SEEING THE SMALL IN THE LONG: RESOLVING WHERE AS WELL AS WHEN FOR TINY EVIDENCE IN LONG VIDEOS

Anonymous authors
Paper under double-blind review

ABSTRACT

Agentic long-video systems have learned to search the timeline, and left one regime unresolved: *seeing the small in the long*. Every one we evaluate exposes a single tool signature, given a time interval return more frames, so it re-samples *when* the model looks and never *where in the frame*. That is insufficient when the queried detail is illegible at the resolution those frames arrive in. To measure that regime we build **FoveaBench**: $1{,}044$ open-ended questions over 454 videos whose answers turn on a target occupying a median 0.31% of the frame of a median 23-minute video. An item is admitted only if a gate model answers *wrongly* from a whole-frame span near the evidence moment and *correctly* from a native-resolution view of it. What the suite exposes is an inversion rather than uniformly low scores: the tool-augmented systems we evaluate score *below* the tool-free model they are built on, though all of them call their tool on every question. An oracle study over the same items prices what focus is worth, and corrects an intuition: the anchoring interval at native resolution is worth 4.2%; the ground-truth box cropped out is worth *less* than the frame it was cut from; that same box rendered as a tracked, dilated, context-preserving view is worth 19.6%. That last view changes several things at once, so it prices the rendering pipeline as a bundle rather than any ingredient of it. We then build **FoveaSeek**, which resolves *when* first and *where in the frame* second: it commits to an interval together with an object hypothesis, localises that hypothesis itself as a policy-emitted box rather than by a detector call, so *where* to look is an action that rollout-level reward can reach, and answers from native-resolution close-ups of the propagated tube. Fine-tuning an 8B backbone reaches 25.2%, against 18.2% for the strongest evaluated baseline; a single-seed reinforcement-learning run under a temporal process reward reaches 31.1%. FoveaSeek demonstrates that the regime is addressable; which ingredient of the rendering bundle carries the gain is left to future measurement.

1 INTRODUCTION

Long-video question answering has reorganised itself around agents that act on the video rather than merely consume it. Rather than fixing one uniform frame budget before the question is read, systems such as LongVT (Yang et al., 2026a) and Video-o3 (Zeng et al., 2026) interleave reasoning with tool calls that re-sample the footage on demand, and their successors extend the idea in several directions (Yang et al., 2026b; Gao et al., 2026; Zhang et al., 2026a; Wang et al., 2025c). A long video sampled uniformly at an affordable budget is mostly irrelevant to any particular question, so letting the model look again beats forcing it to decide once (Figure 1).

What has not changed is the type signature of the tool. Every system we were able to obtain and run exposes one video-cropping operation whose arguments select *when* to look and never *where in the frame*: a denser sample of whole frames comes back, still under the per-frame pixel cap that made the overview affordable; recent successors inherit the same contract (Liu et al., 2026; Li et al., 2026). It re-samples time and nothing else. That is adequate when the evidence is a scene-scale event, since knowing *which* eight seconds matter is then the whole problem. It is not when the evidence is a small object: a target at three tenths of one percent of the frame occupies well under a single visual

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

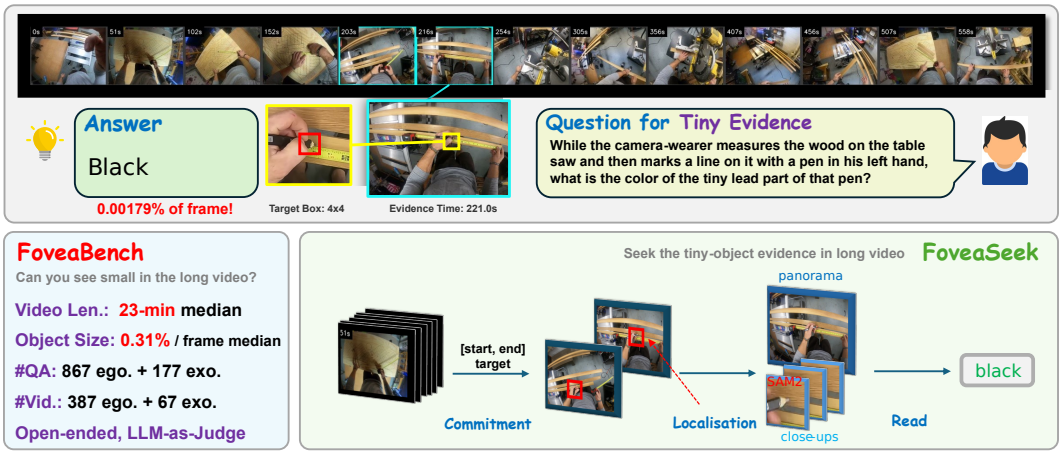


Figure 1: **Top:** one shipped FoveaBench item, uniformly skimmed, as an instance of *seeing the small in the long*: the queried detail is unreadable at that resolution, so more whole frames from the same interval return more of the same unreadable pixels. The target box is stated in the 0–1000 coordinates the agent emits. **Bottom left:** the suite at a glance, 1,044 open-ended questions over 454 videos in two visually opposite domains, with a median target of 0.31% of the frame of a median 23-minute video, scored by a model judge. **Bottom right:** FoveaSeek *commits* to an interval together with an object hypothesis, *localises* that hypothesis into a box, seeds SAM 2 with it to propagate a tube, and *reads* the answer from native-resolution close-ups with a context panorama.

token once the frame is capped at a 160-pixel long side, whichever second it came from. That is the regime of *seeing the small in the long*.

This makes a falsifiable prediction: on data whose evidence is small, a system whose only tool acts on the temporal axis should fail to beat a plain model reading the video directly. The prediction holds sharply. On **FoveaBench**, Conan (Ouyang et al., 2025) scores 13.8%, LongVT-RFT 12.7% and Video-o3 11.3%, each below the tool-free Qwen2.5-VL-7B (Bai et al., 2025b) they are built on, which scores 14.4%; the stronger tool-free backbone, Qwen3-VL-8B (Bai et al., 2025a), reaches 18.2%. Every figure is at each system's own uncapped configuration, which provisions the tool-free models most generously, and all three called their tool on every question, so none of this is a failure to act. They differ in training data, prompt and budget as well as in their tool, so this is an observation and not yet an attribution; the controlled form is Table 2, where only the tool's return value varies inside one trained model.

A benchmark on which strong systems underperform weak ones invites the rebuttal that the questions are simply unanswerable. The next section tests that, holding everything constant except what focus the tool-free base model is granted: the whole video yields 18.2%, the ground-truth box cropped out 23.1%, and the same box rendered as our tool renders it 37.8%. The questions are answerable once the evidence is delivered, and its form moves the answer rate as much as reaching the right moment does.

We therefore build **FoveaSeek**, an agent that resolves *when* the evidence occurs and then *where in the frame* it sits. It commits to an interval *together with an object hypothesis, a description of what it expects to find there*, localises that hypothesis itself on the frames it retrieves, propagates the resulting box into a spatio-temporal tube with SAM 2 (Ravi et al., 2025), and reads the answer from native-resolution close-ups accompanied by a context panorama. Because the box is emitted by the policy rather than by a detector, it is a sequence of actions that rollout-level reward can reach; and because the agent then consumes rendered pixels of the region it chose, its commitment is verifiable from what it subsequently sees. We train by imitation on trajectories whose evidence instants and seed boxes are ground truth, then optimise with group-relative policy optimisation (Shao et al., 2024; Guo et al., 2025) under a process reward on the committed moment.

Our contributions are a benchmark, a diagnosis and a proof of concept. **FoveaBench**, a suite of 1,044 open-ended questions over 454 long videos in which the need for both *when* and *where* is

decided per item by a gate model rather than asserted by a rubric, with a ground-truth target box on every item. Its value is that it is *diagnostic* rather than merely hard: it targets the regime of seeing the small in the long, where a temporal tool does return the pixels carrying the answer but not at a resolution that makes them legible. **A measurement of what focus is worth** on those items, from a five-condition oracle study, which prices the rendering pipeline as a bundle. **FoveaSeek**, an agent that resolves *where* as well as *when*, offered as a demonstration that the regime is addressable: it varies what the tool returns at fixed weights, which is what isolates the interface from the policy consuming it.

2 Related Work

Agentic long-video reasoning with temporal tools. The systems closest to ours give a large multimodal model native access to a video-cropping tool and train it to interleave calls with reasoning. Recent entrants extend rather than revise this design, adding parallel invocation (Yang et al., 2026b), tighter coupling of grounding to question answering (Liu et al., 2026), rewatch actions (Li et al., 2026), self-correcting and reward-incentivised search (Zhang et al., 2026a; Goel et al., 2026), chain-of-tool-thought over egocentric footage (Tian et al., 2026), iterative evidence seeking (Wang et al., 2025c; Zhang et al., 2026d), multi-scale frame selection (Ouyang et al., 2025) and native tool use across audio and video (Chen et al., 2026b). Online hierarchical memory (Ji et al., 2026) and instance-level spatial focusing (Zhao et al., 2026; Gao et al., 2026) depart from that contract rather than extending it. What the rest share is the axis they act on: each re-samples *when* the model looks, and none exposes an argument that selects a region *within* the frame. A denser set of whole frames comes back, each still under the per-frame pixel cap that made the first pass affordable, so asking again changes which seconds are on screen rather than how much of any one is resolvable. Their reward designs point in opposite directions, one paying the tool with a localisation term and the other taxing every call, and neither knob helps if a question is answerable from the global gist. Our gates instead select items on which the gate model benefits from magnified evidence, so the accuracy term pays for the tool and the degeneracy those knobs target is removed at its source (Appendix B).

Spatial zoom and spatio-temporal grounding. Magnifying part of a visual input is well established for single images (Wu & Xie, 2024; Zheng et al., 2026; Huang et al., 2026), with recent work isolating when it helps at all (Ma et al., 2026), addressing grounding brittleness at test time (Jiang et al., 2026) and supervising its use (Dong et al., 2026). Most of it neither searches over time nor maintains identity across frames: a region selected on one frame has to survive the motion and brief occlusions a minutes-long window contains, which is why our tool propagates the box into a tube. A separate line reaches fine detail by *allocating* resolution rather than by naming a region: Q-Frame (Zhang et al., 2025) selects query-relevant frames and spends a fixed visual-token budget unevenly across them, and it is a control our experiments do not include. Two agents are exceptions to the temporal-only reading above: VideoSeeker (Zhao et al., 2026) addresses both axes, and VideoSearcher (Gao et al., 2026) unifies temporal localisation with spatial focusing in one trajectory. Neither ships a suite gate-selected as needing spatial access, and neither has public weights we could run, so the closest spatial baselines are unmeasured here (see Limitations). A parallel literature treats spatio-temporal grounding as the objective rather than the means (Meng et al., 2025; Gu et al., 2025; Chen et al., 2026a; Tu et al., 2026; Wang et al., 2026; Li et al., 2025a; Zhang et al., 2026c; Lu et al., 2026; Feng et al., 2026), and closest to our reward are visual-perception prompting (Lee et al., 2026) and semantic evidence rewards (Xia et al., 2026); ours scores grounding *as an intermediate commitment inside a question-answering rollout* rather than as the output being asked for.

Evaluation suites for long video. General long-video benchmarks (Fu et al., 2025; Zhou et al., 2025; Wu et al., 2024; Wang et al., 2025b) and reasoning-oriented ones (Cheng et al., 2025; Zhu et al., 2026; Zhao et al., 2025; Hu et al., 2025) test whether a model can answer hard questions about long footage, and we use seven to characterise what our specialisation costs. We are not aware of one constructed so that spatial magnification is necessary, nor of one releasing the per-item target-box supervision that would let localisation be scored. That supervision matters for diagnosis: with a reference box a wrong answer separates reaching the wrong moment from attending to the wrong object inside the right one. Evidence-localising suites (Zhang et al., 2026b; Chen et al., 2025) sit on shorter or differently structured material, and the set shipped with LongVT has three columns (video path, question and answer), so localisation is unmeasurable on it. Our own source annotations are

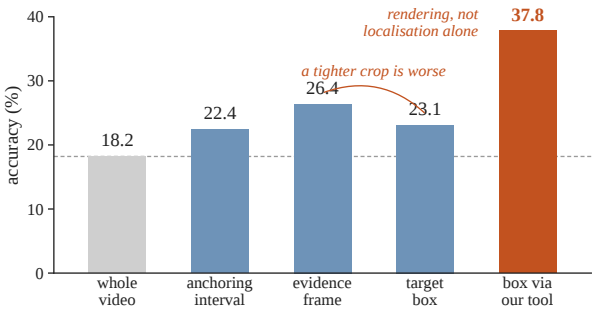


Figure 2: What perfect focus is worth to a tool-free model. Qwen3-VL-8B at its official visual budget, no system prompt, all 1,044 items, under the rubric that produces every other number here. The exact evidence frame at native resolution is worth 8.2%; the ground-truth box, cropped out, is worth less, not more; the same box through our tool is worth 19.6% (Appendix F).

inherited rather than created (Ramanathan et al., 2023; Grauman et al., 2025; Tang et al., 2023; Grauman et al., 2024).

