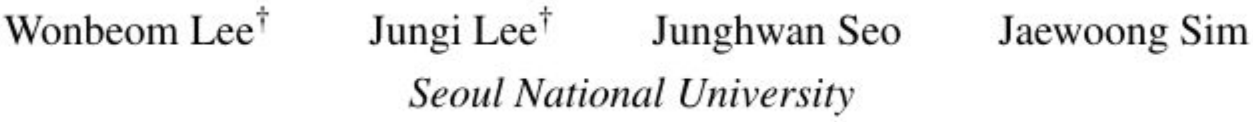
InfiniGen

Efficient Generative Inference of Large Language Models with Dynamic KV Cache Management

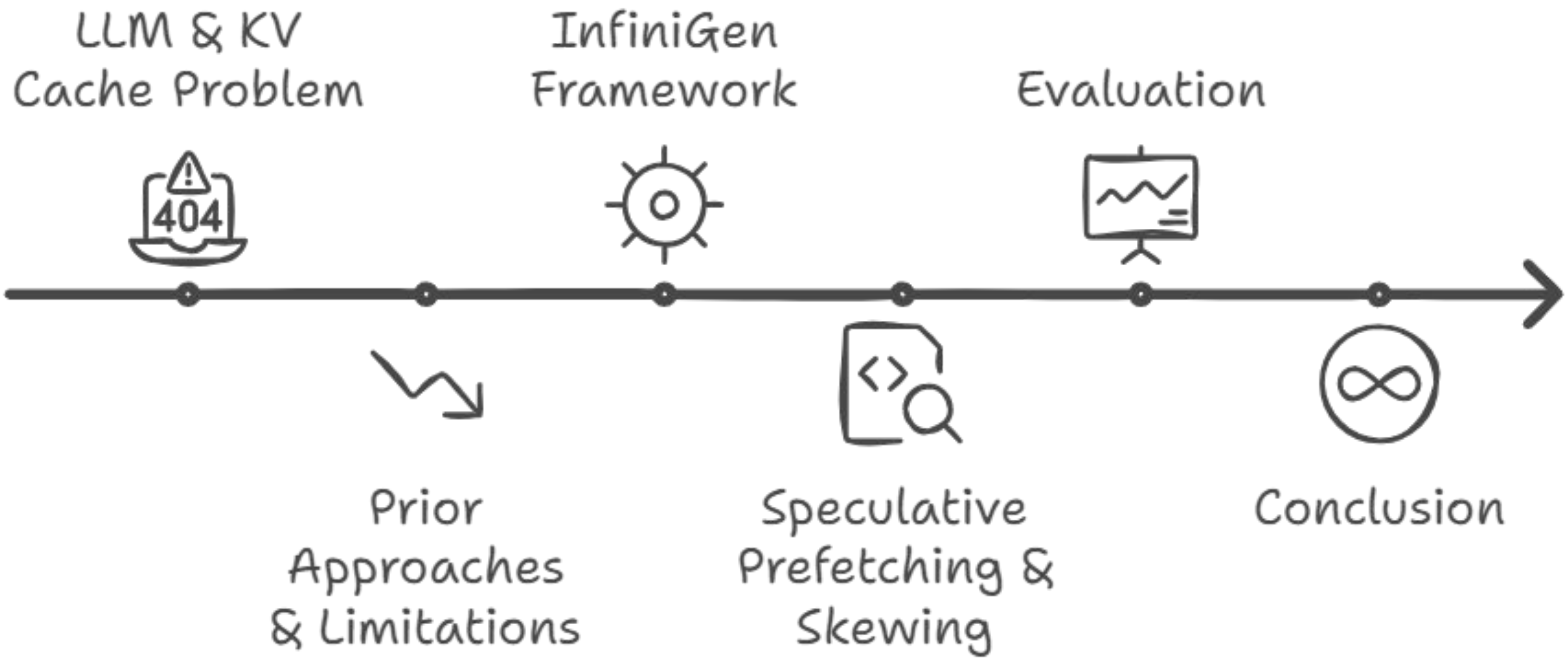


Wonbeom Lee[†] Jungi Lee[†] Junghwan Seo Jaewoong Sim

Seoul National University

110550108 施柏江 110550068 王振倫 313551099 李以恩

Outline



LLM & KV Cache Problem

Prior Approaches & Limitations

