

How Should a Camera-First AI Agent Work?

Xiaofan Wu, Xi Zeng, Miaoxia Chen, Peishan Chen, Shuyan Li,
Chen Wang, Jiyun Yao, Hanyong Zhong, Jiahao Zhu
Chance AI
xiaofan@chance.vision (corresponding author)

Abstract

Camera-first agents must act on the visual world as it unfolds. We argue that the central obstacle for future vision-interaction systems is not recognition ability—multimodal models (vision-language and live/streaming alike) generalize well through language and remain the right tool for semantic reasoning—but **time**. Many visual targets are *ephemeral*: they appear, are query-relevant, and leave within a short window. Binding memory to the same end-to-end path as reasoning means that by the time a monolithic pipeline completes, the evidence is gone. We formalize this as **Ephemeral-Target Interaction** and a **Perceptual Deadline Contract** separating a hard *memory* deadline from a soft *reasoning* deadline, grounded in the bounded, passive nature of sensory and working memory. We propose a **two-deadline architecture**: an always-on, bounded, non-semantic fast path; an **active remembering** mechanism that commits a moment to durable memory before it is overwritten; and a **deferred deliberation** layer (multimodal model + tools). We instantiate the contract on real MOT17 track windows—a consistency check—and calibrate the single-path baseline from **measured** frontier-model latency: even a short answer takes 1.2–9.3 s, and an agentic identify-and-search longer. **Answers can be late; evidence cannot be missed.**

Keywords: camera-first agents; ephemeral targets; memory deadline; deferred deliberation; perceptual deadline; multimodal models (VLM / live)

1 Introduction

Vision-language models have become remarkably capable because language provides a generalizing interface. For semantic recognition, retrieval, and reasoning, a strong VLM is currently the best tool we have, and we expect it to remain the core of the slow, deliberative layer of future vision systems. The difficulty of building a *camera-first agent* lies elsewhere: **in time**.

Consider a camera-first agent interacting with the world in real time. Many of the targets a user cares about are **ephemeral**—a person passing by, a sign glanced at, an object briefly held. Such a target is present only within a short window. If perception and reasoning are bound into a single end-to-end path (see a frame, run a large model, identify, call tools, answer), then by the time the pipeline completes, the target has left and the evidence is gone. The user’s tolerance is asymmetric: a late answer is tolerable; a missed moment is not. We summarize this as: *answers can be late, evidence cannot be missed*.

This reframes what the next-generation intelligent terminal is. We take a **camera-first terminal** to be a device with **always-on but bounded perception**: it receives the visual stream passively and cheaply, while only **actively selected** moments are persisted. Its defining competence is not answering questions about an image but **not missing the moment**—a perception-first counterpart to today’s prompt-first, stateless model-plus-screen devices. This is no longer hypothetical: VisionClaw [Liu et al., 2026] already runs such an agent on Meta Ray-Ban glasses, yet samples perception at ~ 1 Hz and reports a median ~ 12 s action latency—exactly the temporal gap we address.

This asymmetry suggests that the fundamental design question for camera-first agents is not “how do we recognize better?” but “**what should the system do in the moment?**” We argue for a **Perceptual Deadline Contract** that separates two deadlines with different slack: a **hard memory deadline** (lay down the evidence within the target’s window τ) and a **soft reasoning deadline** (the semantic answer may arrive later, computed on frozen evidence).

We instantiate this as a **two-deadline architecture**: an always-on, non-semantic fast path that passively writes a bounded working memory; an **active remembering** mechanism that remembers a moment into durable memory before it is overwritten (event-driven in v1); and a **deferred deliberation** layer (a multimodal model with tools) that reads from memory. The multimodal model remains the slow, deliberative layer: we change *where* and *when* it runs, not whether it is used. The design is motivated by the bounded, automatic nature of sensory and working memory [Sperling, 1960, Atkinson and Shiffrin, 1968, Baddeley and Hitch, 1974, Cowan, 2001] and by reactive-deliberative control [Brooks, 1986]. Perception is itself inference — the visual system performs fast, automatic, non-linguistic inference [Rao and Ballard, 1999, Kersten et al., 2004] — so the contract orders *two kinds of inference by deadline*, not reasoning from non-reasoning.

