ACCURACY IS NOT EVIDENCE: MEASURING AND TRAINING EVIDENCE DEPENDENCE IN VIDEO TEMPORAL GROUNDING

Anonymous authors
Paper under double-blind review

ABSTRACT

In video temporal grounding, a correct interval can come from finding the event, answering blind or fitting one annotation style, and Recall at IoU scores them alike. We measure evidence dependence, how much the answer rests on the event's frames. Our fixed-answer audit edits either the frames inside the annotated interval or, as a placebo, as many frames elsewhere in the same video, and compares the two drops in the annotated answer's likelihood. Supervised fine-tuning and three public checkpoints trained with IoU-reward reinforcement learning raise the accuracy of Qwen2.5-VL-7B by 9 to 20 points, partly by answering blind and by fitting one annotation. Yet their answers rest on the event no more than the base model's. This comparison is a difference of the model's own likelihoods, so we can train on it. A placebo contrast and a bound on the placebo drop rule out learning to react to any edit. With them, evidence dependence rises in all five base models we train, and the weight on clean supervision controls how much. Masking the event's frames makes the trained models decline or miss the event more often than their base; masking as many other frames does not. The dependence moves to another event when the query names it, and holds on a second dataset and under an unseen edit.

1 INTRODUCTION

Video temporal grounding asks a model when a described event happens. Given a video and the query "another person is fixing a doorknob", the model returns an interval. The benchmark counts it correct when its IoU with the annotated interval is at least 0.5 (Recall at IoU 0.5, R@0.5; Gao et al., 2017; Lei et al., 2021). Post-training raises this accuracy by many points, but a model can earn a correct interval in three ways. It can find the event. It can answer blind: with every frame replaced by black, Qwen2.5-VL-7B still returns an interval for 12% of Charades-STA test questions; fine-tuned on 2,400 annotated timestamps, it does so for 78%. And it can fit one annotation style: a fixed window over the first ten seconds reaches two thirds of the base model's accuracy, because benchmarks carry shortcuts (Geirhos et al., 2020) in the length and position of their intervals (Otani et al., 2020; Yuan et al., 2021).

The quantity we want is *evidence dependence*: how much a model's confidence in its answer rests on the frames that contain the event. Finding the event means its frames are enough to place the answer; evidence dependence means they are needed. A model can have the first without the second. Current audits alter the video and regenerate the answer (Otani et al., 2020; Cores et al., 2025; Lee, 2026; Ok & Lee, 2025; AlHamidi et al., 2026). When the regenerated interval changes, the score cannot tell lost evidence from a different guess. A score of how much worse the model does without the evidence also rewards one that does worse whenever anything is removed, a known confound in attribution (Hooker et al., 2019).

Our fixed-answer audit (Section 2; Figure 1, left) scores the annotated interval by teacher forcing, so every readout is the likelihood of the same answer. For the doorknob query, we black out the doorknob-fixing frames or, as a placebo, as many frames elsewhere in the same video. The difference between the two drops in likelihood is zero for a model that reacts equally to every edit, however much each edit hurts it. Our measure, Γ, is how much training changes this difference from the base

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.

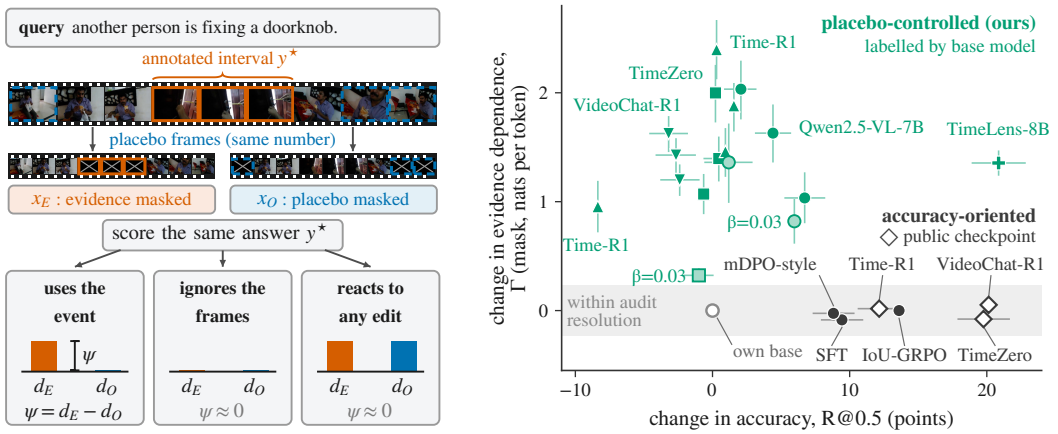

Figure 1: **Accuracy and evidence dependence are separate axes.** Left: the audit scores the same correct answer, the annotated interval $y^\star$, with the event's frames masked (x_E) and with as many frames masked elsewhere, the placebo (x_O). The schematic bars show the two drops in the answer's log-likelihood, d_E and d_O, and their difference ψ for three kinds of model. Right: each point places one post-trained model by its change from its own base (the origin) in R@0.5 (horizontal) and in evidence dependence Γ, the change in ψ (vertical); lines are 95% intervals. The grey band is the audit's resolution, the farthest from zero that the 95% interval of Γ reaches between two seeds of one recipe (Section 2). Accuracy-oriented training (dark grey) moves models right but not out of the band; SFT, IoU-GRPO and mDPO-style are trained from Qwen2.5-VL-7B, and the public checkpoints (open diamonds) are compared with it. The placebo-controlled objective (green) moves them up. Each base model has its own marker shape, its several points differ in α or β, and light-filled points use the smaller margin scale β=0.03 (Section 4).

model, on the questions both models answer correctly. We report it next to the trained model's own placebo drop. Pairs of independently trained seeds of the same recipe, audited against each other, show how far from zero Γ can reach by chance alone. This sets the audit's *resolution*, and only a larger Γ counts as an effect.

The post-training for accuracy that we audit takes all three routes and does not raise evidence dependence (Section 3; Figure 1, right). We audit Qwen2.5-VL-7B after supervised fine-tuning (SFT) and IoU-reward GRPO (IoU-GRPO), and three public checkpoints trained from it by reinforcement learning (TimeZero, Time-R1 and VideoChat-R1). All are more accurate than the base. The public checkpoints' predictions follow the event more often when its frames are moved, so part of their gain comes from finding it. SFT, TimeZero and Time-R1 answer most all-black videos, and against a second annotation the public checkpoints lose their lead over the base. Yet none raises Γ beyond the audit's resolution. For SFT, IoU-GRPO, TimeZero and Time-R1, even the answers they newly get right rest on the event no more than the base's own correct answers. This training can make the event's frames enough to place the answer more often, but it does not make them more needed.

Because the audit's difference comes from the model's own log-likelihoods, we can train on it (Section 4), and the placebo is needed there too. Rewarding the evidence drop alone teaches the model to react to any edit: its own placebo drop rises far above the base's. A contrast against the matched placebo and a bound on the placebo drop remove this solution. We call the result the *placebo-controlled objective*. With it, Γ rises in all five base models we train. The weight α on clean supervision controls the amount: in Qwen2.5-VL-7B, raising it from 0.25 to 1 halves Γ and triples the accuracy gain. Evidence dependence thus becomes a setting of training rather than a by-product.

We then test the trained models on quantities outside the loss (Section 5). When the event's frames are masked, the placebo-controlled Qwen2.5-VL-7B model returns no interval 21.5 points more often than its base, on questions it answers correctly under the placebo mask. Unlike SFT, the placebo-controlled Qwen2.5-VL-7B and Time-R1 models decline queries about another video at least as selectively as their bases. In most trained models the dependence also moves to another event when the query names it, and it holds on QVHighlights and under an edit never seen in training.

Contributions. (i) We introduce the fixed-answer audit, which measures evidence dependence against a matched placebo in the same video (Section 2). (ii) We show that the accuracy-oriented

post-training we audit takes all three routes without raising evidence dependence (Section 3). (iii) We introduce the placebo-controlled objective, which raises evidence dependence in all five base models we train, with the weight on clean supervision setting how much (Section 4). (iv) We show that the trained models decline or miss the event more often than their base when its frames are masked, and that their dependence moves with the query (Section 5).

2 MEASURING EVIDENCE DEPENDENCE

A grounding benchmark checks whether a prediction agrees with the annotation. The fixed-answer audit asks instead how much of the model's confidence in a correct answer disappears when the event's frames are edited, compared with editing as many other frames. For the query "another person is fixing a doorknob", we score the annotated interval on the original video, with the doorknob-fixing frames blacked out, and with as many frames blacked out elsewhere in the same video. A model that uses the event loses confidence under the first edit only; one that reacts to any edit loses it under both (Figure 1, left).

The evidence drop, the placebo drop and Γ. For a question with video x, query q and annotated interval $y^\star$, let $\ell(x) = \log p(y^\star \mid x, q)$ be the model's log-likelihood of the annotated answer, in its template's format and averaged over the answer tokens. Editing the frames inside $y^\star$, the *evidence frames*, gives x_E. Editing the same number of frames drawn at random outside $y^\star$, with the same operator, gives the *placebo* x_O. The evidence drop, the placebo drop and their difference are

$$d_E = \ell(x) - \ell(x_E), \qquad d_O = \ell(x) - \ell(x_O), \qquad \psi = d_E - d_O, \tag{1}$$

in nats per token, with the placebo drop averaged over three draws. The difference ψ says how much more the evidence edit hurts than the placebo edit. We measure the effect of training as the change in ψ from the base model,

$$\Gamma \;=\; \frac{1}{|H_\cap|} \sum_{i \in H_\cap} \big(\psi_i^{\text{trained}} - \psi_i^{\text{base}}\big), \tag{2}$$

where $H_\cap$ holds the questions that both models localise correctly (IoU ≥ 0.5 on the clean video). Γ thus compares two ways of being right on the same questions.

How large a Γ counts. Before any training, the evidence edit already costs a little more than the placebo edit: on the questions it answers correctly, the Qwen2.5-VL-7B base has ψ between $+0.18$ and $+0.23$ nats per token. A Γ of zero means that training left this unchanged. Two seeds of one recipe differ only by chance, so we audit two pairs of seeds of the placebo-controlled Qwen model (Section 4) against each other. Their 95% intervals of Γ reach at most 0.23 nats per token from zero under mask and 0.31 under swap. We call each bound the audit's *resolution* and read a Γ as an effect only when its estimate lies beyond it. For scale, the placebo-controlled models of Section 4 reach a Γ of about $+1$ to $+2$. The likelihood audit uses the first 200 test questions with three placebo draws each. Accuracy (R@0.5) is measured on the full 3,720-question Charades-STA test split, and all intervals are bootstrapped over videos (Appendix A.1).

What is edited and what is scored. We use two edit operators. *Mask* sets frames to black; *swap* replaces them with frames from another video. Mask is the primary operator: Γ without a label refers to it, and under swap we write Γ (swap). The audit teacher-forces each model's answer string directly after the question, without the reasoning trace that TimeZero and Time-R1 write, so the scored answer is the same under every edit; generated predictions include the trace. Each model and its base share one prompt template and sample frames at 2 per second.

Three signatures of evidence use. Each signature rules out another cause of a drop under the evidence edit. First, the trained model's own placebo drop must stay near its base's. A model whose placebo drop rises reacts to any edit, and training on the evidence drop alone produces one (Section 4); we therefore report every trained model's Γ with its own placebo drop. Second, the dependence should move with the query: when the query names another event in the same video, it should move to that event's frames, which Section 5 tests. Third, the rise in ψ should not come from a sharper likelihood on the clean video. SFT sharpens it as much as the placebo-controlled models do, yet its Γ stays at or below zero, so sharpening alone does not raise Γ.

Table 1: **Training for accuracy adds correct answers but not evidence dependence; the placebo-controlled objective adds both.** All models start from Qwen2.5-VL-7B-Instruct. The base row gives levels. Every other entry is a change from the base on the same questions under the same template, except d_O, which is each model's own placebo drop. Accuracy changes are in points, and brackets give 95% intervals. Bold marks a change in points whose interval excludes zero or a Γ beyond the audit's resolution; grey marks a Γ within it.

	Accuracy	Evidence dependence (nats per token)			Regenerated answers (points)	
Model	R@0.5	Γ mask	Γ swap	d_O	evidence share	follows moved event
Qwen2.5-VL-7B (base)	0.480	—	—	−0.06	69%	—
Trained for accuracy						
SFT	**+9.4** [7.9, 11.0]	−0.09 [−0.15, −0.02]	−0.03	0.00	**−14.7** [−19.8, −9.4]	+7.0 [−2.0, 16.5]
IoU-GRPO	**+13.6** [11.9, 15.4]	0.00 [−0.03, 0.04]	+0.02	−0.05	**−6.1** [−11.0, −1.2]	+8.7 [−1.1, 16.8]
mDPO-style	**+8.8** [7.3, 10.4]	−0.03 [−0.09, 0.04]	+0.01	0.00	+1.1 [−3.3, 5.7]	+5.1 [−4.5, 14.1]
TimeZero (public)	**+19.8** [17.9, 21.7]	−0.08 [−0.16, 0.01]	−0.05	−0.03	+1.3 [−4.2, 6.9]	**+16.2** [6.4, 26.9]
Time-R1 (public)	**+12.2** [10.6, 13.8]	+0.02 [−0.02, 0.06]	+0.05	−0.07	−4.4 [−9.5, 0.8]	**+15.3** [4.6, 26.7]
VideoChat-R1 (public)	**+20.1** [18.3, 22.1]	+0.05 [0.02, 0.09]	+0.03	0.00	+8.2 [−0.2, 17.7]	**+20.8** [10.8, 31.5]
Placebo-controlled objective (ours)						
α=0.5	**+4.4** [3.1, 5.7]	**+1.63** [1.36, 1.89]	**+1.41**	0.01	**+15.7** [11.4, 20.3]	+6.1 [−4.3, 16.1]

Two readouts on generated answers. The likelihood audit scores a fixed answer, so we also check what the models generate. We regenerate each model's predictions on the first 1,000 test questions with the evidence frames masked and with as many placebo frames masked (accuracies a_E and a_O). The *evidence share* $1 - a_E/a_O$, the counterpart of ψ for generated answers, is the share of the placebo-masked accuracy lost to the evidence mask. For *follows*, we exchange the event's frames with an equally long block elsewhere in the video and regenerate on $H_\cap$; a prediction follows if its IoU with the annotated interval moved to that block is at least 0.5. Follows tests whether the event's frames are enough to place the answer; the evidence share tests whether they are needed.

3 What post-training accuracy is made of

Every accuracy-oriented model we audit on Qwen2.5-VL-7B-Instruct (Bai et al., 2025b) is 9 to 20 points more accurate than this base, and none raises Γ beyond the audit's resolution (Table 1). Three of them are public checkpoints trained for this task with IoU-reward reinforcement learning: TimeZero and Time-R1, two releases of one method (Wang et al., 2025), and VideoChat-R1 (Li et al., 2025). We train SFT and an mDPO-style preference for clean over corrupted video (Wang et al., 2024) on the same 2,400 Charades-STA pairs as our objective (Section 4), and IoU-GRPO (Shao et al., 2024) on the full Charades-STA training split at 16 frames per video. Because a model learns dependence only on the frames it trains on, and our objective trained on 16 frames raises Γ far less than at 2 per second, the public checkpoints, not our 16-frame IoU-GRPO, show what IoU-reward training does to Γ. Appendix B gives the per-model tables.

The public checkpoints end up where our SFT model does. Other groups trained them on their own data and rewards, yet in Γ they sit next to SFT (Figure 1, right). Our objective, trained on the same base and pairs as SFT, raises Γ instead. On a second backbone, the accuracy-trained TimeLens-8B has $\Gamma = -0.16$ against Qwen3-VL-8B, within the resolution: its evidence drop falls, while its placebo drop stays where it was.

Finding the event. Part of the added accuracy uses the video. When the event's frames are moved, the public checkpoints' predictions follow them 15 to 21 points more often than the base's; for our

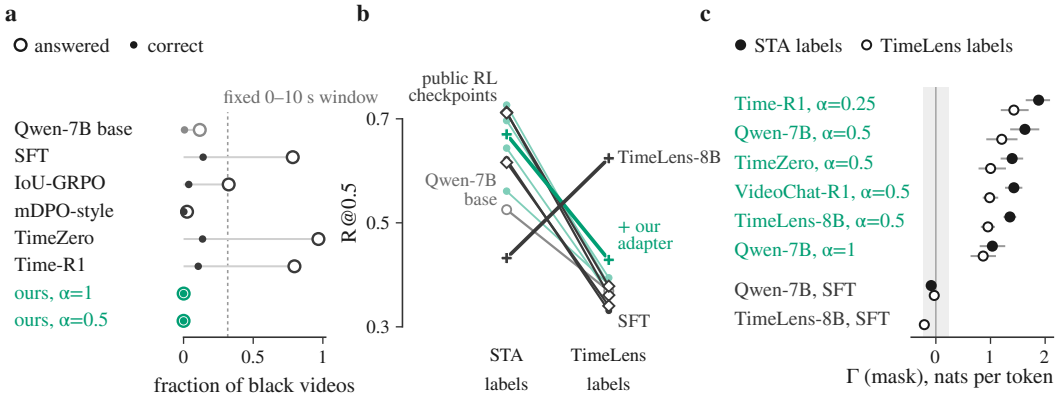

Figure 2: **Part of the added accuracy comes from answering blind or fitting one annotation; Γ keeps its sign under both label sets.** (a) shows, for all-black videos under Time-R1's prompt template, the fraction of questions still given an interval (open circles) and answered correctly (dots); the dashed line is a fixed $(0, 10)$ s window on the real videos, and green rows are our placebo-controlled Qwen models at weight α. (b) scores each model on the questions annotated in both Charades-STA and TimeLens against each label set; the two rankings of the nine 7B configurations barely agree (Spearman 0.15). Open diamonds are public checkpoints; open grey circles are the Qwen2.5-VL-7B base (Qwen-7B); thin lines are our 7B models (green) and SFT (dark grey); thick lines are TimeLens-8B (dark) and our adapter of it trained on the original labels (green). (c) gives Γ (mask) with 95% intervals for each label set as the fixed answer (filled: Charades-STA; open: TimeLens), with the audit's resolution as a grey band; the TimeLens-8B SFT row has only the TimeLens value.

SFT, IoU-GRPO and mDPO-style models the interval of the change includes zero (Table 1). The public checkpoints' answers thus follow the event's frames more often without needing them more.

Answering blind. Several accuracy-trained models also answer when there is nothing to see. With every frame black, SFT still returns an interval for most questions (Section 1). Under Time-R1's prompt template, Time-R1 answers 80% of black videos and TimeZero 97%. IoU-GRPO answers 32% of black videos, while the placebo-controlled Qwen models almost never answer one (Figure 2a). Declining black videos is not evidence dependence: the mDPO-style model almost never answers one, and its Γ stays within the resolution.

