SHORT-MOMENT GROUNDING IN VIDEOLLMS: A SEARCH-SPACE LIMIT RATHER THAN A CAPABILITY LIMIT

Anonymous authors
Paper under double-blind review

ABSTRACT

Video large language models (VideoLLMs) have become the dominant approach to temporal grounding, and their progress is measured by averages over entire benchmarks. We find that these averages hide a near-total failure on short moments, which are rare in standard test sets but often matter most in practice. The failure appears in every model and dataset we examine, and we trace it to two sources: a bias toward predicting spans of a typical length and an imprecise estimate of where the moment lies. The bias is already present before any grounding training, and both sources persist across a wide range of fine-tuning interventions. The same frozen model, however, localizes short moments precisely once its input is narrowed around them, which suggests that the failure lies in how much of the video the model must search rather than in what it can perceive. We turn this finding into a training-free temporal zoom that gives a second, focused look only to moments the model itself predicts to be short, and it substantially improves short-moment grounding. Our results show that VideoLLMs can already do more than their average scores suggest, and that the cheapest progress lies at inference time.

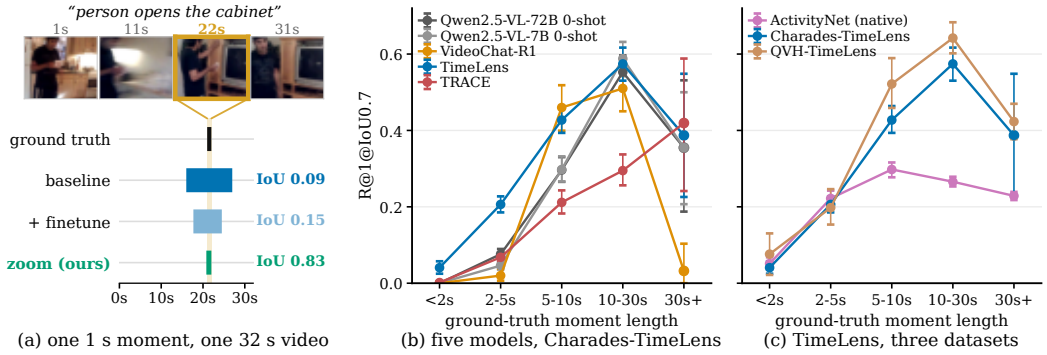
1 INTRODUCTION

Temporal grounding asks a model to return the start-end interval of the event described by a language query, and it underlies how large video collections are searched, moderated, and indexed. The task was long addressed by specialized architectures that score a fixed set of candidate proposals (Zhang et al., 2020; Moon et al., 2026; Mu et al., 2024; Lin et al., 2023). More recently, video large language models have become the dominant approach, emitting the interval directly as text and reaching strong results across the standard benchmarks (Huang et al., 2024a; Ren et al., 2024; Huang et al., 2024b; Zhang et al., 2026b; Li et al., 2025; Guo et al., 2025). Progress is measured by averaging a temporal overlap criterion over a benchmark test split (Gao et al., 2017; Krishna et al., 2017; Lei et al., 2021).

Yet an average over a test split says nothing about which moments a model fails on, and the moments that matter in deployment are often the shortest: a brief unsafe action in a moderation queue, a door opening in surveillance footage, a single gesture in an hour of meeting video. Such moments make up a small fraction of any test set, so a model can report an excellent benchmark average while failing almost completely on them. The most widely used benchmark cannot even reveal this failure, since the native Charades-STA test split (Gao et al., 2017) contains no moment shorter than two seconds. A precise re-annotation of the same videos (Zhang et al., 2026b) includes short moments, but there are too few to move the average by more than a few points. To our knowledge, no VideoLLM evaluation has asked what this leaves open, and we set the goal to answer it: *does a model's accuracy depend on how long the target moment is?*

We reach this conclusion by ruling out the obvious explanations one at a time, and the collapse survives changes to the overlap metric, the annotation boundaries, the queries and videos that accompany short moments, the supply of short training data, and the training itself. Two causes remain that training does not remove. The models pull every predicted span toward a typical few-second length, a bias that the pretrained backbone already carries and that grounding finetuning shifts rather than removes, and they place the moment center about a second away from the truth, an error that

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.



(a) one 1 s moment, one 32 s video (b) five models, Charades-TimeLens (c) TimeLens, three datasets

Figure 1: **The short-moment collapse, at one sample and across the benchmark. (a)** A 1 s moment in a 32 s video: each method's predicted span against ground truth, with its IoU (Secs. 4–5). **(b, c)** Strict recall (R@1@IoU0.7) by ground-truth moment length, across models and across datasets. Per-stratum counts are in Table 1. Error bars: 95% cluster bootstrap by video; hollow markers: $n < 50$.

no finetuning we try brings below a second. Training on genuinely short moments changes which lengths a model predicts far more than where it places them. The same frozen model, however, recovers a sub-second center once the video is cropped to a narrower window, so the failure reflects a limit of the search space rather than a missing capability, and we turn this into a lever that changes no weights and holds across two grounders.

The account rests on evaluations spanning three model families and three datasets, and on finetuning arms that vary corpus, capacity, objective, and architecture. Ablations isolate the narrowed search space as the crop's active ingredient, separating it from added frames, added tokens, and higher per-frame resolution. Accuracy is reported by moment length only in bins five seconds wide, the emitted length distribution is read for simplicity rather than for what it omits, and every zoom variant trains the component that proposes or consumes the window (Appendix B).

Our analysis draws on several model families and datasets, and on fine-tuning interventions that vary the training corpus, model capacity, objective, and architecture. Ablations on the crop further show that its effect comes from narrowing the search space rather than adding frames, tokens, or per-frame resolution. This analysis also sets our work apart from prior studies, which report accuracy by moment length only in bins five seconds wide, read the emitted length distribution for how simple it is rather than for which lengths it leaves out, and train the component that proposes or consumes the zoom window rather than testing a frozen model.

Based on this diagnosis, our main contributions are summarized as follows.

- **Length-stratified evaluation.** We evaluate grounding accuracy separately for each moment length and reveal that VideoLLMs fail almost completely on short moments, a failure that benchmark averages hide.
- **Falsification suite for the prior**: We fine-tune the model with interventions ranging from LoRA adapters to full-parameter updates of the language model, including an RL reward on the center, a center-first output format, and a soft-argmax center head, and none of them brings the center error below a second.
- **A search-space limit and a training-free lever.** We find that the frozen model already localizes short moments with sub-second precision once its input is narrowed, and turn this finding into a length-gated second look that improves short-moment grounding without any training.

2 METHOD

This section fixes the measurement the diagnosis rests on, in two parts: the length strata and diagnostics that every later result is reported within, and the models and benchmarks they are computed on.

2.1 Measurement protocol and diagnostics

Every video–query pair is filed by its ground-truth moment length L_{gt} into five strata: S0 ($<$2 s), S1 (2–5 s), S2 (5–10 s), M (10–30 s) and L ($\geq$30 s). The boundaries are round values, fixed once and applied unchanged to every model, dataset and intervention we report, and S0 is the regime of interest.

We report R@1@IoU$\{0.3,0.5,0.7\}$ and mIoU, but an overlap score cannot say why a prediction missed. The length-compression slope β, fit as $L_{\text{pred}} = \alpha + \beta\, L_{\text{gt}} + \varepsilon$, puts a model's length calibration in one number: $\beta{=}1$ is calibrated, $\beta{<}1$ compressive (Sec. 3). *Emission*, the share of predictions shorter than 2 s, asks whether a model proposes short spans at all, before asking where it puts them; we report it within a stratum or corpus-wide, as stated. *Center error* $|\Delta c| = |\hat{c} - c|$ on span midpoints isolates placement, and is unaffected by span length. We report it in seconds, as a median within each stratum. We report every metric within ground-truth length strata, with 95% cluster-bootstrap (by video) intervals, and follow the official TimeLens-Bench protocol (Zhang et al., 2026b) with one rounding convention fixed per benchmark and center errors computed unrounded throughout (Appendix C). Pooled across the three benchmarks, the sub-2 s population is 1,344 queries: 590 on Charades-TimeLens, 92 on QVHighlights-TimeLens and 662 on native ActivityNet.

2.2 Experimental setup

Models and reproduction. Our primary model is TimeLens-7B, a Qwen2.5-VL grounder (Bai et al., 2025) that interleaves per-frame timestamp text with its visual input. Run through its official pipeline, our reproduction matches the published numbers to within 0.3 points on every metric (Appendix N). We add two further families so that the models we test span training paradigms: VideoChat-R1, which is RL-tuned on the same backbone and evaluated on a length-balanced subsample, and TRACE, a Mistral-7B + CLIP model with structured-token decode evaluated on the full set. Both are validated against their published numbers in Appendix N.

Benchmarks. For each dataset we use the ground truth in which sub-2 s moments exist and are tightly bounded. For Charades and QVHighlights the native ground truth contains no sub-2 s moments at all, so we use precise re-annotations; for ActivityNet the original float annotations already contain them. Native-split results are in Appendix N. Our primary evaluation set is the TimeLens-Bench re-annotation of Charades (Sigurdsson et al., 2016), called *Charades-TimeLens* upstream (Zhang et al., 2026b), which holds 3,363 video–query pairs. For ActivityNet (Krishna et al., 2017) we keep the original float annotations, whose native ground truth is already sub-second on videos with a median length of 117 s, since the TimeLens re-annotation would snap spans to a coarser grid. QVHighlights (Lei et al., 2021) is 150 s of YouTube vlog per clip, and its native windows are unions of 2 s clips, so sub-2 s ground truth exists there only under re-annotation. We therefore use *QVHighlights-TimeLens*, which re-draws one span per query. We write Charades and QVHighlights for short throughout.

3 The Short-Moment Collapse

This section makes moment length the independent variable. It then rules out three explanations of the resulting gap: that shorter moments carry harder queries and longer videos, that the overlap metric manufactures it, and that the boundaries were drawn badly. Short moments are hard twice over: few sampled frames fall inside them, so little evidence for the moment reaches the model, and the target is narrow enough that a small misplacement misses it entirely. Recall falls steadily as moments shorten, in every model and dataset we examine, and the shortest moments are almost never found.

The length gradient. Table 1 reports strict recall by ground-truth length stratum on the TimeLens-Bench re-annotation of Charades, and Fig. 1(b, c) plots the same data. Strict recall climbs across the four strata below 30 s for every model and dataset except ActivityNet, where it peaks in the five-to-ten-second stratum. TimeLens-7B spans 4.1 to 57.4%, so its 28.5% aggregate conceals a $\sim$14$\times$ gap.

Table 1: Strict recall (R@1@IoU0.7, %) by ground-truth moment length: three model families and the raw zero-shot backbone (backbone^0 = Qwen2.5-VL-7B, no grounder-specific finetuning), three datasets (ActivityNet with its original annotations; QVHighlights via the TimeLens-Bench re-annotation).

	Charades-TimeLens				A-Net (native)	QVH-TimeLens
L_{gt}	TimeLens	VC-R1	TRACE	backbone^0	TimeLens	TimeLens
$<$2 s	**4.1** (24/590)	**0.0** (0/300)	**0.2** (1/590)	**0.0** (0/590)	**5.1** (34/662)	**7.6** (7/92)
2–5 s	20.6	2.0	6.8	4.6	22.2	19.8
5–10 s	42.7	46.0	21.2	29.8	29.8	52.2
10–30 s	57.4	51.0	29.5	58.8	26.6	64.2
$\geq$30 s	38.7	3.2	41.9	35.5	22.9	42.3
overall	28.5	24.2	12.6	18.0	24.2	45.1

Generality. The collapse is not a property of one training recipe or one corpus. VideoChat-R1 localizes none of its sub-2 s moments, with no hit even at the lenient IoU 0.3 threshold, and TRACE localizes one. The raw Qwen2.5-VL backbone that TimeLens finetunes is blind to the stratum zero-shot at both strict thresholds, while matching its own finetune on 10–30 s moments (Appendix R). The short-end failure therefore predates the grounder's own finetuning (Sec. 4). The pattern replicates on ActivityNet's original sub-second annotations and, through the re-annotation, on QVHighlights vlogs.

Deconfounding controls. What co-occurs with short moments does not explain the collapse. Three estimators that control for video length, query difficulty and scene clutter, on the same pairs, all return a strongly negative shortness effect at $p < 10^{-10}$. A continuous-length re-specification shows the effect is smoothly graded, at $+1.07$ logits per log-second on Charades, and it survives two query-difficulty proxies. Appendix G names the three estimators and gives the protocols and full outputs. The genuine video-length deconfound comes from ActivityNet, since Charades clips barely vary in duration.

Metric geometry. Because IoU is scale-invariant, a fixed $\sim$1 s boundary error is near-total for a 1 s moment and trivial for a 30 s one, which mechanically steepens the curve. Prop. D gives that steepening in closed form. We therefore re-score with a scale-fair metric, in which each boundary must land within $\alpha \cdot L_{\text{gt}}$ of truth, so tolerance grows with length. The collapse persists monotonically at every α we test (Fig. 2a). IoU exaggerates the steepness; it does not invent the deficit.

Error structure. Short-moment misses are wrong-region errors, so no sharpening of the annotations would repair them. We sort every top-1 prediction into a hit at IoU 0.5, a Type-A error with its midpoint outside the true moment, or a Type-B error with the midpoint inside and the boundaries loose, following the output-format study (Jin et al., 2026); Fig. 2b plots them as shares of all predictions, where Type-A's share falls steadily as moments lengthen. Type-A accounts for 78–83% of sub-2 s misses in all three model–dataset pairs (Appendix H).

4 Mechanism I: the Length-Compression Prior

We profile the *signed* length error for every model and dataset (Fig. 2(c)). The median $\log_2(L_{\text{pred}}/L_{\text{gt}})$ runs from strongly positive at one second, through zero in the three-to-seven-second range, to negative beyond it, so the distribution is pulled toward a preferred $\sim$3–5 s length, the typical training moment. TimeLens emits a span shorter than 2 s in only 3.6% of its predictions against 17.5% of true moments, and on the 590 ultra-short cases predicts a median 3.0 s span for a 1.0 s event. Placement fails separately: the median center error $|\Delta c|$ stays near 1 s at every length below 15 s (Appendix F). Length and placement therefore come apart under training.

The prior in the raw backbone. Whether better finetuning could remove the prior depends on whether the grounder created it. We therefore run the raw Qwen2.5-VL backbone zero-shot through the same protocol (Appendix O). The prior is already there and stronger: at one second the backbone predicts a span about five times too long, where the finetune predicts three, and it emits a sub-2 s span

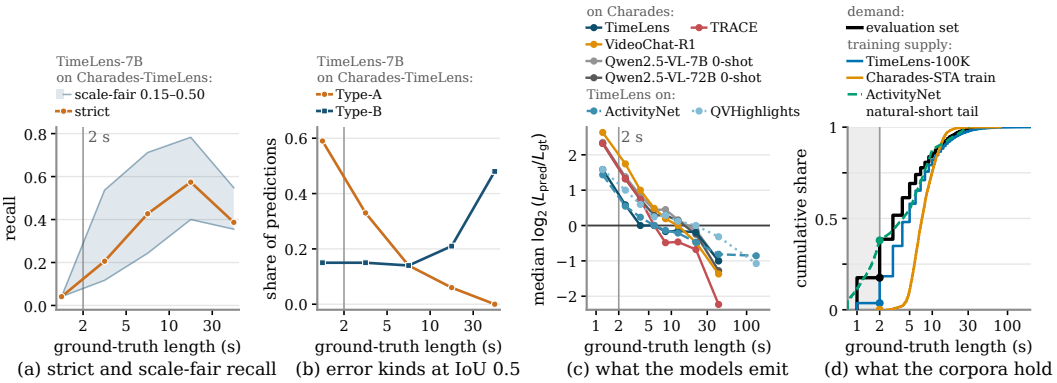


Figure 2: Recall, error kind, emitted length and corpus supply against ground-truth moment length. **(a)** The strict line is IoU ≥ 0.7; scale-fair recall instead requires both boundaries within $\alpha \cdot L_{\text{gt}}$ of truth. **(b)** Error taxonomy at IoU 0.5, as shares of all predictions. **(c)** Median $\log_2(L_{\text{pred}}/L_{\text{gt}})$ by ground-truth length. **(d)** Cumulative ground-truth length distributions, with dots marking each corpus's sub-2 s share.

Table 2: The length-compression prior across models, dataset versions and scales. Each dataset reports β, sub-2 s R@1@IoU0.7 (abbreviated R@.7 in the header), and median center error $|\widetilde{\Delta c}|$ in seconds. Charades-TL is the Charades-TimeLens re-annotation. Every column is a full split; ActivityNet is all 17,505 native pairs, 662 of them sub-2 s.