Contributions. (1) **Problem:** we formalize Ephemeral-Target Interaction and the Perceptual Deadline Contract (hard memory vs. soft reasoning), grounded in bounded sensory/working memory and budget-conditioned remembering. (2) **Formulation:** a deadline curve—memory recall versus dwell τ —and an initial instantiation on real track windows. (3) **System:** a two-deadline architecture (bounded passive store + event-driven active remembering + deferred deliberation), reported as memory recall vs. budget and a bounded-horizon sweep. (4) **Benchmark:** an ETI benchmark on public MOT17 enter/exit tracks with an injected query per episode.

Non-goals. No product lift, no CTR, no claim that multimodal models are the wrong substrate, no human-cognition isomorphism, and no user-experience claim in this version. Query/goal-conditioned selectivity and false-memory are deferred to a v2 selective benchmark.

Scope. Our claim is scoped to current constraints—interaction latency, uplink bandwidth, and inference cost. We do not claim a single model is intrinsically insufficient: with near-free, instantaneous inference, a monolithic design would serve both stages and the contract would be unnecessary.

2 Related Work and Positioning

Single-path VLM agents and streaming video models. Recent camera-first and streaming systems route each frame or segment through one (large) multimodal model, optionally with tool calls. We cut on the **temporal contract**: which computation must complete *within* a target’s window, and which may be deferred.

Realtime, omni, and full-duplex models. A new class of streaming, low-latency multimodal models—GPT-4o Realtime, Gemini Live, Qwen2.5-Omni [Xu et al., 2025], Moshi [Défossez et al., 2024]—provides always-on audio-visual streaming with simultaneous input/output, and is a natural substrate for both the fast perceiver and the deferred deliberator (VisionClaw already runs on Gemini Live). It remains, however, a *single path* with an ephemeral context: continuous perception does not by itself guarantee that a sub-second target is remembered before it is lost, nor that evidence is retrievable for a later, tool-using answer. Our contract is **model-agnostic** and sits above this substrate: it fixes *when* evidence must be remembered and *what* must be persisted, however the backbone is built.

Always-on wearable agents. VisionClaw [Liu et al., 2026] is the closest system: on Meta Ray-Ban glasses it streams egocentric video into Gemini Live and delegates actions to an agentic backend, with reported gains in task time and workload. Two of its own measurements frame our problem: perception is sampled at ~ 1 frame/s, and end-to-end latency has a median ~ 12 s. Under our contract, 1 Hz sampling would leave sub-second windows largely uncovered, and 12 s is far outside the memory window. We treat VisionClaw-style systems as the motivation and formalize the perceptual-timing layer they currently lack.

Cascades, routing, and adaptive computation. Fast/slow and conditional compute have deep precedent: classifier cascades [Viola and Jones, 2001], adaptive computation time [Graves, 2016], early exit [Teerapittayanon et al., 2016], sparse MoE [Shazeer et al., 2017], LLM cascades [Chen et al., 2023, Aggarwal et al., 2023], routing [Ong et al., 2024], per-token depth routing [Raposo et al., 2024], and vision-internal reductions [Rao et al., 2021, Bolya et al., 2023]. Our target differs: **event-level remembering under a hard memory deadline.**

Dual-process and reactive-deliberative architectures. Our fast/slow split echoes dual-process theories [Kahneman, 2011, Evans and Stanovich, 2013]; we use the “System 1/2” label only as a heuristic and make no cognitive-psychology claim [Melnikoff and Bargh, 2018]. Layered fast/slow control dates to subsumption [Brooks, 1986] and three-layer architectures [Gat, 1998]. We revisit it for VLM-era camera-first agents.

Attention, saliency, active vision. Saliency [Itti et al., 1998] and active perception [Bajcsy, 1988, Mnih et al., 2014] establish selective viewing, typically goal-agnostic; our remembering is event-driven in v1 and becomes query/goal-conditioned in v2.

Bounded memory. Iconic/sensory memory is brief and capacity-limited; remembering commits selected content to durable store [Sperling, 1960, Atkinson and Shiffrin, 1968, Baddeley and Hitch, 1974, Cowan, 2001]. Used as design motivation, not isomorphism.

