is the primary uncertainty analysis; the paired clip-level bootstrap isolates finite-test-set noise for the fixed model set. The analyzed $\mathcal{M}$ contains 27 architecturally diverse models: CNNs; the ViT, DeiT, Swin, Twins, XCiT, and GC-ViT families; and three CLIP and IN21k pretrainings. ResNet-101 and ViT-B/16 (IN21k) use the model-specific schedules of App. F.2, since under the common schedule those two checkpoints' clean top-1 inverts with respect to model capacity, the signature of optimization failure.

For this model set, every dynamic point estimate and paired video clip interval lies below $\rho_0 = 0.9$, while the model-set intervals exclude zero (Table 3). The model intervals for drift and motion also lie wholly below 0.9, although motion does so narrowly, with upper endpoint 0.8975; the structure interval reaches 0.91. We reject the threshold for all three families using the pre-specified Fisher-z tests. The model bootstrap supports positive but imperfect transfer without independently placing every interval below 0.9.

Composition and multiplicity. We fixed ρ_0 in advance but did not pre-specify the model-set composition. Across the restrictions in Table 16, temporal structure remains below $\rho_0 = 0.9$ on both video-pretrained 3D sets. We also Holm-correct (Holm, 1979) the three dynamic-family Fisher-z p-values per task. These values test $\mathcal{H}_0 : \rho_*^{(f)} \geq \rho_0$, whereas the permutation p tests only whether $\rho > 0$. On Diving48, all 3 dynamic families remain significant at $\alpha = 0.05$; the largest Holm-adjusted value is $p = 0.043$.

I VIDEO-PRETRAINED 3D BACKBONES: EXTENDED ANALYSIS

The 37-model set spans three pretraining datasets, three frameworks, five 3D-Transformer families, and a VideoMAE scale sweep from 22M to 632M parameters; App. G lists every checkpoint. Table 4 reports the per-family rank correlations behind the Diving48 3D row of Table 1, with the intervals and concordances. Sec. 4.2 reports the temporal-structure selection costs.

Comparing the tx and 3D constructions. We observed that all three dynamic families diverge in the tx set, whereas the stronger 3D models preserve the static ranking under smooth drift but not under temporal structure or motion coupling. Both constructions place the coherent control first and

temporal structure last (Table 1), and both diverge on the non-smooth cross-frame families that the image-style static reference does not fully measure. The effect also holds across an architecturally and pretraining-diverse 3D model set as well as under one temporal-head family.

Motion coupling and temporal structure. The motion-coupling interval $[0.59, 0.88]$ lies entirely below the pre-specified $\rho_0 = 0.9$, giving a second synthetic rank divergence among the 3D models. The structure-family interval $[-0.27, 0.34]$ is also far below 0.9, with a point estimate near zero and near-chance concordance. Motion coupling is a mixed synthetic family: One frame shows blur and brightness modulation at one magnitude but cannot show displacement across frames, so the static reference encodes that family in part. Temporal structure, which no frame shows, is the unambiguous family-level failure at the pipeline level.

Composition sensitivity. Under every clean-accuracy floor and top-K restriction in Table 16, the tests reject $\rho_0 = 0.9$ and the largest observed correlation is 0.64. Restricting either set to the members with the highest clean accuracy raises the structure correlation to about 0.6. Rejection of the pre-specified threshold is stable across the tested restrictions; the near-chance correlation describes the full, clean-accuracy-diverse zoo.

Illustrative model pair. VideoMAE-S (SSv2) and ViViT-B are nearly tied under frame-independent corruption, at 0.841 and 0.840 retention. The static screen selects VideoMAE-S (SSv2) by a 0.001 margin. Under frame-timing corruption produced by drop, repeat, or reorder, however, the two models' retentions are 0.724 and 0.920; the selected model is worse by 0.197. The frame-timing-best model, ViViT-B, ranks only 19th under the static reference.

Ranking rather than retention level. Mean structure-family retention is 0.822 (Table 18), comparable to static retention 0.822. The rank divergence therefore concerns the ranking of models, not uniformly catastrophic performance loss, which motivates the rank-correlation endpoint rather than a retention threshold.

Greater robustness to spatial corruption does not guarantee greater robustness to changes in frame order or timing. Among the 37 video-pretrained 3D models on Diving48, the same static score predicts the drift ranking at $\rho = 0.88$ but is almost unrelated to the temporal-structure ranking, at $\rho = 0.05$. Robustness to temporal structure nevertheless transfers across datasets and predicts which models better tolerate pipeline-produced packet loss and VFR judder.

J SSv2 REPLICATION AND CROSS-DATASET CONSISTENCY

The Diving48 result uses a single motion benchmark chosen because that benchmark suppresses static scene bias. To test whether the conclusion transfers beyond that dataset, we replicate the entire matched-pair construction on SSv2 (Goyal et al., 2017), the canonical temporal-relation benchmark and one of the most widely used video datasets. SSv2 tests the same question under stronger temporal dependence than Diving48: The 174 SSv2 classes are fine-grained interactions whose labels are fixed by the direction and order of motion. For example, distinct classes encode moving something up, moving something down, pushing something from left to right, and placing something behind something, so temporal order carries the class.

We fine-tune the same diverse image-encoder plus tx construction with 17 backbones on SSv2 and run the identical five-family robustness evaluation. Clean top-1 reaches 47.0%, with a mean of 39.8%. Both protocols use identical weights and a shared input budget across an architecturally diverse set with a wide clean-accuracy spread. The 3D set below reaches 64.0% through SSv2-pretrained checkpoints, and the threshold result holds when either set is restricted to the members with the highest clean accuracy (Table 16).

The temporal-structure mismatch also appears on a second motion benchmark. To test whether the static ranking also fails the $\rho_0 = 0.9$ test on another motion task, we repeat the matched-pair evaluation on the temporal-relation benchmark SSv2 (Goyal et al., 2017). On SSv2, the baseline that averages frame predictions without using frame order performs substantially worse than the tx head that uses frame order, with a 16.8-point gap in clean accuracy. This gap shows why frame order

matters for the task, and disrupting the sequence causes substantial damage: At severity 5, frame-repeat retains only 0.58 and frame-shuffle 0.69 of clean accuracy. With the tx head, static→dynamic rank agreement again falls below $\rho_0 = 0.9$ for all three dynamic families. Among 37 video-pretrained 3D models on SSv2, temporal-structure correlation reaches only $\rho = 0.22$. The temporal-structure mismatch appears consistently on both benchmarks, with both image backbones plus a tx head and video-pretrained 3D models.

Order carries the label in this regime. On appearance tasks, an order-invariant avg head performs on par with the tx head, and on Diving48, the avg head nearly matches the order-sensitive one. On SSv2, by contrast, the order-invariant avg baseline reaches only 23.1% clean top-1, compared with 39.8% for the order-sensitive tx, a gap of 16.8 points. This controlled comparison provides operational evidence that order-sensitive aggregation matters on SSv2, without implying that every example requires temporal order. No single image can represent frame drop, repeat, or reorder, and we observed substantial accuracy loss from those corruptions in this regime: At severity 5, frame-repeat retains only 0.58 and frame-shuffle 0.69 of clean accuracy. The worst dynamic operator retains 0.58, in contrast to the far milder effect of that operator on appearance tasks. A single-image benchmark is structurally blind to these failures on a dataset where those failures are most consequential.

The rank divergence replicates. As on Diving48, we observed that the static frame-independent ranking transfers almost perfectly across temporal profiles, at coherent $\rho = 1.00$ in Table 7. The static→dynamic transfer drops well below the pre-specified $\rho_0 = 0.9$ on the temporally dynamic families: drift $\rho = 0.40$, temporal structure $\rho = 0.72$, and motion $\rho = 0.24$. All lie below the pre-specified Fisher-z critical correlation 0.77 for the 17-model set. The Holm correction (App. H) extends to SSv2. On both motion-centric benchmarks, static-corruption robustness provides information about dynamic robustness but does not meet the stipulated fidelity criterion. The gap between the order-invariant baseline and the tx head on SSv2 makes the task's temporal dependence explicit.

The divergence reproduces among video-pretrained 3D models on SSv2. To show the second-dataset replication is not specific to the image-encoder plus tx construction either, we repeat the 3D-among-3D test of Sec. 4 on SSv2. We fine-tune 37 video-pretrained 3D models from the same zoo used on Diving48. That zoo spans the same three frameworks of Hugging Face, Torchvision, and PyTorchVideo; 3D CNNs and 3D Transformers; and the three pretraining datasets K400, SSv2, and K600. The Hugging Face models are VideoMAE and TimeSformer. The SSv2-pretrained checkpoints give the strongest baselines. We evaluate the models over the identical taxonomy. Clean top-1 runs from 24.4% to 64.0%, and the pre-specified Fisher-z critical cutoff is 0.83 at $\rho_0 = 0.9$.