	Charades-TL			Charades-STA	ActivityNet			QVHighlights		
	β	R@.7	$\lvert\widetilde{\Delta c}\rvert$	β	β	R@.7	$\lvert\widetilde{\Delta c}\rvert$	β	R@.7	$\lvert\widetilde{\Delta c}\rvert$
TimeLens-7B	0.60	0.041	1.00	0.53	0.59	0.051	1.83	0.46	0.076	1.00
TimeLens-8B	0.73	0.008	1.00	0.57	0.51	0.069	1.55	0.55	0.033	0.50
VideoChat-R1	0.31	0.000	1.50	0.46	0.45	0.000	4.05	0.39	0.000	6.50
TRACE	0.49	0.002	2.00	0.30	0.40	0.057	2.50	0.28	0.022	6.75
Qwen2.5-VL-7B, zero-shot	0.54	0.000	1.62	0.59	0.26	0.009	2.04	0.59	0.011	2.50
Qwen2.5-VL-72B, zero-shot	0.51	0.000	1.25	0.54	0.19	0.012	1.86	0.50	0.000	1.75

once in 3,363 predictions. The backbone is not simply weaker: its 10–30 s recall is statistically flat against the finetune's (Sec. 3). Nor is the interleaved prompt format responsible: a plain grounding prompt leaves the short-end ratio, β and the empty sub-2 s stratum essentially unchanged. Grounding finetuning therefore does not create the prior; it re-centers it onto the corpus's modal length, moving the zero-crossing from $\sim$12–15 s to $\sim$5 s, without repairing the short end. Prompting does not repair it either: a prompt licensing short answers raises sub-2 s emission by four points while losing overall, and an oracle hint naming the true length bin is absorbed with the center unmoved (Appendix P, Sec. 5).

Length rebalancing. The field's default remedy is to finetune on oversampled short data, and there is none to oversample: the Charades training split's shortest span is 2.5 s, and the grounder's own instruction corpus supplies 3.65% sub-2 s spans at a median of 5.0 s, which is where the re-centered prior lands (Appendix J). Table 3 reports every finetuning arm on Charades against the ground-truth sub-2 s share. Applied there regardless, oversampling the shortest moments available makes the model worse: β collapses and sub-2 s predictions disappear. An equal-sized control on the split's natural distribution is almost as bad, so the oversampling is not what harms it, and this pair cannot say whether the corpus or the finetuning is responsible. The next arm separates them by training on ActivityNet's naturally short moments.

Table 3: Finetuning arms on Charades: sub-2 s emission against the ground-truth share of 17.5%, the length slope β, mIoU and R@1@IoU0.5. **Bold** marks the arms that emit no sub-2 s span at all.

	base	rebal.	control
% preds <2 s (GT 17.5)	3.6	**0.0**	**0.0**
β	0.598	0.195	0.247
mIoU	0.490	0.377	0.393
R@1@IoU0.5	0.556	0.364	0.375

Supply of genuine short moments. We test supply three ways: decode-time recalibration of the emitted span, speed augmentation that manufactures sub-2 s moments by resampling the video clock, and a matched-budget finetune on the ActivityNet Captions short tail, the strongest version since it trains on naturally occurring short moments. Every arm repairs emission, most strongly the ActivityNet arm, where sub-2 s output rises from 10.2 to 54.1% on the ultra-short cases. A same-corpus control bounds what the short moments themselves supply: the identical recipe on ActivityNet's natural lengths reaches 30.2%, from a corpus holding 4% sub-2 s content against the short tail's 38%, and its length calibration is the best we measure, β $0.60 \rightarrow 0.71$. Placement moves far less, and only where the corpus matches. Strict recall on a one-second target needs the predicted center within 0.214 s of the truth, and the share of sub-2 s clips inside that window rises by five points on TimeLens-7B and fifteen on TimeLens-8B (Appendix R). The control leaves that share unmoved at 11%. On ActivityNet, the arm's own corpus, placement does repair: the median center error halves to 0.76 s from 1.83 s, level with the 0.74 s oracle-window ceiling (Table 5). It is paid for elsewhere. Long moments collapse, overall strict recall falls on both corpora, and on ActivityNet the corpus-wide share inside the center window drops from 41 to 34% (Appendix I). The dissociation is not an artifact of adapter capacity, seed, or lineage: full-parameter finetunes, three seeds, TimeLens-8B and the RL-tuned VideoChat-R1 all reproduce it (Appendix T).

Mass without discrimination. Emission and placement could come apart for two reasons: the model learns where the moment is and cannot express it, or it shifts mass toward short spans without learning where they go. We separate these by asking whether a model can tell the right moment from a plausible wrong one. For each sub-2 s clip we take the ground-truth answer and a decoy of the same length placed elsewhere in the video, and compare the total log-probability the model assigns to each. Matching the lengths removes the length prior, so all that distinguishes the two answers is where they point. The frozen grounder prefers the truth by a median 15.6 nats and ranks it above the decoy on 94% of clips. All five finetuned arms fall to roughly 4.6 nats, yet still rank the truth first on 84–93% (Appendix S). No arm improves on the frozen model by either measure, so finetuning does not buy discrimination, and the collapse in margin is largely a change in confidence rather than in ordering.

Generality. Every model, dataset and scale we test compresses: $\beta < 1$ throughout, including the two natively annotated columns (Table 2). An unbiased-twin null rules out regression to the mean (Appendix D). TimeLens and VideoChat-R1 share a backbone yet differ in β, 0.60 against 0.31, so training shapes the bias rather than the backbone fixing it. TimeLens-8B makes the point sharpest. It improves every aggregate metric, mIoU 0.490 to 0.553, yet does not move the sub-2 s center floor, which stays near a second, cuts sub-2 s emission to 1.2%, and lowers sub-2 s strict recall from 0.041 to 0.008: an upgrade the leaderboard sees, and a regression it cannot.

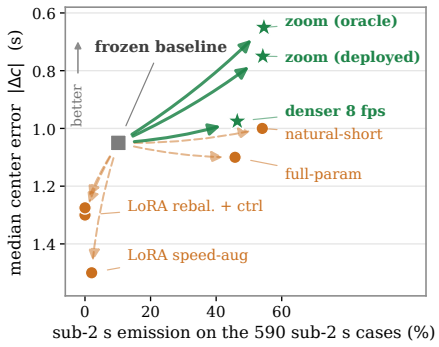
5 Mechanism II: the Center Floor

This section tests the natural conjecture that closing the center floor needs a backbone update, and refutes it in both directions (Table 4). The prior explains the length half of the failure, and every fix that repairs emission, from the dominated remedies (Appendix I) to the denser-sampling second look of Sec. 6, leaves the same residue: a $\sim$1 s absolute center error. A frozen-LLM boundary-regression readout does not extract a sub-second center either (Appendix V).

5.1 Backbone updates

We attack the center error directly with three interventions spanning the design space (Table 4), all LoRA finetunes on a speed-augmented split with genuine sub-2 s moments (implementations in Appendix V). **(1)** A scale-free Gaussian *center reward* on the reinforcement-learning objective (GRPO) changes the emitted span on 5.7% of predictions without improving them, and raising the learning rate a hundredfold degrades every column monotonically rather than helping. **(2)** A *center-first reparameterization* ("centered at C, lasts W") is learned perfectly, yet the center is no more precise when named directly. **(3)** *EC-Sharp*, a soft-argmax expected-center head (Nibali et al., 2018) trained jointly into the backbone, fits sub-second centers on training clips but fails on held-out ones, a train/held-out gap rather than a center that generalizes. Every finetune here matches or degrades the frozen model's sub-2 s center, which remains the best center-localizer we have.

Table 4: Center interventions and the training-free crop, sub-2 s stratum. **Left**: emission against median center error for the supply and density arms and the two levers, with arrows from the frozen baseline; up is better, and orange marks training arms, green the training-free ones. **Right**: the backbone-update ladder, its median center error, overall mIoU and length slope β. Centers here are scored raw, as the finetuned arms are; Table 2 scores Charades-TimeLens on its 1 s annotation grid, where the frozen model reads 1.00.



	sub-2 s $\widehat{\lvert\Delta c\rvert}$ (s)	overall mIoU	β
frozen baseline	1.05	0.490	0.599
finetuning the backbone			
+ GRPO reward, 10^{-6}	1.05	0.475	0.602
+ GRPO reward, 10^{-5}	1.05	0.474	0.584
+ GRPO reward, 10^{-4}	1.45	0.439	0.504
+ (c, w) reparam.	1.30	0.390	0.180
+ EC-Sharp (text)	1.50	0.398	0.310
+ EC-Sharp (head)	1.74	0.362	0.110
readout head	$\geq$1.80	—	—
conditioning the frozen model (training-free)			
zoom, deployed	**0.85**	0.464	**0.622**
zoom, oracle W=8	**0.65**	—	—

5.2 Temporal zoom

Every Sec. 5.1 intervention feeds the model the entire video and asks it to localize, to sub-second precision, a moment occupying ∼3% of a 30 s Charades video. *Temporal zoom* asks whether the residual error is a missing capability or a missing condition. It changes no weights: a coarse pass reads the ordinary predicted center $\hat{c}_0$, and a fine pass re-decodes only a $\pm W/2$ window around it at higher fps where the source allows it, with timestamps read and emitted on the original video's clock (Appendix W).

The oracle ceiling. Table 5 reports the sub-2 s stratum under an oracle window containing the moment, against the frozen model and the two trivial in-window floors of Prop. D. The crop recovers the sub-second center on both datasets: at $W = 8$ s it lifts Charades strict recall five-fold, drops $|\Delta c|$ below the frozen floor, and lifts R@1@IoU0.5 by three-quarters on ActivityNet. The model beats both trivial in-window floors on every metric, so the gain is not the window's geometry (Prop. D). Frame rate is not the cause either: the same 8 fps on the whole video leaves the floor untouched ($|\Delta c| \approx 1$ s, Sec. 6).

Placement without removal. A second route reaches the same center and isolates what produces it. Holding frame count, per-frame resolution, token budget and whole-video support fixed, we change only where frames are sampled: a kernel placed at the true center lowers the median center error from 1.05 to 0.60 s, monotonically in its width (Appendix U). That placement is an oracle, as the crop ceiling is, so this identifies the mechanism rather than supplying a lever. Nothing is removed, so a shorter clip, extra tokens per frame and the absence of distractors are each unnecessary. The realized positional range does narrow as the kernel tightens, so this arm leaves that factor open. A sub-second center is therefore available from the frozen model under the right conditioning. A substitution analysis on the cropped predictions pins what is left: replacing the predicted width with truth lifts R@1@IoU0.7 only to 0.380, while replacing the center lifts it to **0.714**. Which of the two binds is a property of the model. TimeLens-8B places 11% of its sub-2 s centers exactly on truth and emits no one-second span at all, so width is what holds it back there.

We keep $W = 8$ as the deployable default: it leaves margin for coarse-pass error, while wider windows reintroduce the search-space problem (ablations in Appendix E).

5.3 Ablations

Density. The natural alternative to the crop is re-decoding the whole video at 8 fps instead of 2 at the same token budget. While it leaves the center untouched and costs the longer strata their resolution, it repairs the emission half, lifting sub-2 s R@1@IoU0.7 to 0.153, and a resolution-matched

Table 5: *Ceiling*: a $W{=}8$ s oracle window containing the moment, off-center at 0.6. *Trivial floor*: what an agent that *ignores the video* still scores, since the window already contains the moment. We take the *stronger* of Prop. D's two policies per metric, so the comparison is conservative. $\widetilde{\cdot}$ = median. Recalls and mIoU follow each benchmark's annotation resolution; center errors are unrounded. ActivityNet is the full split ($n{=}662$ sub-2 s pairs).

$<$2 s stratum	R@1@IoU0.5	R@1@IoU0.7	mIoU	$\widetilde{\lvert\Delta c\rvert}$
Charades base (full video)	0.259	0.041	0.250	1.05
Charades ceil. $W=8$	**0.386**	**0.202**	**0.334**	**0.65**
best trivial floor	0.000	0.000	0.125	0.80
ActivityNet base (full video)	0.144	0.051	0.216	1.83
ActivityNet ceil. $W=8$	**0.252**	**0.073**	**0.322**	**0.74**
best trivial floor	0.021	0.011	0.166	0.80

control reproduces that lift. The cause is therefore temporal density rather than the resolution trade, which places density on the emission side of the pair and leaves the center floor to the crop.

Component decomposition. Table 6 separates the crop into its parts, reporting the support each intervention sees, the frames it samples, and the emission, center error and strict recall it produces on the same sub-2 s clips. The crop holds 64 frames against the control's 57, so it never had a frame advantage to begin with. Permuting the cropped frames, which changes nothing but their order, sends the center error to 2.85 s, so the window supplies no prior of its own. Redistributing the same budget over the whole video, with nothing removed, recovers the gain in full, and aiming that same kernel 8 s away is worse than sampling uniformly. Doubling the budget changes nothing. What moves the center is where the temporal samples fall relative to the moment, so the zoom re-conditions the input rather than learning a center, which is exactly where the trained heads failed.

6 A Training-Free Lever at Inference

This section turns the two mechanisms into a deployable lever: give a clip a second, concentrated look only when the model's own first pass predicts a short moment. Both point to the same inference-time principle, raising the moment's share of the input.

The router-gated zoom. A clip whose first-pass predicted length falls below $\tau{=}4$ s is re-decoded on the ±4 s window of Sec. 5.2; all others keep their first answer. Routing rates, compute costs and per-metric deltas per benchmark are in Appendix K: Charades routes the largest share of its clips and pays the highest premium, while QVHighlights and ActivityNet each route about a tenth at close to baseline cost. On Charades sub-2 s strict recall rises $3.6\times$, a paired interval of $[+7.5, +13.6]$ points, at a small but real overall cost in mIoU. ActivityNet is the deployment test: the router sees all $17{,}505$ pairs with no ground-truth filter, so its false positives are measured rather than excluded, and its overall aggregate is unchanged. QVHighlights is the same test at a smaller routing rate, lifting sub-2 s R@1@IoU0.5 $0.196 \to 0.337$ while the aggregate holds. Its false positives cost three points of R@1@IoU0.5, where ActivityNet's mis-routed clips gain rather than lose (Appendix K). The gate protects the rest of the benchmark rather than producing that gain: applied to every sub-2 s clip instead, the same fine pass reaches 0.175 on Charades and 0.130 on QVHighlights, on decodes byte-identical to the gated arm's. Removing it altogether buys 17 more short hits on Charades and gives up 314 elsewhere, with the same asymmetry on QVHighlights. The loss begins in the five-to-ten-second stratum rather than at the long end (Appendix K.1). The same configuration transfers to TimeLens-8B, lifting sub-2 s R@1@IoU0.7 $0.008 \to 0.159$ on Charades and $0.033 \to 0.250$ on QVHighlights. There it costs no aggregate at all: overall mIoU rises rather than falls, and lenient recall is unmoved (Appendix K.2).

A frozen threshold. τ was chosen on the sets it is reported on, so we also froze it: selected on a disjoint half of the videos by short strict recall alone, then evaluated once on the other half. On Charades the gain holds, sub-2 s R@1@IoU0.7 rising $0.050 \to 0.137$ with an interval of $[+4.1, +13.3]$ points, and on QVHighlights and ActivityNet the strict change is indistinguishable from zero, as it is under the deployed threshold. What the freeze costs is the aggregate: 2.7 points of overall mIoU on Charades, exceeding the two-point bound we had set in advance.

Table 6: Component decomposition of the crop and of denser sampling, 590 sub-2 s Charades-TimeLens clips on the frozen model, at a 14,336-token budget. Equalizing tokens per frame changes nothing (0.65 s at 16,235 tokens), and the 4 fps rung is in Appendix E. **Bold** marks a center error whose paired by-video interval against the control excludes zero. Window and kernel placement are oracle, as for the crop ceiling. Recalls and emission follow the rounded 1 s grid; center errors are unrounded.

intervention	support	frames	emission	$\widetilde{\lvert\Delta c\rvert}$ (s)	R@1@IoU0.7
uniform 2 fps (control)	whole video	57	9.5%	1.05	0.041
denser sampling, 8 fps	whole video	226	46.4%	0.98	0.153
temporal crop, W=8	window	64	54.1%	**0.65**	0.193
out-of-window frames zeroed	whole video	57	13.1%	**0.85**	0.047
crop with frames permuted	window	64	69.0%	2.85	0.025
foveated sampling, σ=2	whole video	57	52.7%	**0.60**	0.205
aimed 8 s off the moment	whole video	57	35.9%	5.60	0.017

Threshold geometry. Lenient thresholds reward over-long spans that graze the moment, and the router replaces those with tight spans that either hit or miss. On Charades sub-2 s R@1@IoU0.3 therefore falls $0.449 \rightarrow 0.314$. QVHighlights runs a 1 s protocol, where a two-second prediction containing a one-second target scores exactly IoU 0.5: boundary mass carries 85% of its lenient gain, leaving 2.2 points under a strict threshold. ActivityNet's float annotations admit no such ties, so its smaller lenient gain is all margin. Strict-threshold changes on both are within noise, for the reach reason quantified below.