Latency in HCI. Response-time thresholds [Miller, 1968, Nielsen, 1993] motivate an interaction budget; our failure is stronger—lost evidence.

Counter-evidence. (i) *Record everything*: under an unbounded buffer remembering is trivial; we condition every metric on a budget and treat record-everything as the oracle upper bound. (ii) *Representation convergence* [Huh et al., 2024]: semantic-level convergence does not make the perceptual fast path redundant. (iii) *Fast VLMs*: our claim is about the temporal contract, not raw speed.

3 Protocol and Metrics

Task. Ephemeral-Target Interaction (ETI): a target appears at t_{in} , is query-relevant, and leaves at t_{out} with dwell $\tau = t_{\text{out}} - t_{\text{in}}$. The agent must (i) remember within $[t_{\text{in}}, t_{\text{out}}]$ under memory budget B , and (ii) later answer the query from memory. Remembering is hard-deadline; the answer is soft-deadline.

Episode schema. Each episode is $\{t_{\text{in}}, t_{\text{out}}, \text{target_id}, \text{query}, \text{answer_label}, \text{source}, \text{synthetic}\}$. Every episode carries an explicit query; without it, “missed” is undefined.

Budget. All memory metrics are functions of a per-sequence memory budget $B \in (0, 1]$: the fraction of ephemeral entries the gate may persist (a proxy for memory/compute). We keep B (a capture quota) distinct from the passive buffer horizon H (§4.1, §5.7): B bounds how many entries may be persisted, H bounds how long evidence remains available. The primary comparison is at fixed B ; the oracle is unbudgeted and instantaneous.

Metrics (layered). *Problem:* Window Miss Rate (WMR) vs. τ ; latency. *Solution (budget-conditioned):* memory recall; active-remembering rate; time-to-remember; bounded-

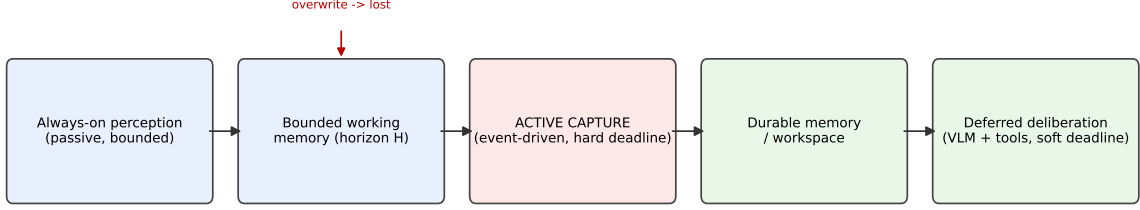


Figure 1: Ours: always-on bounded passive store $\rightarrow$ active remembering (hard deadline) $\rightarrow$ deferred deliberation (soft deadline).

horizon sweep. *Guard*: deferred-answer accuracy | remembered and joint success. False-memory/memory precision require a query with distractors and are deferred to v2.

Baselines. monolithic VLM; always-fast; fixed-rate sampling; classic saliency gate; ours; oracle.

4 Architecture: a two-deadline design

Our architecture instantiates the contract as three stages with a single live budget B (Figure 1).

- **Passive store (always-on, bounded, non-semantic).** Reception is passive: a cheap always-on path automatically writes a *bounded* working memory with horizon H . It is not semantic and not deliberate; $H = \infty$ degenerates to the oracle. *Always-on perception is not always-on recording.*
- **Active remembering (event-driven, hard deadline).** A sub-second gate remembers a moment into durable memory before it is overwritten (v1: on target entry). Remembering must occur within τ and consumes B .
- **Deferred deliberation (soft deadline).** A multimodal model with tools reads durable memory and answers; it may run after the target has left. Its output scores deferred-answer accuracy, not the primary claim.

5 Experiments

Benchmark. Primary: **MOT17 enter/exit** (per-track hard windows; 7 sequences, 796 tracks; median dwell 3.67s; 369 tracks ≤ 3 s) with an injected query. Secondary: LaSOT (re-entry), egocentric sources. Synthetic τ -sweep overlay only as a controlled supplement.