Fitting one annotation. Part of the public checkpoints' lead is fitted to the conventions of one label set. TimeLens re-annotated Charades-STA and reported that model rankings change (Zhang et al., 2026a); 870 test questions appear verbatim in both label sets, so we score the same predictions against each (Figure 2b). Under the original labels the public checkpoints lead the Qwen base, and against the TimeLens labels they fall to its level. TimeLens-8B, trained on TimeLens-100K, does the reverse: it trails its base under the original labels and leads it under the TimeLens labels. Much of this difference is one convention, the length of the interval: keeping TimeLens-8B's predicted centres but setting every length to 8 s raises its R@0.5 under the original labels from 0.417 to 0.678. Our models' Γ keeps its sign and its separation from SFT under either label set (Figure 2c).

The newly correct answers rest on the event no more than the base's. For SFT, IoU-GRPO, TimeZero and Time-R1, ψ on a model's newly correct questions differs from the base's ψ on its own correct answers by -0.08 to $+0.03$ nats per token, and for the placebo-controlled models by $+0.9$ to $+2.0$. The new answers do use the video: regenerated with the evidence frames masked, 24% of IoU-GRPO's new answers stay correct, against 88% with as many placebo frames masked.

4 Training evidence dependence

The objectives of Section 3 raise accuracy but not Γ. Since ψ is computed from the model's own log-likelihoods, we can write it into the loss.

Objective. Training scores the annotated answer by how much more likely the model being trained finds it than the frozen base model does, $r_\theta(x) = \log p_\theta(y^\star \mid x, q) - \log p_{\text{ref}}(y^\star \mid x, q)$, summed

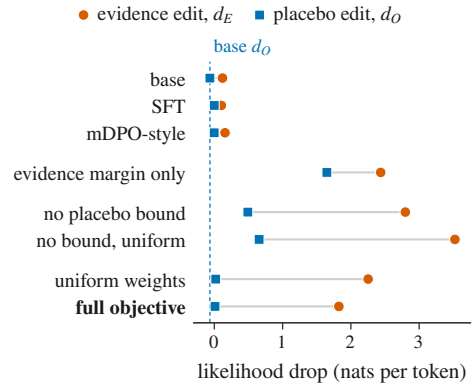

Table 2: **With the evidence margin alone, Γ is a third of what the full objective reaches.** Variants of the full objective (last row of Table 1) are trained from Qwen2.5-VL-7B at α=0.5; *uniform* variants drop the Group-DRO weights. d_O and bold follow Table 1.

	points	nats per token		
Objective	ΔR@0.5	Γ mask	Γ swap	d_O
uniform weights	**+5.4** [4.1, 6.7]	**+1.99** [1.72, 2.24]	**+1.89**	0.02
no placebo bound	**−3.8** [−5.4, −2.1]	**+2.05** [1.75, 2.34]	**+2.34**	0.49
no bound, uniform	**+3.8** [2.4, 5.2]	**+2.62** [2.34, 2.93]	**+2.60**	0.66
evidence margin only	**+6.7** [5.2, 8.3]	**+0.54** [0.34, 0.71]	**+1.09**	1.65

Figure 3: **The placebo contrast, and then the bound, bring the trained model's own placebo drop back to the base's.** The left panel plots, for objectives trained from Qwen2.5-VL-7B on the same pairs and audited on the same questions, the trained model's evidence drop d_E and own placebo drop d_O under masking (nats per token), from the base (top) to the full objective at α=0.5 (bottom); the dashed line is the base's placebo drop.

over the answer tokens. The two margins below sum over tokens, whereas the audit of Section 2 averages over them. As in preference optimisation (Rafailov et al., 2023), measuring against the base means that an edit which already lowers the base model's likelihood earns nothing by itself. The edits x_E and x_O are those of Section 2. Training uses five edit operators: the audit's mask and swap, pixelation, temporal shift and temporal shuffle. For one training pair and one operator, the loss is

$$\mathcal{L}_{\text{op}} = \underbrace{\text{softplus}\big(-\beta\,[\,r_\theta(x) - r_\theta(x_E)\,]\big)}_{\text{evidence margin}} + \underbrace{\text{softplus}\big(-\beta\,[\,r_\theta(x_O) - r_\theta(x_E)\,]\big)}_{\text{placebo contrast}} + \underbrace{\text{relu}\big(d_O^\theta - 1.25\,d_O^{\text{ref}} - 0.10\big)}_{\text{placebo bound}} \tag{3}$$

The evidence margin asks that editing the event's frames lower the likelihood of the annotated answer, so that the answer needs the event. The placebo contrast asks that this edit lower it more than editing as many other frames does, which a model that reacts equally to every edit cannot satisfy. It is the change in ψ from the base, summed rather than averaged over tokens, so the loss acts on Γ directly. The bound compares the audit's token-averaged placebo drops of the trained model, d_O^θ, and of the base, d_O^{ref}. It stays zero while the trained model's placebo drop is at most slightly above the base's, and it keeps the placebo drop down, which the contrast alone does not (Table 2). Its constants come from an earlier experiment in which models trained without it raised their placebo drop far above this limit (Appendix C.2). The full loss adds clean timestamp supervision,

$$\mathcal{L} = \alpha\,\text{NLL}_\theta(y^\star \mid x, q) + \sum_{\text{op}} w_{\text{op}}\,\mathcal{L}_{\text{op}} \tag{4}$$

where α sets how much clean supervision is kept and β sets the scale of both margins (0.1 unless stated). The operator weights w_{op} are Group-DRO weights (Sagawa et al., 2020), which hold the trained model's placebo drop lower than uniform weights do (Table 2). As $\beta \to 0$ without the bound, the loss reduces to timestamp supervision alone, and this is our SFT model. The placebo-controlled models are LoRA adapters (Hu et al., 2022) trained, unless stated, for one pass over the 2,400 pairs.

Without the placebo terms, the model learns to react to any edit. Adding the terms one at a time shows that the optimiser takes the shortcut the placebo was built to exclude until both placebo terms are in the loss (Figure 3, top to bottom). A preference for the clean video over one whose corrupted frames are chosen without the annotation, as in mDPO (Wang et al., 2024), raises accuracy but not Γ (Table 1). Corrupting the annotated frames instead gives the evidence margin alone, the training form of a deletion score. The model then learns to lose likelihood wherever the video is edited.

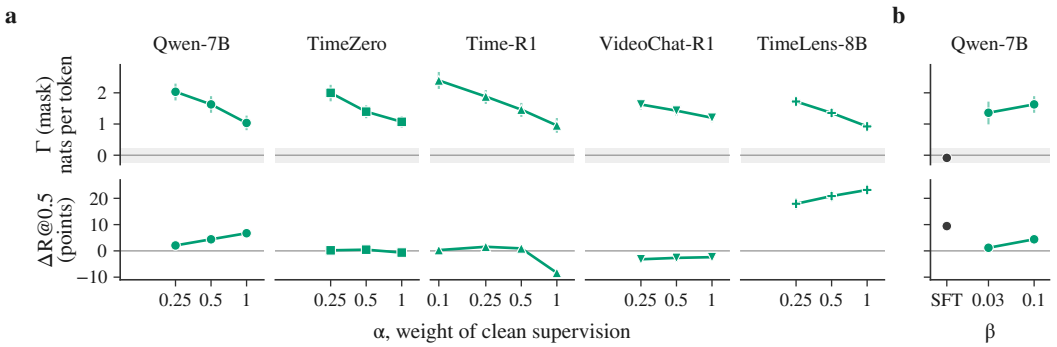

Figure 4: **Evidence dependence falls with the weight on clean supervision; the accuracy change depends on the base.** The top row shows Γ (mask) and the bottom row the change in R@0.5 from the model's own base, with 95% intervals. (a) Each column plots one base against α at β=0.1; TimeLens-8B is built on Qwen3-VL, and the grey band is the audit's resolution. (b) plots Qwen2.5-VL-7B at α=0.5 against β, with SFT at $\beta \to 0$.

Table 3: **The placebo-controlled objective raises evidence dependence in every base model; the change in accuracy varies with the base.** Each base model is paired with its adapter, trained at α=0.5 on the same 2,400 Charades-STA pairs and run under the base's own template. Changes are in points, Γ and d_O are in nats per token, and d_O is given for the adapter and for its base; brackets give 95% intervals.

Base model	base R@0.5	ΔR@0.5	Γ mask	Γ swap	d_O trained / base
Qwen2.5-VL-7B	0.480	+4.4 [3.1, 5.7]	+1.63 [1.36, 1.89]	+1.41	0.01 / −0.06
TimeZero	0.677	+0.5 [−0.7, 1.5]	+1.40 [1.19, 1.60]	+1.31	0.03 / −0.02
Time-R1	0.601	+0.9 [−0.2, 2.0]	+1.46 [1.23, 1.68]	+1.20	0.00 / −0.08
VideoChat-R1	0.668	−2.7 [−4.2, −1.2]	+1.43 [1.27, 1.58]	+0.89	−0.01 / 0.00
TimeLens-8B	0.417	+20.9 [18.9, 22.8]	+1.36 [1.24, 1.47]	+0.88	−0.03 / −0.03

Its placebo drop rises to 1.65 nats per token, two thirds of its evidence drop (Table 2). Adding the placebo contrast raises Γ, but the placebo drop still sits at 0.49, far above the base's. The bound brings it down to 0.01, near the base's level, while the evidence drop remains far above the base's.

The weight on clean supervision controls how much is learned. With both placebo terms, Γ rises in all five base models we train: Qwen2.5-VL-7B, the public TimeZero, Time-R1 and VideoChat-R1, and TimeLens-8B, built on Qwen3-VL (Zhang et al., 2026a; Bai et al., 2025a). At β=0.1 it lies between +0.9 and +2.4, and within every base it falls as α grows (Figure 4). In Qwen2.5-VL-7B, moving α from 0.25 to 1 lowers Γ from +2.03 to +1.03. Over the same range, the accuracy gain rises from 2 to 7 points. The five adapters of Table 3 all keep their own placebo drop within 0.10 nats per token of their base's.

The accuracy change depends on the base; more data raises Γ further. At α=0.5 the accuracy change runs from −3 points on VideoChat-R1 to +21 on TimeLens-8B (Table 3). Checkpoints audited every 400 pairs show Γ still rising at the end of one pass. Doubling the training pairs to 4,800, with everything else fixed, takes the Qwen2.5-VL-7B model to $\Gamma = +2.78$ and a gain of 8.1 points (Appendix C.3).

5 What the trained models do with the evidence

Every test in this section lies outside the loss: it changes the edit operator, the query or the dataset, or it reads the generated answer and the refusal, which the loss never scores. Appendix D gives the per-model tables.

Masking the event costs correct answers; masking elsewhere does not. The placebo-controlled models need the event's frames more than their bases do: every one we regenerate raises its evidence share over its base, by 6 to 29 points (Figure 5c; Table 19). With the evidence frames masked, the R@0.5 of the Qwen model at α=0.5 falls from 0.523 on the clean video to 0.087. With as many frames masked elsewhere, every placebo-controlled model keeps its clean accuracy.

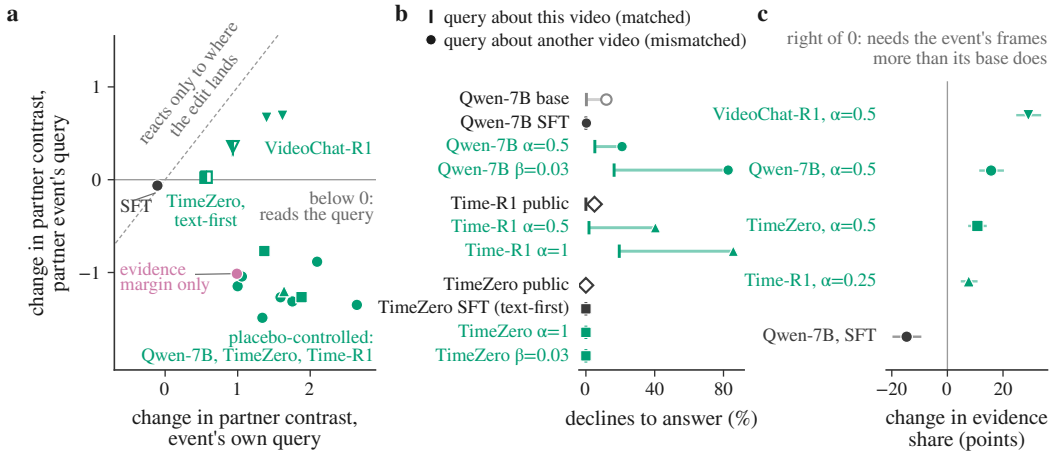

Figure 5: **The placebo-controlled models respond to the query and to the evidence.** (a) Each point gives a model's change from its base (mask, nats per token) in the partner contrast, the drop when the event's frames are masked minus the drop when a partner event's frames are, with the event's own query (horizontal) and with the partner's (vertical). Marker shape gives the base (circle Qwen-7B, square TimeZero, triangle Time-R1, inverted triangle VideoChat-R1); half-filled markers were evaluated text-first, with the query before the video. (b) The tick shows how often a model answers `There are none.` to a query about this video, and the marker how often it does so for one about another video. (c) Each row gives the change from the base in the evidence share, the fraction of the accuracy under the placebo mask that the evidence mask removes, with its 95% interval.

Most of the dependence sits on the framing tokens of tagged answers. Without the event's frames, a model that needs them can decline or place its interval elsewhere. In a tagged answer such as `<answer> 20.80 to 30.00 </answer>`, the two tags and the *onset*, the first token after the opening tag, are the *framing tokens*. The onset commits the model to a timestamp or a refusal. In the tagged templates of Qwen2.5-VL-7B, TimeZero and Time-R1, the placebo-controlled models place most of Γ on the framing tokens. For the Qwen2.5-VL-7B model at α=0.5, Γ is $+0.79$ on the onset and $+0.86$ on the tags, against $+0.07$ on the timestamps. In generation, this shows mostly as refusals for Qwen2.5-VL-7B and as intervals moved off the event for the other two. VideoChat-R1's template asks for a bare interval, so all of its Γ sits on the timestamps; with the event masked, its placebo-controlled model misses it 15 points more often than its base.

The dependence moves with the query. If the dependence belongs to the event the query names, it should move when the query names another event. On videos with two annotated events, A and B, we score A's interval and mask either A's frames or B's. The drop in the answer's log-likelihood under A's mask minus that under B's is the *partner contrast*. We then ask B's query, still scoring A's interval. A model that reads the query is now hurt more by B's edit, so the change in its partner contrast from the base turns negative; one that reacts only to where the edit lands keeps it positive (Figure 5a). The sign reverses in ten of eleven placebo-controlled models from Qwen2.5-VL-7B, TimeZero and Time-R1 tested with both queries. The exception, near zero, is TimeZero evaluated *text-first*, given the query before the video. SFT leaves the contrast unchanged, and in VideoChat-R1 the change under the partner's query shrinks by half to two thirds but keeps its sign.

Placebo-controlled training keeps or widens the base's gap on queries about another video. A model that uses the event should decline a query whose event is not in the video. We pair every question with the annotation sentence of a different video. The Qwen2.5-VL-7B base declines these mismatched queries 12 points more often than matched ones, and its placebo-controlled model at α=0.5 keeps the gap, at 16 points (Figure 5b). On Time-R1 the gap grows from 5 points for the public checkpoint to 39 at α=0.5. Raising α to 1 on Time-R1 or lowering β to 0.03 on Qwen2.5-VL-7B widens it further, while SFT loses the selectivity of its base and declines almost no query.

Unseen edits and a second dataset. Noise, Gaussian blur and patch scrambling never appear in training. Under patch scrambling, three placebo-controlled Qwen2.5-VL-7B and two TimeZero models reach Γ of $+1.12$ to $+1.79$, well beyond the resolution. Under noise and blur their Γ stays positive but mostly within the mask resolution, and every placebo drop stays near zero. On QVHigh-

lights, scored against the TimeLens re-annotation (Zhang et al., 2026a), every placebo-controlled model we evaluate reaches Γ of $+0.92$ to $+1.58$, which still falls with α, and SFT stays at zero.

6 Related work

Grounding models and what their benchmarks reward. The task is defined by Charades-STA (Gao et al., 2017; Sigurdsson et al., 2016), DiDeMo (Hendricks et al., 2017), ActivityNet Captions (Krishna et al., 2017) and QVHighlights (Lei et al., 2021). Specialised architectures (Zhang et al., 2020a;b; Lin et al., 2023) gave way to video LLMs (Lin et al., 2024; Ren et al., 2024; Huang et al., 2024; Bai et al., 2025b; Yan et al., 2025) and reinforcement learning (Shao et al., 2024; Guo et al., 2025) with IoU rewards (Wang et al., 2025; Li et al., 2025; Zhang et al., 2026a). Blind baselines and probes show benchmarks reward priors (Otani et al., 2020; Yuan et al., 2021; Cores et al., 2025; Lei et al., 2023; Xiao et al., 2021; 2024), and debiasing methods reduce them (Lan et al., 2023; Yang et al., 2021; Nan et al., 2021; Jiang et al., 2024). Both work at the benchmark level; we measure how much one model's correct answers rest on the event, and how training changes that.

Audits of evidence use. Deleting inputs is the basic move of perturbation attribution (Fong & Vedaldi, 2017; Petsiuk et al., 2018) and of rationale benchmarks, whose comprehensiveness score is a likelihood drop for a fixed answer (DeYoung et al., 2020). Faithfulness there means reliance rather than plausibility (Jacovi & Goldberg, 2020; Jain et al., 2020). Deletion without a matched control confounds such measurements (Hooker et al., 2019). Video audits that compare answers with and without visual evidence include black-screen gaps (Lee, 2026), frame substitution (Ok & Lee, 2025), dependency filters (Lim et al., 2026), instance-level shortcut scores (Korkut et al., 2026), grounding stress tests (Sethuraman T V et al., 2026) and replay probes (AlHamidi et al., 2026). In grounding, consistency probes move the moment or rephrase the query (Jung et al., 2025). These audits ask whether a regenerated answer changes; the fixed-answer audit asks how much the likelihood of one fixed answer rests on the evidence frames. The one matched control among them, the sham replay of AlHamidi et al. (2026), varies whether an edit happens, whereas our placebo varies the position of an edit that always happens.