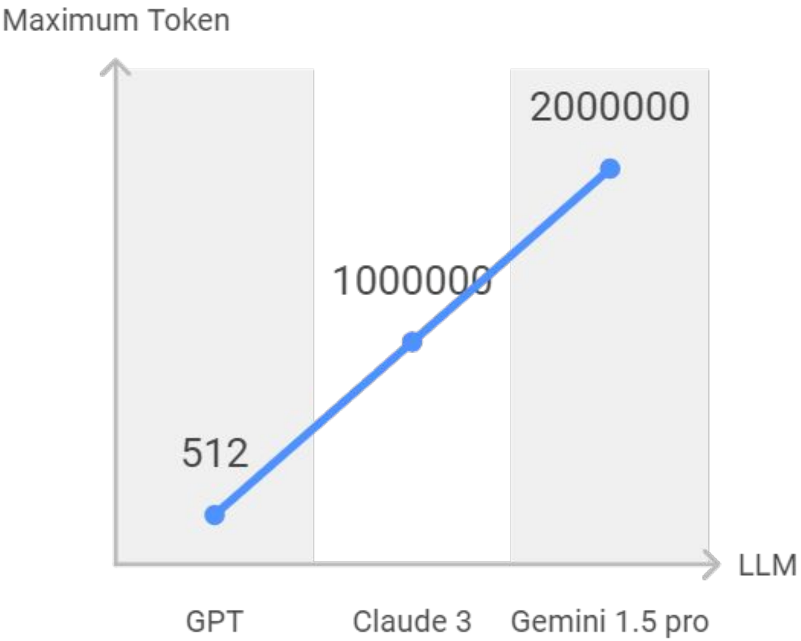
InfiniGen Framework

Speculative Prefetching & Skewing

Evaluation

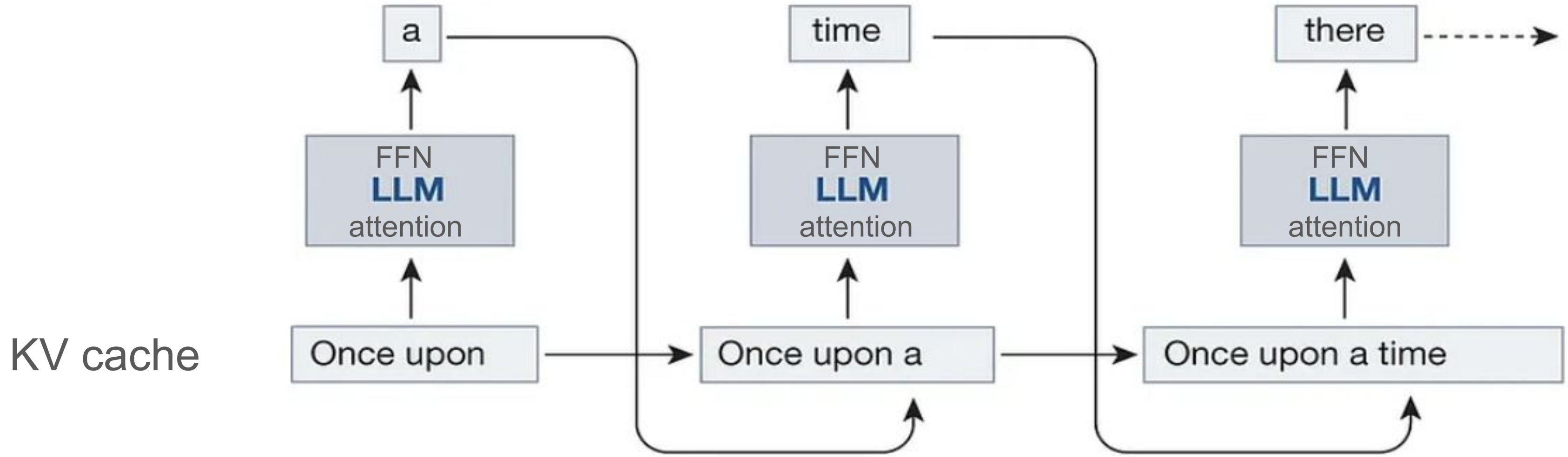
Conclusion

LLM

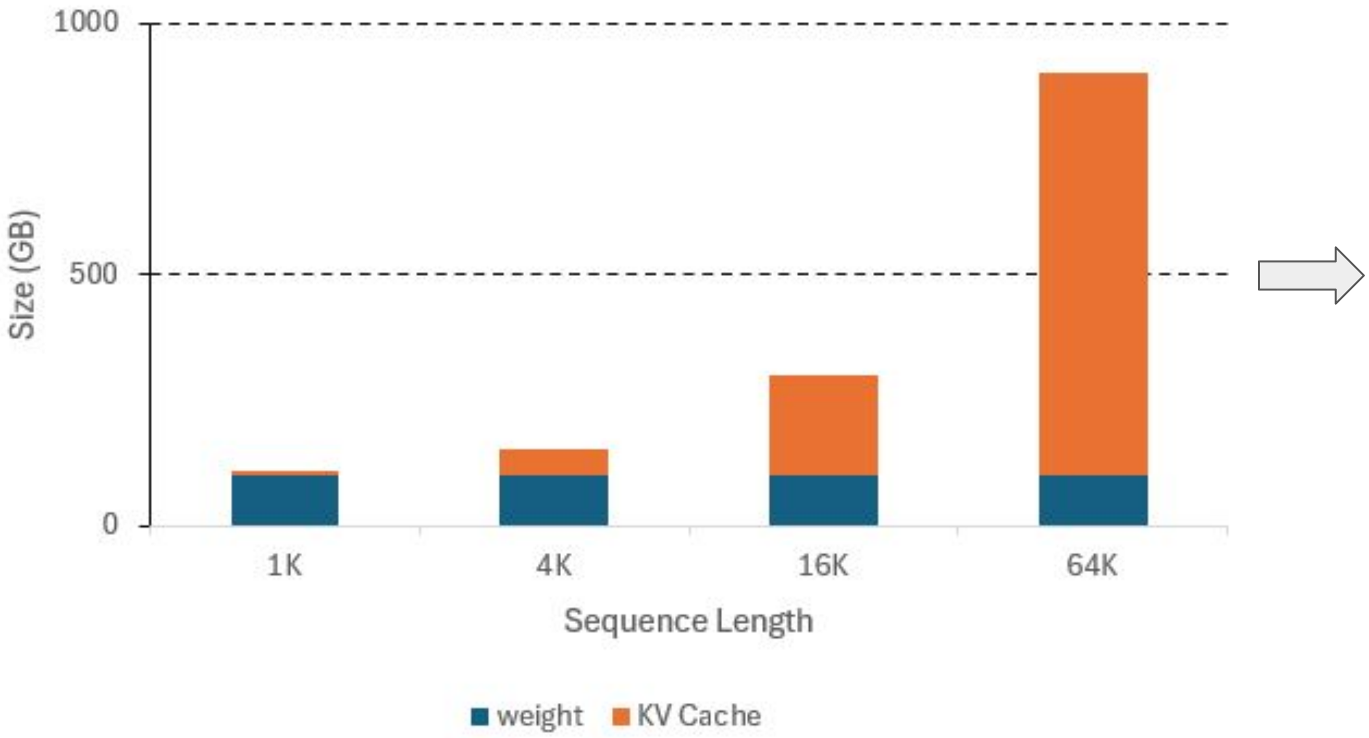


Can handle millions of words and hours of video and audio !

KV Cache

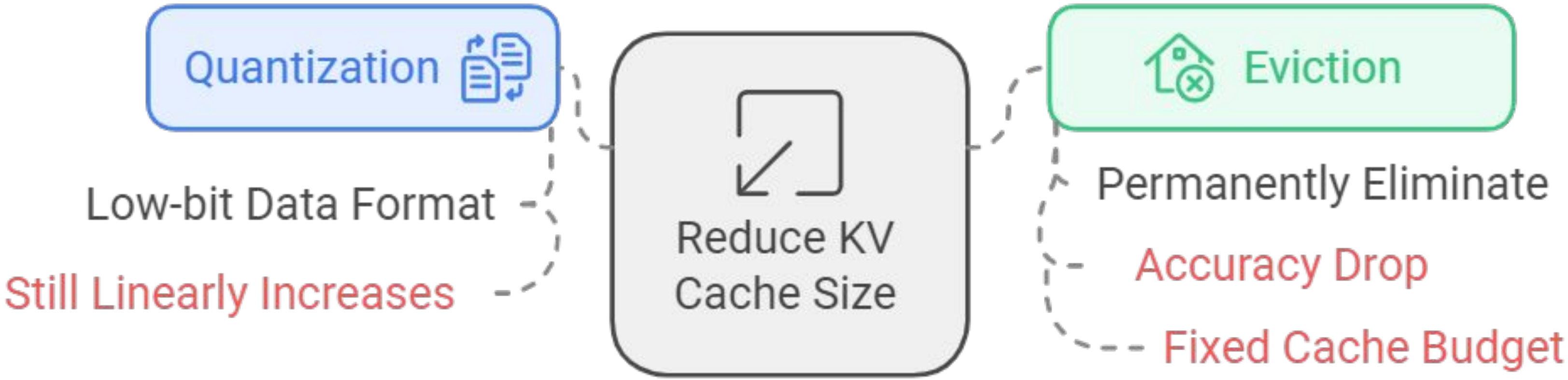

a
time
there
FFN
LLM
attention
FFN
LLM
attention