The density alternative. Denser sampling, gated the same way, is matched or beaten by the zoom router on every Table A6 metric. Across all three regimes the zoom is the better single lever, while denser sampling is regime-specific: it lifts Charades, leaves ActivityNet unchanged, and on QVHighlights falls to 0.043 from a 0.076 baseline. The bound is the source frame rate, and it binds both levers: ActivityNet's 3 fps files turn an 8 fps request into 1.5× density for the denser arm and for the zoom's fine pass alike, so on that corpus the crop narrows the window without adding density. No ground-truth-free gate picking between them beat the better lever, and a per-query oracle gains eight to ten points over it, so what is missing is a gate rather than headroom (Appendix L).

Coarse-pass reach. The zoom can only help where the coarse pass lands near the moment, so its gain is confined to the short-video benchmark and within noise on the two long ones. On ActivityNet the coarse pass lands so far off for about a fifth of sub-2 s moments that no center-anchored window contains them, and what gain remains concentrates in medium-length videos (Appendix K). Long-video reach is therefore gated by the coarse stage, which is also the likelier route to the per-query oracle headroom the mixed-corpus experiment identified (Appendix L).

7 Conclusion

We show that VideoLLM grounders fail almost completely on moments shorter than two seconds, and trace that failure to a limit of the search space rather than a missing capability. It has two causes: a length-compression prior already present in the raw backbone, which grounding finetuning re-centers rather than repairs, and a residual center floor that no finetune we evaluate moves. Neither yields to training: the corpora hold too few short moments and the floor needs conditioning instead. Cropping the frozen model's input supplies that conditioning, and gated by the model's own first pass it raises short-moment recall on all three benchmarks without training. A sub-second center localizer and a stronger first pass remain open. The account is falsifiable: a finetune at any scale that moved the sub-2 s center below the frozen floor would refute it, and none here did. The crop is a bundle, since range, distractors and rescaling are not separated, so "search-space limit" is our reading rather than a factorized mechanism (Appendix A). Reporting by length stratum rather than by corpus average is what makes the deficit visible at all, and the protocol and its conventions are given in Appendix C.

AI USE STATEMENT

We report our use of generative AI under the three categories of the submission form, which together span the required and the recommended disclosure lists of the ICLR 2027 AI Policy for Authors.

Research ideation and execution. We used generative AI to propose and refine hypotheses; to design and critique experiments; to implement the harnesses, the scorer, the temporal-zoom intervention, the analysis code and the data pipelines; to build the manifests and the figures; and to interpret results. The two propositions and the crop-reach bound were derived with AI assistance and checked by hand.

Retrieval and discovery. We used generative AI to search and summarize related literature and to format references.

Writing. We used generative AI to draft and edit the text of this paper and to suggest its structure and its title.

We have reviewed all AI-assisted work. No reported number was produced by a language model: every quantitative claim comes from the scorer we release, which revalidates previously reported landmarks on each run and aborts if one fails to reproduce. We take responsibility for the final content of this work, including all text, claims and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work analyzes published models on public academic benchmarks (Charades, ActivityNet Captions, QVHighlights); no personal data is collected.

REPRODUCIBILITY STATEMENT

All numbers follow the benchmarks' official scoring protocols as stated in Sec. 2, including the timestamp-rounding convention. The scoring conventions are specified in Appendix C.1. The stratified scorer, its reference results for every model in Table 2, and a script that rebuilds the per-stratum splits from each dataset's own release are included in the supplementary material. Appendix T specifies model revisions, decoding settings, and finetuning recipes with seeds; Appendix V the center-probe architectures; and Appendix N the scorer validation against the official implementations.

REFERENCES

Humam Alwassel, Fabian Caba Heilbron, Victor Escorcia, and Bernard Ghanem. Diagnosing error in temporal action detectors. In *European Conference on Computer Vision (ECCV)*, 2018. 15

Shuai Bai et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 3

Shyamal Buch, Cristóbal Eyzaguirre, Adrien Gaidon, Jiajun Wu, Li Fei-Fei, and Juan Carlos Niebles. Revisiting the "video" in video-language understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 15

Yuxin Cao, Wei Song, Derui Wang, Jingling Xue, and Jin Song Dong. Failures to surface harmful contents in video large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026. arXiv:2508.10974. 15

Dazhao Du, Liao Duan, Jian Liu, Tao Han, Yujia Zhang, Eric Liu, Xi Chen, and Song Guo. MLLMs know when before speaking: Revealing and recovering temporal grounding via attention cues. *arXiv preprint arXiv:2605.21954*, 2026. 16

Yue Fan, Xiaojian Ma, Rujie Wu, Yuntao Du, Jiaqi Li, Zhi Gao, and Qing Li. VideoAgent: A memory-augmented multimodal agent for video understanding. In *ECCV*, 2024. 16, 17

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)*, 2017. 1

Silvio Giancola, Anthony Cioppa, Adrien Deliège, Floriane Magera, Vladimir Somers, et al. SoccerNet 2022 challenges results. In *Proceedings of the 5th International ACM Workshop on Multimedia Content Analysis in Sports (MMSports)*, 2022. doi: 10.1145/3552437.3558545. 15

Yongxin Guo, Jingyu Liu, Mingda Li, Qingbin Liu, Xi Chen, and Xiaoying Tang. Trace: Temporal grounding video llm via causal event modeling. In *International Conference on Learning Representations (ICLR)*, 2025. 1, 15

Tanveer Hannan, Md Mohaiminul Islam, Jindong Gu, Thomas Seidl, and Gedas Bertasius. ReVisionLLM: Recursive vision-language model for temporal grounding in hour-long videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. arXiv:2411.14901. 16

James Hong, Haotian Zhang, Michaël Gharbi, Matthew Fisher, and Kayvon Fatahalian. Spotting temporally precise, fine-grained events in video. In *European Conference on Computer Vision (ECCV)*, 2022. 15

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, 2024a. 1, 15

De-An Huang, Shijia Liao, Subhashree Radhakrishnan, Hongxu Yin, Pavlo Molchanov, Zhiding Yu, and Jan Kautz. Lita: Language instructed temporal-localization assistant. In *European Conference on Computer Vision (ECCV)*, pp. 202–218, 2024b. 1, 15

Shengji Jin, Yuanhao Zou, Victor Zhu, Zhengping Ji, and Chen Chen. How should video llms output time? an analysis of efficient temporal grounding paradigms. *arXiv preprint arXiv:2604.08966*, 2026. 4, 17

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)*, 2025. arXiv:2411.12951. 16

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 706–715, 2017. 1, 3

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*, 2023. 15

Jie Lei, Tamara L. Berg, and Mohit Bansal. Qvhighlights: Detecting moments and highlights in videos via natural language queries. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. 1, 3, 15

Xinhao Li, Ziang Yan, Desen Meng, Lu Dong, Xiangyu Zeng, Yinan He, et al. Videochat-r1: Enhancing spatio-temporal perception via reinforcement fine-tuning. *arXiv preprint arXiv:2504.06958*, 2025. 1, 15

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 *ICCV*, 2023. 1, 15

Zachary C. Lipton, Yu-Xiang Wang, and Alexander J. Smola. Detecting and correcting for label shift with black box predictors. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pp. 3122–3130, 2018. 17, 21

Wenqi Liu, Yunxiao Wang, Shijie Ma, Meng Liu, Qile Su, Tianke Zhang, Haonan Fan, Changyi Liu, Kaiyu Jiang, Jiankang Chen, Kaiyu Tang, Bin Wen, Fan Yang, Tingting Gao, Han Li, Yinwei Wei, and Xuemeng Song. VideoTemp-o3: Harmonizing temporal grounding and video understanding in agentic thinking-with-videos. In *International Conference on Machine Learning (ICML)*, 2026a. arXiv:2602.07801. 16

Xiaolin Liu, Yilun Zhu, Xiangyu Zhao, Xuehui Wang, Yan Li, Xin Li, Haoyu Cao, Xing Sun, Shaofeng Zhang, Xu Yang, Zhihang Zhong, and Xue Yang. Moment-Video: Diagnosing temporal fidelity of video MLLMs on momentary visual events. *arXiv preprint arXiv:2606.02522*, 2026b. 15

Ye Liu, Kevin Qinghong Lin, Chang Wen Chen, and Mike Zheng Shou. VideoMind: A chain-of-LoRA agent for temporal-grounded video reasoning. In *International Conference on Learning Representations (ICLR)*, 2026c. arXiv:2503.13444. 16

Yi Liu, Limin Wang, Yali Wang, Xiao Ma, and Yu Qiao. FineAction: A fine-grained video dataset for temporal action localization. *IEEE Transactions on Image Processing*, 31:6937–6950, 2022. 15

WonJun Moon, Sangeek Hyun, SuBeen Lee, and Jae-Pil Heo. Correlation-guided calibration of query dependency for video temporal grounding. *Pattern Recognition*, 174:112984, 2026. 1, 15

Fangzhou Mu, Sicheng Mo, and Yin Li. Snag: Scalable and accurate video grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 1, 15

Aiden Nibali, Zhen He, Stuart Morgan, and Luke Prendergast. Numerical coordinate regression with convolutional neural networks. *arXiv preprint arXiv:1801.07372*, 2018. 6, 17

Ming Nie, Dan Ding, Chunwei Wang, Yuanfan Guo, Jianhua Han, Hang Xu, and Li Zhang. Slow-Focus: Enhancing fine-grained temporal understanding in video LLM. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. 16

NVIDIA. NVIDIA AI blueprint for video search and summarization. https://github.com/NVIDIA-AI-Blueprints/video-search-and-summarization, 2025. 17

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. 15

Seojeong Park, Jiho Choi, Kyungjune Baek, and Hyunjung Shim. Momentmix augmentation with length-aware detr for temporally robust moment retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2026. 16

Long Qian, Juncheng Li, Yu Wu, Yaobo Ye, Hao Fei, Tat-Seng Chua, et al. Momentor: Advancing video large language model with fine-grained temporal reasoning. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024. 15

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)*, 2024. 1, 15

Xubin Ren, Lingrui Xu, Long Xia, Shuaiqiang Wang, Dawei Yin, and Chao Huang. VideoRAG: Retrieval-augmented generation with extreme long-context videos. *arXiv preprint arXiv:2502.01549*, 2025. 17

Marco Saerens, Patrice Latinne, and Christine Decaestecker. Adjusting the outputs of a classifier to new a priori probabilities: A simple procedure. *Neural Computation*, 14(1):21–41, 2002. 17, 21

Sukmin Seo and Geewook Kim. Natural-language temporal grounding in hour-long videos is a search problem: A benchmark and empirical decomposition. *arXiv preprint arXiv:2606.12300*, 2026. 17

Xiaoqian Shen, Min-Hung Chen, Yu-Chiang Frank Wang, Mohamed Elhoseiny, and Ryo Hachiuma. Zoom-zero: Reinforced coarse-to-fine video understanding via temporal zoom-in. *arXiv preprint arXiv:2512.14273*, 2025. 16

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. 3

Te Tao and Shao-Lun Huang. Adavtg-llm: A videollm-based efficient video temporal grounding framework. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 3631–3642, June 2026. 15

Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. VideoAgent: Long-form video understanding with large language model as agent. In *ECCV*, 2024a. 17

Xizi Wang, Feng Cheng, Ziyang Wang, Huiyu Wang, Md Mohaiminul Islam, Lorenzo Torresani, et al. Timerefine: Temporal grounding with time refining video llm. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2026. 15

Yifan Wang, Ziyi Liu, Xiaolong Sun, Jiawei Wang, and Hongmin Liu. Length matters: Length-aware transformer for temporal sentence grounding. *arXiv preprint arXiv:2508.04299*, 2025. 16

Yueqian Wang, Xiaojun Meng, Jianxin Liang, Yuxuan Wang, Qun Liu, and Dongyan Zhao. Hawkeye: Training video-text LLMs for grounding text in videos. *arXiv preprint arXiv:2403.10228*, 2024b. 16

Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13084–13094, 2024. 17

Yongliang Wu, Xinting Hu, Yuyang Sun, Yizhou Zhou, Wenbo Zhu, Fengyun Rao, Bernt Schiele, and Xu Yang. Number it: Temporal grounding videos like flipping manga. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. arXiv:2411.10332. 16

Jinhui Ye, Zihan Wang, Haosen Sun, Keshigeyan Chandrasegaran, Zane Durante, Cristobal Eyzaguirre, Yonatan Bisk, Juan Carlos Niebles, Ehsan Adeli, Li Fei-Fei, Jiajun Wu, and Manling Li. T*: Re-thinking temporal search for long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025. arXiv:2504.02259. 16

Yingsen Zeng, Zepeng Huang, Yujie Zhong, Chengjian Feng, Jie Hu, Lin Ma, and Yang Liu. Distime: Distribution-based time representation for video large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025. 15

Haoji Zhang, Xin Gu, Jiawen Li, Chixiang Ma, Sule Bai, Chubin Zhang, Bowen Zhang, Zhichao Zhou, Dongliang He, and Yansong Tang. Thinking with videos: Multimodal tool-augmented reinforcement learning for long video reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 32903–32914, 2026a. 16

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, 2026b. 1, 3, 15, 17, 27

Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. Learning 2d temporal adjacent networks for moment localization with natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2020. 1, 15

Yuan Zhang, Junwen Pan, Rui Zhang, Xin Wan, Qizhe Zhang, Ming Lu, Qi She, and Shanghang Zhang. ZoomV: Temporal zoom-in for efficient long video understanding. In *Proceedings of the 34th ACM International Conference on Multimedia (MM)*, 2026c. arXiv:2504.01407. 16

Appendix Contents

A Extended Discussion and Limitations **14**

B Extended Related Work **15**

C Scoring Protocol and Length Strata **17**

D Geometry of the Metrics **18**

E Zoom Ablations **19**

F Signed-Error Profile **19**

G Deconfounding Details **19**

H Scale-Fair Recall and Error Taxonomy **20**

I Dominated Remedies: Recalibration and Data Synthesis **20**

J Training-Corpus Length Support (TimeLens-100K) **22**

K Zoom-Router Rows on QVHighlights and Full-Set ActivityNet **22**

L Mixed-Corpus Router Experiment: Protocol and Sensitivity **25**

M Evaluation-Set Composition **26**

N Native Charades-STA and Scorer Validation **26**

O Zero-Shot Backbone Protocol **28**

P Prompt for the Prompting Intervention **28**

Q Oracle-Window Floors, Center-Floor Generality, Parse Rates **29**

R Robustness Checks: Held-Out Settings, Duration Scaling, Annotation Conventions, Equivalence **30**

S Decoy Control: Truth-vs-Decoy Separation **31**

T Finetuning Recipes **31**

U Input-Side Interventions: Decomposing the Crop, and Foveation **33**

V Center-Intervention Implementations **34**

W Zoom Implementation Details **35**

A Extended Discussion and Limitations

Scope of the evidence. The collapse and the length-prior diagnosis are model-agnostic, holding for three model families and every grounder we test (Table 2). The *interventional* evidence is narrower. The strongest training arm, natural-short supply, is measured on two grounders and reproduces on both (Appendix T); everything else is TimeLens-7B: a finetuning ladder from LoRA-scale to full-parameter language-model updates (the vision tower and merger stay frozen), three interventions spanning the objective (an RL center reward among them), the output parameterization and the

architecture, and the crop reversal (Secs. 4–5.1). Neither the corpus nor scale is the boundary: natural short moments repair emission without moving the center (Sec. 4), and a $10\times$ zero-shot control carries the same prior (Table 2). What remains open is *interventions* at that scale, and backbones outside those we test. Three limits remain on the lever itself. Strict recall stays capped by the residual center error, the reach on long videos is set by the first pass, and sampling-based repairs have no diversity to draw on in the grounder we probe, whose decode is peaked enough that sampled outputs are byte-identical even at temperature five (Appendix I). The crop is a bundle: it shortens the search range, raises the share of frames inside the moment, removes distractors, and rescales positions. Frame density is controlled for, since 8 fps on the full video leaves the floor untouched and a resolution-pinned arm holds tokens per frame fixed (Sec. 6), but range, distractors and rescaling are not separated from each other; "search-space limit" is our reading of the crop result, not a factorized mechanism. The title's claim is mechanistic rather than exhaustive (Appendix J), and it is falsifiable. A finetune at any scale that moved the sub-2 s center below the frozen floor would falsify this account; none of the arms here did. Annotation disagreement does not account for the deficit either: 78–83% of short-moment misses are wrong-region errors (Sec. 3), and no re-drawing of boundaries rescues a prediction that lands in the wrong part of the video. We make no claim about where a human-agreement floor lies.

Timestamp-aware decoding. Zoom presumes the model reports *absolute* original-video times for a cropped input, which TimeLens's interleaved per-frame timestamps supply. A plain decoder loses that time base; on VideoChat-R1 the emitted spans ignore the crop's extent entirely. Supplying the timestamps at inference does not fix it: the same model reads them on full videos yet ignores off-origin ones on crops, so the anchoring must be trained in. Agentic retrieve-then-ground pipelines inherit the requirement, since a retrieved clip arrives with no time base.

