PERCEPTION OR PRIOR? AUDITING VIDEO HALLUCINATION MITIGATORS WITH MOTION-BLIND TRAINING CONTROLS

Anonymous authors
Paper under double-blind review

ABSTRACT

Methods for mitigating hallucination in video large language models report their contribution as a delta on a benchmark, and a delta attributes nothing: it shows an intervention helped, not what it taught the model to use. We first show that two unreported choices move such numbers by more than the effects they adjudicate. Blacking out a clip costs Qwen2.5-VL-7B 43.1 points of VideoHallucer pair accuracy while repeating a single frame — position structure intact, motion exactly zero — costs 13.3, so the usual blinded control charges the static scene to the video; and reading an answer by generative parsing rather than per-option likelihood moves Video-LLaVA-7B by 15 points (9.5 versus 24.5, the published 17.8 between them). We then ask what a gain is made of, with a control the evaluation-side literature cannot build: five arms trained on one base model, one preference set and one LoRA-DPO recipe, differing only in the video seen *during training* — full, frame-permuted, black, absent, and order-preserving jitter — with matching machine-verified and every arm evaluated on untouched video. On TempCompass, whose questions are entirely temporal and the only benchmark of four where this recipe reproduces a published-scale gain (+5.2 points), the preregistered ratio-free contrast passes at the least favourable seed (+4.2 (+1.9; p=.006)): the motion-blind arm recovers 106% of the standard arm's gain and the text-only arm 12%. Elsewhere the preregistered denominator gate withholds the ratio, and on two benchmarks a faithful reproduction makes the model *worse*, which we report as a reproduction finding.

1 INTRODUCTION

Methods for mitigating hallucination in video large language models arrive at roughly two per month, split between training-free decoding interventions (Zhang et al., 2025b; Fan et al., 2026; Sun et al., 2025; Gao et al., 2026) and preference-optimisation recipes trained on public video preference data (Zhang et al., 2024; Huang et al., 2025; Li et al., 2025). Almost without exception the evidence offered is a delta on a video hallucination benchmark (Wang et al., 2024; Zhang et al., 2025b; Liu et al., 2024). That delta is the currency in which the field compares methods and decides what to build on.

A parallel diagnostic literature has established that such a delta need not reflect better perception: intervention grids across many video LLMs (Yang et al., 2026), evidence that correctness-only post-training amplifies spatiotemporal shortcuts (Du et al., 2026), decompositions finding large fractions of items answerable without genuine temporal understanding (Feng et al., 2025; Chen et al., 2026a; Li et al., 2026), and a survey recording as settled that preference- and RL-based mitigators gain partly by reshaping language priors (Huang et al., 2026; Xu, 2026). We take all of this as established, and claim none of it.

Two of the choices such a comparison rests on are never stated, and each moves a number by more than the effects it is used to adjudicate.

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.

The blinded control is worth a factor of three. To say how much a video contributed to a score, one compares the model with the video against the model without it — and "without" is a choice. On VideoHallucer, replacing every frame with black costs Qwen2.5-VL-7B 43.1 points of pair accuracy; replacing them with the median frame repeated to the original length, so position structure and duration survive and motion is exactly zero, costs 13.3. The 29.8-point difference is what a *static image* supplies, which the black-frame arm removes along with the motion and therefore charges to the video. Video-LLaVA-7B (Lin et al., 2024) loses 0.4 points under the same substitution: one repeated still frame reaches 98% of its full-video score.

The decision rule is worth more than most published gains. Holding the model, the documents, the frames and the prompt fixed and changing only how an answer is read off moves Video-LLaVA-7B's VideoHallucer pair accuracy from 9.5 to 24.5 — 15 points, with the published 17.8 falling between the two rules, which disagree on 19.2% of items. The mechanism is confirmed rather than assumed: forcing a choice between $P(\text{yes})$ and $P(\text{no})$ partially counteracts a yes-bias, raising accuracy on the hallucinated member of each pair from 0.112 to 0.359, and parse failure is excluded. Benchmarks in this area do not state which rule they use.

Both findings are about the instrument. Neither says what a mitigator's gain is made of, and that question needs a control the evaluation-side literature cannot construct.

What none of it supplies is the quantity a reader needs when one method reports a larger delta than its predecessor: *how much of a reported gain required motion in its training data*. Those diagnostics are evaluation-side — they perturb inputs at test time and thereby characterise models or benchmarks — but a method's gain is a property of the procedure that produced it, and no test-time perturbation isolates that. The missing control is constructive: a model produced by the same recipe, from the same preference pairs, whose *training* video is structurally incapable of carrying motion. It costs a training run that diagnostic papers have no reason to perform and mitigator papers no incentive to report. We supply that measurement.

The population audited, and where we sit. Every such method reports its contribution as a difference of benchmark scores, which attributes nothing: it shows an intervention helped, not what it taught the model to use. We audit a named set rather than a population (§5). **Mirage** (Yin et al., 2026) is the closest methodological ancestor — gains reported for contrastive decoding in image LVLMs are consistent with a shift of the output distribution rather than with reduced hallucination — but it is image-side and decoding-time throughout, as is the concurrent extension of its critique (Bendre et al., 2026). We carry the *critique*, that a benchmark delta need not reflect perception, to video and to *training-based* methods; we do not measure an output-distribution shift, which is Mirage's mechanism and is defined on decoding. The step is not mechanical, since an analysis of a decoding rule has no purchase on a method that changes weights. Two boundaries are worth drawing. **Shuffle-negative preference optimisation** (Li et al., 2025) uses order-corrupted video *constructively*, as the dispreferred branch of a pair; our shuffled arm inverts that role, the permuted video being the only video the model ever sees. And **shuffling at evaluation time** measures a fixed model's sensitivity to frame order, confounding information removal with an off-manifold penalty; we shuffle during *training* and evaluate every arm on untouched video. Relatedly, Lee et al. (2026) diagnose directional *motion blindness* as a capability defect; our motion-blind arm is an experimental control.

A matched-recipe control grid. We train five arms on one base model (Qwen2.5-VL-7B; Bai et al., 2025), one public preference set (LLaVA-Hound-DPO, 17k pairs) and one LoRA-DPO recipe anchored to that release's published hyperparameters (Zhang et al., 2024; Rafailov et al., 2024; Hu et al., 2021), differing *only* in the video seen during training (Figure 1). The load-bearing one is A2, a uniform frame permutation: every static and set-level cue survives while temporal order is destroyed, so whatever it recovers is obtainable without motion-informative training data. A4 sees no visual tokens at all, isolating language-prior recalibration, and A5 applies an order-preserving jitter (measured Kendall τ against the identity: $+0.971$ against -0.005 for A2) that bounds how much of A2's gain is mere augmentation robustness. A recipe digest with per-run data-order and text hashes certifies that the arms differed in pixels and nothing else. A second base with a differently designed temporal position encoding was preregistered as a robustness check and was not trained (§5), so every number here is on one base.

Inference frozen in advance. The verdict on the $\geq 50\%$ claim never passes through a ratio: the primary test is the linear contrast $H_1 : 2\Delta_{\text{A2}} - \Delta_{\text{A1}} > 0$, equivalent to $\text{RGF} \geq 1/2$ whenever $\Delta_{\text{A1}} > 0$, with the recovered-gain fraction reported as an effect size behind a denominator gate (Fieller, 1954). The protocol, the seed handling and every degrade path were registered before any arm was trained, and the design is informative in both directions — a motion-blind arm that *fails* to recover the gain is a quantitative rebuttal of the prior-reshaping narrative and is reported as one. In the event, three preregistered criteria did not pass, and §4.2 states which and why.

Where the question bites. A whole-set result is partly implied in advance: if many items never needed motion, an arm trained without it recovers their share arithmetically. The load-bearing question is what happens on items that *do* need it, and we built a motion-necessary stratum from held-out models to ask it. That analysis is not reported here — the ensemble could not be assembled under our runtime (§4.3) — but one of the four benchmarks, TempCompass, asks only temporal questions, and it is the one on which the recipe reproduces a published-scale gain and on which the preregistered test passes.