Training on evidence. Evidence enters training through counterfactual, attribution and grounding rewards and data (Du et al., 2026; Feng et al., 2025; Wang et al., 2026; Bao et al., 2026; Luo et al., 2025a; Zheng et al., 2026b; Zhang et al., 2026b), explanation constraints (Ross et al., 2017; Selvaraju et al., 2019), and preference or decoding methods built on corrupted visual input (Leng et al., 2024; Favero et al., 2024; Xie et al., 2024; Wang et al., 2024; Wu et al., 2025; Luo et al., 2025b). Two methods add a control, random masks off the evidence (Xu et al., 2026) or non-influential edits in an LLM faithfulness reward (Cheah et al., 2026). The closest precedents put a likelihood drop inside a GRPO reward. TaRO (Zheng et al., 2026a) measures it for sampled reasoning when the frames around the annotated boundaries are shuffled, and Evidence-RL (Huang et al., 2026) for a sampled answer when an image region is neutralised, compared with area-matched other regions. Our objective applies the placebo contrast along time, scores the annotated answer rather than a sampled one, and bounds the model's own placebo drop, which the contrast alone does not keep down (Section 4). The mismatched-query test comes from selective prediction (Geifman & El-Yaniv, 2017; Whitehead et al., 2022).

7 Conclusion

We measured evidence dependence by scoring a fixed answer while editing either the event's frames or as many other frames of the same video. The post-training we audited adds accuracy along all three routes to a correct interval, and none of them raises evidence dependence. The evidence drop alone trains a model to react to any edit. A placebo contrast and a bound on the placebo drop make evidence dependence trainable, and the weight on clean supervision sets how much is learned. Masking the event's frames makes models trained this way decline or miss the event more often than their base; masking as many other frames does not. The audit needs only an answer that can be scored and the location of its evidence, so evidence dependence can be reported, and trained, next to accuracy.

AI USE STATEMENT

We used large language model assistants for two tasks. *Retrieval and discovery:* to search for work on video temporal grounding, perturbation-based evaluation and training on visual evidence, and to summarise candidate papers; the authors read every cited paper and decided what to cite and how to position this work against it. *Writing:* to edit drafts written by the authors. The paper reports many models and tests within the page limit, so most of these edits compress: restructuring sentences, merging statements that were made twice, shortening captions, and keeping terms and notation consistent across sections. The argument, the claims and the organisation of the paper are the authors'. We have reviewed all AI-assisted text and take full responsibility for the content of this paper.

REPRODUCIBILITY STATEMENT

Section 2 and Appendix A define the audit: its question populations, placebo draws and intervals. Section 4 and Appendix C give the objective term by term, and Appendix E.1 gives the inference protocol of each model and the training setup. Appendix E.3 lists the code and the question and training-pair identifiers we release; they are included in the supplementary material.

REFERENCES

Rama AlHamidi, Rasul Khanbayov, Erchin Serpedin, and Hasan Kurban. Counterfactual sensitivity is not repairability: Auditing replay probes for video evidence. *arXiv preprint arXiv:2608.15685*, 2026.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025a. URL https://arxiv.org/abs/2511.21631.

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

Jianzhu Bao, Haozhen Zhang, Kuicai Dong, Bozhi Wu, Sarthak Ketanbhai Modi, Zi Pong Lim, Yon Shin Teo, and Wenya Wang. Learning more from less: Exploiting counterfactuals for data-efficient chart understanding. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 27775–27797, 2026. doi: 10.18653/v1/2026.acl-long.1280. URL https://aclanthology.org/2026.acl-long.1280/.

Yeoktatt Cheah, María Pérez-Ortiz, Noah Y. Siegel, and Oana-Maria Camburu. Training large language models for self-explanation faithfulness. *arXiv preprint arXiv:2607.21090*, 2026.

Daniel Cores, Michael Dorkenwald, Manuel Mucientes, Cees G. M. Snoek, and Yuki M. Asano. Lost in time: A new temporal benchmark for VideoLLMs. In *British Machine Vision Conference (BMVC)*, 2025. URL https://bmva-archive.org.uk/bmvc/2025/assets/papers/Paper_857/paper.pdf.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. ERASER: A benchmark to evaluate rationalized NLP models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4443–4458, 2020. doi: 10.18653/v1/2020.acl-main.408. URL https://aclanthology.org/2020.acl-main.408/.

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. *arXiv preprint arXiv:2605.21988*, 2026. URL https://arxiv.org/abs/2605.21988.

Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. Multi-modal hallucination control by visual information grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14303–14312, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Favero_Multi-Modal_Hallucination_Control_by_Visual_Information_Grounding_CVPR_2024_paper.html.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-R1: Reinforcing video reasoning in MLLMs. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://arxiv.org/abs/2503.21776.

Ruth C. Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 3429–3437, 2017. URL https://openaccess.thecvf.com/content_iccv_2017/html/Fong_Interpretable_Explanations_of_ICCV_2017_paper.html.

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. TALL: Temporal activity localization via language query. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 5267–5275, 2017. URL https://openaccess.thecvf.com/content_iccv_2017/html/Gao_TALL_Temporal_Activity_ICCV_2017_paper.html.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, 2017.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2:665–673, 2020. doi: 10.1038/s42256-020-00257-z. URL https://www.nature.com/articles/s42256-020-00257-z.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638, 2025.

Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 5803–5812, 2017. URL https://openaccess.thecvf.com/content_iccv_2017/html/Hendricks_Localizing_Moments_in_ICCV_2017_paper.html.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://papers.nips.cc/paper_files/paper/2019/hash/fe4b8556000d0f0cae99daa5c5c5a410-Abstract.html.

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. In *International Conference on Learning Representations*, 2022.

Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. VTimeLLM: Empower LLM to grasp video moments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14271–14280, 2024. doi: 10.1109/CVPR52733.2024.01353.

Haojie Huang, Xinlei Yu, Chengming Xu, Zhangquan Chen, Cheng Yang, Qingdong He, Yu Yang, Jiangning Zhang, and Xiaobin Hu. Evidence-RL: Towards evidence-intensive visual reasoning. *arXiv preprint arXiv:2608.08021*, 2026.

Alon Jacovi and Yoav Goldberg. Towards faithfully interpretable NLP systems: How should we define and evaluate faithfulness? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4198–4205, 2020. doi: 10.18653/v1/2020.acl-main.386. URL https://aclanthology.org/2020.acl-main.386/.

Sarthak Jain, Sarah Wiegreffe, Yuval Pinter, and Byron C. Wallace. Learning to faithfully rationalize by construction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4459–4473, 2020. doi: 10.18653/v1/2020.acl-main.409. URL https://aclanthology.org/2020.acl-main.409/.

Xun Jiang, Zhuoyuan Wei, Shenshen Li, Xing Xu, Jingkuan Song, and Heng Tao Shen. Counterfactually augmented event matching for de-biased temporal sentence grounding. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 6472–6481, 2024. doi: 10.1145/3664647.3680948. URL https://doi.org/10.1145/3664647.3680948.

Minjoon Jung, Junbin Xiao, Byoung-Tak Zhang, and Angela Yao. On the consistency of video large language models in temporal comprehension. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13713–13722, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Jung_On_the_Consistency_of_Video_Large_Language_Models_in_Temporal_CVPR_2025_paper.html.

Sena Korkut, María Alejandra Bravo Sarmiento, Sanghwan Kim, and Zeynep Akata. From accuracy to visual dependence: Auditing and filtering modality collapse in traffic VideoQA. *arXiv preprint arXiv:2606.30220*, 2026.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2017.

Xiaohan Lan, Yitian Yuan, Xin Wang, Long Chen, Zhi Wang, Lin Ma, and Wenwu Zhu. A closer look at debiased temporal sentence grounding in videos: Dataset, metric, and approach. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 19(6):218:1–218:23, 2023. doi: 10.1145/3565573. URL https://doi.org/10.1145/3565573.

Jae Joong Lee. Accuracy without grounding: Diagnosing visual dependency dissociation in video LLM benchmarks. *arXiv preprint arXiv:2607.13305*, 2026.

Jie Lei, Tamara L. Berg, and Mohit Bansal. Detecting moments and highlights in videos via natural language queries. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://papers.neurips.cc/paper_files/paper/2021/hash/62e0973455fd26eb03e91d5741a4a3bb-Abstract.html.

Jie Lei, Tamara L. Berg, and Mohit Bansal. Revealing single frame bias for video-and-language learning. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 487–507, 2023.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13872–13882, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Leng_Mitigating_Object_Hallucinations_in_Large_Vision-Language_Models_through_Visual_Contrastive_CVPR_2024_paper.html.

Xinhao Li, Ziang Yan, Desen Meng, Lu Dong, Xiangyu Zeng, Yinan He, Yali Wang, Yu Qiao, Yi Wang, and Limin Wang. VideoChat-R1: Enhancing spatio-temporal perception via reinforcement fine-tuning. *arXiv preprint arXiv:2504.06958*, 2025. URL https://arxiv.org/abs/2504.06958.

Geuntaek Lim, Sungjune Park, Jaeyun Lee, Inwoong Lee, Taeoh Kim, Dongyoon Wee, Minho Shim, and Yukyung Choi. Video-Oasis: Rethinking evaluation of video understanding. In *European Conference on Computer Vision (ECCV)*, pp. 608–626, 2026. doi: 10.1007/978-3-032-37095-2_33.

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. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 5971–5984, 2024.

Kevin Qinghong Lin, Pengchuan Zhang, Joya Chen, Shraman Pramanick, Difei Gao, Alex Jinpeng Wang, Rui Yan, and Mike Zheng Shou. UniVTG: Towards unified video-language temporal grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2794–2804, 2023.

Romy Luo, Zihui (Sherry) Xue, Alex Dimakis, and Kristen Grauman. When thinking drifts: Evidential grounding for robust video reasoning. In *Advances in Neural Information Processing Systems*, volume 38, 2025a. doi: 10.52202/085713-2806. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/788f0d336eeddf698c8d527b1794fca4-Abstract-Conference.html.

Tiange Luo, Ang Cao, Gunhee Lee, Justin Johnson, and Honglak Lee. Probing visual language priors in VLMs. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 41120–41156. PMLR, 2025b. URL https://proceedings.mlr.press/v267/luo25b.html.

Guoshun Nan, Rui Qiao, Yao Xiao, Jun Liu, Sicong Leng, Hao Zhang, and Wei Lu. Interventional video grounding with dual contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2765–2775, 2021.

Hyunjong Ok and Jaeho Lee. TempCore: Are video QA benchmarks temporally grounded? *arXiv preprint arXiv:2509.01167*, 2025.

Mayu Otani, Yuta Nakashima, Esa Rahtu, and Janne Heikkilä. Uncovering hidden challenges in query-based video moment retrieval. In *British Machine Vision Conference (BMVC)*, 2020. URL https://www.bmva-archive.org.uk/bmvc/2020/assets/papers/0306.pdf.

Vitali Petsiuk, Abir Das, and Kate Saenko. RISE: Randomized input sampling for explanation of black-box models. In *British Machine Vision Conference (BMVC)*, 2018.

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. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/a85b405ed65c6477a4fe8302b5e06ce7-Abstract-Conference.html.

Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. TimeChat: A time-sensitive multimodal large language model for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14313–14323, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Ren_TimeChat_A_Time-sensitive_Multimodal_Large_Language_Model_for_Long_Video_CVPR_2024_paper.html.

Andrew Slavin Ross, Michael C. Hughes, and Finale Doshi-Velez. Right for the right reasons: Training differentiable models by constraining their explanations. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pp. 2662–2670, 2017. doi: 10.24963/ijcai.2017/371. URL https://www.ijcai.org/proceedings/2017/371.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=ryxGuJrFvS.

Ramprasaath R. Selvaraju, Stefan Lee, Yilin Shen, Hongxia Jin, Shalini Ghosh, Larry Heck, Dhruv Batra, and Devi Parikh. Taking a HINT: Leveraging explanations to make vision and language models more grounded. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2591–2600, 2019. URL https://openaccess.thecvf.com/content_ICCV_2019/html/Selvaraju_Taking_a_HINT_Leveraging_Explanations_to_Make_Vision_and_Language_ICCV_2019_paper.html.

Sethuraman T V, Savya Khosla, Aditi Tiwari, Vidya Ganesh, Rakshana Jayaprakash, Aditya Jain, Vignesh Srinivasakumar, Onkar Kishor Susladkar, Srinidhi Sunkara, Aditya Shanmugham,

Rakesh Vaideeswaran, Abbaas Alif Mohamed Nishar, Simon Jenni, Rohan Maheshwari, and Derek Hoiem. Stress tests REVEAL fragile temporal and visual grounding in video-language models. *arXiv preprint arXiv:2602.11244*, 2026.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. DeepSeekMath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. URL https://arxiv.org/abs/2402.03300.

Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *European Conference on Computer Vision (ECCV)*, 2016.

Ryota Tanaka, Kyosuke Nishida, Kosuke Nishida, Taku Hasegawa, Itsumi Saito, and Kuniko Saito. SlideVQA: A dataset for document visual question answering on multiple images. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 13636–13645, 2023. doi: 10.1609/aaai.v37i11.26598. URL https://ojs.aaai.org/index.php/AAAI/article/view/26598.

Fei Wang, Wenxuan Zhou, James Y. Huang, Nan Xu, Sheng Zhang, Hoifung Poon, and Muhao Chen. mDPO: Conditional preference optimization for multimodal large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8078–8088. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.emnlp-main.460. URL https://aclanthology.org/2024.emnlp-main.460/.

Ye Wang, Ziheng Wang, Boshen Xu, Yang Du, Kejun Lin, Zihan Xiao, Zihao Yue, Jianzhong Ju, Liang Zhang, Dingyi Yang, Xiangnan Fang, Zewen He, Zhenbo Luo, Wenxuan Wang, Junqi Lin, Jian Luan, and Qin Jin. Time-R1: Post-training large vision language model for temporal video grounding. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-2793. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/7801b29c93b599b8d0c44138596bdeed-Abstract-Conference.html.

Ziyue Wang, Sheng Jin, Zhongrong Zuo, Jiawei Wu, Han Qiu, Qi She, Hao Zhang, and Xudong Jiang. Video-KTR: Reinforcing video reasoning via key token attribution. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=p0sDIEsYG3.

Spencer Whitehead, Suzanne Petryk, Vedaad Shakib, Joseph Gonzalez, Trevor Darrell, Anna Rohrbach, and Marcus Rohrbach. Reliable visual question answering: Abstain rather than answer incorrectly. In *European Conference on Computer Vision (ECCV)*, volume 13696 of *Lecture Notes in Computer Science*, pp. 148–166, 2022. doi: 10.1007/978-3-031-20059-5_9. URL https://www.ecva.net/papers/eccv_2022/papers_ECCV/papers/136960146.pdf.

Shengguang Wu, Fan-Yun Sun, Kaiyue Wen, and Nick Haber. Symmetrical visual contrastive optimization: Aligning vision-language models with minimal contrastive images. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 30284–30297. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.1462. URL https://aclanthology.org/2025.acl-long.1462/.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. NExT-QA: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9777–9786, 2021. URL https://openaccess.thecvf.com/content/CVPR2021/html/Xiao_NExT-QA_Next_Phase_of_Question-Answering_to_Explaining_Temporal_Actions_CVPR_2021_paper.html.

Junbin Xiao, Angela Yao, Yicong Li, and Tat-Seng Chua. Can I trust your answer? Visually grounded video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13204–13214, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Xiao_Can_I_Trust_Your_Answer_Visually_Grounded_Video_Question_Answering_CVPR_2024_paper.html.

Yuxi Xie, Guanzhen Li, Xiao Xu, and Min-Yen Kan. V-DPO: Mitigating hallucination in large vision language models via vision-guided direct preference optimization. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 13258–13273, 2024. doi: 10.18653/v1/2024.findings-emnlp.775. URL https://aclanthology.org/2024.findings-emnlp.775/.

Tianrun Xu, Haoda Jing, Ye Li, Yuquan Wei, Jun Feng, Guanyu Chen, Haichuan Gao, Tianren Zhang, Jing Liu, and Feng Chen. DeFacto: Counterfactual thinking with images for enforcing evidence-grounded and faithful reasoning. In *International Conference on Machine Learning (ICML)*, 2026.

Ziang Yan, Zhilin Li, Yinan He, Chenting Wang, Kunchang Li, Xinhao Li, Xiangyu Zeng, Zilei Wang, Yali Wang, Yu Qiao, Limin Wang, and Yi Wang. Task preference optimization: Improving multimodal large language models with vision task alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29880–29892, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Yan_Task_Preference_Optimization_Improving_Multimodal_Large_Language_Models_with_Vision_CVPR_2025_paper.html.

Xun Yang, Fuli Feng, Wei Ji, Meng Wang, and Tat-Seng Chua. Deconfounded video moment retrieval with causal intervention. In *International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, 2021.

Yitian Yuan, Xiaohan Lan, Xin Wang, Long Chen, Zhi Wang, and Wenwu Zhu. A closer look at temporal sentence grounding in videos: Dataset and metric. In *International Workshop on Human-centric Multimedia Analysis (HuMA), ACM Multimedia*, pp. 13–21, 2021. doi: 10.1145/3475723.3484247.

Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. Span-based localizing network for natural language video localization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 6543–6554, 2020a. doi: 10.18653/v1/2020.acl-main.585. URL https://aclanthology.org/2020.acl-main.585/.

Jun Zhang, Teng Wang, Yuying Ge, Yixiao Ge, Xinhao Li, and Limin Wang. TimeLens: Rethinking video temporal grounding with multimodal LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10419–10429, 2026a. URL https://openaccess.thecvf.com/content/CVPR2026/html/Zhang_TimeLens_Rethinking_Video_Temporal_Grounding_with_Multimodal_LLMs_CVPR_2026_paper.html.

Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. Learning 2D temporal adjacent networks for moment localization with natural language. In *AAAI Conference on Artificial Intelligence (AAAI)*, pp. 12870–12877, 2020b.

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. *arXiv preprint arXiv:2604.05117*, 2026b.

Minghang Zheng, Zihao Yin, Yi Yang, Yuxin Peng, and Yang Liu. Temporal-aware reasoning optimization for video temporal grounding. In *International Conference on Machine Learning (ICML)*, 2026a. URL https://arxiv.org/abs/2606.09248.

Zelin Zheng, Xinyan Liu, Ruixin Li, Antoni B. Chan, Guorong Li, Qingming Huang, and Laiyun Qing. Foresee-to-Ground: From predictive temporal perception to evidence-driven reasoning for video temporal grounding. In *International Conference on Machine Learning (ICML)*, 2026b. URL https://arxiv.org/abs/2605.21973.

APPENDIX OVERVIEW

The appendix follows the four steps of the main text. Appendix A gives the audit of Section 2 in full, including what ψ responds to and how Γ relates to the scale of the likelihood. Appendix B breaks

down the accuracy that post-training adds (Section 3). For the objective of Section 4, Appendix C goes through each term, the earlier experiment its bound comes from, and the settings that control how much dependence is learned. The tests of Section 5 on quantities the loss does not contain are in Appendix D, and protocols, training setup and the release in Appendix E.