B Extended Related Work

From detectors to VideoLLM grounders. Temporal grounding began in the detector era: specialized architectures score or regress candidate spans, from anchor-based 2D-TAN (Zhang et al., 2020) through DETR-style set prediction (Lei et al., 2021; Moon et al., 2026) to scalable single-stage grounders (Mu et al., 2024) and unified grounding pretraining (Lin et al., 2023). VideoLLMs then reframed grounding as text generation: a multimodal LLM reads the video and emits timestamps as plain text (Huang et al., 2024a; Ren et al., 2024; Huang et al., 2024b; Qian et al., 2024; Li et al., 2025; Zhang et al., 2026b), as structured time tokens (Guo et al., 2025; Zeng et al., 2025), or through refinement and auxiliary prediction heads (Wang et al., 2026; Tao & Huang, 2026). Every entry in this now-dominant family reports accuracy as a single aggregate over all moment lengths; that convention is where this paper begins. The aggregate conceals a near-total failure on the shortest moments (Sec. 3). We study three of these grounders (Zhang et al., 2026b; Li et al., 2025; Guo et al., 2025) as analysis targets rather than competitors.

Benchmark audits and short moments. Auditing what grounding benchmarks actually measure also has a history. Otani et al. (2020) showed that detector-era models exploit annotation biases and can score well while barely using the video, and the debiased-TSGV line (Lan et al., 2023) introduced length-discounted recall and causal deconfounding of moment length. Short moments are already known to be hard for detectors: tightening an action-spotting tolerance to a few seconds costs a strong baseline twenty-five points of mAP on unchanged predictions (Giancola et al., 2022; Hong et al., 2022), and the sub-2-second bin is the largest and worst-served in fine-grained localization (Liu et al., 2022; Alwassel et al., 2018). The consequences reach past leaderboards, since under the sixteen-frame sampling common to VideoLLMs, a four-second clip can fall entirely between sampled frames, with reported harmful-content omission rates above ninety percent (Cao et al., 2026; Liu et al., 2026b). Closer to the VideoLLM era, an atemporal probe showed that video-language benchmarks are largely solvable without temporal understanding (Buch et al., 2022). We bring the same scrutiny to generative grounders, where the failure turns out to be different in kind: a length prior in the model's output distribution, present in the raw backbone before any grounding training and re-centered rather than removed by it (Sec. 4). The audit also surfaces a fact about the benchmarks themselves: the native Charades-STA test split contains no sub-2-second moment at all, so the regime where every model fails is invisible to the field's primary leaderboard (Sec. 3).

The blind spot is not confined to Charades. The benchmark assembled specifically for fine-grained temporal understanding merges any clip shorter than five seconds into its predecessor, on the stated grounds that clips too short are unsuitable for temporal grounding (Nie et al., 2024). Where a moment-duration breakdown is reported at all, the finest bin is five seconds wide, and the pattern inside it matches ours: shorter spans are harder, and a reinforcement-learned objective lifts every bin without closing the gap (Shen et al., 2025). The convention reaches the diagnostic literature as well: a benchmark built to probe temporal consistency discards any moment under five seconds before evaluation begins (Jung et al., 2025).

The convention also appears as a resolution floor in the machinery as well as in the ground truth. The training-free numbering overlay that lifts VideoLLM grounding samples frames for moment retrieval once per second (Wu et al., 2025), so a moment of the length we study is carried by one or two frames. The memory that agentic pipelines search is assembled from segments of roughly two seconds (Fan et al., 2024), which is the finest unit such a retriever can return. Keyframe search scores itself against a temporal tolerance of five seconds (Ye et al., 2025), and a recursive spotlight stops subdividing once a sub-event falls below a duration threshold (Zhang et al., 2026c). Each choice is reasonable for the task it was made for; together they place the regime we measure below the resolution of the instruments.

Reading the output distribution. Inspecting what a grounder emits, rather than only how often it is right, has precedent. Wu et al. (2025) plot the most frequent predicted intervals for several VideoLLMs on Charades-STA and find them concentrated on a few modes, in frame indices under a frame-numbering overlay, with the concern being that the spans are too simple rather than that short ones are missing. LA-DETR (Park et al., 2026) decomposes short-moment failure into center and length for detector-era regressors and reports that the predicted center falls inside the ground truth for a minority of short moments. Our measurement differs in three ways that carry the conclusion: it is in absolute seconds on a generative grounder, it compares the emitted length distribution before and after grounding finetuning (Sec. 4), and it re-measures the center after every remedy, which is what separates the emission repair from the center that no update brings below a second.

Short moments have been recognized as hard before. LA-DETR (Park et al., 2026) reports that short moments are mis-centered and mis-sized and repairs them with MomentMix augmentation at training time; LATR (Wang et al., 2025) assigns queries to duration bands with a length-classification loss. Both remedies act on detector training, where length awareness can be built into the matcher. Our diagnosis explains why the corresponding remedy fails for VideoLLMs (the training corpus contains no short moments to learn from, so rebalanced finetuning backfires; Sec. 4), and our lever instead acts at inference time on a frozen model (Sec. 6).

Temporal zoom. Re-reading a predicted window is by now a standard move, and the variants differ in ways that decide what a second pass can recover. SlowFocus (Nie et al., 2024) locates the query-relevant segment from a low-rate pass and samples densely inside it, fusing local detail with global context through a mixing-attention module. HawkEye (Wang et al., 2024b) halves the interval recursively and re-samples each level at a higher rate, and ReVisionLLM (Hannan et al., 2025) recurses over hour-long videos in the same spirit. VITAL (Zhang et al., 2026a) and VideoTempo3 (Liu et al., 2026a) expose clipping as a tool inside a reinforcement-learned reasoning loop; ZoomV (Zhang et al., 2026c) scores candidate windows by the model's own reflected confidence before spotlighting them; Zoom-Zero (Shen et al., 2025) rewards a coarse-to-fine policy for answering correctly from the crop alone; VideoMind (Liu et al., 2026c) expands each candidate moment and crops it for a verifier that accepts or rejects, rather than re-deciding the boundaries. Every one of these trains the component that proposes or consumes the window.

Two further distinctions separate the mechanisms. First, several methods presented as zoom hold the frame rate fixed and redistribute a fixed token budget over the same frames, which raises spatial resolution alone (Shen et al., 2025; Du et al., 2026); a moment shorter than the sampling interval is not recoverable that way. Second, the usual implementation rewrites the cropped clip's timestamps to start at zero and maps the prediction back (Du et al., 2026), where our fine pass leaves the original clock intact (Sec. A).

The closest published system to our lever is the attention-guided read-then-regenerate of Du et al. (2026): training-free, gated by a confidence test, and evaluated on the same three re-annotated

benchmarks with the same grounder family. It differs in where the window comes from, a sparse set of grounding-selective attention heads rather than the model's own coarse prediction, and in what changes inside it, tokens per frame rather than frames per second. Sec. 6 relates its aggregate result to ours. It reports $+0.3$ mIoU on Charades, an aggregate our own router also leaves nearly unchanged, so the two results are compatible: the effect either is built to produce is visible only under length stratification.

Localization as search. Treating localization as search has precedents on both the temporal and the spatial axis. Hour-long grounding has been decomposed into trained retrieve-then-localize pipelines (Seo & Kim, 2026), and in images, V* (Wu & Xie, 2024) equips a multimodal LLM with an explicit visual-search loop that crops toward small details. Temporal zoom (Sec. 5.2) belongs to this family with a twist that matters for the mechanism: where the retrieval pipelines train the search policy, the crop adds no trained component at all, and its success on a frozen model is what points to the center floor being a search-space limit rather than a missing capability. The same budget bottleneck drives the agentic line that pre-indexes long videos by captioning or embedding them chunk by chunk and retrieves the relevant segment at query time (Wang et al., 2024a; Fan et al., 2024; Ren et al., 2025; NVIDIA, 2025). Our analysis bounds this family on the shortest moments: the oracle-window ceiling of Sec. 5.2 is exactly the perfect-retriever case, the retrieved segment must still reach a decoder that carries absolute time (Sec. A), and the zoom is the spectrum's minimal point, with no index and the model's own coarse pass as the retriever.

Finally, our probes borrow from adjacent toolboxes: the expected-center readout is a temporal soft-argmax (DSNT) (Nibali et al., 2018), length recalibration is prior correction under label shift (Saerens et al., 2002; Lipton et al., 2018), and our error taxonomy adapts the split of (Jin et al., 2026) with a center-based cut.

C Scoring Protocol and Length Strata

Scoring protocol. Every number follows the official TimeLens-Bench evaluation protocol (Zhang et al., 2026b): unparseable outputs count as IoU 0 and every pair stays in the denominator. Stored timestamps are rounded to the benchmark's annotation unit before scoring, 1 s for the re-annotations and unrounded for ActivityNet's float ground truth. Rounding matters on short moments, so we score each benchmark at the resolution of its own annotations: the 1 s grid on Charades and QVHighlights, unrounded on ActivityNet's float spans. Two arms are compared at the finest resolution both can express, which is unrounded where both emit fractions and the grid where either emits whole seconds (Table 4, Appendix R). We validate our stratified scorer against the official implementation (Appendices D and N). The protocol pools 1,344 sub-2 s queries across the three datasets. TimeLens-Bench is a re-annotation, so our numbers compare models under a common protocol rather than against native-split leaderboards.

Length-stratified protocol. The five strata are defined in Sec. 2, and Appendix M tabulates every set's composition. We report R@1@IoU0.7 within each bucket with 95% cluster-bootstrap (by video) confidence intervals and per-bucket n and hit counts.

C.1 Scoring conventions

Four conventions govern every number we report. Each is stated because a scorer that departs from one of them silently produces a different table.

The rounding convention is fixed per file. It follows the measured integrality of that file's ground truth rather than the dataset's name. Charades-TimeLens and QVHighlights-TimeLens are annotated on whole seconds, so predictions are read on the 1 s grid before scoring. Native ActivityNet and native Charades-STA carry float annotations, so predictions are read as emitted.

The grid and the statistic are separate choices. The convention above fixes the grid a prediction is read on. It does not round the statistic computed from it, and center errors stay unrounded throughout. A median center error of 1.00 s on QVHighlights is an unrounded statistic over pre-

dictions read on the grid, and reading those same predictions as emitted gives $0.90\,\mathrm{s}$. Both are unrounded statistics, and they differ in the grid beneath them.

Provenance decides whether two arms are comparable. The official runner rounds timestamps as it writes them, while the in-session harnesses store what the model emitted. A file of the first kind has integral stored spans, so the emitted values cannot be recovered from it without re-parsing the generated text. Scoring an arm of one kind against an arm of the other measures the writer rather than the model, so we compare arms at the finest resolution both express.

The annotation lattice can pin a median. Every sub-2 s span in Charades-TimeLens has unit length and integer endpoints, so its centers are half-integers. A model emitting whole seconds then places $|\Delta c|$ on a half-second lattice, where the median steps a full half-second as a few percent of the mass crosses it and can express nothing finer. On such a file we report rates and means alongside, and we use the reachability bound $d_{\max} = L_{\mathrm{gt}}(1-t)/(2t)$, the largest center error from which overlap t stays attainable, which is $0.214\,\mathrm{s}$ for a one-second moment at $t{=}0.7$.

D Geometry of the Metrics

The contraction is not a noise illusion. A naïve predicted-vs-true length plot can show "compression" even from an unbiased but noisy predictor; β is a slope in raw seconds, so a faithful predictor gives $\beta = 1$ however noisy it is. Against a conservative unbiased-twin null (a synthetic predictor that adds the model's full error spread to the *true* length; boundary clipping alone pulls its slope to $\beta \approx 0.92$), the models' β is 0.60 (TimeLens on Charades), 0.48 (TimeLens on ActivityNet), and 0.31 (VideoChat-R1 on Charades), each with a video-clustered bootstrap CI entirely below 0.92: the contraction is learned.

These four elementary statements are used throughout the main text; we collect them with proofs.

Proposition 1 (threshold–center bound). Let the ground-truth span have length L and the prediction length ℓ, with center offset d. Then $\mathrm{IoU} \geq t$ if and only if $tL \leq \ell \leq L/t$ and $d \leq (\ell + L)(1-t)\,/\,(2(1+t))$. *Proof.* When the spans overlap, the intersection is $I = \min\big((\ell{+}L)/2 - d,\ \ell,\ L\big)$ and the union is $\ell + L - I$; $\mathrm{IoU} = I/(\ell + L - I)$ is increasing in I, so $\mathrm{IoU} \geq t \iff I \geq t(\ell + L)/(1+t)$. If $d \leq |L-\ell|/2$ (one span contains the other's center region), $I = \min(\ell, L)$ and the condition reduces to $\min(\ell, L)/\max(\ell, L) \geq t$, i.e. the width window; the two conditions meet exactly at $d = |L-\ell|/2$. Otherwise $I = (\ell{+}L)/2 - d$, giving the stated bound. □

For $\ell = L$ this is $d \leq L(1-t)/(1+t)$: at $t{=}0.5$ a tolerance of $L/3$, at $t{=}0.7$ of $0.18L$: on a 1 s moment, $0.33\,\mathrm{s}$ and $0.18\,\mathrm{s}$ respectively. Strict thresholds therefore tighten *linearly in* L, which is the mechanical component of the short-moment collapse quantified by the scale-fair re-scoring in Sec. 3.

Proposition 2 (grid alignment). Let ground truth be the unit integer span $[k, k{+}1]$ and the prediction a unit-length span with center offset $d < 1$. Rounding both predicted endpoints to the nearest integer maps the prediction to $[k, k{+}1]$ exactly when $d < 1/2$, so rounded IoU is the indicator $\mathbf{1}[d < 1/2]$. *Proof.* A unit span with center offset $d \in [0, 1/2)$ has endpoints in $[k - \frac{1}{2}, k + \frac{1}{2}) \times [k + \frac{1}{2}, k + \frac{3}{2})$, which round to k and $k{+}1$. □

Unrounded, Prop. 1 gives acceptance $d \leq 1/3$ at $t{=}0.5$ and $d \leq 0.18$ at $t{=}0.7$; rounding enlarges both acceptance regions to $d < 1/2$, by $1.5\times$ and $2.8\times$ respectively, for grid-aligned ground truth. This is the mechanism behind the protocol-sensitivity note of Sec. 2 and the reason strict recall benefits most.

Proposition 3 (window floors). The whole-window policy on an oracle window of width W containing a moment of length L scores $\mathrm{IoU} = L/W$ deterministically; its recall at threshold t over a stratum is therefore $\Pr[L \geq tW]$. *Proof.* Intersection L, union W. □

For the sub-2 s stratum ($L < 2$), $W{=}8$ gives a floor of 0 at $t \geq 0.25$, while $W{=}2$ gives $\Pr[L \geq 1] = 1$ at $t{=}0.5$: tighter crops make the trivial policy saturate the lenient metric, which is why $W \leq 3$ is excluded as geometrically degenerate in Sec. 5.

Lemma 4 (censoring attenuates the length slope). Let $\hat{y} = \mathrm{clip}_{[0,D]}(y + \varepsilon)$ with $\mathbb{E}[\varepsilon \mid y] = 0$. Then the OLS slope of $\hat{y}$ on y is at most 1, with equality only if clipping is almost surely inactive. *Proof.* $g(y) = \mathbb{E}[\mathrm{clip}(y + \varepsilon) \mid y]$ is 1-Lipschitz because clip is, so $\mathrm{Cov}(g(y), y) \leq \mathrm{Var}(y)$ and the population slope $\mathrm{Cov}(\hat{y}, y)/\mathrm{Var}(y) \leq 1$; strict whenever the clip binds with positive probability. □

This is the unbiased-twin null of Sec. 3: boundary clipping alone pulls a faithful predictor's slope to ≈ 0.92 on our data, so the models' $\beta = 0.31$–0.60 cannot be regression to the mean.

E Zoom Ablations

Table A1 varies the zoom's knobs at the oracle-window ceiling (moved here from Sec. 6 for space). Width is mild (± 3 points over $W \in [4, 12]$ s; $W \leq 3$ is excluded because trivial in-window policies saturate the lenient metrics there). The fine pass's frame rate decomposes the gain: a $W{=}8$ crop at the *original* 2 fps already recovers most of the lenient gain (R@1@IoU0.5 0.400), so the crop rather than the density does that work, while strict recall peaks at 8 fps ($0.144 \rightarrow 0.202$) and then falls as further density costs per-frame resolution. The two components map onto the two mechanisms: the crop supplies the missing search-space condition of Sec. 5, and the in-window density supplies the emission repair of Sec. 4; the deployed zoom addresses both. Doubling the token budget changes nothing: the gains come from reallocating a fixed budget rather than from spending more. Placement within the window is mild (0.419 vs 0.386 R@1@IoU0.5 at 0.4 vs 0.6). The deployed knobs (the router threshold τ, and the form of the second look, crop against denser re-sampling) are ablated in Table A6.