FFN
LLM
attention
KV cache
Once upon
Once upon a
Once upon a time

Memory



GPU Memory Capacity

Prior Approaches & Limitations



Quantization

Low-bit Data Format

Still Linearly Increases

Reduce KV Cache Size

Eviction

Permanently Eliminate

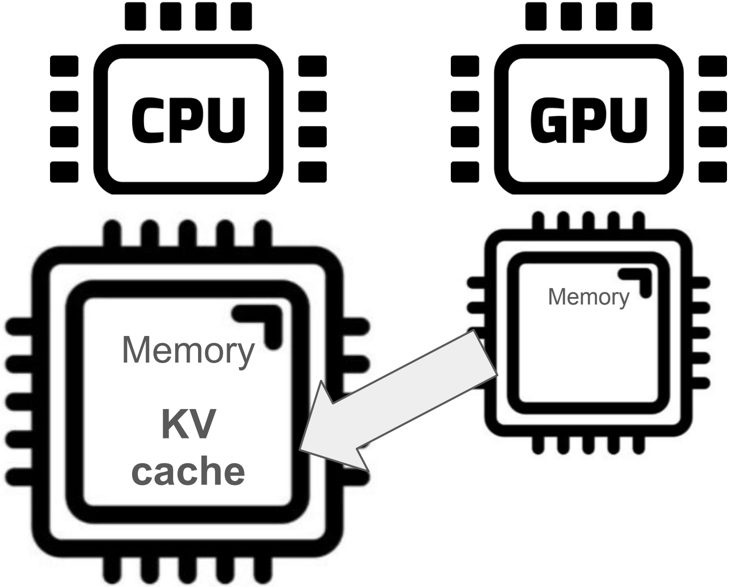
Accuracy Drop

Fixed Cache Budget

Not an effective solution in millions of tokens !