A MEASURING EVIDENCE DEPENDENCE

A.1 ESTIMAND, POPULATIONS AND RESOLUTION

Readouts. Figure 6 shows the audit on one question, and Table 4 collects the notation. d_E, d_O, ψ and Γ are defined in Section 2. The generation readouts are R@0.5, which counts a prediction as correct when its IoU with the annotation is at least 0.5 and counts an unparsed output as a miss; the parse rate; strict and wide abstention; and coverage, which is $1-$strict abstention, together with R@0.5 on the answered questions. They also include the black-video answering rate and its R@0.5, the evidence share and follows (Appendix D.4; Table 19), and the mismatched-query abstention change (Table 21).

Populations. The likelihood audit runs on the first 200 test questions (79 videos), with three placebo draws per question. Only questions a model localises correctly are scored, and Γ uses $H_\cap$, whose size $n_\cap$ ranges from 55 to 143 on Charades-STA and from 39 to 82 on QVHighlights; Tables 6 and 17 print it per row. We varied the question set in two ways. With one question set held fixed across four placebo-controlled Qwen2.5-VL-7B models (n=87), every sign is preserved. With each contrast computed on each model's own full hit set instead of the intersection, Γ moves by at most 0.10 and keeps its sign in every recomputed mask contrast except IoU-GRPO's, which goes from 0.00 to -0.01. On the 200 audit questions the Qwen base reaches R@0.5 0.560; Table 1 gives its value on the full split.

Placebo construction. The placebo edits as many frames as the evidence set, drawn uniformly from outside the annotated interval, averaged over three draws. It is defined by position alone, so its pool includes the frames of any other annotated event in the video. Each run draws its own placebo frames; the behavioural audit uses one set of draws for all models.

Training seeds and resolution. Intervals are 95% bootstrap intervals over videos, with 4,000 resamples for the accuracy changes against the base, the evidence share and follows, and 2,000 for every other quantity. Two independent training seeds of the Qwen model at α=1 agree in Γ ($+1.03$ $[0.80, 1.27]$ and $+1.13$ $[0.92, 1.34]$) and in accuracy (gains of 6.7 and 6.5 points in R@0.5), and they give the same query-crossed decomposition (Table 16). Audited against each other, with one seed as the base, they give $\Gamma = +0.06$ $[0.02, 0.10]$ under mask and $+0.15$ $[-0.01, 0.31]$ under swap, and an accuracy resolution of 1.1 points in domain and 4.9 out of domain. At α=0.5 the first seed gives Γ $+1.63$ $[1.36, 1.89]$ and a gain of 4.4 $[3.1, 5.7]$ points, and a second seed gives $+1.48$ $[1.24, 1.71]$ and 5.5 $[4.2, 6.8]$. Audited against the first, the second seed gives $\Gamma = -0.15$ $[-0.23, -0.08]$ under mask and -0.05 $[-0.20, 0.12]$ under swap. The audit's resolution (Section 2) is the largest absolute interval end over the two pairs. The tests of the 16-frame experiment (Appendix C.2) are Holm-corrected (Holm, 1979).

A.2 WHAT ψ RESPONDS TO

Table 5 compares ψ or Γ with controls that each differ from the audit in one respect.

A.3 Γ AND THE SCALE OF THE LIKELIHOOD

Γ does not follow the change in clean-video likelihood that training produces. Table 6 sets each trained model's Γ next to that change and gives two versions of Γ that do not depend on scale.

Units and scale. Likelihoods in the audit are token means in nats. In the objective, the evidence margin and the placebo contrast compare log-likelihoods summed over the answer tokens, while the placebo bound, like the audit, uses the token-mean placebo drop, and the clean-supervision term is a token-mean NLL. Under sequence sums, Γ scales by roughly the answer length (≈ 18 tokens).

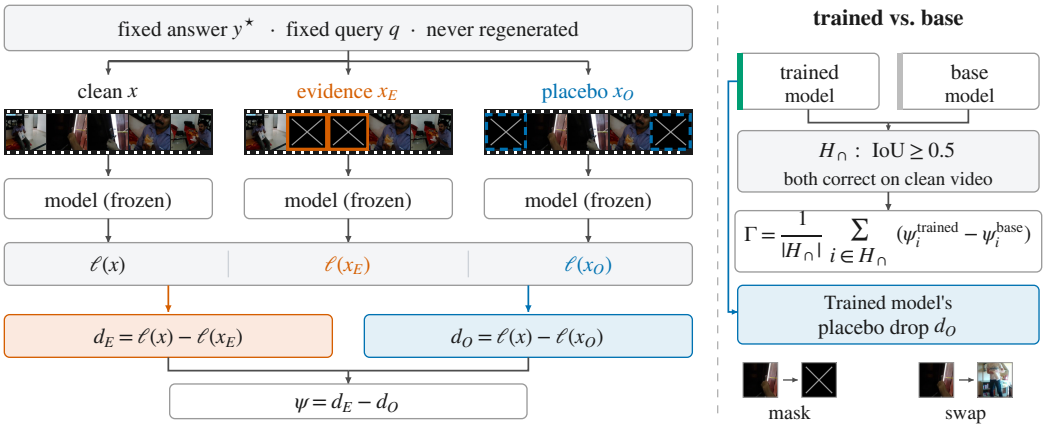

Figure 6: **On each question, the audit scores one fixed answer while editing either the evidence frames or as many other frames.** The annotated answer $y^\star$ and the query q are held fixed and scored by teacher forcing. Editing the frames inside the annotated interval gives x_E; editing as many frames outside it with the same operator gives the placebo x_O. The drops of the answer's log-likelihood under the two edits are d_E and d_O, and $\psi = d_E - d_O$. The figure shows one mask draw; the audit averages d_O over three draws. The right panel shows how a trained model is compared with its base: Γ is the mean trained-minus-base change in ψ over $H_\cap$, the questions both models localise on the clean video (IoU ≥ 0.5), and the trained model's placebo drop d_O is reported next to it, under the mask and the swap operator.

Table 4: **This table collects the notation used throughout the paper.** Likelihood quantities are token means in nats, and behavioural quantities are fractions of questions. In every table, brackets give 95% bootstrap intervals over videos.

Symbol	Meaning
$y^\star, x, q$	annotated interval, video, query
x_E, x_O	video with the evidence frames edited; with as many frames outside $y^\star$ edited (placebo)
d_E, d_O	drop of the answer's log-likelihood under x_E and under x_O (mean of three placebo draws)
$\psi = d_E - d_O$	evidence dependence of one model on one question
$H_\cap, n_\cap$	questions both the trained model and its base localise (IoU ≥ 0.5); their number
Γ or Γ (mask); Γ (swap)	mean trained-minus-base change in ψ on $H_\cap$ under mask (primary) and under swap
sign fraction	fraction of $H_\cap$ on which the trained model's ψ exceeds its base's; scale-free companion of Γ
Γ_c	mean trained-minus-base change on $H_\cap$ in the bounded index $c = (d_E - d_O)/(\lvert d_E\rvert + \lvert d_O\rvert + 0.05)$; scale-free companion of Γ
$\pi(q)$	partner contrast: drop in the log-likelihood of $y_i^\star$ when the question's evidence frames E_i are masked minus the drop when a partner event's frames E_j are masked, under query q (Appendix D.2)
$\psi_{\text{own}}, \psi_{\text{cross}}$	$\pi(q_i)$ and $\pi(q_j)$: the partner contrast under the question's own query and under the partner's query; Δ: trained minus base, mean over the partnered questions of $H_\cap$
Γ_{spec}	$\Delta\psi_{\text{own}} - \Delta\psi_{\text{cross}}$: the part of the change in the partner contrast that depends on which query is asked
$\Gamma_{\text{agn}}, \Gamma_{\text{sem}}$	trained-minus-base change in ψ under the partner's query (placebo drawn outside both events); $\Gamma - \Gamma_{\text{agn}}$
evidence share	$1 - a_E/a_O$, with a_E and a_O the R@0.5 of predictions regenerated with the evidence frames masked and with as many placebo frames masked (first 1,000 test questions)
follows	fraction of $H_\cap$ on which the prediction, regenerated after the event's frames are exchanged with an equally long block elsewhere, has IoU ≥ 0.5 with the annotated interval moved to that block
$\alpha, \beta, w_{\text{op}}$	objective weights: clean supervision, margin scale, per-operator weights

Scale-free versions and clean likelihood. The two scale-free versions are the fraction of questions on which $\psi^{\text{trained}} > \psi^{\text{base}}$, and the mean change Γ_c in a bounded contrast index $c = (d_E - d_O)/(|d_E| + |d_O| + 0.05)$. Correlations are over the 32 post-trained models on Charades-STA in the table. SFT lowers the clean-video NLL as much as the placebo-controlled models do (from 1.38 to 0.39 nats per token) while its Γ is at or below zero, and across the 32 models Γ is at most weakly related to the change in clean NLL (Spearman -0.20). The sign fraction orders the models as Γ does (Spearman 0.84). It is 0.73 or higher for every placebo-controlled model in the table, 0.60 or lower for SFT, IoU-GRPO, TimeZero, Time-R1 and TimeLens-8B, and 0.71 for VideoChat-R1, whose small Γ ($+0.05$) is spread over many questions. The bounded index Γ_c (Spearman 0.59 with Γ) also penalises the placebo drop relative to the evidence drop, and so separates the models trained without placebo control more sharply. The evidence-margin-only model has $\Gamma = +0.54$ but $\Gamma_c = -0.34$ $[-0.45, -0.22]$, and the two models trained without the placebo bound lose most of their Γ_c. Under Γ_c, SFT's interval contains zero, and TimeLens-8B against Qwen3-VL-8B stays negative (-0.42 $[-0.65, -0.23]$; sign fraction 0.29).

Table 5: ψ **grows with how much of the evidence is edited, and** Γ **keeps its sign across placebo shapes, question sets, query wordings and unseen operators.** Each row names a property of the audit, the control that ψ or Γ is compared with, which differs from the audit in that one respect, and the result. Rows marked * were run in the 16-frame setup of Appendix C.2.

Property	Compared with	Result
Ring of adjacent frames*	frames just outside the interval	drop lies between the evidence and placebo drops
Edit severity*	edits removing more or less of the evidence	drop rises monotonically with severity
Sub-step jitter*	each evidence frame replaced by its neighbour half a sampling step away, and likewise at the placebo	difference of the two drops $+0.004$ for the base
Contiguous placebo	a placebo of consecutive frames	Γ $+1.33$ to $+1.65$ for the placebo-controlled Qwen2.5-VL-7B, TimeZero and VideoChat-R1 at α=0.5 and Time-R1 at α=0.25, -0.07 to $+0.03$ for the others
Selection on $H_\cap$	a fixed cohort; each model's own hit set	signs kept (Appendix A.1)
Query crossed to another event	the partner event's query	Section 5; Table 16
Query paraphrased	a second human-written query	contrast retained (Appendix D.2)
Operators unseen in training	noise, Gaussian blur, patch scrambling	15/15 signs kept (Table 15)

A.4 Single questions, a larger base and other inputs

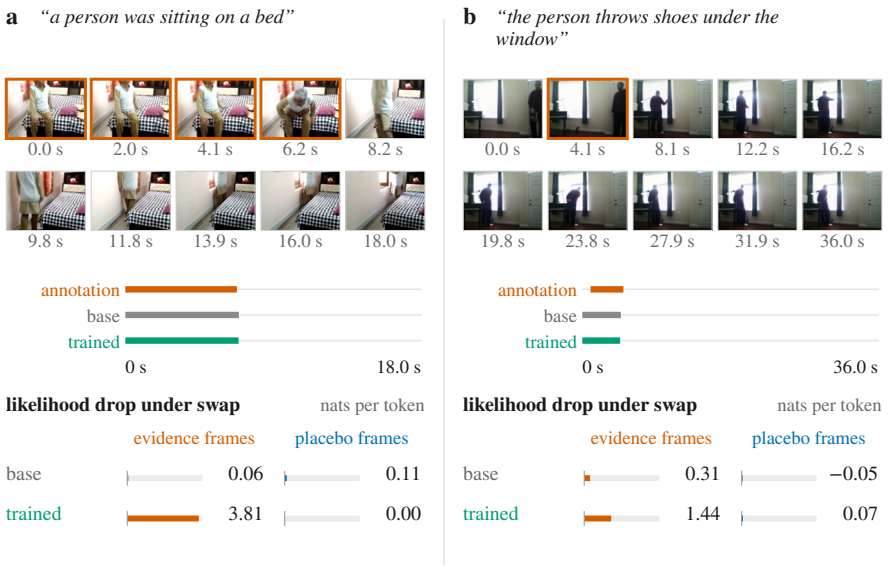

Figure 7: **Two models give the same correct interval for different reasons.** Panels (a) and (b) each show a test question on which the Qwen2.5-VL-7B base and its placebo-controlled model (α=0.5) predict the same interval, with IoU above 0.7. The top rows show the native frames with the annotated ones outlined, followed by the annotated interval and both predictions. The bars below give the drop in the answer's log-likelihood (nats per token) when the evidence frames are swapped out and when as many frames outside the interval are swapped the same way. In (a), for the same correct interval, the base loses 0.06 nats per token under the evidence swap and its trained model 3.81, with placebo drops of 0.11 and 0.00.

A larger base. At 32B, the Qwen2.5-VL-32B base localises at R@0.5 0.475, answers 16% of black videos and has $\psi = +0.18$ $[0.16, 0.20]$, at the level of the Qwen2.5-VL-7B base (Section 2).

Table 6: Γ **does not follow the change in clean-video NLL.** Each row pairs a model with its base: our trained models, the three public checkpoints against the Qwen2.5-VL-7B base (the pairs of Table 1), and TimeLens-8B against Qwen3-VL-8B. All values use the mask operator on $H_\cap$, the $n_\cap$ questions both models localise. The columns give the clean-video negative log-likelihood (NLL) of the annotated answer in nats per token for the base and the trained model, then Γ (mask), its sign fraction (the fraction of $H_\cap$ on which ψ rises) and Γ_c, the change in the bounded index $c = (d_E - d_O)/(|d_E| + |d_O| + 0.05)$, with its 95% interval. Text-first rows put the query before the video.

Configuration	$n_\cap$	clean NLL base / trained	Γ (mask)	sign frac.	Γ_c
Charades-STA					
Qwen2.5-VL-7B, IoU-GRPO	104	1.35 / 1.39	0.00	0.49	+0.02 [−0.09, 0.12]
Qwen2.5-VL-7B, SFT	112	1.38 / 0.39	−0.09	0.40	0.00 [−0.14, 0.13]
Qwen2.5-VL-7B, evidence margin only	104	1.36 / 0.40	+0.54	0.74	−0.34 [−0.45, −0.22]
Qwen2.5-VL-7B, placebo-controlled α=0.25	101	1.34 / 0.59	+2.03	1.00	+0.35 [0.25, 0.44]
Qwen2.5-VL-7B, placebo-controlled α=0.5	112	1.38 / 0.46	+1.63	0.97	+0.40 [0.30, 0.51]
Qwen2.5-VL-7B, placebo-controlled α=0.5, β=0.03	89	1.35 / 0.39	+1.36	0.73	+0.26 [0.14, 0.38]
Qwen2.5-VL-7B, placebo-controlled α=0.5, 8,434 pairs, trained on 16 frames	89	1.34 / 2.05	+0.23	0.83	+0.19 [0.09, 0.29]
Qwen2.5-VL-7B, placebo-controlled α=1	102	1.35 / 0.40	+1.03	0.80	+0.28 [0.18, 0.37]
Qwen2.5-VL-7B, placebo-controlled α=1 (second seed)	100	1.35 / 0.39	+1.13	0.88	+0.28 [0.20, 0.38]
Qwen2.5-VL-7B, placebo-controlled α=1, β=0.03	105	1.37 / 0.38	+0.82	0.83	+0.34 [0.23, 0.46]
Qwen2.5-VL-7B, uniform weights	103	1.35 / 0.48	+1.99	1.00	+0.34 [0.26, 0.45]
Qwen2.5-VL-7B, no placebo bound	87	1.35 / 0.39	+2.05	1.00	+0.11 [−0.02, 0.22]
Qwen2.5-VL-7B, no bound, uniform	98	1.35 / 0.41	+2.62	1.00	+0.14 [0.04, 0.26]
Qwen2.5-VL-7B, two-stage, α=0.5	84	1.35 / 0.39	+2.10	1.00	+0.36 [0.25, 0.49]
Qwen2.5-VL-7B, two-stage, β=0.03	86	1.36 / 0.39	+2.75	0.98	+0.38 [0.28, 0.49]
TimeZero, placebo-controlled α=0.25	133	1.36 / 0.54	+2.00	0.99	+0.39 [0.30, 0.47]
TimeZero, placebo-controlled α=0.5	135	1.36 / 0.49	+1.40	0.99	+0.35 [0.27, 0.44]
TimeZero, placebo-controlled α=0.5, β=0.03	131	1.36 / 0.39	+0.32	0.91	+0.27 [0.17, 0.36]
TimeZero, placebo-controlled α=0.5, text-first	143	1.41 / 0.47	+1.29	0.98	+0.58 [0.45, 0.71]
TimeZero, placebo-controlled α=1	130	1.36 / 0.43	+1.07	0.93	+0.37 [0.27, 0.47]
Time-R1, placebo-controlled α=0.1	130	1.30 / 0.85	+2.40	1.00	+0.37 [0.27, 0.46]
Time-R1, placebo-controlled α=0.25	131	1.34 / 0.58	+1.88	1.00	+0.42 [0.31, 0.54]
Time-R1, placebo-controlled α=0.5	127	1.29 / 0.42	+1.46	0.98	+0.32 [0.23, 0.42]
Time-R1, placebo-controlled α=1	102	1.31 / 0.39	+0.95	0.86	+0.25 [0.14, 0.34]
VideoChat-R1, placebo-controlled α=0.25	136	0.58 / 0.58	+1.63	0.99	+0.42 [0.33, 0.51]
VideoChat-R1, placebo-controlled α=0.5	132	0.57 / 0.42	+1.43	1.00	+0.44 [0.36, 0.53]
VideoChat-R1, placebo-controlled α=1	129	0.58 / 0.38	+1.20	1.00	+0.40 [0.32, 0.48]
TimeLens-8B, placebo-controlled α=0.5	85	1.98 / 0.49	+1.36	1.00	+0.64 [0.46, 0.83]
TimeZero (public)	103	1.35 / 1.36	−0.08	0.42	−0.11 [−0.25, 0.03]
Time-R1 (public)	101	1.35 / 1.28	+0.02	0.60	−0.02 [−0.10, 0.06]
VideoChat-R1 (public)	101	0.66 / 0.56	+0.05	0.71	0.00 [−0.13, 0.12]
TimeLens-8B (public) against Qwen3-VL-8B	85	0.51 / 2.00	−0.16	0.29	−0.42 [−0.65, −0.23]
Charades, TimeLens labels					
Qwen2.5-VL-7B, SFT	55	1.22 / 0.40	−0.03	0.51	+0.05 [−0.09, 0.18]
Qwen2.5-VL-7B, placebo-controlled α=0.5	57	1.21 / 0.57	+1.21	0.91	+0.41 [0.26, 0.57]
Qwen2.5-VL-7B, placebo-controlled α=1	58	1.21 / 0.41	+0.87	0.90	+0.41 [0.29, 0.54]
TimeZero, placebo-controlled α=0.5	63	1.30 / 0.60	+1.00	0.92	+0.50 [0.37, 0.64]
Time-R1, placebo-controlled α=0.25	62	1.23 / 0.79	+1.43	0.98	+0.53 [0.40, 0.67]
VideoChat-R1, placebo-controlled α=0.5	57	0.54 / 0.46	+0.98	0.98	+0.48 [0.36, 0.59]
TimeLens-8B, SFT	75	0.23 / 0.44	−0.21	0.05	−0.13 [−0.17, −0.09]
TimeLens-8B, placebo-controlled α=0.5	70	0.24 / 0.49	+0.95	0.96	+0.11 [0.04, 0.17]
QVHighlights					
Qwen2.5-VL-7B, SFT	52	1.18 / 0.36	−0.02	0.46	+0.04 [−0.03, 0.13]
Qwen2.5-VL-7B, placebo-controlled α=0.5	43	1.16 / 0.88	+1.30	1.00	+0.27 [0.17, 0.38]
Qwen2.5-VL-7B, placebo-controlled α=1	39	1.13 / 0.63	+1.06	0.95	+0.19 [0.11, 0.28]
TimeZero, placebo-controlled α=0.5, text-first	51	1.16 / 0.80	+0.92	1.00	+0.27 [0.18, 0.37]

Beyond video. The audit also applies to inputs divided into other units. On document QA, with SlideVQA (Tanaka et al., 2023) pages in place of frames, Qwen2.5-VL-7B and Qwen3-VL-8B lose 2.00 and 2.61 nats per token of the annotated answer's likelihood when the evidence pages are masked and 0.00 and 0.06 when as many other pages are.