Threshold tests on SSv2. Within this 3D model set, the static frame-independent ranking again transfers on the static-magnitude coherent family, with $\rho = 0.99$ and CI $[0.95, 1.00]$, while the dynamic families separate in Table 6. For temporal structure, $\rho = 0.22$, CI is $[-0.07, 0.48]$, concordance 0.58 is near chance on this model set, and we reject $\rho_0 = 0.9$ with a one-sided $p < 10^{-3}$. We also reject the threshold for drift, at $\rho = 0.60$, but not for motion, at $\rho = 0.85$ with CI $[0.68, 0.92]$. Thus, the tests reject 2 of the three dynamic families at the pre-specified threshold. As on Diving48, temporal structure is the core divergence, appearing on two datasets and under both constructions. On the more strongly order-dependent SSv2, even temporal drift diverges among the 3D models, whereas on Diving48, the strong space-time models absorb drift at $\rho = 0.88$ (Table 4); conversely, we reject the threshold for motion on Diving48 but not on SSv2. The two partly represented families thus trade places across datasets, as families between the two ends of the representability axis may, while the coherent control is the stable cross-dataset success and temporal structure the stable cross-dataset failure.

Cross-dataset transfer of the video-aware measurement. We evaluate the identical 37-model 3D zoo on SSv2 and Diving48, enabling a comparison of cross-dataset stability for each robustness measurement. For each family, we correlate the family's retention ranking measured on Diving48 against the same family's ranking on SSv2; this correlation is video→video transfer. We contrast that transfer with the image-style static reference on Diving48 predicting the same SSv2 criterion, which is static→cross transfer (Table 5). Both predictors cross datasets, although predictor-specific interactions with dataset shift remain possible.

Table 5: Cross-dataset rank transfer for the same 37 3D models on Diving48 and SSv2. Video→video correlates each family's ranking across datasets; static→cross uses the Diving48 static reference to predict the SSv2 dynamic-family ranking. Williams p compares these dependent correlations. Model-bootstrap CIs and permutation p-values accompany the video→video estimates.

	video→video transfer			static→cross	
Family	ρ	model-set CI	permutation p	ρ	Williams p
Clean accuracy	0.43	[0.12, 0.67]	0.0028		
Frame-independent	0.92	[0.81, 0.96]	0.0002		
Temporally coherent	0.90	[0.77, 0.96]	0.0002		
Temporally drifting	0.36	[−0.001, 0.65]	0.0130	0.48	0.9462
Temporal structure	0.67	[0.36, 0.86]	0.0002	0.27	0.0180
Motion-coupled	0.71	[0.49, 0.84]	0.0002	0.82	0.9438

Table 6: Static→video rank correlation by family among 37 video-pretrained 3D models on SSv2. The critical cutoff is 0.83. Coherent corruption transfers; the tests reject $\rho_0 = 0.9$ for temporal structure and drift but not motion.

Family	ρ	model-set CI	reject $\rho_0 = 0.9$	concord.
Temporally coherent	0.99	[0.95, 1.00]	✘	0.96
Temporally drifting	0.60	[0.31, 0.81]	✔	0.72
Temporal structure	0.22	[−0.07, 0.48]	✔	0.58
Motion-coupled	0.85	[0.68, 0.92]	✘	0.83

Table 7: Static→video rank correlation by corruption family on the second motion benchmark, SSv2, with 17 tx models, in the columns of Table 3 together with a permutation p. As on Diving48, the dynamic families fall below $\rho_0 = 0.9$ and the critical cutoff under the pre-specified rule on a second, more strongly order-dependent dataset. Cross-model score spreads are 0.077 for the static reference and 0.020, 0.014, and 0.015 for the drifting, temporal-structure, and motion-coupled families.

Video score	ρ	model CI	clip CI	τ_b	reversals	permutation p
Temporally coherent	1.000	[1.00, 1.00]	[0.99, 1.00]	1.000	0.00	< 0.001
Temporally drifting	0.402	[−0.17, 0.81]	[0.17, 0.66]	0.294	0.35	0.056
Temporal structure	0.718	[0.34, 0.93]	[0.49, 0.83]	0.544	0.23	0.001
Motion-coupled	0.238	[−0.35, 0.76]	[−0.05, 0.51]	0.191	0.40	0.178
Overall	0.971	[0.88, 0.99]	[0.92, 0.99]	0.868	0.07	< 0.001

Temporal-structure robustness transfers more strongly than the static reference. We observed that temporal-structure robustness measured on Diving48 predicts temporal-structure robustness on SSv2 at $\rho = 0.67$, with CI $[0.36, 0.86]$ and permutation $p = 0.0002$. The static reference predicts the same criterion at $\rho = 0.27$; here $\Delta\rho = 0.40$, CI is $[-0.003, 0.78]$, and Williams $p = 0.0180$. Notably, temporal-structure robustness transfers more strongly across these benchmarks than clean accuracy, whose $\rho = 0.43$. The dependent-correlation gap is significant under the rank-scale Hotelling–Williams test on the identical 37-model set, with $p = 0.0180$; the model-bootstrap $\Delta\rho$ interval is $[-0.003, 0.78]$, with a lower bound near zero.

K Cross-validation against pipeline-produced corruptions

The two cross-validation axes differ in representability: A single frame can represent the compression axis and cannot represent the temporal-structure axis.

Table 8: Codec cross-validation on the compression axis. We correlate robustness under synthetic JPEG or the image-style static reference with robustness under low-bitrate H.264 and H.265, reporting model-set bootstrap CIs and pairwise concordance.

Model set	Predictor → codec	$\lvert\mathcal{M}\rvert$	ρ	model-set CI	concord.
UCF101 (avg)	synthetic JPEG	17	0.89	[0.64, 0.99]	0.89
	static reference		0.69	[0.26, 0.89]	0.74
Diving48 (tx)	synthetic JPEG	27	0.93	[0.78, 0.98]	0.91
	static reference		0.92	[0.78, 0.96]	0.87

K.1 Compression axis: external validity of synthetic compression

We cross-validate the controlled taxonomy against pipeline-level corruption. We re-encode every video clip through low-bitrate H.264 and H.265 video codecs (Wiegand et al., 2003; Sullivan et al., 2012) using libav. This re-encoding produces the discrete-cosine-transform block, deblocking, and motion-compensated artifacts of deployed bandwidth-limited closed-circuit television, dashcam, and streaming pipelines rather than the artifacts of a hand-written filter. Lossy compression and transmission loss are also the two distortion classes the video-quality literature identifies as dominant in deployed video pipelines (Seshadrinathan et al., 2010), which is why we cross-validate against lossy compression and transmission loss.

We evaluate the ability of synthetic JPEG robustness to predict codec robustness. We observed a strong relation: Across the UCF101 matched set of 17 backbones, the synthetic-JPEG ranking tracks the codec ranking with Spearman $\rho = 0.89$ and pairwise concordance 0.89. The image-style static reference, which aggregates thirteen spatial types of which compression is one, predicts the same codec ranking at $\rho = 0.69$, so an image-style measurement predicts the codec ranking better the more closely that measurement matches the damage the pipeline produces. On the Diving48 set of 27 tx models, the cross-validation holds at $\rho = 0.93$ for synthetic JPEG and 0.92 for the static reference (Table 8). The synthetic compression family ranks models similarly to the codec pipelines, providing evidence of external validity and supporting the use of the controlled taxonomy to study temporal structure.

K.2 Temporal-structure axis: pipeline-produced corruptions and controls

The two pipeline-level corruptions. The synthetic taxonomy implements the temporal-structure family only with index-level tensor operations such as hold-last frame drop, freeze, and resampling. We complement these operators with two temporal corruptions produced by a codec and transport pipeline. For transport packet loss, we encode each video clip with an H.264 encoder that uses periodic key frames as in a low-latency stream, drop a controlled fraction of the transport packets, and let the decoder's error-concealment path (Wang & Zhu, 1998; Stockhammer et al., 2003) reconstruct the stream. The resulting decoding errors can propagate across subsequent frames until an intact refresh, as in a deployed lossy stream. For container-level VFR, we mux the frames with non-uniform presentation timestamps in a VFR container and resample at a fixed playback rate. A fixed-rate consumer then holds the last-shown frame across long inter-frame gaps, producing decoder-side judder and freeze. Fig. 5 shows both pipelines beside the index-level operator whose external validity the two pipelines check.