Offloading

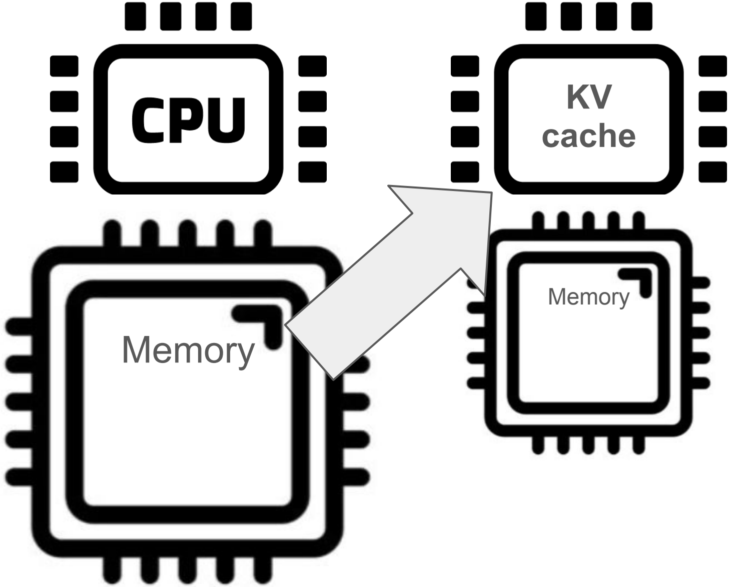
InfiniGen use the abundant CPU memory capacity to manage the KV cache