B What post-training accuracy is made of

B.1 Every configuration against its base

Table 7 pairs the trained configurations of Sections 3 and 4 with the base each was trained from. It extends Table 1 with coverage, accuracy on answered questions, both edit operators and $n_\cap$, and is the reference for the trained configurations of Section 5. Coverage equals the parse rate in every

Table 7: **The full audit matrix on Charades-STA compares every trained configuration with the base it was trained from.** Each block of rows belongs to one base model. A base row gives its R@0.5 and, in the column *ans.*, its R@0.5 on the questions it answers; a trained row gives the paired changes of these two values from its base in points, the second on the questions the trained model answers. A † marks a change that is not paired by question but is the difference of the two models' values. The column *ord.* gives the message order, video first (v) or query text first (t); *cov.* is the coverage, 1−strict abstention; Γ (mask) and Γ (swap) are in nats per token on the $n_\cap$ questions both models localise. The last row is a constant $(0, 10)$ s window, which uses no model.

Model / objective	ord.	R@0.5 / Δ	cov.	ans. / Δ	Γ (mask)	Γ (swap)	$n_\cap$
Qwen2.5-VL-7B							
base	v	0.480	0.994	0.482	—	—	—
placebo-controlled α=1	v	+6.7 [5.2, 8.2]	0.937	+9.5	+1.03	+1.00	102
placebo-controlled α=0.5	v	+4.4 [3.1, 5.7]	0.957	+6.1	+1.63	+1.41	112
placebo-controlled α=0.5, β=0.03	v	+1.2 [−0.6, 2.9]	0.831	+9.6	+1.36	+1.16	89
placebo-controlled α=0.25	v	+2.1 [0.8, 3.2]	0.959	+3.4	+2.03	+1.84	101
placebo-controlled α=1, β=0.03	v	+6.0†	0.926	+10.2†	+0.82	+0.72	105
SFT	v	+9.4 [7.9, 11.0]	1.00	+9.4	−0.09	−0.03	112
IoU-GRPO	v	+13.6 [11.9, 15.4]	0.999	+13.4†	0.00	+0.02	104
mDPO-style	v	+8.8 [7.3, 10.4]	0.997	+8.9	−0.03	+0.01	112
Qwen2.5-VL-7B: objective ablation at α=0.5							
uniform weights	v	+5.4 [4.1, 6.7]	0.974	+6.4	+1.99	+1.89	103
no placebo bound	v	−3.8 [−5.4, −2.1]	0.802	+5.7	+2.05	+2.34	87
no bound, uniform	v	+3.8 [2.4, 5.2]	0.938	+6.6	+2.62	+2.60	98
evidence margin only	v	+6.7 [5.2, 8.3]	0.959	+8.8†	+0.54	+1.09	104
Qwen2.5-VL-7B: training budget							
placebo-controlled α=0.5, 4,800 pairs	v	+8.1 [6.7, 9.5]	0.981	+8.7	+2.78	+2.58	105
placebo-controlled α=0.5, 2,400 pairs, trained on 16 frames	v	+2.5†	0.975	+3.6†	+0.06	+0.04	103
placebo-controlled α=0.5, 8,434 pairs, trained on 16 frames	v	−5.7†	0.986	−5.3†	+0.23	+0.24	89
TimeZero							
base	v	0.677	1.00	0.677	—	—	—
base, text-first	t	0.705	1.00	0.705	—	—	—
placebo-controlled α=0.5	v	+0.5 [−0.7, 1.5]	0.999	+0.5	+1.40	+1.31	135
placebo-controlled α=0.25	v	+0.2†	1.00	+0.2†	+2.00	+1.66	133
placebo-controlled α=0.5, text-first	t	−2.6†	1.00	−2.6†	+1.29	+1.13	143
placebo-controlled α=1	v	−0.6†	1.00	−0.6†	+1.07	+0.92	130
placebo-controlled α=0.5, β=0.03	v	−1.0 [−2.1, 0.1]	1.00	−1.0	+0.32	+0.23	131
Time-R1							
base	v	0.601	1.00	0.601	—	—	—
placebo-controlled α=0.25	v	+1.6†	0.996	+1.8†	+1.88	+1.66	131
placebo-controlled α=0.5	v	+0.9 [−0.2, 2.0]	0.982	+2.0†	+1.46	+1.20	127
placebo-controlled α=0.1	v	+0.3†	0.998	+0.4†	+2.40	+2.20	130
placebo-controlled α=1	v	−8.4†	0.805	+4.2†	+0.95	+0.99	102
VideoChat-R1							
base	v	0.668	1.00	0.668	—	—	—
placebo-controlled α=0.5	v	−2.7 [−4.2, −1.2]	1.00	−2.7	+1.43	+0.89	132
placebo-controlled α=0.25	v	−3.2 [−4.6, −1.8]	1.00	−3.2	+1.63	+0.92	136
placebo-controlled α=1	v	−2.4 [−3.8, −0.9]	1.00	−2.4	+1.20	+0.85	129
TimeLens-8B							
base	v	0.417	1.00	0.417	—	—	—
placebo-controlled α=0.25	v	+17.9 [15.9, 20.0]	1.00	+17.9	+1.72	+1.15	84
placebo-controlled α=0.5	v	+20.9 [18.9, 22.8]	1.00	+20.9	+1.36	+0.88	85
placebo-controlled α=1	v	+23.2 [21.4, 25.1]	1.00	+23.2	+0.92	+0.66	87
constant window $(0, 10)$ s, no model	—	0.317	1.00	0.317	—	—	—

configuration, because every output without an interval is the refusal string. At β=0.03 the model declines more often and, on the questions it answers, gains the most over its base among the Qwen models at α=0.5. Γ (swap) tracks Γ (mask), mostly at lower magnitude.

A second backbone. TimeLens-8B (Zhang et al., 2026a), post-trained for accuracy from Qwen3-VL-8B-Instruct on TimeLens-100K, the automatically annotated training set released with the TimeLens benchmark, is audited against Qwen3-VL-8B-Instruct under the TimeLens template, with the video first, on 200 questions ($n_\cap = 85$). Its Γ (mask) is -0.16 $[-0.26, -0.08]$ (Section 3), and its Γ (swap) is -0.14 $[-0.20, -0.09]$. The change comes from the evidence side: on $H_\cap$ the evidence drop falls and the placebo drop stays where it was (Table 13).

B.2 THE NEWLY CORRECT ANSWERS

Table 8 breaks IoU-GRPO's gain down on the *newly correct* set, the questions that the trained model gets right on the clean video and its base gets wrong, with two placebo-controlled models for comparison. IoU-GRPO's newly correct questions are solvable by the constant window less often than all questions are (26% against 32%), and 4.5% of them survive a black video. Regenerated with the evidence frames masked, IoU-GRPO's newly correct answers lose 64 points $[54, 72]$ more than with matched placebo masking. Table 9 compares, on the audited questions, a model's ψ on its newly correct questions with the base model's ψ on the questions the base gets right. For SFT, the newly correct questions carry less evidence dependence than the base's own correct answers ($+0.11$ against $+0.18$). For IoU-GRPO, TimeZero and Time-R1 the interval of the difference includes zero, and for the placebo-controlled models the difference is positive and larger (Section 3 gives the range); these sets are unpaired and hold 11 to 28 newly correct questions each.

Table 8: **Most of IoU-GRPO's newly correct answers survive masking as many placebo frames, and most are lost when the evidence frames are masked.** $|\mathrm{NC}|$ is the number of questions a trained model localises on the clean video and its base does not. The other columns give the percentage of that set that the constant $(0, 10)$ s window also solves (on all questions it solves 32%), that stays correct on an all-black video, and that stays correct when regenerated with the evidence frames or as many placebo frames masked. The last two columns are given for IoU-GRPO only, on the 181 of its newly correct questions that are among the 1,000 questions of the behavioural audit. Brackets give 95% intervals.

Objective	$\lvert\mathrm{NC}\rvert$	const-win (%)	black (%)	evid-mask (%)	placebo-mask (%)
IoU-GRPO	666	26.0 [22.0, 29.7]	4.5 [2.9, 6.2]	24.3 [17.6, 31.7]	87.8 [83.0, 92.1]
placebo-controlled α=0.5	273	19.8 [14.7, 25.3]	0.0 [0.0, 0.0]	—	—
placebo-controlled α=1	445	24.7 [19.9, 29.5]	0.0 [0.0, 0.0]	—	—

Table 9: ψ **on the newly correct questions stays within** 0.1 **nats per token of the base's after accuracy training and far exceeds it after placebo-controlled training.** All values come from the likelihood audit with the mask operator on 200 questions and are in nats per token. $|\mathrm{NC}|$ is the number of questions the trained model localises and its base does not, and $\psi^{\mathrm{trained}}(\mathrm{NC})$ is the trained model's ψ on them. ψ^{base} is the base's ψ on the questions the base localises, in the audit paired with that model. The column *difference* is the first minus the second; the two sets are not paired, and brackets give 95% intervals.

Configuration	$\lvert\mathrm{NC}\rvert$	$\psi^{\mathrm{trained}}(\mathrm{NC})$	ψ^{base} (base hits)	difference
Qwen2.5-VL-7B, SFT	23	+0.11 [0.06, 0.15]	+0.18 [0.12, 0.23]	−0.07 [−0.13, −0.01]
Qwen2.5-VL-7B, IoU-GRPO	38	+0.15 [0.09, 0.23]	+0.23 [0.16, 0.30]	−0.08 [−0.17, 0.02]
Qwen2.5-VL-7B, placebo-controlled α=0.25	11	+2.22 [1.63, 2.83]	+0.23 [0.16, 0.30]	+1.99 [1.41, 2.61]
Qwen2.5-VL-7B, placebo-controlled α=0.5	14	+1.75 [1.11, 2.34]	+0.18 [0.12, 0.23]	+1.57 [0.93, 2.14]
Qwen2.5-VL-7B, placebo-controlled α=1	26	+1.16 [0.84, 1.55]	+0.23 [0.16, 0.30]	+0.93 [0.60, 1.32]
Qwen2.5-VL-7B, placebo-controlled α=1 (second seed)	28	+1.24 [0.87, 1.66]	+0.23 [0.16, 0.30]	+1.01 [0.63, 1.44]
TimeZero (public)	39	+0.26 [0.19, 0.32]	+0.23 [0.16, 0.30]	+0.03 [−0.08, 0.12]
Time-R1 (public)	35	+0.22 [0.15, 0.30]	+0.23 [0.16, 0.30]	−0.01 [−0.11, 0.10]

B.3 ANSWERING BLIND

Table 10 gives the all-black readout by configuration, dataset and template. Under the VideoChat-R1 and TimeLens-8B templates no model refuses a black video. On QVHighlights the Qwen base answers no black video, while TimeZero answers 94.3% of them and its placebo-controlled model 16.7% (Figure 11b).

B.4 ACCURACY AND THE ANNOTATION

Figure 2b–c draws on the re-annotation matrix of Table 11, which scores the same predictions under each label set on 870 questions and measures Γ against the re-annotated intervals. On these 870 questions TimeLens-8B's R@0.5 differs from its base Qwen3-VL-8B's by -19.9 $[-23.5, -16.4]$ points under the original labels and by $+8.7$ $[5.4, 12.1]$ under the TimeLens labels (paired by question). On all 3,363 re-annotated test questions the difference is $+8.9$ $[7.3, 10.6]$ points. On the 870 questions, our placebo-controlled adapter of TimeLens-8B, trained on the original labels, does the reverse of TimeLens-8B: it leads Qwen3-VL-8B under the original labels and trails it under the TimeLens labels. The three public checkpoints lead the Qwen2.5-VL-7B base by 9 to 19 points under the original labels, and the spread of the nine 7B configurations shrinks from 20 points under

Table 10: **On an all-black video, SFT and the public checkpoints keep answering.** The placebo-controlled Qwen2.5-VL-7B models almost never answer, and those trained from TimeZero and Time-R1 answer less often than their bases. Every frame is set to zero; duration and video token count are kept. The column *data* names the question set: Charades-STA (STA), QVHighlights (QVH) or Charades-STA with TimeLens labels (C-TL). *ans.* is the fraction of questions still given an interval, and *black R@0.5* is the R@0.5 over all questions. *abst.* is the strict (s) or wide (w) abstention rate: the strict criterion counts only the verbatim output `There are none.`, and the wide criterion counts any refusal wording (Appendix E.1). Text-first rows put the query before the video.

Model / objective	data	ans.	black R@0.5	abst. (s)	abst. (w)
Charades-STA					
Qwen2.5-VL-7B, base, text-first	STA	0.578	0.092	0.289	0.290
Qwen2.5-VL-7B, base	STA	0.116	0.007	0.348	0.878
Qwen2.5-VL-7B, placebo-controlled α=0.5, β=0.03	STA	0.002	0.000	0.498	0.994
Qwen2.5-VL-7B, placebo-controlled α=0.5	STA	0.000	0.000	0.065	0.982
Qwen2.5-VL-7B, placebo-controlled α=1	STA	0.001	0.000	0.051	0.994
Qwen2.5-VL-7B, placebo-controlled α=0.5, 8,434 pairs, trained on 16 frames	STA	0.001	0.000	0.519	0.998
Qwen2.5-VL-7B, uniform weights	STA	0.000	0.000	0.042	0.993
Qwen2.5-VL-7B, no placebo bound	STA	0.000	0.000	0.029	0.994
Qwen2.5-VL-7B, no bound, uniform	STA	0.000	0.000	0.025	0.990
Qwen2.5-VL-7B, IoU-GRPO	STA	0.324	0.037	0.301	0.669
Qwen2.5-VL-7B, mDPO-style	STA	0.025	0.002	0.733	0.970
Qwen2.5-VL-7B, SFT	STA	0.784	0.140	0.200	0.216
TimeZero, base, text-first	STA	0.940	0.189	0.000	0.000
TimeZero, base	STA	0.968	0.136	0.004	0.307
TimeZero, placebo-controlled α=0.25	STA	0.474	0.085	0.180	0.782
TimeZero, placebo-controlled α=0.5, β=0.03	STA	0.530	0.098	0.127	0.814
TimeZero, placebo-controlled α=0.5, text-first	STA	0.706	0.172	0.077	0.709
TimeZero, placebo-controlled α=0.5, trained text-first, evaluated video-first	STA	0.591	0.110	0.074	0.846
TimeZero, placebo-controlled α=0.5, trained video-first, evaluated text-first	STA	0.551	0.109	0.119	0.656
TimeZero, placebo-controlled α=0.5	STA	0.303	0.052	0.312	0.844
TimeZero, placebo-controlled α=1, text-first	STA	0.933	0.233	0.000	0.000
TimeZero, placebo-controlled α=1	STA	0.439	0.085	0.178	0.795
TimeZero, SFT, text-first	STA	0.938	0.253	0.000	0.000
TimeZero, two-stage (SFT, then placebo-controlled), text-first	STA	0.924	0.230	0.000	0.001
Time-R1, base, text-first	STA	0.762	0.131	0.086	0.087
Time-R1, base	STA	0.796	0.105	0.123	0.272
Time-R1, placebo-controlled α=0.1	STA	0.495	0.062	0.230	0.513
Time-R1, placebo-controlled α=0.25	STA	0.049	0.011	0.178	0.959
Time-R1, placebo-controlled α=0.5	STA	0.487	0.042	0.214	0.547
Time-R1, placebo-controlled α=1	STA	0.184	0.010	0.463	0.807
VideoChat-R1, base, text-first	STA	0.948	0.211	0.000	0.000
VideoChat-R1, base	STA	1.00	0.153	0.000	0.000
VideoChat-R1, placebo-controlled α=0.5	STA	1.00	0.120	0.000	0.000
VideoChat-R1, placebo-controlled α=1	STA	1.00	0.092	0.000	0.000
VideoChat-R1, SFT, text-first	STA	0.961	0.262	0.000	0.000
VideoChat-R1, two-stage (SFT, then placebo-controlled), text-first	STA	0.941	0.239	0.000	0.000
TimeLens-8B, base	STA	1.00	0.152	0.000	0.000
TimeLens-8B, placebo-controlled α=0.5	STA	1.00	0.179	0.000	0.000
QVHighlights					
Qwen2.5-VL-7B, base	QVH	0.000	0.000	0.298	0.984
Qwen2.5-VL-7B, placebo-controlled α=0.5	QVH	0.003	0.000	0.174	0.987
TimeZero, base	QVH	0.943	0.013	0.008	0.719
TimeZero, placebo-controlled α=0.5	QVH	0.167	0.006	0.432	0.968
Charades, TimeLens re-annotation					
TimeLens-8B, base	C-TL	1.00	0.134	0.000	0.000
TimeLens-8B, placebo-controlled α=0.5	C-TL	1.00	0.133	0.000	0.000

the original labels to 6 under the TimeLens labels (Table 11). Against the re-annotated intervals, TimeLens-8B's Γ (mask) relative to the same base is $+0.01$ $[-0.03, 0.05]$ (Table 11). Table 12 is the window-length substitution. The constant is the typical span of the label set, so the substitution measures how much accuracy one number about the annotation carries.

