# Temporal Resolution Is Duration-Dependent in Generative Video Grounding

**Anonymous authors**
Paper under double-blind review

## Abstract

Temporal resolution is usually treated as uniformly beneficial in video temporal grounding. We show instead that, for a strong generative grounder, its marginal value is signed and duration-dependent. After averaging over sampling phase, increasing density from 4 to 8 fps improves one-second events, degrades medium-duration events, and crosses zero at 2.3 s, while the aggregate effect is null. Controlled interventions narrow the source: the short-event gain does not require new visual content, since duplicating byte-identical frames reproduces and exceeds the gain from real densification, while granularity-matched permutation of the model's interleaved time markers largely attenuates it, so readable temporal addressing drives the effect. Matched fine-tunes with different duration compositions also shift the density-response crossing, consistent with a learned duration-sensitive prior. We therefore introduce Zoom-VTG, an inference-time, duration-gated coarse-to-fine policy that applies a denser second pass only when the coarse prediction is short. On held-out Charades TEST-400, it improves one-second mIoU by $+0.172$, and the gain transfers without retuning to ActivityNet and QVHighlights. On the complete, descriptively-accounted 3,363-query population, one-second mIoU improves by $+0.137$ while aggregate mIoU changes by only $+0.001$. These results indicate that temporal resolution should be allocated and evaluated conditionally on event duration, rather than treated as a single global hyperparameter.

## 1 Introduction

Video temporal grounding — localizing the span of a video that matches a natural-language query (Gao et al., 2017; Krishna et al., 2017; Lei et al., 2021) — is typically evaluated with a single aggregate score, while temporal resolution is implicitly treated as a uniformly beneficial resource. We find that these two conventions hide a surprising interaction: increasing temporal density can improve short-event localization while degrading medium-duration events in the same frozen generative grounder (Zhu et al., 2026). The reason is that temporal resolution changes not only how much visual evidence the model receives but how finely it is temporally *addressed* — and the response tracks the latter, not the former.

The underlying scale problem is substantial: one-second events reach only $0.34$ mIoU against $0.76$ for events longer than ten seconds (Appendix Table 14), and are $17.5\%$ of Charades queries under the tight TimeLens-Bench convention (Appendix A.5.1). Part of this deficit is a measurement effect: fixed absolute boundary error is penalized more heavily by IoU at short durations (Section 4.1). But the more consequential modeling result is that the natural intervention — providing more temporal detail — does not help uniformly. Doubling the sampling rate from 4 to 8 fps improves one-second events while degrading medium-duration ones, in the same model at the same time. Averaged over sampling phase, the marginal value of density crosses zero at 2.3 s (Figure 1): positive below the crossing, negative over roughly 4–9 s, and undetectable beyond $\sim$13 s; the two sides cancel so completely that the aggregate effect is null ($-0.004$, $p$=$0.67$). Temporal resolution is not a monotonic resource, and a single score cannot say so.

We next ask what this signed response depends on. Additional visual evidence is not necessary: duplicating byte-identical frames reproduces and even exceeds the phase-averaged short-event gain from real densification, while disrupting frame-to-marker correspondence largely removes that benefit (Section 4.2), showing that readable temporal addressing, not raw visual content, modulates the



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.

effect. The response is also training-modifiable: matched fine-tunes differing in duration composition attenuate the short-event density response and shift its zero-crossing earlier in all three seeds (Section 5.1). Together, these interventions support a learned duration-sensitive response rather than a simple "more frames is better" account.

These pieces form one causal chain: density changes the readable time markers, which produces a *coordinated* shift in predicted center and extent — either alone is worse than nothing, both together recovers the gain (Section 4.2) — and only that coordinated shift is duration-dependent in sign. Zoom-VTG is this chain's consequence, not an independent method.

This diagnosis suggests an inference rule: spend density only where its measured marginal value is positive. We introduce Zoom-VTG, a duration-gated coarse-to-fine policy that runs a frozen grounder once at low density and invokes a higher-density second pass only when the first pass's own predicted span is short (Section 5.2). The temporal zoom operation itself builds on established coarse-to-fine re-decoding (Nie et al., 2024); our contribution is to derive its routing criterion from the measured density–duration interaction. On held-out TEST-400, Zoom-VTG improves one-second mIoU by $+0.172$; on the complete Charades population (descriptive accounting, includes dev-300), it reallocates accuracy toward short events ($+0.137$) while aggregate mIoU moves only $+0.001$ — illustrating why a single benchmark average can miss large changes in temporal behavior.

**Contributions.** Each claim below names its isolating intervention and endpoint. Three endpoints were pre-specified before the runs that test them (Section 3).

1. **Temporal density has a signed, duration-dependent effect.** *Intervention:* 4→8 fps and its reverse, phase-averaged over a twelve-arm sweep. *Endpoint:* short-minus-long difference-in-differences (DiD) $+0.077$ ($p$=0.0037), zero-crossing near 2.3 s; the aggregate effect is null ($-0.004$) (Section 4.1).
2. **The short-event benefit does not require new visual evidence and is strongly modulated by temporal addressing.** *Intervention:* frame duplication (new temporal positions, byte-identical pixels), crossed with timestamp semantics. *Endpoint:* $\leq$1 s mIoU increases by $+0.098$ under duplication despite no new visual content, and that gain falls to $+0.011$ (ns) after matched time-marker permutation — both default-phase, using the granularity-matched pair; phase-averaged duplication is $+0.108$ (Section 4.2).
3. **The density response is training-modifiable.** *Intervention:* matched fine-tunes differing in duration composition. *Endpoint:* arm$\times$density on $\leq$1 s falls by $0.051$, with the crossing shifting $1.6$ s earlier in all three seeds, supporting a learned duration-prior account (Section 5.1).
4. **A duration gate converts the measured interaction into a useful allocation policy.** *Intervention:* inference-time **Zoom-VTG** against rate-matched random and self-uncertainty routing, which it outperforms at the same pass-2 firing rate. *Endpoint:* $\leq$1 s mIoU $+0.172$ on held-out TEST-400 ($+0.137$ on all 3,363 queries, aggregate $+0.001$) — illustrating why duration-stratified reporting is necessary (Sections 5–6).