3 Which Axis of Focus Carries the Information?

Before describing how FoveaBench is built we report the measurement that motivates it: when the evidence is a small object appearing briefly in a long video, how much accuracy does perfect focus buy, and does the form in which it is delivered matter? We answer on the tool-free base model itself, Qwen3-VL-8B at its official visual budget, over all 1,044 items under the same three-round rubric used everywhere else, rendering five information states for the identical question and reference answer. The endpoint is the whole video, what a tool-free system on a realistic budget sees, which scores 18.2%. The *anchoring interval* shipped with each item, sixty frames of it at native resolution, yields 22.4%; the single annotated evidence frame yields 26.4%; the ground-truth target box, a median 89×78 pixels, cropped out, yields 23.1%; and that same box seeded into our own tool, so SAM 2 propagates it into a tube and the model receives dilated close-ups together with context panoramas in which the region is outlined, yields 37.8%. All four gains are significant under 5,000 paired bootstrap resamples, and none saturates (Table 7): handed the exact instant at native resolution, with the search problem removed, the base model still reaches only 26.4%, so **search is a substantial bottleneck and rendering still matters after it**. It does not, however, bound what *allocating* resolution can buy: such a scheme could pair this frame with others inside the same budget, which is the control we do not run. Two readings follow, the second of which is the hinge.

A tighter crop is not a better crop. The ground-truth box scores 23.1%, *below* the whole frame it was cut from: at a median 89×78 pixels the target is worth about a hundred visual tokens once the processor's pixel floor up-samples it (Appendix D), and stripping away its surroundings removes the context that identifies it. Resolution and context are a pair, which is why our tool dilates its close-ups well past the box and returns them alongside a context panorama in which the region is outlined.

The rendering pipeline is worth more than perfect localisation. The same box cropped directly gives 23.1%; through our tool it gives 37.8%, a further 14.7%, wider than perfect localisation buys on its own, and with no appeal to training, since the model receiving both renderings is the untrained base. The two differ in several ways at once: dilation, the context panorama, multiple close-ups, propagation, and the budget buying them. This prices the rendering pipeline as a bundle rather than isolating an ingredient of it.

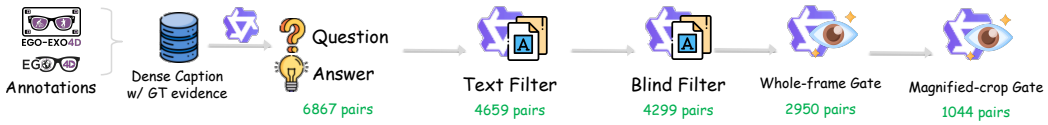


Figure 3: The FoveaBench funnel, counts pooled over both splits. Questions are generated per chapter from the inherited annotations and a dense caption carrying the ground-truth evidence; two conventional filters follow; the last two gates keep an item only if the gate model answers *wrongly* from whole frames and *correctly* once the target is magnified.

4 FoveaBench

FoveaBench exists to make one regime measurable: *seeing the small in the long*, where the evidence is a target of a few hundred pixels inside footage far too long to read at full resolution. General suites contain such items only incidentally and, to our knowledge, certify none of them.

The suite is 1,044 open-ended short-answer questions over 454 long videos in two visually opposite domains, and every question ships a ground-truth target box together with the instant at which it was annotated. The egocentric split contributes 867 questions over 387 head-mounted Ego4D videos, with targets from PACO's part-level labels and from EgoTracks (Ramanathan et al., 2023; Grauman et al., 2025; Tang et al., 2023); the exocentric split contributes 177 questions over 67 fixed-tripod Ego-Exo4D takes, with targets from that dataset's third-person instance masks (Grauman et al., 2024), which are segmentation-quality ground truth. Targets occupy a median 0.31% of the frame of a median 23-minute video; Appendix F gives the full statistics.

Every question pins its target by the *action performed with it* and then asks for a fine visual detail, and the suite is gate-selected so three requirements hold together: the action is a narrated, locally unique event, so the agent must skim to find the moment; identification proceeds by interaction rather than description, so it must watch that moment to know which object is meant; and at skim resolution the queried detail falls well under one visual token, so it must magnify to read it. The generator chooses which of an entity's annotations yields a well-posed question, so what survives is set by the gates rather than a hand-chosen taxonomy. Answers are free-form short text, 690 distinct strings, and the twenty most frequent cover 18.8% of items, coverage rather than a reachable score. The attribute annotations the source corpora supply are dominated by colour, which carries 939 of the 1,044 questions (Appendix F).

4.1 The Two Gates that Constitute the Contribution

One source-agnostic funnel serves both domains, with only two front-end files differing. Questions are generated per chapter with the whole-video outline supplied so uniqueness can be enforced, and the four filters of Figure 3 follow. The first two are conventional, culminating in a **blind filter** that discards anything answerable from language priors with no video at all.

The last two filters are gates, and they are the contribution: each is decided by a model rather than asserted by a rubric, and every stage of the funnel is run by one self-hosted Qwen3-VL-235B-A22B-Instruct-FP8 *gate model*. Both gates receive the same non-visual context: a *coarse* description of the scene and the action taking place, carrying no object-level detail and in particular not the dense caption the generator worked from. The only thing that varies between the two gates is therefore how the evidence is rendered. The **whole-frame gate** adds a span around the evidence moment as *full frames* at a 336-pixel long side and keeps an item only if the gate model gets it *wrong*, selecting items for which that model indicates magnification is necessary. The **magnified-crop gate** adds one full frame together with a native-resolution crop of the target box and a sentence stating where that crop sits, and keeps the item only if the gate model is now *right*, selecting items for which it indicates magnification is sufficient. Both prompts forbid abstention: a gate permitting "I cannot tell" measures willingness to answer rather than legibility. Their conjunction is what admits an item, each judged at the rendering its gate saw, so an item passing only one is discarded.

What the gates license. The magnified-crop gate admits an item when a native-resolution view of the target makes the answer available, which is close to what our own tool returns, so the suite

is not neutral between spatial interfaces. Both verdicts are one model's at one rendering rather than proofs; the oracle study checks them against a second. Any spatial interface should gain here, and which gains most is a question to settle *inside* the suite.

Validation and caveats. A manual audit of 100 stratified candidates found the reference answer correct for 97 and magnification necessary for 95, both holding for 93. Two caveats belong with the benchmark rather than a footnote: the whole-frame gate probed a fixed span that for 16.5% of egocentric items excludes the annotated frame, and the magnified-crop gate is mildly self-selecting (Appendix I). A control against language priors sits inside the funnel: every candidate was put to the 235B gate model *with the question alone and no video*, and any it answered correctly was discarded, 137 of 1,983 in the egocentric funnel, so leakage to a text-only reader is removed by construction rather than measured; a constant-answer strategy also gets little from the suite (Appendix F).

5 FoveaSeek

5.1 Resolving Evidence in Three Stages

FoveaSeek resolves *when* the evidence occurs and then *where in the frame* it sits. The agent starts from a deliberately cheap *overview* in which the target is gate-selected as unreadable at that resolution and converts it into a state where the answer is legible, committing to the moment and then to the region, acting on the video in between, so the second commitment is made from evidence the first obtained.

Stage 1: the commitment. From the overview the agent commits to an *interval* together with an *object hypothesis*, a referring description of what it expects to find there, phrased with the action context that pins it down. Naming the object separates the procedure from temporal search: the agent is not asking merely to see the interval again, it is declaring a hypothesis whose two halves are both checked downstream. The video is re-read inside that interval at a resolution sufficient to locate the target, though not to read its detail.

Stage 2: the localisation. On those frames the agent localises its object hypothesis, emitting a *box* in the backbone's own multi-frame grounding format, so no new output convention must be learned. We are not aware of a counterpart to this step in long-video agents built on the temporal contract, and it is a decision made *by the policy inside the dialogue* rather than delegated to a module. That is what makes it reachable by training: the box is a sequence of actions the reward scores, which a frozen detector's output would not be.

Stage 3: the tube. That box seeds SAM 2 (Ravi et al., 2025), which propagates the region through the interval into a spatio-temporal *tube*. The video is read a final time as native-resolution *close-ups* of the tube, each accompanied by a *context panorama* in which the region is outlined. Two properties are deliberate: the agent consumes *rendered pixels* of the region it chose rather than a box and a confidence, so its commitment is verifiable from what it then sees; and propagation rather than single-frame cropping preserves identity across motion and brief occlusion when tracking holds, with frames in which the target is not visible skipped.

The stages compose into a loop rather than a pipeline: having seen the tube, the agent may judge its commitment wrong and re-commit, for at most four tool calls before it must answer, which is why the reward scores the *final* commitment rather than the first. When propagation fails the agent is told so and may answer or re-commit, a path present in the training data so the fallback is learned rather than improvised. All three stages are reached through a single action taking an interval and an object hypothesis, at a measured 8,448 overview visual tokens against Video-o3's declared 16,384, plus the close-ups each call returns (Appendix D).

5.2 Rewarding the Commitment, Not Only the Answer

Imitation teaches the shape of this procedure but not how to commit well, since a demonstration only ever shows a commitment that was already correct. We therefore score the commitment itself,

rating a rollout as

$$R = \underbrace{a}_{\text{answer}} + \underbrace{f}_{\text{format}} + \beta \underbrace{w}_{\text{when}}, \qquad \beta = 1.5,$$

where $a \in \{0, 1\}$ is a binary judge verdict under the benchmark's own construction rubric, f rewards well-formed output, and w scores the committed interval against the certified one, computed only on rollouts that called the tool, so answering directly cannot collect it. An answer-only signal cannot distinguish a correct answer reached from the right evidence from one reached despite the wrong evidence. Where in the frame to look carries no term of its own: sampling, cropping and propagation are not differentiable, so no loss flows back through the rendered pixels, and what the box tokens receive is the rollout-level advantage, since they are actions in the sequence the answer reward scores.

One property is load-bearing: w is **monotone in distance** rather than a thresholded overlap. A one-dimensional intersection-over-union is zero for most rollouts, so a group in which nothing overlaps carries *no temporal signal* under group-relative optimisation — in precisely the regime where finding the moment is the bottleneck, only a and f still vary. Appendix E gives the term in full.

6 Training

All experiments start from Qwen3-VL-8B-Instruct (Bai et al., 2025a) and adapt it with low-rank adaptation (Hu et al., 2022), freezing the vision tower and the projector, for one epoch over $12{,}866$ rows. Adapting rather than fully fine-tuning keeps the base recoverable: with the adapter disabled the model is byte-identical to it. Enabled, it does cost something, which Table 3 bounds against the quoted external rows rather than against the backbone itself. Reinforcement learning then uses group-relative policy optimisation (Shao et al., 2024; Guo et al., 2025) over $1{,}367$ items across 763 videos, each carrying a certified tube, its annotated instant and native frame dimensions. There is zero question overlap with supervised training and zero video overlap with the benchmark; Appendix M gives every hyperparameter.

Our tool trajectories are not written free-hand by a teacher. The evidence instant and the seed box are ground truth from the inherited annotations, so the teacher only writes prose, and only if it can answer the question *from the student's exact rendered view and without seeing the reference*; a wrong answer marks the trajectory not executable and discards it, since teaching a model to produce an answer its tool did not reveal is teaching it to hallucinate. Yield is $1{,}391$ egocentric trajectories over 539 videos from $4{,}577$ candidates, plus 190 exocentric ones up-sampled to the mixture's $1{,}771$ `fovea` rows (Appendix F).

Our supervised mixture is $12{,}866$ rows against LongVT's 247.9K and the 418K corpus on which Video-o3 builds, so we trained on far less and make no data-efficiency claim; Appendix B decomposes the three corpora by provenance and Appendix G gives our own composition row by row. What distinguishes our $1{,}771$ tool rows is not volume but grounding: each is seeded by a ground-truth instant and target box, and each survives the executability gate.

7 Experiments

Every method runs through *its own repository's* code path, prompts and budget. Competitors use their published settings; the base models use their libraries' defaults with no quota constraint, making them the most generously provisioned systems here, and Appendix C re-runs them under Video-o3's declared budget, under which our margin grows. Significance statements rest on paired tests (Appendix A). Accuracies are percentages of the $1{,}044$ items; every difference between two accuracies is stated in percentage points.

7.1 Main Results

Table 1 places each method beside the backbone it was built on, which is the comparison in which a tool can be said to have paid for itself. Each of them loses ground against its own base: Conan by 0.6%, LongVT-RFT by 1.7% and Video-o3 by 3.1%. Only the largest of the three is individually certified: the paired bootstrap of Appendix A separates Video-o3 from that backbone at $p = .017$,

Method	ego	exo	**all**	Δ base
built on Qwen2.5-VL-7B				
Qwen2.5-VL-7B *(base)*	14.5	13.6	14.4	—
Conan	13.1	16.9	13.8	−0.6
LongVT-RFT	12.1	15.8	12.7	−1.7
Video-o3	10.7	14.1	11.3	−3.1
built on Qwen3-VL-8B				
Qwen3-VL-8B *(base)*	17.5	21.5	18.2	—
FoveaSeek-SFT	**24.9**	**26.6**	**25.2**	**+7.0**
FoveaSeek-SFT+RL†	31.3	30.5	31.1	+12.9

Table 1: FoveaBench accuracy in percent, and what each method gains or loses *against the backbone it was built on*, which removes the backbone as a confound: the three tool-augmented systems all end below the Qwen2.5-VL-7B they start from despite calling their tool on every question, while ours moves the other way. Scored by a binary judge under the benchmark's own rubric, three-round majority vote over all 1,044 items, empty predictions counted wrong; Appendix J reports judge agreement. †Preliminary single-seed run, reported as exploratory; the bold imitation row is what we rest on.