Dependent-correlation and paired-bootstrap support. The two correlations share the pipeline-level temporal criterion, so the difference between the two is an overlapping dependent correlation comparison and must be tested as such; overlap between the two marginal CIs is not a valid test of that difference. We apply Williams's T_2 test (Williams, 1959; Steiger, 1980) to the average-rank variables. Spearman correlation is the Pearson correlation of those ranks. The Student-t calibration of that test is a parametric approximation, not an exact finite-sample test for ranks. The test compares the two correlations with a common index,

$$\mathcal{H}_0 : \rho_{\text{synth,pipe}} = \rho_{\text{static,pipe}} \quad \text{against} \quad \mathcal{H}_A : \rho_{\text{synth,pipe}} > \rho_{\text{static,pipe}},$$

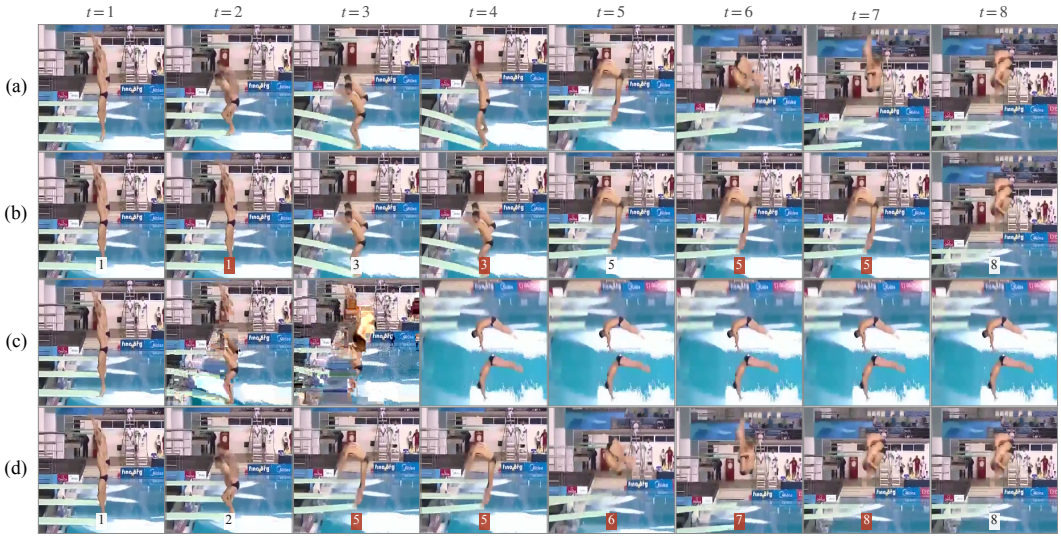


Figure 5: Eight consecutive Diving48 frames under severity-5 temporal-structure corruption. (a) Clean input. (b) Synthetic hold-last frame drop, with undamaged source frames. (c) H.264 packet loss, whose decoder concealment creates block artifacts and shifts the timeline. (d) Container-level VFR, where fixed-rate playback holds the previous frame. Digits in (b) and (d) give source-frame indices and are tinted when displaced; packet-loss content has no resampling-rule index.

and rejects $\mathcal{H}_0$ with $t = 2.86$, 34 degrees of freedom, and one-sided $p = 0.0036$. A nonparametric paired model bootstrap of the difference agrees after resampling the model set, giving $\Delta\rho = 0.44$ with 95% percentile CI $[0.07, 0.80]$ excluding zero and one-sided bootstrap tail fraction 0.0083.

Other constructions and the compression-adjusted residual. On the tx and UCF101 constructions, the raw gaps are small, at 0.83 compared with 0.80 on Diving48-tx and 0.75 compared with 0.61 on UCF101 (Table 9), and the dependent-correlation tests do not reach significance. After normalization by the matched compression baseline described below, the Diving48-tx residual gap also becomes significant. The two residual correlations are 0.80 for the synthetic measurement and 0.57 for the static measurement, with Williams $p = 0.0094$. The primary 3D comparison is supported by both the Williams test and bootstrap interval. For the Diving48-tx residual, the bootstrap $\Delta\rho$ CI $[-0.05, 0.56]$ narrowly includes zero, with one-sided tail fraction 0.0648, so the evidence comes from the Williams test.

Per-corruption analyses. Among the 3D models, the gap is significant for packet loss alone. The synthetic correlation is $\rho = 0.68$, compared with the static correlation 0.32, with Williams $p = 0.0269$ raw and 0.0203 residual. The correlation gap is also significant for VFR alone: The synthetic correlation is 0.80, compared with the static correlation 0.30, with $p = 0.0013$ raw and 0.0001 residual. These per-type tests are unadjusted; the Diving48-tx replication concerns the pooled residual criterion.

The appearance-regime control. On the appearance task, we observed that pipeline-level temporal corruption is near-harmless, with mean retention 0.975. This result supports the interpretation that the appearance-regime dynamic null reflects the task rather than the synthetic operators.

Compression-baseline adjustment. Pipeline-level packet loss and VFR are produced by an H.264 pipeline, so each mixes a temporal perturbation with the compression artifacts of the encoding and decoding roundtrip. Since the image-style static reference predicts compression well, this component could inflate the static→pipeline correlation. The codec cross-validation in Table 8 gives static $\rho = 0.92$ on Diving48. Because ratios and ranks combine components nonlinearly, the direction of that inflation must be checked empirically. We adjust for that component with a matched compression-only baseline: the identical roundtrip with the temporal perturbation switched off, using 0% packet loss and uniform frame timing. The baseline is near-lossless on the evaluation video clips, with

Table 9: Cross-validation on the temporal-structure axis. The criterion combines H.264 packet-loss corruption and VFR judder. The synthetic temporal-family predictor and image-style static reference are compared against this shared pipeline criterion. Among the 3D models, the correlations with the pipeline criterion are 0.78 and 0.34, with Williams $p = 0.0036$.

Model set	Predictor → pipeline	$\lvert\mathcal{M}\rvert$	ρ	model-set CI	concord.
UCF101 (avg)	synthetic temporal-structure	17	0.75	[0.44, 0.94]	0.79
	static reference		0.61	[0.17, 0.87]	0.73
Diving48 (tx)	synthetic temporal-structure	27	0.83	[0.56, 0.95]	0.84
	static reference		0.80	[0.55, 0.93]	0.81
Diving48 (3D)	synthetic temporal-structure	37	0.78	[0.57, 0.90]	0.79
	static reference		0.34	[0.02, 0.60]	0.62

mean retention 0.995. We normalize each pipeline score by this baseline to obtain the temporal-only residual, whose mean is 0.804. This ratio removes a model-specific compression baseline under a multiplicative decomposition.

The adjustment sharpens the divergence. We observed that the divergence not only survives this adjustment but sharpens: After normalization by the compression component, the static reference's correlation falls from 0.34 to $\rho = 0.25$, with CI $[-0.07, 0.52]$, while the synthetic temporal family's correlation holds at $\rho = 0.84$, with CI $[0.67, 0.91]$, as Table 10 shows. The dependent-correlation difference test is correspondingly stronger, with $\Delta\rho = 0.59$, one-sided Williams $p = 0.0001$, and bootstrap $\Delta\rho$ CI $[0.25, 0.93]$. Because this residual is a ratio of two retention ratios, we also verify an additive residual computed as full minus baseline. The additive residual gives the same conclusion: The synthetic correlation is $\rho = 0.84$, compared with the static correlation 0.24, with Williams $p = 0.0001$. The synthetic temporal measurement predicts the pipeline ranking better under both the ratio and additive residual definitions. Together these adjustments show that the gap between the static and temporal measurements is not explained solely by the model-specific codec baseline.

Codec and transport tests showed the same pattern. The static ranking predicts codec-compression robustness at $\rho = 0.92$, but predicts packet-loss and VFR-judder robustness at only $\rho = 0.57$ after controlling for compression damage, across the same 27 Diving48 tx models. The static reference reaches only $\rho = 0.25$ after the same compression control, much lower than the synthetic temporal-structure score's $\rho = 0.84$, among the 37 video-pretrained 3D models.

Interpreting the coefficient. The pre-specified $\rho_0 = 0.9$ of Eq. 4 is a substitution standard, asking whether a cheap measurement can stand in for an expensive one, and we do not claim substitution for the synthetic temporal family. The claim in Sec. 4 is comparative, namely that one predictor ranks models better than the incumbent against a shared criterion, which is what the Williams test above evaluates. Applied post hoc and symmetrically, the same pre-specified rule is nonetheless informative: Against the 37-model critical cutoff 0.83, we do not reject $\rho_* \geq \rho_0$ for the compression-controlled synthetic temporal measurement at 0.84; we reject the null hypothesis for the raw pooled measurement at 0.78 and for the image-style static reference on both criteria. Two references fix the scale of those coefficients. The prediction crosses a measurement gap, from an index-level tensor operation to a codec and transport stack with a decoder error-concealment path, and an architectural gap over 37 backbones drawn from three frameworks and spanning 22M to 632M parameters. Clean accuracy, the quantity practitioners routinely select on, transfers between our two motion datasets over an identical zoo at only $\rho = 0.43$ (Table 5), below the temporal-structure and motion-coupling coefficients reported there.