Protocol. For each episode, run a policy with budget B ; log memory events and the deferred answer; report memory metrics as functions of B and τ . Deterministic; no fitting.

E1–E3 (design). E1: deadline curve for the monolithic baseline. E2: ours vs. baselines (memory recall, TTR, compute; bounded-horizon sweep). E3: ablations (deferred). v1 ships E1 and the B - and horizon-sweeps as the labeled self-check (§5.7); the rest is future work.

Table 1: Memory recall on 796 real MOT17 tracks (scheduler simulation). Ours is event-driven at $B=1$.

policy	memory recall	TTR (s, conf.)	recall @ $\tau = 0.5/1/2/4$ s			
monolithic ($L=1.5$ s)	0.681	1.5	0.000	0.042	1.000	1.000
monolithic ($L=5$ s)	0.430	5.0	0.000	0.000	0.000	0.801
monolithic ($L=12$ s)	0.249	12.0	0.000	0.000	0.000	0.464
fixed sampling (2 fps)	0.957	0.30	0.755	1.000	1.000	1.000
saliency gate (goal-agnostic)	0.500	0.05	0.460	0.442	0.464	0.539
Ours (event-driven)	1.000	0.05	1.000	1.000	1.000	1.000
oracle	1.000	0.00	—	—	—	—

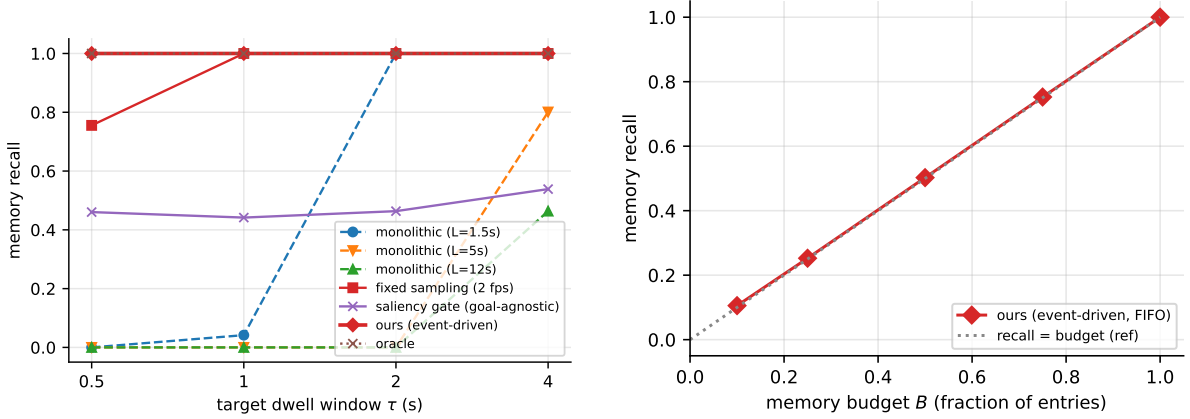


Figure 2: Left: deadline curve (memory recall vs dwell τ): the monolithic family collapses below its latency; ours/oracle are flat at 1. Right: memory recall vs memory budget B .

5.1 Contract check on real windows

Perception is treated as free here. Per-track windows are real (MOT17; 796 tracks, $369 \leq 3$ s are the ephemeral core); with perception free, the check isolates the contract’s *scheduling* logic. TTR is a configured gate latency. Table 1 and Figure 2.

Reading. By construction, *monolithic* recall is exactly the fraction of tracks longer than L (the dwell survival function), event-driven remembering is 1.0 (perception is free), goal-agnostic salience is a median split (~ 0.5), and the FIFO budget curve is $\text{recall} = B$. The bounded-horizon sweep (lazy gate, period 0.5 s) shows memory recall falling $0.957 \rightarrow 0.364$ as the horizon shrinks from 0.5 s to 0.25 s (Figure 3).