KV cache offloading

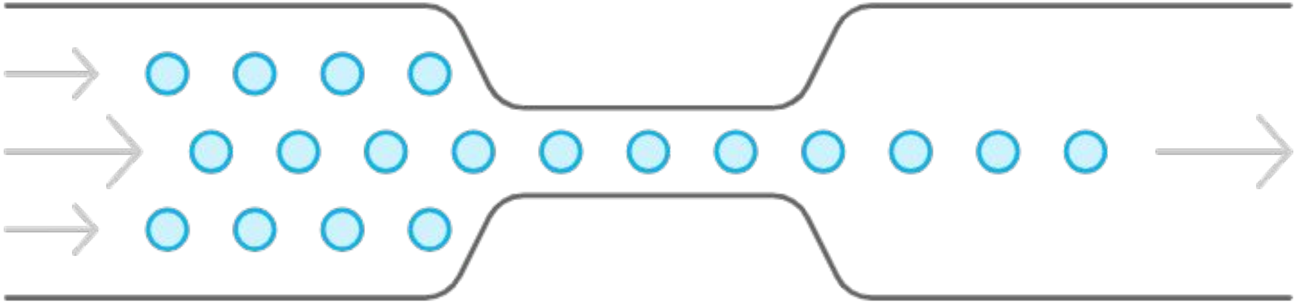
Offloading

InfiniGen use the abundant CPU memory capacity to manage the KV cache



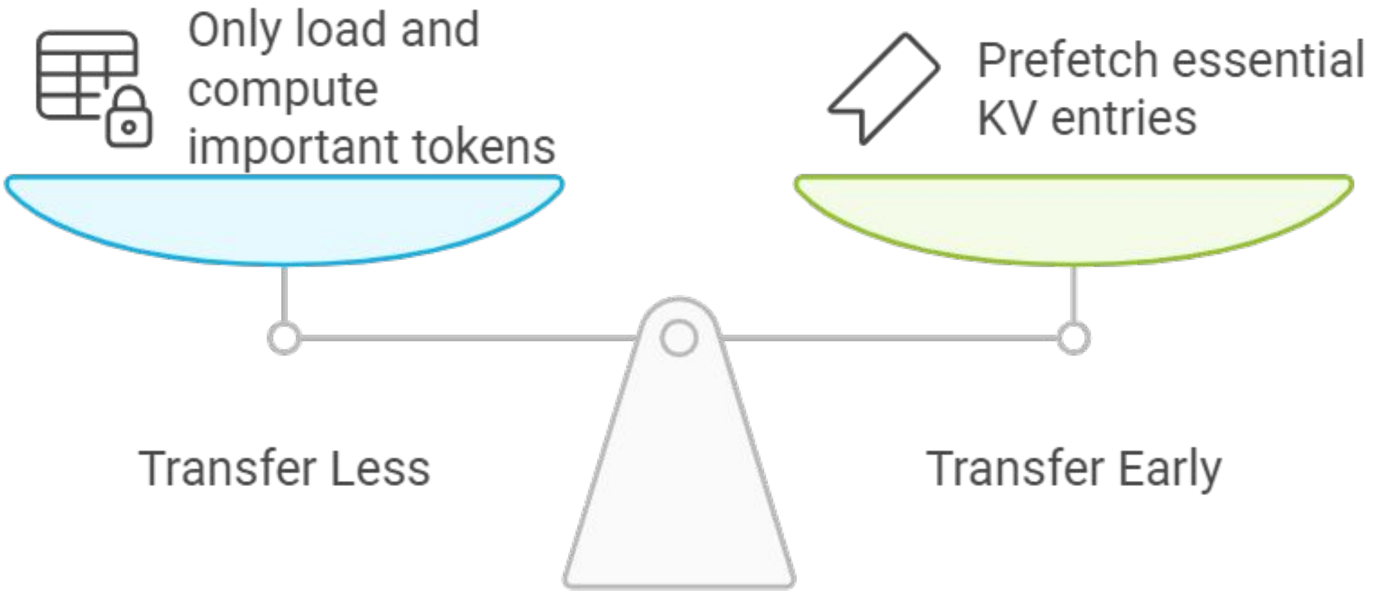
KV cache offloading

Offloading



Significant slowdown due to the limited PCIe bandwidth

Data Transfer



Only load and compute important tokens

Prefetch essential KV entries

Transfer Less

Transfer Early

Speculative Prefetching

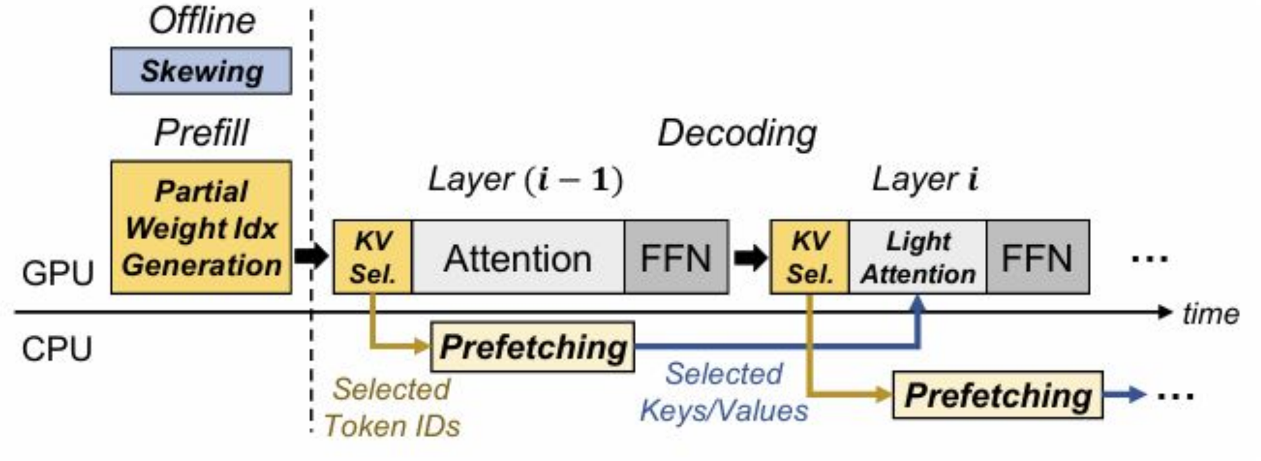


Figure 8: Operation flow of the prefetching module of InfiniGen.

But how to predict the important tokens?