Configuration	agent sees	all	Δ
Qwen3-VL-8B (no SFT)	—	18.2	−12.9
No re-reading	overview only	19.6	−11.5
Temporal only	whole frames	22.1	−9.0
No propagation	one close-up	27.9	−3.2
Full (ours)	**tube close-ups**	**31.1**	—

Table 2: Ablations at fixed weights, so only what the agent may retrieve varies. The opening row is the untrained backbone; in the rest the weights are the tube-trained ones, at the exploratory RL checkpoint: *Temporal only* is that policy evaluated under whole-frame returns, reproducing the prevailing contract *inside our own model* rather than a temporal-only agent trained as one, which a matched training run would be. *No propagation* keeps the crop geometry as the tool renders it, dilation and context view included, but returns one close-up instead of the tube, so it removes tracking together with the several close-ups the tube supplies and bounds propagation's contribution rather than isolating it.

while the 0.6% and 1.7% gaps are tested in the same block and fall within noise, so we do not claim them one by one. What the set supports is the direction, not each margin. Three groups, three recipes, all within 2.5% of one another and all calling their tool on every item, which is consistent with the shared type signature being what limits them; Table 2 is where that variable is actually controlled. Our own procedure, applied to Qwen3-VL-8B, moves the other way: 7.0% from imitation alone and 12.9% once the commitments are optimised.

7.2 Which Part of the Tool Carries the Gain?

Figure 2 removes the search problem: ground-truth evidence, no training. Table 2 keeps it, fixes the trained weights and varies only what the tool returns, so it measures how far performance depends on the observations the policy was trained to expect. It does not separate the interface from that training, nor isolate coupling.

Each row closes an objection. Denying the agent any re-reading leaves it near its untrained starting point, so the gain depends on what it may retrieve and not on the weights alone (Ma et al., 2026). Returning whole frames from the committed interval, the prevailing contract executed by our own weights, recovers only a fraction of the full gain, the trained-agent counterpart of the asymmetry Figure 2 measures on the untrained one. Removing propagation costs 3.2%, so within this bundle tracking and the extra close-ups it supplies are the smaller term and the close-ups themselves the larger; which ingredient of a close-up carries it is not separated here. Figure 2 also sizes what remains: the untrained backbone handed the annotated box reaches 37.8% against our agent's 31.1%.

Method	Base	VideoMME w/o subs	MLVU M-Avg	LVBench Avg	LVideoB Avg	VMMMU Overall	MMVU M-Avg	MMR-V Overall
Qwen2.5-VL *(no tool)*	7B	65.1	70.2	45.3	56.0	47.4	61.3	32.4
Conan	7B	60.5	63.4	39.2	56.6	—	—	42.7
LongVT-RFT	7B	64.3	—	41.3	—	45.4	—	—
Video-o3	7B	**66.5**	**72.1**	**47.6**	**60.5**	51.7	**67.2**	**44.7**
FoveaSeek-SFT	8B	62.1	65.6	41.4	51.1	51.7	49.4	35.3
FoveaSeek-SFT+RL	8B	64.8	64.3	46.4	54.3	**53.4**	53.5	38.6

Table 3: General long-video ability, in percent, each suite under its own official extractor; dashes mark results the cited paper does not report. Competitor numbers are quoted from Video-o3's comparison table (Zeng et al., 2026), so they are at least internally consistent; our two rows are measured by us. Columns are *not* budget-matched, and Qwen3-VL-8B is not measured here, so what they bound is the specialisation cost against the quoted rows rather than against the backbone itself. Appendix K adds four further rows and every configuration.

Those systems differ in weights as well as in what they are told, so 6.7% is the distance between them, not a decomposition of search error.

The cost of the budget trade. Table 3 evaluates on seven public suites under their own official extractors: FoveaSeek stays within 2% of Video-o3 on Video-MME and LVBench and ahead on Video-MMMU, but trails on the rest, by the widest margin on MMVU. Those runs are not budget-matched and we claim no efficiency from them; Appendix K gives the allowances behind them.

8 Analysis

Reinforcement learning adds 5.9% over the imitation checkpoint, from 25.2% to 31.1%, improving *both* domains and leaving the policy ahead of every external system we measure by between 12.9% and 21.0% (Appendix K). We tune β against the shaping term's share of the *within-group* reward variation rather than its mean (Appendix H).

Limitations. Four things this paper does not establish. **No spatial or multi-resolution baseline is evaluated**: the closest spatially-aware agents ship no public weights, and a query-aware resolution allocator such as Q-Frame (Zhang et al., 2025) is the control that would say whether naming a region is necessary at all; the oracle bound above is the nearest thing we have to it and not a substitute for running it. We claim a working spatial procedure, not the best one. **Localisation is not scored in aggregate**: every item ships a target box, yet we report localisation case-wise only, so whether reinforcement learning improves spatial grounding or only temporal search is untested, as is the policy-emitted box against a detector-supplied one. The certified quantities are the target box and the annotated instant, not the shipped interval, so agreement with that interval is not a localisation metric (Appendix F). **Gate verdicts are relative to the gate model**, of the same family as the student, and the final gate admits exactly the items a native-resolution view answers, so the suite is not neutral between spatial interfaces; every result rests on one 8B backbone over two collections from one family. **The answer-prior control is a construction filter rather than a measurement**: candidates a text-only reader answered correctly were removed instead of scored, so the accuracy owed to visual evidence rather than to a prior fitted on the shipped suite is still not separated. Licences bar redistributing the footage, so a release comprises the manifest, funnel code, gate verdicts and harness.

9 Conclusion

Temporal tools re-sample time and nothing else, and on evidence this small that axis buys little: the anchoring interval at native resolution is worth 4.2%, the same box through the full rendering bundle 19.6%. The consequence is counter-intuitive: the tool-augmented systems we evaluate score below the base models they are built on while calling their tool on every question. An agent resolving *when* and then *where* reaches 25.2% against 18.2% by imitation alone, and *which* kind of focus pays is what the suite now exists to settle.

AI USE STATEMENT

Generative models are a component of the research method described here, and their use is disclosed in the body rather than only in this statement. A self-hosted Qwen3-VL-235B-A22B-Instruct-FP8 runs every stage of the FoveaBench funnel, both admission gates and the question generator (Appendix F). The accuracies reported in this paper are graded by GPT-5.5 through a commercial API, three independent runs combined by majority; the cross-family agreement check of Appendix J uses Claude Sonnet-5 by the same route. A locally served 30B instruct model at temperature zero grades the answer term of the reinforcement-learning reward and the development gates (Appendix E, Appendix J). The reliability of that judge is measured rather than assumed: Appendix J reports within-judge and cross-family agreement, and the human audit of Appendix I checks the gate verdicts against annotators. We also used generative AI tools for prose editing and for drafting and checking LaTeX. Beyond the model judge disclosed above, which produces every accuracy we report, we did not use generative AI tools to generate experimental results or as an author of the research ideas. We have reviewed all AI-assisted work, and we take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

FoveaBench is built entirely on top of previously released annotation corpora (Ramanathan et al., 2023; Grauman et al., 2025; Tang et al., 2023; Grauman et al., 2024); we collected no new footage and ran no new human-subject data collection. The egocentric sources carry consent and de-identification procedures established by their original releases, and we inherit rather than relax them. The licences on that footage bar redistribution, so what we release is the item manifest, the construction funnel with its per-item gate verdicts, and the evaluation harness, and the videos themselves are obtained by users from the upstream sources under those sources' own terms (Appendix N). The human validation of Appendix I was an annotation audit of already-released material by the authors rather than a study on human subjects. We see no application of this work that raises safety concerns beyond those already attached to long-video question answering in general, though we note that a system trained to resolve small detail in long footage is, by construction, better at recovering fine-grained content from recordings of people, and the licence terms above are the mechanism by which we keep that capability tied to already-consented material.

REPRODUCIBILITY STATEMENT

The construction funnel, its gate prompts and its per-item verdicts are described in Appendix F and released as part of the supplement described in Appendix N, which also states exactly which artifacts the licences allow us to ship. Every prompt used at every stage is reproduced verbatim in Appendix L: the question generator, the text filter, the blind filter and the two admission gates of the construction funnel, together with the agent system prompt, the user turn, the grounding instruction, the observation envelopes and the judge rubric, and the judge protocol together with its leakage controls and its measured agreement rates is in Appendix J. Training hyperparameters for both the supervised and the reinforcement-learning stages are in Appendix M, the supervised mixture is broken down by provenance in Appendix G, how the shaping coefficient is set is in Appendix H, and the visual budget accounting that underlies the budget-aligned re-test is in Appendix C and Appendix D. The paired-bootstrap procedure behind every significance statement in the main text is given in Appendix A, and every accuracy we report is recomputed from the per-item judge verdicts shipped with the harness rather than transcribed, so each table can be regenerated end to end. Appendix K and the Limitations paragraph state which comparisons are budget-matched and which are not, so that a reproduction attempt inherits the same caveats we do.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025a. doi: 10.48550/arxiv.2511.21631. URL https://arxiv.org/pdf/2511.21631.

Shuai Bai, Keqin Chen, Xuejing Liu, et al. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025b. doi: 10.48550/arxiv.2502.13923. URL https://arxiv.org/pdf/2502.13923.

Kai Chen, Ming Dai, Wenxuan Cheng, and Wankou Yang. Scanfocus: A coarse-to-fine framework for spatio-temporal video grounding. In *ECCV*, 2026a. URL https://arxiv.org/pdf/2607.13421.

Qirui Chen, Shangzhe Di, and Weidi Xie. Grounded multi-hop videoqa in long-form egocentric videos. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(2):2159–2167, 2025. doi: 10.1609/aaai.v39i2.32214. URL https://ojs.aaai.org/index.php/AAAI/article/download/32214/34369.

Yu Chen, Caorui Li, Ziyu Xiong, et al. Omnireasoner: Thinking with long audio-video via native tool use. *arXiv preprint arXiv:2607.19339*, 2026b.

Junhao Cheng, Yuying Ge, Teng Wang, Yixiao Ge, Jing Liao, and Ying Shan. Video-Holmes: Can MLLM think like Holmes for complex video reasoning? *arXiv preprint arXiv:2505.21374*, 2025. doi: 10.48550/arxiv.2505.21374. URL https://arxiv.org/pdf/2505.21374.

Qihua Dong, Gozde Sahin, Pei Wang, Zhaowei Cai, Robik Shrestha, Hao Yang, and Davide Modolo. Visual reasoning through tool-supervised reinforcement learning. In *CVPR Findings*, 2026. URL https://arxiv.org/pdf/2604.19945.

Kaituo Feng, Kaixiong Gong, Bohao Li, et al. Video-R1: Reinforcing video reasoning in MLLMs. In *NeurIPS*. Curran Associates, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/8eb3976840e08b80dda9667562574246-Abstract-Conference.html.

Kaituo Feng, Manyuan Zhang, Hongyu Li, et al. Onethinker: All-in-one reasoning model for image and video. In *CVPR*, 2026. doi: 10.48550/arxiv.2512.03043. URL https://arxiv.org/pdf/2512.03043.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, et al. Video-MME: The first-ever comprehensive evaluation benchmark of multi-modal LLMs in video analysis. In *CVPR*, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Fu_Video-MME_The_First-Ever_Comprehensive_Evaluation_Benchmark_of_Multi-modal_LLMs_in_CVPR_2025_paper.html.

Zhenkun Gao, Yicheng Bao, Jinlong Peng, et al. Videosearcher: Empowering video deep research with multi-tool agentic reasoning via reinforcement learning. *arXiv preprint arXiv:2607.02927*, 2026. URL https://arxiv.org/pdf/2607.02927.

Harsh Goel, S. P. Sharan, Sahil Shah, et al. Incentivizing vision language models to search for long video question answering. In *ECCV*, 2026.

Kristen Grauman, Andrew Westbury, Lorenzo Torresani, et al. Ego-exo4d: Understanding skilled human activity from first- and third-person perspectives. In *CVPR*, 2024. doi: 10.1109/CVPR52733.2024.01834. URL https://dblp.org/rec/conf/cvpr/GraumanWTKMAABB24.

Kristen Grauman, Andrew Westbury, Eugene Byrne, et al. Ego4D: Around the world in 3,600 hours of egocentric video. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2025. doi: 10.1109/TPAMI.2024.3381075. URL https://dblp.org/rec/journals/pami/GraumanWBCCFGHJKLLMNRRR25.

Xin Gu, Haoji Zhang, Qihang Fan, et al. Thinking with bounding boxes: Enhancing spatio-temporal video grounding via reinforcement fine-tuning. *arXiv preprint arXiv:2511.21375*, 2025. doi: 10.48550/arxiv.2511.21375. URL https://arxiv.org/pdf/2511.21375.

Daya Guo, Dejian Yang, Haowei Zhang, et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638, 2025. doi: 10.1038/s41586-025-09422-z. URL https://www.nature.com/articles/s41586-025-09422-z.pdf.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Yuanhan Zhang, Xiang Yue, Bo Li, and Ziwei Liu. Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos. *arXiv preprint arXiv:2501.13826*, 2025. doi: 10.48550/arXiv.2501.13826. URL https://dblp.org/rec/journals/corr/abs-2501-13826.