## 2 Related work

**Video temporal grounding** has progressed from proposal-based and span-regression models (Gao et al., 2017; Lin et al., 2023; Moon et al., 2023) to generative grounders that emit timestamps as text (Huang et al., 2023; Ren et al., 2024; Guo et al., 2025; Li et al., 2025; Zhu et al., 2026), which is the regime we study. **Duration and short-moment bias** have been studied through dataset/evaluation analysis, short-moment augmentation, and duration-aware representations (Otani et al., 2020; Park et al., 2026; Zeng et al., 2026). We instead study duration as a conditioning variable for the *response to temporal density*, including when additional density becomes harmful. **Adaptive temporal computation** uses coarse-to-fine inference, frame/token selection, or temporal refinement to allocate computation non-uniformly (Nie et al., 2024; Fan et al., 2026; Zhang et al., 2025b; Wang et al., 2026; Chen et al., 2026; Shen et al., 2025; Wang et al., 2025), motivating non-uniform allocation but with different routing signals, training paradigms, or task formulations. We ask a different question of these established tools (Nie et al., 2024; Chen et al., 2026; Fan et al., 2026; Zhang et al., 2026): *when* added density helps versus hurts as a function of event duration, deriving the routing signal directly from that interaction while controlling sampling phase and separating token multiplicity from

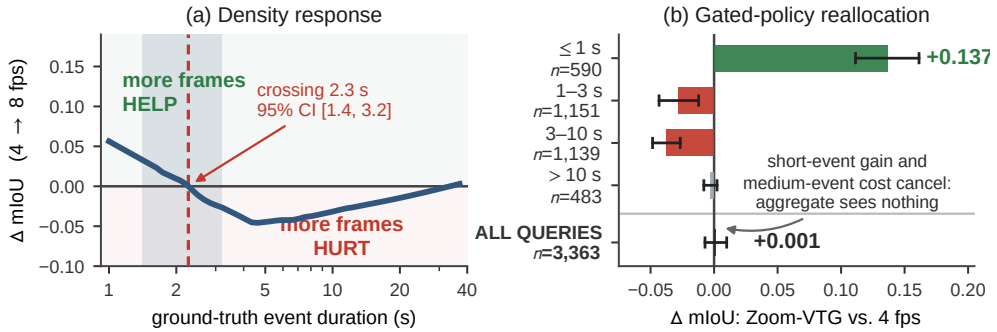



Figure 1: **Temporal density has a signed value that depends on event duration. (a)** Its marginal value against ground-truth duration (phase-averaged), crossing zero at 2.3 s (95% CI $[1.4, 3.2]$) — help below, hurt over 4–9 s, and no detectable effect beyond $\sim$13 s. **(b)** Gated Zoom-VTG on the complete 3,363-query population (accounting; includes dev-300): one-second events gain $+0.137$, medium bins lose, aggregate moves $+0.001$ ($p$=0.75) — invisible to the scalar the field reports.

timestamp semantics. Our comparisons hold the pass-2 firing rate fixed and vary the routing signal (Section 6; Appendix A.4.1).

## 3 Setup

We study TimeLens2-4B (Zhu et al., 2026), a Qwen3-VL-based (Bai et al., 2025) generative grounder fine-tuned for temporal grounding, on TimeLens-Bench (Zhang et al., 2025a) re-annotations of Charades (Gao et al., 2017), ActivityNet (Krishna et al., 2017), and QVHighlights (Lei et al., 2021). It is a strong recent generative grounder reported to match or beat much larger open models (Qwen3-VL 235B-A22B) and proprietary systems, including GPT-5 and Gemini 2.5 Pro (Zhu et al., 2026; OpenAI, 2026; Comanici et al., 2025) (full seven-benchmark comparison in Appendix A.5.8, Table 29).

The TimeLens-Bench annotations are deliberately tight: 17.5% of Charades events are $\leq$1 s, compared with 0.0% under the original Charades-STA convention. For shared queries, 95% of one-second intervals are nested inside the corresponding Charades-STA interval, so the short-event population we study exists only under the tighter convention (Appendix A.5.4).

Unless stated otherwise, results use a frozen 400-query Charades test split, TEST-400, containing *100 queries in each duration bin* and disjoint from the dev set used to select $p, T$. This bin-balanced design is appropriate for estimating a duration-dependent interaction, but its aggregate behaves like the duration-balanced summary in Section 6 rather than the natural Charades mixture. Decoding is greedy, making every inference arm deterministic and paired. Confidence intervals use 20,000-resample bootstraps clustered by video because multiple queries can share a video. Within each comparison, we apply Holm correction (Holm, 1979) across that comparison's family of duration-bin tests (e.g. the four bins for one density intervention), rather than across unrelated comparisons.

Three endpoints were pre-specified as *primary* before the runs that test them, matching claims 1, 2, and 4: (i) the short-minus-long difference-in-differences (DiD), defined as the $\leq$1 s 4$\rightarrow$8 fps $\Delta$mIoU minus the $>$10 s $\Delta$mIoU; (ii) the $\leq$1 s delta under frame duplication, which adds no new visual evidence; and (iii) $\leq$1 s mIoU for the gated policy versus uniform sampling on held-out TEST-400, measuring the practical consequence. Claim 3 and the remaining analyses were not pre-specified as primary endpoints, and are reported as supporting evidence for these same claims rather than as independent tests. The full claim-by-claim hierarchy and protocol are given in Appendix A.5.7, Table 28, and Appendix A.3.1.

