ENCODED BUT UNREAD: LOCATING AND RECOVERING WITHIN-EVENT MOTION IN VIDEO LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Video large language models (Video-LLMs) answer questions about a video from sparsely sampled frames that a vision encoder turns into visual tokens for a language model to read. Questions about motion, such as how long an expression is held or whether a movement pauses, can only be answered by comparing what the model sees at different moments. Whether current Video-LLMs make this comparison, where the information is lost when they do not, and what recovery costs remain unclear. We introduce WARPPAIRS, a diagnostic benchmark of time-warp minimal pairs: each item re-indexes the frames of a real face, generated or action clip, so that positive and negative differ only in temporal structure and the label follows exactly from the edit. Six Video-LLMs perform at or near chance on all five edit types. The loss has two regimes. Below a threshold set by the temporal receptive field of a visual token, the sampler's frame stride times the encoder's temporal merge, a local scramble is attenuated for a linear probe on the tokens and invisible to the model, and the threshold moves with the architecture and with the sampling stride as predicted. Above the threshold the motion is recoverable from the tokens but absent from the answer: a linear classifier on difference statistics of the visual tokens reaches 72–90% balanced accuracy out of fold on three video sources, whereas the language model remains at chance. Supplying the missing comparison as *temporal-difference (TD) tokens*, computed from the model's own visual tokens, recovers the readout at a cost that depends on where the tokens enter and on the base model: at the decoder input they cut the fine-tuning steps to a fixed criterion up to six-fold at the same final accuracy, whereas the same signal added residually to the visual tokens is left at low weight by the model and saves nothing. The recovered readout transfers to unseen actors but only partly to other edits and sources. On controlled edits, these results separate a temporal-resolution scale set by sampling and encoding from a readout failure in the language model, and show that supplying the comparison saves most where frames-only fine-tuning stays longest in a constant-answer phase.

1 INTRODUCTION

Video large language models (Video-LLMs) such as Qwen2.5-VL (Bai et al., 2025b), InternVL3.5 (Wang et al., 2025) and LLaVA-OneVision (Li et al., 2024) have made rapid progress on video question answering, yet fine-grained motion understanding remains weak despite strong results on content-centric benchmarks. MotionBench (Hong et al., 2025), FAVOR-Bench (Tu et al., 2026) and TempCompass (Liu et al., 2024) report low scores on motion questions, and SpookyBench (Upadhyay et al., 2026) finds them at 0% where humans reach 98%. These models share a pipeline: a sampler selects a sparse set of frames, a vision encoder maps each frame (or pair of frames) to visual tokens, and a language model reads the token sequence together with the question. The question is at which of these stages the motion is lost.

Recent diagnostic studies have begun to ask *where* temporal information is lost. Probing classifiers on video representations recover event order almost perfectly while the model fails, which places the failure in the language model (Li et al., 2025a). Models fail to tell before from after (Bagad et al., 2023) and barely distinguish a video from its reversal unless trained to (Xue et al., 2025); tracing

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.

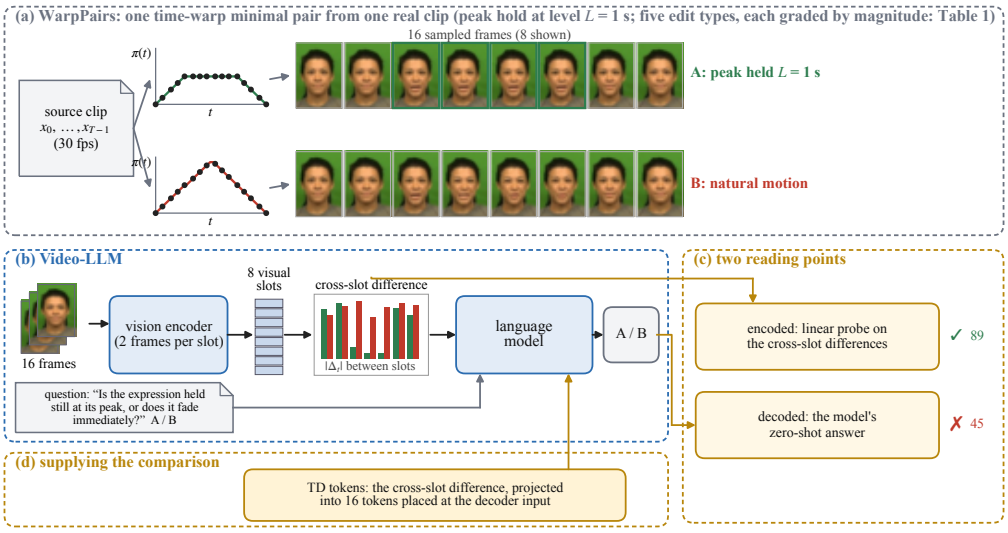

Figure 1: **Overview.** (a) A time-warp minimal pair on one CREMA-D clip, one of the five edit types (Table 1) at one of its levels: the same source frames re-indexed by two maps $\pi(t)$, one holding the apex for 1 s (A) and one with natural motion of the same length (B); only π differs, and the label is a property of π. (b) The pipeline on this pair; bars: magnitude of the first difference between consecutive visual slots of Qwen2.5-VL-7B for A (green) and B (red). (c) The two reading points: a linear probe on cross-slot differences of the visual tokens (*encoded*) and the zero-shot answer (*decoded*), with their balanced accuracy on peak hold. (d) TD tokens supply the same comparison at the decoder input.

places the loss at the projector (Han et al., 2026), and some models read order from the positional encoding rather than from the frames (Rahman, 2026). All of these studies examine *ordering* signals, such as the direction of time or the order of two events. The quantities that distinguish a genuine smile from a posed one are instead kinematic properties *within* a single event: its duration, its speed and the asymmetry between its onset and offset (Ekman & Friesen, 1982; Dibeklioğlu et al., 2012). Existing motion benchmarks annotate natural video, where labels are noisy and content cues leak, or synthesise pixels far from the training distribution. Whether within-event motion meets the same bottleneck, or an earlier one of its own, therefore remains open.

We address these questions with WARPPAIRS, a controlled diagnostic benchmark of *time-warp minimal pairs* (Fig. 1): because a time-warp only re-indexes the frames of a real clip, the pixels remain natural and the label, a property of the warp, is exact by construction. Each positive has a length-matched negative from the same clip. The five edit types (Table 1; examples in Fig. 5), a held peak, a mid-clip freeze, a uniform speed change, an onset–offset asymmetry and a local scramble, are each graded by magnitude and applied to face, generated and action video. On this benchmark we pose three research questions.

(RQ1) Does the sampling and encoding stage itself limit which temporal edits can be recovered from the visual tokens? We derive the temporal receptive field of a visual token from the sampler's native-frame stride and the encoder's temporal merge, and test its prediction on three architectures with two different receptive fields, and on one of them by changing the sampling stride alone. It does, and the threshold moves as predicted. **(RQ2)** Beyond that threshold, is within-event motion encoded in the tokens the language model receives, and is it decoded? We fit a linear probe on cross-token difference statistics of those tokens (*encoded*) and compare it with the model's zero-shot answer (*decoded*); four control tasks place the two on one scale. It is encoded but not decoded. **(RQ3)** If so, what does recovery cost, and does the cost depend on where the missing computation is supplied? We define the *recovery cost* as the fine-tuning steps needed to reach a fixed accuracy criterion under a fixed LoRA budget, a time-to-accuracy measure (Coleman et al., 2019), and compare it across input forms. It does: supplied at the decoder input the comparison cuts the cost up to six-fold; fused into the visual tokens it saves nothing. Our contributions are as follows:

- We introduce WARPPAIRS, a controlled diagnostic benchmark of time-warp minimal pairs on real video, with exact, graded labels for within-event kinematics, on which six Video-LLMs are at or near chance on all five edit types (§3, §5.2).
- We show that the failure has two regimes. Below a *detectability threshold* set by the receptive field, a local scramble is attenuated for our probes and invisible to the model; the threshold moves with the architecture and with the stride as predicted, and the shortest edits of the other types are those that fine-tuning recovers least (§5.3). Above it, within-event motion is *encoded but not decoded*: a linear probe separates the edits on three video sources while the language model answers at chance, and controls locate the probe's advantage in a comparison across temporal slots (§5.4).
- We quantify the recovery cost in the second regime, which depends on where the comparison is supplied: temporal-difference tokens at the decoder input reduce it two- to six-fold on Qwen2.5-VL at the same final accuracy, the same signal fused into the visual tokens gives no saving, and across three base models the saving is associated with the length of a constant-answer phase of frames-only training (§5.5–§5.7).
- We show that the recovered readout is narrower than the representation: under held-out edit rules and on two non-face sources the probe still reads the motion while the fine-tuned readout stays tied to the trained edit rule and domain (§5.7).

2 RELATED WORK

Temporal benchmarks for Video-LLMs. Benchmarks differ in whether their labels are exact. Annotated motion benchmarks (Hong et al., 2025; Tu et al., 2026; Liu et al., 2024; Zhang et al., 2024; Liu et al., 2026; Chen et al., 2026) report low scores but do not isolate the failure with exact labels; controlled diagnostics take exact labels from synthetic or concatenated video (Johnson et al., 2017; Girdhar & Ramanan, 2020; Kwan et al., 2026; Ghazanfari et al., 2026; Ahn et al., 2026), and minimal pairs (Warstadt et al., 2020; Thrush et al., 2022) have reached video (Krojer et al., 2025; Li et al., 2026). TempCompass builds such pairs by reversing or re-timing real videos (Liu et al., 2024), the arrow of time is detectable from motion features (Pickup et al., 2014), and self-supervised learning uses shuffled, reversed or re-timed clips as pretext tasks (Misra et al., 2016; Wei et al., 2018; Benaim et al., 2020; Yao et al., 2020). WARPPAIRS writes these and two edits without prior use as index maps on real clips, with a length-matched negative from the same clip, and studies the kinematics of a single event instead of the order of several.

Temporal resolution and frame sampling. The frame budget is ablated (Hong et al., 2025; Liu et al., 2026; Li et al., 2025b) rather than derived: SpookyBench finds models at 0% at every rate from 1 to 30 fps and concludes that sampling frequency does not explain their gap to humans (Upadhyay et al., 2026), and PAS (Sun et al., 2026a) stabilises the encoding against small changes of frame rate or sampling offset. None computes what an architecture can resolve (§4.1).

Probing temporal information in Video-LLMs. Prior work on where temporal information is lost concerns ordering signals. Probing places the loss of event order in the language model (Li et al., 2025a), and the arrow of time is poorly read even by current models (Xue et al., 2025; Han et al., 2026); contrastive supervision of the video tokens against reordered counterfactuals improves order reasoning (Shi et al., 2026), a study of claim deference separates whether two events were sampled from whether their order is trusted (Cao et al., 2026), and gradient-free layer merging repairs temporal reasoning lost in multimodal adaptation (Fu et al., 2026). We apply the same linear probes (Alain & Bengio, 2016; Belinkov, 2022) with control tasks (Hewitt & Liang, 2019) to within-event kinematics, separate that bottleneck from the loss below the detectability threshold, and measure what it costs to supply the comparison.

Motion-enhanced Video-LLMs. Three routes add motion to a Video-LLM. Train-free approaches render it into the frames: object spotlight and motion blur in MotionSight (Du et al., 2026), tracked-point trajectories in Motion-as-Prompt (Sun et al., 2026b). Data-centric approaches fine-tune on motion instruction data (MotionVid-QA and MotionChat, Du et al., 2026). Architectural approaches add a motion pathway: GMoT (Wang et al., 2026) computes adjacent-frame differences and injects them into the visual tokens through a gate, and COMET (Zhu et al., 2026) fuses a second encoder over Taylor frame differences into the appearance tokens by cross-attention. Codec

Table 1: The five edit types of WARPPAIRS. Each positive and its negative come from the same clip and have the same length; the level is the edit magnitude. The last column gives prior use of the same perturbation in video learning or evaluation; – where we found none.

Edit type	Positive: index map π	Negative	Levels	Prior use
peak hold	repeats the apex frame of an expression for L s	natural motion of the same length	L = 0.25, 0.5, 1 or 2 s	–
freeze	repeats one frame for L s in mid-clip	natural motion of the same length	L = 0.25, 0.5, 1 or 2 s	–
speed	subsamples or repeats frames uniformly (faster or slower)	the unedited clip (third class)	$\times\frac{1}{2}$, $\times 2$	Benaim et al. (2020); Liu et al. (2024)
asymmetry	plays a segment forward, then in reverse, at different speeds	the two speeds swapped	speed ratio 1.5, 2, 3 or 4	reversal: Wei et al. (2018)
scramble	shuffles frames within windows of w native frames	the unedited clip	w = 2, 4, 8, 16 or 32	Misra et al. (2016)

motion vectors (Sarkar et al., 2026), landmark-guided features for faces (Chaubey et al., 2026) and pose tokens for bodies (Markhorst et al., 2026) are further motion inputs. Our temporal-difference tokens are the simplest architectural instance: they difference the model's own tokens and add no new input; we measure what the addition costs rather than propose a design.

3 WARPPAIRS: TIME-WARP MINIMAL PAIRS ON REAL VIDEO

Task definition. Let a source clip be $x_0, \ldots, x_{T-1}$ at 30 fps. A time-warp is an index map $\pi : \{0, \ldots, T'-1\} \to \{0, \ldots, T-1\}$ giving the edited clip $y_t = x_{\pi(t)}$; each edit type is a parametrised family of such maps and the question asks about a property of π. A negative is the same clip under a map π' that preserves duration but not the property, so length, identity, expression and background are identical across classes and only temporal structure differs. Questions are two- or three-way multiple choice, answered by letter. Table 1 defines the five edit types, negatives and levels; Fig. 5 shows the frames of each and Appendix A the prompts.

