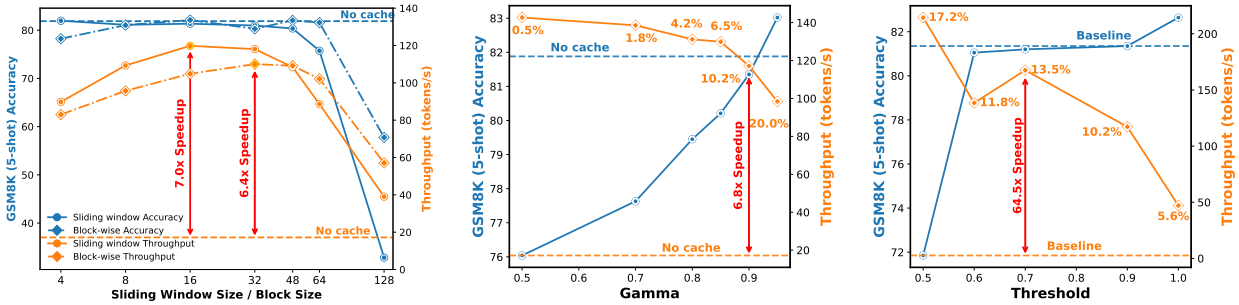

(a) Impact of Sliding window and block size (b) Cache update frequency. (c) Confident-aware decoding.

Figure 3: Ablation study and analysis of our proposed method. (a) Ablation study of our sliding window mechanism compared to block-wise decoding. (b) Analysis of cache update frequency under varying γ. The blue and orange lines represent accuracy and throughput, respectively. The numbers along the lines indicate the frequency of cache updates, assuming no baseline. (c) Analysis of cache update frequency under confident-aware decoding with varying ϵ.

Table 4: Impact of attention threshold on accuracy and speedup under GSM8K (5-Shot) for LLaDA and LLaDA1.5 with generation length of 512.

Model	No Cache	Fast-dLLM	Elastic-Cache (Ours) $\gamma = 0.5$	$\gamma = 0.7$	$\gamma = 0.8$	$\gamma = 0.85$	$\gamma = 0.9$	$\gamma = 0.95$
LLaDA	77.10	74.83	71.57	73.46	74.30	74.68	77.71	76.72
	3.6 (1.0×)	44.0 (12.2×)	109.9 (30.5×)	108.7 (30.2×)	103.9 (28.9×)	99.1 (27.5×)	91.5 (25.4×)	75.5 (21.0×)
LLaDA-1.5	81.35	80.82	76.04	77.63	79.45	80.21	81.35	83.02
	2.6 (1.0×)	36.8 (14.2×)	142.7 (54.9×)	138.6 (53.3×)	131.2 (50.5×)	129.9 (50.0×)	117.2 (45.1×)	98.4 (37.8×)

Table 5: Comparison between Elastic-Cache and Fast-dLLM when varying few-shots and generation length.

(a) Impact of few-shots on Accuracy and Speedup Under GSM8K (generation length of 1024) for LLaDA.

Model.	3-shot	5-shot	8-shot
Fast-dLLM	73.77	76.04	75.36
	28.5 (1.0×)	25.0 (1.0×)	20.8 (1.0×)
Elastic-Cache	75.13	75.21	75.28
	185.3 (6.5×)	169.8 (6.8×)	143.9 (6.9×)

(b) Impact of generation length on Accuracy and Speedup Under GSM8K (5-Shot), $\gamma = 0.8$ for LLaDA.

Model.	256	512	1024
Fast-dLLM	77.94	74.83	76.04
	53.7 (1.0×)	44.0 (1.0×)	25.0 (1.0×)
Elastic-Cache	78.24	77.71	75.21
	58.0 (1.1×)	91.5 (2.1×)	169.8 (6.8×)

Cache (Ma et al., 2025), and DeepCache (Ma et al., 2024). Orthogonal accelerations exploit parallel/non-AR generation (Gu et al., 2017; Xiao et al., 2023), block-wise diffusion (Arriola et al., 2025), fast sampling (Chen et al., 2023), test-time scaling (Ramesh & Mardani, 2025), and consistency models (Kou et al., 2024). However, most rely on temporal heuristics or fixed thresholds, leaving attention patterns underused. **Our Perspective.** We close this gap with attention-aware and layer-aware caching for diffusion LLMs: tracking most-attended tokens and depth-varying KV dynamics to guide recomputation, complementary to interval-based (Ma et al., 2025) and confidence-based (Wu et al., 2025) policies and compatible with the broader acceleration toolkit (Ainslie et al., 2023; Su et al., 2024; Touvron et al., 2023a;b; Dubey et al., 2024; Gu et al., 2017; Xiao et al., 2023; Arriola et al., 2025; Chen et al., 2023; Ramesh & Mardani, 2025; Kou et al., 2024).

6 Conclusion

We presented **Elastic-Cache**, a training-free, architecture-agnostic policy that makes KV caching in diffusion LLMs adaptive along two axes: *when* to refresh (via an attention-aware drift test) and *where* to refresh (via a depth-selective update starting at a learned boundary layer). By block-caching distant MASK tokens, reusing shallow-layer caches, and refreshing only when the most-attended token indicates meaningful state change, Elastic-Cache removes large amounts of redundant QKV work. Across decoding steps, this yields substantial latency reductions with negligible impact on generation quality, addressing a key deployment bottleneck for diffusion decoders. Looking ahead, we plan to refine drift thresholds with learned predictors, formalize guarantees linking attention patterns to KV drift, and explore interplay with speculative decoding or other hardware-aware scheduling, extending the same principles to autoregressive LLMs and multimodal diffusion frameworks.