F Signed-Error Profile

Table A2 is the full per-bin signed-error profile summarized by the compression-fingerprint figure of the main paper (TimeLens, Charades): the median length ratio crosses zero near 3–7 s (compression at both extremes) while the absolute center error stays ≈1 s at every length below 15 s (2 s beyond), so the *relative* center error explodes as moments shorten.

G Deconfounding Details

Estimators. The three confounds are $c_1 = \log$ video duration, $c_2 =$ query difficulty, $c_3 =$ moments per second of video. Generalized estimating equations (GEE): cluster-robust logistic regression grouped by video ($n{=}3{,}363$ pairs, 1,313 videos), outcome = hit at IoU ≥ 0.7: shortness log-odds -1.45 (95% CI $[-1.65, -1.25]$), $p < 10^{-10}$. MixedLM: per-video random intercept, outcome = continuous IoU, shortness coefficient -0.20 (95% CI $[-0.23, -0.17]$, $p < 10^{-10}$). CEM: coarsened exact matching on a $3 \times 3 \times 3$ quantile grid over (c_1, c_2, c_3): 27 matched strata cover all 3,363 pairs (nothing pruned); within matched cells, long beats short by $+0.32$ R@1@IoU0.7, with a 95% video-clustered bootstrap interval of $[+0.27, +0.38]$. Coarsening into three bins per covariate cuts the short-versus-long standardised mean difference on (c_1, c_2, c_3) from 0.38, 0.16 and 0.54 to 0.15, 0.13 and 0.15 within matched cells. The imbalance the matching exists to remove is therefore much reduced, though it stays above the conventional threshold of 0.1 on every covariate, so this estimator is corroboration rather than a clean design.

Graded effect and units. The continuous-length GEE (video-clustered, log-duration term included) at the strict threshold gives $+1.07$ logits per log-second on Charades ($p < 10^{-10}$) and $+0.34$ on ActivityNet ($p \approx 1\mathrm{e}{-06}$). Adding a quadratic log-length term leaves the linear coefficient's CI strictly positive on both datasets (Charades $+2.54$ $[2.14, 2.95]$; ActivityNet $+1.88$ $[1.15, 2.61]$; concave, no turning point): the rise is graded and monotone, not a step.

Query-difficulty proxy robustness. c_2 was originally query word count; we recomputed it as a grounder-independent *CLIP span-spread* difficulty and re-ran all three estimators: GEE odds ratio

Table A1: Zoom ablations at the oracle-window ceiling (Charades sub-2 s stratum, n=590; official scoring convention). One knob varies per block; the default row repeats for reference. $W \leq 3$ is excluded: trivial in-window policies saturate the lenient metrics there. Deployed knobs (τ; crop vs. denser second look) are in Table A6.

knob	setting	R@1@IoU0.5	R@1@IoU0.7	mIoU
width W	4 s	0.439	0.231	0.363
	8 s (default)	0.386	0.202	0.334
	12 s	0.378	0.185	0.326
fine-pass fps	2	0.400	0.144	0.360
	8 (default)	0.386	0.202	0.334
	16	0.344	0.195	0.314
	32	0.319	0.185	0.289
token budget	14,336 (default)	0.386	0.202	0.334
	28,672	0.386	0.195	0.331
placement	0.6 (default)	0.386	0.202	0.334
	0.4	0.419	0.231	0.364

Table A2: Signed error vs. moment length ($\widetilde{\cdot}$ = median; TimeLens-7B, Charades). $\log_2(L_{\text{pred}}/L_{\text{gt}}) > 0$ means *too long*.

L_{gt}	n	$\widetilde{\log_2 \frac{L_{\text{pred}}}{L_{\text{gt}}}}$	% long	$\widetilde{\lvert\Delta c\rvert}$	$\widetilde{\frac{\lvert\Delta c\rvert}{L_{\text{gt}}}}$	R@1@IoU0.7
$[0, 2)$	590	$\mathbf{+1.58}$	90	1.00	**1.00**	4.1
$[2, 3)$	709	$+0.58$	62	1.00	0.50	12.8
$[3, 5)$	765	$+0.00$	48	1.00	0.25	27.8
$[5, 7)$	427	$+0.00$	33	1.00	0.20	36.5
$[7, 10)$	329	-0.17	35	1.00	0.12	50.8
$[10, 15)$	265	-0.15	26	1.00	0.09	58.9
$[15, +)$	278	-0.22	19	2.00	0.09	54.0

0.24, MixedLM -0.20, CEM $+0.31$ (unchanged conclusions), although the two proxies are near-uncorrelated ($\rho = -0.09$).

Video-length deconfound. Charades clips are nearly constant duration (∼30 s), so c_1 barely varies there; the genuine video-length deconfound comes from ActivityNet, where the continuous $\log L_{\text{gt}}$ effect on recall remains positive and graded (slope $+1.3$, $p \approx 1\text{e}{-17}$) with a log-duration term in the model.

H Scale-Fair Recall and Error Taxonomy

Table A3: Scale-fair recall (TimeLens, Charades), the numeric form of the band in Fig. 2a: a hit requires both boundaries within $\alpha \cdot L_{\text{gt}}$ of ground truth, so tolerance grows with length. The collapse is monotone across the four buckets below 30 s at every α; the ≥30 s dip is the prior's other tail.

α	$[0, 2)$	$[2, 5)$	$[5, 10)$	$[10, 30)$	$[30, \infty)$
0.15	0.041	0.117	0.243	0.400	0.355
0.25	0.041	0.191	0.447	0.559	0.387
0.50	0.041	0.537	0.712	0.783	0.548

I Dominated Remedies: Recalibration and Data Synthesis

Two remedies that respect the support gap repair *emission* but not accuracy; both are dominated by the levers of Sec. 6 of the main paper on every axis and are reported here for completeness.

(A) Decode-time length recalibration (training-free). A wrapper rescales each emitted span's *length* around a fixed midpoint by quantile-matching the predicted-length distribution onto the true-length distribution (fit on a calibration split, 5-fold by video, no leakage). Framed properly this is

Table A4: Error taxonomy at IoU 0.5, as shares of all predictions summing to one within each cell: hit, Type-A (predicted midpoint outside the moment), Type-B (midpoint inside, loose boundary).

	TimeLens (Charades)			TimeLens (ANet)			VideoChat-R1		
L_{gt}	hit	A	B	hit	A	B	hit	A	B
<2 s	0.26	0.59	0.15	0.08	0.76	0.16	0.00	0.78	0.22
2–5 s	0.53	0.33	0.15	0.44	0.45	0.11	0.24	0.39	0.37
5–10 s	0.72	0.14	0.14	0.49	0.28	0.23	0.79	0.13	0.08
10–30 s	0.74	0.06	0.21	0.49	0.33	0.18	0.73	0.07	0.20
≥30 s	0.52	0.00	0.48	0.31	0.25	0.44	0.10	0.00	0.90

prior recalibration / label-shift correction (Saerens et al., 2002; Lipton et al., 2018). It restores much of the length distribution (β 0.60 $\rightarrow$ 0.68) and lifts short strict recall through the same shorter-span mechanism as denser sampling, but by less and at a larger mid/long redistribution cost. Its limit: even with length corrected, short strict recall stays capped, because the binding constraint is the $\sim$1 s center error a length-only fix cannot touch.

(B) Speed-augmentation finetune (data synthesis). Temporally speeding up training clips ($2\times$, $4\times$, by physically re-encoding) yields 4,060 genuine sub-2 s moments. Finetuning on them restores short *emission* (0.0% $\rightarrow$ 2.0% on sub-2 s cases; 0.8% of all predictions), which is consistent with the support-gap diagnosis, but not short *accuracy*: S0 R@1@IoU0.5 (0.047) stays far below the frozen baseline (0.259), β stays compressed (0.23), and spans remain mis-centered ($|\Delta c| = 1.50$ s). Genuine short data is necessary but not sufficient.

(C) Natural-short-corpus finetune (the strongest corpus test). A matched-budget LoRA (identical recipe to the Table 3 arms) trained on the ActivityNet Captions train short tail: 3,913 pairs with the speed-augmented arm's length mix (38% sub-2 s, all 1,487 genuine natural sub-2 s segments; the longer strata seed-sampled to match proportions). *Cross-corpus (Charades, n=3363):* sub-2 s emission 10.2 $\rightarrow$ 54.1% on the ultra-short cases (3.6 $\rightarrow$ 29.1% corpus-wide), median sub-2 s $|\Delta c|$ essentially unchanged, 1.05 $\rightarrow$ 1.00 s scored unrounded on both sides, R@1@IoU0.7 0.041 $\rightarrow$ 0.156 (paired 95% CI $[+8.1, +14.8]$ points; span-shortening arithmetic on unmoved centers, cf. the denser-sampling mechanism), R@1@IoU0.5 also up (0.259 $\rightarrow$ 0.324, CI $[+2.3, +10.7]$), overall R@1@IoU0.7 0.285 $\rightarrow$ 0.231 (CI $[-7.0, -3.7]$), 30 s+ collapses ($\widetilde{|\Delta c|}$ 8.5 $\rightarrow$ 14.5 s), β 0.60 $\rightarrow$ 0.30. *In-distribution (ActivityNet, full native split, n=17,505):* sub-2 s R@1@IoU0.7 0.051 $\rightarrow$ 0.151 (CI $[+7.2, +12.7]$) and R@1@IoU0.5 0.144 $\rightarrow$ 0.320 (CI $[+13.9, +21.4]$), median sub-2 s $|\Delta c|$ 1.83 $\rightarrow$ 0.76 s (CI $[-1.32, -0.69]$), level with the zoom ceiling's 0.74 s on the same clips, and the share inside the strict center window doubles, 13 $\rightarrow$ 26% (CI $[+9.5, +16.7]$). The cost is paid in the long strata: 30 s+ R@1@IoU0.7 collapses 0.229 $\rightarrow$ 0.092 (CI $[-14.7, -12.6]$), overall R@1@IoU0.7 falls 0.242 $\rightarrow$ 0.200 (CI $[-4.9, -3.6]$), and the corpus-wide share inside the center window falls 41.2 $\rightarrow$ 34.3% (CI $[-7.6, -6.0]$). Natural short support repairs emission everywhere. In distribution it repairs placement too, down to the oracle crop's median, which the crop reaches without touching the long strata; cross-corpus that placement gain does not transfer, and either way the lose-lose trade returns.

(D) Natural-distribution control (the same corpus without the short tail). The control that (C) never had: identical recipe, budget and corpus, drawn at ActivityNet's own length distribution instead of its short tail, so the training mix holds 4.0% sub-2 s content (155 segments) against (C)'s 38% (1,487). Matched from artifacts rather than intent: 200 steps, global batch 64, lr 10^{-4}, r=64, and 3.2808 epochs against (C)'s 3.2822. *Cross-corpus (Charades, n=3363):* sub-2 s emission 10.2 $\rightarrow$ 30.2% on the ultra-short cases, so the short tail itself supplies under half of (C)'s emission gain. β rises 0.60 $\rightarrow$ 0.71, the best length calibration of any arm we run and the only one that moves β toward 1, ahead of the explicit recalibrator of (A). R@1@IoU0.7 rises 0.041 $\rightarrow$ 0.080 while R@1@IoU0.5 *falls* 0.259 $\rightarrow$ 0.166, which is the signature of tighter spans on centers that did not improve: over-long spans graze a short moment and collect credit at the lenient threshold, tight ones in the wrong place do not. Overall R@1@IoU0.7 0.285 $\rightarrow$ 0.264 and mIoU 0.490 $\rightarrow$ 0.441.

Placement does not follow emission. The share of sub-2 s clips centered inside the 0.214 s strict window is 11.2% for the frozen model and 11.0% here, and the mean $|\Delta c|$ rises 2.67 $\rightarrow$ 3.27 s. We

quote those rates rather than the median because the frozen model's spans reach us rounded, putting every frozen $|\Delta c|$ on a 0.5 s lattice where a median pins and cannot express a small change.

One account this control rules out. Both ActivityNet arms were expected to raise emission partly by teaching sub-second output granularity, since the frozen predictions on file carry whole-second endpoints on every clip while these arms carry fractional ones. The frozen file is the only one written by the official runner, which rounds on write, and the recovered emissions behind it are fractional on 97.5% of clips. The granularity account is therefore an artifact of provenance, and the emission gain here is not attributed to it.

Decode determinism under grounding SFT. Sampling-based repairs (best-of-N, self-consistency, sampled multi-hypothesis proposals) presume the decoder has answer diversity to offer. The frozen grounder does not: drawing sampled decodes (K=5, top-p 0.95) at temperatures 1.0, 1.5 and 5.0 on sub-2 s Charades clips yields outputs byte-identical to the greedy decode in every case we probe (95 decodes over 11 clips, 8 from the far tail and 3 from a smoke set; not every clip at every temperature), including clips whose prediction is $\sim$8 s off; a text-only prompt through the same generation path does vary, confirming sampling is active. Grounding SFT has collapsed the answer distribution to a point, so for this model candidate diversity for any multi-hypothesis search must come from elsewhere (beam diversification, re-prompting, or external signals). The unchanged greedy decode of the GRPO arm (main text, Sec. 5.1) is consistent with rollouts drawn from this near-degenerate policy.

J Training-Corpus Length Support (TimeLens-100K)

Figure 2(d) shows the support gap as cumulative length distributions; the numbers below are the support-gap measurement of Sec. 4 of the main paper, computed directly on the grounder's own instruction corpus.

One reading the main-text caption leaves out: the natural-short ActivityNet tail (dashed) is the one corpus whose short mass exceeds the evaluation demand, and finetuning on it still leaves the center floor unmoved (arm (C) above). From the released TimeLens-100K annotation file ($19{,}466$ videos, $96{,}586$ grounded spans across seven sources), $3{,}525$ spans (3.65%) are sub-2 s and *no span is shorter than* 1 *s* (corpus minimum 1.00 s); the 5th percentile is 2.0 s and the median is 5.0 s, the modal length onto which the grounder's SFT re-centers the backbone prior. For comparison, the sub-2 s share of the Charades evaluation set is 17.5% ($590/3{,}363$), and the Charades-STA training split (the base of the rebalance corpora of Table 3) contains zero.

Table A5: Sub-2 s support in TimeLens-100K by source. No source exceeds 7.5%, and the corpus contains no span under 1 s.

source	spans	sub-2 s	share
cosmo_cap	47,482	818	1.7%
internvid_vtime	23,007	1,724	7.5%
didemo	14,636	586	4.0%
queryd	7,532	346	4.6%
hirest (3 subsets)	3,929	51	1.3%
all	96,586	3,525	3.65%

K Zoom-Router Rows on QVHighlights and Full-Set ActivityNet

Table A7 completes the benchmark coverage for the deployable two-pass zoom (coarse pass $\rightarrow$ route if the *predicted* length is $< \tau$=4 s $\rightarrow \pm 4$ s window re-decoded at 8 fps). The ActivityNet run is the strongest deployment test in the paper: the router sees the *entire* validation set ($17{,}505$ pairs) with no ground-truth length filter, so its false positives are measured rather than excluded. It routes $1{,}837$ pairs (10.5%), of which 951 have true moments $\geq$4 s; measured on those 951 alone the mis-routed clips gain rather than lose, R@1@IoU0.5 $0.103 \rightarrow 0.121$, so the no-regression property survives contact with the router's own mistakes. On QVHighlights the gate likewise sees the full split: it routes 151 of 1541 pairs, 95 of them with true moments $\geq$2 s, and those mis-routed clips do lose

Table A6: Deployable routers on all three benchmarks: a clip gets a second, concentrated look only when the first-pass *predicted* length is short ($L_{\text{pred}} < \tau$). The router-gated zoom (± 4 s window re-decoded at 8 fps) deploys on all three datasets (41.5% / 9.8% / 10.5% routed); the denser-sampling rows ablate the form of the second look. On the grid benchmarks the lenient sub-2 s column cannot separate a hit from a one-second-too-wide envelope: no IoU falls in $(0.5, 0.7)$ there. ActivityNet's denser arm covers all 662 sub-2 s pairs but only 2,002 overall, so its aggregate cells are omitted rather than compared across populations. ActivityNet is the *full* validation set with no ground-truth filter, so the router's false positives are measured: the 951 routed clips whose true moment is $\geq$4 s lose nothing (rows below). Short recall rises on every dataset; the aggregate is unchanged on QVHighlights and ActivityNet and costs $\sim$2 mIoU points on Charades.