Speculative Prefetching

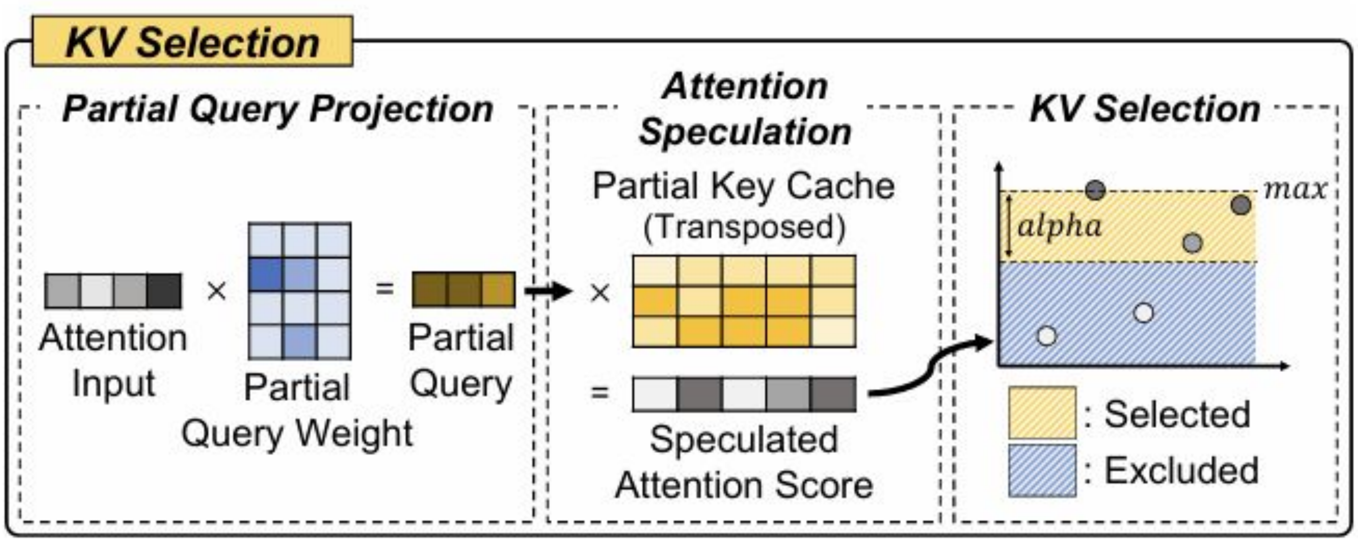
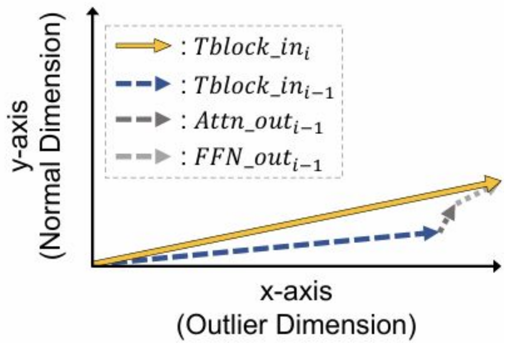
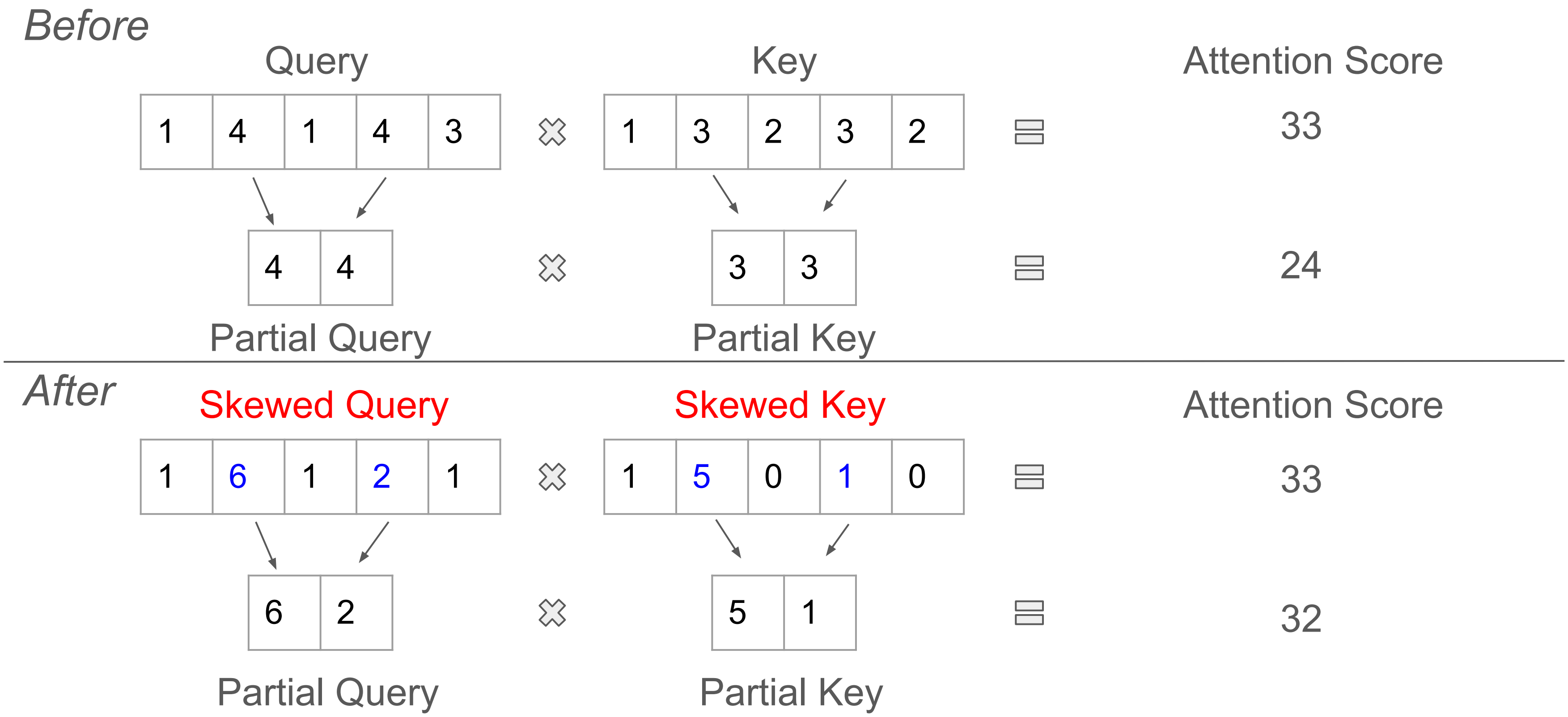


Figure 10: Attention score speculation in the decoding stage.