## 4 The density–duration effect

### 4.1 The effect is signed and duration-dependent

Increasing temporal density has a signed effect whose direction depends on event duration. Raising the sampling rate from 4 to 8 fps improves one-second events while degrading medium-duration

events; lowering it reverses the pattern. After averaging over sampling phase, the continuous response crosses zero at 2.3 s (95% CI $[1.4, 3.2]$; Figure 1, left), remaining positive past the minimum one-second span ($+0.059$ at 2.1 s).

**Why short-event mIoU is intrinsically sensitive to boundary error.** Before explaining the signed density response, we separate a simpler scale effect. For the illustrative translation case with total absolute boundary error $\varepsilon$, IoU $\approx (L - \varepsilon/2)/(L + \varepsilon/2)$ grows with event length $L$, so an even roughly duration-invariant absolute error is penalized more heavily at short durations; a sampling grid of width $1/f$ bites hardest when $L$ approaches that scale. Median boundary error is indeed flat across bins even as mIoU rises from 0.34 to 0.76 (Appendix A.1.1). This explains part of the short-event accuracy gap, but not why density *harms* medium-duration events.

**Sampling phase substantially biases the naive estimate, but not the signed interaction.** The conventional zero-offset comparison overstates the short-event benefit because the default 4 fps grid is the *worst* of eight tested phases on one-second events (0.3622 versus a phase mean of 0.3946). We therefore average over eight 4 fps and four 8 fps phases (Table 1). This correction reduces the $\leq$1 s effect from $+0.0947$ to $+0.0563$ and eliminates the apparent aggregate gain ($+0.0186 \rightarrow -0.0038$, $p$=0.67), but strengthens rather than removes the central conclusion: the short-minus-long DiD remains $+0.0772$ ($p$=0.0037), positive at short and negative at medium durations. We therefore use phase-averaged values as the canonical density result throughout.

Table 1: **Phase averaging removes the apparent aggregate benefit of density, but not the duration interaction.** The 4→8 fps effect measured naively (default phase in both arms) and after averaging eight 4 fps and four 8 fps phases on TEST-400; 20,000-resample video-clustered bootstraps. Positive $\Delta$ favors 8 fps; corrected effects significant at $p < 0.05$ are bold. Short–long DiD is the $\leq$1 s effect minus the >10 s effect.

| Bin | Naive $\Delta$ | Corrected $\Delta$ | 95% CI | $p$ |
|---|---|---|---|---|
| $\leq$1 s | $+0.0947$ | $+$**0.0563** | $[+0.008, +0.102]$ | 0.026 |
| 1–3 s | $+0.0351$ | $+0.0002$ | $[-0.037, +0.038]$ | 0.98 |
| 3–10 s | $-0.0398$ | $-$**0.0507** | $[-0.079, -0.025]$ | $< 10^{-4}$ |
| >10 s | $-0.0158$ | $-$**0.0210** | $[-0.039, -0.005]$ | 0.008 |
| **short − long DiD** | $+0.1105$ | $+$**0.0772** | $[+0.026, +0.127]$ | 0.0037 |
| ALL (aggregate) | $+0.0186$ | $-0.0038$ | $[-0.022, +0.014]$ | 0.67 |

**Where the interaction replicates, and where it does not.** The interaction is not confined to Charades: ActivityNet reproduces the signed short-versus-long response (DiD $+0.1098$, $p$=0.0009) under its original annotations too, QVHighlights is inconclusive because its one-second population is small, and it disappears under the looser original Charades-STA labels (Appendix A.5.4). Every comparison holds its labels fixed, so the interaction transfers across datasets, but its magnitude and detectability depend on annotation convention.

### 4.2 What temporal information drives the response?

**New visual content is not required.** Density could act by supplying new pixels or new temporal addresses. Duplicating every frame doubles the visual-token count over byte-identical pixels, yet produces a $+0.1084$ phase-averaged $\leq$1 s gain — larger than the $+0.0563$ gain from real 4→8 fps densification ($+0.0980$ vs. $+0.0947$ at default phase; Appendix A.1.1) — so the benefit cannot be explained simply by observing additional visual content. The gain survives not just without new pixels but with *less* visual information: at `temporal_patch_size`=2 native tokens merge two distinct frames, duplicated tokens merge one frame with itself and carry no intra-pair motion. Duplication does change both token multiplicity and the number of time-marker groups, however, so we intervene directly on the second channel next (Figure 2).

**A factorial over the two temporal channels.** Qwen3-VL writes an interleaved textual marker (`<X.X seconds>`) before each temporal group. We cross token multiplicity (1× vs. duplicated) with timestamp semantics (correct vs. permuted: the same multiset of times, but with the frame-to-time mapping destroyed). Two pre-specified gates confirm both manipulations reach the model, and the first is informative on its own: scaling every marker by 10× stretches predicted centers by 10.21× (87% of queries within 5–15×), so the model uses the markers as a quantitative time coordinate rather than ordinal cues; destroying chronological frame order costs $-0.367$ (Appendix A.1.3).