Table 11: **A second annotation of the same videos reorders the configurations by accuracy.** Measured against it, every placebo-controlled model keeps a positive Γ and no accuracy-oriented model rises beyond the audit's resolution. The upper block scores one run's predictions on the 870 questions present in both label sets under each set of labels, where STA denotes the original Charades-STA labels and TL the TimeLens labels. The lower block gives Γ (mask) and Γ (swap) in nats per token, with 95% intervals, measured with the re-annotated intervals as the fixed answer, together with the number $n_\cap$ of questions both models localise and the fraction of them on which ψ rises.

Model / objective	R@0.5, STA labels	R@0.5, TL labels
Qwen2.5-VL-7B, base	0.525	0.369
Qwen2.5-VL-7B, placebo-controlled α=0.5	0.561	0.382
Qwen2.5-VL-7B, SFT	0.622	0.331
TimeZero, base	0.713	0.361
TimeZero, placebo-controlled α=0.5	0.726	0.380
Time-R1, base	0.616	0.340
Time-R1, placebo-controlled α=0.25	0.644	0.364
VideoChat-R1, base	0.711	0.378
VideoChat-R1, placebo-controlled α=0.5	0.697	0.394
Qwen3-VL-8B, base of TimeLens-8B	0.631	0.537
TimeLens-8B (public checkpoint)	0.432	0.624
TimeLens-8B, placebo-controlled α=0.5	0.670	0.429

Evidence dependence measured against the re-annotated intervals

Model / objective	Γ (mask)	Γ (swap)	$n_\cap$	sign frac.
Qwen2.5-VL-7B, placebo-controlled α=0.5	+1.21 [0.92, 1.49]	+1.06 [0.77, 1.37]	57	0.912
Qwen2.5-VL-7B, placebo-controlled α=1	+0.87 [0.63, 1.10]	+1.04 [0.77, 1.33]	58	0.897
Qwen2.5-VL-7B, SFT	−0.03 [−0.08, 0.01]	+0.04 [0.00, 0.08]	55	0.509
TimeZero, placebo-controlled α=0.5	+1.00 [0.78, 1.28]	+0.95 [0.74, 1.20]	63	0.921
Time-R1, placebo-controlled α=0.25	+1.43 [1.19, 1.70]	+1.19 [0.90, 1.52]	62	0.984
VideoChat-R1, placebo-controlled α=0.5	+0.98 [0.86, 1.14]	+0.68 [0.51, 0.84]	57	0.982
TimeLens-8B against Qwen3-VL-8B	+0.01 [−0.03, 0.05]	+0.03 [−0.01, 0.07]	78	0.564
TimeLens-8B, placebo-controlled α=0.5	+0.95 [0.82, 1.08]	+0.64 [0.53, 0.75]	70	0.957
TimeLens-8B, SFT	−0.21 [−0.24, −0.18]	−0.17 [−0.21, −0.13]	75	0.053

Table 12: **Keeping each prediction's centre and replacing only its length reproduces the largest accuracy gain in the audit matrix.** The predictions come unchanged from the Charades-STA run, and only each interval's length is replaced, either by a constant 8 s or by the length of the annotated interval. The columns give R@0.5 and mIoU, the mean IoU with the annotation, for the predictions as made (*as pred.*) and after the substitution (*subst.*).

Configuration	length substituted	R@0.5		mIoU	
		as pred.	subst.	as pred.	subst.
TimeLens-8B, placebo-controlled α=0.5	constant 8 s	0.626	0.649	0.540	0.519
TimeLens-8B, placebo-controlled α=0.5	length of the annotated interval	0.626	0.644	0.540	0.562
TimeLens-8B, base	constant 8 s	0.417	0.678	0.434	0.542
TimeLens-8B, base	length of the annotated interval	0.417	0.685	0.434	0.593

C TRAINING EVIDENCE DEPENDENCE

C.1 THE OBJECTIVE, TERM BY TERM

Figure 8 traces one training pair through Equations 3–4.

Components per operator. Table 13 decomposes the paired estimates of the trained models into d_E and d_O for trained model and base, per operator. The full objective moves d_E and leaves d_O at the base model's level; removing the bound, or the bound and the placebo contrast together, moves d_O instead. The Group-DRO weights cut d_O by a quarter to a third when the bound is removed. With the bound, the two weightings differ by 0.01 in d_O in the first seed: it is 0.02 with uniform weights (Table 2) and 0.01 with Group-DRO (Table 1). In a further pair of runs sharing one seed, Group-DRO (the run of Figure 9) keeps the placebo drop lower than uniform weights do, at 0.05

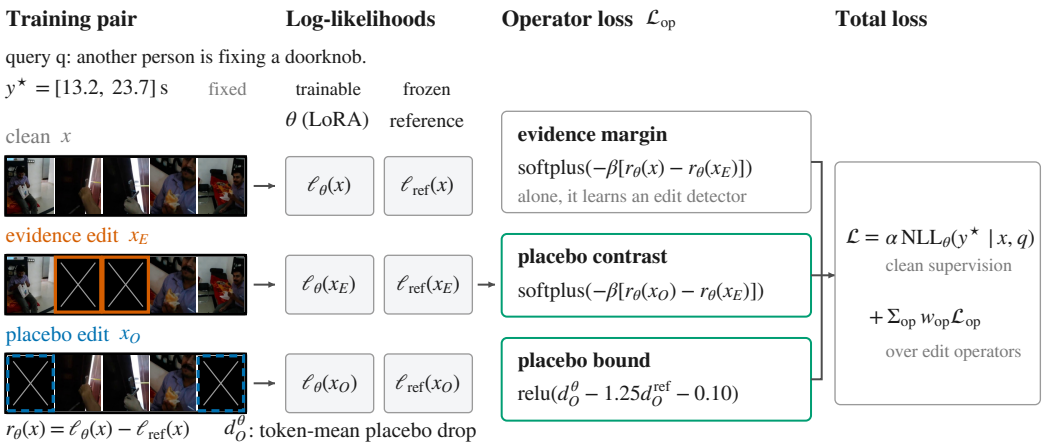

Figure 8: **The placebo-controlled objective scores each training pair on the clean video and on both edits.** The trainable model and the frozen base score the fixed annotated answer on the clean video x, the evidence edit x_E and the placebo edit x_O. Their log-likelihood differences r_θ enter the evidence margin and the placebo contrast, the placebo drops of both models enter the placebo bound, and the total loss adds clean supervision to the operator losses (Equations 3–4).

against 0.12 nats per token, with the base at -0.07. In that pair the Group-DRO run reaches Γ (mask) $+1.69$ and the uniform run $+1.97$ $[1.77, 2.19]$, with gains of 5.8 $[4.5, 7.0]$ and 6.5 $[5.1, 7.8]$ points; audited against each other, the two runs give a Γ (mask) of -0.31 $[-0.38, -0.24]$.

C.2 Where the placebo bound comes from

The form and constants of the bound in Equation 3 come from an earlier experiment in which every model saw 16 frames at 224 px under one prompt (the *16-frame setup*). In it, a Qwen2.5-VL-7B model trained with the two margins of Equation 3 and uniform weights but no placebo bound raised Γ (Holm-adjusted $p \leq 0.0004$ over the experiment's four tests) while its placebo drop rose from 0.01 to 3.22 nats per token, far above 1.25 times the base drop plus 0.10, and its R@0.5 fell from 0.260 to 0.203. Transfers of the same objective to the three public checkpoints repeated the pattern. Each raised Γ while its placebo drop rose to 0.46 (TimeZero), 4.37 (VideoChat-R1) and 2.40 (Time-R1) nats per token against limits near 0.1, and their accuracy changed by $+2.5$, $+7.2$ and $+12.5$ points. This limit became the bound of the objective. In the same setup an IoU-reward DPO model shifted the likelihood scale itself, its clean-video NLL rising from 0.72 to 2.27; the objective therefore measures the evidence and placebo terms against the frozen base.

C.3 Clean-supervision weight, base and data

Dose. Table 3 gives one configuration per base model, and Table 14 the α and β grids behind Figure 4. Within every base, Γ (mask) falls with each step in α; from the smallest to the largest weight it drops by 0.43 (VideoChat-R1) to 1.45 (Time-R1). The sign fraction is at least 0.73 for every placebo-controlled model trained at 2 frames per second, and at least 0.97 among those at β=0.1 and $\alpha \leq 0.5$. The two models trained on 16 frames (Training frames, below) have sign fractions of 0.69 and 0.83.

Budget. The Qwen2.5-VL-7B model trained on 4,800 pairs raises ψ on every question of its $H_\cap$ (sign fraction 1.00; Table 14). Its placebo drop is 0.11 nats per token, against -0.07 for its base (Table 13).

Training dynamics. Figure 9 audits checkpoints of a further run of the placebo-controlled Qwen2.5-VL-7B model at α=0.5 and of two SFT runs. Γ rises throughout the single pass and has not levelled off at its end ($+1.69$ $[1.45, 1.92]$ after 2,400 pairs), in line with the larger Γ after 4,800 pairs (Section 4); the trained model's placebo drop stays below 0.06 nats per token. SFT's Γ stays between -0.09 and -0.06 from the first checkpoint on.

Table 13: **Per edit operator, the placebo-controlled objective moves the evidence drop and keeps the placebo drop within 0.2 nats per token of the base model's; without the placebo bound the placebo drop rises.** For the mask and the swap operator, the columns give the evidence drop d_E and the placebo drop d_O of the trained model (tr) and of its base on the same questions, in nats per token, followed by Γ; under mask they also give the sign fraction, the fraction of questions on which ψ rises. Text-first rows put the query before the video.

	mask						swap				
Model / objective	d_E^{tr}	d_O^{tr}	d_E^{base}	d_O^{base}	Γ	sign frac.	d_E^{tr}	d_O^{tr}	d_E^{base}	d_O^{base}	Γ
Qwen2.5-VL-7B											
placebo-controlled α=0.25	2.29	0.01	0.18	−0.07	+2.03	1.00	1.98	0.03	0.05	−0.06	+1.84
placebo-controlled α=0.5, β=0.03	1.63	0.02	0.18	−0.07	+1.36	0.730	1.31	0.01	0.07	−0.07	+1.16
placebo-controlled α=0.5	1.82	0.01	0.12	−0.06	+1.63	0.973	1.55	0.04	0.06	−0.04	+1.41
placebo-controlled α=1, β=0.03	1.00	−0.01	0.13	−0.06	+0.82	0.829	0.82	0.00	0.06	−0.04	+0.72
placebo-controlled α=1	1.28	0.01	0.17	−0.06	+1.03	0.804	1.17	0.04	0.07	−0.06	+1.00
placebo-controlled α=0.5, 4,800 pairs	3.13	0.11	0.18	−0.07	+2.78	1.00	2.81	0.12	0.04	−0.06	+2.58
placebo-controlled α=0.5, 8,434 pairs, trained on 16 frames	0.37	−0.12	0.19	−0.06	+0.23	0.831	0.26	−0.10	0.05	−0.06	+0.24
placebo-controlled α=0.5, 2,400 pairs, trained on 16 frames	0.23	−0.07	0.18	−0.07	+0.06	0.689	0.12	−0.06	0.07	−0.06	+0.04
uniform weights	2.25	0.02	0.18	−0.07	+1.99	1.00	2.06	0.06	0.04	−0.06	+1.89
no placebo bound	2.80	0.49	0.19	−0.07	+2.05	1.00	2.77	0.30	0.06	−0.07	+2.34
no bound, uniform	3.52	0.66	0.18	−0.07	+2.62	1.00	3.17	0.45	0.05	−0.07	+2.60
IoU-GRPO	0.20	−0.05	0.18	−0.07	0.00	0.490	0.05	−0.07	0.04	−0.06	+0.02
mDPO-style	0.16	0.00	0.12	−0.06	−0.03	0.500	0.11	0.00	0.06	−0.04	+0.01
SFT	0.10	0.00	0.12	−0.07	−0.09	0.402	0.07	0.00	0.06	−0.04	−0.03
evidence margin only	2.44	1.65	0.18	−0.07	+0.54	0.740	2.59	1.39	0.05	−0.07	+1.09
two-stage (SFT, then placebo-controlled), α=0.5	2.38	0.04	0.19	−0.06	+2.10	1.00	1.96	0.06	0.06	−0.07	+1.77
two-stage (SFT, then placebo-controlled), β=0.03	2.99	−0.01	0.19	−0.06	+2.75	0.977	2.27	0.07	0.06	−0.06	+2.09
TimeZero											
placebo-controlled α=0.25	2.22	0.04	0.16	−0.02	+2.00	0.992	1.76	0.03	0.08	0.00	+1.66
placebo-controlled α=0.5, β=0.03	0.50	0.00	0.16	−0.02	+0.32	0.908	0.33	0.02	0.09	0.00	+0.23
placebo-controlled α=0.5, text-first	1.45	0.01	0.14	−0.01	+1.29	0.979	1.25	0.03	0.12	0.03	+1.13
placebo-controlled α=0.5	1.61	0.03	0.16	−0.02	+1.40	0.985	1.45	0.06	0.08	0.00	+1.31
placebo-controlled α=1	1.24	−0.01	0.15	−0.02	+1.07	0.931	1.03	0.03	0.09	0.00	+0.92
Time-R1											
placebo-controlled α=0.1	2.49	−0.18	0.18	−0.08	+2.40	1.00	2.19	−0.16	0.05	−0.09	+2.20
placebo-controlled α=0.25	2.07	−0.02	0.14	−0.07	+1.88	1.00	1.82	0.04	0.08	−0.04	+1.66
placebo-controlled α=0.5	1.72	0.00	0.19	−0.08	+1.46	0.976	1.34	0.00	0.05	−0.09	+1.20
placebo-controlled α=1	1.22	0.02	0.18	−0.08	+0.95	0.863	1.15	0.02	0.05	−0.09	+0.99
VideoChat-R1											
placebo-controlled α=0.25	1.78	−0.02	0.18	0.00	+1.63	0.993	1.08	0.07	0.11	0.01	+0.92
placebo-controlled α=0.5	1.60	−0.01	0.18	0.00	+1.43	1.00	1.01	0.01	0.12	0.01	+0.89
placebo-controlled α=1	1.39	0.00	0.19	0.01	+1.20	1.00	0.99	0.02	0.13	0.01	+0.85
TimeLens-8B											
public checkpoint against Qwen3-VL-8B	0.07	−0.02	0.22	−0.03	−0.16	0.294	0.01	−0.03	0.15	−0.03	−0.14
placebo-controlled α=0.25	1.79	−0.05	0.09	−0.03	+1.72	1.00	1.18	−0.04	0.02	−0.04	+1.15
placebo-controlled α=0.5	1.43	−0.03	0.08	−0.03	+1.36	1.00	0.90	−0.03	0.02	−0.04	+0.88
placebo-controlled α=1	1.00	−0.03	0.08	−0.03	+0.92	1.00	0.70	−0.02	0.02	−0.04	+0.66

Training frames. Dependence is learned on the frames the model is audited on. Two models trained in the 16-frame setup (Appendix C.2) and audited at 2 frames per second reach a Γ (mask) of $+0.06$ when trained on 2,400 pairs and $+0.23$ when trained on all 8,434 pairs, the latter with an R@0.5 5.7 points below its base; both are far below the models trained at 2 frames per second (Tables 13 and 7).

Two-stage training. Training Qwen2.5-VL-7B with SFT first and the objective second gives a Γ (mask) of $+2.10$ $[1.77, 2.41]$ at R@0.5 0.490, and with β=0.03 it gives $+2.75$ $[2.29, 3.22]$ at 0.531. The placebo drops of the two models under mask are 0.04 and -0.01 nats per token.

Table 14: **Within every base, evidence dependence falls monotonically as the clean-supervision weight rises.** Each row compares one trained model with its base. ΔR@0.5 is the change in points, and † marks a difference of the two accuracies that is not paired by question; Γ (mask) and Γ (swap) are in nats per token on the $n_\cap$ questions both models localise, and the sign fraction is the fraction of them on which ψ rises. Rows other than SFT use β=0.1 unless they name another value, and brackets give 95% intervals.

Setting	ΔR@0.5	Γ (mask)	Γ (swap)	$n_\cap$	sign frac.
Qwen2.5-VL-7B					
α=0.25	+2.1 [0.8, 3.2]	+2.03 [1.76, 2.29]	+1.84 [1.58, 2.10]	101	1.00
α=0.5	+4.4 [3.1, 5.7]	+1.63 [1.36, 1.89]	+1.41 [1.18, 1.63]	112	0.973
α=1	+6.7 [5.2, 8.2]	+1.03 [0.80, 1.27]	+1.00 [0.79, 1.23]	102	0.804
α=0.5, β=0.03	+1.2 [−0.6, 2.9]	+1.36 [0.99, 1.72]	+1.16 [0.81, 1.50]	89	0.730
SFT	+9.4 [7.9, 11.0]	−0.09 [−0.15, −0.02]	−0.03 [−0.08, 0.01]	112	0.402
α=0.5, 4,800 pairs	+8.1 [6.7, 9.5]	+2.78 [2.54, 3.03]	+2.58 [2.28, 2.86]	105	1.00
TimeZero					
α=0.25	+0.2†	+2.00 [1.73, 2.26]	+1.66 [1.40, 1.91]	133	0.992
α=0.5	+0.5 [−0.7, 1.5]	+1.40 [1.19, 1.60]	+1.31 [1.11, 1.50]	135	0.985
α=1	−0.6†	+1.07 [0.89, 1.25]	+0.92 [0.75, 1.08]	130	0.931
α=0.5, β=0.03	−1.0 [−2.1, 0.1]	+0.32 [0.26, 0.39]	+0.23 [0.18, 0.27]	131	0.908
Time-R1					
α=0.1	+0.3†	+2.40 [2.12, 2.67]	+2.20 [1.88, 2.50]	130	1.00
α=0.25	+1.6†	+1.88 [1.64, 2.09]	+1.66 [1.42, 1.87]	131	1.00
α=0.5	+0.9 [−0.2, 2.0]	+1.46 [1.23, 1.68]	+1.20 [1.00, 1.38]	127	0.976
α=1	−8.4†	+0.95 [0.72, 1.19]	+0.99 [0.75, 1.21]	102	0.863
VideoChat-R1					
α=0.25	−3.2 [−4.6, −1.8]	+1.63 [1.46, 1.79]	+0.92 [0.75, 1.08]	136	0.993
α=0.5	−2.7 [−4.2, −1.2]	+1.43 [1.27, 1.58]	+0.89 [0.74, 1.04]	132	1.00
α=1	−2.4 [−3.8, −0.9]	+1.20 [1.05, 1.34]	+0.85 [0.72, 1.00]	129	1.00
TimeLens-8B					
α=0.25	+17.9 [15.9, 20.0]	+1.72 [1.60, 1.85]	+1.15 [0.96, 1.32]	84	1.00
α=0.5	+20.9 [18.9, 22.8]	+1.36 [1.24, 1.47]	+0.88 [0.76, 1.01]	85	1.00
α=1	+23.2 [21.4, 25.1]	+0.92 [0.83, 1.01]	+0.66 [0.57, 0.76]	87	1.00

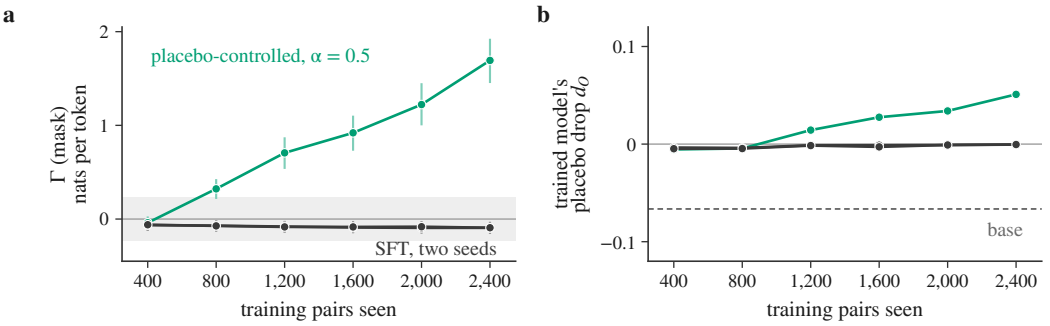

Figure 9: **Evidence dependence grows throughout training; SFT's stays within the audit's resolution from the start.** Checkpoints are taken every 50 steps (400 pairs) of one placebo-controlled run (α=0.5, β=0.1) and of two SFT runs from Qwen2.5-VL-7B, and each checkpoint is audited against the base. (a) The curves give Γ (mask) with 95% intervals, and the grey band marks the audit's resolution. (b) The curves give the trained model's placebo drop on the same questions, and the dashed line marks the base's.

