DLite: History-Free Speculative Decoding for Agentic AI

Research group paper discussion

Paper by Jiaming Xu et al.

arXiv:2610.23787v1 16 October 2026

arxiv.org/abs/2610.23787v1

The thesis: let the target model carry the history

Problem

An agent repeatedly grows its context. A small drafter must both understand that history and maintain a second historical KV cache.

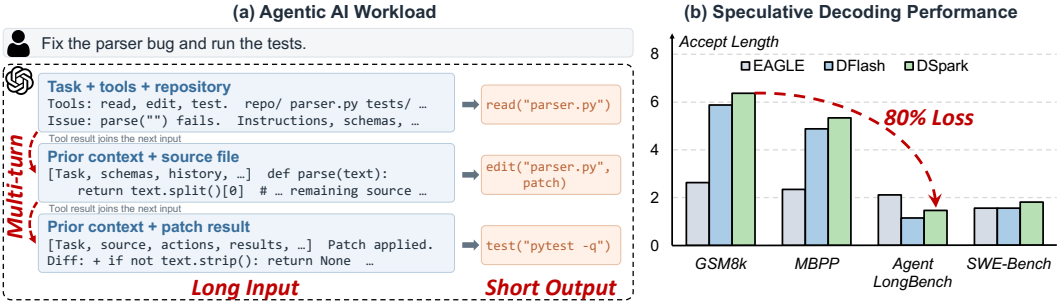
DLite

Reuse target-model hidden states; predict in a fixed local window; restore only the missing within-block causal dependence.

Qwen3-8B SWE-bench: 2.74× decoding speedup versus 1.60× for DSpark.

Paper, Secs. 3–5; Table 3, PDF p. 8. Speedups are relative to vanilla decoding.

Agent turns grow the input but often produce short outputs



- Existing drafters can fall to roughly 1–2 output tokens per verification round on agent tasks.
- Repeated short turns make draft-state maintenance frequent.

Paper, Fig. 1 and Sec. 3.1, PDF pp. 2–4. Figure reproduced from the authors.

Speculative speedup depends on quality *and* draft cost

Draft → verify → commit

A draft model proposes N tokens. The target verifies them together and commits the accepted prefix, with a target-model continuation token at rejection.

$$\text{Speedup} \approx \frac{\text{Time}_{M_T}(1)\,\tau}{\text{Time}_{M_T}(N) + \text{Time}_{M_D}(N)}$$

- τ: mean tokens produced per target verification; larger is better.
- M_T: target model; M_D: drafter. Draft latency can erase gains in τ.

Paper, Sec. 2.1, Eq. (1), Fig. 2(a), PDF p. 3. The equation is an approximation.

A historical draft KV cache creates a second long-lived state

Model-side cost

A limited-capacity drafter independently reconstructs the growing context; its acceptance length falls on agent workloads.

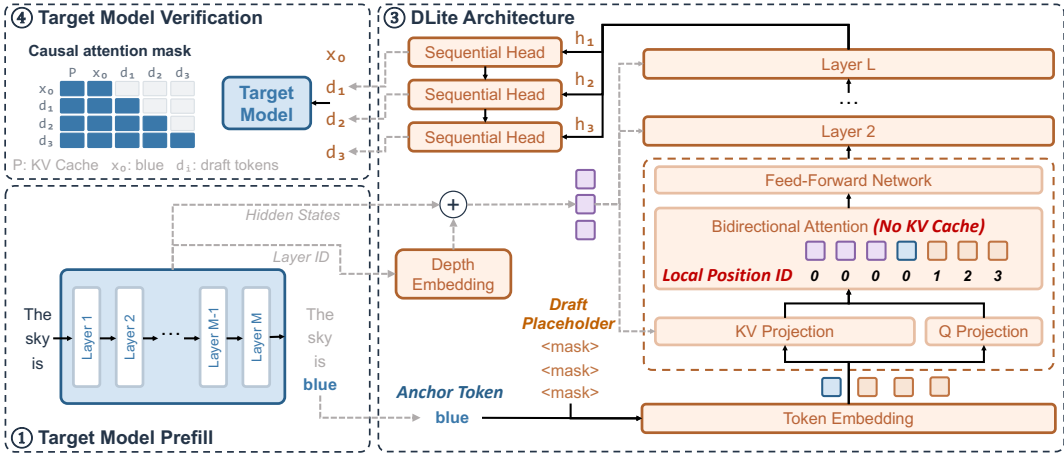
System-side cost

Accepted tokens update draft KV, and prefix reuse preserves it across turns. A DeepSeek-V4 MTP/SGLang example attributes 10–20% of inference latency to this work.