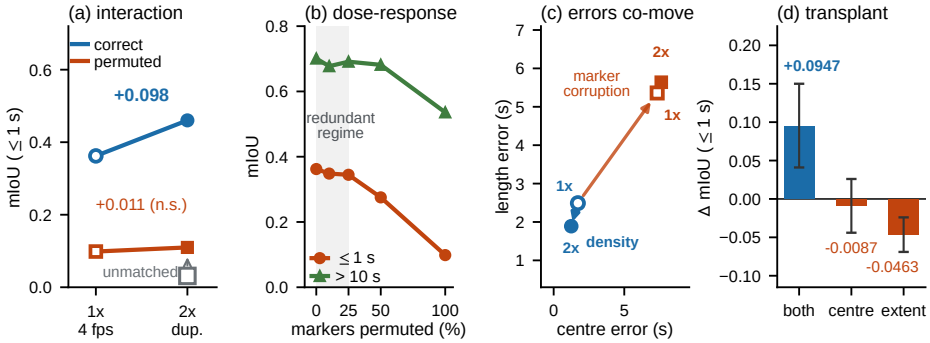



Figure 2: **The short-event density benefit needs no new pixels and is strongly modulated by readable temporal addressing** (Charades TEST-400, $n$=400). Open markers are $1\times$ tokens, filled the duplicated $2\times$ stream; blue is correct time markers, rust permuted. **(a)** Duplication gains $+0.098$ with correct markers, $+0.011$ (ns) after matched permutation. **(b)** Marker corruption hurts short events most. **(c)** Placement and extent errors co-move. **(d)** In the default-phase diagnostic, only the joint center+extent transplant recovers the $+0.0947$ gain.

**Readable time markers exert the larger behavioral influence.** Corrupting the markers costs $-0.2637$ at $1\times$ tokens and $-0.4289$ at $2\times$ (Holm $p$=0.0006), three to four times the gain from the corresponding density manipulation, and duplication's $+0.0980$ gain falls to $+0.0113$ (ns) under granularity-matched permutation. Without readable time the model falls back toward much longer spans ($+2.9$ s, $+3.4$ s). Additional temporal positions produce a detectable short-event benefit when paired with readable temporal addresses, but not under the matched marker permutation (Appendix A.1.3).

**The channel is offset-tolerant but sensitive to temporal order and precision.** Corrupting one property at a time separates three roles: a uniform half-step offset is free ($-0.008$, $p$=0.45), 0.5 s quantization costs short events alone ($-0.055$, $p$=0.0003), while permuting order does the real damage, scaling with dose ($-0.014/-0.087/-0.264$ at 10/50/100%). It is redundant rather than brittle: corrupting a quarter of the markers costs 0.020 mIoU (Appendix A.5.9).

**The response is more consistent with sub-interval selection than action retrieval.** Density could help the model *find* the action, or refine the sub-interval within an action already found. A provenance mismatch separates these: among 873 queries with identical video and text, the tight TimeLens span nests inside the looser Charades-STA span in 95% of one-second cases, so the loose span provides a proxy for action retrieval while the tight span probes sub-interval selection. Doubling density leaves the retrieval proxy unchanged in every bin (largest shift $+0.037$, all ns) while tight-span IoU tracks the duration interaction ($+0.066$ at $\leq$1 s, $-0.034$ at $>$10 s, both significant): the response is more consistent with sub-interval selection than with action retrieval (Table 4; alternative definitions in Appendix A.1.2).

**Density also changes the geometry of the predicted interval.** At default phase, increasing density contracts predictions by roughly 0.4–0.7 s through the first three duration bins, primarily by moving the predicted start later (Table 2). This helps short events, whose 4 fps predictions are substantially over-long, but overshoots in the 3–10 s regime, where baseline lengths are already close to correct. This contraction explains the direction of the default-phase short-event gain and medium-event loss, but not the complete phase-averaged response (Appendix A.1.2).

Nor does density act through span length alone. Center and extent change jointly: marker corruption damages both ($+5.615$ s, $+2.870$ s), and transplanting only the 8 fps center or only the 8 fps extent onto the 4 fps prediction makes $\leq$1 s performance *worse* ($-0.0087$, $-0.0463$), whereas moving both together produces the full $+0.0947$ default-phase gain (Figure 2c–d). We therefore interpret density as producing a coordinated re-localization rather than an independent correction to either placement or extent (Appendix A.1.2).

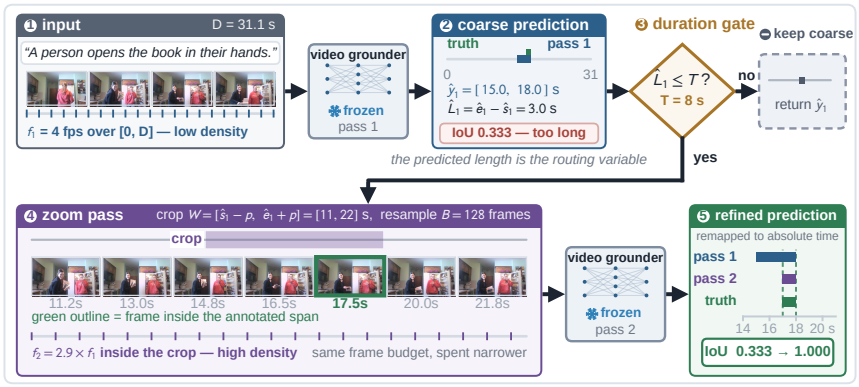



Figure 3: **Zoom-VTG run end-to-end on one Charades query** (URK3G, a $\leq$1 s event). One frozen model, one prompt, run twice: a coarse pass, then a high-density pass inside a padded crop, but only when the predicted length is at or below $T$ (72% of TEST-400 queries fire; the rest return the coarse span after a single pass). All frames and numbers are read off the pipeline. The $+0.667$ is this query's gain; Appendix Table 14 gives the population effect, including the medium-bin costs not shown here.