Cross-validation evidence. Together, these checks provide evidence that the synthetic taxonomy tracks the pipeline-level ranking on both the compression and temporal-structure axes. The static–temporal gap reappears under pipeline-produced codec and transport corruption, is supported by the rank-scale Williams test and model-bootstrap interval in the primary 3D comparison, and survives the compression control.

Table 10: Compression-controlled temporal-only residual. Pipeline retention is normalized by a matched compression-only baseline with 0% loss and uniform timing. Among the 3D models, the synthetic temporal family reaches $\rho = 0.84$, compared with 0.25 for the image-style static reference.

Model set	Predictor → temporal-only	$\lvert\mathcal{M}\rvert$	ρ	model-set CI	concord.
UCF101 (avg)	synthetic temporal-structure	17	0.65	[0.24, 0.87]	0.73
	static reference		0.53	[0.13, 0.77]	0.70
Diving48 (tx)	synthetic temporal-structure	27	0.80	[0.58, 0.91]	0.81
	static reference		0.57	[0.19, 0.84]	0.72
Diving48 (3D)	synthetic temporal-structure	37	0.84	[0.67, 0.91]	0.83
	static reference		0.25	[−0.07, 0.52]	0.58

L PRESCRIPTION DETAILS

This appendix records the per-cell evidence behind Sec. 4.3. A cell is one model set evaluated against one pipeline-level corruption family. Each cell compares two predictors of the same criterion, the image-style static reference and the blend of Eq. 7 at $\lambda = 1/2$, so the difference between the two predictors is an overlapping dependent correlation and is tested with the Williams T_2 test of App. K.2.

What the probe battery contains. The battery is fixed in advance by the representability rule rather than selected per cell, and comprises every corruption type of the two families built around cross-frame behavior that one frame does not fully represent, at the single severity $q = 3$ under the video protocol: frame drop, frame repeat, frame shuffle, VFR, and temporal jitter from the temporal-structure family, and camera shake, time-varying blur, and time-varying brightness from the motion-coupled family. Those eight conditions are aggregated by Eq. 2 within each family and then averaged across the two families, so neither family is overweighted by having more types. The image sweep the battery is compared against is the frame-independent family's thirteen spatial types over the full severity grid $\{1, 3, 5\}$. Where a head makes a condition degenerate, as an order-invariant consensus does for frame shuffle, that condition contributes no cross-model variance and the battery is unaffected in rank terms.

Where the probe helps. On pipeline-level temporal corruption, we observed that the blend improves prediction in every one of the 4 matched cells. The improvement is largest where the static reference is weakest, moving from $\rho = 0.34$ to 0.74 on the 37-model 3D zoo, for a one-sided p-value of <0.001, and smallest where the static reference already works, moving from 0.61 to 0.71 on the appearance-regime UCF101 set, where the gain is not significant at $p = 0.254$.

The gain follows representability. On motion-coupled pipeline corruption the blend improves 4 of 5 cells, with 1 significant. The smaller gain is consistent with the representability rule: One frame already shows part of a motion-coupled corruption (App. I), so the static reference leaves less for the probe to add, whereas temporal structure, which no frame shows, is where the probe adds most.

Representable-corruption control. The instrument targets corruptions that a single frame cannot represent. For pipeline-level H.264 and H.265 compression, which is representable in one frame, the blend does not improve prediction: mean $\Delta\rho = -0.20$.

The blending weight is not tuned. Pooled fidelity rises from 0.63 at $\lambda = 0$ to 0.78 at $\lambda = 1/2$ and is flat across the moderate blends. We report $\lambda = 1/2$ rather than the sweep's argmax for that reason, so the reported setting is not a selected operating point. For deployment, the weight should reflect how much of the expected corruption involves frame timing or other cross-frame behavior that one frame cannot fully show.

Cost. The battery is eight conditions at one severity. Under the video protocol, the battery needs roughly $1/5$ of the forward passes of the full corruption-by-severity image sweep, and Fig. 6 places each evaluation battery on the plane of evaluation cost against fidelity to non-representable pipeline-level corruption. Adding the probe to the image sweep therefore increases the sweep's forward-pass

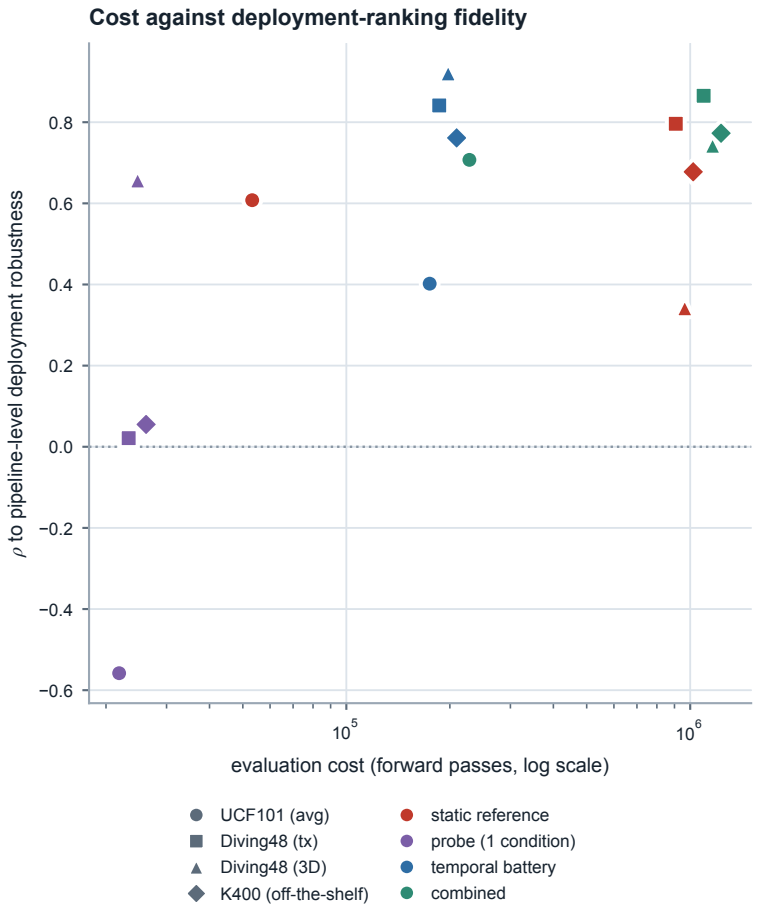


Figure 6: Cost against fidelity for model selection. Each point is one evaluation battery on one cell, with evaluation cost in forward passes on the horizontal log axis and the rank correlation to non-representable pipeline-level robustness on the vertical axis. On the three zoos scored under the video protocol, the one-severity temporal battery Pareto-dominates the image sweep for that target, and the blend of Eq. 7 further improves fidelity at the cost of both instruments, which is the cost side of the measurement asymmetry reported in Sec. 4.

cost by roughly $1/5$ when the image sweep and probe both use the video protocol. On every zoo scored under the video protocol, the battery alone reaches higher fidelity than the image sweep at that lower cost, so the image sweep is Pareto-dominated for that target, and blending the two rankings further improves fidelity at the cost of both instruments. On UCF101, where the image sweep is a single-frame protocol and pipeline-level temporal corruption is near-harmless, the image sweep remains the cheaper instrument. Across 8 model-set constructions, a single probe condition predicts the synthetic frame-timing ranking at median $\rho = 0.83$ while the full image sweep reaches median 0.40 at far greater cost. Because this single-condition result uses a synthetic target, the prescription recommends the more comprehensive eight-condition battery.

Scope. The prescription is a screening instrument for model selection before deployment and is not a substitute for testing in the deployment environment. The prescription's evidence is the matched cells above, and the prescription's validity beyond those cells rests on the representability rule rather than on a guarantee.

Table 11: Static→video rank correlation among 27 public K400 checkpoints from three frameworks, evaluated with native heads and no fine-tuning by us. The critical cutoff is 0.81 for this checkpoint set.

Family	ρ	model-set CI	reject $\rho_0 = 0.9$	concord.
Temporally coherent	1.00	[0.98, 1.00]	✘	0.99
Temporally drifting	0.93	[0.80, 0.98]	✘	0.90
Temporal structure	0.45	[0.10, 0.73]	✔	0.67
Motion-coupled	0.85	[0.66, 0.94]	✘	0.84

M Off-the-shelf operating-point replications

The two off-the-shelf replications use an appearance-leaning K400 zoo and a temporally engaged SSv2 zoo, with each checkpoint scored at the published operating point without fine-tuning by us.