Identifiability. The label is exact with respect to the edit; whether it is identifiable from the edited clip alone differs by type. A repeated frame, a permuted window and an onset–offset asymmetry leave a signature in the frames. A uniform speed change does not: without a reference it can be confused with a naturally faster or slower performance, so we report SPEED but exclude it from every cost comparison. PEAK HOLD and FREEZE are, by construction, repeated-frame detection at the sampled points, and a logistic regression on pixel differences of the sampled frames separates them (Table 22), so the benchmark tests whether a signal present in the pixels is read.

Sources, splits and sizes. Every source clip yields one item per edit type, of one of its answer classes, so items per edit type equal clips (Table 5). Faces come from CREMA-D (Cao et al., 2014) and Aff-Wild2 (Kollias & Zafeiriou, 2019), split by actor into 1,215 training and 404 test clips (2,020 test items from 85 actors), and from RAVDESS (Livingstone & Russo, 2018) (336 clips, every actor held out); generated video from VideoPhy-2 (Bansal et al., 2026) (300 clips) and human actions from UCF101 (Soomro et al., 2012) (399 clips) are test sets only. A fixed 400-item subset of the faces test (202 PEAK HOLD, 198 FREEZE) serves the checkpoint evaluations of §4.3; there is no separate validation split. A two-frame task shows one face at two times D native frames apart and asks whether it moved.

4 METHOD

A chance score on WARPPAIRS is compatible with three failures: the sampler never captured the edit, the encoder did not represent it, or the language model does not read a representation that holds it. Only the third is a readout failure; §4.1–4.3 separate the three.

4.1 THE TEMPORAL RECEPTIVE FIELD OF A VISUAL TOKEN

Two architecture parameters fix how many native frames one visual token summarises. The sampler selects n frames spread uniformly over the clip at no more than f fps, giving a native-frame stride s ($s = 4$–6 for 16 frames on the 2–3 s scramble items at 8 fps). The encoder merges m consecutive sampled frames into one temporal slot: $m = 2$ for Qwen2.5-VL and Qwen3-VL, whose patch embedding spans two consecutive frames, and $m = 1$ for InternVL3.5, which encodes each frame on its own. The temporal receptive field of a token is therefore $R = m\,s$ native frames: 8–12 for Qwen2.5-VL and 4–6 for InternVL3.5 at 16 frames. A permutation confined to a window $w < R$ changes which native frames a sample lands on but not which span a token summarises, and the perturbation it induces is of the order of the natural variation within $R/30$ s. We treat R as an empirical time scale, the visual token's analogue of the motion-aliasing limit of classical video analysis (Shechtman et al., 2005), and it yields a prediction before any measurement: scrambles with $w < R$ are attenuated and should go undetected, and detectability should rise as w grows to cover whole slots. The prediction concerns attenuation, not impossibility: a merged token could in principle encode change within its span, so whether an attenuated edit is recoverable must be measured. The same arithmetic applies to stills: a 0.25 s PEAK HOLD spans 7.5 native frames, inside R on Qwen2.5-VL, and a 0.5 s hold fifteen.

4.2 ENCODED VERSUS DECODED

Two scores are computed on the same benchmark: the *decoded* score is the model's zero-shot answer, and the *encoded* score is a logistic regression on the visual tokens the model feeds to its language model, after the vision tower and the temporal merge. We compute first and second differences Δ_t, Δ_t^2 across temporal slots, summarise each slot by magnitude and cosine-change statistics of $|\Delta_t|$ and $|\Delta_t^2|$, and pool the per-slot values over the clip (Appendix B). The probe is fitted with five-fold cross-validation grouped by actor on faces, by clip on VideoPhy-2 and by source video on UCF101, so every probe number is a held-out prediction. Its features add a comparison across slots and non-linear summaries of it, so we also report a strictly linear control on the mean consecutive difference vector alone. Four controls calibrate the two scores. *Shuffled labels*, a control task (Hewitt & Liang, 2019) with labels permuted across items, bound what the probe can fit by itself. *No differencing* applies the same statistics to the tokens themselves and tests whether differencing is needed. *Single slot* uses the mean token of one slot and makes no comparison. Two *static tasks* on the 286 CREMA-D test clips (actor sex and emotion) ask about a property that a single frame carries, so the decoded score should track the encoded one; this gives the scale of an ordinary readout gap.

4.3 RECOVERING THE COMPARISON UNDER A FIXED BUDGET

If what is missing is the comparison and not the information, supervision should recover it, and the cost should depend on whether the comparison is supplied or has to be learned. We fine-tune with LoRA (Hu et al., 2021) (r=16) on the language model under an identical objective and budget (Appendix B), varying only the form in which the comparison is supplied. *Frames*: the 16 sampled frames only. *Temporal-difference (TD) tokens*: seven first-difference statistics per consecutive slot pair, the magnitude and cosine-change terms of §4.2, computed from the model's own visual tokens and projected into K=16 tokens placed beside the frames at the first decoder layer; no external signal enters. *Trajectory text*: OpenFace-3.0 (Hu et al., 2025) action-unit, landmark and gaze curves written out as a short textual summary (Appendix B), a comparison computed outside the model. *Fused TD*: the same statistics added residually *into* the visual tokens, the placement of GMoT without its gate schedule. Controls are a train-free difference-map prompt, evaluation at 4 and 1 frames, further seeds and two further base models. We report steps to criterion, the first 600-step checkpoint at which the mean balanced accuracy over PEAK HOLD and FREEZE on a 400-item subset reaches 70, and the final accuracy on the full test (the 808 PEAK HOLD and FREEZE items of the faces test).

Table 2: Zero-shot balanced accuracy on the faces test (404 items per edit type, Table 5; chance 50, 33 for speed; †: one letter on at least 95% of items). All sources in Table 13.

Model	asymmetry	peak hold	speed	freeze	scramble
Qwen2.5-VL-7B	45.0†	44.7	33.3	51.6	49.9
Qwen3-VL-8B	44.6†	50.0	33.3	50.2	50.1
InternVL3.5-8B	50.2	50.0	33.3	51.3	49.5
LLaVA-OneVision-7B	49.7	52.8	36.1	52.1	49.3
VideoLLaMA3-7B	51.8	51.6	33.0	48.2	50.0
MiniCPM-V-4.5	50.3	50.0	33.3	48.4	49.8

5 EXPERIMENTS

5.1 EXPERIMENTAL SETUP

Models. Qwen2.5-VL-7B-Instruct (Bai et al., 2025b) for probing, fine-tuning and the two-frame task; Qwen3-VL-8B-Instruct (Bai et al., 2025a) for fine-tuning and the two-frame task; InternVL3.5-8B (Wang et al., 2025) for probing, the receptive-field test (§5.3) and the comparison of §5.7; LLaVA-OneVision-7B (Li et al., 2024) for the receptive-field test and zero-shot; VideoLLaMA3-7B (Zhang et al., 2025) and MiniCPM-V-4.5 (Yu et al., 2025) zero-shot only. Unless stated, 16 frames are spread uniformly over the clip at no more than 8 fps. **Training.** LoRA at learning rate 10^{-4} (sweep in Appendix D.3), in two regimes: one epoch over all five edit types (6,075 steps) or three epochs on the named types, PEAK HOLD and FREEZE unless stated (7,290 steps; 10,935 for the other three types), with checkpoints every 600 steps. TD tokens add under 2% to the cost of a training step. The schedule is cosine with warm-up over the optimiser steps, except in the learning-rate and rank sweeps of Appendix D.3, which hold the rate constant. **Metric.** Balanced accuracy in percent (mean of per-class recalls; chance 50, or 33 for SPEED). Its 95% half-width is ±4–5 points at n=404 and ±3–4 on the 808-item PEAK HOLD+FREEZE test (Appendix E.1); differences under 5 points are not interpreted. Throughout, † marks a constant answer, at least 95% of an edit type's items given the same letter (Table 21).

5.2 ZERO-SHOT RESULTS

Table 2 reports zero-shot balanced accuracy for six Video-LLMs on the faces test and Table 13 (Appendix E) on every source. Every faces cell lies within or below its confidence interval of chance, as do all but two non-face cells, and accuracy does not rise with edit magnitude (Appendix D.1). On the two-frame task (Fig. 3, right) Qwen2.5-VL-7B reports motion only once the two frames are at least about 8 native frames (270 ms) apart.

5.3 RQ1: A DETECTABILITY THRESHOLD SET BY THE RECEPTIVE FIELD

The probe rises only outside the receptive field. The prediction of §4.1 holds (Fig. 2). On Qwen2.5-VL, where windows of $w = 2, 4, 8$ native frames lie inside the receptive field and 16 and 32 exceed it, the probe scores 50, 50 and 47 inside, each within its confidence interval of chance, and 65 and 73 outside, while the zero-shot answer stays at 49–51 throughout. InternVL3.5-8B and LLaVA-OneVision-7B encode one frame per slot, so their receptive field is a single stride and $w = 8$ should already be visible: their probes sit at 59–67 at $w = 2, 4$, below the 73–82 they reach at $w = 8, 16, 32$, and both answer at chance at every window. On each architecture the largest increase occurs where the receptive field predicts, as a graded step rather than a sharp cut-off; an actor-clustered paired bootstrap puts it at +18 [9, 26] on Qwen2.5-VL, +14 [5, 22] on InternVL3.5 and +9 [−0.5, 17] on LLaVA-OneVision (Appendix C). Changing the stride alone moves the threshold (bottom row): with 32 frames (R=4–6) the same probe reads w=8 at 83 and w=16, 32 at 87, and with 8 frames (R=16–24) only w=32 stays visible, at 62.

Below the threshold, fine-tuning recovers little. Trained on scramble with asymmetry and speed, the frames-only model reaches 59 and 64 at $w = 16, 32$ but stays at 51–55 for $w = 2, 4, 8$ (Fig. 6). The shortest stills are also recovered least: 55 on a 0.25 s FREEZE, which the sampled

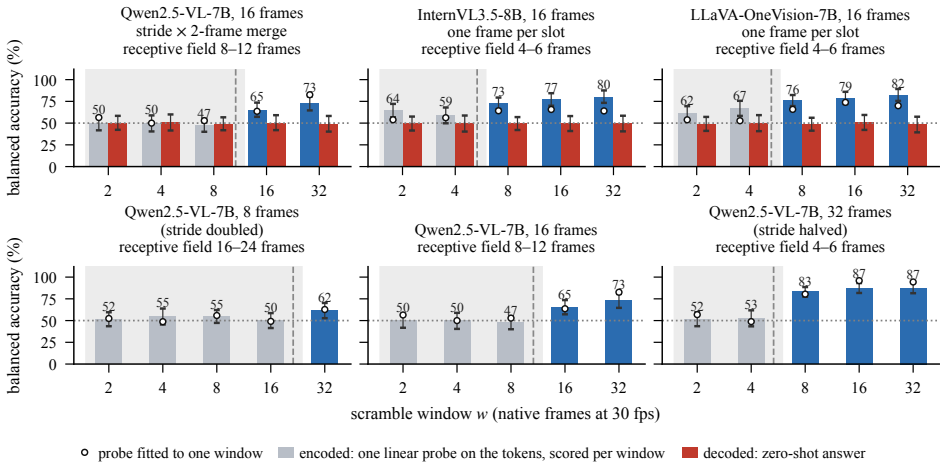

Figure 2: **The detectability threshold** (404 SCRAMBLE items of the faces test: 33–56 positives per window and 195 shared negatives; 95% intervals). Per-window balanced accuracy for one linear probe fitted on all windows (*encoded*) and for the zero-shot answer (*decoded*); dots: a probe fitted to that window alone. Top: three architectures at 16 frames, the receptive field of one token shaded. Bottom: Qwen2.5-VL at 8, 16 and 32 frames, where only the stride changes.

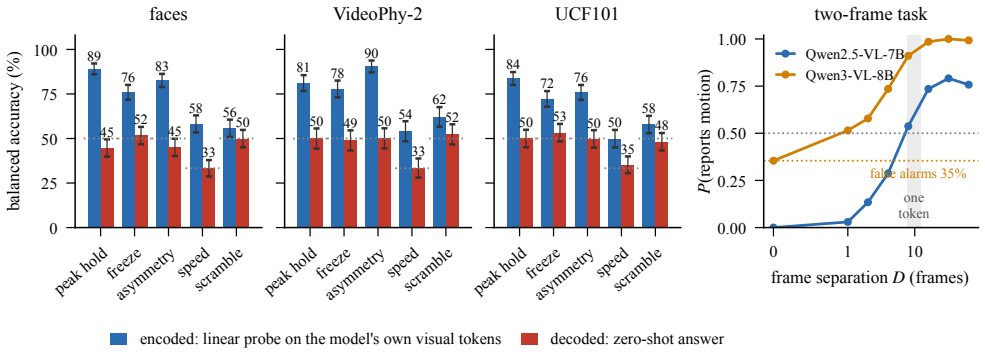

Figure 3: **Encoded versus decoded** (Qwen2.5-VL-7B). Per edit type, balanced accuracy of the linear probe on difference statistics of the model's own visual tokens (*encoded*, out of fold) against the zero-shot answer (*decoded*), on three sources. Right: the two-frame task for the two Qwen models.

pixels already fail to separate (Appendix E.2), and 67 on a 0.25 s PEAK HOLD, against 62–99 at 0.5 s and above. The threshold attenuates an edit without erasing it, and the attenuated edit is the one that supervision recovers least.