		$[0, 2)$ s		overall	
		R@1@IoU0.5	R@1@IoU0.7	R@1@IoU0.7	mIoU
Charades n=3363	baseline	0.259	0.041	0.285	0.490
	zoom router τ=4	**0.271**	**0.146**	**0.294**	0.471
	denser router τ=4	0.234	0.119	0.289	0.464
	denser router τ=5	0.268	0.134	0.287	0.459
QVHighlights n=1541	baseline	0.196	0.076	0.449	0.572
	zoom router τ=4	**0.337**	**0.098**	0.450	0.573
	denser router τ=4	0.250	0.043	0.449	0.572
ActivityNet (full) n=17505	baseline	0.144	0.051	0.242	0.422
	zoom router τ=4	**0.172**	0.053	0.244	0.423
	denser router τ=4	0.156	0.050	—	—

3.1 points of R@1@IoU0.5, so its unchanged aggregate is measured rather than a consequence of scoping. What limits the gain is the coarse pass rather than the fine decode: window containment on routed ActivityNet clips is 852 full, 559 partial and 426 miss, so on 426 of 1,837 routed clips the window the fine pass is given does not contain the moment at all, and on over half it does not contain it fully. Conventions as in the main text: QVH rows follow the official 1 s-unit protocol, ActivityNet rows its official unrounded scoring; center errors are unrounded throughout.

Table A7: Deployable zoom-router rows on the two remaining benchmarks (cf. the Charades router table in Sec. 6 of the main paper). QVHighlights: full split, no ground-truth filter (router false positives included). ActivityNet: full validation set, no ground-truth filter (router false positives included). The ActivityNet 2–5 s stratum (n=1200, 508 routed) moves within noise (mIoU 0.381 $\rightarrow$ 0.378) and is omitted for space. Aggregates are in Table A6.

	stratum / arm	mIoU	R@0.3	R@1@IoU0.5	R@1@IoU0.7	$\widetilde{\lvert\Delta c\rvert}$
QVH	<2 s coarse (n=92)	0.250	0.370	0.196	0.076	0.90
	<2 s router	**0.277**	0.348	**0.337**	**0.098**	**0.75**
A-Net	<2 s coarse (n=662)	0.216	0.313	0.144	0.051	1.83
	<2 s router	**0.226**	**0.338**	**0.172**	0.053	1.81
	$\geq$4 s coarse (n=15643)	0.434	0.610	0.427	0.254	7.72
	$\geq$4 s router	0.434	0.610	0.428	0.256	7.70

Freezing the router's threshold. The threshold above was chosen on the sets it is reported on. We therefore repeated the experiment with it frozen: for each benchmark we split the videos in half, selected τ on one half by short strict recall alone, and evaluated once on the other. Charades keeps its gain, sub-2 s R@1@IoU0.7 0.050 $\rightarrow$ 0.137 with a paired interval of $[+4.1, +13.3]$ points, and pays 2.7 points of overall mIoU, which exceeds the two-point bound we set in advance. QVHighlights and ActivityNet stay inside that bound, at -0.3 and -0.1 points, but their strict changes, $+0.027$ and $+0.003$, have intervals covering zero. Their held-out short strata hold 37 and 329 clips against Charades' 300, so the QVHighlights result is uninformative rather than negative. Selection also chose a different threshold on each benchmark, 3.5, 3.5 and 2.0, which is what one expects if the right threshold tracks a corpus's length distribution, and is a further reason not to read the gate as one setting that travels.

K.1 The Ungated Arm and the Gate's Protection

Section 6 gates the second pass on the first pass's predicted length. Removing that gate is the ablation that prices it. We run it on every clip of both benchmarks rather than on a ground-truth-selected subset, so the cost falls where a deployment would meet it.

Table A8: Zooming every clip, against the frozen model and the deployed τ=4 s gate, by ground-truth length stratum. Recalls and mIoU on the shared 1 s grid, one population per benchmark and one convention throughout. Median centre errors are omitted: every arm here is fully on the 0.5 s lattice, where a median cannot express a small change. The two benchmarks are not pooled, since 57.8% of QVHighlights clips lie beyond the crop's reach against 12.5% of Charades'.

		R@1@IoU0.7			mIoU		
L_{gt}	n	frozen	gated	ungated	frozen	gated	ungated
Charades-TimeLens							
<2 s	590	0.041	0.146	0.175	0.250	0.261	0.291
2–5 s	1474	0.206	0.182	0.195	0.465	0.423	0.423
5–10 s	756	0.427	0.435	0.320	0.605	0.595	0.505
10–30 s	512	0.574	0.574	0.113	0.666	0.664	0.381
≥30 s	31	0.387	0.387	0.032	0.584	0.582	0.317
overall	3363	0.285	0.294	0.206	0.490	0.471	0.411
QVHighlights-TimeLens							
<2 s	92	0.076	0.098	0.130	0.250	0.277	0.324
2–5 s	268	0.209	0.201	0.328	0.464	0.456	0.509
5–10 s	207	0.527	0.531	0.420	0.595	0.596	0.515
10–30 s	525	0.632	0.632	0.055	0.689	0.689	0.285
≥30 s	449	0.419	0.419	0.018	0.555	0.555	0.169
overall	1541	0.449	0.450	0.145	0.572	0.573	0.324

What zooming every clip costs. It buys 17 more strict hits on Charades' sub-2 s stratum and gives up 297 across the benchmark; on QVHighlights it buys 3 and gives up 469. Overall R@1@IoU0.7 falls 8.8 points on Charades, $[-10.2, -7.4]$, and 30.4 on QVHighlights, $[-33.2, -27.7]$. The loss is not confined to the long end: Charades gives up 87 strict hits in the five-to-ten second stratum and 236 of 294 in the ten-to-thirty, and QVHighlights loses 303 of 332 mid-length hits and 180 of 188 long ones. The gated arm leaves the two longest strata of both benchmarks unchanged, which is what the gate is for.

Reach, and how well the gate tracks it. A window of width W over a moment of length L_{gt} caps overlap at W/L_{gt}, so IoU t is unreachable beyond $L_{\text{gt}} = W/t$, or 11.4 s at the strict threshold with $W = 8$. The bound is close but not hard, since a cropped decode is not confined to its window: emissions leave the crop on 8.1% of Charades clips and 30.8% of QVHighlights ones, and exceed W outright on 26.3% of QVHighlights emissions, which is how the longest stratum clears a zero ceiling on a handful of clips. The gate declines most of what the bound places out of reach, 93.6% of the 420 Charades clips above it and 98.1% of QVHighlights' 891, and of the clips it does route only 27 of 1,396 and 17 of 151 are unreachable. A threshold on predicted length is therefore a serviceable proxy for a bound on true length, and it is cheaper: zooming every clip costs 1.41× the gated arm's decodes on Charades and 1.82× on QVHighlights.

Threshold selection and where the gain lands. Two notes on the main-paper router table (Sec. 6). First, the gains land on the strict metric: on a ∼1.5 s moment, collapsing an over-long span onto roughly the right center clears the 0.7 bar far more often even with the center unmoved; lenient short recall moves little at either threshold (0.259 → 0.271 at τ=4, 0.268 at τ=5). Second, τ was selected on the same full set (no held-out calibration), so the configuration is exploratory; the held-out router experiment is Sec. L.

K.2 The Router on a Second Grounder

Section 6 reports the router on TimeLens-7B. Run unchanged on TimeLens-8B it behaves in the same direction and costs less, which is the narrowest useful test of whether the lever is a property of one checkpoint.

Table A9: The deployed router at τ=4 s on TimeLens-8B, against its own first pass. Columns are R@1@IoU{0.3,0.5,0.7} and mIoU, on the shared 1 s grid. ActivityNet is absent rather than negative: the re-annotated videos were not on disk at the time of this run, and the native split has no 8B first pass.

	sub-2 s stratum				whole benchmark			
	R@.3	R@.5	R@.7	mIoU	R@.3	R@.5	R@.7	mIoU
Charades-TimeLens, 38.4% routed at 1.38× compute								
first pass	0.527	0.312	0.008	0.279	0.764	0.626	0.357	0.553
router τ=4	0.527	0.469	0.159	0.382	0.764	0.651	0.389	0.570
QVHighlights-TimeLens, 12.2% routed at 1.12× compute								
first pass	0.489	0.380	0.033	0.279	0.801	0.714	0.557	0.655
router τ=4	0.500	0.446	0.250	0.405	0.799	0.714	0.579	0.664

The 8B's own crop ceiling, and how it differs from the 7B. The oracle-window rows of Table 5 on the 8B, sub-2 s stratum of Charades: the first pass reaches R@1@IoU0.5 0.312, R@1@IoU0.7 0.008, mIoU 0.279; the crop on the model's own centre reaches 0.568, 0.185, 0.426; an oracle $W = 8$ window reaches 0.646, 0.217, 0.489. Two differences from the 7B, both in the lever's favour: the 8B pays no aggregate price where the 7B gives up 1.9 points of overall mIoU, and it keeps lenient recall instead of trading it for strict, so the threshold geometry of Sec. 6 does not bite. The frozen threshold behaves the same way, lifting held-out sub-2 s R@1@IoU0.7 0.013 → 0.180 while overall mIoU rises 0.554 → 0.571, against the 2.7 points the 7B freeze costs.

The binding constraint is emission, not the centre. The 8B first pass already places 11.2% of sub-2 s centres inside the strict window, but emits no span of one second in 590 clips, its modal emission being two seconds. The router lifts sub-2 s emission 1.2 → 20.8%, and that is where its strict recall comes from, so the two grounders fail the same stratum through different halves of the same pair. Dropping the gate reaches 0.185 against the gated 0.159, the same direction and magnitude as the 7B's 0.175 against 0.146. The 8B emits whole-second endpoints on every record of all six files, its stored spans agreeing with its own decoded answers throughout, so unlike the 7B this is a property of the model rather than of the runner that wrote the file, and putting both arms on the 1 s grid moves no prediction. Its $|\widetilde{\Delta c}|$ is therefore floored at 0.50 s and is not comparable across models at that resolution.

L Mixed-Corpus Router Experiment: Protocol and Sensitivity

Table A10: Lever selection on a mixed corpus (Charades + ActivityNet, held-out by-video split, gate calibrated on the other half). Denser sampling helps only Charades and fails ActivityNet outright; the zoom helps both. The ratio-router picks the denser span on 82% of clips and inherits that failure. ActivityNet contributes 100 sub-2 s clips here rather than the 17,505 pairs of Table A6: this comparison needs both levers, and its denser arm was run only on the 500-pair subset used for cross-arm ActivityNet comparisons. *All-length cells for the zoom arms cover only the clips a zoom prediction exists for, a set fixed by ground-truth length, so they are not deployable rates; every other cell is.

	<2 s R@1@IoU0.7			all-len
arm	agg	Charades	ActivityNet	R@1@IoU0.7
baseline	0.045	0.050	0.018	0.284
always-denser	0.138	**0.163**	0.000	0.276
always-zoom	0.144	0.157	**0.073**	0.302*
ratio-router	0.135	0.160	0.000	0.285
oracle-lever	**0.239**	**0.267**	**0.091**	0.397*

Protocol. Corpus: Charades (full, n=3363; 590 sub-2 s) + ActivityNet (n=500; 100 sub-2 s), pooled and scored with the official convention. Zoom predictions cover the sub-2 s strata (Charades 590/590; ActivityNet 97/100); uncovered queries fall back to baseline, as in deployment. Videos are split deterministically into calibration and test halves (by video, no leakage); gates are fit on calibration (maximizing sub-2 s R@1@IoU0.7 subject to no regression in all-length R@1@IoU0.7) and frozen before scoring the test half.

Why the ratio gate fails. The sub-2 s moment-to-video ratio distributions overlap almost entirely across regimes: Charades $\hat{r}$ p10/50/90 $= 0.057/0.097/0.225$; ActivityNet $= 0.013/0.068/0.294$. The calibrated router pins the zoom threshold at the grid floor and routes on the denser threshold alone, picking the denser span on 82% of clips. Whether the overlap causes that we have not shown: refitting the same objective against zoom values scored on the grid selects the top of the threshold range instead, and the router that results helps both sources.

Why a length-only gate fails. Gating on predicted length alone (route to zoom when $L_{\text{pred}} < 4$ s, ignoring video duration) cannot tell *which* lever a short-predicted query needs: on the test split it scores sub-2 s R@1@IoU0.7 0.042 against the baseline's 0.045, and drops R@1@IoU0.5 $0.234 \rightarrow 0.141$ by sending short-video queries to zoom, against the ratio-router's 0.135 and the per-query oracle's 0.197.

Why a duration gate also fails. Video duration separates the *datasets* (Charades p95 $= 41.7$ s vs. ActivityNet median $= 120.5$ s) but not the *queries*: ActivityNet's own sub-2 s moments are duration-mixed (p10/50/90 $= 11/89/227$ s), and routing short-video queries to denser sampling fails the all-length no-regression guard (calibration all-length R@1@IoU0.7 $0.284 \rightarrow 0.262$ at every duration threshold tried, 90–150 s), because denser sampling trades spatial resolution and costs mid/long moments.

Threshold sensitivity. Across the calibration grid ($\tau_{\text{zoom}} \in [0.01, 0.2]$, $\tau_{\text{dense}} \in [0.05, 1.0]$), sub-2 s R@1@IoU0.7 varies smoothly in τ_{dense} (from 0.021 at 0.05 to 0.125 at ≥ 0.5) and is insensitive to τ_{zoom}; configurations with $\tau_{\text{dense}} \geq 0.5$ violate the no-regression guard. The zoom-side gain concentrates in ActivityNet's medium-length videos: on the $\leq$90 s half, zoom lifts sub-2 s mIoU $0.157 \rightarrow 0.226$ (R@1@IoU0.7 $0.000 \rightarrow 0.075$); on the $>$90 s half it is flat ($0.180 \rightarrow 0.189$), bounded by coarse-stage containment.

M Evaluation-Set Composition

Table A11 gives the length-stratified composition of every evaluation set in the paper (moved here from the main text for space).

Table A11: Length-stratified composition of every evaluation set (S0 $<$2 s / S1 2–5 s / S2 5–10 s / M 10–30 s / L $\geq$30 s). The two Charades rows share the same 1,313 videos; only the annotation layer differs, and the native split's empty S0/L columns against the re-annotation's 590 and 31 are the benchmark-blindness finding of main-text Sec. 3 in tabular form. "GT": native annotations vs. the TimeLens-Bench re-annotation; each dataset uses the finest-granularity ground truth available.

evaluation set	GT	S0	S1	S2	M	L	pairs	videos
Charades-STA (native, scoreable)	native	0	457	2372	850	0	3679	1313
Charades-TimeLens (primary)	re-ann.	590	1474	756	512	31	3363	1313
VC-R1 length-balanced subsample	re-ann.	300	300	300	300	31	1231	614
ActivityNet (full native split)	native	662	1784	2424	5393	7242	17505	4917
stratified subset	native	100	100	100	100	100	500	192
QVHighlights-TimeLens	re-ann.	92	268	207	525	449	1541	1511

N Native Charades-STA and Scorer Validation

Reproduction numbers. TimeLens-7B through its official pipeline: R@0.3/0.5/0.7, mIoU $= 70.77/55.63/28.46/49.00$ vs. published $70.5/55.6/28.4/48.8$ (within 0.3 points on every metric).

TRACE on native Charades-STA reproduces its published R@0.5/0.7, mIoU of $40.3/19.4/38.7$ to within 0.4 points; VideoChat-R1 matches its published native R@1@IoU0.5 within 2.5 points, the gap traced to harness differences.

Native-split support. The native Charades-STA test split contains $3{,}679$ scoreable pairs (41 of the official $3{,}720$ are excluded because 21 video files are missing from the distribution), of which 457 / $2{,}372$ / 850 fall in $[2, 5)/[5, 10)/[10, 30)$ s and *zero* fall below 2 s or at or above 30 s. The TimeLens-Bench (Zhang et al., 2026b) re-annotation of the same video corpus contains $3{,}363$ pairs with 590 (17.5%) below 2 s and 31 at or above 30 s.

Run validation. TRACE evaluated on the native split through our pipeline reproduces its published metrics (R@0.5/R@0.7/mIoU $= 40.3/19.5/39.1$ vs. published $40.3/19.4/38.7$), VideoChat-R1's native run matches its published R@1@IoU0.5 within 2.5 points (0.692 vs. $\sim$0.717; the residual gap traces to harness and frame-sampling differences rather than parsing, since every output parses, Sec. Q), and our TimeLens reproduction matches its paper to within 0.3 points on every metric ($70.77/55.63/28.46/49.00$ vs. $70.5/55.6/28.4/48.8$). The stratified scorer reproduces the official evaluator's overall numbers on every prediction file before stratifying, which validates the per-stratum numbers as official-convention.

Annotation-style note. TimeLens-Bench was produced by the TimeLens authors, and TimeLens benefits most from its precise boundaries (its 10–30 s R@1@IoU0.7 rises $0.299 \rightarrow 0.574$ from native to re-annotated, vs. VideoChat-R1's $0.486 \rightarrow 0.510$), so *cross-model* per-stratum comparisons partly reflect annotation-style match; the within-model collapse and all interventional results are unaffected.