The target model has already encoded the historical context.

Paper, Sec. 3, PDF pp. 3–4. The 10–20% measurement is for the reported setup.

DLite drafts locally while the target retains its KV cache



Paper, Fig. 3 and Sec. 4.1, PDF pp. 4–5. Figure reproduced from the authors.

Target features carry context into a fixed local window

- Extract multiple target-layer states at the current position.
- Add a learned source/depth embedding:

$$z_t^{(l)} = h_t^{(l)} + e_{\text{depth}}^{(l)}.$$

- Present layers as separate sequence elements.

Local prediction input

Target features
+ current anchor token
+ K draft placeholders

- Local position IDs reset each round.
- No historical *draft* KV is stored.

Paper, Sec. 4.1.1, Eq. (2), Fig. 3, PDF pp. 4–5; Appendix A.2, PDF p. 13.

Parallel drafting is missing the within-block prefix

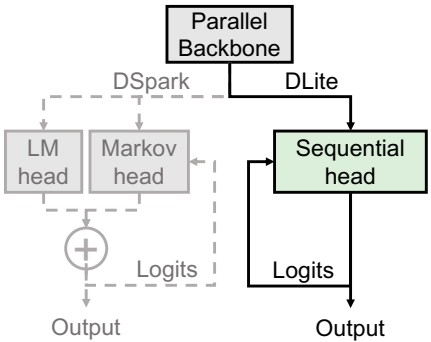
$$p_T(x_{t+T} \mid x_{<t+T}) = p_T(x_{t+T} \mid x_{<t}, x_{t:t+T-1}).$$

- Parallel backbone computes h_T before positions $< T$ are sampled.
- Target hidden states supply historical context $x_{<t}$.
- A light recurrent state supplies the generated block prefix:

$$s_T = g_\phi(s_{T-1}, E(x_{t+T-1})), \qquad q_{\theta,\phi}(x_{t+T} \mid h_T, s_T) \approx p_T(x_{t+T} \mid x_{<t+T}).$$

Paper, Sec. 4.1.2, Eqs. (3)–(5), PDF p. 5. E is the token embedding.

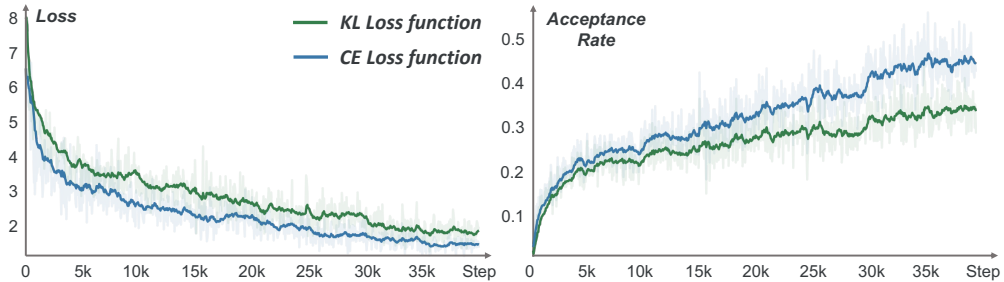
The sequential head works before vocabulary projection



- Low-rank RNN-style head updates only within a draft block.
- DLite refines hidden states before applying the LM head.
- This avoids an intermediate $D \times V$ projection at every draft position (D: hidden size, V: vocabulary size).

Paper, Fig. 4(a), Sec. 4.1.2, PDF pp. 5–6. Cropped author figure.

Direct target-model KL struggles in this architecture



- In the paper's direct-distillation comparison, CE yields a higher acceptance rate than KL.
- The target predicts autoregressively over history; the backbone predicts in parallel from compressed target states.