5.4 RQ2: ENCODED BUT NOT DECODED

A linear probe reads what the answer does not. Above the threshold, a linear classifier on difference statistics of the tokens the language model receives separates PEAK HOLD, FREEZE and ASYMMETRY on all three sources, at 72–90 (on faces PEAK HOLD 89, FREEZE 76, ASYMMETRY 83), while the language model answers at chance (Fig. 3; Table 19, Appendix E.1). The advantage rests on a comparison across temporal slots and not on differencing as such: the no-differencing control, the same statistics on the tokens themselves pooled over slots, reaches the same level (89, 78 and 76), whereas the single-slot control, which makes no comparison across slots, falls to 65, 64 and 48. Shuffled labels collapse the probe to 51–55, and a strictly linear probe on the mean token difference already recovers 68–74, so much of the signal is linearly present in the token differences.

Table 3: What each input form recovers (faces test, 404 items per edit type, seed 0; balanced accuracy, chance 50, 33 for speed). Three-epoch rows are trained on peak hold + freeze (those two columns) or on asymmetry + speed + scramble (the other three); every row uses the settings of §5.1. Trajectory text is extracted from the full-rate clip and was trained only in the one-epoch regime. Best per column in bold.

Input form	asymmetry	peak hold	speed	freeze	scramble
trajectory text, 1 epoch on all five edit types	79.5	70.6	57.4	57.9	**63.3**
frames only, 3 epochs	83.1	88.2	**64.4**	**76.1**	56.2
TD tokens, 3 epochs	**85.0**	**92.7**	61.4	73.9	57.3

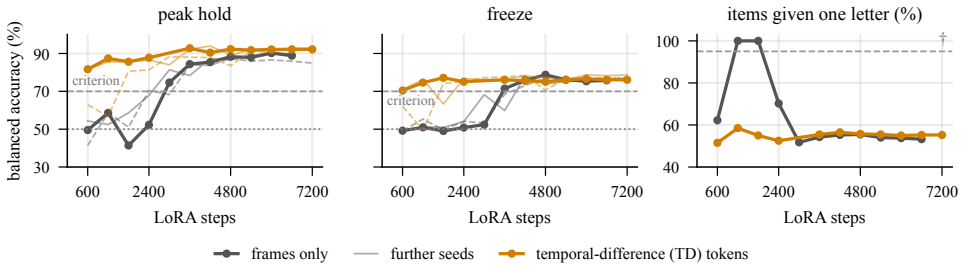

Figure 4: **Steps to criterion by input form** (Qwen2.5-VL-7B, 400-item subset of the faces test, three epochs on PEAK HOLD and FREEZE; thick: seed 0; thin: seed 1; dashed: seed 2). Left and centre: balanced accuracy on the two edit types; right: the share of items answered with one letter.

The static control gives the gap a scale. On actor sex the zero-shot model is right on every item, while the difference statistics reach 39 and the single-slot control 97. The controls thus identify an operation that the zero-shot readout does not use (Table 19), though not why it is unused. A probe on the hidden state at the answer position falls between the two (Fig. 8).

5.5 RQ3: The Cost of Recovery Depends on Where the Comparison Enters

Fine-tuning reaches the accuracy the probe already read. Trajectory text recovers ASYMMETRY (80) and moves PEAK HOLD to 71 within one epoch (Table 3). Three epochs on frames, one adapter for PEAK HOLD+FREEZE and one for the other three types, recover every edit type the probe separates, at PEAK HOLD 88, FREEZE 76 and ASYMMETRY 83 against the probe's 89, 76 and 83, with no † cell, and TD tokens reach the same endpoint (93 / 74 / 85). SPEED reaches 64 against a chance of 33, and SCRAMBLE only 56, two thirds of whose positives lie below the threshold. The fine-tuned model thus arrives at the accuracy the probe had already read from the tokens, and its endpoint has no constant-answer cell.

TD tokens cut the steps to criterion two- to six-fold. The frames-only model stays at chance through 2,400 steps, leaves chance at 3,000 (75 / 52), reaches criterion at 3,600 (84 / 71) and 90 / 75 at 6,000 (Fig. 4). Its early training is a constant-answer phase: one letter on all 400 items at 1,200 and 1,800 steps, and both letters on both edit types only from 3,000 onwards (right panel). The TD-token model has no such phase: it reaches 82 / 70 at 600 steps and plateaus by 1,200 (87 / 75). TD tokens reduce the steps to criterion from 3,600 and 3,000 with frames to 600 on two seeds, factors of six and five, and from 3,600 to 1,800 on a third, a factor of two (Appendix D.2). The saving is in cost alone: on the full test the endpoints agree within 5 points on all three seeds (Table 17), and at the only other stable setting, a lower rate with a higher rank, the gap is unchanged (Appendix D.3).

Placement matters. The fused-TD model, which adds the difference residually *inside* the visual stream like GMoT (Wang et al., 2026) without its gate, reaches criterion at 3,600 steps on both seeds and ends where the other input forms end (Table 18). The model keeps the added term at a tenth of the token scale, and removing it at test time changes at most 10 of the 400 subset answers from 3,000 steps on (Appendix D.5): in this placement the signal is not read, and it costs as many steps as frames.

5.6 Comparison with Alternative Remedies

Alternative remedies do not recover the readout. Table 12 (Appendix E) compares, on the faces test, larger frame budgets, chain-of-thought, a difference-map prompt, a trajectory overlay after Motion-as-Prompt, motion blur after MotionSight (Appendix G), and the tuned models MotionChat and Face-LLaVA. Doubling or quadrupling the frame budget changes no edit type (PEAK HOLD 47 and 46, against 45). The prompts render the motion into the pixels (Fig. 9) yet leave every cell but one (the difference map on SCRAMBLE, 56) within or below its interval of chance, as do the two tuned models. Nor does our fine-tuning carry to natural motion: on FAVOR-Bench (Tu et al., 2026) both frames-only adapters leave the overall score where the base model has it (51.5 against 51.7), and the TD-token adapter, trained at 16 frames and read at 32, falls to 48.1 (Appendix E).

5.7 Generalisation and Replication

The representation transfers; the readout only in part. Trained on faces, both input forms transfer to held-out RAVDESS actors, at 77–79 / 69–70 (Table 15). Transfer to other domains depends on the edit type: on generated video and human actions FREEZE carries over at 68–75, while PEAK HOLD degrades to 52 and 53 with frames and to 60 and 51 with TD tokens. A probe fitted within those sources reads PEAK HOLD at 81 and 84 (Table 15), so their tokens carry the edit but the faces-trained readout does not transfer; both readouts also degrade when fewer frames are sampled at test time (Table 16).

The saving varies with the constant-answer phase of frames-only training. On InternVL3.5-8B frames reach criterion at 4,800 steps on both seeds, TD tokens never on one seed and at 1,800 on the other, and both input forms enter the phase. On Qwen3-VL-8B frames need 1,800–3,600 steps on three seeds and TD tokens 1,200–1,800 (Table 17). The saving therefore varies with how long frames-only training stays in that phase, which the temporal merge alone does not predict: five- to six-fold where the frames-only model stays in it until 1,800 steps and the TD-token model leaves chance at once (Qwen2.5-VL, two seeds), 1.5- to two-fold where the frames-only phase ends by 1,200 (Qwen3-VL, two seeds) or the TD-token model is itself slow to leave chance (the third seed of each Qwen model), inconsistent across seeds where both input forms enter the phase for longer (InternVL3.5), and absent for the fused-TD model of §5.5. Label smoothing, balanced accumulation windows and resuming from the TD-token adapter remove neither the phase nor the cost, and a linear probe at the answer position before any training orders the Qwen conditions by their steps to criterion, but not InternVL3.5 (Appendix D.4).

Held-out edit rules transfer only in part. On three edit rules no adapter was trained on (near-still holds and freezes, and a linear deceleration; Appendix D.6), the probe still reads the edits but the adapters reach only 62–64 with frames and 57–59 with TD tokens on the near-still edits, against 74–93 on exact repeats (Table 3), and chance on the deceleration (Table 11).

6 Conclusion

On WarpPairs, within-event motion fails in Video-LLMs in two regimes. Below a scale set by sampling and encoding the signal is attenuated in the tokens; above it, a linear probe reads the motion while the zero-shot answer does not. Supplying the comparison at the decoder input makes recovery several times cheaper where frames-only fine-tuning stalls in a constant answer that the supplied comparison avoids, but the recovered readout stays narrower than the representation. **Limitations.** The probe is supervised and given the cross-slot comparison, so it shows that the motion can be extracted from the tokens, not that the language model extracts it unaided. Peak hold and freeze test repeated-frame detection rather than motion understanding in general, and the recovered readout stays tied to the trained edit rule and domain. One of the hand-chosen TD statistics, the share of unchanged positions, flags a repeated frame directly. The link between the constant-answer phase and the saving is an association, not a demonstrated cause. The threshold is shown only on the controlled edits (Appendix F), and all fine-tuning is at LoRA scale.

AI USE STATEMENT

In this work we used generative AI tools for research execution and to aid and polish the writing. Under author supervision they were used to implement the benchmark generator, the probes and the training and evaluation scripts, to give feedback on the design of experiments, to help interpret results, and to improve the clarity and grammar of text written by the authors. We have not used generative AI tools to formulate the research questions, to generate any data (every item of WARPPAIRS is produced by a deterministic index map over a public clip, and its label follows from the map), to label data or judge the validity of any result, or to find related work. Translation and qualitative data analysis are not applicable to this work. We have reviewed all AI-assisted work: AI-written code was checked against held-out controls (the constant-letter baselines of Table 20 reproduce the accuracy of a model that collapses to one letter, shuffled labels bound what the probe can fit, and every number in the text was recomputed by the authors from the evaluation outputs it summarises). We take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work involves no new data collection from human subjects. The face sources (CREMA-D, Aff-Wild2 and RAVDESS), the generated videos (VideoPhy-2) and the action videos (UCF101) are publicly released research corpora used under their licences, and the actors in the face corpora consented to research use when the corpora were recorded. WARPPAIRS is released as index maps and clip identifiers over these corpora together with its generator, so no video is redistributed. The only attributes of a person that the paper reads from a clip are actor sex and the acted emotion category, both used as positive controls for the probe on labels the corpora publish; no identity or health inference is made or enabled by the benchmark. The temporal-difference tokens are a diagnostic input form rather than a deployable capability, and we see no path to misuse beyond that of the base models.

REPRODUCIBILITY STATEMENT

WARPPAIRS, its generator and the code for every experiment reported here are released with the paper, together with the seeds and configurations. The benchmark construction is specified in §3 and Appendix A.3, the probe protocol in §4.2, and the fine-tuning settings in §5.1.

REFERENCES

Daechul Ahn, Yura Choi, Hyeonbeom Choi, Seongwon Cho, San Kim, and Jonghyun Choi. What happens when: Learning temporal orders of events in videos. In *WACV*, pp. 2786–2796, 2026.

Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*, 2016.

Piyush Bagad, Makarand Tapaswi, and Cees G. M. Snoek. Test of time: Instilling video-language models with a sense of time. In *CVPR*, pp. 2503–2516, 2023.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025a.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report. *arXiv preprint arXiv:2502.13923*, 2025b.

Hritik Bansal, Clark Peng, Yonatan Bitton, Roman Goldenberg, Aditya Grover, and Kai-Wei Chang. VideoPhy-2: A challenging action-centric physical commonsense evaluation in video generation. In *ICLR*, pp. 118456–118470, 2026.

Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48:207–219, 2022.

Sagie Benaim, Ariel Ephrat, Oran Lang, Inbar Mosseri, William T. Freeman, Michael Rubinstein, Michal Irani, and Tali Dekel. SpeedNet: Learning the speediness in videos. In *CVPR*, pp. 9919–9928, 2020.

Houwei Cao, David G. Cooper, Michael K. Keutmann, Ruben C. Gur, Ani Nenkova, and Ragini Verma. CREMA-D: Crowd-sourced emotional multimodal actors dataset. *IEEE Transactions on Affective Computing*, 5:377–390, 2014.

Yuxin Cao, Wei Song, Jingling Xue, and Jin Song Dong. Caved or convinced: Temporal sampling gates claim deference in video large language models. *arXiv preprint arXiv:2608.03160*, 2026.

Ashutosh Chaubey, Xulang Guan, and Mohammad Soleymani. Face-LLaVA: Facial expression and attribute understanding through instruction tuning. In *WACV*, pp. 2648–2660, 2026.

Junzhe Chen, Siyuan Meng, Yuxi Chen, Man Zhao, Wenyao Gui, and Xiaojie Guo. TOC-Bench: A temporal object consistency benchmark for video large language models. *arXiv preprint arXiv:2605.09904*, 2026.

Cody Coleman, Daniel Kang, Deepak Narayanan, Luigi Nardi, Tian Zhao, Jian Zhang, Peter Bailis, Kunle Olukotun, Chris Ré, and Matei Zaharia. Analysis of DAWNBench, a time-to-accuracy machine learning performance benchmark. *ACM SIGOPS Operating Systems Review*, 53:14–25, 2019.

Hamdi Dibeklioğlu, Albert Ali Salah, and Theo Gevers. Are you really smiling at me? Spontaneous versus posed enjoyment smiles. In *ECCV*, pp. 525–538, 2012.

Yipeng Du, Tiehan Fan, Kepan Nan, Rui Xie, Penghao Zhou, Xiang Li, Jian Yang, Zhenheng Yang, and Ying Tai. MotionSight: Boosting fine-grained motion understanding in multimodal LLMs. In *ICLR*, volume 2026, pp. 1423–1447, 2026.

Paul Ekman and Wallace V. Friesen. Felt, false, and miserable smiles. *Journal of Nonverbal Behavior*, 6(4):238–252, 1982.

Zihang Fu, Haonan Wang, Jian Kang, Kenji Kawaguchi, and Jiaying Wu. Lost in adaptation: Layer-selective recovery of temporal reasoning in video-language models. *arXiv preprint arXiv:2604.11399*, 2026.