What we do not claim. We do not claim the first intervention grid on hallucination benchmarks: Yang et al. (2026) built that instrument and we use it. We do not claim to discover that mitigator gains draw on language priors — that is received wisdom (Huang et al., 2026; Du et al., 2026; Xu, 2026); our contribution is to quantify it per method under one ruler. We offer no additive causal decomposition of a gain into perception and prior components: a permuted frame sequence is off-manifold for a model with temporal position encodings, so training on it may cost something that has nothing to do with the information removed. A5 is the control for that, not an appeal to a correction we did not apply. We do not audit published mitigators as a population. And "motion-blind" here names an experimental control, not a model capability deficit as in Lee et al. (2026).

2 METHOD

Three quantities carry the audit. An *interventional sensitivity profile* (ISP) is a configuration's vector of accuracies across the evaluation-side intervention arms, with each arm's paired delta against `full`. The *recovered-gain fraction* is $\text{RGF}(\text{arm}) = \Delta_{\text{arm}}/\Delta_{\text{A1}}$, each Δ a clean-evaluation gain over the shared base. The *motion-necessary set* (MNS) of a benchmark is the subset of items a held-out ensemble answers with the video but not with a static reference input. None is decomposed into additive causal shares: shuffling frames feeds an off-manifold input to a model with temporal position encodings, so a delta mixes information removal with an out-of-distribution artifact whose shares are not identified. Profiles are sensitivity vectors, negative entries included, not budgets that sum to a gain.

2.1 MATCHED-RECIPE CONTROL ARMS

Reported video-DPO gains are not comparable across papers because base model, preference data, hyperparameters and evaluation rule all move at once. This design fixes every one and moves one thing: five arms train on one base from one public preference set with one LoRA-DPO recipe (Hu et al., 2021; Rafailov et al., 2024) whose hyperparameters are verbatim from a published one (Zhang et al., 2024), differing only in the video seen during training (Figure 1).

A2 is the load-bearing control: a uniform frame permutation destroys temporal order while leaving every frame, object and scene intact, so a model trained on it cannot have derived any preference signal from temporal order, and whatever gain it recovers is obtainable from a recipe whose training video carried no motion. The scope of that statement is exact: every arm is *evaluated* on untouched video, so what is measured is the fraction of a gain obtainable without motion-informative training data, not the fraction obtainable without motion at inference (§5). A4 removes the modality rather than corrupting it, and is implemented as such — an empty frame list would have become A3 the moment downstream code read "no frames" as "black frames". A5 answers the objection that A2's gain might be augmentation robustness rather than content: it perturbs the frame sequence while preserving order, at Kendall $\tau = +0.971$ against the identity with maximum displacement 2, against -0.005 and unbounded displacement for A2. The two are not matched in strength — A2's

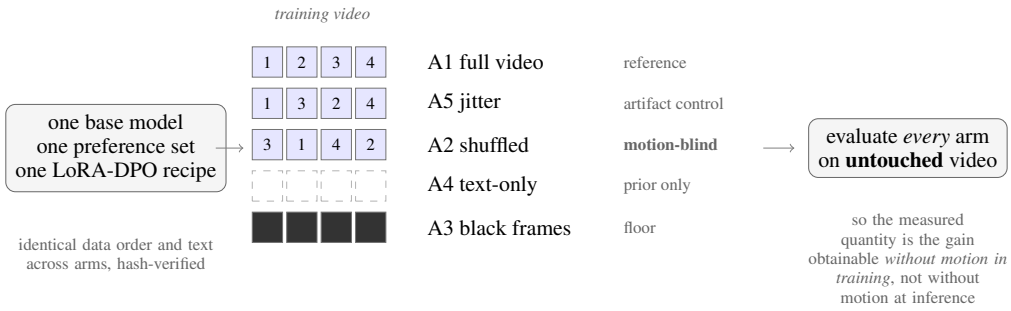


Figure 1: The control. Five arms share a base model, a preference set, a recipe, a data order and a text stream; they differ only in the video seen *during training*. A2 keeps every frame and destroys only their order, so a model trained on it cannot have derived a preference signal from temporal order. A5 perturbs the frame sequence while preserving order, which is what separates "motion was unnecessary" from "noisy training data helps". A3 keeps the token structure of watching a video with none of its content, and is the floor. Every arm is then evaluated on untouched video.

perturbation is far larger on both axes — so Δ_{A5} bounds the augmentation component only under the assumption that a weaker perturbation does not buy more (Appendix A.1).

2.2 Why the arms are matched, and how a machine checks it

Every claim here is a between-arm contrast, so an arm differing in anything but its training video would make its contrast a mixture of the intervention and whatever else drifted — invisible in a loss curve, since a run trained at the wrong learning rate looks entirely normal. Matching is therefore a machine-checked gate that must pass before any headline number is computed. The recipe is reduced to a receipt, the SHA256 of its canonical JSON with `arm` and `seed` removed. But configuration equality shows only that runs were *launched* alike, so each run also hashes what passed through the loader, and two invariants are checked rather than assumed: within a seed all arms must produce an identical `data_order_hash`, and across all runs including the text-only one the `text_hash` must be identical. Either violation would corrupt every contrast in the paper while both the loss curve and the accuracy table looked normal. Appendix D.2 lists the full check set and the defects it caught here — all of them silent.

2.3 The evaluation-side profile, and why the reference is frozen-frame

Grids of this kind are established instrumentation (Yang et al., 2026; Feng et al., 2025; Chen et al., 2026a); we claim no novelty for the grid and apply it to *mitigators* rather than to models. The one choice not inherited is which arm is the reference for "what the video contributed". The obvious answer, `blind`, is measurably the wrong one — it deletes the static scene along with the motion and charges both to the video (§4.4) — so the preregistered rule selects, per benchmark, the single blinded arm with the highest pooled blinded accuracy, which is the hardest bar for calling an item motion-necessary. It selects `frozen_frame`.

2.4 Frozen inference protocol

The protocol was preregistered and hash-pinned with the module implementing it; amendments are new dated files, never edits. The $\geq 50\%$ claim is adjudicated by the ratio-free linear contrast $H_1 : 2\Delta_{\mathrm{A2}} - \Delta_{\mathrm{A1}} > 0$, item-level, video-clustered paired bootstrap, one-sided at $\alpha = 0.05$ — equivalent to $\mathrm{RGF} \geq 0.5$ under $\Delta_{\mathrm{A1}} > 0$ and never dividing. RGF is an effect-size report, and a benchmark reports one only if its standard-arm gain has a point estimate of at least $+4.0$ points and a 95% CI lower bound above $+2.0$; below that gate it contributes absolute differences and no ratio, which keeps the threshold claim from quietly becoming unfalsifiable when the reference arm's gain is small. Seeds are a fixed sensitivity factor — three is not an inference sample — so we never bootstrap over them and the *least favourable* seed is the reported bound (Appendix A.2).

2.5 THE MOTION-NECESSARY SET

A whole-set result is close to arithmetically implied by existing work (Feng et al., 2025; Chen et al., 2026a): the informative question is what happens on items that do require motion. The stratum built to ask it is defined by *held-out* models spanning several vision–temporal families (Zhu et al., 2025; Zhang et al., 2025d;a; Li et al., 2024; Wang et al., 2025), with the audit base excluded — were the model whose gains are being decomposed to help define the stratum those gains are conditioned on, the stratum would partly describe that model. Each item yields Bernoulli observations under `full` and under the frozen blinded reference and enters the stratum when the gap exceeds τ at BH-FDR level q (Benjamini & Hochberg, 1995), under a *paired permutation* null: an exact binomial on pooled counts would treat repeated measurements as independent and admit too many items, in the direction that flatters our own claim. Items are admitted on one prompt-replicate set and the effect estimated on the other, which severs selection-on-noise. The ensemble could not be assembled under our runtime and the conditional analysis is not reported (§4.3); the construction is described here and released because it is part of the contribution even unrun.

3 EXPERIMENTAL SETUP

All control arms train from **Qwen2.5-VL-7B-Instruct** (Bai et al., 2025) on the public **LLaVA-Hound-DPO** 17k preference set (Zhang et al., 2024), with the visual pathway configured identically on the training and evaluation sides (32 frames, evenly spaced) — profiles are measured at evaluation time and read as evidence about training, so a tokenisation mismatch would enter every one. The preregistered second base (Zhu et al., 2025) was not trained. The recipe is frozen in a pre-registration before any non-pilot arm runs and each value carries its provenance: the hyperparameters that define the method are upstream's verbatim ($\beta = 0.1$, three epochs as planned — 2.65 realised, §5 — effective batch 128), the sequence limit and batch factorisation are recomputed because upstream's are Video-LLaVA's, and only the LoRA geometry (r=16, α=32, language tower only) and learning rate ($1\mathrm{e}{-5}$) are ours, having no upstream counterpart. That split is what makes A1 a hyperparameter-faithful reproduction rather than a recipe of our own invention. Whether it is *representative* of published recipes was to be tested, not assumed, and that test could not be run (§5). One ordering detail decides whether A2 means what it claims: frames are subsampled to the 32-frame budget *first* and the arm's frame plan applied *second*, since uniform subsampling of a permuted sequence re-imposes part of the original ordering.