Paper, Fig. 4(b), Sec. 4.2, PDF p. 6. Cropped author figure.

Two-stage training bridges the conditioning mismatch

1. Warm-up

Train an isomorphic, history-aware teacher with token CE. Then distill its distribution into history-free DLite with forward KL.

2. Stable stage

Remove the teacher. Align DLite directly to the target model using token CE and a target–draft L_1 distribution loss.

Training-only teacher: no historical draft KV at inference.

Paper, Sec. 4.2, PDF p. 6; Appendix A.1, PDF p. 12.

The two objectives supervise different gaps

$$\mathcal{L}_{\text{warmup}} = \sum_{i=1}^{K} w_i D_{\text{KL}}(p_i^H \| q_i), \qquad w_i = e^{-(i-1)/\gamma}, \quad \gamma = 4.$$

$$\mathcal{L}_{\text{stable}} = \lambda_{\text{CE}} \mathcal{L}_{\text{CE}} + \lambda_{\text{TV}} \sum_i w_i \| p_i^T - q_i \|_1.$$

- p_i^H: history-aware teacher; p_i^T: target; q_i: DLite.
- Teacher CE: 8 epochs; warm-up KL: 2; stable stage: 6.
- $\lambda_{\text{CE}} = 0.1$, $\lambda_{\text{TV}} = 1$. The paper's "TV" formula uses L_1 without a $1/2$ factor.

Paper, Sec. 4.2, Eqs. (6)–(7), PDF p. 6; Appendix A.1, PDF p. 12.

Evaluation separates acceptance, decoding speed, and TPS

- **Targets:** Qwen3-4B/8B on NVIDIA H800; $K = 7$ for all methods.
- **DLite:** 5-layer backbone; 12/14 target hidden states; rank 320/512 for 4B/8B.
- **Tasks:** general math/code/chat; four LongBench v2 categories; four agent benchmarks.
- **Comparators:** vanilla decoding; EAGLE-3, DFlash, DSpark; DFlash2/DFlow in Table 1.

Metric reading

τ = tokens per target verification; speedup = decoding rate / vanilla rate; TPS = aggregate token throughput at a stated concurrency.

Paper, Sec. 5.1, PDF p. 7; Appendix A.1, PDF p. 12.

On standard tasks, acceptance mostly improves modestly

Target	Benchmark	DSpark τ	DLite τ
Qwen3-4B	GSM8K	6.27	**6.38**
Qwen3-4B	MBPP	5.48	**5.64**
Qwen3-8B	LiveCodeBench	5.43	**5.66**
Qwen3-8B	Alpaca	3.64	**3.78**

- Removing draft history does not lower acceptance against DSpark in Table 1.
- Exception to "best against everyone": 4B MT-Bench has DFlash2 4.14, DLite 3.94.

Paper, Table 1, Sec. 5.2, PDF p. 7. Higher τ is better.

Long context gives larger gains in acceptance and speed

Target	LongBench v2 task	DSpark	DLite
4B	Multi-Doc QA	1.81× / 2.20	2.18× / **2.74**
4B	Code Repo Ud.	2.02× / 2.44	2.66× / **3.14**
8B	Long-dialog Hist.	3.02× / 4.01	3.96× / **4.77**

Each cell: decoding speedup / acceptance length τ

The advantage appears on all four reported LongBench v2 categories for both target sizes.

Paper, Table 2, Sec. 5.2, PDF p. 8. Speedup is versus vanilla decoding.

Agent workloads show the strongest relative speed gains

Target	Agent task	DSpark	DLite
4B	AgentLongBench	1.45× / 1.84	2.07× / **2.51**
4B	SWE-bench	1.64× / 1.89	2.68× / **3.15**
8B	SWE-bench	1.60× / 1.94	2.74× / **3.19**

- 8B SWE-bench: $(2.74/1.60 - 1) \approx 71\%$ higher speedup ratio than DSpark.
- Figures describe model decoding; BFCL tool execution time is excluded.

Paper, Table 3, Sec. 5.2, PDF p. 8; Appendix A.2, PDF pp. 13–14.