Sara Ghazanfari, Siddharth Garg, Prashanth Krishnamurthy, and Farshad Khorrami. SYNCR: A cross-video reasoning benchmark with synthetic grounding. *arXiv preprint arXiv:2605.08412*, 2026.

Rohit Girdhar and Deva Ramanan. CATER: A diagnostic dataset for compositional actions and temporal reasoning. In *ICLR*, 2020.

Peitao Han, Fei Cheng, Lis K. Pereira, Qianying Liu, and Shigeru Kitazawa. Tracing the arrow of time: Diagnosing temporal information flow in Video-LLMs. *arXiv preprint arXiv:2605.07568*, 2026.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *EMNLP-IJCNLP*, pp. 2733–2743, 2019.

Wenyi Hong, Yean Cheng, Zhuoyi Yang, Weihan Wang, Lefan Wang, Xiaotao Gu, Shiyu Huang, Yuxiao Dong, and Jie Tang. MotionBench: Benchmarking and improving fine-grained video motion understanding for vision language models. In *CVPR*, pp. 8450–8460, 2025.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

Jiewen Hu, Leena Mathur, Paul Pu Liang, and Louis-Philippe Morency. OpenFace 3.0: A lightweight multitask system for comprehensive facial behavior analysis. In *IEEE FG*, pp. 1–11, 2025.

Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *CVPR*, pp. 2901–2910, 2017.

Dimitrios Kollias and Stefanos Zafeiriou. Expression, affect, action unit recognition: Aff-Wild2, multi-task learning and ArcFace. *arXiv preprint arXiv:1910.04855*, 2019.

Benno Krojer, Mojtaba Komeili, Candace Ross, Quentin Garrido, Koustuv Sinha, Nicolas Ballas, and Mahmoud Assran. A shortcut-aware video-QA benchmark for physical understanding via minimal video pairs. *arXiv preprint arXiv:2506.09987*, 2025.

Hohin Kwan, Hongyu Li, Ray Zhang, Manyuan Zhang, Xianghao Kong, Anyi Rao, Jiahao Xie, and Si Liu. Video-MME-Logical: A controlled diagnostic benchmark for video temporal-logical reasoning. *arXiv preprint arXiv:2606.27828*, 2026.

Baiqi Li, Kangyi Zhao, Ce Zhang, Chancharik Mitra, Jean de Dieu Nyandwi, and Gedas Bertasius. TimeBlind: A spatio-temporal compositionality benchmark for video LLMs. *arXiv preprint arXiv:2602.00288*, 2026.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. LLaVA-OneVision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

Lei Li, Yuanxin Liu, Linli Yao, Peiyuan Zhang, Chenxin An, Lean Wang, Xu Sun, Lingpeng Kong, and Qi Liu. Temporal reasoning transfer from text to video. In *ICLR*, volume 2025, pp. 74070–74102, 2025a.

Yixuan Li, Changli Tang, Jimin Zhuang, Yudong Yang, Guangzhi Sun, Wei Li, Zejun Ma, and Chao Zhang. Improving LLM video understanding with 16 frames per second. *arXiv preprint arXiv:2503.13956*, 2025b.

Xiaolin Liu, Yilun Zhu, Xiangyu Zhao, Xuehui Wang, Yan Li, Xin Li, Haoyu Cao, Xing Sun, Shaofeng Zhang, Xu Yang, Zhihang Zhong, and Xue Yang. Moment-Video: Diagnosing temporal fidelity of video MLLMs on momentary visual events. *arXiv preprint arXiv:2606.02522*, 2026.

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. TempCompass: Do video LLMs really understand videos? In *Findings of ACL*, pp. 8731–8772, 2024.

Steven R. Livingstone and Frank A. Russo. The Ryerson audio-visual database of emotional speech and song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English. *PLoS ONE*, 13(5):e0196391, 2018.

Thomas Markhorst, Zhi-Yi Lin, Jouh Yeong Chew, Jan van Gemert, and Xucong Zhang. Fine-grained human motion understanding with language models. *arXiv preprint arXiv:2606.20888*, 2026.

Ishan Misra, C. Lawrence Zitnick, and Martial Hebert. Shuffle and learn: Unsupervised learning using temporal order verification. In *ECCV*, pp. 527–544, 2016.

Lyndsey C. Pickup, Zheng Pan, Donglai Wei, YiChang Shih, Changshui Zhang, Andrew Zisserman, Bernhard Schölkopf, and William T. Freeman. Seeing the arrow of time. In *CVPR*, pp. 2035–2042, 2014.

Farrukh Rahman. What does a temporal benchmark score measure? decomposing channel use in video VLM evaluation. *arXiv preprint arXiv:2607.12304*, 2026.

Sayan Deb Sarkar, Rémi Pautrat, Ondrej Miksik, Marc Pollefeys, Iro Armeni, Mahdi Rad, and Mihai Dusmanu. CoPE-VideoLM: Leveraging codec primitives for efficient video language modeling. *arXiv preprint arXiv:2602.13191*, 2026.

Eli Shechtman, Yaron Caspi, and Michal Irani. Space-time super-resolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27:531–545, 2005.

Yumeng Shi, Quanyu Long, Yin Wu, and Wenya Wang. The shape of time: Video-token contrast for temporal understanding in VideoLMs. In *EMNLP*, 2026.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. UCF101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

Bowen Sun, Yujun Cai, Ming-Hsuan Yang, Hang Wu, and Yiwei Wang. PAS: A training-free stabilizer for temporal encoding in video LLMs. In *CVPR*, pp. 14471–14480, 2026a.

Xikai Sun, Kebin Liu, Haotian Wang, Li Liu, Xu Wang, and Yunhao Liu. Motion-as-Prompt: Enhancing motion reasoning in multimodal large language models via motion-guided cross-frame visual prompting. *arXiv preprint arXiv:2608.11655*, 2026b.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *CVPR*, pp. 5228–5238, 2022.

Chongjun Tu, Lin Zhang, Peng Ye, Xianfang Zeng, Wei Cheng, Gang Yu, Tao Chen, et al. FAVOR-Bench: A comprehensive benchmark for fine-grained video motion understanding. *Advances in Neural Information Processing Systems*, 38, 2026.

Ujjwal Upadhyay, Mukul Ranjan, Zhiqiang Shen, and Mohamed Elhoseiny. Time blindness: Why video-language models can't see what humans can? In *CVPR*, pp. 30906–30918, 2026.

Taorui Wang, Wei Xia, Hui Ma, Zijia Song, Jiayu Zhang, Zeheng Wang, Yong Xu, and Zitong Yu. GMoT: Gated motion-aware tokenization for fine-grained micro-gesture video reasoning with multimodal LLMs. In *ACM MM*, 2026.

Weiyun Wang, Zhangwei Gao, Lixin Gu, et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the ACL*, 8:377–392, 2020.

Donglai Wei, Joseph J. Lim, Andrew Zisserman, and William T. Freeman. Learning and using the arrow of time. In *CVPR*, pp. 8052–8060, 2018.

Zihui Xue, Mi Luo, and Kristen Grauman. Seeing the arrow of time in large multimodal models. In *NeurIPS*, 2025.

Yuan Yao, Chang Liu, Dezhao Luo, Yu Zhou, and Qixiang Ye. Video playback rate perception for self-supervised spatio-temporal representation learning. In *CVPR*, pp. 6547–6556, 2020.

Tianyu Yu, Zefan Wang, Chongyi Wang, et al. MiniCPM-V 4.5: Cooking efficient MLLMs via architecture, data, and training recipe. *arXiv preprint arXiv:2509.18154*, 2025.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. VideoLLaMA 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025.

Jianrui Zhang, Mu Cai, and Yong Jae Lee. Vinoground: Scrutinizing LMMs over dense temporal reasoning with short videos. *arXiv preprint arXiv:2410.02763*, 2024.

Chenghua Zhu, Zhaolu Kang, Qifan Shi, Siyan Wu, Kehan Jiang, Lei Wei, Lianyu Hu, Guangyuan Dong, Mingbo Yang, Rui Lu, and Guibo Luo. COMET: Contrastive motion-enhanced temporal reasoning for video multimodal large language models. In *ACM MM*, 2026.

Appendix Overview

Appendix A gives the construction, prompts and statistics of WarpPairs, Appendix B the settings of the probe, the input forms and fine-tuning. Appendices C and D report additional results on the detectability threshold and recovery, Appendix E full result tables and confidence intervals, Appendix F the threshold on a natural motion benchmark, and Appendix G what the train-free prompts render.

A WarpPairs: Construction and Statistics

A.1 Dataset Construction

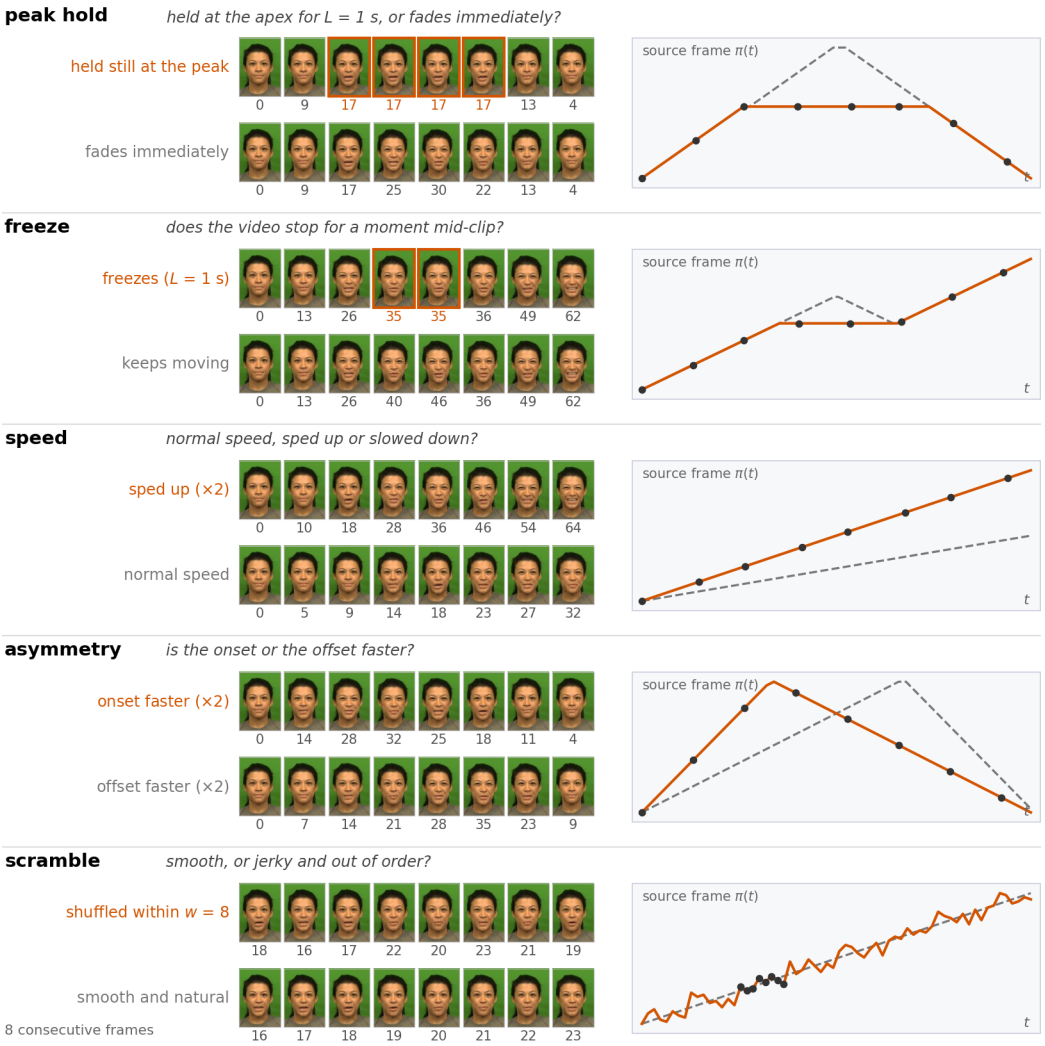

Figure 5: **The five edit types of WarpPairs** on one CREMA-D clip. For each type, eight of the 16 frames the model samples from each answer class, with the source frame under each (repeated source frames outlined), and the index maps $\pi(t)$ of the two classes (orange: the upper class; dashed grey: the lower class; dots: the frames shown). For scramble, eight consecutive output frames are shown instead, since a permutation within w=8 native frames is not visible at the sampling stride; this is the attenuation of §4.1. Edit magnitudes are one level of each type: L=1 s for hold and freeze, a speed ratio of 2, and w=8.

Every item of §3 re-times the first 2–3 s of a source clip, resampled to 30 fps, by an index map, so the two answer classes of an edit type share identity, expression, background and duration. Faces come from CREMA-D and Aff-Wild2, split by actor between train and test, and from RAVDESS, whose actors are all held out; generated video from VideoPhy-2 and human actions from UCF101. A PEAK HOLD item plays the first 0.6 s forward, then a middle segment of L seconds, then the first 0.6 s in reverse; the positive repeats the apex frame throughout the middle segment, the negative fills it with natural motion of the same length from the continuation of the same clip. A FREEZE item inserts the same still or natural segment at the midpoint of the otherwise unedited clip. An ASYMMETRY item plays the first 1.2 s forward at one subsampling rate and in reverse at another,

and its two classes differ in which half is faster. A SCRAMBLE item shuffles the frames of the whole clip within windows of w native frames, with the unedited clip as the negative shared by all windows. A SPEED item subsamples or repeats the frames of the first 1.2 s, with the unedited 1.2 s as the third class. Answer options are shuffled independently for every item, so a constant letter scores the fixed fraction of items whose correct letter it is (Table 20). The faces test holds 404 items per edit type; the RAVDESS, VideoPhy-2 and UCF101 tests hold 1,680, 1,500 and 1,995 items and follow the same construction.