Xinyu Huang, Yuhao Dong, Weiwei Tian, Bo Li, Rui Feng, and Ziwei Liu. Mgpo: Thinking with images via multi-turn grounding-based reinforcement learning. In *Findings of ACL*, pp. 383–399, 2026. doi: 10.18653/v1/2026.findings-acl.18. URL https://aclanthology.org/2026.findings-acl.18.pdf.

Yixin Ji, Fanghua Ye, Juntao Li, et al. Homer: Understanding long-form videos with hierarchical memory and agentic reasoning. *arXiv preprint arXiv:2607.02588*, 2026.

Zheng Jiang, Yiming Chen, Nan He, Jiahui Chen, Chaoyang Li, Houde Qian, and Lifeng Sun. Test-time scaling over perception: Resolving the grounding paradox in thinking with images. *arXiv preprint arXiv:2604.11025*, 2026. URL https://arxiv.org/pdf/2604.11025.

Daeun Lee, Shoubin Yu, Yue Zhang, and Mohit Bansal. Visioncoach: Reinforcing grounded video reasoning via visual-perception prompting. *arXiv preprint arXiv:2603.14659*, 2026. URL https://arxiv.org/pdf/2603.14659.

Hongyu Li, Jinyu Chen, Ziyu Wei, Shaofei Huang, Tianrui Hui, Jialin Gao, Xiaoming Wei, and Si Liu. Llava-st: A multimodal large language model for fine-grained spatial-temporal understanding. In *CVPR*, 2025a. URL https://openaccess.thecvf.com/content/CVPR2025/html/Li_LLaVA-ST_A_Multimodal_Large_Language_Model_for_Fine-Grained_Spatial-Temporal_Understanding_CVPR_2025_paper.html.

Xinhao Li, Ziang Yan, Desen Meng, et al. VideoChat-R1: Enhancing spatio-temporal perception via reinforcement fine-tuning. *arXiv preprint arXiv:2504.06958*, 2025b. doi: 10.48550/arXiv.2504.06958. URL https://dblp.org/rec/journals/corr/abs-2504-06958.

Yixuan Li, Guangzhi Sun, Yudong Yang, and Chao Zhang. video-salmonn-r^3: Learning to rewatch, reask, and reanswer for efficient video understanding. *arXiv preprint arXiv:2606.24477*, 2026. URL https://arxiv.org/pdf/2606.24477.

Wenqi Liu, Yunxiao Wang, Shijie Ma, et al. Videotemp-o3: Harmonizing temporal grounding and video understanding in agentic thinking-with-videos. *arXiv preprint arXiv:2602.07801*, 2026. doi: 10.48550/arXiv.2602.07801. URL https://arxiv.org/pdf/2602.07801.

Zhu Lu, Tiantian Geng, Yangye Chen, Teng Wang, Ping Lu, and Feng Zheng. R-AVST: Empowering Video-LLMs with fine-grained spatio-temporal reasoning in complex audio-visual scenarios. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(9):7627–7635, 2026. doi: 10.1609/aaai.v40i9.37704. URL https://doi.org/10.1609/aaai.v40i9.37704.

Yan Ma, Weiyu Zhang, Tianle Li, Linge Du, Xuyang Shen, and Pengfei Liu. What does vision tool-use reinforcement learning really learn? disentangling tool-induced and intrinsic effects for crop-and-zoom. *arXiv preprint arXiv:2602.01334*, 2026. URL https://arxiv.org/pdf/2602.01334.

Jiahao Meng, Xiangtai Li, Haochen Wang, et al. Open-o3-video: Grounded video reasoning with explicit spatio-temporal evidence. *arXiv preprint arXiv:2510.20579*, 2025. doi: 10.48550/arxiv.2510.20579. URL https://arxiv.org/pdf/2510.20579.

Kun Ouyang, Yuanxin Liu, Linli Yao, Yishuo Cai, Hao Zhou, Jie Zhou, Fandong Meng, and Xu Sun. Conan: Progressive learning to reason like a detective over multi-scale visual evidence. *arXiv preprint arXiv:2510.20470*, 2025.

Vignesh Ramanathan, Anmol Kalia, Vladan Petrovic, et al. Paco: Parts and attributes of common objects. In *CVPR*, pp. 7141–7151, 2023.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, et al. SAM 2: Segment anything in images and videos. In *ICLR*, 2025. doi: 10.48550/arxiv.2408.00714. URL https://arxiv.org/pdf/2408.00714.

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

Guangming Sheng, Chi Zhang, Zilingfeng Ye, et al. HybridFlow: A flexible and efficient RLHF framework. In *EuroSys*, 2025. doi: 10.1145/3689031.3696075. URL https://arxiv.org/abs/2409.19256.

Hao Tang, Kevin J. Liang, Kristen Grauman, et al. Egotracks: Long-term egocentric visual object tracking. In *NeurIPS*, 2023.

Shulin Tian, Ruiqi Wang, Hongming Guo, et al. Ego-R1: Agentic chain-of-tool-thought for ultra-long egocentric video reasoning. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2026. doi: 10.1109/TPAMI.2026.3697367. URL https://pubmed.ncbi.nlm.nih.gov/42202198/.

Xuezhen Tu, Jingyu Wu, Fangyu Kang, Qingpeng Nong, Kaijin Zhang, Chaoyue Niu, and Fan Wu. Bridging time and space: Decoupled spatio-temporal alignment for video grounding. *arXiv preprint arXiv:2604.08014*, 2026. URL https://arxiv.org/pdf/2604.08014.

Jiankang Wang, Zhihan Zhang, Zhihang Liu, Yang Li, Jiannan Ge, Hongtao Xie, and Yongdong Zhang. Spacevllm: Endowing multimodal large language model with spatio-temporal video grounding capability. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(12): 9912–9920, 2026. doi: 10.1609/aaai.v40i12.37956. URL https://doi.org/10.1609/aaai.v40i12.37956.

Shijian Wang, Jiarui Jin, Xingjian Wang, et al. Video-thinker: Sparking "thinking with videos" via reinforcement learning. *arXiv preprint arXiv:2510.23473*, 2025a.

Weihan Wang, Zehai He, Wenyi Hong, et al. Lvbench: An extreme long video understanding benchmark. In *ICCV*, 2025b. doi: 10.1109/ICCV51701.2025.02131. URL https://dblp.org/rec/conf/iccv/WangHHCZQDGHXDT25.

Ziyang Wang, Honglu Zhou, Shijie Wang, et al. Active video perception: Iterative evidence seeking for agentic long video understanding. *arXiv preprint arXiv:2512.05774*, 2025c. doi: 10.48550/arxiv.2512.05774. URL https://arxiv.org/pdf/2512.05774.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. In *NeurIPS Datasets and Benchmarks*, 2024. doi: 10.48550/arxiv.2407.15754. URL https://arxiv.org/pdf/2407.15754.

Penghao Wu and Saining Xie. V*: Guided visual search as a core mechanism in multimodal LLMs. In *CVPR*, 2024.

Sheng Xia, Zhengqin Lai, Tianxiang Jiang, Kanghui Tian, Shoujun Zhou, Bin Li, and Yi Wang. Ser: Learning to ground video reasoning with semantic evidence rewards. *arXiv preprint arXiv:2606.24726*, 2026. URL https://arxiv.org/pdf/2606.24726.

Zuhao Yang, Sudong Wang, Kaichen Zhang, et al. Longvt: Incentivizing "thinking with long videos" via native tool calling. In *CVPR*, 2026a. doi: 10.48550/arxiv.2511.20785. URL https://arxiv.org/pdf/2511.20785.

Zuhao Yang, Kaichen Zhang, Sudong Wang, et al. Paravt: Taming the tool prior paradox for parallel tool use in agentic video reinforcement learning. *arXiv preprint arXiv:2605.20342*, 2026b. URL https://arxiv.org/pdf/2605.20342.

Xiangyu Zeng, Zhiqiu Zhang, Yuhan Zhu, et al. Video-o3: Native interleaved clue seeking for long video multi-hop reasoning. *arXiv preprint arXiv:2601.23224*, 2026. doi: 10.48550/arXiv.2601.23224. URL https://dblp.org/rec/journals/corr/abs-2601-23224.

Ce Zhang, Ziyang Wang, Yulu Pan, et al. Searching videos as trees: Self-correcting agents for grounded long video QA. *arXiv preprint arXiv:2607.16189*, 2026a. URL https://arxiv.org/pdf/2607.16189.

Mingfang Zhang, Jingjing Pan, Ashutosh Kumar, et al. Cast-bench: Benchmarking causal chain-grounded spatio-temporal reasoning for video question answering. In *CVPR*, 2026b. URL https://arxiv.org/pdf/2605.23216.

Shaojie Zhang, Jiahui Yang, Jianqin Yin, et al. Q-Frame: Query-aware frame selection and multi-resolution adaptation for Video-LLMs. In *ICCV*, 2025.

Tianshu Zhang, Yan Wang, Ji Qi, and Lijie Wen. Efficient spatio-temporal grounding with multimodal large models via second-level tracking and rl verification. *arXiv preprint arXiv:2606.29023*, 2026c. URL https://arxiv.org/pdf/2606.29023.

Wenhao Zhang, Kuanwei Lin, Xuyi Yang, et al. Eflow: Learning evidence flow for long-video reasoning with adaptive reflection. *arXiv preprint arXiv:2607.00867*, 2026d.

Yilun Zhao, Haowei Zhang, Lujing Xie, et al. Mmvu: Measuring expert-level multi-discipline video understanding. In *CVPR*, pp. 8475–8489, 2025. doi: 10.1109/cvpr52734.2025.00793. URL https://doi.org/10.1109/cvpr52734.2025.00793.

Yiming Zhao, Yu Zeng, Wenxuan Huang, et al. Videoseeker: Incentivizing instance-level video understanding via native agentic tool invocation. *arXiv preprint arXiv:2605.16079*, 2026. URL https://arxiv.org/pdf/2605.16079.

Ziwei Zheng, Michael Yang, Jack Hong, Chenxiao Zhao, Guohai Xu, Le Yang, Chao Shen, and Xing Yu. Deepeyes: Incentivizing "thinking with images" via reinforcement learning. In *ICLR*, 2026. doi: 10.48550/arxiv.2505.14362. URL https://arxiv.org/pdf/2505.14362.

Junjie Zhou, Yan Shu, Bo Zhao, et al. Mlvu: Benchmarking multi-task long video understanding. In *CVPR*, 2025. doi: 10.48550/arxiv.2406.04264. URL https://arxiv.org/pdf/2406.04264.

Kejian Zhu, Zhuoran Jin, Hongbang Yuan, et al. Mmr-v: What's left unsaid? a benchmark for multimodal deep reasoning in videos. In *ICLR*, 2026. doi: 10.48550/arxiv.2506.04141. URL https://arxiv.org/pdf/2506.04141.

A Full Paired-Bootstrap Comparisons

Table 4 gives the paired comparisons the significance statements of the main text rest on, each with its confidence interval, over the $n = 1{,}024$ items graded under every system appearing in it. The later entrants added to Table 1 of the main paper were scored under the same rubric, and their accuracies are reported there.

B Competitor Reward Designs

The two closest systems shape their rewards in opposite directions, and the pair of facts is what motivates our own design. LongVT pays the tool, adding a temporal intersection-over-union term weighted equally with answer correctness, which forces its reinforcement-learning corpus to carry window annotations and therefore caps that corpus at roughly 1.7K items. Video-o3 taxes the tool, multiplying accuracy by a factor below one whenever a tool is called and providing no compensating localisation term at all, which frees it to scale reinforcement learning to a far larger pool whose only requirement is a verifiable answer. ParaVT, which is LongVT's own follow-up, diagnoses the failure mode of the taxing regime as a skip-tool reward shortcut. Our inference is that neither knob addresses the underlying problem: when a question is answerable from the global gist, a tax makes skipping optimal and a bounty buys tool calls without buying accurate ones, because the accuracy point was already available without the tool. Construction gates that select items on which the gate model benefits from magnified target evidence remove that degeneracy at its source, which is why our own reward can spend its localisation weight where the gradient is genuinely missing rather than buying tool calls that were never in doubt.