(a) Input Similarity

Skewing



Before

Query

1	4	1	4	3

⊗

Key

1	3	2	3	2

=

Attention Score

33

4	4

Partial Query

⊗

3	3

Partial Key

=

24

After

Skewed Query

1	6	1	2	1

⊗

Skewed Key

1	5	0	1	0

=

Attention Score

33

6	2

Partial Query

⊗

5	1

Partial Key

=

32

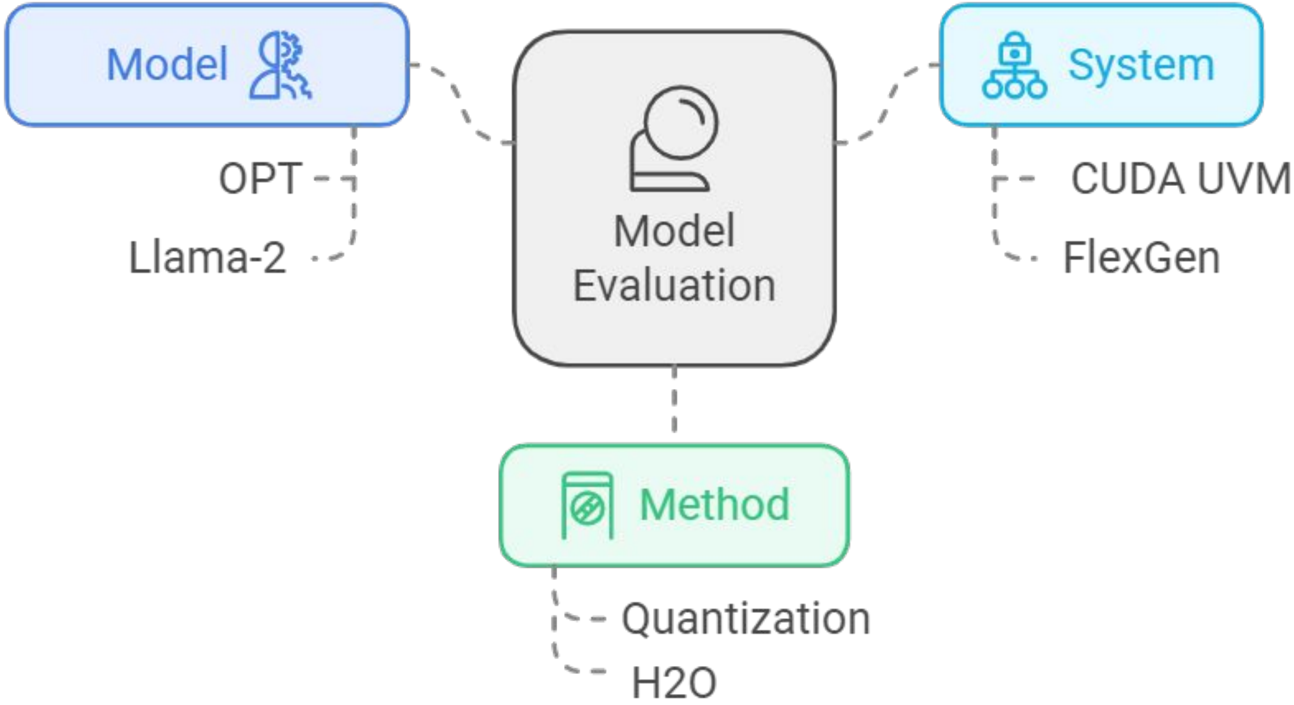
Skewing

How does it remain unchanged?