A.2 PROMPTS

Table 4 lists the prompt of each edit type: face sources use wording that refers to an expression, generated and action sources wording that refers to a motion; the answer strings are identical across sources. The model receives the sampled frames, the question and the lettered options, and is asked to answer with the letter.

Table 4: Prompt templates for each edit type. Faces: CREMA-D, Aff-Wild2 and RAVDESS; generic: VideoPhy-2 and UCF101.

Edit type	Faces	Generic	Options
peak hold	In this clip the face makes an expression and then returns to how it started. Is the expression held still at its peak for a visible pause before it fades, or does it start fading immediately?	This clip plays a motion forward, then reverses it. At the turning point, does the video hold still for a moment, or reverse immediately?	Held still at the peak / Fades immediately
freeze	Does the video freeze at some point, so that the face stops moving completely for a moment?	Does the video freeze (stop moving completely) for a moment in the middle of the clip?	Yes, it freezes / No, it keeps moving
speed	Is this video of a person playing at normal speed, sped up, or slowed down?	Is this video playing at normal speed, sped up, or slowed down?	Normal speed / Sped up / Slowed down
asymmetry	In this clip the face makes an expression and then returns to how it started. Is the expression's onset (appearing) faster or slower than its offset (fading away)?	This clip plays a motion forward and then the same motion in reverse. Is the forward part faster or slower than the reverse part?	Onset is faster / Offset is faster
scramble	Does the motion in this video look smooth and natural, or jerky and out of order?	Is the motion in this clip smooth and natural, or jerky and out of order?	Smooth and natural / Jerky and out of order

A.3 DATASET STATISTICS

Table 5: Sources, splits and sizes of WARPPAIRS. A clip is one source segment; every clip yields one item per edit type (one of its answer classes, assigned near-evenly), so items per edit type equal clips. Groups are the units of the train/test split and of the probe's cross-validation folds: actors for the face corpora, source videos for UCF101, one per clip for VideoPhy-2. The 400-item subset used for checkpoint evaluation holds 202 PEAK HOLD and 198 FREEZE items on 288 of the faces test clips (198 CREMA-D, 90 Aff-Wild2).

Source	Content	Split rule	Train clips (groups)	Test clips (groups)	Test items
CREMA-D	acted facial expressions	by actor	794 (75)	286 (16)	1,430
Aff-Wild2	in-the-wild faces	by actor	421 (218)	118 (69)	590
faces test (the two above)			1,215 (293)	404 (85)	2,020
RAVDESS	acted facial expressions	every actor held out	–	336 (6)	1,680
VideoPhy-2	generated video	by clip	–	300 (300)	1,500
UCF101	human actions	by source video	–	399 (277)	1,995

Each edit type's level is the magnitude of its edit: hold or freeze duration in seconds (negatives length-matched to the positive of the same level), onset/offset speed ratio, scramble window in

Table 6: Composition of the faces test: items per edit type and level, split by answer class (e.g. fades / held; offset / onset faster; smooth / jerky). Level = hold or freeze duration (s), onset/offset speed ratio, scramble window (native frames); for speed the columns 0.5, 1 and 2 are the speed factor (slowed, unchanged, sped up); – marks a level that is not defined for that edit type, not a missing experiment. The scramble column 1 holds its 195 unedited negatives, shared by all windows. RAVDESS 1,680, VideoPhy-2 1,500 and UCF101 1,995 items follow the same construction.

Edit type	0.25	0.5	1	1.5	2	3	4	8	16	32	total
asymmetry	–	–	–	52/45	86/80	43/34	34/30	–	–	–	404
peak hold	44/48	53/55	48/54	–	42/60	–	–	–	–	–	404
speed	–	134	116	–	154	–	–	–	–	–	404
freeze	52/49	40/57	46/44	–	63/53	–	–	–	–	–	404
scramble	–	–	195	–	46	–	33	56	39	35	404

native frames, and for SPEED the speed factor, $\times\frac{1}{2}$ or $\times 2$ against the unedited clip; a separate graded set of 532 items adds slow-downs of $\times\frac{1}{3}$ and $\times\frac{1}{4}$ (Appendix D.1).

B Implementation Details

Probe. For every consecutive pair of temporal slots the first difference of the visual tokens is summarised by five statistics over spatial positions: the mean and maximum of $|\Delta_t|$, the mean and maximum of the cosine change between the two slots, and the share of positions whose cosine change is below 0.02. Each statistic is pooled over the clip by its mean, minimum, maximum, standard deviation and log ratio of first-half to second-half mean. The second-difference features repeat this for Δ_t^2 and add the ratio of jerk to speed. A logistic regression is fitted on the standardised features under five-fold grouped cross-validation, grouped by actor on faces (85 groups), by clip on VideoPhy-2 (300) and by source video on UCF101 (277); every reported probe score is a balanced accuracy on held-out folds. The strictly linear control uses only the mean of the consecutive difference vectors.

TD tokens. Seven statistics per consecutive slot pair are computed from the same visual tokens: the log mean, log maximum and log standard deviation of $|\Delta_t|$ over positions, and the mean, maximum, standard deviation and below-threshold share of the cosine change. The sequence is normalised, passed through two one-dimensional convolutions of kernel 5 and width 512 with GELU activations, pooled by attention onto $K{=}16$ learned queries, and projected to the decoder width. The 16 tokens replace 16 placeholder positions beside the visual tokens at the input of the first decoder layer. The projector is trained jointly with the LoRA adapter. On InternVL3.5-8B the same statistics are taken over its `<IMG_CONTEXT>` tokens, one frame per slot.

Fused TD. This input form adds the same seven statistics residually into the video tokens of the later slot, projected and norm-matched with a learnable gain that starts at a tenth of the token scale. No new tokens are added; this is the placement of GMoT's motion token (Wang et al., 2026). The data, rank, rates, budget and schedule are unchanged.

Answer-position probe before training. For each base model and input form, every item of the 400-item subset (202 PEAK HOLD, 198 FREEZE) is passed once through the training input pipeline with the base weights, the adapter at zero and the projector at a random initialisation, the seven statistics standardised by the training-set mean and standard deviation. At each decoder layer a logistic regression ($C{=}0.01$, standardised features) is fitted on the hidden state at the last prompt position under the actor-grouped five-fold split, and the balanced accuracy on held-out folds is reported at its best layer. Along a training run the same probe is applied to each checkpoint with that checkpoint's adapter and projector.

Trajectory text. OpenFace-3.0 action-unit, landmark and gaze curves are extracted from the full-rate clip and rendered as a fixed-format text of at most 600 characters listing, for eight action units, the intervals in which each is active and its peak intensity; whether smile, brow height and mouth opening rise, fall or stay steady over the clip; and the facial motion energy of the two halves with the time of its peak. The text is prepended to the question.

Training details. Table 7 lists the hyperparameters shared by every fine-tuning experiment. A step is one training item and gradients are accumulated over eight items per optimiser update, so a 600-step checkpoint is 75 updates. Learning-rate and rank sweeps change only the named setting, and the two interventions of Appendix D.4 change only the loss at the answer position (label smoothing of 0.1 spread over the option letters) or the item order (answer letters interleaved so every window of eight items is balanced).

Table 7: Training hyperparameters.

Adapter	LoRA, rank 16, α=32, dropout 0.05, on the query, key, value, output, gate, up and down projections of every decoder layer
Optimiser	AdamW, no weight decay, learning rate 10^{-4}, cosine schedule with 50 warm-up updates
Batch	one item per step, gradient accumulation over 8 steps
Epochs	1 over all five edit types (6,075 steps) or 3 over the named types (7,290 steps for peak hold + freeze, 10,935 for the other three)
Input	16 frames spread uniformly over the clip at no more than 8 fps
Checkpoints	every 600 steps, evaluated on the fixed 400-item subset
Decoding	greedy; the first option letter in the output is the answer

C Additional Results on the Detectability Threshold

This section adds the scramble-only fine-tuning and its size-matched control, the per-window scoring behind Fig. 2, the probe at three strides and the bootstrap intervals cited in §5.3.

Scramble-only fine-tuning recovers little below the threshold. Above the threshold the probe finds the scramble encoded (65 and 73 at w=16, 32, Fig. 2) while the zero-shot model reads it at chance, the signature of a readout loss. Fourteen epochs at a constant learning rate on scramble minimal pairs whose positives lie only above the threshold (7,224 steps, close to the 7,290 that recover PEAK HOLD and FREEZE) end at 62.8 (w=16) and 72.6 (w=32) and leave the sub-threshold windows at chance (49.7, 57.2 and 53.7 at w=2, 4, 8). The fine-tuned model does not collapse to a constant answer (56/44 by letter, 55/45 by class).

A size-matched control. Against the frames-only model trained on all 2,430 items this could be read as a further regime, encoded above the threshold yet not recoverable; but the above-threshold set holds 516 items, so steps were matched and data were not. On 516 PEAK HOLD and FREEZE items at the same 7,224 steps the frames-only model reaches 84.1 and 68.9, against 90.9 and 77.4 for the same schedule on all 2,430 items; against that baseline the scramble is no longer a separate case: the w=32 window, which spans four tokens, sits above the size-matched FREEZE at 72.6, while w=16, which spans two, sits below it at 62.8, both inside the intervals the smaller item sets carry ($\pm$6 at n=198, $\pm$10 at $n\approx$81). What survives the control is the threshold with a gradient: sub-threshold windows stay at chance at the budgets we use, and above the threshold the cost rises as the permutation crosses fewer token spans.

How a window is scored, and the same model at three strides. Every per-window number in Fig. 2 and here is a balanced accuracy for that window: the mean of the share of its positives called scrambled and the share of the 195 shared negatives called natural. For the pooled probe, one classifier fitted on all windows, the second term is common to every window, so a window is undetected when its positives are called scrambled at the probe's false-alarm rate. On Qwen2.5-VL at 32 frames that rate is 25%; the w=2, 4 positives are called scrambled on 28% and 30% of items (52 and 53), those of w=8, 16, 32 on 91–100% (83–87).

Each halving of the stride makes the next window visible with no change to the model, the items or the probe. The same 404 items and probe on Qwen2.5-VL with 8, 16 and 32 sampled frames (receptive fields of 16–24, 8–12 and 4–6 native frames) give, per window w = 2, 4, 8, 16, 32, 52, 55, 55, 50, 62 at 8 frames; 50, 50, 47, 65, 73 at 16; and 52, 53, 83, 87, 87 at 32. A classifier fitted to one window alone (its positives against all 195 negatives, the same actor-grouped folds) gives 52, 49, 56, 49, 63 at 8 frames; 56, 50, 53, 64, 83 at 16; and 57, 49, 80, 96, 94 at 32. At 32 frames and w=2 it calls 9 of 46 positives scrambled and 182 of 195 negatives natural. On InternVL3.5-8B and

Table 8: Balanced accuracy by level (edit magnitude) for the zero-shot base model and the fine-tuned frames model, Qwen2.5-VL-7B, faces test. – marks a level that is not defined for that edit type; the speed columns are the magnitude of the speed factor in the graded speed set (×2: slowed to $\frac{1}{2}$ or sped up to 2; ×3 and ×4: slowed to $\frac{1}{3}$ and $\frac{1}{4}$), each scored with the unedited items (Appendix D.1).

Edit type / model	0.25	0.5	1	1.5	2	3	4	8	16	32	×2	×3	×4
asymmetry / zero-shot	–	–	–	40.9	45.8	43.0	51.9	–	–	–	–	–	–
asymmetry / fine-tuned on frames	–	–	–	72.2	82.4	94.7	86.4	–	–	–	–	–	–
peak hold / zero-shot	48.1	46.2	46.4	–	37.6	–	–	–	–	–	–	–	–
peak hold / fine-tuned on frames	66.7	88.0	99.0	–	96.4	–	–	–	–	–	–	–	–
speed / zero-shot	–	–	–	–	–	–	–	–	–	–	33.3	50.0	50.0
speed / fine-tuned on frames	–	–	–	–	–	–	–	–	–	–	54.9	48.5	52.7
freeze / zero-shot	46.7	55.5	54.9	–	51.4	–	–	–	–	–	–	–	–
freeze / fine-tuned on frames	55.1	62.4	94.6	–	95.2	–	–	–	–	–	–	–	–
scramble / zero-shot	–	–	–	–	50.3	–	50.7	49.2	50.5	49.2	–	–	–
scramble / fine-tuned on frames	–	–	–	–	51.2	–	55.4	54.6	58.5	63.7	–	–	–

LLaVA-OneVision-7B it gives 54, 56, 64, 66, 64 and 54, 53, 66, 74, 70. Below the threshold this per-window probe stays within its confidence interval of chance at every stride; above it, the smaller training set costs accuracy relative to the pooled fit.

Four of the five bootstrap intervals for the step across the predicted boundary exclude zero; the LLaVA-OneVision one includes it (lower bound −0.5). A paired bootstrap over the 85 actors (1,000 resamples, the pooled probe refitted on each) gives +18 [9, 26] between w=8 and 16 on Qwen2.5-VL at 16 frames, +30 [18, 39] between w=4 and 8 at 32 frames, +11 [0.5, 21] between w=16 and 32 at 8 frames, +14 [5, 22] on InternVL3.5-8B and +9 [−0.5, 17] on LLaVA-OneVision-7B, the last two between w=4 and 8.