5.2 Measured frontier-VLM latency (calibration)

The useful latency of a monolithic pass is **estimated time-to-answer** = **TTFB** + **output_tokens / throughput** (+ tool calls). TTFB alone is a floor. From OpenRouter’s aggregated per-endpoint stats (last 30 min, request-weighted; text generation), estimated time-to-answer for a 64-token answer is 1.2–9.3 s (Table 2); low-throughput models pay most. This is a **lower bound**: OpenRouter’s stats exclude *image upload* and *image-token prefill*; a frame is ~ 100 KB (~ 0.04 – 0.16 s at 5–20 Mbps) plus image tokens (hundreds to low-thousands, model-dependent). Both terms are positive, so the single-path failure is conservative. An agentic identify-and-search is longer still (VisionClaw: ~ 12 s median) [Liu et al., 2026, OpenRouter, 2026].

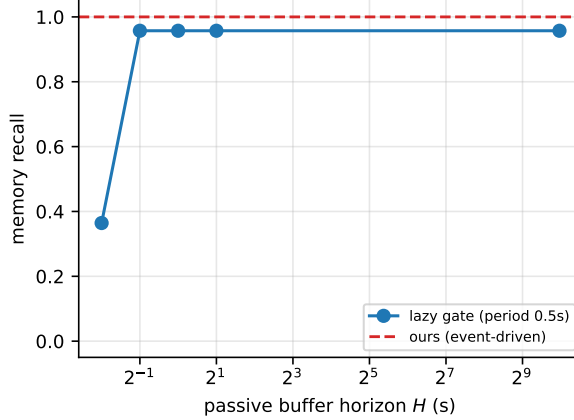


Figure 3: Bounded-horizon sweep: memory recall vs passive buffer horizon H (lazy gate, period 0.5 s) vs the horizon-robust event-driven remembering.

Table 2: Frontier-VLM latency (OpenRouter, last 30 min, request-weighted).

model	TTFB p50 (s)	throughput (tok/s)	est. time-to-answer, 64 tok (s)
google/gemini-3.5-flash-lite	0.73	127.5	1.23
openai/gpt-5.6-terra	1.47	82.6	2.25
anthropic/claude-sonnet-5	1.63	70.3	2.54
openai/gpt-6-sol	1.97	64.2	2.96
anthropic/claude-opus-5.5	5.50	65.7	6.47
qwen/qwen3-vl-8b-instruct	3.92	12.0	9.25

6 Limitations

- **Contingent on current costs.** The argument assumes a non-negligible latency/bandwidth/compute budget; with near-zero inference latency a single-path design would suffice—a judgment for the current environment, not a universal law.
- **Oracle-perception simulation.** Policies run over real track windows but no pixels/detector/model are run; the scheduler is the object, not perception. The empirical section is a *labeled self-check*, not a system evaluation.
- **Latency is a floor.** Text-generation stats exclude image upload and image prefill; real vision latency is higher.
- **Answer metric unrun.** MOT17 has no attribute labels; deferred-answer accuracy is not reported here.
- **Query-conditioning/false-memory deferred** to a v2 selective benchmark.

Ethics and privacy. Selective processing is selective data: the always-on fast path runs locally, and only actively remembered moments are persisted or uploaded, reducing exposure relative to continuous recording. This work uses public datasets only and no personal user data.

7 Conclusion

The binding constraint for future camera-first agents is not recognition quality but **time**. Many targets are ephemeral; a single end-to-end path loses their evidence before it can be used. We

formalized this as Ephemeral-Target Interaction and a Perceptual Deadline Contract, and proposed a two-deadline architecture: a bounded passive working memory, an active remembering gate that remembers within the hard window, and a deferred VLM-and-tools layer that answers within the soft window. On real track windows, event-driven remembering recovers every ephemeral target while a monolithic pass misses all those shorter than its latency. The contribution is a measurable contract and a system that respects it: *answers can be late; evidence cannot be missed*.