D What the trained models do with the evidence

D.1 Unseen edit operators

Training uses five edit operators (Appendix E.1), among them the audit's mask and swap. Table 15 audits three placebo-controlled Qwen2.5-VL-7B models and two TimeZero models under additive noise, patch scrambling and Gaussian blur, none of which occurs in training. Fourteen of the fifteen intervals exclude zero, and all six values of the Qwen2.5-VL-7B models under noise and Gaussian blur, $+0.07$ to $+0.16$, lie within the mask resolution.

D.2 The partner event's query

Definitions. Take a question i of the 200 audit questions, with query q_i, annotated interval $y_i^\star$ and evidence frames E_i (event A of Section 5). Its partner j (event B) is another annotation of the same

Table 15: **Under edit operators never used in training, every placebo-controlled model keeps a positive Γ, large under patch scrambling and small under noise and Gaussian blur.** Each model is audited against its base on the same 200 questions. For each operator, the table gives Γ with its 95% interval on the $n_\cap$ questions both models localise, and the trained model's placebo drop d_O, both in nats per token.

		noise		patch scrambling		Gaussian blur	
Model / objective	$n_\cap$	Γ	d_O	Γ	d_O	Γ	d_O
Qwen2.5-VL-7B							
placebo-controlled α=0.5	100	+0.15 [0.07, 0.26]	−0.02	+1.52 [1.23, 1.82]	0.09	+0.16 [0.08, 0.25]	0.00
placebo-controlled α=0.5, β=0.03	88	+0.15 [0.02, 0.30]	0.00	+1.34 [1.00, 1.71]	0.03	+0.07 [0.01, 0.16]	0.00
placebo-controlled α=1	100	+0.07 [−0.01, 0.14]	−0.01	+1.12 [0.87, 1.38]	0.04	+0.08 [0.02, 0.14]	0.00
TimeZero							
placebo-controlled α=0.25	134	+0.28 [0.16, 0.40]	−0.03	+1.79 [1.55, 2.05]	0.04	+0.27 [0.16, 0.39]	−0.01
placebo-controlled α=0.5	134	+0.23 [0.13, 0.35]	0.00	+1.29 [1.11, 1.48]	0.04	+0.23 [0.14, 0.34]	0.00

video with a different sentence, whose frames E_j do not overlap E_i; among several, it is the one whose frames are centred farthest from E_i. Every quantity below scores the same answer $y_i^\star$ by teacher forcing, as the token-mean log-likelihood $\ell(x;q) = \log p(y_i^\star \mid x, q)$ under a query q, and uses the mask operator. With x_{E_i} and x_{E_j} the video with all frames of E_i or all frames of E_j masked, the *partner contrast* under query q is

$$\pi(q) = \big[\ell(x;q) - \ell(x_{E_i};q)\big] - \big[\ell(x;q) - \ell(x_{E_j};q)\big] = \ell(x_{E_j};q) - \ell(x_{E_i};q), \tag{5}$$

which is positive when masking the question's own event costs more than masking the partner's. We write $\psi_{\text{own}} = \pi(q_i)$ for the question's own query and $\psi_{\text{cross}} = \pi(q_j)$ for the partner's. Both are partner contrasts, not the placebo difference ψ of Section 2. Table 16 gives their trained-minus-base means and the difference of the two,

$$\Delta\psi_{\text{own}} = \frac{1}{n}\sum_i \big(\psi_{\text{own},i}^{\text{trained}} - \psi_{\text{own},i}^{\text{base}}\big), \qquad \Gamma_{\text{spec}} = \Delta\psi_{\text{own}} - \Delta\psi_{\text{cross}}, \tag{6}$$

with $\Delta\psi_{\text{cross}}$ defined in the same way, over the n questions that both models localise (IoU ≥ 0.5 on the clean video) and that have a partner. The sign fraction is the fraction of these questions on which the per-question term of Γ_{spec} is positive. A model that reads the query raises ψ_{own} and lowers ψ_{cross}, so $\Delta\psi_{\text{cross}} < 0$ and $\Gamma_{\text{spec}} > \Delta\psi_{\text{own}}$; a model that responds only to where the edit lands changes both alike, so $\Gamma_{\text{spec}} \approx 0$. The right block of the table replaces the partner edit with the placebo. There $\psi(q) = d_E(q) - d_O(q)$ is the difference of Section 2 under query q, with the placebo frames drawn from outside both E_i and E_j so that the placebo never touches the partner event. Γ is the trained-minus-base mean of $\psi(q_i)$, and Γ_{agn} is that of $\psi(q_j)$: the part of Γ that remains when the query names another event. The rest, $\Gamma_{\text{sem}} = \Gamma - \Gamma_{\text{agn}}$, is the part that depends on the question's own query. Γ_{spec} and Γ_{sem} thus measure the query-dependent part against two different controls, the partner event's frames and the placebo frames. Figure 5a plots $\Delta\psi_{\text{own}}$ (horizontal) against $\Delta\psi_{\text{cross}}$ (vertical). Because of its placebo pool, this block leaves out the questions with fewer frames outside $E_i \cup E_j$ than in E_i, and its Γ differs from the Γ of the other tables.

An edit inside the annotated interval changes two things: the frames the query describes, and the frames at the timestamps the answer names. Figure 10 and Table 16 separate them with the partner contrast, holding the scored answer $y_i^\star$ and the edited frames fixed and changing only the query. Three accounts predict different contrasts under q_j. A model that reads the query should now be hurt more by the partner edit, which reverses the contrast. A model that detects damage at the answer's timestamps gives the same contrast under both queries. A low likelihood under q_j shrinks the contrast toward zero, but only reading the query reverses it. For every placebo-controlled Qwen and Time-R1 model with a crossed run, for video-first TimeZero and for the evidence-margin-only model, the crossed contrast reverses. In the three VideoChat-R1 models it shrinks but stays positive, and text-first TimeZero sits near zero. Γ_{spec} is positive in every placebo-controlled model ($+0.5$ to $+4.0$; sign fractions 0.79–1.00) and zero for SFT (-0.04 $[-0.18, 0.09]$). The query-agnostic part Γ_{agn} is at most 0.28 in absolute value for every placebo-controlled Qwen and TimeZero model, and about a quarter to a third of Γ for VideoChat-R1, whose models respond to the edit's position as well as to the query.

Reworded queries. Re-asking each question through a second human-written description of the same interval retains the contrast (0.98 $[0.94, 1.01]$ for mask, 1.02 $[0.96, 1.09]$ for swap, n=132; the interval of each original-against-paraphrase difference contains zero), while a different event's query removes it.

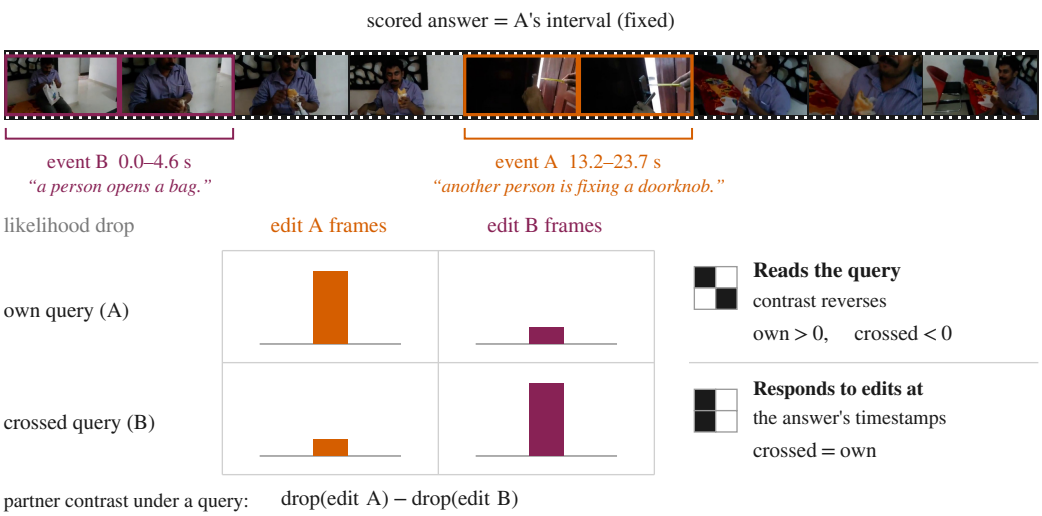

Figure 10: **The query-crossed design scores the same answer under the question's own query and under a partner event's query.** Both conditions score the annotated interval of event A and use the same two edits: masking A's frames, or masking the frames of a partner event B of the same video. Only the query changes, from A's own to B's. A model that reads the query shows a contrast that reverses under B's query; a model that responds to edits at the answer's timestamps shows the same contrast under both queries. The bars are schematic; Figure 5a and Table 16 give the measurements.

Table 16: **The audited contrast moves with the query: Γ_{spec} is positive in every placebo-controlled model and zero in SFT.** All values use the mask operator and are in nats per token; Appendix D.2 defines them. The left block uses the n questions that both models localise and that have a partner event, another annotated event of the same video. The partner contrast scores the question's own interval and is the likelihood drop when its event's frames are masked minus the drop when the partner event's frames are masked. The left block gives the trained-minus-base change of this contrast under the question's own query ($\Delta\psi_{\text{own}}$) and under the partner event's query ($\Delta\psi_{\text{cross}}$), and $\Gamma_{\text{spec}} = \Delta\psi_{\text{own}} - \Delta\psi_{\text{cross}}$ with the fraction of questions on which it is positive. The right block draws the placebo from outside both events, which leaves out 8 to 11 of these questions per row, and splits Γ into the part that remains under the partner event's query (Γ_{agn}) and the rest ($\Gamma_{\text{sem}} = \Gamma - \Gamma_{\text{agn}}$). Brackets give 95% intervals, and text-first rows put the query before the video.

		Query crossed				Placebo pool split		
	n	$\Delta\psi_{\text{own}}$	$\Delta\psi_{\text{cross}}$	Γ_{spec}	sign frac.	Γ	Γ_{agn}	Γ_{sem}
Qwen2.5-VL-7B								
placebo-controlled α=0.25	51	+1.75	−1.31	+3.07 [2.30, 3.75]	1.00	+1.52	+0.24 [0.12, 0.45]	+1.29
placebo-controlled α=0.5	46	+1.59	−1.27	+2.86 [1.99, 3.63]	0.96	+1.21	+0.20 [0.04, 0.41]	+1.01
placebo-controlled α=0.5, β=0.03	43	+1.34	−1.49	+2.83 [1.89, 3.70]	0.91	+0.86	0.00 [−0.12, 0.15]	+0.87
placebo-controlled α=1	48	+1.00	−1.15	+2.15 [1.40, 2.73]	0.92	+0.65	−0.03 [−0.15, 0.11]	+0.69
placebo-controlled α=1 (second seed)	48	+1.06	−1.04	+2.10 [1.47, 2.62]	0.98	+0.70	−0.01 [−0.10, 0.12]	+0.71
two-stage, α=0.5	42	+2.10	−0.88	+2.98 [2.30, 3.60]	1.00	+1.74	+0.19 [−0.01, 0.44]	+1.56
two-stage, β=0.03	44	+2.64	−1.35	+3.99 [3.04, 4.90]	0.98	+2.18	+0.28 [0.03, 0.57]	+1.90
evidence margin only	49	+0.99	−1.01	+2.00 [1.22, 2.61]	0.94	+0.46	−0.35 [−0.76, 0.20]	+0.81
SFT	46	−0.11	−0.06	−0.04 [−0.18, 0.09]	0.59	−0.19	−0.11 [−0.21, −0.03]	−0.07
TimeZero								
placebo-controlled α=0.25	59	+1.88	−1.27	+3.15 [2.40, 3.86]	0.98	+1.70	+0.25 [0.13, 0.41]	+1.44
placebo-controlled α=0.5	59	+1.37	−0.77	+2.14 [1.64, 2.64]	1.00	+1.16	+0.13 [0.02, 0.29]	+1.04
placebo-controlled α=0.5, text-first	58	+0.56	+0.03	+0.54 [0.33, 0.71]	0.88	+0.56	+0.13 [0.06, 0.21]	+0.43
Time-R1								
placebo-controlled α=0.25	56	+1.64	−1.20	+2.84 [2.21, 3.41]	0.98	+1.41	+0.20 [0.07, 0.35]	+1.21
VideoChat-R1								
placebo-controlled α=0.25	58	+1.62	+0.69	+0.93 [0.61, 1.18]	0.81	+1.65	+0.49 [0.20, 0.77]	+1.16
placebo-controlled α=0.5	56	+1.40	+0.67	+0.73 [0.47, 0.95]	0.84	+1.42	+0.49 [0.20, 0.76]	+0.93
placebo-controlled α=0.5, text-first	62	+0.93	+0.35	+0.58 [0.34, 0.80]	0.79	+1.03	+0.25 [0.14, 0.37]	+0.78

D.3 A SECOND DATASET

QVHighlights. Table 17 and Figure 11 evaluate the Charades-trained configurations on QVHighlights, scored against the TimeLens re-annotation for both accuracy and the audit.

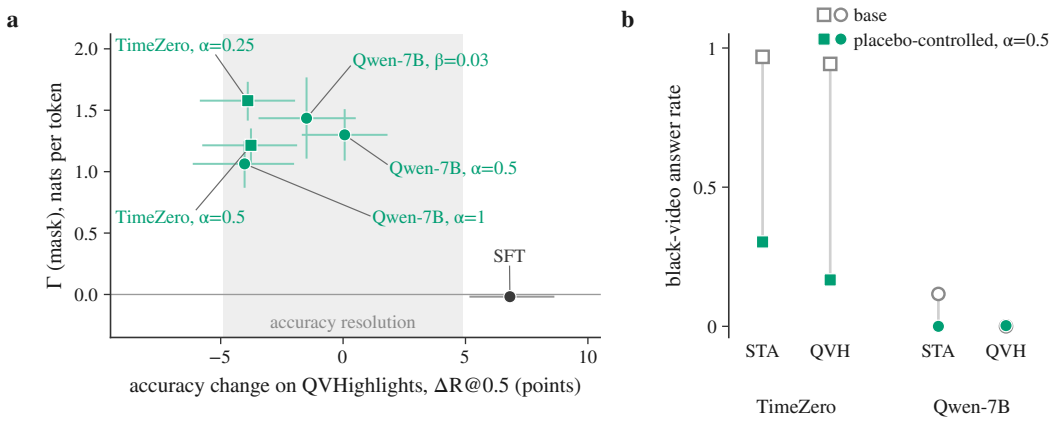

Figure 11: **On QVHighlights, the placebo-controlled models keep a positive Γ (mask) while SFT stays at zero, and the placebo-controlled TimeZero model answers far fewer black videos than its base.** (a) Each Charades-trained model is placed by its paired change in R@0.5 on QVHighlights (horizontal) and its Γ (mask) (vertical), with 95% intervals; the grey band marks the out-of-domain accuracy resolution (Appendix A.1). *Qwen-7B* is Qwen2.5-VL-7B. Green points are placebo-controlled models, trained at β=0.1 unless the label gives β; the β=0.03 model uses α=0.5. (b) For each base model on Charades-STA (STA) and QVHighlights (QVH), the black-video answer rate of the base (open) is set against that of its placebo-controlled model at α=0.5 (filled); squares are TimeZero and circles Qwen-7B.

Table 17: **On QVHighlights, every placebo-controlled model keeps a positive Γ (mask) and SFT stays at zero.** All models are trained on Charades-STA and evaluated on QVHighlights with the TimeLens re-annotation (1,541 questions). ΔR@0.5 is the paired change from the model's base in points, and Γ (mask) and Γ (swap) are in nats per token on the $n_\cap$ questions both models localise. The column *ord.* gives the message order, video first (v) or query text first (t); *parse* is the share of outputs that parse to an interval, and *abst.* is the strict abstention rate, the share of outputs that are verbatim `There are none.`

Model / objective	ord.	R@0.5	parse	abst.	ΔR@0.5	Γ (mask)	Γ (swap)	$n_\cap$
Qwen2.5-VL-7B								
base	v	0.319	0.992	0.008	—	—	—	—
base, text-first	t	0.097	0.925	0.075	—	—	—	—
placebo-controlled α=0.5	v	0.319	0.988	0.012	+0.1 [−1.7, 1.8]	+1.30 [1.09, 1.51]	+1.03	43
placebo-controlled α=1	v	0.278	0.977	0.023	−4.0 [−6.1, −2.0]	+1.06 [0.87, 1.25]	+1.08	39
placebo-controlled α=0.5, β=0.03	v	0.304	0.698	0.302	−1.5 [−3.5, 0.5]	+1.44 [1.11, 1.77]	+1.28	40
SFT	v	0.387	1.00	0.000	+6.8 [5.2, 8.6]	−0.02 [−0.07, 0.03]	+0.02	52
TimeZero								
base	v	0.518	1.00	0.000	—	—	—	—
base, text-first	t	0.379	1.00	0.000	—	—	—	—
placebo-controlled α=0.25	v	0.479	1.00	0.000	−3.9 [−5.9, −2.0]	+1.58 [1.42, 1.73]	+1.52	82
placebo-controlled α=0.5	v	0.480	0.999	0.001	−3.8 [−5.8, −1.9]	+1.21 [1.07, 1.35]	+1.10	78

A re-split of Charades-STA. Charades-CD (Yuan et al., 2021) re-splits the same benchmark so that the test-ood moment distribution no longer matches training; the constant window falls from 0.294 on test-iid to 0.150 on test-ood (Table 18). The test sets partly reuse the original training split, on which TimeZero and VideoChat-R1 were trained, so we compare each model's change from iid to ood. In each of the four bases that have an SFT model, SFT loses the most.