Why ActivityNet uses the native annotations. We evaluate Charades and QVHighlights through their TimeLens-Bench re-annotations but ActivityNet through its *original* annotations, an instrument choice made by one rule: use the finest-granularity ground truth available per dataset. The TimeLens re-annotation of ActivityNet is fully snapped to a 1 s grid: all $4{,}500$ of its spans have integer starts and integer lengths, its minimum span is exactly 1.00 s, and it contains no sub-second ground truth: its whole sub-2 s stratum is 564 spans of exactly 1 s. Native ActivityNet Captions carries float, sub-second spans. For a study whose measurand is a center error of 0.6–1 s, a ground-truth grid with ± 0.5 s quantization is an invalid instrument; the native annotations are the only valid one. (For Charades and QVHighlights the re-annotations are, conversely, the only instruments with sub-2 s support at all: the native splits contain zero such moments.) The instrument choice is not what creates the collapse: run on the 1 s-snapped ActivityNet re-annotation itself, the frozen model shows the same profile: sub-2 s mIoU 0.153 against 0.463 overall, strict recall 0.017, with the monotone length gradient intact ($4{,}140$ of $4{,}500$ pairs scoreable; attrition from missing videos is proportional across strata). The failure appears on every annotation layer we can obtain, chosen or not.

Release. The stratified scorer, its reference results for every model in Table 2, and a script that rebuilds the per-stratum splits from each dataset's own release are included in the supplementary material. No videos and no third-party annotations are redistributed. The rebuild is deterministic and pins the exact evaluation pairs, including the per-stratum ActivityNet subsample (100 per stratum) and the balanced VideoChat-R1 subsample (300 per stratum), so every table in the paper is recomputable from a model's raw predictions.

Native-split length profiles. Table A12 gives the full native-split profiles for all three models (n per stratum: $457/2372/850$). All three look healthy in aggregate, none shows the graded monotonic collapse visible under precise annotation, and each profile has a different shape: the coarse native boundaries scramble the length signal rather than merely hiding the tails.

Confidence intervals. All stratum CIs are 95% percentile intervals from a cluster bootstrap that resamples *videos* with replacement ($2{,}000$ replicates), respecting the within-video correlation of query outcomes. For the sub-2 s row of main-text Table 1, with exact Clopper–Pearson intervals for the zero-count cells: TimeLens $[2.5, 5.7]$, VideoChat-R1 $[0, 1.2]$, TRACE $[0, 0.5]$, zero-shot backbone $[0, 0.6]$, ActivityNet $[0, 3.3]$, QVHighlights $[2.2, 13.0]$.

Table A12: Native Charades-STA length profiles (n=3679 scoreable pairs; no sub-2 s or $\geq$30 s stratum exists). Flat or non-monotonic in every model, unlike the steep graded collapse under precise annotation.

	TimeLens		VideoChat-R1		TRACE	
L_{gt}	R@1@IoU0.5	R@1@IoU0.7	R@1@IoU0.5	R@1@IoU0.7	R@1@IoU0.5	R@1@IoU0.7
2–5 s	0.361	0.123	0.551	0.330	0.425	0.267
5–10 s	0.383	0.116	0.707	0.490	0.440	0.181
10–30 s	0.571	0.299	0.726	0.486	0.288	0.196
overall	0.423	0.159	0.692	0.469	0.403	0.195

O Zero-Shot Backbone Protocol

The zero-shot control (Secs. 4–5 of the main paper) runs the raw `Qwen2.5-VL-7B-Instruct` checkpoint, the backbone TimeLens finetunes, through the *identical* evaluation pipeline: the same interleaved per-frame-timestamp input construction, the same 2 fps / token budget, the same greedy decode, and the same official scoring. Only the weights differ between this run and TimeLens under the same construction, and the native-prompt control below shows the result is not specific to either prompt format. All 3,363 outputs parse (100%; the backbone grounds zero-shot without any format failure). Key numbers: sub-2 s emission $1/3{,}363$ predictions; sub-2 s R@1@IoU0.5 = R@1@IoU0.7 = 0.000 $(0/590)$; median $\log_2(L_{\text{pred}}/L_{\text{gt}})$ from $+2.35$ at $[0, 2)$ s to -0.18 at $[15, \infty)$ s with the zero-crossing near 12–15 s; length-compression slope $\beta = 0.54$; overall R@0.3/0.5/0.7/mIoU $= 0.596/0.347/0.180/0.399$; 10–30 s R@1@IoU0.7 $= 0.588$ (statistically flat against TimeLens's 0.574: paired by-video 95% CI $[-3.4, +6.1]$ points).

Native-prompt control. Re-running the zero-shot backbone with a plain grounding prompt (no interleaved per-frame timestamps; same 2 fps, greedy decode, 100% parse) preserves the prior while lowering overall quality: $[0, 2)$ median $\log_2$ ratio $+2.32$ (vs. $+2.35$ interleaved), $\beta = 0.517$ (vs. 0.54), sub-2 s R@1@IoU0.7 again $0/590$ (R@1@IoU0.5 0.008), sub-2 s emission 4.5% of predictions (6.8% on the ultra-short cases; ground truth 17.5%), overall R@0.5/0.7/mIoU $= 0.173/0.075/0.193$, median $|\Delta c| = 5.75$ s, zero-crossing near 5–7 s. The interleaved timestamps help grounding quality but do not create the length prior: it is format-invariant.

P Prompt for the Prompting Intervention

Verbatim prompts. Every decode uses the timestamp prefix followed by the grounding instruction; {} is the query slot.

Base prompt (all main results):

> *You are given a video with multiple frames. The numbers before each video frame indicate its sampling timestamp (in seconds). Please find the visual event described by the sentence '{}', determining its starting and ending times. The format should be: 'The event happens in <start time> - <end time> seconds'.*

Short-license prompt (the query-agnostic intervention of Sec. 4; differences from the base in the middle sentences):

> *You are given a video with multiple frames. The numbers before each video frame indicate its sampling timestamp (in seconds). Please find the visual event described by the sentence '{}', determining its precise starting and ending times. Events vary widely in length; some last only one or two seconds. Report tight boundaries that match the event's true duration — do not widen a brief event to a typical interval. The format should be: 'The event happens in <start time> - <end time> seconds'.*

Changing only this prompt shifts sub-2 s emission on the 590 ultra-short cases from 10.2% to 14.2% and lowers overall R@1@IoU0.7 from 28.5% to 27.4% (paired by-video 95% CI on the drop: $[-1.7, -0.3]$ points, McNemar-exact $p = 0.006$).

Per-query oracle length hint. The strongest conditioning short of the answer itself: for each query we append a hint naming the ground-truth length bin ("Note: the target event lasts under 2 seconds.", with the paper's strata as bins) to the otherwise unchanged grounding prompt, on the frozen model, full test set (3,363 pairs, identical decode protocol, official scoring). The hint is absorbed: sub-2 s emission on the ultra-short cases rises sharply ($10.2 \rightarrow 17.8\%$), β decompresses ($0.60 \rightarrow 0.69$), and long moments are rescued outright (30 s+ median $|\Delta c|$ $8.5 \rightarrow 1.25$ s, R@1@IoU0.5 $0.516 \rightarrow 0.774$; 10–30 s R@1@IoU0.7 $0.574 \rightarrow 0.617$). The sub-2 s stratum does not improve: median $|\Delta c|$ $1.00 \rightarrow 0.95$ s, R@1@IoU0.7 $0.041 \rightarrow 0.036$ (paired by-video 95% CI on the change $[-2.2, +1.1]$ points), and R@1@IoU0.5 drops $0.259 \rightarrow 0.110$ (CI $[-18.5, -11.6]$ points). The hinted spans are shorter but land on the same $\sim$1 s-off centers, losing the incidental overlap that over-long spans had. Where the failure is length mis-calibration (long moments), an oracle length fixes it; where it is localization (short moments), it does not, which is the dissociation that motivates Sec. 5 of the main paper.

Q Oracle-Window Floors, Center-Floor Generality, Parse Rates

Trivial floors under the oracle window. A fixed-fraction oracle window leaks geometry: with the moment center placed at a known fraction of the window, an agent that ignores the video entirely can still score. We therefore evaluate two trivial policies on the exact windows used in each run (predict the whole window, and predict a median-length span at the window center) and compare them with the model (Table A13). The floor we report is the better of the two *per metric*, so one floor row can mix policies. On Charades every sub-2 s moment is exactly 1 s on the integer grid, so the window policy gives IoU $= L/W = 1/8$, the mIoU floor 0.125; the $\widetilde{|\Delta c|}$ floor 0.80 s is the center policy, which sits $0.1W$ from the true center because the moment center is placed at fraction 0.6. Neither reaches IoU $= 0.5$: the window policy caps at $L/W = 0.25$, and two 1 s spans offset by 0.8 s overlap $0.2/1.8 = 0.11$, which is why both recalls are 0.000. A floor is informative only if it can be beaten, and two things break that: a window narrow enough that a blind guess clears IoU $= 0.5$ unaided, and a placement that puts the moment *at* the window center, which makes the center policy exactly right. Our W=2 and W=3 settings fail the first at *every* placement we checked: even at fraction 0.60 the center policy clears R@0.5 on all 590 pairs, because a 1 s span offset by only 0.2–0.3 s still holds IoU > 0.5. They additionally used fraction 0.50, which makes the center policy exact ($|\Delta c|$=0, strict recall above 0.89). At W=8 and W=12 with fraction 0.60 neither failure applies. At W=8 and W=12 (fraction 0.60) the model beats both floors on every metric, on both datasets. At W=4 the center policy attains higher mean overlap than the model (0.378 vs. 0.336) and a smaller median center error, yet scores 0.000 at both recall thresholds: a 1 s span at a fixed offset is consistently close and never correct, while the model clears IoU $= 0.5$ on a quarter of the stratum.

Table A13: Oracle-window trivial floors, computed per record from the stored crop windows (sub-2 s strata; Charades n=590, ActivityNet n=100). "window" = predict the entire window; "center" = predict a median-length span at the window center. **Scored unrounded**, so the model columns here are lower than the same runs in Table 5, which follows the integer-grid convention Charades-TimeLens calls for; centre errors are unrounded throughout and agree across both.

setting (place)	policy	R@1@IoU0.5	R@1@IoU0.7	mIoU	$\widetilde{\lvert\Delta c\rvert}$
Cha. W=2 (.50)	model / window / center	0.485 / 1.000 / 0.920	0.112 / 0.000 / 0.919	0.437 / 0.500 / 0.945	0.35 / 0.00 / 0.00
Cha. W=3 (.50)	model / window / center	0.329 / 0.000 / 0.900	0.063 / 0.000 / 0.893	0.372 / 0.333 / 0.902	0.45 / 0.00 / 0.00
Cha. W=4 (.60)	model / window / center	0.253 / 0.000 / 0.000	0.071 / 0.000 / 0.000	0.336 / 0.250 / 0.378	0.55 / 0.40 / 0.40
Cha. W=8 (.60)	model / window / center	0.263 / 0.000 / 0.000	0.081 / 0.000 / 0.000	0.305 / 0.125 / 0.105	0.65 / 0.80 / 0.80
Cha. W=12 (.60)	model / window / center	0.259 / 0.000 / 0.002	0.068 / 0.000 / 0.002	0.292 / 0.084 / 0.032	0.75 / 1.20 / 1.20
ANet W=8 (.60)	model / window / center	0.180 / 0.000 / 0.010	0.030 / 0.000 / 0.000	0.297 / 0.163 / 0.137	0.71 / 0.80 / 0.80

Center-floor generality. From the main-table prediction files, the sub-2 s median center error is 1.00 s for TimeLens, 1.50 s for VideoChat-R1, 2.00 s for TRACE, and 1.62 s for the zero-shot backbone, falling only to 1.25 s at 72B. Both zero-shot rows emit fractional timestamps, and read on a common 1 s grid each of them lands at 1.50 s, so the improvement at the larger scale does not survive matching the resolution: an absolute center error comparable to or exceeding the moment length is a property of every grounder we test rather than of one model. The *interventional* ladder (main-text Table 4) is measured on TimeLens only.

Parse rates. Every prediction file behind the main tables parses completely: 0 format failures out of 3,363 records each for TimeLens, TRACE, and the zero-shot backbone (both prompt formats), 1,231 for VideoChat-R1, 1,541 for QVHighlights, and 17,505 for ActivityNet. The zero cells in main-text Table 1 are genuine misses rather than harness artifacts.

R Robustness Checks: Held-Out Settings, Duration Scaling, Annotation Conventions, Equivalence

Held-out confirmation of the lever settings. The frame rate and window width were selected on the same set they are reported on, so we confirm them on data untouched by selection: a deterministic by-video two-fold split (hash parity), with fps $\in \{2, 8\}$ and $W \in \{4, 8, 12\}$ selected on one half by sub-2 s strict recall and scored on the other. Both folds independently select fps 8 and $W{=}8$. Held-out sub-2 s strict recall: denser sampling 0.160 and 0.144 against baselines of 0.035 and 0.047 (full-set values 0.153 and 0.041 lie between the folds); the oracle-crop median $|\Delta c|$ is 0.65 s in both folds, identical to the full set. The headline settings and numbers survive held-out evaluation.

Center error vs. input duration. A constant *relative* decoding precision would predict median $|\Delta c|$ proportional to input duration. Measured on the same sub-2 s clips (oracle windows of $2/3/4/8/12$ s and the $\sim$30 s full video), median $|\Delta c|$ is $0.35/0.45/0.55/0.65/0.75/1.05$ s: growth is strongly sublinear (near-logarithmic), and the proportional account over-predicts the full-video error five-fold. Relative precision *improves* as the input grows, consistent with a search-space account and inconsistent with a fixed relative-precision decode.

Envelope vs. core: conditioning the re-annotation on the native spans. For the 872 re-annotated Charades pairs whose query text exactly matches a native Charades-STA pair on the same video, 95.2% of sub-2 s re-annotated moments lie entirely inside their native counterpart's span, which is median $5.4\times$ longer with centers 0.8 s apart; nesting declines monotonically with re-annotated length (to 0% at $\geq$30 s). The native split's missing sub-2 s stratum is therefore an annotation-granularity convention, event envelopes against action cores of the *same* events, and not an absence of short events; and the re-annotation is anchored to the same events rather than introducing new targets. The capability the stratified protocol measures is finding the core inside the envelope.

Equivalence testing and multiplicity. The backbone-vs-finetune comparison at 10–30 s is an equivalence claim, so we test it as one: paired on 512 shared queries over 437 videos, the strict-recall difference (backbone $-$ finetune) is $+0.014$ with a 90% by-video cluster-bootstrap CI of $[-0.026, +0.053]$. The non-inferiority margin is five points, an order of magnitude below the $4.1 \rightarrow 57.4$ spread across strata that this paper is about, so a gap the test tolerates is negligible against the effect under study. At that margin the backbone is *non-inferior* (the direction the claim requires); two-sided equivalence narrowly fails on the upper side only because the backbone trends slightly better. The zero-count cells span many independent videos ($0/590$ over 412 videos; $0/300$ over 208), and the

Equivalence of the centre claims. The centre non-improvements are equivalence claims as well, and their margin follows from the metric rather than from a judgement. With ground truth of length L, a prediction centred d away can still reach IoU of t at some length only while $d \leq L(1-t)/(2t)$, so on a stratum whose targets are one second long, strict recall is unreachable beyond 0.214 s. Tested at that margin, the frozen model and the natural-short adapter have the same median centre error, and the paired interval is degenerate at zero. Degeneracy in a tied and quantized statistic is not evidence that two distributions agree, so the median is not where this question should be read. The distribution is indeed not unchanged. The share of sub-2 s clips centred inside that window, both arms on the shared 1 s grid, rises by 5.3 points on TimeLens-7B, with an interval of $[+2.1, +8.6]$, and by 14.9 points on TimeLens-8B, $[+11.4, +18.4]$. Scored unrounded instead, the same gains are 3.2 and 11.9 points, so the direction survives the convention this section is most exposed to and the effect is the smaller of the two. improvement comes from, and on TimeLens-7B the mean centre error moves the other way. What these tests support is that no arm reaches sub-second placement. They do not show that placement is unmoved.

conservative cluster-level exact upper bounds (0.9% and 1.8%) leave the collapse conclusion unchanged. Headline p-values are reported as $p < 10^{-10}$ rather than at implausible precision.

S Decoy Control: Truth-vs-Decoy Separation

What it measures. Sec. 4 separates two things a finetune can change: which lengths the model *emits*, and whether it can tell the right moment from a plausible wrong one. This control bounds the second. For each clip we score the ground-truth answer and a same-length decoy answer under teacher forcing and take the difference in nats,

$$\text{sep} \;=\; \sum \log P(\text{truth}) \;-\; \sum \log P(\text{decoy}),$$

summed over the *whole* answer sentence ("The event happens in $X.X$ - $Y.Y$ seconds."), not the numerals alone. The decoy has the same span length as the truth, so the length prior cannot prefer one over the other; digit counts are not matched, but the same pairs are scored for every arm, so that difference is common to all of them and cannot produce the collapse in separation we report. What differs between the two answers a model is asked to score is *where* they point.