## 5 From diagnosis to duration-conditioned inference

The preceding section establishes a signed density–duration response and constrains, without fully identifying, the mechanism: it requires no new visual content, is strongly modulated by temporal addressing, and manifests as a coordinated change in placement and extent. Two questions follow: is the response learned, and can it be exploited at inference time without touching the weights? The first motivates a targeted training intervention, the second our inference-time method, Zoom-VTG.

### 5.1 The density response is modifiable through training

Matched fine-tunes differing in training duration composition change not just baseline short-event accuracy but the sensitivity to density itself. Across three seeds, balancing the duration distribution reduces the short-event $4\to8$ fps gain by $0.051$ mIoU and shifts the zero-crossing $1.6$ s earlier. The replication unit is the training seed, not the query, so at $n$=3 we report sign agreement rather than an interval: the arm$\times$density interaction on $\leq$1 s (mean $\mathbf{-0.0509}$) and the crossing shift (mean $\mathbf{-1.591}$ s) share their sign across all three seeds (Appendix A.1.4).

The response also predates the released model's reinforcement-learning stage: on the supervised-fine-tuned (SFT) parent the phase-averaged DiD is $+0.0672$ (CI $[+0.015, +0.117]$, $p$=0.0099), close to the released model's canonical response, and one-second span emission rises $9\% \to 29\%$ under the SFT parent and $7\% \to 29\%$ after RL (Appendix A.1.4). These interventions show that the density response is not fixed by architecture alone and is consistent with learned duration-sensitive behavior.

**A targeted training-side check gives consistent but weaker evidence.** We additionally train a LoRA calibration, **Temporal Resolution Debiasing** (TRD), that explicitly penalizes disagreement between coarse- and high-density predictions via an answer-token $\mathrm{KL}(p_{\text{low}} \| \mathrm{sg}[p_{\text{high}}])$ term. On Charades, it reduces the short-minus-long interaction $D$ in all three seeds at native phase, through the leading-edge channel the diagnosis predicts, although the effect is smaller and seed-heterogeneous after phase correction. We therefore treat this as secondary evidence that the response can be altered by training, rather than as a proposed training method; the full objective, ablations, and regime-dependent transfer are in Appendix A.1.4.

### 5.2 Zoom-VTG: duration-conditioned density allocation

The experiments above show that the response to temporal density is both signed and modifiable, but eliminating it through retraining is unnecessary for inference: we can instead exploit the response as measured, allocating additional density only to queries for which it has positive expected value, with no weight updates at all. Our contribution is not a new zoom operator, but a principled criterion for *when* temporal zoom should be used: Zoom-VTG turns the measured response into a routing rule, invoking established coarse-to-fine re-decoding (Nie et al., 2024; Chen et al., 2026) *conditionally* on the measured interaction rather than unconditionally as in prior work. A frozen grounder runs a

coarse first pass and a higher-density second pass only when the first pass indicates density will help. No checkpoint modification is required, only a grounder able to express the target short-span regime (Appendix A.4.2).

**The gate.** Let $r(L, \rho)$ denote expected grounding quality for an event of true duration $L$ decoded at temporal density $\rho$; Sections 4–5.1 characterize this response, including the sign change of $r(L, \rho_2) - r(L, \rho_1)$ at the measured crossing $L^\star \approx 2.3$ s. For motivation only, a second pass at higher density $\rho_2 > \rho_1$ is worth its compute whenever its expected marginal benefit exceeds its cost, i.e. whenever

$$G^\star(q) = \mathbf{1}\big[\mathbb{E}_{L|\hat{y}_1,q}[r(L, \rho_2) - r(L, \rho_1)] > \lambda C_2\big] \tag{1}$$

($C_2$: pass-2 cost; $\lambda$: the accuracy–compute trade-off) — an ideal rule we never estimate, since $r(L, \rho)$ and $\lambda C_2$ are not directly observed. True duration is likewise unobserved at inference time, so the deployed gate instead thresholds the first-pass predicted length $\hat{L}_1 = \hat{e}_1 - \hat{s}_1$, a monotone proxy for it, on a single value $T$ selected on held-out data:

$$G(q) = \mathbf{1}\Big[\hat{L}_1 \le T\Big]. \tag{2}$$

$G$ is a practical approximation to $G^\star$, not a calibrated estimate of the posterior or the ideal utility. Importantly, $T$ is *not* set equal to the 2.3 s ground-truth crossing $L^\star$: the gate observes predicted, not true, duration, and pass-1 predictions are systematically over-long for short events, so $T$ must exceed $L^\star$ to select the intended regime. We select $T$ on disjoint dev-300; the resulting $T$=8 s gate separates queries with median *true* length 2.0 s against 14.0 s when closed (Appendix A.5.6).

**Inference procedure.** Pass 1 grounds the full video at rate $f_1$, giving $\hat{y}_1 = [\hat{s}_1, \hat{e}_1]$ and its length $\hat{L}_1$. If the gate is closed, the method returns $\hat{y}_1$ after a single forward pass. If it opens, the method pads $\hat{y}_1$ by $p$ seconds on each side — widening to a minimum width $w_{\text{min}}$, shifted inward at the video boundaries — to obtain a window $W = [w_s, w_e]$, and spends its fixed budget $B$ inside that shorter window. This raises the effective rate to $f_2 = |F_2|/|W| \gg f_1$ without increasing the pass-2 budget: density comes from narrowing the window rather than enlarging $B$. Because the crop carries no absolute offset, the second prediction is mapped back as $[w_s + \hat{s}_2,\ w_s + \hat{e}_2]$. A pass-2 parse failure falls back to $\hat{y}_1$; a pass-1 failure returns the evaluator's parse-failure sentinel. The complete branching logic is Algorithm 1 (Appendix A.2.1).