References

- Pranjal Aggarwal, Aman Madaan, Ankit Anand, et al. Automix: Automatically mixing language models. *arXiv preprint arXiv:2310.12963*, 2023.
- Richard C. Atkinson and Richard M. Shiffrin. Human memory: A proposed system and its control processes. In *Psychology of Learning and Motivation*, volume 2, pages 89–195. 1968. VERIFY.
- Alan D. Baddeley and Graham Hitch. Working memory. *Psychology of Learning and Motivation*, 8:47–89, 1974. VERIFY.
- Ruzena Bajcsy. Active perception. *Proceedings of the IEEE*, 76(8):966–1005, 1988. VERIFY.
- Daniel Bolya et al. Token merging: Your vit but faster. In *ICLR*, 2023. arXiv:2210.09461.
- Rodney A. Brooks. A robust layered control system for a mobile robot. *IEEE Journal on Robotics and Automation*, 2(1):14–23, 1986. doi: 10.1109/JRA.1986.1087032.
- Lingjiao Chen, Matei Zaharia, and James Zou. Frugalgpt: How to use large language models while reducing cost and improving performance. *arXiv preprint arXiv:2305.05176*, 2023.
- Nelson Cowan. The magical number 4 in short-term memory: A reconsideration of mental storage capacity. *Behavioral and Brain Sciences*, 24(1):87–114, 2001. VERIFY.
- Alexandre Défossez et al. Moshi: a speech-text foundation model for real-time dialogue. *arXiv preprint arXiv:2410.00037*, 2024.
- Jonathan St. B. T. Evans and Keith E. Stanovich. Dual-process theories of higher cognition: Advancing the debate. *Perspectives on Psychological Science*, 8(3):223–241, 2013. doi: 10.1177/1745691612460685.
- Erann Gat. On three-layer architectures. In *Artificial Intelligence and Mobile Robots*. AAAI Press, 1998. VERIFY.
- Alex Graves. Adaptive computation time for recurrent neural networks. *arXiv preprint arXiv:1603.08983*, 2016.
- Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*, 2024.
- Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, 1998. VERIFY.
- Daniel Kahneman. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, 2011.

- Daniel Kersten, Pascal Mamassian, and Alan Yuille. Object perception as bayesian inference. *Annual Review of Psychology*, 55:271–304, 2004. doi: 10.1146/annurev.psych.55.090902.142005.
- Xiaoan Liu, DaeHo Lee, Eric J. Gonzalez, Mar Gonzalez-Franco, and Ryo Suzuki. Visionclaw: Always-on ai agents through smart glasses. *arXiv preprint arXiv:2604.03486*, 2026. cs.HC; CC-BY-4.0; verified live.
- David E. Melnikoff and John A. Bargh. The mythical number two. *Trends in Cognitive Sciences*, 22(4):280–293, 2018. doi: 10.1016/j.tics.2018.02.001.
- Robert B. Miller. Response time in man-computer conversational transactions. *Proceedings of the AFIPS Fall Joint Computer Conference*, pages 267–277, 1968. doi: 10.1145/1476589.1476628.
- Volodymyr Mnih, Nicolas Heess, Alex Graves, and Koray Kavukcuoglu. Recurrent models of visual attention. *arXiv preprint arXiv:1406.6247*, 2014.
- Jakob Nielsen. *Usability Engineering*. Morgan Kaufmann, 1993. VERIFY.
- Isaac Ong et al. Routellm: Learning to route llms with preference data. *arXiv preprint arXiv:2406.18665*, 2024.
- OpenRouter. Openrouter model latency statistics (last 30 minutes, request-weighted p50). <https://openrouter.ai/models>, 2026. accessed 2026-10-10.
- Rajesh P. N. Rao and Dana H. Ballard. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, 2(1):79–87, 1999. doi: 10.1038/4580.
- Yongming Rao et al. Dynamicvit: Efficient vision transformers with dynamic token sparsification. In *NeurIPS*, 2021. arXiv:2106.02034.
- David Raposo et al. Mixture-of-depths: Dynamically allocating compute in transformer-based language models. *arXiv preprint arXiv:2404.02258*, 2024.
- Noam Shazeer et al. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*, 2017.
- George Sperling. The information available in brief visual presentations. *Psychological Monographs: General and Applied*, 74(11):1–29, 1960. VERIFY.
- Surat Teerapittayanon, Bradley McDanel, and H. T. Kung. Branchynet: Fast inference via early exiting from deep neural networks. *arXiv preprint arXiv:1709.01686*, 2016.
- Paul Viola and Michael Jones. Rapid object detection using a boosted cascade of simple features. In *IEEE CVPR*, 2001. VERIFY.
- Jin Xu et al. Qwen2.5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025.