Comparison	Δ [95% CI]	p
Ours − Qwen3-VL-8B	$+13.6$ [$+10.4$, $+16.9$]	$< 10^{-4}$
Ours − Qwen2.5-VL-7B	$+17.0$ [$+13.7$, $+20.3$]	$< 10^{-4}$
Ours − LongVT-RFT	$+18.8$ [$+15.5$, $+22.0$]	$< 10^{-4}$
Ours − Video-o3	$+20.2$ [$+17.0$, $+23.4$]	$< 10^{-4}$
Qwen3-VL-8B − LongVT-RFT	$+5.2$ [$+2.3$, $+8.1$]	.0008
Qwen3-VL-8B − Video-o3	$+6.6$ [$+4.0$, $+9.4$]	$< 10^{-4}$
Qwen2.5-VL-7B − Video-o3	$+3.2$ [$+0.8$, $+5.7$]	.0136
LongVT-RFT − Video-o3	$+1.5$ [-1.2, $+4.1$]	.298
deficits against the shared backbone, $n = 1{,}041$		
Qwen2.5-VL-7B − Conan	$+0.6$ [-2.0, $+3.2$]	.688
Qwen2.5-VL-7B − LongVT-RFT	$+1.6$ [-1.1, $+4.4$]	.255
Qwen2.5-VL-7B − Video-o3	$+3.1$ [$+0.6$, $+5.6$]	.017

Table 4: Paired bootstrap, $B = 20{,}000$. The upper block runs over the $n = 1{,}024$ items graded under every system it compares, which is why its margins differ slightly from Table 1's, computed over all $1{,}044$; the counter-intuitive part is that at each system's own published configuration both tool-augmented systems fall significantly below a tool-free base model, and the two are not separable from each other. The lower block drops our own rows so that Conan enters the intersection, and therefore carries its own n. It certifies the three deficits of Table 1 individually, and only Video-o3's survives: the Conan and LongVT-RFT deficits are within noise, so the inversion is supported as a property of the *set* of tool-augmented systems, none of which exceeds its backbone, rather than as three separate margins.

Supervised corpora by provenance. Our supervised mixture is $12{,}866$ rows against LongVT's 247.9K and the 418K conventional corpus on which Video-o3 builds. Decomposing by *provenance* changes what those totals count: both competitors are dominated by borrowed material. LongVT's tool-augmented split is $19{,}161$ rows, or 7.7% of its supervised corpus, and of Video-o3's 173K agentic trajectories only the $14{,}893$ self-collected ones are not repurposed from public sets, or 8.6%. The one axis on which the scales meet is reinforcement learning, at our $1{,}367$ items against LongVT's $1{,}667$.

C Visual-Budget-Aligned Re-Test

The main table follows each system's own published configuration, which leaves the two tool-free base models the most generously provisioned systems present: at their libraries' defaults they receive approximately $115{,}000$ overview visual tokens, against Video-o3's declared $16{,}384$, LongVT's implied $24{,}576$, and our measured $8{,}448$. We therefore re-ran both base models under Video-o3's declared budget as a separate configuration, so that the comparison is available under both conventions. This appendix is the only place a capped base-model figure appears; everywhere else in the paper the base models are reported at their uncapped defaults, which is the convention under which base-model numbers are normally quoted. Table 5 gives the outcome.

Two observations follow. Our comparison against the strongest baseline strengthens under alignment, from $+13.2\%$ to $+16.3\%$ on this subset, since it is the base models whose accuracy falls when the overview budget is capped. And the secondary observation that a tool-free base exceeds a tool-augmented system holds under both conventions against Video-o3, at $p = 0.005$ once aligned, while against LongVT it is significant at each system's own configuration and not separable once budgets are equalised. We therefore state that observation in the main text with the qualifier *at each system's own published configuration*, which is the convention under which competitor numbers are normally reported. The alignment is to Video-o3's declared $16{,}384$ tokens, while LongVT's own specification is $24{,}576$, so the budgets here are equalised against one competitor's figure rather than held identical across all four.

Comparison	own settings	aligned 16k
Qwen3-VL-8B	18.2	15.1 (p=.005)
Qwen2.5-VL-7B	14.4	12.6 (p=.089)
Ours − Qwen3-VL-8B	+13.2	**+16.3** ($p<10^{-4}$)
base − Video-o3	+6.8	+3.8 (p=.005)
base − LongVT	+5.4 (p=.0003)	+2.3 (p=.103)
Qwen2.5@16k − LongVT	—	0.0 (p=1.0)

Table 5: Base models re-run under Video-o3's declared overview budget, over the $n = 1{,}038$ items scored in both configurations; the upper block reports accuracies and the lower block differences, both in percent. On this subset our margin over Qwen3-VL-8B is $+13.2\%$ rather than the $+12.9\%$ of Table 1 in the main paper, and it *grows* to $+16.3\%$ under the aligned convention, since the base models lose accuracy when their token budget is capped.

D The Agent Loop and Its Visual Budget

The coarse-periphery, sharp-fovea split is what makes the budget close. The overview pass decodes the video at 2 fps capped at 768 frames with a 160-pixel long side, then keeps every second frame for an effective 1 fps capped at 384 frames. At roughly 22 tokens per frame this costs a measured $8{,}448$ visual tokens. For comparison, LongVT's specification implies $24{,}576$ overview tokens, Video-o3's paper declares $16{,}384$, and the two base models at their libraries' defaults receive approximately $115{,}000$. We spend a smaller overview budget than the allowances LongVT and Video-o3 report and than either base model's library default, and spend the savings where the evidence actually is, on close-ups of an object occupying three tenths of a percent of the frame.

One engineering detail is load-bearing for that budget. The backbone's image processor carries a minimum pixel-area floor of $65{,}536$, which *up-samples* a 160×90 skim frame to roughly 66 tokens and would therefore make the overview cost three times what it should and overflow the context. We ship a shadow processor directory in which that floor is lowered to $4{,}096$, with weights symlinked, and load it identically for training, inference, rollout, and rejection sampling. The floor binds only on small images, so the 768- and 512-pixel tool crops sit above it and are untouched, which is the intended asymmetry. Because the field is written into every saved checkpoint's preprocessor configuration, intermediate checkpoints evaluate under the training protocol automatically.

E Process Reward in Full

The temporal term. Let $\hat{W}$ be the interval the agent committed to on its last call and $W^\star$ the certified one, and let δ be the distance between their centres. Then

$$w = \frac{\tau}{\tau + \delta} \cdot \min\left(1,\ \frac{|W^\star|}{|\hat{W}|}\right),$$

$$\tau = \max(60\,\mathrm{s},\ 0.05 \cdot \mathrm{duration}),$$

a near-miss score times a span-precision factor.

Three functional forms are available, and what decides between them is the median *within-group* standard deviation, since under group-relative optimisation that is the only scale the advantage can see. A one-dimensional intersection-over-union is identically zero for the large majority of rollouts, so a group in which nothing overlaps yields no descent direction. An interval gap saturates instead: any overlap at all scores zero, so a one-second graze and a direct hit are indistinguishable. A centre distance is non-zero everywhere and keeps resolving *inside* the overlap, and on that measure it separates rollouts several times more strongly than the interval-gap form. The gap is one of functional form and not of weighting: raising the coefficient on the interval-gap form leaves it well below what the centre distance reaches unaided.

The span-precision factor takes over the one thing overlap did well, punishing a window widened until it cannot miss, without reintroducing a term that dies at zero overlap: a thirty-second window against a twenty-second certified one scores 0.67, a whole-video window 0.016. Without it, a request

spanning the entire video would place its centre near the certified centre on a good fraction of items and collect most of the near-miss score for nothing. The tolerance τ answers the same concern. A thirty-second value suits evidence windows a few tens of seconds long; on videos of twenty minutes, where misses are minutes rather than seconds, it floors the near-miss score almost everywhere and stops separating *getting warmer* from *hopeless*. Sixty seconds is therefore the floor, with the $0.05 \cdot$ duration term taking over on longer videos, so the tolerance tracks the scale on which misses actually occur.

The rendering pipeline in full. The agent loop caps tool calls at four rounds. The overview is sampled at 2 frames per second to at most 768 frames with a 160-pixel long side, then subsampled every second frame to at most 384. A committed interval is clamped to between 8 and 30 seconds and decoded at 2 frames per second. SAM 2 is seeded with *the policy's own box* at the window frame nearest the box's stated time, that frame being the first at or after $\lfloor (t - s) \cdot 2 \rfloor$ on which the target is visible; when the policy emits boxes for several frames, the seed is taken from the one whose time it states first. Close-ups are cut at 1 frame per second, at least eight and at most twenty-four per call, each a square dilated $2.5\times$ about the propagated box with 46 pixels of padding. Context panoramas are emitted at least once every eight seconds, two to four per call, with the region outlined in red and a 448-pixel long side. Whole-frame returns, used by the *Temporal only* ablation, are sampled at 2 frames per second to at most twelve frames with a 448-pixel long side, under the same observation envelope. Frames on which the tracker reports the target invisible are skipped rather than rendered.

The remaining terms. There is deliberately **no accuracy gate** on the shaping term: gating it on a correct final answer, as one competing recipe does, removes the gradient exactly where committing to the moment is the problem. The **format term** is two distinct components of 0.05, one for a non-empty answer span and one for a reasoning span, capped at 0.10 against an answer term of 1.0; it is deliberately a minority weight, and is non-zero on every rollout observed, so it never becomes the bottleneck. The **judge is binary**, matching the reported metric exactly: a locally served 30B instruct model at temperature zero, allowed to emit a single letter, under wording taken from the benchmark's own construction rubric. A three-level scale awarding half credit is deliberately not used, since it rewards answers earning 0.5 in the loop and 0.0 on the benchmark. Answer extraction takes the *last* non-empty answer span, since the system prompt itself contains a literal empty answer tag that a first-match search would return instead. A judge that times out returns no verdict, is counted, and prints a warning, so that a run degrading toward a pure process reward is visible while it happens rather than inferred afterwards.

Scoring whatever the model chose to do. One reward function serves every rollout, and it reads the call the model actually made rather than anything decided for it in advance. Whichever tool name appears is parsed and the interval it requested is scored; a rollout that answers without calling anything receives the answer and format terms alone, and no process credit is reachable from that path. Parsing only our own tool name would score a rollout that genuinely called the other one as though it had never acted at all, which is indistinguishable from the policy having stopped calling the tool and would corrupt any tool-use rate read off the run.

F Benchmark Statistics, Question Generation and Trajectory Construction

Teacher and student are served at the same pixel floor throughout trajectory construction, so the teacher never describes a sharper image than the student receives.

What the suite is not answerable from. The 1,044 items carry 690 distinct gold answers. The single most frequent covers 33 of them, so a predictor that always emits it scores 3.2%. The twenty most frequent cover 18.8% between them, which is the *coverage* figure of Table 6 and not a score any prior-only predictor can reach, since it would still have to choose among the twenty; the tool-free backbone's 18.2% is therefore about $5.8\times$ an answer-frequency baseline rather than level with a ceiling. A common-answer advantage nonetheless exists, and Table 8 reports it rather than leave it to be recomputed from the release. Separately, every shipped item carries a question-only verdict from the 235B gate model, which saw the question and no video: it is correct on 0 of the 1,044, refusing

	ego	exo	all
Questions	867	177	1044
Videos	387	67	454
Video duration (median, s)	1540	492	1389
Target area (median)	0.333%	0.264%	**0.312%**
Target area (min)	0.0017%	0.0090%	0.0017%
Distinct answers	546	159	690
Top-20 answer coverage	21.7%	21.5%	**18.8%**

Table 6: FoveaBench statistics, recomputed from the shipped manifest. Targets occupy a median 0.31% of the frame of a 23-minute video, which at a 160-pixel long side leaves a median target well under a single visual token in a skim frame. *Top-20 answer coverage* is the share of items whose answer falls in the twenty most frequent answers. It is a dispersion statistic, not a ceiling on an answer-prior predictor: a constant most-frequent-answer baseline is bounded by the frequency of a *single* answer, far below this figure.

What the model is given	ego	exo	**all**	Δ	p
the whole video	17.5	21.5	18.2	—	—
the anchoring interval	21.6	26.6	22.4	$+4.2$	.007
the evidence frame	26.6	25.4	26.4	$+8.2$	$< 10^{-4}$
the target box, cropped	23.3	22.0	23.1	$+4.9$	.004
that box, through our tool	**38.6**	**33.9**	**37.8**	$\mathbf{+19.6}$	$\mathbf{< 10^{-4}}$

Table 7: Figure 2 of the main paper in full, per domain. Qwen3-VL-8B at its official visual budget with no system prompt, all $1{,}044$ items, same three-round rubric as everywhere else; Δ and p are against the whole-video endpoint, paired over items with $5{,}000$ bootstrap resamples. The window condition supplies sixty frames at native resolution and the frame condition a single one, both of which exceed what any evaluated temporal tool returns. The two splits agree on the endpoints, the whole video worst and our tool best, but not in between: the window is the weakest non-tool condition on the egocentric split and the strongest on the exocentric one, consistent with its fixed camera making a single instant less decisive.

on 982 and answering wrongly on 62. That figure is produced by the funnel, which discarded the candidates this reader answered correctly, so it bounds literal leakage rather than measuring a prior fitted to the shipped suite.

	gold in top-20	gold outside	gap
Qwen2.5-VL-7B	18.9	13.3	$+5.6$
Conan	16.8	13.1	$+3.7$
LongVT-RFT	13.3	12.6	$+0.7$
Video-o3	16.8	10.0	$+6.8$
Qwen3-VL-8B	33.2	14.7	$+18.4$
FoveaSeek-SFT	31.4	24.4	$+7.0$
FoveaSeek-SFT+RL	44.1	28.3	$+15.8$

Table 8: Accuracy in percent, split by whether the reference answer is among the twenty most frequent in the suite. Every system does better on common answers; among the stronger rows the smallest dependence is FoveaSeek-SFT's. A question-only baseline, which would separate this from what the visual evidence supplies, is not run here and is listed as a limitation. The gap is computed before rounding, so it need not equal the difference of the two rounded columns.