Decoy construction. The decoy is deterministic rather than sampled, so the control is reproducible without a seed. For a video of duration D and a ground-truth span (g_0, g_1) of length L, the decoy is centred at the reflection of the true centre about the video midpoint, $c' = D - (g_0{+}g_1)/2$. Where that span would overlap the truth, which happens for centred moments, it falls back to the roomier end of the video, $(0, L)$ or $(D{-}L, D)$. The result is clamped to $[0, D{-}L]$ and rounded to 0.1 s. One decoy per clip.

Sample and arms. $n = 200$ paired sub-2 s clips from Charades-TimeLens. The five finetuned arms are the rebalanced arm, the natural-distribution control, speed-augmented, speed-augmented (gentle), and full-parameter natural-short, all evaluated as merged full-weight checkpoints.

Table A14: Truth-vs-decoy separation in nats, $n = 200$ paired sub-2 s Charades clips. Intervals are 95% percentile bootstrap on the mean ($4{,}000$ resamples, seed 0). Every finetuned arm's interval is disjoint from the frozen model's. *Truth first* is the percentage of clips on which the truth outscores the decoy, a rank statistic insensitive to changes in a model's overall confidence; a paired bootstrap on the difference from frozen ($4{,}000$ resamples) excludes zero for the rebalanced, speed-augmented and full-parameter arms and includes it for the other two.

arm	median	mean	95% CI (mean)	truth first
frozen	**15.62**	15.12	$[13.62, 16.64]$	**94.0**
rebalanced	4.55	4.08	$[3.51, 4.61]$	86.5
natural control	4.34	4.30	$[3.80, 4.79]$	92.5
speed-augmented	4.92	4.63	$[4.01, 5.26]$	84.0
speed-augmented (gentle)	5.42	5.63	$[4.98, 6.26]$	90.5
full-parameter natural-short	4.44	4.53	$[4.02, 5.04]$	89.0
five arms pooled	4.64	4.63	—	88.5

Reading. The frozen grounder prefers the truth over a same-length wrong answer by a median 15.6 nats. Every finetuned arm falls to roughly 4.6, and no arm's interval overlaps the frozen model's, so the collapse is not a sampling accident. Training relocates the model's probability mass without improving its ability to tell the right moment from a plausible wrong one, which is why emission moves and accuracy does not.

T Finetuning Recipes

The rebalanced and natural-control corpora are matched in volume and differ only in length distribution. The speed-augmented clips come from $2\times$ and $4\times$ physical re-encoding of Charades training clips; the merged corpus is 38% sub-2 s. GRPO uses group size $G{=}8$ and center-reward width $\sigma{=}1.0$. The full-parameter arms carry no adapter, unfreeze the language model while keeping the vision tower and merger frozen, and run ZeRO-3 with gradient checkpointing at a peak of 181 GB

Table A15: Every finetuning arm. Shared across all of them: TimeLens-7B, 200 steps, global batch 64, 1–2×B200, DeepSpeed ZeRO-2/3, sdpa or flash-attention-2. Adapter arms are LoRA r=64, α=64. The GRPO arms train on the standard span format rather than the (c, w) reparameterization, since the reward is sequence-level and reads the center as $(s+e)/2$.

arm	update	corpus	lr
rebalanced	LoRA	Charades-STA train, length-rebalanced	10^{-4}
natural control	LoRA	Charades-STA train, natural lengths	10^{-4}
speed-augmented	LoRA	+4,060 genuine sub-2 s clips	10^{-4}
natural-short (arm C)	LoRA	ActivityNet Captions short tail	10^{-4}
center-first reparam.	LoRA	Charades-STA train	10^{-4}
EC-Sharp	LoRA	Charades-STA train	10^{-4}
GRPO ×3	LoRA	Charades-STA train	10^{-6}, 10^{-5}, 10^{-4}
full-param natural-short	LM only	ActivityNet short tail	10^{-5}
full-param speed-aug.	LM only	speed-augmented	10^{-5}
full-param merged	LM only	natural + speed-aug., 14,513 recs	10^{-5}

on one B200; their learning rate is lower because the adapter arms' 10^{-4} destabilizes full-parameter training.

The arms trained, and the largest rate is an optimization limit. GRPO diagnostics confirm the runs moved: KL to the frozen policy grew to 0.05–0.09, group-relative advantages stayed non-degenerate (reward standard deviation 0.20–0.22), and at 10^{-4} the training reward rose from 0.44 to 0.50 across 150 steps. That arm is the informative one, since its reward improved throughout while its greedy decode *worsened* on every reported column, so the objective reaches the deployable decode and moves it the wrong way. Beyond 10^{-4} the recipe collapses rather than degrading further: at 3×10^{-4} the group-relative reward variance reaches exactly zero by step 5, identical rollouts carry identical reward, the advantage vanishes, and the objective supplies no gradient for the remaining 145 steps. Completion length falls from 16.5 tokens to 5.1 by step 20 and the KL rises to 18.4. We therefore treat 10^{-4} as the largest usable rate: the ladder is bounded above by optimization failure rather than by an unexplored hyperparameter.

Replication and the full-parameter verdict. The rebalanced arm is replicated across three seeds, giving $\beta = 0.195/0.194/0.223$, sub-2 s emission 0.0% in every run and overall mIoU 0.373–0.382, so the backfire is systematic rather than run-to-run variance. The full-parameter arms return the adapters' verdict: sub-2 s median $|\Delta c|$ of 1.10/1.35/1.15 s against the frozen 1.05 s, emission repaired only by the natural-short corpus (45.8%), and overall mIoU at 0.374–0.393 against the frozen 0.476.

The natural-short arm on a second grounder. The natural-short ActivityNet arm was repeated on TimeLens-8B with its TimeLens-7B counterpart's corpus, adapter rank, step budget and learning rate. Sub-2 s emission rises from 1.2 to 54.9%, close to the 10.2 to 54.1% the same arm produces on the smaller backbone, and sub-2 s R@1@IoU0.7 rises 0.008 → 0.239. The length prior compresses to $\beta = 0.295$ from 0.726, against that arm's 0.301. The costs reproduce as well: overall strict recall falls from 0.357 to 0.252, and mIoU from 0.553 to 0.425. The center does not reach sub-second placement, and it is not unmoved either. Read on the 1 s grid, which puts the arm's fractional timestamps and the frozen model's whole seconds at one resolution, the two share a median of 1.00 s and the paired interval is degenerate at zero, so the median is not where this question can be read; the arm's unrounded 0.70 s is a resolution artifact rather than a shift. The mean absolute center error moves the wrong way, 2.41 → 2.46 s. What does improve is the window strict recall actually depends on: the share of sub-2 s clips centred within 0.214 s of the truth rises by 14.9 points, [+11.4, +18.4] (Appendix R), while across their first eight deciles the two distributions differ by at most 0.35 s. We do not read placement from a 0.5 s threshold, whose sign depends on the convention: that share falls from 47.1 to 42.4% at each model's own resolution and rises to 49.5% on the grid.

The arm across training paradigms. The natural-short arm was repeated on VideoChat-R1, an RL-tuned grounder on the same backbone family, with arm C's corpus, adapter rank, step budget and learning rate. Emission is repaired from a standing start: the frozen model emits no sub-2 s

span at all, the arm emits 41.2%, and sub-2 s R@1@IoU0.7 rises $0.000 \rightarrow 0.115$ on a stratum the frozen model fails completely. The costs reproduce, with overall mIoU falling $0.385 \rightarrow 0.330$ and β falling $0.279 \rightarrow 0.179$. Placement improves on this grounder and still does not reach the second: the median sub-2 s $|\Delta c|$ moves $1.95 \rightarrow 1.45$ s with a paired cluster-bootstrap interval of $[-0.60, -0.30]$ s over 412 videos, and $P(|\Delta c| \leq 0.5)$ rises from 14.6 to 24.6%. The mean worsens $3.10 \rightarrow 3.67$ s once predicted centers are clamped to the clip, against $3.58 \rightarrow 7.39$ s unclamped, where a handful of predictions place the moment hundreds of seconds past the end of a 30 s video; the same clamp leaves every TimeLens figure in this appendix unchanged, neither backbone emitting a prediction that leaves the clip. A grounder that begins by localizing nothing in the stratum, gains 41 points of emission and moves its center by half a second therefore stops at the barrier every other update stops at.

U Input-Side Interventions: Decomposing the Crop, and Foveation

The crop conflates several changes. It shortens the clip, removes out-of-window content, narrows the positional range and, at a fixed token budget, raises the tokens each remaining frame receives. Foveation separates them. For each clip we take the frame count and per-frame resolution that the uniform 2 fps sampler uses on that clip, so the token budget matches exactly, and we then choose *when* to sample: frame times follow a density kernel centered on the moment, with support over the entire video, and the frame metadata makes the model emit truthful absolute timestamps. Nothing is removed and no frame is cheaper or dearer than in the uniform arm.

Table A16: Input-side interventions on the frozen model, all on the 590 sub-2 s Charades-TimeLens clips. Every row in the lower two blocks inherits the uniform arm's per-clip frame count and per-frame resolution, so the token budget is identical and each kernel has support over the whole video. The crop block uses the 14,336-token default except the matched-token row, which uses 16,235 so tokens per frame equal the control's. Window and kernel placement are oracle. Emission and recalls follow the rounded 1 s grid; center errors are unrounded. All columns are restricted to the sub-2 s stratum, so these mIoU values are sub-2 s and not comparable with the corpus-wide figures elsewhere. †On 542 of 590 clips: for the other 48 the window already spans essentially the whole video, so widening the positional range is undefined and the run refuses it rather than reporting a no-op.

	emission	R@1@IoU0.5	R@1@IoU0.7	mIoU	$\widetilde{\lvert\Delta c\rvert}$
decomposing the crop					
uniform 2 fps (control)	9.5%	0.259	0.041	0.251	1.05
temporal crop, $W{=}8$	54.1%	0.383	0.193	0.328	**0.65**
matched tokens per frame	53.7%	0.408	0.210	0.350	**0.65**
out-of-window frames zeroed	13.1%	0.341	0.047	0.296	0.85
frames permuted	69.0%	0.053	0.025	0.054	2.85
positions widened†	7.0%	0.035	0.006	0.056	5.25
redistributing the same budget, full video support					
uniform (control)	9.2%	0.263	0.031	0.245	1.05
Gaussian $\sigma{=}8$	26.1%	0.327	0.095	0.280	0.85
Gaussian $\sigma{=}4$	43.1%	0.361	0.175	0.323	0.70
Gaussian $\sigma{=}2$	**52.7%**	**0.393**	**0.205**	**0.346**	**0.60**
two-level 80/20	39.0%	0.378	0.154	0.321	0.70
aiming the kernel off the moment					
$\sigma{=}4$, displaced 2 s	40.3%	0.369	0.166	0.325	0.70
$\sigma{=}2$, displaced 2 s	48.6%	0.371	0.180	0.327	0.65
$\sigma{=}2$, displaced 8 s	35.9%	0.042	0.017	0.054	5.60

Three rows carry the identification. Permuting the cropped frames leaves their content, resolution, count and positional range untouched and sends the center error past the control, so the gain is not a prior the window supplies. Matching tokens per frame changes nothing, and the premise behind that control does not hold on these videos in the first place: an 8 s window at 8 fps holds 64 frames against the control's 57, so at a fixed budget the crop has slightly *fewer* tokens per frame rather than

more. And aiming a σ=2 kernel 8 s away from the moment is worse than sampling uniformly, which is what rules out non-uniformity by itself as the explanation.

Widening the positional range while keeping the cropped pixels also collapses, which says that positions must stay consistent with content. It does not show that a narrower positional range is the mechanism, since the manipulation misstates when every frame occurred.

The center error is monotone in the kernel width. Since the frame count, the per-frame resolution, the token budget and the positional range are all held fixed, and no content is withheld, the ordering cannot be attributed to clip length, resolution, token allocation or distractor removal. What remains is where the samples fall.

Two limits are worth stating. The kernel is placed using the true center, so this identifies the mechanism in the same sense the oracle crop does and is not a deployable method. And at a fixed budget, concentrating samples on the moment necessarily thins them elsewhere, so this design cannot separate "more samples inside the moment" from "fewer samples outside it"; the two move together by construction.

The last row displaces the kernel 2 s off the true center. It matches the on-center arm of the same width, which establishes that the model is not merely echoing the kernel's position back as its answer. It does not establish that placement matters, because at σ=4 a 2 s displacement is half a standard deviation and the kernel still covers the moment.

V Center-Intervention Implementations

EC-Sharp: the jointly trained expected-center head. The head reads one layer's hidden states (default: the last; d=3584). The per-frame visual states V ($T\times d$) are the means of each frame's patch-token states, and the query vector h_q is the hidden state at the last prompt token, the position whose output predicts the first answer token, so training and inference read the same position. Scoring is one scaled dot product: $q = W_q\,\mathrm{LN}(h_q)$, $k_i = W_k\,\mathrm{LN}(V_i)$, and $p = \mathrm{softmax}(kq^\top/\sqrt{d})$ over the T frames (in fp32). The center is the expectation $\hat{c} = \sum_i p_i t_i$ over normalized frame times $t_i = (i{+}0.5)/T$, the spread is $\hat{\sigma} = (\sum_i p_i(t_i - \hat{c})^2)^{1/2}$, and a width branch predicts $\hat{w} = \sigma(W_w\,\mathrm{LN}([h_q; \bar{V}]))$ from the query and the mean frame state. All three are scaled by the clip duration to seconds and trained with

$$\mathcal{L} = \mathrm{CE}_{\mathrm{text}} + 1.0\,\mathrm{Hub}_{1.0}(\hat{c},\, c^*) + 0.5\,\mathrm{Hub}_{0.5}(\log\hat{w},\, \log w^*) + 0.5\,\mathrm{Hub}_{0.5}(\hat{\sigma},\, w^*/2),$$

where Hub_δ is the Huber loss with threshold δ (seconds). The third term is the anti-collapse constraint: a soft-argmax head can otherwise spread p uniformly, whose expectation lands mid-video regardless of the query, so the spread is tied to the ground-truth half-width. The span cross-entropy is kept so the backbone's grounding stays intact while the auxiliary terms reshape the per-frame representation. The head trains jointly with the same LoRA recipe as the other Table 4 arms, so gradients flow into the backbone; two readouts of the resulting adapter are evaluated, the head's $\hat{c}$ ($|\Delta c| = 1.74$ s) and the co-trained text decode (1.50 s). During training the head fits the centers of training clips to 0.02–1.0 s; the failure is out-of-sample. A frozen twin of the identical architecture (backbone untouched; layers 14/21/28 swept) stops at $|\Delta c| \approx 2.3$ s, which isolates joint backbone training as the variable this comparison tests.

Frozen readout heads. Each variant trains a small estimator on cached last-layer hidden states of the frozen grounder (fp16) and regresses the moment with L1 or gIoU losses; the model's weights never change. The deployed versions read the *post-answer* state: the model first greedily decodes its answer, and the estimator reads the hidden state that has, by then, computed the decided moment. Three architectures: an MLP on the single post-answer token state ($|\Delta c| = 1.82$ s; 1.87 s when speed-augmented sub-2 s clips are added with 5× up-weighting, so short training data changed nothing); the post-answer state as the query of one cross-attention layer over the clip's visual-token states average-pooled to 256 vectors (4.88 s; the added capacity hurt); and an MLP on the mean of all visual-token states. An earlier variant read the *pre-generation* state, the last input token before any decoding, and failed hardest (3.0 s): that state has not yet computed the moment, which the autoregressive decode constructs token by token. Every variant stays at $|\Delta c| \geq 1.8$ s against the frozen model's own text decode at 1.05 s, the basis of the ladder's censored row.

W ZOOM IMPLEMENTATION DETAILS

Windowed decode. The crop is injected through the `video_start`/`video_end` fields of the Qwen video-loading utility (`qwen-vl-utils`), so frames are sampled only inside $[\hat{c}_0 - W/2,\ \hat{c}_0 + W/2]$ while the timestamps the model reads and emits remain absolute original-video seconds; no offset bookkeeping is applied, and a manual audit found $< 5/100$ outputs that looked window-relative.

Window-width ablation (Charades ceiling, sub-2 s, n=590). W=4: $\widetilde{|\Delta c|} = 0.55$ s, R@1@IoU0.5 0.253, R@1@IoU0.7 0.071; W=8: 0.65 s, 0.263, **0.081**; W=12: 0.75 s, 0.259, 0.068. Robust across $W \in [4, 12]$ s; W=8 maximizes strict recall, and settings tighter than W=4 are geometry-dominated (Sec. Q). Deployed two-pass on Charades (window seeded by the model's own coarse pass): R@1@IoU0.5 0.217, R@1@IoU0.7 0.059, $|\Delta c| = 0.75$ s: the center tightens while short-video recall stays flat-to-negative, consistent with the regime analysis of Sec. 6 of the main paper.

Long-video containment. On ActivityNet, widening the realistic window (W=8/16/24) raises GT containment (60/74/79%) but not recall: the moment fills too little of a wide crop. Iterative wide-to-narrow zoom improves the median center error (1.46 s) but not recall; the residual frontier is a stronger coarse/proposal stage.