Task and head controls for the low-signal regime. The same operator and severity that leave model-zoo mean retention at or above 0.83 on the appearance task remove almost two thirds of clean accuracy on Diving48 (Sec. 3), indicating that the contrast is not explained by operator strength alone. The order-invariant consensus head also contributes to low signal under reordering, but the head alone does not explain the low-signal regime: Order-sensitive, video-pretrained 3D models on K400 yield a model-discriminating temporal-structure measurement, for which we reject the proxy criterion at $\rho = 0.45$. We describe the near-zero-spread task–head combination as low-signal and non-diagnostic.

M.1 Public K400 checkpoints, no fine-tuning

The matched evaluations above use models fine-tuned at a common single-crop operating point. To test whether the static→dynamic divergence depends on that operating point, we repeat the 3D-among-3D rank test on 27 public K400 checkpoints evaluated at the published operating points: Each model keeps the original 400-class classification head from the checkpoint and is scored on the official K400 validation split, with no fine-tuning by us. The set spans both 3D CNNs and 3D Transformers and the same three frameworks of Torchvision, PyTorchVideo, and Hugging Face, including a VideoMAE scale sweep from 22M to 632M parameters; because the three frameworks ship three different 400-class permutations, each model's native head is remapped to a single canonical class order by class name. A clean-accuracy gate requires top-1 ≥ 0.15, well above the 0.0025 chance level, and verifies the remap of every model. The checkpoints' clean top-1 runs from 36.8% to 82.8%, with a mean of 66.7%, at the published operating points.

Among these 27 public checkpoints, the pre-specified Fisher-z critical cutoff is 0.81 at $\rho_0 = 0.9$. We observed that the static frame-independent ranking transfers almost perfectly on the static-magnitude coherent family, with $\rho = 1.00$ and CI $[0.98, 1.00]$ (Table 11). On this appearance-leaning K400 zoo, the dynamic divergence is attenuated, consistent with the appearance-regime results: Drift and motion coupling remain well predicted by the static ranking. Temporal drift has $\rho = 0.93$ and CI $[0.80, 0.98]$; motion has $\rho = 0.85$ and CI $[0.66, 0.94]$. We do not reject $\rho_0 = 0.9$ for either family, whereas drift and motion coupling both diverged on the more motion- and order-dependent fine-tuned Diving48 and SSv2 sets. But the divergence does not vanish: The failure concentrates on precisely the family a single frame cannot represent, temporal structure, where the static ranking falls to $\rho = 0.45$. The CI is $[0.10, 0.73]$, concordance is 0.67, and we reject $\rho_0 = 0.9$ with one-sided $p < 10^{-3}$.

We reject the pre-specified $\rho_0 = 0.9$ criterion for 1 of the three dynamic families. Because no weight here was tuned by us, the temporal-structure divergence cannot be an artifact of our matched fine-tuning operating point. The complete per-family rank correlation is in Table 11.

M.2 Public SSv2 checkpoints on a temporally engaged task

The K400 zoo above is appearance-leaning, so dynamic corruption discriminates the K400 models weakly. For a temporally engaged comparison, we use a second off-the-shelf zoo of public SSv2

Table 12: Static→video rank correlation among 16 public SSv2 checkpoints, scored with native 174-class heads and no fine-tuning by us. The set spans temporal CNNs and space-time Transformers; the critical cutoff is 0.77. Coherent corruption and drift transfer, temporal structure diverges, and motion is a boundary case whose CI reaches 0.9 although the one-sided test rejects the threshold.

Family	ρ	model-set CI	reject $\rho_0 = 0.9$	concord.
Temporally coherent	0.99	[0.93, 1.00]	✘	0.98
Temporally drifting	0.90	[0.68, 0.96]	✘	0.85
Temporal structure	−0.47	[−0.81, 0.09]	✔	0.34
Motion-coupled	0.76	[0.37, 0.94]	✔	0.79

checkpoints, fine-tuned by the original authors rather than by us. SSv2's classes are defined by the direction and order of motion.

We repeat the rank test among video models on 16 public SSv2 checkpoints. The set comprises cleanly loadable, native-SSv2-head video recognizers from two standard public ecosystems, the Hugging Face Hub and the MMAction2 model zoo (Contributors, 2020). Other available checkpoints use K400 heads or self-supervised backbones without an SSv2 classifier and therefore do not meet the native-head inclusion criterion. We score the selected models at each model's published operating point, keeping the original 174-class head, evaluated on the official SSv2 validation split, with no fine-tuning by us.

The set spans several architecture families. Six are 2D temporal-CNN recognizers: TSN (Wang et al., 2016), Temporal Relation Network (Zhou et al., 2018), and Temporal Shift Module (Lin et al., 2019) at two depths and two video clip lengths. Ten are space-time Transformers: MViTv2, TimeSformer, VideoMAE, and Video Joint Embedding Predictive Architecture 2 (V-JEPA 2) (Assran et al., 2025). Unlike the K400 zoo, this set also spans several distinct pretraining paradigms: supervised training for the 2D CNNs, TimeSformer, and MViTv2; masked autoencoding for VideoMAE; and joint-embedding self-supervision for V-JEPA 2. Checkpoint sizes run from a 22M VideoMAE-Small to the 1B-parameter V-JEPA 2 ViT-g.

A per-model clean-accuracy gate requires each checkpoint to reach $0.70\times$ the checkpoint's published top-1, far above the 0.006 chance rate over 174 classes, and verifies every native-head remap. Clean top-1 runs from 35.5% to 71.6%, with a mean of 57.5%, at the published operating points.

The result follows the same pattern as the other evaluations in Table 12. The static frame-independent ranking transfers perfectly on the static-magnitude coherent family, with $\rho = 0.99$ and CI $[0.93, 1.00]$. The static reference ranks models more consistently on drift and motion coupling than on temporal structure. Temporal drift transfers, with $\rho = 0.90$ and CI $[0.68, 0.96]$, and we do not reject $\rho_0 = 0.9$. Motion is an intermediate case, as a partly represented family should be, with $\rho = 0.76$ and CI $[0.37, 0.94]$ whose upper end still reaches the bar, consistent with the attenuation seen in the larger 3D cells.

But we observed that on temporal structure, the one family a single frame cannot represent, the static ranking again drops to $\rho = -0.47$. The CI is $[-0.81, 0.09]$, concordance 0.34 is near chance, and we reject $\rho_0 = 0.9$ with one-sided $p < 10^{-3}$ at critical cutoff 0.77, even at these published operating points: The correlation estimate and the entire bootstrap interval for that correlation lie far below the null boundary. Model inclusion is based on clean loadability and native-head compatibility, independently of any robustness ranking.

The practitioner cost is concrete: A static-guided pairwise choice can miss by up to 0.17 in frame-timing retention, and the static-best model ranks seventhof 16 on temporal structure. Because no weight here was tuned by us and the task is strongly order-dependent, this evaluation provides the temporally engaged off-the-shelf control. Across the matched tx-head, video-pretrained 3D fine-tuned, and off-the-shelf evaluations on Diving48, SSv2, and K400, temporal structure is the consistent failing family and the one a static benchmark cannot measure directly. The complete per-family rank correlation is in Table 12.

N DENSE ONLINE REGIME: VOS LABEL-PROPAGATION PROTOCOL

Protocol. On DAVIS-2017 (Pont-Tuset et al., 2017), we evaluate semi-supervised VOS by training-free label propagation, introduced for self-supervised correspondence by Vondrick et al. (2018) and since the established protocol for probing a frozen encoder's spatio-temporal correspondence (Jabri et al., 2020; Caron et al., 2021). Each backbone is used only as a fixed feature extractor with no fine-tuning, so the regime reuses the matched-pair philosophy: We change the task and protocol, not the weights.

Descriptors, propagation, and metric. For frame t, we extract dense, ℓ_2-normalized descriptors at stride 16. We use the patch-token grid for ViT and DeiT and the stride-16 stage exposed through `features_only` for CNNs, with a long side of 640. The first-frame ground-truth mask is downsampled to the feature grid and propagated by soft k-nearest-neighbor voting against a memory bank comprising the first frame as a permanent clean anchor and the seven most recent predicted frames. Affinities are cosine similarities restricted to a spatial neighborhood of radius 12. We select the top $k = 10$, apply softmax weighting at temperature 0.07, and use the result to vote on object labels, which are then written back into the bank. Quality is the region similarity $\mathcal{J}$, the primary DAVIS metric, averaged over objects and frames. That metric is also called mask intersection-over-union (IoU). We omit the boundary $\mathcal{F}$-measure because that metric does not affect the relative robustness transfer we study.