Attribute mix and target size. Of the $1{,}044$ queried targets, 939 are asked about by colour, 12 by material, and 93 carry no attribute label and are asked about by identity. The suite is dominated by colour identification, inherited rather than chosen: the attribute vocabularies of its sources annotate colour far more densely than any other property. Accuracy does not fall as targets shrink (Table 9).

target area at the annotated instant	$< 0.1\%$	0.1–0.3%	0.3–1%	$\geq 1\%$
items	240	277	319	208
Qwen3-VL-8B	20.4	22.7	14.7	14.9
FoveaSeek-SFT	24.6	22.0	30.1	22.6
FoveaSeek-SFT+RL	31.2	31.8	30.7	30.8

Table 9: Accuracy in percent by target size. Every item in a bin enters that bin's denominator and an item carrying no graded verdict is counted wrong, the convention Table 1 uses, so each row here aggregates to its headline figure. The profile is flat, which is most plausibly what gate selection produces rather than a counterexample to the premise: an item is admitted only when a whole-frame view fails and a magnified view succeeds, and that test is applied per item at its own scale. Selection on one model's outcomes does not force another model's accuracy to be flat, so this is an explanation of the flatness rather than a guarantee of it. The gradient itself is what the whole-frame gate measures, not what the shipped set retains.

The shipped interval. Each item's interval is the *anchoring* interval the generator emitted to mark the action the question is posed against. It is not a certified evidence interval: the instant at which the target's box was annotated falls inside it for 56.9% of egocentric items and 34.5% of exocentric ones. The exocentric figure is the less meaningful of the two, since those targets are tracks sampled at one representative instant rather than single annotated frames. The certified quantities are therefore the target box and that instant, which is why the oracle row above is labelled by what it is, and why agreement with the interval is not used anywhere as a measure of localisation quality.

Question form and generator constraints. A representative shipped item reads: *"While the camera-wearer opens a bag on the floor and removes clothes from it, what is the identity of the small part of the ladder that has a light green and light orange colour?"* with reference answer `rail`. Two alternative forms are excluded by construction: pure state-change questions degenerate into vacuous outputs on sources carrying no interaction annotations for parts, and pure appearance questions are answerable from a single frame and therefore require no temporal search at all. The generator receives every annotation the candidate entity actually carries and selects among them, under constraints that are structural rather than semantic: the target must be tagged `tiny` or `small`, its parent object must appear in the narrated actions, whatever is asked must be grounded in a real annotation rather than invented, no timestamp may appear in the question or answer, and questions are capped at 50 words with answers at 25. All language-model stages of the funnel call one self-hosted Qwen3-VL-235B-A22B-Instruct-FP8 gate model. Both gates are prompted with the same coarse scene-and-action description, which is deliberately not the dense caption the generator worked from: it situates the clip without naming object-level detail, so it cannot carry the queried attribute. The two gates therefore differ only in what pixels accompany that context: the whole-frame gate adds frames spanning the evidence moment at a 336-pixel long side, the magnified-crop gate adds one full frame plus a native-resolution crop of the target box and a sentence locating that crop. An item survives the first gate only when context and whole frames together are insufficient, so anything the context does give away is removed there rather than shipped. The **text filter** applies three text-level rules before the blind filter: the evidence must be unique within the video, the question must require the image, and the target must be grounded in the annotations. Every stage runs the gate model at temperature zero. The funnel's five stages and their settings are: generation over 357 videos yielding $1{,}096$ qualifying segments; the text filter over $2{,}858$ candidates in 569 segments; the blind filter over $1{,}983$ candidates, of which 137 are discarded as answerable without the video; the **whole-frame gate** over the surviving $1{,}846$, sampling the span at 1.0 frame per second to at most eight frames with a 336-pixel long side; and the **magnified-crop gate** over $1{,}285$, pairing one full frame at a 768-pixel long side with a square crop dilated $2.5\times$ around the target box plus 46 pixels of padding, resampled to between 512 and 896 pixels.

Trajectory synthesis. Trajectory synthesis proceeds in three stages. Training questions are drawn from the disjoint training video pool and the student's exact information state is rendered first, the same skim frames, the same locate view, the same panoramas the deployed agent will see. SAM 2 is then seeded with the ground-truth box at the annotated frame and propagated bidirectionally, rendering exactly what the tool returns at inference time and emitting the multi-frame grounding target

at the locate instants. Only then does the teacher write prose: a first reasoning span composed from the real skim, with the annotated evidence moment supplied as a target to justify rather than labelled as ground truth, and a second span covering evidence, integration and self-reflection composed from the actual tool result, after which a validity check discards logically inconsistent assemblies.

Three prompt-level constraints govern the recipe. Every tool's observation envelope must be *textually distinct*: when two tools return observations opening with the same words, the model follows the majority continuation in the training data and skips the grounding turn entirely, so the tracker never fires. The grounding instruction appended to the locate view is load-bearing: without it the model answers immediately after the locate view, whereas a teacher-forced prefix probe confirms it grounds correctly whenever the instruction is present, so what the instruction supplies is the cue rather than the capability. And the agent declares a single tool: with two, selection follows the shape of the prompt rather than the question, which is leakage from the data source rather than a decision worth training, and no benchmark item requires a second.

G SUPERVISED MIXTURE

Table 10 gives the composition of the supervised mixture, recounted from the shipped data file. The mixture is built around one invariant: the model is never shown a demonstration in which a tool is available and then ignored. Rows that present a tool carry it out, rows where answering directly is correct present no tool to skip, and rows destined for general benchmarks carry no system block whatsoever, which makes the model's public-benchmark condition byte-identical to the base model's own. What the model does at inference is its own decision; what the mixture controls is only which decisions it has ever seen made. The reported corpus contains 12,866 rows, of which 5,680 are general chain-of-thought material serving anti-degradation.

Harness	rows	%	tool declared
`general`	5680	44.1	none
`crop_rft`	1913	14.9	temporal
`fovea`	1771	13.8	**coupled**
`crop_longvt`	1502	11.7	temporal
localisation skill	2000	15.5	none
Total	12866	100	4 system blocks

Table 10: Verified supervised mixture, grouped by the system block each row presents. Every row that presents a tool goes on to use it, so no demonstration of declaring-then-skipping exists; the `general` rows carry no system block at all, so the model's condition on public benchmarks is byte-identical to the base model's. Training on the two competing temporal dialects alongside our own is deliberate, since it keeps the surface syntax of a call from being what distinguishes them. Localisation skill data comes from a repack of the LLaVA-ST spatio-temporal grounding corpus (cited in the main paper) teaching the grounding turn's output format.

One further convention applies to every reported checkpoint: the user turn is phrased consistently, so that every in-domain user turn ends with the same sentence emitted by one shared function. Nothing dispatches on that sentence, and it exists only so that the training and evaluation conditions match.

H SETTING THE SHAPING COEFFICIENT

The quantity that decides a shaping term's influence is not its mean but its share of the *within-group* standard deviation, because group-relative optimisation normalises within a group, so a component that is near-constant across one contributes nothing to the advantage however large its mean. Concretely, for a group G of rollouts sharing a prompt, with total reward $R_i = a_i + f_i + \beta w_i$ as in Appendix E, the reported share is

$$s(G) \;=\; \frac{\beta\,\sigma_G(w)}{\sigma_G(R)}, \qquad \sigma_G(x) \;=\; \sqrt{\tfrac{1}{|G|}\sum_{i\in G}\left(x_i-\bar{x}_G\right)^2},$$

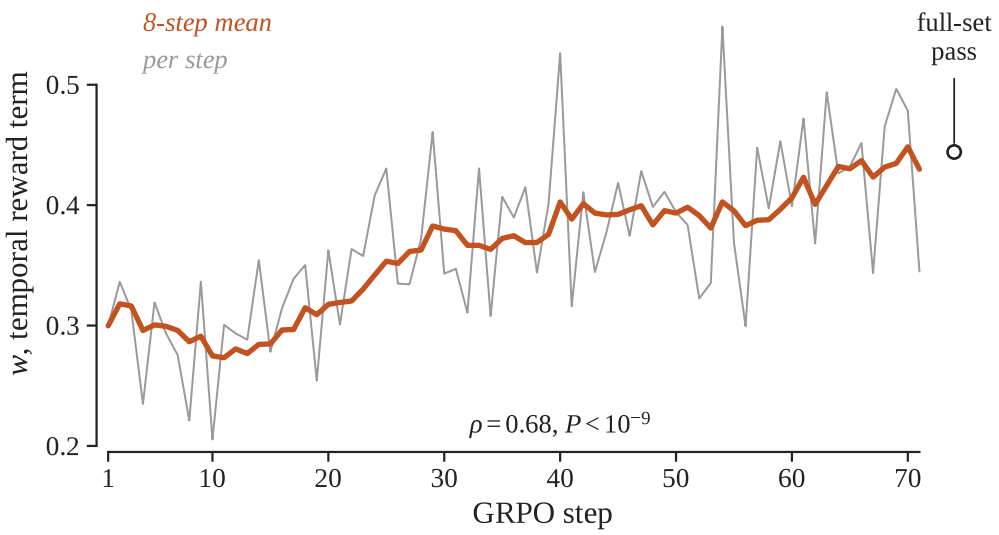


Figure 4: The temporal term w over the 71 optimiser steps of the reported reinforcement run, at $\beta = 1.5$ and with no spatial term in the reward. Grey is the per-step batch mean and the heavy line an 8-step running mean; the open marker is a single pass over all 1,367 reinforcement items after the run, which reaches $w = 0.444$ against 0.300 at the first step, with the median distance from the certified centre falling from 266 to 149 s. Spearman ρ is over step against per-step mean. What this establishes is that the shaping term is *live* rather than decorative: the policy moves in the direction the term rewards.

averaged over the groups of one logging interval. It is a ratio of standard deviations and *not* an additive decomposition of the reward's variance: a, f and w covary, so the three terms' shares need not sum to one, and s measures how much of the spread the optimiser sees is attributable to the shaping term rather than how much gradient it supplies. The loss curve does not expose that share: a term contributing a negligible fraction of the gradient is indistinguishable there from one carrying the run. Every reward process therefore reports, at a fixed rollout interval, the answer and process means, the window hit rate, the median centre distance, and above all the process term's share of the within-group standard deviation, with a target band of 30–45%: below it the term is decorative, above it the answer stops leading.

The coefficient $\beta = 1.5$ is chosen against that quantity rather than against the reward's magnitude, and leaves the temporal term accounting for roughly two fifths of the measured within-group reward variation, the answer carrying the rest. That puts the two terms on the same order as the equally weighted accuracy and grounding terms of the one competing recipe that pays for localisation at all, while reaching it through a term that is non-zero everywhere rather than one that vanishes on the rollouts where the gradient is most needed.

Figure 4 tracks the term's mean over the reported run. It moves, monotonically enough over 71 steps to give $\rho = 0.68$ at $P < 10^{-9}$, so the shaping term is live rather than decorative. The within-group share that defines the band is logged at the same interval rather than plotted here. Because w is itself part of what the optimiser maximises, what the curve measures is the term's purchase on the policy; the accuracy of Table 1 of the main paper is scored independently, by the benchmark's own judge.

I Human Validation

Because the construction funnel is automatic, we validated a stratified sample of candidate items by hand. The sample is 100 items drawn under a fixed random seed, stratified by domain in proportion

to the split sizes, giving 75 egocentric and 25 exocentric items. One reviewer inspected each item and answered two questions. First, shown the annotated evidence moment and the target box, is the shipped reference answer correct? Second, shown the video at a 224-pixel long side, is magnifying the target *necessary* to determine the answer, or is it already legible at that scale? The second question is the one that matters for the benchmark's premise, and posing it at a fixed low resolution makes it a judgement about legibility rather than about the reviewer's own acuity.

Criterion	ego	exo	all
Reference answer correct	74/75	23/25	97/100
Magnification necessary	71/75	24/25	95/100
Both satisfied	**71/75**	**22/25**	**93/100**

Table 11: Manual validation of 100 stratified candidate items. Items failing either criterion were discarded during construction and are not part of the shipped 1,044.

Table 11 reports the outcome. The reference answer was judged correct for 97 of 100 items, and magnification was judged necessary for 95; both criteria held simultaneously for 93, at $71/75$ on the egocentric split and $22/25$ on the exocentric one. The check is a precision measurement on the funnel: one reviewer, sampling construction candidates, with items failing either criterion removed before the set was shipped. Its residual concentrates in the exocentric split, where three of twenty-five items failed at least one criterion, which tracks how that split's attribute labels are produced: back-filled by a model rather than inherited from human annotation.

Two construction caveats. Two properties of the funnel qualify how strongly the needs-both-axes property holds, and we state them here in full because the main text only summarises them. First, the **whole-frame gate probed a fixed span** centred on the generator's evidence instant rather than the frame at which the target box was annotated. For 16.5% of egocentric items that span excludes the annotated frame, so for those items the gate certifies that the answer is unobtainable from whole frames *near* the evidence moment rather than exactly at it, and the needs-zoom property is established more weakly than for the remaining 83.5%. The exocentric split is unaffected, since its evidence instants and mask annotations come from the same timeline. Second, the **magnified-crop gate is mildly self-selecting**: it discards items the gate model still cannot answer even at native resolution, and those items have systematically larger targets than the ones it keeps, at a median 114 against 86 pixels on the long side. The retained set is therefore biased slightly towards smaller targets, in the direction that makes the suite harder rather than easier, but it does mean the reported median target area is not an unbiased estimate of the candidate pool's. Both are properties of the shipped funnel, and every per-item gate verdict is released so that either can be recomputed under a different probe.