**Instantiation.** For TimeLens2-4B we use $f_1$=4 fps, $B$=128, $B_{\text{full}}$=2048, $T$=8 s, and $p$=4 s, and $w_{\text{min}}$=8 s ($2p$). Only $p$ and $T$ are fitted, on a disjoint dev-300 split, then frozen for everything that follows. We keep pass 1 at the released model's native 4 fps rather than 8 fps: the denser setting raises mean input frames by 58% while cutting the mean zoom factor available inside the crop from 2.64× to 1.36×, leaving less density for the second pass to gain. Pass 1 is deliberately a localization-and-routing pass, not a high-density final prediction; remaining settings are in Appendix A.5.7.

**Gate calibration and robustness.** The gate is not sensitive to a narrow choice of $T$: every threshold tested on dev-300 improves over pass 1, with an interior optimum at $T$=8 s (**0.6093**). Frozen on TEST-400, this gate achieves 0.6081 versus 0.5759 for pass 1, closely matching a ground-truth-duration oracle at the same threshold (0.6097); the much larger gap to the per-query best-pass oracle (0.6714) places remaining headroom in predicting pass-2 benefit, not in the duration proxy itself (Appendix A.5.6). Unconditional pass 2 helps below $T$ (+0.0446) and hurts above it (−0.0454), exactly decomposing the gain into the open group's advantage and the avoided closed-group cost.

## 6 Results

All Zoom-VTG results below use $p, T$ selected once on dev-300 in Section 5.2 and then frozen, with no refitting by dataset or duration bin; the population result later in this section is descriptive accounting, not a further held-out test.

**Held-out accuracy and temporal allocation.** Table 2 places every density baseline on held-out TEST-400: global 10.55 fps edges out overall mIoU at 49% more frames, unconditional two-pass keeps the short-event gain but damages long events, and gating recovers most of that loss more cheaply (exact frame-matched control below; not mIoU-specific, Appendix Table 30).

**The reallocation, at full population size.** On all 3,363 Charades queries (Appendix Table 14, population accounting), one-second events gain +0.1366 while the aggregate moves only +0.0014

Table 2: **Held-out accuracy and temporal allocation, TEST-400.** Same frozen checkpoint and prompt throughout. Frames/query is the nominal allocation.

| Method | Frames/ query | All | ≤1 s | 1–3 s | 3–10 s | >10 s |
|---|---|---|---|---|---|---|
| Global 4 fps | **120.2** | 0.5759 | 0.3622 | 0.5670 | **0.6736** | **0.7007** |
| Global 8 fps | 240.8 | 0.5944 | 0.4569 | 0.6020 | 0.6338 | 0.6849 |
| Global 10.55 fps | 317.8 | **0.6125** | 0.5025 | **0.6193** | 0.6369 | 0.6914 |
| Unconditional two-pass | 248.2 | 0.5955 | 0.5341 | 0.5835 | 0.6110 | 0.6534 |
| **Gated Zoom-VTG** | 212.7 | 0.6081 | **0.5342** | 0.5745 | 0.6263 | 0.6974 |

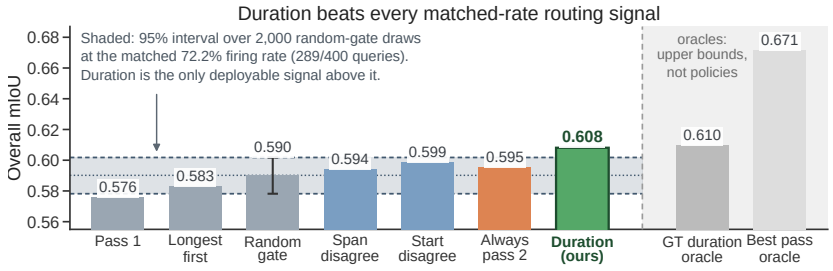



Figure 4: **Predicted duration is the only tested deployable routing signal that beats chance.** Gated policies matched at 72.2% pass-2 firing (pass-1-only/ungated/oracle unmatched). Band is the 95% random-gate interval. Inverting duration falls below random, as the signed response predicts.

($p$=0.75), from opposing duration-specific effects rather than indiscriminate span shortening: predicted length falls 3.32 → 2.49 s while R@$1_{.7}$ rises 0.058 → 0.303 on the same queries (localization-safety and per-bin failure rates: Appendix A.3.2).

**The sign change is not specific to mIoU, or to the gated policy.** Under the 4→8 fps intervention itself, R@$1_{.7}$ rises 0.060→0.230 on one-second events but falls 0.530→0.450 on 3–10 s events, and median boundary error splits the same way. These are default-phase diagnostics; the phase-averaged mIoU analysis remains primary (Appendix A.5.8).

**What duration-aware reporting reveals.** The reallocation is a property of the *predictions*, not the metric — but the summary statistic decides whether it is visible. Writing $m_b$ for mIoU in duration bin $b$ of the four bins $\mathcal{B}$: the **Duration Performance Profile** is the vector $(m_b)_{b\in\mathcal{B}}$ itself (Appendix Table 14); **duration-balanced mIoU** is $|\mathcal{B}|^{-1}\sum_b m_b$, dropping the population weights through which the conventional aggregate inherits the benchmark's duration mixture; and **WorstDur** is $\min_b m_b$, the duration analogue of worst-group reporting (Sagawa et al., 2020). On identical predictions the conventional aggregate reports +0.001, duration-balanced mIoU +0.017, and WorstDur +0.137 (Appendix A.3.4). We report the profile as primary: the method reallocates rather than uniformly improves, and localizing a one-second event precisely is not interchangeable with equal IoU off a twenty-second one.

**Routing-signal ablation: predicted duration versus the tested alternatives.** The gate fires on 72% of TEST-400 queries, so a policy spending compute on *any* 72% might do as well. Holding the firing rate fixed and varying only the signal (Figure 4; Appendix A.3.5), the duration gate is the only tested deployable signal above the random-gate interval, and beats both self-uncertainty gates built from the same $K$=8 sample cache. Adaptive compute alone does not explain the gain.