Corruptions and the static reference. Corruptions are applied to the frame stream exactly as in the classification regimes: a representative spatial operator per perceptual group for the photometric families, plus the drift, structure, and motion families. The first-frame annotation remains clean because the annotation is the task's given reference. The static reference is the frame-independent family.

Rank transfer under corruption. Clean mask IoU reaches 0.518 over the 28 frozen backbones. We score per-family IoU retention as before, with the frame-independent family as the reference. We observed three recurring findings in Table 13. Static-corruption robustness transfers essentially perfectly, at coherent $\rho = 0.99$. Temporally dynamic corruptions become consequential, often severe vulnerabilities the image-style static reference cannot represent: At severity 5, frame-shuffle retains only 0.41 and camera-shake 0.58 of clean IoU, in contrast to the far milder effect of these corruptions on appearance classification, because propagation depends on temporal continuity. The static→dynamic rank transfer weakens: Temporal structure falls to $\rho = 0.00$ and motion coupling to $\rho = 0.66$, with permutation $p = 0.0002$. Both estimates are below $\rho_0 = 0.9$ and the $n = 28$ critical correlation 0.82. Drift barely perturbs propagation here, with retention 0.975, and still transfers at $\rho = 0.83$.

What the dense-task evidence shows. The structure-family correlation estimate is centered near zero, the model- and clip-level intervals for that correlation lie far below ρ_0, and the permutation test finds no positive association, while rank agreement drops moderately under motion coupling. The dense task thus reproduces the motion-task conclusion through an independent mechanism based on correspondence propagation rather than a video clip aggregator. With $n = 28$ frozen backbones, after Holm correction across the three dynamic families, the tests reject the proxy for 2 families: temporal structure, whose video ranking has near-chance concordance with the static ranking, and motion coupling. Drift, which every frame shows at some magnitude, is the one dynamic family that still transfers, so the graded split recurs under a mechanistically different task. Fig. 7 isolates the stateful mechanism in one transient-fault diagnostic: Once clean pixels resume, a memoryless first-frame control recovers immediately, whereas the seven-frame predicted-mask memory continues to propagate the corrupted mask state.

O SCOPE AND LIMITATIONS

Interpretive scope. The representability rule is a graded design principle: The more of a corruption one frame shows, the more of the video ranking the image-style score preserves. The seven settings of Table 1 support the rule on the evaluated corruption axes, the positive codec control and the packet-loss and VFR tests show the rule in concrete pipelines, and Theorem 1 establishes that per-frame

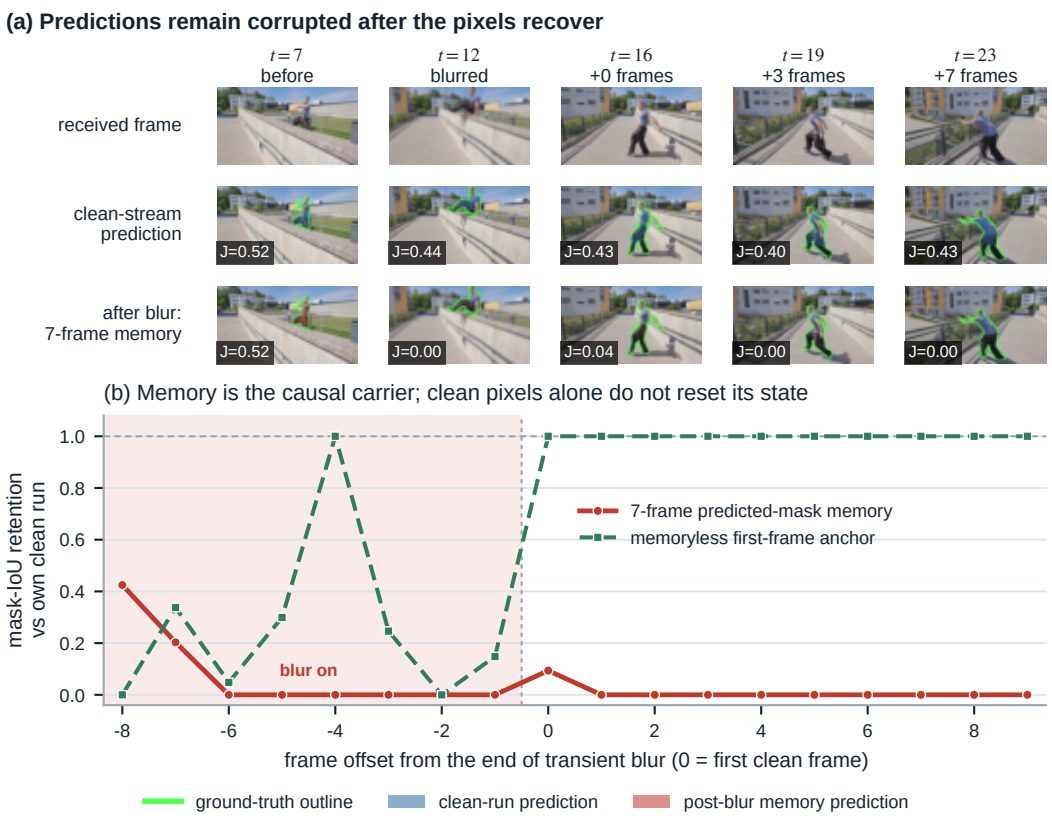


Figure 7: A transient visual fault becomes a persistent state error on the DAVIS-2017 parkour sequence under frozen ResNet-50 label propagation. This sequence was selected as the largest ten-frame post-corruption retention gap among all 30 validation sequences. (a) Severity-5 Gaussian blur affects only frames 8 through 15; after clean pixels resume, the seven-frame memory continues to propagate an erroneous mask. (b) IoU is normalized to each propagator's clean-input run. The memoryless control immediately returns to 1.0, whereas the seven-frame memory averages 0.009 over the next ten frames. This selected worst case illustrates state contamination and is not an estimate of prevalence.

Table 13: Static→video rank correlation by corruption family on the dense online DAVIS VOS task with 28 frozen backbones, in the columns of Table 3 together with a permutation p. Static-magnitude corruption transfers nearly perfectly, with coherent $\rho = 0.991$, while temporal-structure and motion fall below $\rho_0 = 0.9$. Cross-model score spreads are 0.032 for the static reference and 0.022, 0.018, and 0.018 for the drifting, temporal-structure, and motion-coupled families.

Video score	ρ	model CI	clip CI	τ_b	reversals	permutation p
Temporally coherent	0.991	[0.96, 1.00]	[0.94, 0.99]	0.936	0.03	< 0.001
Temporally drifting	0.829	[0.62, 0.93]	[0.54, 0.86]	0.661	0.17	< 0.001
Temporal structure	0.004	[−0.40, 0.38]	[−0.13, 0.41]	−0.005	0.50	0.506
Motion-coupled	0.658	[0.37, 0.83]	[0.36, 0.79]	0.476	0.26	< 0.001
Overall	0.929	[0.81, 0.97]	[0.80, 0.96]	0.794	0.10	< 0.001

marginals cannot guarantee the correct video ranking without assumptions on temporal coupling. The transport, codec, and analytics literature cited in App. F.1 establishes that temporal-structure faults occur in deployed pipelines, and our experiments establish what such faults do to model rankings. We do not estimate how often each fault occurs in a given deployment, so the weight placed on temporal

measurement should follow the fault mix of the target pipeline. A deployment may contain other couplings, recovery mechanisms, or feedback loops and should be evaluated directly.

Empirical coverage. Our coverage is broad but not exhaustive. We study video clip classification on UCF101, HMDB51, Diving48, and SSv2 and dense online segmentation on DAVIS, using matched backbone pools, 37 video-pretrained 3D models, and two public checkpoint zoos. These settings do not cover every video task, architecture, acquisition system, or transport stack. The reported correlations also describe finite candidate sets: Changing the candidate pool can change a rank coefficient even when the underlying condition scores are unchanged. We therefore report model-set and clip-level uncertainty and composition-sensitivity analyses rather than treating any coefficient as universal.

Use of the probe. The temporal probe in Sec. 4.3 is a screening instrument, not a safety certification or a substitute for deployment testing. The probe is validated against matched pipeline-level cells, and the temporal-structure component of the probe is stronger than the motion-coupled component. Transfer to an unmeasured deployment depends on whether the probe covers that deployment's dominant cross-frame failure modes.

P Appearance-regime results, collinearity, and motion-coupling cost

Tables 15 and 16 collect the supporting robustness checks referenced from the main text: the UCF101 measurement-collinearity matrix behind the mechanical half of the appearance-regime transfer (App. P.1) and the baseline-strength sensitivity of the video-pretrained 3D temporal-structure result (App. H). Fig. 8 visualizes the appearance-regime rank correlations of Table 14.