J Scoring Protocol and Leakage Control

Judging runs for two full rounds over every system; the 101 items on which the rounds disagree then receive a third round applied only to those disagreements, which is equivalent to full three-round majority voting while re-judging only the small minority that needed it. That third round splits 55 to 46, close to random, which is what genuine ambiguity looks like. The rubric is binary rather than three-level for two reasons: a three-level rubric awarding partial credit shows 18–26% self-disagreement against the 0.77% of the binary one below, and it pulls policies toward vague answers that earn half credit in the training loop and nothing on the benchmark. Answer extraction takes the contents of the answer span when present and otherwise the remaining prose after stripping reasoning, tool-call and placeholder spans, applied identically to every system, since feeding a full chain of thought to a judge systematically under-scores tool-using competitors. Because verdicts are content-addressed and exclude the question text, deleting items from the manifest is safe and requires no re-inference, whereas *editing* an item would silently reuse a stale verdict; a per-item question fingerprint guards against this by withholding modified items from scoring until they are re-judged.

Comparison	agree %	κ	direction
same judge, round 1 vs 2	99.2	**.9812**	symmetric, p=.73
two judges, three rounds	97.4	**.9351**	25:2, $p<10^{-4}$
two judges, one round	97.6	.9397	—

Table 12: Judge agreement on identical predictions. The binary rubric is near-deterministic within a judge and still strong across families; what cross-family disagreement remains has a consistent direction, so absolute scores are mildly judge-dependent. Every comparison we report is scored under one common protocol, which removes the judge as a source of difference between rows; it does not by itself exclude a judge scoring some systems differently from others.

How much the judge matters. Every self-measured FoveaBench accuracy in this paper is produced by the same model-judge protocol, so we measure what that choice costs. Table 12 reports the comparisons, all on the same predictions. Within a single judge, two independent rounds agree on 99.2% of items at $\kappa = 0.9812$, and their disagreements are symmetric under a McNemar test at $p = 0.73$, so round-to-round variation is noise rather than drift. Across judges from different families agreement is 97.4% at $\kappa = 0.9351$, which is high but not neutral: the disagreements run 25 to 2 in one direction at $p < 10^{-4}$, so one judge is systematically the more permissive of the two by a small margin.

Two things follow. Absolute accuracies shift slightly with the choice of judge, so a number here should not be set against one scored elsewhere by a different judge; every system in every table of this paper is scored by the same judge under the same rubric, so every within-paper comparison is made under one protocol. And the cross-family gap is not an artefact of the three-round protocol: repeating the comparison with a single round on each side moves κ by under 0.005, rather than collapsing as it would if majority voting were doing the work.

Benchmark and training videos are disjoint at packet-build time rather than by post-hoc filtering, with benchmark items drawn from PACO *test* videos and training items from PACO *train* videos; PACO's validation split is discarded entirely because 258 of its 270 videos overlap the training split. The assembler checks every self-built video identifier against the evaluation manifest and reports zero leaks, and question identifiers consumed by supervised trajectories are removed from the reinforcement-learning pool.

K Public-Benchmark Configurations

Table 13 extends Table 3 of the main paper with four further single-turn rows, and restores the FoveaBench block alongside them so that every system measured on the suite appears in one place. The overview allowances behind it are not close: Qwen evaluates at 640–768 frames per video, Video-o3 declares $16{,}384$ visual tokens, ours is capped at 384 frames at a 160-pixel long side for $8{,}448$. We have not controlled that variable, so the comparison is context rather than a measured cost of specialisation. Those figures are overview only; our tool calls add close-ups, so end-to-end cost is matched in neither direction. What the suites do show is that this allocation is not uniformly advantageous outside FoveaBench, where it funds the close-ups behind a 12.9% margin; the cause is not isolated. The coupled tool is the right instrument when the evidence is small, not a free upgrade when it is not.

Video-MME is run *without* subtitles, which is the protocol of the comparison table the competitor numbers are quoted from, so that column is like-for-like across every row. MMVU is scored with its official language-model judge for both multiple-choice and open-ended items. The remaining suites use their own released extractors. The generation budget is a maximum of $4{,}096$ new tokens throughout.

L Prompts for Every Stage

All text below is reproduced verbatim from the single module that every producer and consumer imports, so the strings shown are byte-identical to those used in supervised-data construction,

Method	Base	VideoMME w/o subs	MLVU M-Avg	LVBench Avg	LVideoB Avg	VMMMU Overall	MMVU M-Avg	MMR-V Overall	**FoveaBench** ego	exo	**all**
Single-turn video models, no tool											
Qwen2.5-VL (Bai et al., 2025b)	7B	65.1	70.2	45.3	56.0	47.4	61.3	32.4	14.5	13.6	14.4
Video-R1 (Feng et al., 2025)	7B	61.4	—	—	—	52.4	63.8	36.3	11.0	23.7	13.1
VideoChat-R1 (Li et al., 2025b)	7B	63.6	68.7	—	58.2	—	—	36.1	10.4	22.6	12.5
Video-Thinker (Wang et al., 2025a)	7B	—	—	37.0	—	—	—	—	8.1	19.8	10.1
Tool-augmented, temporal resampling only											
Conan (Ouyang et al., 2025)	7B	60.5	63.4	39.2	56.6	—	—	42.7	13.1	16.9	13.8
LongVT-RFT (Yang et al., 2026a)	7B	64.3	—	41.3	—	45.4	—	—	12.1	15.8	12.7
Video-o3 (Zeng et al., 2026)	7B	**66.5**	**72.1**	**47.6**	**60.5**	51.7	**67.2**	**44.7**	10.7	14.1	11.3
FoveaSeek-SFT	8B	62.1	65.6	41.4	51.1	51.7	49.4	35.3	24.9	26.6	25.2
FoveaSeek-SFT+RL	8B	64.8	64.3	46.4	54.3	**53.4**	53.5	38.6	**31.3**	**30.5**	**31.1**

Table 13: General long-video ability across seven public suites and FoveaBench, in percent; the public columns abbreviate Video-MME, MLVU, LVBench, LongVideoBench, Video-MMMU, MMVU and MMR-V, and dashes mark results the cited paper does not report. Competitor numbers are quoted from the comparison table of Video-o3 (Zeng et al., 2026) rather than from each system's own paper, so they are at least internally consistent with one another; the two FoveaSeek rows are measured by us with each benchmark's own official extractor. The public columns are *not* budget-matched: each row carries its own paper's frame allowance and prompt. Qwen3-VL-8B, the backbone our two rows are built on, is not measured on them, so what these columns bound is the specialisation cost against the quoted rows rather than against the backbone itself. The **FoveaBench** block is measured by us, as no published number exists for it: every system runs its own repository's code path at its own published configuration with only judge and rubric held identical, so the tool-free base model gets its library's uncapped default and is the most generously provisioned row here. It carries the three single-turn systems absent from Table 1, which is why the widest margin quoted in the Analysis section is read off this table.

reinforcement-learning rollout, and evaluation. Line breaks inside a block are ours, for typesetting only.

Agent system prompt. One tool is declared and no direct-answer licence appears anywhere in it.

```
You are a helpful assistant for long-video
understanding and reasoning.

You may call functions to inspect the video
before answering.

<tools>
{"type":"function","function":{"name":
"fovea_locate","description":"Retrieve
full-view frames sampled from a chosen time
window of the video, to find a small target
region. After you localize the target on the
returned frames, a tracker follows it through
the whole window and returns close-up crops
plus context panoramas.","parameters":{"type":
"object","properties":{"start_time":{"type":
"number","description":"Window start in
seconds"},"end_time":{"type":"number",
"description":"Window end in seconds, must be
> start_time"},"target":{"type":"string",
"description":"Short referring description of
the object to find, with the action context
that pins it down"}},"required":["start_time",
"end_time","target"]}}}
</tools>

Protocol: think inside <think></think>. Call
the tool with <tool_call>{"name":
"fovea_locate","arguments":{...}}</tool_call>.
When frames are returned, output the grounding
of the target for each frame as a fenced JSON
list [{"time": T, "bbox_2d": [x1,y1,x2,y2],
"label": "..."}] with coordinates normalized
to 0-1000, skipping frames where the target is
not visible. After the tracker returns
close-ups, give the final short answer inside
<answer></answer>.
```

User turn. Every in-domain user turn ends with one shared sentence emitted by a single function. The suffix is not a switch, since the decision to call remains the model's, but it is a cue the model conditions on, so an evaluation harness must reproduce it for the deployed condition to match the trained one.

```
<skim frames>
Question: {question}
Think first, call the fovea tool to zoom into
the clue segment, then answer.
```

Grounding instruction, appended to the locate view. Without this block the fine-tuned model answers immediately after the locate view instead of grounding, while a teacher-forced probe confirms correct grounding whenever it is present, so what the block supplies is the cue rather than the capability.

```
Given the query "{target}", for each frame,
detect and localize the visual content
described by the given textual query in JSON
format. If the visual content does not exist
in a frame, skip that frame. Use exactly the
timestamp shown before each frame as the
"time" value. Output Format: [{"time": {t0},
"bbox_2d": [x_min, y_min, x_max, y_max],
"label": ""}, ...].
```

Observation envelopes. The two turns of a single call must be textually distinct: when both open with the same words, the model learns to skip the grounding turn entirely.

```
# skim, first user turn
Frames sampled evenly across one long
egocentric video:

# turn 1 of the call: the locate view
Frames sampled from [{a}s - {b}s]:

# turn 2: what the tracker returned
Tracked the target through [{a}s - {b}s] at
native resolution. Context views (target
marked with a red box):
Close-up crops of the target:

# turn 2 when propagation fails
Tracker could not follow the target from that
grounding. Answer from what you have, or try
a different window.
```

Judge prompt. The same binary rubric scores the benchmark, the ablations, and the reinforcement-learning reward, so the metric and the training signal cannot disagree about what counts as correct. Benchmark scoring batches items and content-addresses verdicts, in the form below. The reward path serves the identical rubric from a locally hosted 30B instruct model at temperature zero, restricted to a one-letter reply, because it has to return a verdict inside the rollout loop.

```
For each numbered item, decide whether the
CANDIDATE answer means the same as the
REFERENCE answer for that question. Accept
synonyms and wordings that are correctly more
or less specific than the reference. A
candidate that refuses, is empty, or says it
cannot tell is NOT correct.
Return STRICT JSON only, exactly one verdict
per item, nothing else:
{"verdicts": [{"id": <n>, "correct": true or
false}]}

Items:
{items}
```

Whole-frame gate prompt (funnel stage 4). The gate sees the sampled frames and this text. Abstention is explicitly forbidden, so the verdict measures legibility rather than willingness to answer.

```
These frames are consecutive samples from one short moment of a video.
Question: {q}

Answer with a short phrase only, no explanation. Commit to your single
best guess even if you are unsure; never say you cannot tell.
```

Magnified-crop gate prompt (funnel stage 5). The gate sees the full frame, the native-resolution crop, and this text; `loc`, `cx`, `cy` and `span` are filled per item.

```
The first image is the full video frame, for context. The second image
is a zoomed-in crop of that same frame, enlarged around one small
object. In the full frame that object sits near the {loc} -- about
{cx:.0f}% from the left and {cy:.0f}% from the top, and the crop spans
roughly {span:.0f}% of the frame width. Use the full frame to place the
object and the crop to read its fine detail.
Question: {q}

Answer with a short phrase only, no explanation. Commit to your single
best guess even if you are unsure; never say you cannot tell.
```

Blind (question-only) filter prompt (funnel stage 3). No video is supplied. Any candidate answered correctly here is discarded.

```
Answer this question about a video you have NOT seen. Use only prior
knowledge and common sense. If you genuinely cannot know without seeing
the video, reply exactly: UNKNOWN.

Question: {q}

Answer with a short phrase only.
```

Text filter prompt (funnel stage 2). Applied before the blind filter; it enforces the three text-level rules.

```
You are filtering candidate questions for a long-video benchmark.

You are given, as text, the SAME description of the video that the
question-writer saw:
  * OUTLINE - one line per chapter of the whole video, in order.
    * SEGMENT - one chapter in detail: the narrated actions, hand-object
  events and the
        annotated entities inside it. Each entity carries its annotated
    appearance.
Then a numbered list of candidate QUESTIONS, each with its answer and
the object it is
about.

Keep a question ONLY if it breaks NONE of the three rules below;
otherwise drop it.

RULE 1 - UNIQUE EVIDENCE. To answer, a viewer must look at one specific
moment. That
moment must be UNIQUE in the whole video. Scan the OUTLINE: if the same
kind of
activity, object or place appears in more than one chapter, and the
question does not
say what makes THIS occurrence different, a viewer could not know which
moment to look
at -> break.

RULE 2 - NEEDS THE IMAGE. The answer must require actually seeing that
moment. If the
answer can be produced from the question text plus everyday common sense
-- a default
or overwhelmingly likely value (a knife blade is metal, a sink is white)
-- then seeing
the video is not needed -> break.

RULE 3 - GROUNDED. The answer's content must be supported by the SEGMENT
description
above: the annotated appearance of the named object, or a narrated fact.
If the answer
asserts something the description does not contain -- invented, guessed,
a wrong
attribute type (e.g. giving a transparency when asked for material), or
"not specified"
```

```
-> break.

Judge EVERY question. Return STRICT JSON, nothing else:
{"verdicts": [
  {"id": <question number>, "keep": true or false,
   "violated": [<rule numbers broken; empty list if kept>],
   "reason": "<one short clause>"}
]}
```