Benchmarks and the decision rule. After a pre-committed availability check the suite is Video-Hallucer (1,952 items in native pairs; Wang et al., 2024), EventHallusion (409 items, scored by item because the staged release yields no polarity pair; Zhang et al., 2025b), TempCompass-temporal (order, direction and speed only; Liu et al., 2024), and DyBench (1,507 counterfactual pairs; Du et al., 2026); MoHallBench was not located and is dropped under the frozen rule. A pair scores 1 only if *both* polarities are correct, which is immune to inflation by a constant yes-bias (Wang et al., 2024; Chen et al., 2026b). Every number on the critical path is a deterministic teacher-forced per-option log-likelihood, argmax over options, with no sampling, generation, judge model or external API (Zhang et al., 2025c); the log-softmax is computed in float32 even under bfloat16 weights, since accumulating bf16 log-probabilities over a long video prompt can flip an argmax between close options. Pinning the rule is not stylistic — our own measurements make it worth 15 points of pair accuracy with everything else fixed (§4.4).

Evaluation arms. The profile spans seven arms, from `full` through progressively stronger removals of temporal and visual information to `blind`. Every arm is a pure frame plan computed deterministically from a hash of (global seed, item id, arm), with permutations resampled until non-identity and interventions keyed by *video* so paired questions see identical pixels. The training-side arms call the same planner, so the training-side shuffle is provably the same operation as the evaluation-side one rather than a second implementation asserted to agree.

Appendix D gives the recipe table and the derivation of each recomputed value, the matching checks and the defects they caught, the reversal whitelist and exclusions, the blinded reference's selection and stability analysis, the 800-clip synthetic corpus in which motion-necessity is known by construction, the audited-subject list (Appendix D.7), and the compute arrangement.

Table 1: Every number the design produces. Rows are benchmarks; the first five columns are each arm's paired gain over the shared base in percentage points; the last two are the preregistered verdict. GATE is the denominator gate, which requires $\Delta_{A1} \geq +4.0$ with a 95% lower bound above $+2.0$ before a benchmark may report a ratio. The contrast is $2\Delta_{A2} - \Delta_{A1}$ with its one-sided 95% lower bound and bootstrap p; by the frozen seed rule this is the least favourable seed, not the mean. Arms differ only in the video seen during training; all are evaluated on untouched video. Per-seed values and intervals are in Appendix B.2.

Benchmark	A1 *full*	A5 *jitter*	A2 *shuffled*	A4 *text-only*	A3 *blind*	$2\Delta_{A2} - \Delta_{A1}$ (LB; p)	GATE
VideoHallucer	+1.7	+1.6	+1.5	+0.6	-0.8	+1.3 (-0.6; p=.130)	withheld
EventHallusion	-1.7	-1.2	-2.2	-0.7	-2.4	-3.4 (-6.1; p=.995)	withheld
TempCompass-T	+5.2	+5.2	+5.5	+0.6	+0.6	+4.2 (+1.9; p=.006)	pass
DyBench	-2.7	-2.9	—	—	-2.1	n/a	withheld
Pooled	+0.6	—	+0.8	—	—	+0.9 (-0.8; p=.142)	—

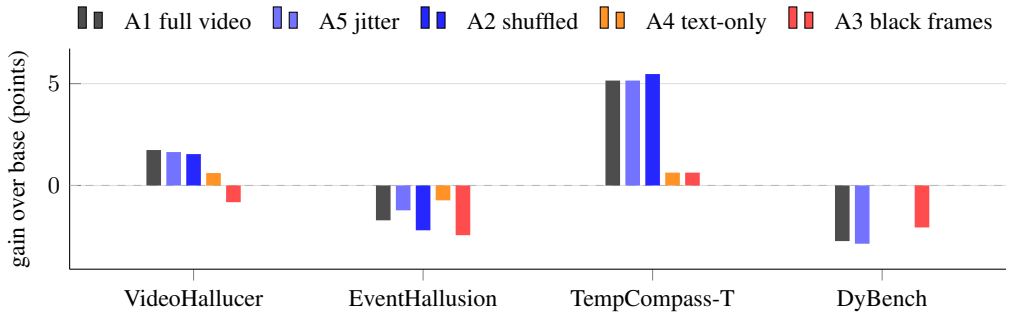


Figure 2: What each arm gains over the shared base, in percentage points, at seed 0. The arms differ only in the video seen *during training*; all are evaluated on untouched video. Where the recipe produces a gain at all, the arm trained on randomly ordered frames (A2) tracks the arm trained on real video (A1), while the arm that saw no video (A4) does not and the black-frame floor (A3) can fall below the untrained base. Headline seed; per-seed ranges in Appendix B.2. DyBench shows three bars rather than five: its shuffled and text-only evaluations did not complete, and the cells are left absent rather than imputed.

4 RESULTS

Every accuracy comes from one frozen decision rule and every quantity we interpret is a *contrast between arms* under it, in which a calibration offset common to the arms cancels. Intervals are video-clustered paired bootstraps, $B = 1000$, the arms sharing each resample so the Fieller covariance is valid (Fieller, 1954; Efron & Tibshirani, 1993); multi-seed arms report the *least favourable* seed (Appendix B.2).

4.1 WHAT THE GAIN IS MADE OF

Table 1 is the result. It is five absolute differences, each against the same untuned base, measured under one frozen decision rule.

The ordering is the result, and it is the same wherever the recipe produces a gain at all. **Permuting time costs almost nothing**: on TempCompass the arm whose training video carried no temporal order recovers 106% of the standard arm's gain, and on VideoHallucer 88%. **Removing vision does**: the text-only arm recovers 12% and 35% respectively, so the gain is not a recalibrated language prior and what survives a permutation is what a single frame shows. **The floor behaves like a floor**: A3, which keeps the token structure of watching a video while removing all visual content, lands at +0.6 and -0.8, at or below the untuned base — a design that produced gains whatever it was fed would not do that. And A5, which perturbs the frame sequence as visibly as A2 but preserves order, tracks

A1 at +5.2, so training-time perturbation *per se* buys nothing and A2's gain is not augmentation robustness.

On two benchmarks the recipe produces no gain at all. EventHallusion's standard arm is -1.7 and DyBench's is -2.7: a faithful reproduction of a published video-DPO recipe makes the model *worse* on two video hallucination benchmarks. There is then no gain to decompose, and every share in those columns is withheld rather than computed. We checked the obvious explanation for EventHallusion and it does not hold — the yes-rate barely moves (0.472 to 0.469) while accuracy falls — so this is not a label-distribution artifact. We report it as a reproduction finding.

4.2 What the preregistered test says, and what it does not

The preregistered verdict on the $\geq 50\%$ claim is the ratio-free linear contrast $2\Delta_{A2} - \Delta_{A1} > 0$ (§2.4). Table 1 reports it, and two preregistered criteria fail. We state both rather than reporting around them.

On TempCompass the preregistered test passes. That benchmark is the only one of the four whose questions are explicitly temporal — order, direction and speed — and the only one on which the recipe reproduces a published-scale gain: Δ_{A1} =+5.2 points, inside the $+3$ to $+8$ band the reproduction criterion names. The denominator gate passes on all three seeds, so a ratio is reportable here; the contrast is +4.2 (+1.9; p=.006) at the least favourable seed, and A2 recovers 106% of A1's gain (89–106% across seeds) while A4 recovers 12%. A model whose training video carried no temporal order recovers the whole of a gain measured on temporal questions.

Elsewhere the gate withholds the ratio, and on two benchmarks there is no gain to divide. VideoHallucer's standard arm gains +1.7 points, below the $+4.0$ the gate requires, so it reports absolute differences only and its contrast (+1.3 (-0.6; p=.130)) does not reach significance at the least favourable seed — the cause is visible in Figure 2: a test defined against a small Δ_{A1} has little power. On EventHallusion and DyBench the standard arm is *negative*, so no share of it is meaningful and every ratio there is withheld rather than computed. This is the preregistered degrade path, not a choice made after seeing the data.