P.1 Appearance regime: static transfer and little damage from dynamic corruption

Table 14 reports Spearman's correlation $\rho^{(f)}$ between the static and video rankings for each corruption family over the 17-model UCF101 set, with both a model-set bootstrap CI and a paired clip-level bootstrap CI; Fig. 8 visualizes both intervals. The results fall into two groups, described below: spatial corruptions applied at constant strength, and corruptions that vary in strength over time or disrupt the frame sequence or motion.

The static ranking transfers when corruption strength stays constant. Frame-independent corruption samples a new random corruption pattern for each frame, whereas temporally coherent corruption reuses the same pattern throughout the clip. Both corruption families keep corruption strength fixed. In both cases, the static ranking transfers almost perfectly: The respective correlations are $\rho = 0.98$ and $\rho = 0.99$, with reversal fractions 0.04 and 0.03 and the intervals in Table 14. These families reduce accuracy by roughly 13% to 14% relative to clean performance, with mean retention 0.868 and 0.865, respectively. The frame-independent score spread is 0.072, so the high correlations occur where models differ substantially in how much accuracy each model retains. Temporal coherence here means that the same corruption pattern persists across frames; this persistence barely changes the static ranking. Models that are more robust to independently sampled spatial corruption remain more robust when the pattern persists. The same pattern of observations recurs on HMDB51. The image and frame-independent video scores both measure robustness to the same spatial corruptions with the same model weights, differing in whether predictions come from one frame or are averaged across frames. The coherent comparison additionally tests whether sharing the corruption pattern across frames changes the ranking.

Dynamic corruptions barely reduce accuracy or separate models here. Rank agreement drops for the three dynamic families: Drifting corruption gives $\rho = 0.35$ with CI $[-0.18, 0.72]$, temporal-structure corruption gives $\rho = 0.42$ with CI $[-0.06, 0.74]$, and motion-coupled corruption gives $\rho = 0.20$ with CI $[-0.31, 0.61]$. The reversal fraction rises toward 40%. If retention scores are nearly equal across models, choosing a different model offers little robustness benefit. We count a weak correlation between the static and video rankings as a proxy failure only when video corruption

Table 14: Image-to-video rank correlation by corruption family over the 17-model UCF101 set, in the columns of Table 3 together with a permutation p. The predictor is the image-style static reference of Sec. 2, which is the literal single-frame image score in this appearance regime. The static-magnitude families above the gap transfer almost perfectly, while the temporally dynamic families below the gap carry near-zero cross-model score spread, at 0.010, 0.002, and 0.006 against 0.072 for the frame-independent family, which identifies that cell as low-signal and non-diagnostic rather than as a proxy failure (App. P.1). Fig. 8 visualizes the same coefficients with both intervals.

Video score	ρ	model CI	clip CI	τ_b	reversals	permutation p
Frame-independent	0.978	[0.88, 1.00]	[0.94, 0.99]	0.912	0.04	< 0.001
Temporally coherent	0.988	[0.91, 1.00]	[0.94, 0.99]	0.941	0.03	< 0.001
Temporally drifting	0.353	[−0.18, 0.72]	[0.10, 0.52]	0.250	0.38	0.085
Temporal structure	0.419	[−0.06, 0.74]	[0.02, 0.62]	0.265	0.37	0.044
Motion-coupled	0.201	[−0.31, 0.61]	[−0.23, 0.38]	0.162	0.42	0.205
Overall	0.983	[0.91, 1.00]	[0.93, 0.99]	0.926	0.04	< 0.001

produces clear retention differences between models. We fixed this rule before computing any rank correlations (App. E.1). Here, retention scores are nearly equal across models, with mean retention ≥ 0.976. The three score spreads are 0.010, 0.002, and 0.006, respectively, roughly 7 to 36 times smaller than the frame-independent spread. The model-set CIs are wide and all straddle zero, showing how much the estimated correlations vary with the models included in the UCF101 evaluation set. These low correlations alone do not show that image-based selection would choose a meaningfully less robust model in this setting. Sec. 3 tests the static ranking where temporal corruption causes substantial losses that differ across models.

The small performance differences here reflect the combination of the task and a head that averages predictions without using frame order. The same operators cause large losses on Diving48, and temporal-structure corruption separates video-pretrained 3D models that use frame order even on the appearance-leaning K400 task; there, we reject the hypothesis that the static ranking meets the required agreement with the video ranking (App. M).

P.2 Collinearity, HMDB51, and motion-coupling cost

The high static correlation is partly mechanical. The matched-pair construction may account for part of the $\rho \approx 0.98$ correlation for the static families in Table 14. Under a fixed set of weights and an order-invariant head, the single-frame image score s^{img}, the frame-independent video score, and the coherent video score are three measurements of nearly the same quantity at different T; the cross-model Spearman correlations among those three scores are indeed near one. The correlation between s^{img} and the frame-independent score is 0.98, while the correlation between the frame-independent and coherent scores is 0.99; see Table 15. We therefore do not interpret the frame-independent transfer as an independent discovery. The informative finding is the coherent result: Changing the temporal realization from independent to shared does not break the static ranking because coherence lowers all models alike.

The scope of this caveat is the appearance regime. The caveat rests on a fixed set of weights and an order-invariant head, under which the three static measurements effectively coincide, and therefore does not extend to the motion regime of Sec. 3, whose head is order-sensitive and whose task cannot be solved from one frame. The coherent row there is read as the control described in App. O rather than as a near-duplicate measurement.

Cross-dataset replication (HMDB51). The static-transfer side of the appearance-regime result is not an artifact of UCF101. On HMDB51 with 10 backbones, we observed that both the frame-independent and coherent rankings transfer at $\rho = 1.00$ and $\rho = 1.00$, respectively. The dynamic-family coefficients are less stable over this smaller model set. Temporal structure has $\rho = 0.45$ and is not used as standalone proxy-failure evidence.

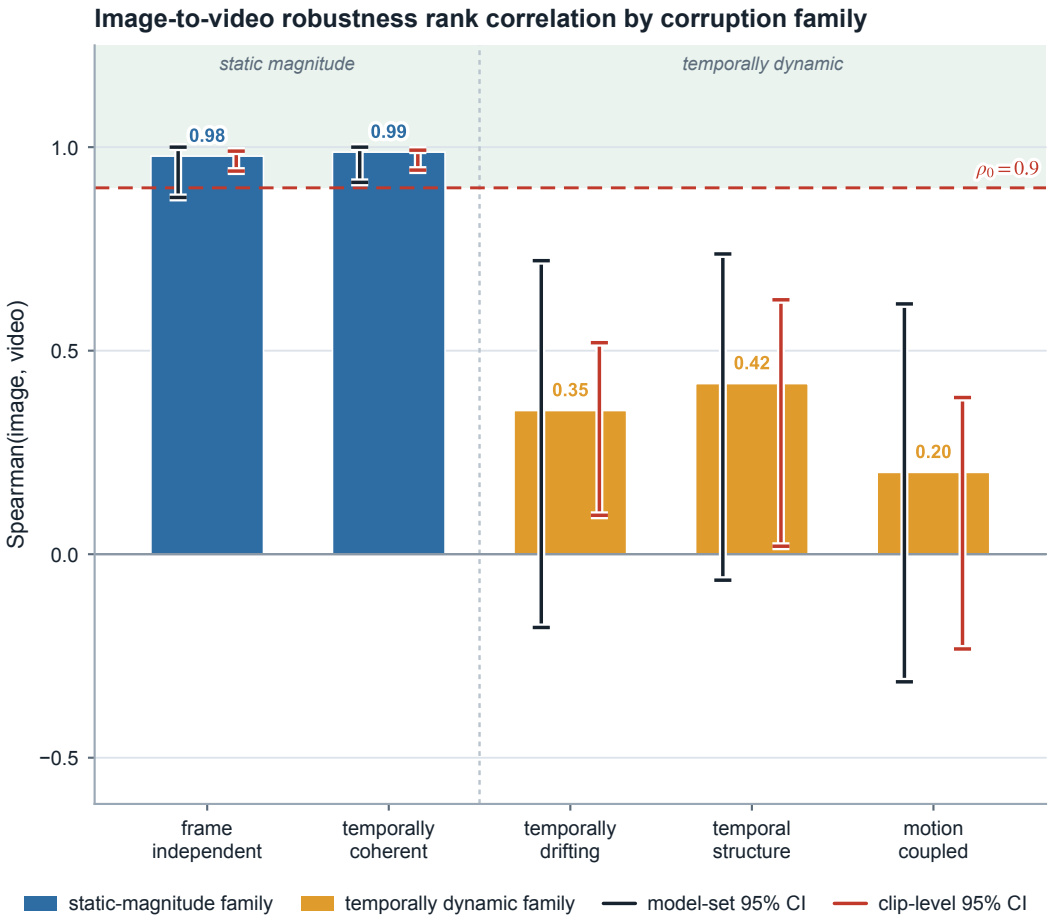


Figure 8: Spearman static→video rank correlation by family on UCF101. Bars are point estimates, colored by temporal stationarity. Dark left whiskers give model-set 95% CIs and red right whiskers give paired clip-level 95% CIs. The dashed line marks $\rho_0 = 0.9$ and the tinted region marks the passing range. Static families clear the threshold; dynamic families have low estimates and very small cross-model score spread, identifying a low-signal cell.

