

A Experimental Details

A.1 Throughput Measurement

Throughout the experiments, we measure the maximum reachable prefilling and decoding throughput of JetLM and the baselines on a single H100 GPU. This is achieved by adjusting the chunk size in chunk-prefilling [63] to maximize the decoding batch size without sacrificing the prefilling throughput. We list the optimized batch size and the corresponding chunk size for each model in Table 7. The prefilling context length is 64K.

Model	Batch Size	Chunk Size
Qwen2.5-1.5B	32	4,096
H2O Danube-1.8B	16	4,096
Llama3.2-1B	32	4,096
MiniCPM-2.8B	2	2,048
Pythia-2.8B	1	16,384
Smollm2-1.7B	4	16,384
StableLM2-1.6B	4	16,384
Mamba2-2.7B	128	1,024
RWKV7-1.5B	256	2,048
Rec.Gemma-2B	128	512
Gemma3-1B	64	4,096
Gemma2-2.6B	32	2,048
Hymba-1.5B	64	512
Zamba2-1.2B	8	8,192
JetLM-2B	256	1,024
JetLM-3.7B	128	512

Table 7: Hyper-Parameters in Efficiency Measurement.

A.2 Compute Resources

We train JetLM-2B with 16 NVIDIA H100 GPUs and JetLM-3B with 32 NVIDIA H100 GPUs.

B Broader Impacts

JetLM matches the accuracy of SOTA full-attention models while achieving up to $13\times$ higher decoding throughput, which can serve as the basis for long-reasoning models to reduce their deployment costs. It also helps the research of large-scale reinforcement learning, where decoding takes a large proportion of time. Furthermore, PostNAS offers a lightweight strategy to evaluate newly designed model architectures, which can accelerate the development of next-generation foundation models. A potential risk of JetLM is being used to synthesize human-like texts on the Internet, which is hard to detect. This can be migrated by adding watermarks [90] to model-generated contents.

C Safeguards

We will provide detailed guidelines for the released model to avoid users’ misuse. We will also add safety filters [91] to our online demo.

D Licenses for Existing Assets

The existing assets used in our paper and the licenses are:

- Qwen2.5 [4]: Qwen Research License, available for research use. We will mention that our model starts from inheriting the weights from Qwen2.5 models in the final model release.

- 530 • Redstone-QA [57]: MIT License, available for research use.
- 531 • Megamath [58]: Apache License, available for research use.
- 532 • Nemotron-CC [56]: The developers allow users to use it for pre-training or fine-tuning
- 533 language models, including both closed-source and open-source models.