**Fast-only, slow-only, and duration-adaptive inference.** Table 2's global-4 fps, unconditional-two-pass, and gated-Zoom-VTG rows are exactly a fast-only / slow-only / adaptive comparison: the unconditional slow path (SlowFocus's mechanism (Nie et al., 2024)) improves over fast-only but damages long events, while routing improves further on fewer frames. Since 96% of the ≤1 s bin is gated anyway, gating retains the entire short-event benefit of unconditional two-pass (+0.0001 difference) while recovering most of its long-event cost (>10 s: 0.6534 → 0.6974, $p$=0.002; aggregate 0.5955 → 0.6081, $p$=0.012) on 14% fewer frames.

Against the stricter exact frame-matched single-pass uniform control (Appendix A.3.2), the result is again reallocation, not free accuracy: +0.062 on ≤1 s at $n$=100, −0.035 on 3–10 s, indistinguishable in aggregate (+0.008, ns). It is not paid for on the remaining systems axes: at equal frames, Zoom-

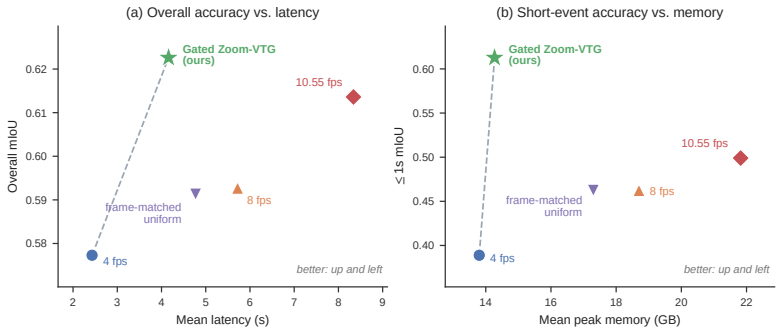



Figure 5: **The short-event gain costs less than dense global sampling** ($n$=100 systems-profiled subset; *dashed*: non-dominated frontier; memory is mean per-query CUDA peak). Gated Zoom-VTG is non-dominated on both pairs, and the exact frame-matched uniform baseline — which by construction receives the same number of sampled images — sits strictly inside the frontier.

VTG has lower measured latency and peak memory than the frame-matched uniform baseline, and is non-dominated on both accuracy–cost pairs (Figure 5; Appendix A.3.6).

**The benefit of density is non-monotonic even within a fixed proposal window.** Varying only the pass-2 budget at a fixed proposal window separates context narrowing from densification: a crop-only re-decode at the pass-1 rate recovers only a third to a half of the one-second gain out of sample ($+0.036$ of $+0.104$ on ActivityNet, CI crossing zero; $+0.046$ of $+0.099$ on QVHighlights). The gain then grows with budget to 128 frames but *falls* at 256 ($+0.059$ / $+0.061$): even inside a fixed window, density is not monotonically beneficial (Appendix A.3.4).

**Transfer and scope.** With $p, T, B$ still frozen from Charades, Table 3 (left) shows the one-second gain replicating on ActivityNet and QVHighlights, though the aggregate gain does not travel with it (Appendix A.3.3); on TEST-400 the policy beats global 10.55 fps on one-second mIoU by $+0.0317$ (tied, $n$=100) at 33% fewer frames, and by $+0.0433$ on the full population ($n$=590; Appendix A.3.2). The interaction itself is not unique to TimeLens2 (Table 3, right): InternVL3.5-4B reproduces it in 3/3 seeds after TimeLens language LoRA (relative ordering only, since it shares the Qwen3 LM family), and non-Qwen Gemma 4 goes from a null zero-shot DiD to positive in 2/2 seeds under the same adaptation — currently requiring TimeLens adaptation. Specialists lacking short-span output cannot show the interaction either way (Appendix A.4.2), consistent with a short-span-support scope condition rather than an architecture-specific one.

Table 3: **Transfer and scope.** *Left:* with hyperparameters frozen from Charades, the targeted short-event gain transfers to all three domains; aggregate gain does not. *Right:* the signed interaction appears across adapted model stacks, but every positive replication uses TimeLens supervision. Different estimands, not pooled.

*Frozen Zoom-VTG transfer*

| Domain | Δ ≤ 1 s | Δ all |
|---|---|---|
| Charades | **+0.172*** | +0.032* |
| ActivityNet | **+0.091*** | +0.012 |
| QVHighlights | **+0.089*** | +0.018 |

*Adapted density-interaction support*

| Stack | Seeds | Short–long DiD |
|---|---|---|
| TimeLens2 | canonical | **+0.077** |
| InternVL3.5 | 3/3 | **+0.092** ± 0.020 |
| Gemma 4 LoRA | 2/2 | **+0.126** ± 0.034 |

*95% CI excludes zero; ±=SD across seeds (Appendix A.4.2).

## 7 Conclusion

Temporal resolution in generative video grounding is not a monotonic resource: phase-averaged densification improves short events, degrades medium ones, and is neutral in aggregate. This benefit needs no new visual content, is modulated by readable temporal addressing, and is training-modifiable.

Zoom-VTG converts this into an inference-time rule that spends a denser pass only when predicted duration is short (Appendix A.2.1), gaining on held-out Charades and transferring to ActivityNet and QVHighlights, offset on the full population by medium-duration losses ($+0.137$ shortest bin, $+0.001$ aggregate), motivating a **Duration Performance Profile** (Appendix Table 14). More broadly, this is duration-conditioned reallocation, not a global hyperparameter, and duration-stratified, phase-controlled evaluation should become standard practice.