Motion-coupling selection cost. The cost here is subtler: The static and dynamic rankings are pairwise discordant on a fraction 0.21 of pairs. Concordance is 0.79, well short of a perfect proxy, and the conditional tail shortfall reaches 0.064. Yet the single top-1 static pick happens to coincide with the motion-best model, giving selection regret 0.000. The motion divergence is therefore a pairwise and tail phenomenon rather than a top-1 selection failure, a measurable but milder cost than temporal structure's. Even among video-pretrained models, the divergence persists and, for temporal structure, is costly in deployment terms, not merely below a bar.

Table 15: Cross-model Spearman correlations among the static measurements on UCF101, which form the mechanical half: The single-frame image score, the frame-independent video score, and the coherent video score are nearly collinear. The dynamic families are not. The rules and the row gap partition the matrix into static and dynamic blocks, so the near-unit upper-left block and the far weaker off-diagonal block can be read off directly.

	static measurements			dynamic families		
	s^{img}	f.indep	coher.	drift	struct	motion
s^{img}	1.00	0.98	0.99	0.35	0.42	0.20
f.indep	0.98	1.00	0.99	0.34	0.43	0.20
coher.	0.99	0.99	1.00	0.34	0.43	0.21
drift	0.35	0.34	0.34	1.00	0.60	0.92
struct	0.42	0.43	0.43	0.60	1.00	0.65
motion	0.20	0.20	0.21	0.92	0.65	1.00

Table 16: Baseline-strength sensitivity of the 3D temporal-structure result. We recompute the static→temporal-structure correlation under clean-accuracy floors and top-K restrictions. Higher-clean-accuracy subsets raise the correlation to about 0.6, but every tested subset remains below the pre-specified $\rho_0 = 0.9$ threshold; the largest observed ρ is 0.64.

Set	restriction	$\lvert\mathcal{M}\rvert$	min clean	ρ_{struct}	concord.
Diving48-3D	full set	37	0.465	+0.05	0.52
	top-25	25	0.649	+0.24	0.59
	top-20	20	0.707	+0.43	0.66
	top-15	15	0.728	+0.64	0.73
SSv2-3D	full set	37	0.244	+0.22	0.58
	clean ≥ 0.40	22	0.400	+0.58	0.74
	top-20	20	0.403	+0.59	0.74
	top-15	15	0.427	+0.63	0.78

Table 17: Per-model Diving48 clean top-1 and per-family video robustness retention scores. The frame-independent column is the static reference. Models are sorted by clean top-1, and the rows are spaced in groups of five so that a row can be tracked across the retention columns.

		video retention by corruption family				
Backbone (tx)	clean (%)	f.indep (ref)	coherent	drift	struct	motion
ViT-B/16 (CLIP-OA)	78.6	0.868	0.867	0.951	0.893	0.962
ViT-B/16 (IN21k)	75.9	0.896	0.895	0.963	0.893	0.974
ViT-B/16	75.3	0.887	0.885	0.951	0.894	0.960
ViT-B/16 (CLIP)	73.7	0.882	0.881	0.963	0.903	0.981
XCiT-S12	73.5	0.795	0.794	0.939	0.886	0.955
ViT-S/16	72.7	0.851	0.850	0.940	0.890	0.959
DeiT3-M	71.3	0.873	0.873	0.966	0.891	0.974
DeiT3-S	70.7	0.818	0.814	0.914	0.877	0.931
Twins-SVT-S	69.0	0.835	0.832	0.954	0.886	0.962
GC-ViT-T	68.1	0.839	0.838	0.965	0.884	0.974
Swin-S	67.2	0.807	0.806	0.914	0.884	0.934
ViT-Ti/16	66.9	0.806	0.806	0.928	0.875	0.942
RegNetY-080	64.5	0.662	0.660	0.925	0.877	0.940
RegNetY-016	64.4	0.719	0.717	0.924	0.882	0.948
Swin-T	63.8	0.732	0.730	0.883	0.881	0.919
DeiT3-B	63.8	0.840	0.840	0.909	0.882	0.951
EfficientNet-B1	62.7	0.717	0.714	0.895	0.878	0.920
RegNetY-040	58.3	0.599	0.594	0.871	0.880	0.921
ConvNeXt-T	58.0	0.786	0.784	0.894	0.878	0.942
ResNeXt-50	57.1	0.715	0.714	0.869	0.849	0.908
EfficientNet-B3	55.2	0.661	0.658	0.912	0.890	0.943
ResNet-34	54.8	0.629	0.626	0.866	0.841	0.895
EfficientNet-B0	54.2	0.631	0.632	0.898	0.883	0.919
ResNet-101	51.5	0.645	0.645	0.896	0.848	0.921
ResNet-18	48.7	0.656	0.654	0.853	0.856	0.905
ResNet-50	47.7	0.690	0.690	0.892	0.860	0.925
ConvNeXt-S	40.3	0.774	0.771	0.848	0.856	0.905

Table 18: 3D video-pretrained backbones on Diving48. The input length follows each model's configuration; native-32 wrappers upsample the same matched 16-frame content. The static column is the frame-independent reference named in Table 17. Among the dynamic families, temporal structure has the lowest retention in all but one model, as the structure column shows, while drift and motion are largely absorbed in retention, though the static→dynamic rank transfer still breaks for structure and motion (Table 4). Models are sorted by clean top-1, and the rows are spaced in groups of five as in Table 17.

		video retention by corruption family			
3D model	clean (%)	static	drift	structure	motion
VideoMAE-H	84.1	0.952	0.981	0.861	0.980
VideoMAE-L	79.9	0.928	0.970	0.834	0.970
X3D-L	79.5	0.892	0.971	0.790	0.962
Slow-R50	77.9	0.852	0.957	0.803	0.951
VideoMAE-B	76.5	0.902	0.957	0.816	0.955
EfficientX3D-XS	75.5	0.753	0.937	0.788	0.934
EfficientX3D-S	75.1	0.757	0.941	0.774	0.926
X3D-M	74.4	0.829	0.955	0.782	0.947
TimeSformer-B (SSv2)	74.0	0.870	0.958	0.860	0.963
VideoMAE-B (SSv2)	73.5	0.894	0.961	0.751	0.943
X3D-S	73.5	0.777	0.938	0.794	0.945
TimeSformer-B (K600)	73.3	0.903	0.967	0.890	0.973
TimeSformer-B	73.1	0.895	0.965	0.889	0.976
VideoMAE-B (short)	72.9	0.906	0.965	0.796	0.948
MViT-B 32×3	72.8	0.901	0.980	0.811	0.972
X3D-XS	72.0	0.742	0.941	0.774	0.937
VideoMAE-B (short, SSv2)	71.1	0.890	0.961	0.765	0.934
Video-Swin-B	71.0	0.856	0.970	0.901	0.968
MViTv2-S	70.9	0.923	0.979	0.838	0.968
Video-Swin-S	70.7	0.819	0.950	0.888	0.951
R(2+1)D-R50	69.8	0.794	0.927	0.803	0.926
ViViT-B	68.1	0.840	0.960	0.920	0.965
I3D-R50	67.6	0.826	0.928	0.821	0.936
Video-Swin-T	67.0	0.784	0.955	0.885	0.960
C2D-R50	64.9	0.742	0.882	0.857	0.902
MViTv1-B	64.3	0.886	0.959	0.796	0.949
MC3-18	63.0	0.606	0.877	0.879	0.921
VideoMAE-S	62.1	0.862	0.945	0.782	0.940
VideoMAE-S (SSv2)	59.9	0.841	0.925	0.724	0.913
MViT-B 16×4	59.9	0.878	0.961	0.794	0.955
R(2+1)D-18	57.3	0.569	0.851	0.832	0.878
SlowFast-16×8-R101	54.0	0.805	0.938	0.793	0.895
SlowFast-R101	53.5	0.801	0.955	0.841	0.902
SlowFast-R50	52.4	0.779	0.932	0.808	0.892
CSN-R101	51.0	0.795	0.898	0.828	0.907
S3D	50.6	0.713	0.863	0.846	0.884
R3D-18	46.5	0.647	0.867	0.804	0.889