And the reproduction criterion holds on one benchmark of four. A1 lands in the published band on TempCompass and below it elsewhere. A hyperparameter-faithful reproduction of a published recipe therefore reproduces its effect size on the benchmark closest to what the recipe targets and not on the others, which we record as a partial pass rather than adjusting the band.

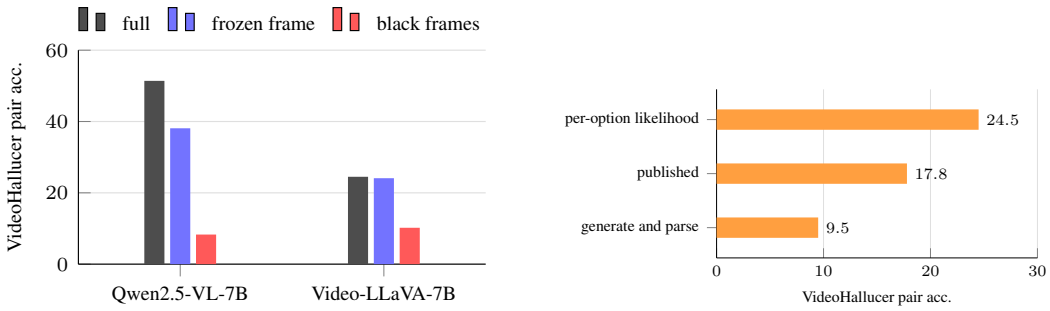
What survives all three is Table 1: the ordering of five absolute differences, each with its own interval, which is what this design was built to measure.

4.3 The conditional analysis is not reported

A whole-set result is close to arithmetically implied by existing benchmark decompositions (Feng et al., 2025; Chen et al., 2026a; Zhang et al., 2026): if most items never needed motion, an arm trained without it recovers their share arithmetically. The informative question is what happens on items that *do* require motion, and §2.5 describes the stratum built to ask it. We do not report that analysis: the stratum is induced from held-out models, one of the five named in the protocol load under the runtime available to us, and we do not patch a model's released code to make it run — our fork would then be an unstated implementation choice of exactly the kind under audit. A stratum from a fraction of the preregistered ensemble is not the preregistered object, and reporting a conditional number against it would be the substitution this paper exists to object to. The definition code, the permutation test, the split-replicate protocol and the synthetic corpus are released so the analysis can be run when the ensemble is complete.

4.4 The two measured findings in full

Figure 3 gives the numbers behind the two results stated in §1; the full seven-arm profile and the per-rule item table are in Appendix B.6 and B.5. Two readings beyond the threefold gap: EventHallusion's small deltas under the same substitution are an artifact of a near-uniformly negative label distribution rather than evidence that the video contributes two points there — a blinded Qwen



(a) What "without the video" means.

(b) One model, one benchmark, two ways to read an answer.

Figure 3: Two measured choices, neither usually reported, each larger than the effects it is used to adjudicate. **(a)** Blacking out a clip costs Qwen2.5-VL-7B 43.1 points; replacing every frame with the median frame repeated — position structure and duration intact, motion exactly zero — costs 13.3. The difference is what a still image supplies, which the black-frame arm removes along with the motion and charges to the video. Video-LLaVA-7B loses 0.4 points under the same substitution: one still frame reaches 98% of its full-video score. **(b)** Holding model, documents, frames and prompt fixed and changing only how the answer is read off moves the score by 15 points, with the published figure between the two rules. The benchmark does not state which rule it used.

collapses to a 0.012 yes-rate and lands on the correct answer by construction, which is the empirical case for selecting the strongest blinded arm rather than defaulting to blind. And a criterion of the form "recover a published number within ± 2 points" is not executable when an unstated implementation choice moves the result by an order of magnitude more than the tolerance; we record it as *not evaluable* rather than adopting whichever rule lands closer. The two rules also expose different amounts of language prior — the generative rule the unconditioned prior, the likelihood rule what survives a forced comparison — so for a paper about separating perception from prior the measurement rule is not neutral with respect to the quantity measured.

4.5 The training-free leg is not reported

The preregistered scope also included training-free mitigators: temporal contrastive decoding (TCD; Zhang et al., 2025b), whose released repository ships the benchmark and the evaluator but no implementation of the method, and STEAR (Fan et al., 2026), whose published description leaves four implementation choices open. We built the TCD audit — the counterpart is itself a frame plan, so shuffled, reversed and blind counterparts come from the same planner as the evaluation arms, and the question is whether a gain tracks temporal destruction at all — but did not run it within the compute available here. It is released rather than reported. One property of the method surfaced while building it and is worth stating regardless: a frame budget is a ceiling, not a count, so asking for an 8-frame counterpart from a clip already shorter than eight frames returns the original frames, the correction is identically zero, and the method silently reduces to its own baseline while every row still carries its label. VideoHallucer's clips run 4–69 frames, so this is not a corner case; degeneracy must be tested on the *realised* model inputs rather than on the requested budget, since a guard on the budget passes trivially.

5 What we would ask, and what we cannot claim

Three changes would cost almost nothing. State the decision rule, since an unqualified accuracy is not comparable across papers. Report the blinded reference used, and prefer one that is out-of-distribution-calibrated — a repeated static frame rather than black frames — so the static scene is not counted as a contribution of the video. And, for training-based mitigators, run the recipe once more on shuffled frames: one additional run, no new data and no new hyperparameters, converting an unfalsifiable claim about perception into a number. We release the stratum, the intervention tooling and the control-arm checkpoints so that the third is a download rather than a project.

This paper's argument is that unstated implementation choices move reported numbers by more than the effects those numbers support. That obliges us to state ours, including the ones that cost us a preregistered check. Appendix C gives the evidence behind each.

Three preregistered criteria did not pass. The reproduction gate (A1's clean gain within the published band), the denominator gate (no benchmark clears $+4.0$ points, so no ratio is reported anywhere), and the threshold test at the least favourable seed. Each is reported as failing in §4.2 rather than adjusted. A fourth, a harness-level requirement to reproduce a published score within ± 2 points, is recorded as *not evaluable*: its precondition does not hold, because the decision rule alone moves that score by 15 points and the benchmark does not state which rule it used.

What "motion-blind" scopes. A2's training video carries no temporal order, but A2 is *evaluated* on untouched video. So the measurement is the fraction of a gain obtainable without motion-informative training data, not the fraction obtainable without motion at inference. Those are different claims and only the first is licensed. In particular, preference optimisation on order-uninformative video could elicit temporal competence the base model already holds, and such a gain would still require motion at test time.

The arms stop short of three epochs. Every arm trained for 350 of 396 planned optimizer steps, the final segment having deadlocked identically for every arm. All stopped at the *same* step with byte-identical data consumption, so the *differences* are comparisons at one training budget. The *ratio* is weaker: equal truncation makes the two gains commensurable, not their quotient invariant to where training stops.

Permuted training video is off-manifold, and the correction for it was not applied. A model with temporal position encodings has never seen a permuted sequence, so A2 may lose something that has nothing to do with the temporal information removed — which would make A2's gain an underestimate and the paper's conclusion conservative rather than inflated. A5 bounds this from one side: it is off-manifold in the same way while preserving order. The preregistered calibration, an artifact floor estimated per question type on a corpus where answerability is known by construction, was built and released but not run, so no delta reported here is artifact-corrected. Note also that the floor as designed is an *evaluation-side* quantity, measured on models scored under a shuffled input; transferring it to a training-side manifold problem is an assumption we do not make here.

Scope, and what "matched" does and does not mean. The audited set is our reproduction plus the methods whose artifacts an availability check found public; mitigators here appear at roughly two per month, so this is a snapshot, not a survey. One planned benchmark had no locatable release; one arm is unscorable for one model; EventHallusion's near-uniformly negative questions leave it scored by item rather than by pair, so the bipolar bias protection degrades there. On the training side there is one preference dataset and one base — the preregistered second base was not trained — with LoRA adapters rather than full-parameter DPO, and the LoRA geometry and learning rate have no published counterpart and are our judgement. The preregistered defence of treating A1 as representative, a profile comparison against VistaDPO (Huang et al., 2025), could not be run because its artifacts were never released; we execute the preregistered downgrade rather than arguing around it, so no sentence here extrapolates from A1 to published video-DPO recipes as a class. "Matched recipe" is therefore a statement about the arms — identical recipe digest, data-order hash and text hash, verified before any headline number — not a claim of identity with any published run. All scoring is deterministic multiple-choice and bipolar binary, so free-form captioning and user-facing helpfulness lie outside this audit.