D Additional Results on Recovery

D.1 Accuracy by Edit Magnitude

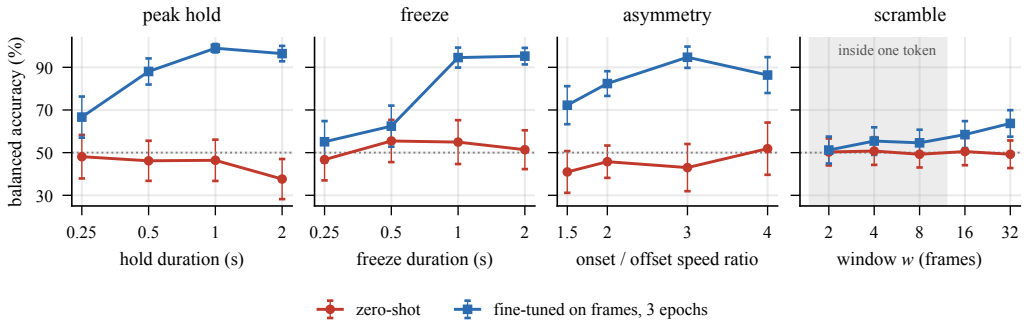

Figure 6: **Accuracy by edit magnitude before and after fine-tuning** (faces test; zero-shot Qwen2.5-VL-7B versus the frames-only model after three epochs; 95% intervals). Shaded: scramble windows inside the receptive field.

A slope of accuracy over edit magnitude appears only after fine-tuning. On the graded SPEED set of Appendix A.3 (532 items, three-way, chance 33.3) the base model answers "normal speed" on every item, which is 33.3 balanced and a recall of 0 on every off-normal level. The recovered frames readout reaches 60 balanced, and its recall of a slow-down is monotone in the magnitude of the warp (64, 75 and 83 for $r = \times\frac{1}{2}, \frac{1}{3}, \frac{1}{4}$, with a ×2 speed-up at 79), while it recognises unchanged speed on only 22% of items; the per-level balanced cells of Table 8, each of which merges the unchanged items, therefore sit near 50. The zero-shot base model shows no slope on any edit type; the recovered frames readout rises with magnitude on PEAK HOLD, FREEZE and ASYMMETRY, and on SCRAMBLE

stays within noise of chance inside the receptive field (51, 55, 55 at w=2, 4, 8) and rises only outside it (59, 64), consistent with the detectability threshold of §5.3.

D.2 Steps to Criterion Across Criteria and Seeds

Steps to first reach a criterion on the fixed 400-item subset (cosine schedule, three seeds; §5.1):

	frames			TD tokens			factor		
criterion	seed 0	seed 1	seed 2	seed 0	seed 1	seed 2	seed 0	seed 1	seed 2
PEAK HOLD $\geq$70	3,000	3,000	2,400	600	600	1,800	5	5	1.3
mean of the two edit types $\geq$70	3,600	3,000	3,600	600	600	1,800	6	5	2
mean $\geq$70, held for two checkpoints	3,600	4,200	3,600	600	600	1,800	6	7	2
PEAK HOLD $\geq$80 with FREEZE $\geq$73	4,200	4,200	4,200	1,200	1,200	1,800	3.5	3.5	2.3

On the first two seeds the first three rows put the factor between five and seven. The fourth gives 3.5 on both because it requires FREEZE to clear 73, and the TD-token model reads 70.4 and 70.9 there at its first checkpoint, so two to three points on the noisier edit type cost it one checkpoint, 600 against 1,200; the frames-only model needs 4,200 steps to satisfy the same pair and is unaffected. On the third seed the TD-token model reaches every criterion at 1,800 steps and the frames-only model at 2,400 to 4,200, factors of 1.3 to 2.3. The multiple therefore ranges from 1.3 to seven across the four criteria and three seeds, and the main text quotes the mean criterion, two to six.

The shape of the two curves, rather than any single crossing, is what holds across seeds (Fig. 7, first column). Before it leaves chance, the frames-only model answers one letter to every item of both edit types for one or two checkpoints on every seed (1,200 and 1,800 steps on seed 0, 1,800 on seed 1, 600 and 1,200 on seed 2). The TD-token model stays below 95% one-letter answers at every checkpoint of every seed; its largest share is 85% of FREEZE items, at 1,800 steps on seed 1 and 600 on seed 2. The frames-only model has to leave a constant-answer solution and the TD-token model never enters one, and Appendix D.4 finds the same association on two further base models.

D.3 Learning Rate and Rank

Table 9: Learning-rate and rank sweeps on Qwen2.5-VL-7B (1.5 epochs on peak hold + freeze at a constant rate; last row: the three-epoch cosine runs of the main text, three seeds). Steps to the mean-70 criterion on the 400-item subset and the endpoint on the full test (peak hold / freeze). Unstable: the run falls to a constant or non-letter answer during training. –: not run.

		frames only		TD tokens	
learning rate	rank	criterion	endpoint	criterion	endpoint
10^{-3}	16	unstable	–	–	–
3×10^{-4}	16	unstable	–	unstable	–
10^{-4}	64	unstable	–	unstable	80 / 69
3×10^{-5}	64	3,600	88 / 77	600	92 / 75
10^{-4}, cosine (main text)	16	3,000–3,600	87–89 / 76	600–1,800	91–93 / 74–76

Learning rate. The multiple is not an artefact of the rate. A sweep on the frames-only model (1.5 epochs, same subset, constant rate) finds 10^{-3} and 3×10^{-4} unusable, since both end in non-letter output (PEAK HOLD 0), the latter after sitting at 59 / 54 through 2,400 steps. Of the constant rates swept only 3×10^{-5} is stable, and the TD-token model at 3×10^{-4} collapses the same way; the main text uses 10^{-4} with cosine annealing, between the two. At the stable lower rate the ordering is unchanged: criterion is reached at 3,600 steps with frames and 600 with TD tokens (rank 64, below), the same 3,000–3,600 against 600 that the annealed cosine schedule at 10^{-4} gives on two of three seeds.

Rank. Nor does the multiple depend on the rank. At a constant rate, rank 64 (α=128) at 10^{-4} is not stable for either input form; both collapse to a constant or a non-letter answer, and the TD-token adapter ends at 80 / 69 on the full test, below the rank-16 endpoints of 91–93 / 74–76. At 3×10^{-5} rank 64 is stable and the ordering matches rank 16. Frames answer one letter through 1,800 steps,

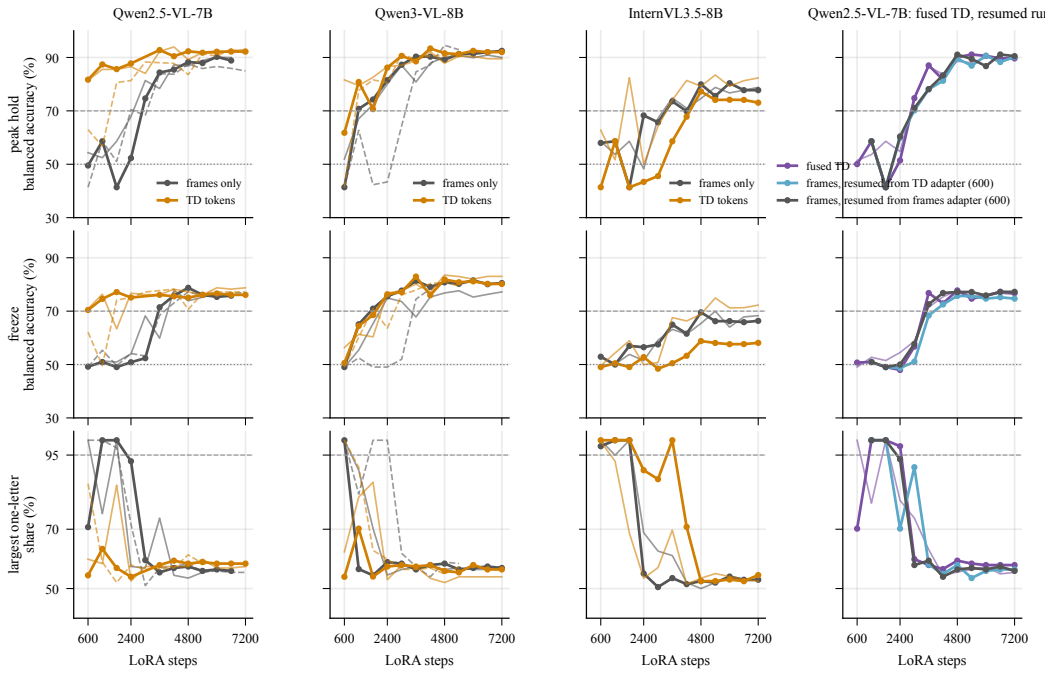

Figure 7: Every fine-tuning run of the recovery experiments on the 400-item subset (three epochs on PEAK HOLD and FREEZE, checkpoints every 600 steps). Rows: balanced accuracy on the two edit types and the largest share of one edit type's items answered with the same letter (dashed line: the 95% constant-answer mark). Columns: the three base models with frames only and TD tokens (thick: seed 0; thin: seed 1; dashed: seed 2), and on Qwen2.5-VL the fused-TD model (two seeds) and frames-only training resumed from the 600-step TD-token and frames-only adapters.

reach 73 / 57 at 3,000 and criterion at 3,600 (85 / 76) and end at 88 / 77 on the full test, the same 3,000–3,600 steps to criterion as rank 16 at 10^{-4}; TD tokens reach criterion at 600 steps (84 / 75, both letters from the first checkpoint) and end at 92 / 75. The cost multiple at the best stable rate is six.

D.4 THREE BASE MODELS

Qwen3-VL-8B, like Qwen2.5-VL, merges two sampled frames into one temporal slot. Under the same data, rank, rate, frame budget and schedule, its frames-only model reaches the mean-70 criterion at 1,800, 2,400 and 3,600 steps on three seeds and its TD-token model at 1,200, 1,200 and 1,800 (Table 17; Fig. 7, second column), and on every seed the two input forms end within 5 points of each other on the full test (87.9–92.2 / 77.1–82.3 with frames, 90.2–91.5 / 79.5–81.2 with TD tokens). The frames-only model answers one letter to every item of at least one edit type at 600 steps on every seed, and on the third seed to every item at 1,800 and 2,400 as well; the TD-token model does so only on the third seed, at 600 steps.

Qwen2.5-VL and Qwen3-VL merge the same two frames per token, yet the TD tokens are worth a factor of two to six on the first and 1.5 to two on the second; what differs between the two models is how late the frames-only model leaves the constant-answer phase. The Qwen2.5-VL frames-only model answers one letter to every item of both edit types for one or two checkpoints on every seed and needs 3,000–3,600 to reach the criterion, whereas the Qwen3-VL frames-only model leaves the phase by 1,200 on two of three seeds and reaches it at 1,800 or 2,400. On InternVL3.5-8B, the third case, both input forms enter the phase and the order of the two is not stable across seeds (below). Across the three base models and the fused-TD model of §5.5, conditions that leave the phase later tend to need more steps, though not seed by seed, and the temporal merge alone does not account for the difference.

The phase is foreshadowed before training. A measure taken at step 0 orders the conditions. A logistic regression is fitted per decoder layer on the hidden state at the answer position, on the 400-item subset under the actor-grouped split, with the base weights, the adapter at zero and the projector at a random initialisation (Appendix B); Table 10 gives its best layer. With frames only, the answer position separates PEAK HOLD at 71 and FREEZE at 65 on Qwen2.5-VL and at 76 and 71 on Qwen3-VL; with TD tokens it separates them at 87–91 and 77–80 on both models under three initialisations of the projector; with fused TD at 70 and 66, the frames-only values. The five conditions order by this measure as they order by their median steps to criterion: the two TD-token conditions (600 and 1,200 steps), then Qwen3-VL frames (2,400), then Qwen2.5-VL frames and fused TD (3,600). The measure separates input forms and base models, not seeds: the three initialisations read 89–91 / 77–78 on Qwen2.5-VL and 87–89 / 77–80 on Qwen3-VL, while the third seed of each reaches criterion at 1,800 steps rather than 600 or 1,200. Nor does it carry over to InternVL3.5-8B, where frames read 82 / 71 and TD tokens 98 / 83 before training, the highest values of any condition, yet frames need 4,800 steps and TD tokens 1,800 or never. Linear availability at the answer position before training thus accompanies the fast route on the two Qwen models and is not sufficient for it on InternVL3.5. Along the Qwen2.5-VL frames-only run the same probe reads 73 / 66–68 at 600, 1,200, 1,800 and 2,400 steps, the checkpoints of the constant-answer phase, 80 / 64 at 3,000, where the model leaves chance, and 87 / 79 at 3,600, where it reaches criterion; the TD-token model reads 90 / 78 at its 600-step checkpoint, the values it had before training. On the Qwen models the comparison that the TD tokens supply is linearly available at the answer position before any training; with frames it becomes available only as the phase ends, and both reach criterion at the same level.

Table 10: Linear probe at the answer position before training (best layer, balanced accuracy on the 400-item subset, actor-grouped folds) against steps to the mean-70 criterion on three seeds (two for fused TD and InternVL3.5). Qwen TD-token rows: three random initialisations of the projector.

base model	input form	peak hold	freeze	steps to criterion
Qwen2.5-VL-7B	frames only	71	65	3,600 / 3,000 / 3,600
Qwen2.5-VL-7B	fused TD	70	66	3,600 / 3,600
Qwen2.5-VL-7B	TD tokens	90, 89, 91	77, 77, 78	600 / 600 / 1,800
Qwen3-VL-8B	frames only	76	71	1,800 / 2,400 / 3,600
Qwen3-VL-8B	TD tokens	88, 87, 89	80, 77, 77	1,200 / 1,200 / 1,800
InternVL3.5-8B	frames only	82	71	4,800 / 4,800
InternVL3.5-8B	TD tokens	98	83	not reached / 1,800

These comparisons rest on two or three seeds per base model, and three models is a small sample. They establish the negative claim that the temporal merge alone does not predict the saving; what they support is weaker and correlational, that the saving is associated with avoiding a prolonged constant-answer phase, and that a probe taken before training orders the conditions the same way. The phase and the slow learning occur together in every case here, so these results neither identify one as the cause of the other nor rule out a contribution from the merge.