Question generator prompt (funnel stage 1). The system message is identical for every generation call.

```
You write questions for a long-video benchmark.

Every question COMBINES an ACTION and an APPEARANCE, so that answering
it forces a
viewer to do all three of these:
  (1) SKIM the long video and find the right STRETCH of it,
    (2) WATCH what the camera-wearer does there, to tell WHICH small
  object is meant --
           the object is pinned down by the ACTION performed with it,
      never by naming its
           look, so a single glance at a frame is not enough to know
      which object to read,
  (3) ZOOM in on that small object and read a FINE VISUAL detail of it.

You are shown, as text:
    * OUTLINE - one line per chapter of the whole video, in order. Use
  it only to judge
    whether a moment is UNIQUE.
    * SEGMENT - one chapter in detail: the narrated actions, hand-object
  events and the
       annotated entities inside it, all in time order. Each entity is
    marked ENTITY <id>,
       tagged with its size [tiny]/[small]/[medium]/[large], its name
    (its identity), and any
       annotated appearance it has (colour, material, transparency).
    Objects with no annotated
       appearance still have a name -- ask about their identity. The
    narrated actions tell you
       what is DONE with each object. Write about THIS segment; the
    outline is context.

For each question produce:

    anchor_interval  a SCENE-LEVEL span, at least 20 and typically
  30-120 seconds, inside
                                    this segment, that the opening
                  clause describes: a recognisable stretch
                                    findable by skimming. It must be
                  UNIQUE: check the OUTLINE, and if a
                                    similar stretch happens elsewhere,
                  add what separates this one. Do NOT
                  make it short.

    evidence_time    the timestamp of the small object the question asks
  about, inside
                  anchor_interval.

Build each question like this:
- Open with the scene-level anchor clause.
- Pin down the small target by the ACTION tied to it -- what the
camera-wearer does with
    it, or with the object it is part of (picks it up, uses it, opens
  it, sets it down...).
    This is how the reader knows WHICH object, and it makes them watch
  rather than glance.
    Do NOT identify the object by naming what it is, its colour or its
  material.
- Then ask about the object, drawing on what the data actually supports:
       * its IDENTITY -- what the object is (always fair: you know its
    name);
       * one or more of its annotated visual attributes -- colour,
    transparency, material.
    BOTH kinds are welcome: a SINGLE-attribute question (e.g. just the
  colour) is just as
    good as a COMBINED one (e.g. colour+transparency, identity+colour,
  identity+material).
    Prefer to write a MIX -- some single, some combined -- and you may
```

```
  ask more than one
    question about the same object using different attributes. Every
  part you ask must be
  settleable only by looking closely at full resolution.

Three example shapes (do not reuse the content):
    "While the camera-wearer picks up the spray bottle from the cabinet
  and sprays the boots,
   what colour is its small cap?"  -> single (colour).
    "While the camera-wearer sweeps the floor and then picks it up from
  the table to trim the
      plant, what is that small tool and what colour are its
   finger-holes?"  -> identity+colour.
    "As the camera-wearer lifts a bottle from the cabinet and pours from
  it, what colour is its
      small cap and is the cap see-through or solid?"  ->
   colour+transparency.
The action says which object; the identity/attributes are the fine
details you must zoom for.

Rules:
- The target MUST be [tiny] or [small]; never a [medium] or [large]
object. Prefer a target
    whose object appears in the narrated actions, so the action anchor
  is strong -- but if no
    narrated action names it, you may still anchor the question on what
  the camera-wearer is
    doing at that time, near that object. A slightly weaker action
  anchor is fine; a missing
  question is not.
- target_entity MUST be one of the ENTITY <id>s listed in this SEGMENT.
- Only ask what the data supports. IDENTITY is always fair. Colour /
transparency / material
    may be asked ONLY if that attribute is listed in the object's
  annotated appearance. If an
    object shows NO annotated appearance, ask about its IDENTITY (not
  its colour/material).
    Never state a value that is not in the data; do not guess. Do NOT
  ask about pattern.
    Material alone is often guessable -- if you use it, combine it with
  the identity or colour.
- The [tiny]/[small] size tag and the word "annotated" are internal
labels for you, not
    things a viewer can see. Never ask about an object's size, and never
  use the word
    "annotated" or the size tag, in the question. Ask only about what is
  visible on screen.
- Make clear WHICH object you mean; other objects are present and a
reader must not be able
  to answer about a different one.
- The answer states the queried identity and/or attributes as a short
natural phrase.
- Do NOT put the answer's value, a timestamp, a clock time, a frame
number, or "first/second
    half" in the QUESTION. Put no timestamps or frame numbers in the
  ANSWER either.

Produce between {MIN_QA} and {MAX_QA} questions for this segment. Aim
for the HIGH end: every
[tiny]/[small] entity listed is usable, and one entity can support
several questions asking
different things (its identity, its colour, its material, a
combination). Work down the entity
list rather than stopping after the first one or two.

Returning an empty list must be RARE. If this segment lists even one
[tiny] or [small] entity
with a name or an annotated attribute, write questions about it. Only
return an empty list if
the segment genuinely lists no usable small object at all. A merely
ordinary question is still
worth writing -- later stages do the filtering, not you.

Return STRICT JSON, nothing else:
{{"questions": [
    {{"question": "...", "answer": "...", "target_entity": "<id from
  SEGMENT>",
   "anchor_interval": [start, end], "evidence_time": t,
      "anchor_description": "the stretch, in one line, as a skimmer
   would recognise it",
   "why_unique": "why no other stretch of this video matches"}}
```

```
]}}
```

M TRAINING HYPERPARAMETERS

Supervised fine-tuning loads the backbone through the shadow processor of Appendix D and adapts it with low-rank adaptation at rank 64, scaling factor 128 and dropout 0.05, applied to all linear projections, with both the vision tower and the multimodal projector frozen so that only the language side is trained. The learning rate is 10^{-4} and the run covers one epoch over $12{,}866$ rows, which is 161 optimiser steps, checkpointing every 32. Adapting rather than fully fine-tuning is a design choice rather than an economy: with the adapter disabled the weights are byte-identical to the base model, so whatever the adaptation costs in general video ability is recoverable rather than baked in. With the adapter enabled it does cost something, and Table 13 of the main paper reports it.

Reinforcement learning runs on a fork of verl (Sheng et al., 2025) $0.6.1$ (commit `d62da495`, 2025-11-14), entered at `verl.trainer.main_ppo`. Each step draws 32 prompts with 8 rollouts each for 256 retained sequences, with rejection sampling enabled, a learning rate of 10^{-6}, no KL loss term, clip ratios of 0.2 and 0.3, and prompt and response limits of $18{,}432$ and $8{,}192$ tokens. The rollout engine caps a sequence at six generation rounds, which is its own turn budget and is separate from the agent loop's cap of four tool calls. Because rejection sampling discards groups whose rollouts all agree, the reward function scores more sequences than the 256 retained for each update. Steps take approximately 715 seconds on 32 accelerators. The optimiser premises are inherited from a public recipe; what we add on top of them is the process term of Appendix E.

N REPRODUCIBILITY

What this supplement contains. The supplement ships the item manifest with its boxes and tubes, the construction funnel together with every per-item gate verdict it emitted, and the evaluation harness with the judge prompt it serves and the analysis code behind the reported tables. Source videos are not ours to redistribute under the licences of the datasets they are drawn from, so the manifest identifies each item by its upstream video identifier and interval and is applied to footage obtained from those sources directly.

Inference, training, and rollout share the same functions for the locate view, the grounding instruction, and every observation envelope, imported by the supervised-data builder, the rollout engine, and the evaluation harness alike; the assembler asserts byte-identity between the system prompt in supervised rows and the one the rollout engine loads. The shadow processor with its lowered pixel floor is loaded in all four settings and is written into every saved checkpoint's preprocessor configuration, so intermediate checkpoints evaluate under the training protocol automatically; we verify that field before every evaluation. Skim caches key on the video identifier rather than the question identifier, since public benchmarks ask three to fifteen questions per video and per-question caching would re-decode a ninety-minute file a dozen times; caches are built in a hashed staging directory and renamed into place, because two shards can reach the same unrendered video simultaneously and a half-written cache is indistinguishable from a complete one, and any render missing more than ten percent of its frames is refused rather than published. SAM 2 is served as a separate GPU-side HTTP service during rollout: the rollout coordinator holds no CUDA context, because the inference workers sit behind an external executor. Total compute for the reported artefacts is four hours twenty-nine minutes on 40 accelerators for supervised training, approximately fourteen hours on 32 accelerators for the seventy-one reinforcement-learning steps of the reported run, and approximately forty-five minutes on 40 accelerators per full FoveaBench evaluation of one checkpoint.

O QUALITATIVE EXAMPLES

Success and failure cases. Figure 5 shows four graded episodes of the reported checkpoint, two correct and two incorrect, drawn from the prediction log and scored by the same judge that produces every number in this paper. They were selected to span the localisation regimes, and together they illustrate the two mechanisms the analysis section discusses.

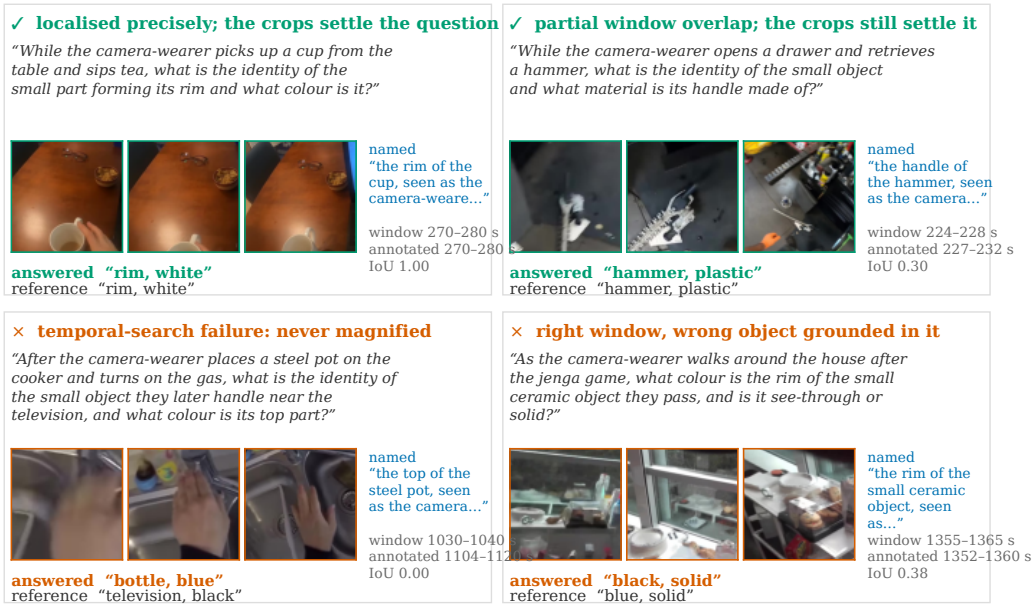


Figure 5: Four graded episodes of the reported FoveaSeek checkpoint on FoveaBench, two correct and two incorrect, taken from the prediction log and scored by the same binary judge used throughout the paper. Each panel shows the object the agent named, the interval it committed to against the certified one, and three of the close-ups its own tube returned. Top left: the committed interval coincides with the certified one and the close-ups settle the question. Top right: the interval misses the certified one entirely, yet the crops still contain the queried part, and the answer is accepted although it is less specific than the reference. Bottom left: a temporal-search failure, where inspecting the wrong moment means the queried object is never magnified and the agent reports a different one. Bottom right: the interval overlaps the certified one, but the agent grounds and tracks the wrong small object inside it, so the tube returns close-ups of a distractor. The two failure modes are distinct, and both are diagnosable only because every item ships a reference target box.

The two successes bracket the localisation regimes. In the first the committed interval coincides with the annotated one, and the close-ups the tube returns settle the question directly. In the second the agent commits to 224–$228\,\mathrm{s}$ against an annotated window of 227–$232\,\mathrm{s}$, so the two overlap only in part; the tube nonetheless carries the named object across that overlap, and the close-ups it returns are enough to read what the handle is made of.

The failures separate cleanly into the two error modes we would expect from a coupled tool, and neither is a refusal to act. In the third episode the agent inspects 1030–$1040\,\mathrm{s}$ where the evidence sits at 1104–$1120\,\mathrm{s}$, so the queried object is never magnified and it describes a different one; this is a temporal-search failure that the spatial axis cannot rescue, and it is the residual that a localisation-weighted reward is meant to attack. In the fourth the interval does overlap the certified one, but the agent grounds and tracks the wrong small object inside it, so the tube delivered close-ups of a distractor and the colour it reports belongs to that distractor. That second mode is the more instructive one: it shows that the remaining headroom is not solely a matter of finding the moment, and that grounding precision inside a correct window is an independent axis of failure. Both modes are diagnosable only because every item ships a reference target box.