AI USE STATEMENT

Generative AI tools were used in this work for code authoring and for editing prose, and their output was verified rather than trusted. All analysis code was written with AI assistance and is covered by a unit-test suite that the authors reviewed; several defects introduced in that process are documented in Appendix D.2, including one that attached LoRA adapters to a tower the recipe excludes and one that advanced a data cursor per rank rather than per group. Neither was visible in a loss curve, and both were caught by machine checks written for that purpose. AI tools were also used to draft and revise

sections of this manuscript; every claim, number and citation was checked against the underlying artifacts by the authors. Generative AI was not used to generate research ideas, to produce any experimental result, or to author any number reported in this paper: every figure in every table comes from the frozen analysis code applied to recorded model outputs. We take responsibility for the final content of this work.

ETHICS STATEMENT

This study evaluates publicly released models on publicly released benchmarks using a publicly released preference dataset, and involves no human subjects and no human annotation at any point — the validation stack is anchored on a procedurally generated corpus rather than on annotator judgement. The audited methods are identified by name because auditing them is the contribution; we have tried to state the scope of each finding precisely, since a result that bounds what a method's benchmark evidence establishes is not a claim that the method is without value, and we say so explicitly in §5. Where a preregistered check could not be run because an artifact was never released, we report it as unrun rather than substituting a favourable alternative (§5).

REPRODUCIBILITY STATEMENT

The elements a reader would need to reproduce or contest this work are frozen in advance and released. Two pre-registrations fix the inference protocol — the linear contrast, the denominator gate, the Fieller construction, the seed rule and every degrade path — and the training recipe with each hyperparameter's provenance; both were hash-pinned together with the modules implementing them before any non-pilot arm ran, and amendments are new dated files rather than edits. The stratum threshold and FDR level were to be frozen in a third file before any conditional number was computed; that file does not exist, so those two constants are reported as chosen rather than preregistered (§5). Arm matching is not asserted but machine-verified: every run emits a recipe digest over its canonical configuration together with hashes of the data order and text actually consumed, and a verification pass must exit zero before any number in this paper is computed (§2.2, Appendix D.2). Scoring is deterministic throughout — no sampling, no generation, no judge model and no external API on the critical path — so a re-run on the same inputs gives the same numbers. We release the intervention tooling, the analysis code with its test suite, the motion-necessary stratum, and the control-arm adapters, each bound to its arm label by a hash of its own weights so that a mislabelled checkpoint is detectable rather than plausible.

REFERENCES

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, 2025. URL https://arxiv.org/abs/2502.13923.

Arnav Bendre, Guneesh Gupta, Kavish Grover, Chayan Aggarwal, and Shreyansh Modi. Rethinking cd: A reproducibility study and extension on the ineffectiveness of contrastive decoding at mitigating object hallucinations in mllms, 2026. URL https://arxiv.org/abs/2607.25196.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995.

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, 2026a. URL https://arxiv.org/abs/2605.09904.

Sihan Chen, Jiale Li, Jianghang Lin, and Mengyuan Liu. Mohallbench: A benchmark for motion hallucination in video large language models, 2026b. URL https://arxiv.org/abs/2607.01117.

Dazhao Du, Jian Liu, Jialong Qin, Tao Han, Bohai Gu, Fangqi Zhu, Yujia Zhang, Eric Liu, Xi Chen, and Song Guo. Learning spatiotemporal sensitivity in video llms via counterfactual reinforcement learning, 2026. URL https://arxiv.org/abs/2605.21988.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall/CRC, 1993.

Linfeng Fan, Yuan Tian, Ziwei Li, and Zhiwu Lu. Stear: Layer-aware spatiotemporal evidence intervention for hallucination mitigation in video large language models, 2026. URL https://arxiv.org/abs/2604.03045.

Bo Feng, Zhengfeng Lai, Shiyu Li, Zizhen Wang, Simon Wang, Ping Huang, and Meng Cao. Breaking down video llm benchmarks: Knowledge, spatial perception, or true temporal understanding?, 2025. URL https://arxiv.org/abs/2505.14321.

E. C. Fieller. Some problems in interval estimation. *Journal of the Royal Statistical Society: Series B (Methodological)*, 16(2):175–185, 1954.

Yuansheng Gao, Wenbin Xing, Jiahao Yuan, Kaiwen Zhou, Han Bao, Zonghui Wang, and Wenzhi Chen. Multitop: Learning to patch visual tokens to mitigate hallucinations in video large multimodal models, 2026. URL https://arxiv.org/abs/2606.11792.

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, 2021. URL https://arxiv.org/abs/2106.09685.

Haojian Huang, Haodong Chen, Shengqiong Wu, Meng Luo, Jinlan Fu, Xinya Du, Hanwang Zhang, and Hao Fei. Vistadpo: Video hierarchical spatial-temporal direct preference optimization for large video models, 2025. URL https://arxiv.org/abs/2504.13122.

Yiyang Huang, Yitian Zhang, Yizhou Wang, Mingyuan Zhang, Liang Shi, Huimin Zeng, and Yun Fu. Distorted or fabricated? a survey on hallucination in video llms, 2026. URL https://arxiv.org/abs/2604.12944.

Jongseo Lee, Hyuntak Lee, Sunghun Kim, Sooa Kim, Jihoon Chung, and Jinwoo Choi. Which way did it move? diagnosing and overcoming directional motion blindness in video-llms, 2026. URL https://arxiv.org/abs/2605.22823.

Baiqi Li, Kangyi Zhao, Ce Zhang, Chancharik Mitra, Jean de Dieu Nyandwi, and Gedas Bertasius. Timeblind: A spatio-temporal compositionality benchmark for video llms, 2026. URL https://arxiv.org/abs/2602.00288.

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, 2024. URL https://arxiv.org/abs/2408.03326.

Shicheng Li, Lei Li, Kun Ouyang, Shuhuai Ren, Yuanxin Liu, Yuanxing Zhang, Fuzheng Zhang, Lingpeng Kong, Qi Liu, and Xu Sun. Temple: Incentivizing temporal understanding of video large language models via progressive pre-sft alignment, 2025. URL https://arxiv.org/abs/2503.16929.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection, 2024. URL https://arxiv.org/abs/2311.10122.

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?, 2024. URL https://arxiv.org/abs/2403.00476.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model, 2024. URL https://arxiv.org/abs/2305.18290.

Yiming Sun, Mi Zhang, Feifei Li, Geng Hong, and Min Yang. Smartsight: Mitigating hallucination in video-llms without compromising video understanding via temporal attention collapse, 2025. URL https://arxiv.org/abs/2512.18671.

Yi Wang, Xinhao Li, Ziang Yan, Yinan He, Jiashuo Yu, Xiangyu Zeng, Chenting Wang, Changlian Ma, Haian Huang, Jianfei Gao, Min Dou, Kai Chen, Wenhai Wang, Yu Qiao, Yali Wang, and Limin Wang. Internvideo2.5: Empowering video mllms with long and rich context modeling, 2025. URL https://arxiv.org/abs/2501.12386.

Yuxuan Wang, Yueqian Wang, Dongyan Zhao, Cihang Xie, and Zilong Zheng. Videohallucer: Evaluating intrinsic and extrinsic hallucinations in large video-language models, 2024. URL https://arxiv.org/abs/2406.16338.

Zekun Xu. When does a video-language model stop watching? reward strength controls the formation and reversal of visual shortcuts in multimodal rlvr, 2026. URL https://arxiv.org/abs/2606.22043.

Junqi Yang, Yuecong Min, Jie Zhang, Shiguang Shan, and Xilin Chen. Infact: A diagnostic benchmark for induced faithfulness and factuality hallucinations in video-llms, 2026. URL https://arxiv.org/abs/2603.11481.

Hao Yin, Guangzong Si, and Zilei Wang. The mirage of performance gains: Why contrastive decoding fails to mitigate object hallucinations in mllms?, 2026. URL https://arxiv.org/abs/2504.10020.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, Peng Jin, Wenqi Zhang, Fan Wang, Lidong Bing, and Deli Zhao. Videollama 3: Frontier multimodal foundation models for image and video understanding, 2025a. URL https://arxiv.org/abs/2501.13106.