A: Offline modification of the query/key weights using SVD

$$(Q \times A) \times (A^T \times K^T) = Q \times K^T$$

Experimental Setup



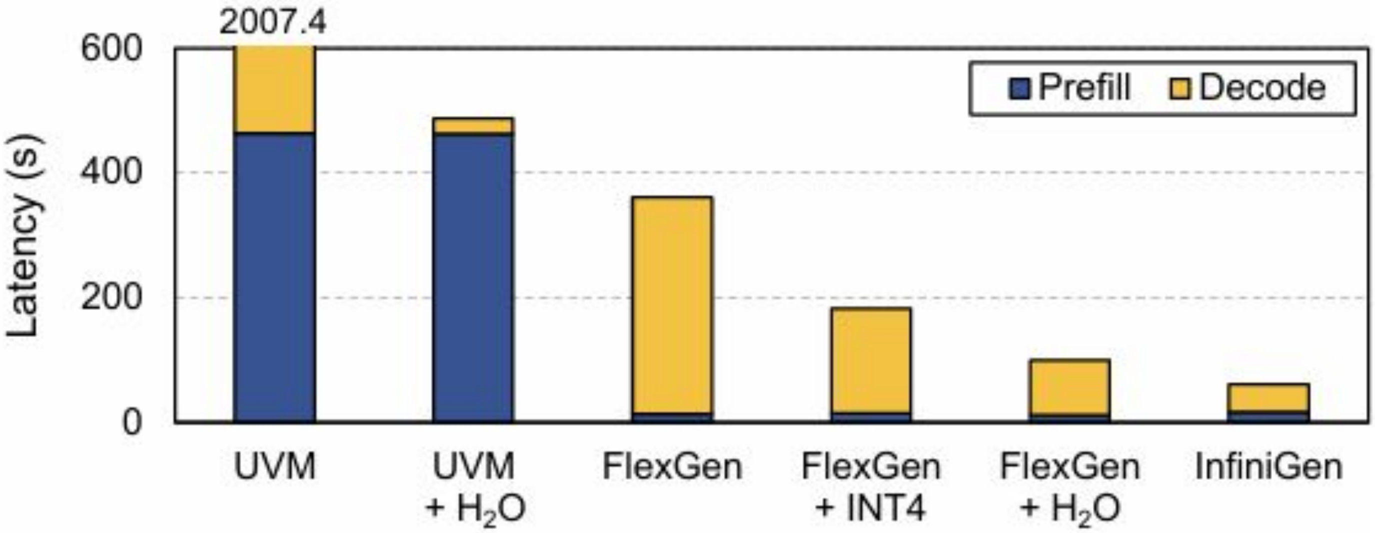
Model Evaluation

Model
- OPT
- Llama-2

System
- CUDA UVM
- FlexGen

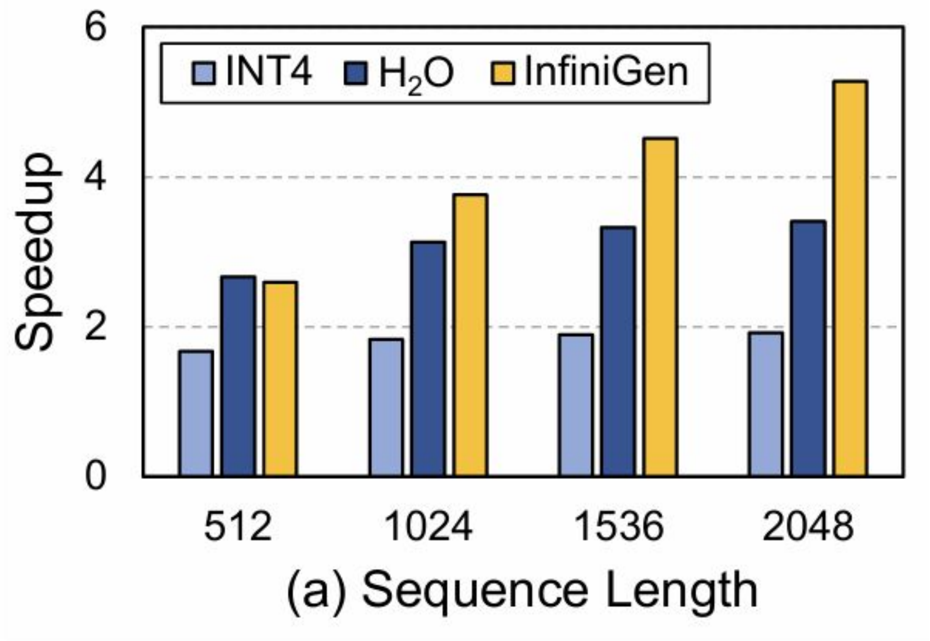
Method
- Quantization
- H2O

Evaluation - Performance



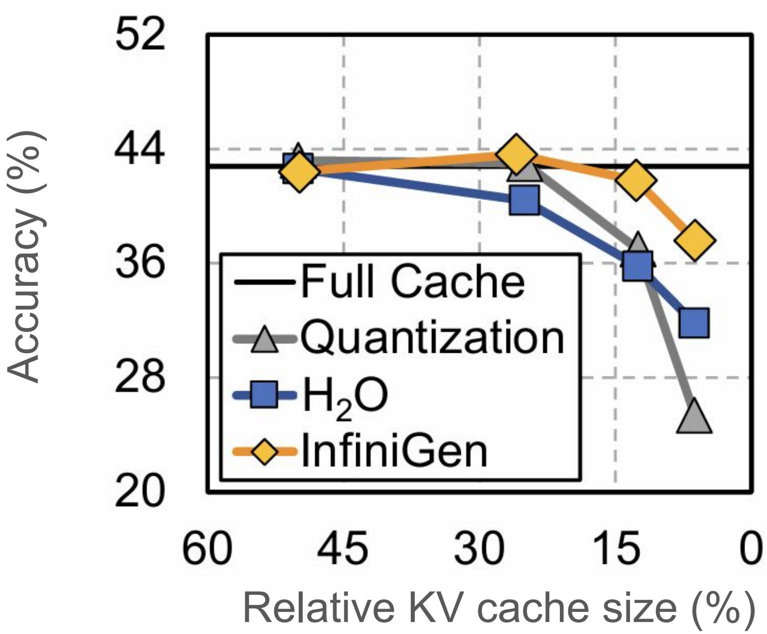
InfiniGen greatly improves the overall performance !

Evaluation - Performance



InfiniGen improves performance with longer sequence!

Evaluation - Accuracy



InfiniGen offers significantly better accuracy!

Conclusion

Problem

- KV cache size

Solution

- Speculative Prefetching
- Skewing

Result

- 3 times faster while preserving accuracy

Thank you

for listening