D.4 Generated answers

Evidence share and follows. The evidence share is computed on the first 1,000 test questions. Each model regenerates its answer on the clean video, with the evidence frames edited, and with as many frames outside the annotated interval edited (one placebo draw per question, the same for every model), under mask and under swap. With a_E and a_O the R@0.5 of the evidence-edited and the placebo-edited runs, the share is $1 - a_E/a_O$, a ratio of two accuracies over all 1,000 questions rather than a mean over questions. Follows is computed on the likelihood audit's 200 questions. For each question a model localises, the block of evidence frames is exchanged with a block of as many consecutive frames that does not overlap it, placed at random, and the model regenerates its answer on the exchanged video. The answer follows the event if its IoU with the annotated interval, shifted by the offset between the two blocks and clipped to the video, is at least 0.5; questions with no room

Table 18: **Across the shift to the Charades-CD test-ood split, SFT loses more accuracy than the other objectives on Qwen2.5-VL-7B.** The table uses the test-iid (823 questions, 333 videos) and test-ood (3,375 questions, 1,442 videos) splits of Charades-CD (Yuan et al., 2021). The re-split shifts the moment distribution, so the constant ten-second window falls from iid to ood; the column *ood−iid* is each model's change across that shift, in points. On Qwen2.5-VL-7B, SFT's change is lower than IoU-GRPO's by 4.3 $[0.8, 7.5]$ points and lower than the placebo-controlled model's by 5.5 $[1.7, 9.1]$; questions are paired within each split, and the 95% intervals resample videos. Text-first rows put the query before the video.

Model / objective	R@0.5 iid	R@0.5 ood	ood−iid
constant $(0, 10)$ s window (no model)	0.294	0.150	−14.4
Qwen2.5-VL-7B			
base	0.480	0.449	−3.1
SFT	0.622	0.524	−9.8
IoU-GRPO	0.635	0.580	−5.5
placebo-controlled α=0.5	0.498	0.455	−4.3
TimeZero			
base	0.710	0.643	−6.7
base, text-first	0.744	0.663	−8.1
SFT, text-first	0.727	0.628	−9.9
placebo-controlled α=0.5	0.705	0.626	−7.9
Time-R1			
base	0.612	0.591	−2.2
placebo-controlled α=0.5	0.628	0.581	−4.7
placebo-controlled α=0.1	0.618	0.585	−3.3
VideoChat-R1			
base	0.708	0.640	−6.8
base, text-first	0.761	0.684	−7.6
SFT, text-first	0.759	0.659	−10.0
placebo-controlled α=1	0.689	0.596	−9.3
TimeLens-8B			
base	0.413	0.423	+1.0
SFT	0.629	0.577	−5.2
placebo-controlled α=0.5	0.617	0.596	−2.1

for such a block are left out. For both readouts the tables report each trained model's difference from its base in points, the evidence share over all 1,000 questions and follows over $H_{\cap}$, with 95% intervals from a bootstrap over videos.

Regenerated predictions. Table 19 gives the full readouts on generated answers behind Section 5, with the same 1,000 questions regenerated under clean, evidence-edited and placebo-edited versions of each operator. With as many frames masked outside the interval, every placebo-controlled model stays at or above its clean accuracy; masking the evidence frames removes most of it.

Public checkpoints. We also regenerated TimeZero and Time-R1 with their full reasoning, masking either the evidence or the placebo frames. For both, the difference in evidence share from the Qwen base has an interval that includes zero, while every placebo-controlled model of these bases raises the evidence share over its base (Tables 1 and 19).

D.5 Where the dependence sits

Table 20 splits the answers lost to the evidence mask into refusals and wrong intervals, and Γ over the onset, timestamp and tag tokens of the scored answer. Among the tagged templates, TimeZero's placebo-controlled model places the most Γ on the timestamps, still far less than VideoChat-R1 with its bare interval. On the clean video the placebo-controlled models of the tagged templates write the onset token with probability 0.63 to 0.96. SFT writes it with probability 1.00 and places no Γ on it.

Table 19: **In regenerated answers, every placebo-controlled model loses a larger share of its correct answers to evidence edits than its base does.** Under the mask, SFT and IoU-GRPO lose a smaller share. Each block compares a trained model with its base on the same 1,000 questions, edited frames and placebo draws. The first six columns give R@0.5 on the clean video, with the evidence frames (E) or as many frames outside the interval (O) masked or swapped, and on all-black versions of these questions. The Δ columns give trained-minus-base changes in points in two readouts (Appendix D.4). The evidence share is $1 - a_E/a_O$, with a_E and a_O the R@0.5 under the evidence edit and under the placebo edit, and is given under swap and under mask. *Follows* is the fraction of the audit questions both models localise on which the prediction, regenerated after the event's frames are exchanged with another block, has IoU ≥ 0.5 with the moved interval; Table 1 gives it for the Qwen2.5-VL-7B models. Table 10 gives answering and refusal on black videos over the full test split. Brackets give 95% intervals.

	clean	mask E	mask O	swap E	swap O	black	Δshare swap	Δshare mask	Δfollows
Qwen2.5-VL-7B, placebo-controlled α=0.5 (video-first)									
base	0.489	0.170	0.543	0.230	0.551	0.006	—	—	—
trained	0.523	0.087	0.558	0.182	0.556	0.000	+9.0 [5.2, 12.9]	+15.7 [11.4, 20.3]	—
Qwen2.5-VL-7B, IoU-GRPO (video-first)									
base	0.489	0.170	0.543	0.230	0.551	0.006	—	—	—
trained	0.637	0.250	0.668	0.278	0.685	0.047	+1.2 [−2.9, 5.5]	−6.1 [−11.0, −1.2]	—
Qwen2.5-VL-7B, mDPO-style (video-first)									
base	0.489	0.170	0.543	0.230	0.551	0.006	—	—	—
trained	0.596	0.190	0.630	0.217	0.629	0.005	+7.2 [3.7, 10.9]	+1.1 [−3.3, 5.7]	—
Qwen2.5-VL-7B, SFT (video-first)									
base	0.489	0.170	0.543	0.230	0.551	0.006	—	—	—
trained	0.612	0.285	0.620	0.267	0.621	0.091	−1.3 [−5.3, 3.0]	−14.7 [−19.8, −9.4]	—
TimeZero, placebo-controlled α=0.5, trained video-first, evaluated text-first									
base	0.704	0.228	0.736	0.200	0.735	0.198	—	—	—
trained	0.694	0.181	0.735	0.171	0.713	0.108	+3.2 [0.8, 5.8]	+6.4 [3.3, 9.5]	+3.1 [−3.9, 9.6]
TimeZero, placebo-controlled α=0.5 (video-first)									
base	0.700	0.218	0.709	0.262	0.721	0.143	—	—	—
trained	0.690	0.145	0.726	0.213	0.724	0.049	+6.9 [3.9, 10.0]	+10.8 [7.5, 14.2]	—
Time-R1, placebo-controlled α=0.25 (video-first)									
base	0.616	0.232	0.650	0.287	0.648	0.093	—	—	—
trained	0.614	0.188	0.671	0.266	0.657	0.043	+3.8 [0.9, 6.8]	+7.7 [4.8, 10.8]	−3.2 [−10.4, 3.9]
VideoChat-R1, placebo-controlled α=0.5, trained video-first, evaluated text-first									
base	0.714	0.178	0.739	0.197	0.723	0.222	—	—	—
trained	0.679	0.050	0.741	0.146	0.706	0.200	+6.6 [4.2, 9.1]	+17.3 [13.7, 21.0]	−2.3 [−10.4, 5.9]
VideoChat-R1, placebo-controlled α=0.5 (video-first)									
base	0.687	0.231	0.707	0.273	0.691	0.203	—	—	—
trained	0.661	0.026	0.730	0.154	0.681	0.116	+16.9 [13.4, 20.3]	+29.1 [24.8, 33.7]	—

D.6 Queries about another video

Table 21 pairs, question by question and for eleven configurations of three bases, abstention on the matched query with abstention on the annotation sentence of a different video. None of the four TimeZero configurations abstains on either query, and each answers mismatched queries at the accuracy of a constant window.

E Reproduction

E.1 Protocols and training setup

Inference protocols. Every model generates at 2 frames per second with greedy decoding. TimeLens-8B, Qwen3-VL-8B and our adapters of TimeLens-8B generate with TimeLens's released pixel budget of $14{,}336 \times 32 \times 32$ and are audited at $3{,}584 \times 28 \times 28$; every other model uses $3{,}584 \times 28 \times 28$ in both. Against Qwen2.5-VL-7B (Section 3), TimeZero and Time-R1 run with the Qwen base under Time-R1's prompt template with the video before the text, and VideoChat-R1 with the Qwen base under VideoChat-R1's own template. When a public checkpoint is the base of a trained model (Sections 4 and 5), both run under the checkpoint's released prompt text with the video placed first; rows marked text-first use the order of TimeZero's and VideoChat-R1's released code. TimeLens-8B uses its own. The scored answer string follows the answer format of each template: under Time-R1's template it is `<answer> 20.80 to 30.00 </answer>`, as in our training targets. Appendix E.2 evaluates every base under both orders. Abstention is recorded separately from parse failure: the strict criterion counts only the verbatim string `There are none.`, and the wide criterion counts any refusal wording.

Table 20: **Evidence dependence shows in whether a model answers and where it places the interval.** Each entry is a trained model minus its base, with the evidence frames masked. The left block starts from the regenerated answers that are correct under the placebo mask (1,000 questions) and gives the change in the share that returns no interval and, among the intervals returned, in the share that is wrong. The right block splits Γ over the tokens of the scored answer, using a separate token-level audit of 200 questions; the *onset* is the first token after the opening tag. Bold marks a change whose 95% interval excludes zero.

	Lost answers become (points)		Where Γ sits (nats per token)		
Model	no interval	wrong interval	onset	timestamps	tags
Trained for accuracy, against the Qwen2.5-VL-7B base					
SFT	**−14.9*** [−18.5, −11.6]	**−8.7** [−13.2, −4.3]	−0.01	−0.01	−0.12
TimeZero (public)	**−14.9*** [−18.5, −11.6]	+4.6 [−0.5, 9.7]	−0.02	+0.04	−0.09
Time-R1 (public)	**−12.1** [−15.3, −9.0]	−0.6 [−5.6, 4.5]	+0.03	+0.01	−0.02
VideoChat-R1 (public)	n/a	−0.1 [−7.2, 7.2]	n/a	+0.06	n/a
Placebo-controlled, against the base it was trained from					
Qwen2.5-VL-7B, α=0.5	**+21.5** [16.7, 26.3]	**+10.4** [5.4, 15.4]	+0.79	+0.07	+0.86
TimeZero, α=0.5‡	**+0.8** [0.1, 1.6]	**+6.6** [3.7, 9.7]	+0.44	+0.34	+0.62
Time-R1, α=0.25	**+3.2** [1.4, 5.1]	**+6.6** [3.7, 9.5]	+1.06	+0.07	+0.83
VideoChat-R1, α=0.5‡	n/a	**+15.0** [11.8, 18.4]	n/a	+1.40	n/a

*Both models return an interval for every question under the mask, so both changes equal minus the base's rate. ‡The left block is evaluated with the text before the video, as released, and the right block with the video first. The entries marked n/a belong to VideoChat-R1, whose template fixes its tags and the bracket after the opening tag, leaving only the timestamps free, and has no way to decline. Brackets give 95% intervals over videos.

Table 21: **Given the query of another video, the placebo-controlled Qwen2.5-VL-7B and Time-R1 models decline more often than their bases, and no TimeZero configuration declines.** Each question is asked once with its own annotation sentence as the query and once with the sentence of another video, which gives n pairs. *abst.* is the strict abstention rate, the share of outputs that are verbatim `There are none.`, and R@0.5 is measured on the same pairs. Each Δ is the foreign-query value minus the own-query value, in points, with its 95% interval. Text-first rows put the query before the video.

Model / objective	n	abst. own	abst. foreign	Δ abst.	R@0.5 own	ΔR@0.5
Qwen2.5-VL-7B, base	1000	0.003	0.119	+11.6 [7.8, 15.9]	0.485	−33.0 [−37.8, −28.2]
Qwen2.5-VL-7B, placebo-controlled α=0.5, β=0.03	1000	0.164	0.827	+66.3 [60.7, 71.4]	0.528	−47.1 [−51.4, −42.7]
Qwen2.5-VL-7B, placebo-controlled α=0.5	1000	0.052	0.210	+15.8 [11.1, 20.7]	0.523	−40.5 [−44.8, −36.1]
Qwen2.5-VL-7B, SFT	1000	0.000	0.003	+0.3 [0.0, 0.8]	0.599	−32.3 [−37.3, −27.7]
TimeZero, base	1000	0.000	0.000	+0.0 [0.0, 0.0]	0.694	−38.2 [−43.1, −33.4]
TimeZero, placebo-controlled α=0.5, β=0.03	1000	0.000	0.000	+0.0 [0.0, 0.0]	0.687	−38.2 [−43.1, −33.6]
TimeZero, placebo-controlled α=1	1000	0.000	0.000	+0.0 [0.0, 0.0]	0.688	−38.7 [−43.8, −33.7]
TimeZero, SFT, text-first	1000	0.000	0.000	+0.0 [0.0, 0.0]	0.680	−41.1 [−46.0, −36.2]
Time-R1, base	1000	0.000	0.050	+5.0 [2.5, 8.0]	0.616	−37.3 [−42.2, −32.4]
Time-R1, placebo-controlled α=0.5	1000	0.018	0.404	+38.6 [32.6, 44.9]	0.637	−48.4 [−52.9, −43.8]
Time-R1, placebo-controlled α=1	1000	0.194	0.857	+66.3 [60.5, 71.6]	0.542	−49.2 [−53.7, −44.7]

Training. All trained models are LoRA adapters (Hu et al., 2022). All except IoU-reward GRPO have rank 16 (LoRA alpha 32, dropout 0.05) on the attention and MLP projections and are trained with AdamW at a constant learning rate of 5×10^{-6} and eight pairs per update. These models, apart from the budget, frame and two-stage runs of Appendix C.3, make one pass over the same 2,400 pairs from the Charades-STA training split (8,434 in the full pool), with the same trainer and schedule. Each pair is edited by two of five operators, chosen among those valid for it, least-used first. Mask replaces the selected frames with black, and swap replaces them with frames from another training video. Pixelation averages each selected frame over a 16×16 grid of blocks, and temporal shift replaces each selected frame with the one half the video away. Temporal shuffle permutes the selected frames, never as the identity; it is used only where the query or the pair makes the order of frames meaningful and at least two frames are selected. The operator weights w_{op} are Group-DRO-style weights (Sagawa et al., 2020) over nine groups that cross the operator with whether the query makes the order of frames meaningful (a state change or a temporal order, against a generic action), normalised over the pair's two operators and updated with step size 0.05 on each group's loss relative to its running mean.

Baselines and cost. The SFT model is the same trainer with the evidence and placebo terms disabled ($\beta \to 0$, no bound). The IoU-reward GRPO model follows Shao et al. (2024) with a format and an IoU reward. It uses LoRA rank 64 and a KL weight of 0.04 and runs 700 steps of four questions with eight sampled answers each, on 16 frames at 224 px; the questions are drawn from the 12,383 Charades-STA training questions whose annotated window holds at least one of the 16 sampled frames (of the split's 12,408). Our IoU-GRPO trains on 16 frames per video, not at the audit's 2 frames per second, and a model learns dependence only on the frames it trains on (Training frames, Appendix C.3); the public checkpoints therefore stand for IoU-reward training in Section 3. The mDPO-style baseline follows the conditional preference of Wang et al. (2024): the evidence margin of Equation 3, with the edited frames a contiguous quarter of the video placed at random without the annotation, at the same α=0.5, β=0.1, operators and schedule, with uniform weights and no placebo terms. Each pair of the placebo-controlled objective takes ten forward passes, in which the trained and the frozen model each score the clean video and the evidence and placebo edits of two operators; our SFT, run in the same trainer, scores the clean video and the two evidence edits under both models (six passes). On the same base, a pass of SFT over the 2,400 pairs takes 0.66 to 0.68 of the time of a pass of the placebo-controlled objective at α=1 (TimeZero, VideoChat-R1 and TimeLens-8B).

E.2 Message order

Table 22 evaluates every base under both orders, with its answering rate and accuracy on black videos. Message order moves accuracy in different directions for different bases: text-first moves TimeZero and VideoChat-R1 up by 2.8 points each, Time-R1 down by 6.9 and Qwen2.5-VL-7B down by 24.1. Part of TimeZero's $+2.8$ comes from the prior: its black-video R@0.5 rises from 0.136 to 0.189. For a trained adapter the order changes R@0.5 by $+1.0$ $[-3.9, 5.8]$ points on 200 questions.

Table 22: **Message order moves base models in family-specific directions, and it also moves their accuracy on a black video.** Each base model is evaluated under both message orders, with the video first (video-first) or the query text first (text-first), and everything else identical. ΔR@0.5 is text-first minus video-first, paired per question, in points, with its 95% interval; *black ans.* and *black R@0.5* are the answering rate and the accuracy on all-black videos under the same order. A $^\circ$ marks the order the checkpoint ships with.

Model	order	R@0.5	ΔR@0.5	black ans.	black R@0.5
Qwen2.5-VL-7B	video-first$^\circ$	0.480	—	0.116	0.007
	text-first	0.238	−24.1 [−26.5, −21.9]	0.578	0.092
TimeZero	video-first	0.677	—	0.968	0.136
	text-first$^\circ$	0.705	+2.8 [1.3, 4.1]	0.940	0.189
Time-R1	video-first$^\circ$	0.601	—	0.796	0.105
	text-first	0.533	−6.9 [−8.8, −4.8]	0.762	0.131
VideoChat-R1	video-first	0.668	—	1.000	0.153
	text-first$^\circ$	0.697	+2.8 [1.2, 4.5]	0.948	0.211

E.3 Release

The supplementary material contains the trainer and the evaluation and scoring code; the ids of the 8,434-pair pool (the 2,400 training pairs are its first 2,400), the 200-question audit set, the 1,000-question generation set, the partner-event pairs, the reworded queries and the mismatched-query pairing; the placebo draws of the behavioural audit; and the scripts that produce every table and figure.