Jiacheng Zhang, Yang Jiao, Shaoxiang Chen, Na Zhao, Zhiyu Tan, Hao Li, Xingjun Ma, and Jingjing Chen. Eventhallusion: Diagnosing event hallucinations in video llms, 2025b. URL https://arxiv.org/abs/2409.16597.

Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Lmms-eval: Reality check on the evaluation of large multimodal models, 2025c. URL https://arxiv.org/abs/2407.12772.

Ruohong Zhang, Liangke Gui, Zhiqing Sun, Yihao Feng, Keyang Xu, Yuanhan Zhang, Di Fu, Chunyuan Li, Alexander Hauptmann, Yonatan Bisk, and Yiming Yang. Direct preference optimization of video large multimodal models from language model reward, 2024. URL https://arxiv.org/abs/2404.01258.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Llava-video: Video instruction tuning with synthetic data, 2025d. URL https://arxiv.org/abs/2410.02713.

Yuxuan Zhang, EunJeong Hwang, Huaisong Zhang, Penghui Du, Yiming Jia, Dongfu Jiang, Xuan He, Shenhui Zhang, Ping Nie, Peter West, and Kelsey R. Allen. Watch before you answer: Learning from visually grounded post-training, 2026. URL https://arxiv.org/abs/2604.05117.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models, 2025. URL https://arxiv.org/abs/2504.10479.

A METHOD IN FULL

A.1 A5: MEASURING THAT IT PRESERVES ORDER, AND THE TWO DECISIONS THAT REQUIRED

A5 exists to separate two explanations of A2's gain, so its claim to preserve temporal order has to be measured rather than asserted — an A5 that quietly behaved like A2 would make the augmentation bound vacuous in the direction that flatters this paper.

Over 300 items at 32 frames, A5's frame plans reach a Kendall rank correlation of $+0.971$ against the identity ordering (minimum $+0.943$) with maximum frame displacement 2, while A2's sit at -0.005 (maximum $+0.383$) with unbounded displacement: two orders of magnitude apart in temporal preservation. Preservation is global rather than monotone — local inversions are the mechanism of the arm — so the testable property is a rank correlation, not an ordering predicate.

Obtaining that separation required two decisions that are easy to get wrong in the direction of our own conclusion:

- **Only non-overlapping pairs are transposed.** A sequential swap loop carries a frame arbitrarily far right; we measured displacements of 4 — a global reordering in local disguise, which would have blurred A5 toward A2 and weakened the augmentation bound.
- **Jitter plans are resampled until non-identity.** On short clips the identity plan is drawn often enough to matter (25.9% of samples at two frames), and an identity plan makes that sample's A5 literally A1 — biasing the augmentation bound toward zero, again toward this paper's own conclusion.

A.2 INFERENCE PROTOCOL: THE REMAINING CONSTANTS

Unbounded intervals. When $g = z^2\mathrm{Var}(\Delta_{\mathrm{A1}})/\Delta_{\mathrm{A1}}^2 \geq 1$ the denominator is not significantly nonzero, the Fieller confidence set is unbounded, and *no interval is reported*; the caller falls back to absolute deltas. Reporting a very wide interval instead would present an unbounded set as a bounded one.

Why the ratio is not tested directly. Testing $\mathrm{RGF} \geq 0.5$ by a percentile bootstrap of the ratio was rejected on simulation, not taste: with a planted true RGF, a naive percentile-bootstrap ratio interval is unusable near this study's operating point — its width explodes while it still returns a finite, reportable-looking interval. The linear contrast $2\Delta_{\mathrm{A2}} - \Delta_{\mathrm{A1}} > 0$ is equivalent under $\Delta_{\mathrm{A1}} > 0$ and never divides.

Pooling and the headline. The headline is the pooled RGF's one-sided 95% lower bound ($z = 1.645$), phrased "at least $X\%$", pooled unit-count weighted at the bootstrap-sample level over all benchmarks, gated or not. Per-benchmark bootstraps use distinct seeds so that pooling at the sample level is valid.

The denominator gate's degrade path. Below the gate a benchmark reports absolute paired deltas only and still enters the pooled RGF. The degrade path is preregistered so the verdict still runs, through the ratio-free contrast, rather than the threshold claim quietly becoming unfalsifiable when the reference arm's gain is small.

B RESULTS IN FULL

B.1 CLEAN GAINS OF ALL FIVE ARMS

B.2 PER-SEED RANGES

Seed is a fixed sensitivity factor, never a random effect, and we never bootstrap over seeds: three is not a valid inference sample. The main text reports the *least favourable* seed — the one arguing least strongly for H_1 — and the range across seeds is given here. Reporting the mean instead would let

Table 2: Clean accuracy of the base model and paired gain Δ (percentage points vs. the same base) per control arm. Metrics: pair accuracy (VideoHallucer, DyBench), item accuracy (EventHallusion), multiple-choice accuracy (TempCompass-temporal). A3 is a format/floor control, read qualitatively only — no summary quantity in this paper derives from it.

Arm	Seeds	VideoHallucer	EventHallusion	TempCompass-T	DyBench
Qwen2.5-VL-7B (base, absolute)	—	51.4	67.2	39.40	66.70
A1 standard-DPO (Δ)	3	+1.7	-1.7	+5.2	-2.7
A2 shuffled-frame DPO (Δ)	3	+1.5	-2.2	+5.5	—
A3 blind-frame DPO (Δ)	1	-0.8	-2.4	+0.6	-2.1
A4 text-only DPO (Δ)	2	+0.6	-0.7	+0.6	—
A5 jittered DPO (Δ)	1	+1.6	-1.2	+5.2	-2.9

one strong seed carry a claim the design says a single seed cannot carry; the implementation offers no option to do so.

Only the headline seed is resolved here. The seed-resolved pooled Δs and the seed-resolved contrast are not tabulated: the analysis that produced them for the headline seed was not repeated per seed within the compute available. We state that rather than printing a range we did not compute. The rule itself is not conditional on the outcome — it is fixed in the preregistration and implemented as the only selection the code offers — so the main text's numbers are the least favourable seed by construction, not by choice after seeing the three.

B.3 Rank-order decomposition

Table 3: Preregistered rank-order decomposition, in percentage points, at the headline seed. Each entry is an arithmetic difference of the per-arm gains in Table 1. No additive share is claimed: the three sources mixed into Δ_{A2} are not separately identifiable, so this is an ordering of bounds, not a variance decomposition. The $\max$ in the last column is taken per benchmark on the point estimates, as preregistered. Two cells are absent because DyBench's shuffled and text-only evaluations did not complete; the pooled row is absent for the same reason, since pooling across an incomplete benchmark set would not be the preregistered quantity. The preregistered clustered-bootstrap intervals for these *derived* differences were not produced within the compute available, so the table is point estimates only, and the ordering argued in the main text rests on the contrast in Table 1 rather than on this table.

Benchmark	Δ_{A4} *prior*	$\Delta_{A5} - \Delta_{A1}$ *augmentation, upper bd.*	$\Delta_{A2} - \max(\Delta_{A4}, \Delta_{A5})$ *static visual, lower bd.*
VideoHallucer	+0.6	-0.1	-0.1
EventHallusion	-0.7	+0.5	-1.5
TempCompass-T	+0.6	0.0	+0.3
DyBench	—	-0.1	—

Only TempCompass-T's row is interpretable, for the reason the main text gives: it is the only benchmark whose denominator gate passes. There the augmentation upper bound is zero to the reported precision — order-preserving jitter buys nothing over real video — and the static-visual lower bound is positive and small, while the prior term is a sixth of the full gain. On the three benchmarks whose gate is withheld the rows are printed for completeness and carry no claim.

B.4 Why the missing single cell cannot change the reference arm

The preregistered rule picks, per benchmark, the single blinded arm with the highest pooled blinded accuracy from $\{\text{blind}, \text{single}, \text{frozen}\}$, ties broken toward the most position-structure-preserving arm. frozen is highest in every model $\times$ benchmark cell where it can be compared.

Table 4: Video-LLaVA-7B on VideoHallucer (1,952 items, 976 pairs) under two decision rules. Every pair is a basic question answered yes plus a hallucinated question answered no, and pair accuracy requires both. Per-item agreement between the rules is 80.8% (375 of 1,952 items differ).