The phase is not a product of the output distribution or of the batch composition. Two interventions on the frames-only model test whether it can be removed without any temporal input: label smoothing of 0.1 over the option letters at the answer position, and an item order in which every accumulation window holds equal numbers of each answer letter, with everything else unchanged (Appendix B). Both models still answer one letter on 98–100% of the items of each edit type, at 1,200 and 1,800 steps with label smoothing and at 1,200, 1,800 and 2,400 steps with balanced windows. They reach the criterion at 3,600 and 4,200 steps and end at 87.7 / 74.8 and 86.5 / 74.9 on the full test, against 3,600 steps and 88.2 / 76.1 without the intervention. Among the frames-only models the phase is avoided by none, and among the input forms only by the one that supplies the comparison; this strengthens the association without identifying the comparison as the cause of the faster learning.

The saving belongs to the input, not to the state the optimiser reaches. A third test resumes frames-only training (no TD channel) from the TD-token adapter at 600 steps, where that model had already reached criterion, and, as a control, from the frames-only adapter at 600 steps, with the same data order and schedule from that point. Both runs answer one letter to every item at 1,200 and 1,800 steps, reach the criterion at 3,600 steps (78 / 68 and 78 / 73 on the subset) and end at 89.1 / 74.8

Table 11: Held-out edit rules on the same 404 source clips (balanced accuracy; chance 50). No Video-LLM adapter was trained on any of them; fine-tuned rows are seed 0. Pixels and probe are read as in Table 22, out of fold under the actor-grouped split used throughout; the fine-tuned rows are the three-epoch peak hold + freeze models of Table 3.

stage	near-still peak hold	near-still freeze	deceleration
sampled pixels	74	60	63
visual-token statistics (probe)	76	67	66
zero-shot answer	48	53	50
fine-tuned on frames	64	62	52
fine-tuned with TD tokens	57	59	50

and 89.3 / 76.3 on the full test, the same course as training from scratch. The adapter that reads the comparison from TD tokens therefore does not read it from frames once the tokens are removed.

On Qwen2.5-VL a second seed reproduces both input forms at the endpoint (88.5 / 75.6 against 88.2 / 76.1 with frames and 92.6 / 75.8 against 92.7 / 73.9 with TD tokens, Table 18), and FREEZE is the noisier of the two edit types in every seed, moving by several points between neighbouring checkpoints. A third seed reaches the criterion at 3,600 steps with frames and at 1,800 with TD tokens (Appendix D.2); they end at 87.1 / 76.3 and 91.0 / 74.9.

InternVL3.5-8B. The constant-answer phase is present without a temporal merge, and here TD tokens are neither a reliable shortcut nor a reliable cost (Fig. 7, third column). TD tokens on this model use the same seven per-slot-pair statistics of its own <IMG_CONTEXT> tokens, one frame per slot, with sixteen placeholder tokens replaced at the input of the first decoder layer, the projector at 10^{-4}, and the same LoRA, data, budget and schedule. Frames reach the mean criterion at 4,800 steps on both seeds, after answering one letter to 83–100% of one edit type's items at 600 steps on both and through 1,800 on the first, and end at 78 / 68 and 78 / 69 on the full test. TD tokens answer one letter to 71–100% of one edit type's items at every checkpoint through 4,200 steps on the first seed, never reach the criterion and end at 73 / 57, the reverse of the order on Qwen2.5-VL; on the second seed they cross it at 1,800 steps, fall back to chance at 2,400, hold it from 3,600 and end at 81 / 72. What is stable across seeds is that both input forms enter the phase and that frames need 4,800 steps.

D.5 Fusing the Difference into the Visual Tokens

Placed inside the visual stream, the difference costs as many steps as frames (Fig. 7, fourth column). The fused-TD model of Appendix B answers one letter to every item of at least one edit type for two checkpoints on each seed, reaches the mean-70 criterion at 3,600 steps on both (87 / 77 and 87 / 72 on the subset), against 600–1,800 for the decoder-input form and 3,000–3,600 for frames, and ends at 89.4 / 75.6 and 90.1 / 74.6 on the 808-item test (Table 18). The fused term is carried at a small weight and the adapter learns around it. The gain that scales the statistics before they are added to the video tokens starts at 0.10 and ends at 0.11 on both seeds, and evaluating a checkpoint with the fused term removed changes the answer on at most 10 of the 400 subset items at every checkpoint of the first seed, on at most 8 from 3,000 steps onwards on the second, and on 15 and 8 of the 808 test items. Inside the visual stream the same signal is neither amplified nor read; the model reaches the frames-only endpoint by the frames-only route.

D.6 Held-Out Edit Rules

Transfer to held-out rules is partial and confined to the longer stills. The three held-out rules are built from the same 404 clips: a held peak and a mid-clip pause whose still segment advances one native frame every four instead of repeating one, so that no two consecutively sampled frames are identical, and a linear deceleration against a steady pace of the same mean speed. With frames the PEAK HOLD+FREEZE models score 46 and 60 at 0.25 and 0.5 s against 67 and 82 at 1 and 2 s. Neither transfers to the deceleration, which changes the edit and the question together, although the tokens carry it. The adapters trained on the other three edit types read 47 / 51 / 54 (frames) and 50 /

Table 12: Alternative remedies on the faces test (balanced accuracy). Train-free motion prompts, larger frame budgets, a motion-instruction-tuned model, a face model, and our fine-tuning on the minimal pairs (three epochs, seed 0; shaded). 404 items per edit type; chance 50, 33 for speed. Best value per column in bold. ‡: Face-LLaVA answers in prose, so it is scored by length-normalised option likelihood (more favourable to it than generation) with its own frame sampling.

Model or remedy	asymmetry	peak hold	speed	freeze	scramble
Qwen2.5-VL-7B, 16 frames, no training	45.0†	44.7	33.3	51.6	49.9
+ chain-of-thought prompt	46.3	52.4	32.5	47.9	47.9
+ 32 frames (at most 16 fps)	44.7†	47.2	33.3	50.6	50.0
+ 64 frames (at most 30 fps)	44.5†	45.8	33.3	52.0	49.2
+ difference-map prompt	45.0†	51.2	32.0	50.4	56.1
+ point-trajectory overlay (after Motion-as-Prompt)	44.8†	50.5	33.8	51.7	52.6
+ motion blur (after MotionSight)	45.0†	50.9	33.2	52.3	49.7
MotionChat (tuned on MotionVid-QA motion instructions)	49.5	50.2	33.3	50.5	50.1
Face-LLaVA (face model with landmark-guided features)‡	48.1	49.3	33.3	50.9	46.7
our fine-tuning: frames only, 3 epochs	83.1	88.2	**64.4**	**76.1**	56.2
our fine-tuning: TD tokens, 3 epochs	**85.0**	**92.7**	61.4	73.9	**57.3**

Table 13: Zero-shot balanced accuracy on every source (16 frames, at most 8 fps). Chance: 50 (33 for speed).

Source / model	asymmetry	peak hold	speed	freeze	scramble
faces/Qwen2.5-VL-7B	45.0†	44.7	33.3	51.6	49.9
faces/Qwen3-VL-8B	44.6†	50.0	33.3	50.2	50.1
RAVDESS/Qwen2.5-VL-7B	52.1†	51.5	33.3	52.6	50.0
RAVDESS/Qwen3-VL-8B	52.4†	50.0	33.3	50.0	51.1
VideoPhy-2/Qwen2.5-VL-7B	50.1	50.0	33.5	48.9	52.3
VideoPhy-2/Qwen3-VL-8B	50.4†	50.1	32.1	50.0	50.5
UCF101/Qwen2.5-VL-7B	49.8	50.0	35.3	53.3	48.2
UCF101/Qwen3-VL-8B	51.9†	52.0	34.1	50.0	49.2
faces/InternVL3.5-8B	50.2	50.0	33.3	51.3	49.5
RAVDESS/InternVL3.5-8B	50.0	50.0	33.3	54.0	50.0
VideoPhy-2/InternVL3.5-8B	50.4	51.2	33.3	53.3	47.1
UCF101/InternVL3.5-8B	49.6	58.8	33.3	54.5	47.4
faces/LLaVA-OneVision-7B	49.7	52.8	36.1	52.1	49.3
faces/VideoLLaMA3-7B	51.8	51.6	33.0	48.2	50.0
faces/MiniCPM-V-4.5	50.3	50.0	33.3	48.4	49.8
faces/MiniCPM-V-4.5, 3-frame packing	48.8	50.0	33.3	48.4	49.0
VideoPhy-2/LLaVA-OneVision-7B	50.0	50.7	33.7	51.9	50.7
VideoPhy-2/VideoLLaMA3-7B	49.7	53.8	30.6	53.6	50.5
VideoPhy-2/MiniCPM-V-4.5	49.5	50.0	31.7	53.2	49.7
UCF101/LLaVA-OneVision-7B	50.0	49.7	38.9	49.8	49.0
UCF101/VideoLLaMA3-7B	50.0	51.5	36.7	53.6	49.5
UCF101/MiniCPM-V-4.5	50.4	50.7	33.1	54.8	49.5

49 / 46 (TD tokens) on the three held-out rules of §5.7, so the partial transfer in Table 11 is specific to the PEAK HOLD+FREEZE adapters.

E Full Results and Confidence Intervals

Table 12 gives the alternative remedies of §5.6 and Table 15 the transfer results of §5.7. Table 14 is the external anchor: FAVOR-Bench is a public five-way motion benchmark, and none of the three fine-tuned adapters transfers to it; the TD-token adapter, trained at 16 frames and read here at 32, loses 3.6 points overall. Table 16 varies the number of frames at test time; both readouts degrade with the budget and are at chance with one frame, the TD-token one by construction, since there is no difference to take. Table 17 gives the steps to criterion and endpoints of every seed on the three base models, and Table 18 the endpoints of every seed of the three input forms on Qwen2.5-VL.

Table 14: FAVOR-Bench close-ended accuracy (stratified 1,064-question subset, 5-way, chance 20; 32 frames at most 4 fps) for the base model, the base model after our fine-tuning (three epochs, seed 0, trained at 16 frames; shaded) on two groups of edit types with frames only and on peak hold + freeze with TD tokens, and MotionChat. AS action sequence, CM camera motion, HAC holistic action, MAD multi-agent dynamics, NSM non-subject motion, SAD single-action detail. Best per column in bold.

Model	AS	CM	HAC	MAD	NSM	SAD	all
Qwen2.5-VL-7B, no training	59.5	38.0	**53.0**	52.5	54.7	**54.5**	**51.7**
+ our fine-tuning on peak hold + freeze (frames only)	**61.0**	40.0	51.0	50.0	**62.5**	52.0	51.5
+ our fine-tuning on asymmetry + speed + scramble (frames only)	57.0	**40.5**	52.0	53.5	56.2	53.0	51.5
+ our fine-tuning on peak hold + freeze (TD tokens)	55.0	35.0	50.5	52.0	42.2	50.0	48.1
MotionChat	52.5	37.0	**53.0**	**55.5**	40.6	45.5	48.2

Table 15: Transfer of the fine-tuned model, trained on faces and tested on held-out actors (RAVDESS) and on two non-face sources (peak hold / freeze; three epochs, cosine schedule; balanced accuracy, chance 50). Items per edit type: 404 faces, 336 RAVDESS, 300 VideoPhy-2, 399 UCF101. Seeds as labelled; TD tokens were run with one seed on these sources. The last row gives the linear probe on the base model's tokens, fitted out of fold within each source, as a reference for what those tokens carry.

	faces		RAVDESS		VideoPhy-2		UCF101	
Input form	peak hold	freeze	peak hold	freeze	peak hold	freeze	peak hold	freeze
frames only, seed 0	88.2	76.1	79.3	69.1	51.8	69.9	52.7	68.4
frames only, seed 1	88.5	75.6	76.7	70.1	51.4	68.5	50.7	68.6
TD tokens, seed 0	92.7	73.9	76.6	69.9	59.6	75.3	51.2	72.6
linear probe on the base model's tokens	89.1	75.9	–	–	81.1	77.8	83.7	72.2

Table 16: Frame-budget dependence of the fine-tuned model (seed 0; full test, 808 items; chance 50): 16 / 4 / 1 frames at test time (peak hold / freeze).

Input form, training data	16 frames		4 frames		1 frame	
	peak hold	freeze	peak hold	freeze	peak hold	freeze
frames only, 3 epochs on peak hold + freeze	88.2	76.1	50.4	49.5	50.0	45.7
TD tokens, 3 epochs on peak hold + freeze	92.7	73.9	42.9	47.8	50.0	50.0

Table 17: Steps to criterion on three base models under identical data, LoRA rank, learning rate, frame budget and schedule. Criterion: the first 600-step checkpoint whose mean peak-hold and freeze balanced accuracy reaches 70 on the fixed 400-item subset. Largest same-letter share: over all checkpoints, the largest share of one edit type's items given the same letter (50 balanced, 100 a constant answer). Final accuracy: peak hold / freeze on the full faces test.