### Ethics statement

This work uses only existing public academic benchmarks (Charades, ActivityNet, QVHighlights) and their published re-annotations; it collects no new human-subject data. The primary inference-time experiments evaluate released, frozen models; separately identified training-side experiments (duration-composition fine-tunes, TRD, and cross-architecture adaptation) use only these same public benchmarks and introduce no new capability beyond the diagnostic scope of this paper. Improved fine-grained video temporal grounding carries the same dual-use profile as the broader video-understanding literature it builds on (e.g. surveillance applications), and we are not aware of a risk specific to this paper's contribution beyond that shared baseline.

### Reproducibility statement

Frozen index lists with checksums, decoding configuration, bootstrap/clustering procedures, and pre-specified endpoints are described in Appendix A.5.7; the scripts regenerating the reported numbers and figures are included in the supplementary material. All interventions score against unchanged ground-truth annotations, with a fingerprint over (index, start, end) identical across every arm. Row-level predictions and complete analysis artifacts are provided for TEST-400 and the controlled intervention arms. For the benchmark-wide 3,363-query population accounting, the supplementary material provides the frozen aggregate results and analysis configuration rather than row-level prediction files; reproducing or re-filtering that population analysis (e.g. by excluded index) requires rerunning inference.

### AI use statement

The authors used generative AI tools to polish and edit portions of the manuscript and as coding assistants for implementation support, debugging, and refactoring. The authors designed the study, selected the experiments and analyses, interpreted the results, inspected and tested AI-assisted code and outputs, and checked every reported number against the frozen experimental artifacts. No generated text, result, or artifact was accepted without author review. The authors take full responsibility for the manuscript and its claims.

## References

Shuai Bai et al. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025.

Kai Chen, Ming Dai, Wenxuan Cheng, and Wankou Yang. ScanFocus: A coarse-to-fine framework for spatio-temporal video grounding. *arXiv preprint arXiv:2607.13421*, 2026.

Gheorghe Comanici et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Yue Fan, Wenqi Zhao, Jiawei Xu, Kunyu Chen, and Boyang Li. GroundVTS: Visual token sampling in multimodal large language models for video temporal grounding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. TALL: Temporal activity localization via language query. In *IEEE International Conference on Computer Vision*, 2017.

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*, 2025.

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

Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. VTimeLLM: Empower LLM to grasp video moments. *arXiv preprint arXiv:2311.18445*, 2023.

Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *IEEE International Conference on Computer Vision*, 2017.

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*, 2021.

Zeqian Li, Shangzhe Di, Zhonghua Zhai, Weilin Huang, Yanfeng Wang, and Weidi Xie. Universal video temporal grounding with generative multi-modal large language models. In *Advances in Neural Information Processing Systems*, 2025.

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 *IEEE/CVF International Conference on Computer Vision*, 2023.

WonJun Moon, Sangeek Hyun, SangUk Park, Dongchan Park, and Jae-Pil Heo. Query-dependent video representation for moment retrieval and highlight detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023.

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*, 2024.

OpenAI. OpenAI GPT-5 system card, 2026. URL `https://arxiv.org/abs/2601.03267`.

Mayu Otani, Yuta Nakashima, Esa Rahtu, and Janne Heikkilä. Uncovering hidden challenges in query-based video moment retrieval. In *British Machine Vision Conference*, 2020. doi: 10.5244/C.34.84.

Seojeong Park, Jiho Choi, Kyungjune Baek, and Hyunjung Shim. MomentMix augmentation with length-aware DETR for temporally robust moment retrieval. In *IEEE/CVF Winter Conference on Applications of Computer Vision*, 2026.

Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. TimeChat: A time-sensitive multimodal large language model for long video understanding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *International Conference on Learning Representations*, 2020.

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.

Shihao Wang, Guo Liu, Han Zhao, and Fan Yu. VideoITG: Multimodal video understanding with instructed temporal grounding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Xizi Wang, Feng Cheng, Ziyang Wang, Huiyu Wang, Md Mohaiminul Islam, Lorenzo Torresani, Mohit Bansal, Gedas Bertasius, and David Crandall. TimeRefine: Temporal grounding with time refining video LLM, 2025. URL `https://arxiv.org/abs/2412.09601`.

Yuhui Zeng, Xinyu Mao, Xiaokun Liu, Xin Tao, Jinfa Huang, Jiayi Ji, and Xiawu Zheng. TimePLE: Rethinking temporal representation for video temporal grounding, 2026. URL `https://arxiv.org/abs/2607.23951`.

Jun Zhang et al. TimeLens: Rethinking video temporal grounding with multimodal LLMs. *arXiv preprint arXiv:2512.14698*, 2025a.

Shaojie Zhang, Jiahui Yang, Jianqin Yin, Zhenbo Luo, and Jian Luan. Q-Frame: Query-aware frame selection and multi-resolution adaptation for video-LLMs. In *IEEE/CVF International Conference on Computer Vision*, 2025b.

Zhengbo Zhang, Mark He Huang, Zhigang Tu, and Ming-Hsuan Yang. DART: Difficulty-adaptive routing for zero-shot video temporal grounding, 2026. URL `https://arxiv.org/abs/2607.00672`.

Yuhan Zhu, Changlian Ma, Xiangyu Zeng, Xinhao Li, Zhiqiu Zhang, Songze Li, Jun Zhang, Tianxiang Jiang, Yuandong Yang, Ziang Yan, Zikang Wang, Xinyu Chen, Haoran Chen, Shaowei Zhang, and Limin Wang. TimeLens2: Generalist video temporal grounding with multimodal LLMs. *arXiv preprint arXiv:2607.17423*, 2026.