Decision rule	yes-rate	basic acc.	hallucinated acc.	Pair acc.
Generate and parse	0.925	0.967	0.112	**9.5**
Per-option log-likelihood	0.738	0.834	0.359	**24.5**
Published figure	—	—	—	17.8

Qwen2.5-VL cannot score single at all: a one-frame clip violates its preprocessor's two-frame minimum. That gap cannot change the selection, for a reason independent of the missing measurement. single and frozen carry the same visual information — the median frame and nothing else — while frozen additionally preserves the frame count and duration the model was trained on. Video-LLaVA puts that out-of-distribution penalty at 10.1 points (24.1 vs. 14.0), so a single score cannot exceed its frozen counterpart, and Qwen's frozen already leads blind by 29.8. Routing one-frame clips through the image path instead would have scored them under a different tokenisation from Video-LLaVA's, which is worse than an honest gap.

B.5 The decision-rule diagnostic

B.6 The full intervention profile

Table 5: Measured intervention profile of two base models. full: unmodified clip; blind: every frame black; single: the median frame alone; frozen: the median frame repeated to the original length. Metrics are pair accuracy (VideoHallucer) and item accuracy (EventHallusion).

Benchmark	Model	full	blind	single	frozen	Δ vs. blind	Δ vs. frozen
VideoHallucer	Qwen2.5-VL-7B	51.4	8.3	n/a	38.1	43.1	**13.3**
	Video-LLaVA-7B	24.5	10.2	14.0	24.1	14.2	**0.4**
EventHallusion	Qwen2.5-VL-7B	67.2	65.3	n/a	65.8	2.0	1.5
	Video-LLaVA-7B	42.5	38.1	28.6	39.9	4.4	2.7

C Limitations in detail

C.1 Why the reproduction criterion has no evaluable precondition

The preregistered gate required reproducing published clean scores within ± 2 points. Its precondition is a published value measured on the same data under a stated decision rule. Neither half holds on VideoHallucer.

Holding model, documents, visual preprocessing and prompt fixed and changing only whether the answer is generated and parsed or chosen by per-option log-likelihood moves Video-LLaVA-7B's (Lin et al., 2024) pair accuracy from 9.5 to 24.5 — 15 points, an order of magnitude more than the tolerance — with the published 17.8 falling between the two rules, and the benchmark does not state which rule it used. The mechanism is measured rather than assumed, and parse failure is excluded as an explanation: only 11 of $1{,}952$ generations were unparseable (0.56%), and a strict and a lenient parser agreed on every probe row (§4.4). Separately, duplicate Temporal questions were removed from the corpus after the published figures were produced, so the two are not measured on the same data.

The unevaluable verdict is therefore a property of the criterion's preconditions on this benchmark, not a finding about our pipeline, and nothing on our side resolves it without the reference implementation. Recording it as *not evaluable* rather than selecting whichever rule lands within tolerance is the point: the alternative is the exact degree of freedom this paper is about.

C.2 Residual circularity in the stratum

The MNS is not curated: it is the output of a procedure applied to held-out model families, with twenty Bernoulli observations per item and condition, a paired permutation null, BH-FDR correction (Benjamini & Hochberg, 1995), and selection on one prompt-replicate set with estimation on the other.

Two problems survive that design. Per-item power at twenty observations is structurally limited, so the stratum carries error and no conclusion is drawn about any individual item. And the definition stays circular in a way the design shrinks but cannot remove — what the procedure certifies is that held-out contemporary models fail on an item without motion, which is not the same proposition as the item intrinsically requiring motion. Two external anchors bound the divergence without closing it: the synthetic calibration set, where motion-necessity is known by construction, and DyBench's counterfactual pairs (Du et al., 2026).

C.3 The single-frame gap

Qwen2.5-VL cannot score single-frame clips: its utility layer requires at least two frames, and a one-frame video raises rather than being padded. We record the gap rather than working around it, since routing single frames through the image path would tokenise them differently from the video path used in every other cell. The blinded-reference selection does not depend on the missing measurement, for the reason given in Appendix B.4.

C.4 Realised versus preregistered subject list

The preregistered subjects were TCD, STEAR and VistaDPO, with CRPO as a conditional fourth. The 2026-07 availability audit inverted the last two: VistaDPO's repository carries open TODOs for both its training data and its weights, a checked-in subset described as reference data, and an empty linked dataset, while the CRPO checkpoints were found released. The realised count is therefore three rather than four. The conditional matched standard/shuffled retraining pair on VistaDPO's recipe was not run either: a half-reproduction on incomplete data is the apples-to-oranges comparison the preregistration exists to exclude.

C.5 Reversal coverage

The reversed arm covers only the label-remappable whitelist, not every item. Speed questions are excluded because time reversal does not change their answer, and question-type by intervention combinations that are not semantically determinate are excluded rather than guessed.

D Experimental setup in full

D.1 Recipe provenance, value by value

Why the sequence limit was recomputed rather than inherited. At our visual budget Qwen2.5-VL emits roughly 3072 video tokens for a 32-frame clip ($602{,}112$ pixels per frame after the processor's clamp, divided by 28^2 per patch, by 4 for the 2×2 spatial merge, and by 2 for the temporal merge). Upstream's 2048-token cap therefore leaves negative room for the prompt, and every step would raise rather than train. The visual budget is what must not move, since it is matched to the evaluation side, so the text cap is what gets recomputed. The longest prompt we measured in the training set is 4610 tokens; the cap is set at 8192 so that no pair is truncated. This matters more than it appears: a length filter would have been the obvious alternative, and it would have been fatal to the design, because A4 carries no visual tokens and so would have passed a filter that dropped pairs in every other arm — silently unmatching the arms along the one axis the paper rests on.

Why the batch factorisation changes while the effective batch does not. Upstream's $4 \times 4 \times 8$ assumed 2048-token sequences; ours run to several thousand, and a DPO step performs four forward passes (policy and reference, chosen and rejected). We therefore use $1 \times 16 \times 8$, which holds the effective batch at upstream's 128 while fitting activation memory. Under eight-way data parallelism

Table 6: The LoRA-DPO recipe by provenance. UPSTREAM is verbatim from the published LLaVA-Hound-DPO script (Zhang et al., 2024); RECOMPUTED values are those whose upstream counterpart does not transfer; OURS have no upstream counterpart, since upstream trains all parameters on Video-LLaVA whereas we train LoRA adapters (Hu et al., 2021) on Qwen2.5-VL. The split is what licenses A1 as a representative reproduction: the hyperparameters defining the method are upstream's, and the values we chose are those forced by the change of base model and parameterisation. Appendix D.1 derives each RECOMPUTED entry.

Parameter	Value	Provenance
β / epochs / effective batch	0.1 / 3 / 128	UPSTREAM
Warmup / schedule / weight decay	0.1 / linear / 0	UPSTREAM
Grad. checkpointing, bf16	on	UPSTREAM
Max sequence length	8192	RECOMPUTED; upstream's 2048 is Video-LLaVA's and leaves negative room for the prompt here
Per-device $\times$ accum. $\times$ GPUs	$1 \times 16 \times 8$	RECOMPUTED to hold the effective batch at upstream's 128
LoRA r / α / dropout	16 / 32 / 0.05	OURS
LoRA scope	language tower only	OURS; the projection names also occur in the vision tower, so the scope is pinned by a path pattern, not a name list (Appendix D.1)
Learning rate	$1e{-5}$	OURS; upstream's $5e{-7}$ is a full-parameter rate. Fixed on the A1 pilot, then frozen
Frames per clip	32	OURS; matched to the evaluation side

each rank takes a strided slice of the flat order, so the group consumes the same pairs in the same order as a single process would.

Why the LoRA scope is a path pattern rather than a name list. The recipe targets the language tower's linear projections (`q,k,v,o,gate,up,down`) and places no adapter on the vision tower, matching upstream, which freezes its multimodal projector. Those projection names, however, also occur inside the vision tower, and the adapter library matches module paths by name, so supplying the list alone attaches adapters to both towers — 47.6M trainable parameters where the language tower alone is 40.4M. That is not merely a departure from the recipe. A4 sends no visual tokens, so its vision-tower adapters would never receive a gradient, and A4 would train a strictly smaller parameter set than every other arm while every loss curve looked normal. The scope is therefore pinned as a path pattern that excludes the vision tower, the resolved pattern enters `recipe_digest` (the name list alone never said which tower), and the trainer refuses to start if any adapted module path names the vision tower.