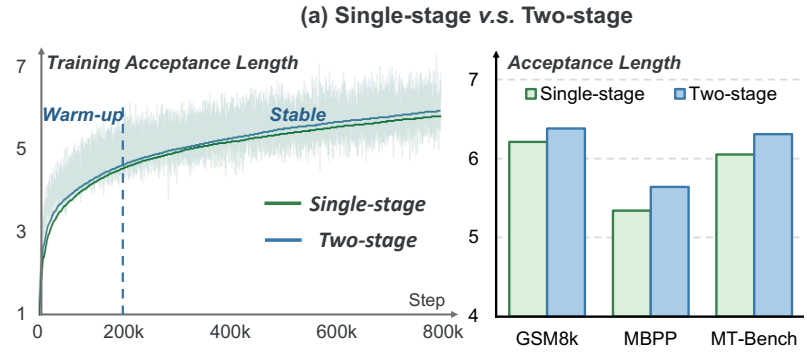
At concurrency 64, DLite retains a throughput edge

Target	Task	DSpark TPS	DLite TPS	Relative gain
4B	Alpaca	13,548	**14,300**	5.6%
4B	LiveCodeBench	11,309	**12,075**	6.8%
8B	LiveCodeBench	5,621	**5,758**	2.4%

- Total TPS reflects both acceptance quality and draft/system overhead.
- The table supports a throughput benefit; it does not isolate head latency.

Paper, Table 4, Sec. 5.2, PDF p. 9. One tiny exception: 4B MT-Bench at concurrency 16.

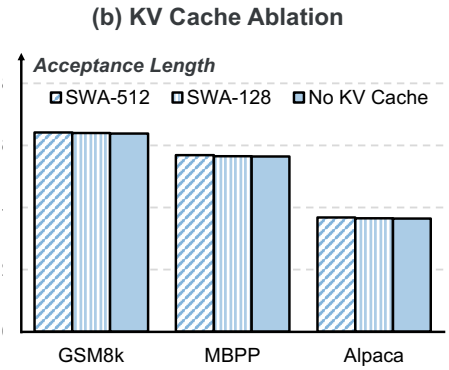
The matched teacher improves final acceptance



- Two-stage training exceeds single-stage training through the stable phase.
- Final acceptance improves on GSM8K, MBPP, and MT-Bench.

Paper, Fig. 5(a), Sec. 5.3, PDF p. 9. Cropped author figure.

After training, a draft KV window adds little acceptance



- Compare no draft KV with 128- and 512-token sliding KV windows.
- Final τ is similar on GSM8K, MBPP, and Alpaca.
- The authors infer target features can substitute for draft history after sufficient training.

Paper, Fig. 5(b), Sec. 5.3, PDF pp. 8–9. Cropped author figure.

A moderate feature count and head rank are sufficient

Design choice	Configuration	Avg. τ
Target hidden states	7 / **14** / 36	3.14 / **3.30** / 3.27
Composition	feature / **sequence**	3.04 / **3.30**
Sequential-head rank	256 / **512** / 4096	3.16 / **3.30** / 3.33

- The 8B design study uses a common 80K-sample subset.
- Larger designs give diminishing acceptance gains with more computation.

Paper, Appendix A.2, Table 5, PDF p. 13. Average over Alpaca, GSM8K, MBPP, Multi-Doc QA.

Takeaways and measured scope

1. **Architectural idea:** remove growing draft history while reusing contextual target states.
2. **Quality mechanism:** parallel local backbone plus low-rank within-block causal head.
3. **Optimization:** matched history-aware teacher, then direct CE + L_1 target alignment.
4. **Evidence:** strong decoding gains on reported Qwen3-4B/8B long-context and agent tasks.

Deployment question

The serving stack must expose target hidden states; reported speedups measure decoding, not full agent task completion or tool time.

Paper, Secs. 4–6 and Appendix A.2, PDF pp. 4–10, 13–14.

Paper and discussion pointers

Primary source

Jiaming Xu et al., *DLite: History-free Speculative Decoding for Agentic AI*, arXiv:2610.23787v1, 2026.

https://arxiv.org/abs/2610.23787v1

- Fig. 3: what information moves from target to drafter?
- Fig. 5: when can target features replace draft KV?
- Table 3: which agent traces give the largest relative gain?