Base model	Input form	Frames per token	Largest same-letter share	Steps to criterion	Final accuracy
Qwen2.5-VL-7B	frames only	2	100	3,600	88.2 / 76.1
Qwen2.5-VL-7B	TD tokens	2	63	600	92.7 / 73.9
Qwen2.5-VL-7B, seed 1	frames only	2	100	3,000	88.5 / 75.6
Qwen2.5-VL-7B, seed 1	TD tokens	2	85	600	92.6 / 75.8
Qwen2.5-VL-7B, seed 2	frames only	2	100	3,600	87.1 / 76.3
Qwen2.5-VL-7B, seed 2	TD tokens	2	85	1,800	91.0 / 74.9
Qwen3-VL-8B	frames only	2	100	1,800	92.2 / 80.5
Qwen3-VL-8B	TD tokens	2	70	1,200	91.5 / 80.5
Qwen3-VL-8B, seed 1	frames only	2	100	2,400	87.9 / 77.1
Qwen3-VL-8B, seed 1	TD tokens	2	86	1,200	90.5 / 81.2
Qwen3-VL-8B, seed 2	frames only	2	100	3,600	91.2 / 82.3
Qwen3-VL-8B, seed 2	TD tokens	2	100	1,800	90.2 / 79.5
InternVL3.5-8B	frames only	1	100	4,800	78.0 / 67.5
InternVL3.5-8B	TD tokens	1	100	not reached	72.5 / 57.3
InternVL3.5-8B, seed 1	frames only	1	100	4,800	77.8 / 68.7
InternVL3.5-8B, seed 1	TD tokens	1	99	1,800	81.1 / 71.9

Table 18: Every seed of the three input forms on Qwen2.5-VL-7B (3 epochs on peak hold + freeze; full test, 808 items; balanced accuracy, chance 50).

Input form, training data	peak hold	freeze
frames only, seed 0	88.2	76.1
frames only, seed 1	88.5	75.6
frames only, seed 2	87.1	76.3
TD tokens, seed 0	92.7	73.9
TD tokens, seed 1	92.6	75.8
TD tokens, seed 2	91.0	74.9
fused TD, seed 0	89.4	75.6
fused TD, seed 1	90.1	74.6

Table 19: Encoded versus decoded, with the four controls (Qwen2.5-VL-7B, faces test, balanced accuracy; chance 50, 33 for speed, 17 for emotion, 50 for sex). Every probe cell is an out-of-fold prediction under the same actor-grouped five-fold split, over the full test set: 404 items per edit type, 286 per static task. Linear Δ is the strictly linear control: the per-slot mean of $v_{t+1} - v_t$, no magnitudes and no maxima. No differencing uses per-slot token magnitudes and their variation over the clip; single slot uses one temporal slot only; shuffled bounds what the probe fits on its own. The two static tasks ask about a property that a single frame carries. Their labels are constant within each actor, so under the actor-grouped split a probe whose features lack the attribute can fall below chance (39.1 for sex).

Edit type	linear probe on the visual tokens						zero-shot
	Δ statistics	$\Delta{+}\Delta^2$ statistics	linear Δ	no differencing	single slot	shuffled labels	answer
peak hold	90.3	89.1	74.0	88.5	65.4	52.3	44.7
freeze	76.3	75.9	67.6	77.9	64.4	51.2	51.6
asymmetry	69.2	82.6	72.2	75.8	48.0	54.5	45.0
speed	54.8	58.2	52.4	53.4	48.9	36.2	33.3
scramble	51.9	55.8	55.4	54.5	52.6	52.6	49.9
static emotion (6-way)	–	25.3	40.0	23.6	35.3	17.8	29.9
static actor sex (2-way)	–	39.1	68.4	47.4	96.8	50.0	100.0

Table 20: Constant-letter baselines (balanced accuracy, faces test). Options are shuffled per item, so a model that always answers one letter scores the fixed fraction of items whose correct letter it is. A fine-tuned adapter that collapses to one letter reproduces this number exactly.

Edit type	always A	always B	always C
asymmetry	55.0	45.0	0.0
peak hold	45.5	54.5	0.0
speed	26.4	37.2	36.5
freeze	48.3	51.7	0.0
scramble	50.4	49.6	0.0

E.1 Confidence Intervals and Answer Distributions

Actor-clustered and binomial intervals differ by less than one point. Balanced accuracy is the mean of per-class recalls; its 95% half-width is $1.96\sqrt{\sum_c p_c(1-p_c)/n_c}/k$ for k classes: $\pm$4–5 points at n=404 per edit type on the faces test, $\pm$3–4 on the 808-item PEAK HOLD+FREEZE test, $\pm$6–7 per edit type on the 400-item subset. Items from one actor are not independent, so we also bootstrap by actor (85 actors in the 404-item faces test, 1,000 resamples). The actor-clustered half-widths are $\pm$3.9 / 4.1 for the zero-shot PEAK HOLD / FREEZE cells, $\pm$3.4 / 3.7 for the fine-tuned frames endpoint and $\pm$2.7 / 3.3 for the TD-token endpoint, against binomial $\pm$4.8 / 4.6, 3.1 / 3.8 and 2.5 / 3.5 on the same items.

Table 21: Answer-letter distribution (%) for the settings and edit types where it is not spread: one letter at ≥65%, non-letter output at ≥5%, or the three-way SPEED type. Every other setting and edit type answers each letter between 35 and 65% of the time (SPEED: no letter above 65%).

Model or remedy	edit type	A	B	C	D	E	F	unparsed	n
zero-shot Qwen2.5-VL-7B	asymmetry	–	100	–	–	–	–	–	404
zero-shot Qwen2.5-VL-7B	speed	30	35	35	–	–	–	–	404
zero-shot Qwen3-VL-8B	asymmetry	1	99	–	–	–	–	–	404
zero-shot Qwen3-VL-8B	speed	30	35	35	–	–	–	–	404
chain-of-thought prompt	speed	29	36	35	–	–	–	0	404
difference-map prompt	asymmetry	–	100	–	–	–	–	–	404
difference-map prompt	speed	26	39	35	–	–	–	–	404
point-trajectory overlay (after Motion-as-Prompt)	asymmetry	0	100	–	–	–	–	–	404
point-trajectory overlay (after Motion-as-Prompt)	scramble	83	17	–	–	–	–	–	404
point-trajectory overlay (after Motion-as-Prompt)	speed	27	39	34	–	–	–	–	404
motion blur (after MotionSight)	asymmetry	–	100	–	–	–	–	–	404
motion blur (after MotionSight)	freeze	69	31	–	–	–	–	–	404
motion blur (after MotionSight)	speed	34	33	33	–	–	–	–	404
32 frames (at most 16 fps)	asymmetry	0	100	–	–	–	–	–	404
32 frames (at most 16 fps)	speed	30	35	35	–	–	–	–	404
64 frames (at most 30 fps)	asymmetry	0	100	–	–	–	–	–	404
64 frames (at most 30 fps)	speed	30	35	35	–	–	–	–	404
MotionChat	speed	30	35	35	–	–	–	–	404
InternVL3.5-8B	freeze	81	19	–	–	–	–	–	404
InternVL3.5-8B	speed	30	35	35	–	–	–	–	404
zero-shot Qwen2.5-VL-7B	static emotion	3	8	–	10	79	1	–	286

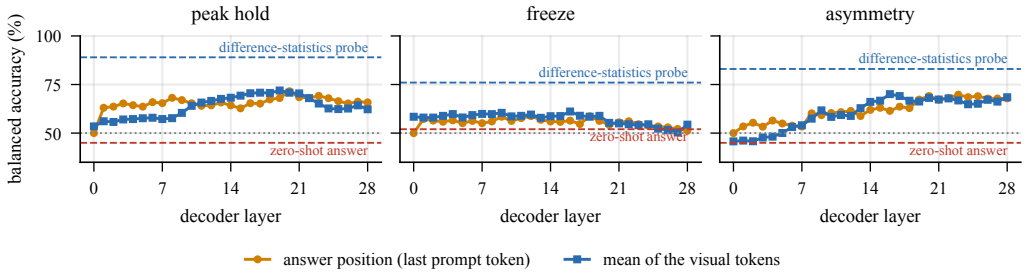

Figure 8: **Linear probe by decoder layer** (Qwen2.5-VL-7B, faces test, 404 items per edit type, zero-shot prompt). At every layer a logistic regression is fitted, under the actor-grouped five-fold split of §4.2, on the hidden state at the answer position (the last prompt token) and on the mean of the visual-token positions. Dashed lines: the difference-statistics probe of Table 19 and the zero-shot answer.

E.2 Where the Signal Is Lost, Stage by Stage

The decoder performs the cross-slot comparison only in part, and least on the edit type whose signature is a short still in mid-clip. Fig. 8 asks how much of the edit reaches the answer position. The hidden state at the answer position separates PEAK HOLD at 63–72 and ASYMMETRY at 53–70 depending on the layer, and FREEZE at no more than 59; the mean of the visual-token positions rises through the middle layers to 72, 70 and 61 on the three edit types. Both stay well below the difference-statistics probe on the same tokens (89 / 83 / 76).

The two short stills behave differently by stage. Table 22 reads the same 404 faces items per edit type at four points: the sampled pixels (a logistic regression on six statistics of the mean absolute difference between consecutive sampled frames, 64×64 grey: minimum, mean, maximum, standard deviation, share below 0.002 and log ratio of the two halves; fitted under the same actor-grouped folds as the probe), the visual-token statistics (§4.2), the zero-shot answer and the answer after three epochs on frames. On a 0.25 s FREEZE neither the pixel-difference baseline (50, against 69, 92 and 98 at 0.5, 1 and 2 s) nor the fine-tuned model (55, then 62, 95 and 95) separates the pair, so the sampled pixels carry no signal these six statistics detect. On a 0.25 s PEAK HOLD the sampled pixels carry a clear signal (93, then 95, 98 and 99) that the fine-tuned model uses only in part (67, then 88,

99 and 96). That second gap lies between the sampled pixels and the answer, and the table does not distinguish whether it arises in the encoder or in the readout; the first indicates only that these six statistics find no signal, not that no signal is there.

Table 22: The same 404 faces items per edit type read at four stages (balanced accuracy; chance 50, 33 for SPEED). Pixels: logistic regression on sampled-frame differences. Tokens: the linear probe on $\Delta+\Delta^2$ statistics of the visual tokens, out of fold on all 404 items. Fine-tuned: frames only, three epochs on the named types (ASYMMETRY, SPEED and SCRAMBLE from the model trained on those three).

stage	peak hold	freeze	asymmetry	speed	scramble
sampled pixels	96	78	82	–	68
visual-token statistics (probe)	89	76	83	58	56
zero-shot answer	45	52	45	33	50
fine-tuned answer	88	76	83	64	56

F THE THRESHOLD ON A NATURAL MOTION BENCHMARK

The threshold does not predict where more frames help on natural video. The receptive field $R = m\,s$ predicts that a larger frame budget should help exactly on the clips whose motion is faster than one token spans. We test this prediction outside WARPPAIRS on FAVOR-Bench (Tu et al., 2026), on a stratified subset of 2,064 close-ended questions over 1,185 videos (a larger sample than Table 14), with 8, 16, 32 and 64 frames spread uniformly over each clip, so that R per clip is m times the clip duration over the budget. For every video we measure a motion time scale τ, the time over which the mean absolute pixel difference reaches half its saturation value (median 0.4 s), and ask whether accuracy rises where R falls below τ.

On Qwen2.5-VL-7B accuracy rises with the budget from 45.8 at 8 frames to 51.1 at 64, but is flat across bins of $\log_2(\tau/R)$, and the within-item gain when R crosses below τ (+2.2 [+0.1, +4.4] points, paired over items) is no larger than the same statistic with τ permuted across videos (+3.2 [+1.0, +5.3]). The budget gain is larger on long clips, where R never falls below τ (+6.5 [+4.0, +9.2] from 8 to 64 frames), than on short clips, where it almost always does (+3.5 [+0.4, +6.5]). Qwen3-VL-8B shows the same pattern (48.5 to 55.2 across budgets; +3.9 [+1.7, +6.1] against a permuted +3.6 [+1.3, +5.9]; +8.2 on long against +3.7 on short clips) and InternVL3.5-8B a smaller one (47.1 to 50.4; +1.1 [−0.6, +2.8] against +1.5; +4.5 against +1.5). On natural video with annotated questions the benefit of more frames therefore comes from coverage of long clips, not from resolving motion inside one token, and the threshold of §5.3 is established only on the controlled edits.

G QUALITATIVE EXAMPLES OF THE TRAIN-FREE PROMPTS

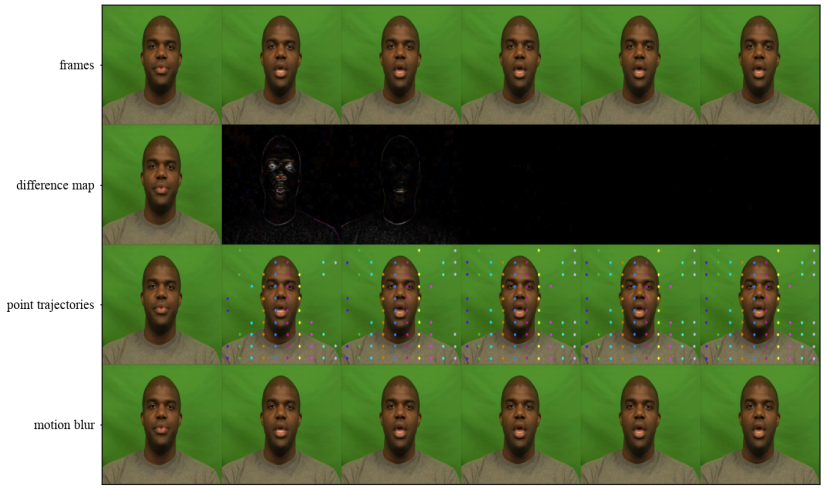

Figure 9: One PEAK HOLD item: raw frames; difference maps; Motion-as-Prompt-style trajectories (a grid of points tracked through the full-rate video since the previous sampled frame, drawn as a line ending in a dot); MotionSight-style motion blur (mean of the full-rate frames since the previous sample). After the apex the face is still.