Learning rate. Upstream's $5e{-7}$ is a full-parameter rate and does not transfer to LoRA. Three pilot rates were run on a quarter of the data; mean loss over the final 50 steps decreased monotonically with rate ($5e{-6}$: 0.642; $1e{-5}$: 0.610; $2e{-5}$: 0.566). Lower DPO training loss is not better past a point — it reflects harder separation of the preference pair, and the fastest rate reached a minimum of 0.052 on a quarter of the data, which we read as over-separation rather than as better learning. We fixed $1e{-5}$ and froze it. The pre-registered criterion for the rate is not training loss but held-out clean gain within the published band; that criterion is evaluated on the A1 run itself and reported in §4.1 whether or not it passes.

D.2 THE MATCHING CHECKS, AND THE DEFECTS THEY CAUGHT

The pre-registered (arm, seed) composition is `a1:3, a2:3, a3:1, a4:2, a5:1` on Qwen2.5-VL and `a1:1, a2:1` on InternVL3-8B as its own group. The verification pass checks: a single `recipe_digest` across every run entering a comparison; the realised (arm, seed) multiset; the digest's presence in every segment of a resumed run; identical data-order hashes across arms at the same seed; and identical text hashes across all arms. The digest covers the SHA256 of every

source file that can change what training does, which is a wider set than it first appears — the loss, the optimiser, the learning- rate schedule, the loss mask and truncation rule, and the frame planner all live in different files, and an earlier version of the digest covered only three of six.

The checks are reported here because they were not decorative. Each of the following was caught by one of them or by the invariant that a flag exists to protect, and none would have appeared in a loss curve or an accuracy table:

- **Adapters on the vision tower.** Detailed above. Surfaced only because the distributed reducer objected to A4's unused parameters; the loss curves of all ten runs were normal.
- **A data cursor counted per rank.** Under eight-way data parallelism the trainer advanced its cursor once per locally consumed pair while the resume path read that number as a position in the global order, so a segment that had collectively reached position 8960 recorded 1120 and the next segment re-consumed seven eighths of what it had just seen. The stated three epochs would not have been three epochs. The consumption-plan check caught it.
- **A learning rate restored as zero.** Constructing the schedule after loading optimiser state clobbers the restored rate to zero, because the scheduler's constructor writes into the optimiser's parameter groups while its `load_state_dict` does not. The first update of every resumed segment would have moved nothing, while the logged rate stayed correct. The number of lost updates equals the number of resumes, which is set by preemption and therefore *differs per arm* — an uncontrolled segmentation difference of exactly the kind the design forbids.
- **Segment-dependent hashes.** Rolling hashes cover only what one segment consumed, so two arms reaching the same cursor through different segment boundaries — routine under preemption — produced different digests and the check reported a mismatch that was not real. The invariant is now a prefix hash over the plan, which is segmentation-invariant.

D.3 Evaluation arms: the whitelist and the exclusions

Reversal is restricted by a frozen whitelist: direction items whose gold option has exactly one antonym among the options, and two-option order items. Labels are remapped by rule. Speed items are excluded, since time reversal does not change a speed answer. Every other question-type $\times$ intervention combination that is not semantically determinate is excluded rather than guessed, and coverage is reported per benchmark.

EventHallusion ships no native pairs. Its polarity pairs would come from a deterministic within-video zip of gold-yes with gold-no questions in file order, leftovers scoring as singletons; on the staged release its questions are near-uniformly negative, so the zip yields no pair at all and the benchmark is scored by item accuracy. A pre-registered switch to template re-asking fires if the bias-index zero point measured on the synthetic calibration set (§D.5) deviates from zero by more than 2.0 points.

One gap is recorded rather than worked around: Qwen2.5-VL cannot score `single_frame`, whose clips hold one frame while its video utility requires two, and the image path would have scored them under a tokenisation incomparable with the other model's.

D.4 The motion-necessary stratum in full

The blinded reference. The MNS predicate compares full-video accuracy against a *single* blinded reference arm, not a per-item maximum, which would bias the contrast downward. The reference is chosen per benchmark on the pilot as the blinded arm with the highest pooled blinded accuracy — that is, the hardest bar for calling an item motion-necessary. That rule selects `frozen_frame`, which leads in every model $\times$ benchmark cell where it can be compared, and is then frozen for the study. How much the choice changes the measured video contribution is reported in §4.4.

Stability and filtering. We report leave-one-family-out Jaccard stability, with LLaVA-Video and LLaVA-OneVision counted as one family. A leave-family-out filter drops ill-posed items — those on which every held-out family is at chance even with full video; it excludes, never relabels, and

results are reported with it on and off. A Qwen-included definition is reported as a sensitivity version beside the held-out headline.

D.5 SYNTHETIC CALIBRATION

No part of the validation stack uses human annotation. Its anchor is 800 procedurally rendered clips (16 frames at 8 fps, 192×96 canvas) in which motion-necessity is known *by construction*: at most three uniquely coloured shapes with explicit appearance and disappearance times and velocities, under constraints that keep every question well-posed on the untouched clip (a minimum visible span, a minimum displacement below which two frames look identical, a factor-of-1.5 margin before one object counts as faster, and a per-frame displacement cap so an object never degenerates into unrelated flashes). Questions span seven information kinds: static appearance, object count, presence, motion direction, temporal order, motion speed, and event-anchored temporal existence.

The oracle is *derived, not tabulated*. Each question records the source frames and orderings its answer depends on, and answerability under an arm follows from intersecting that record with the arm's frame-plan structure, so no arm is ever named in the oracle. A per-arm lookup table could not be caught being wrong, because the table would *be* the expectation; derivation instead yields consequences a table would miss, such as a gold-no existence question requiring full frame coverage — an absence cannot be certified from a subset — so frozen-frame destroys it although nothing about it is temporal.

The corpus supplies four otherwise unmeasurable quantities: the intervention–oracle answerability matrix; the out-of-distribution artifact floor, estimated *per question type* on items the oracle marks unanswerable under both compared arms and subtracted from the shuffled-arm conclusions (stratified rather than global whenever the two floors differ by more than 2 points); the zero point of the bipolar bias index; and a constructive precision/recall anchor for τ, with DyBench's paired counterfactuals as the second, non-synthetic anchor.

D.6 COMPUTE

The campaign runs as containerised cluster array jobs in four stages: CPU-only materialisation of intervention variants; single-GPU sharded inference emitting per-shard JSON lines; multi-GPU DPO training; CPU analysis. Segmentation is validated by running one configuration continuously and in two halves and requiring identical data-order and text hashes; loss trajectories agree to tolerance but not bitwise across a restart, which we state rather than paper over.

Inference is resumable for a measured reason: Qwen2.5-VL's dynamic resolution makes peak host memory vary widely within one benchmark (per-video pixel limits between 235k and 602k were observed), and one blind-arm run was killed at 1520 of 1952 items under a 64 GB allocation before completing under 192 GB. Resuming appends, skipping items already present, and refuses to append to an output written under a different model, evaluation arm, or adapter — with thirteen adapters across several arms, merging two conditions into one file is a likely mistake rather than an exotic one, and it would leave no seam anywhere in the file.

D.7 AUDITED SUBJECTS

The scope is deliberately narrow: our matched-recipe reproduction A1, plus methods whose artifacts a pre-registered availability check found public as of 2026-07. Those are temporal contrastive decoding (Zhang et al., 2025b), transcribed from its paper because the released repository ships the benchmark and the evaluator but no implementation of the method — a fact stamped into every scored row's provenance; STEAR (Fan et al., 2026), run only under its released layer-intervention implementation; and the CRPO checkpoints (Du et al., 2026), evaluated against their own base and never across bases.

The pre-registered list also named VistaDPO (Huang et al., 2025) as a released-checkpoint subject. The availability check found neither its weights nor its complete training data released — both are open items in its repository and its linked dataset is empty — so it is audited nowhere in this paper, and the conditional matched standard/shuffled retraining pair on its recipe was not run either: a half-reproduction on incomplete data is the apples-to-oranges comparison the pre-registration exists

to exclude. The same check found the CRPO checkpoints released, so CRPO takes the released-checkpoint seat. We nowhere claim to audit published mitigators as a population.