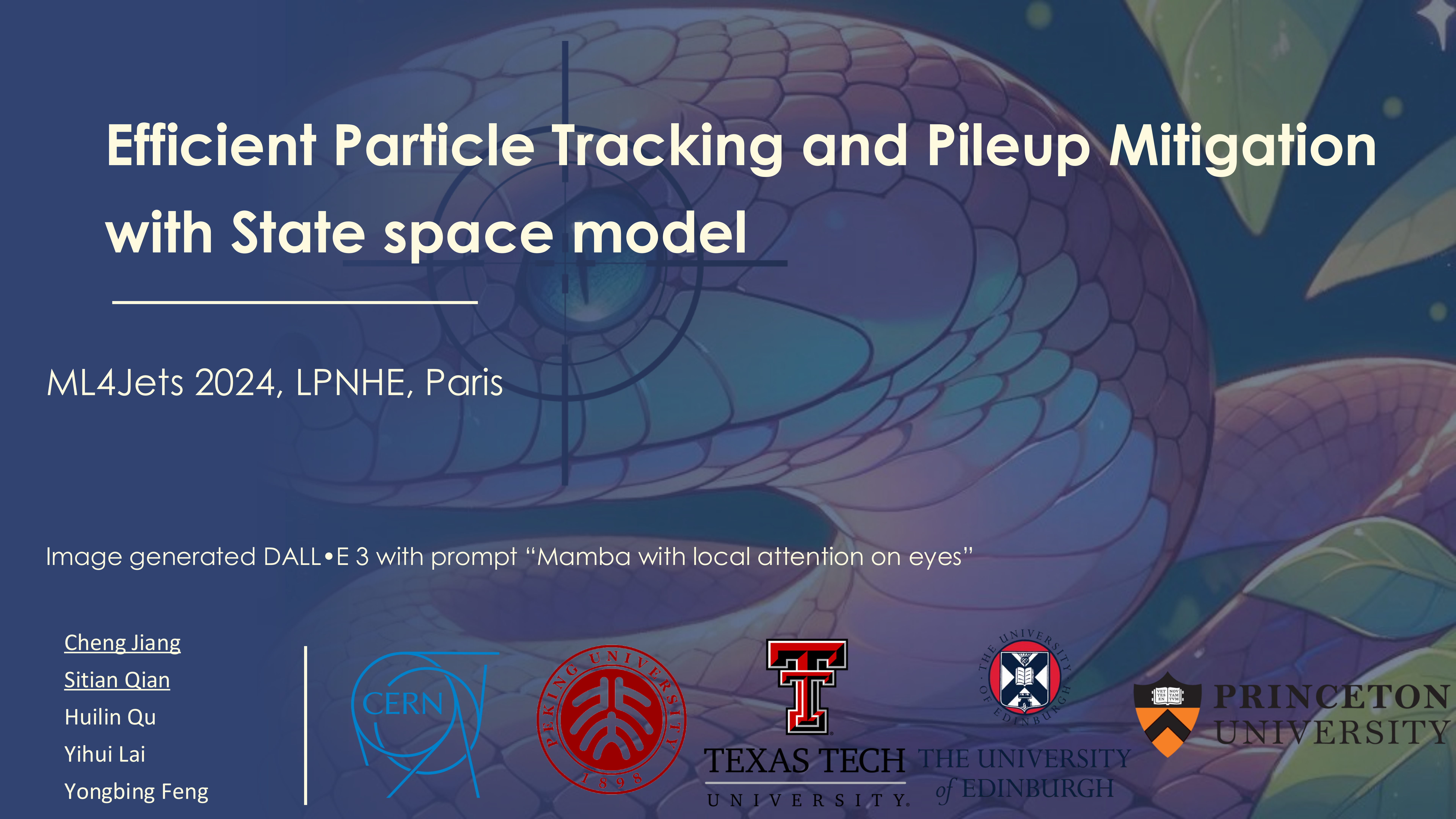


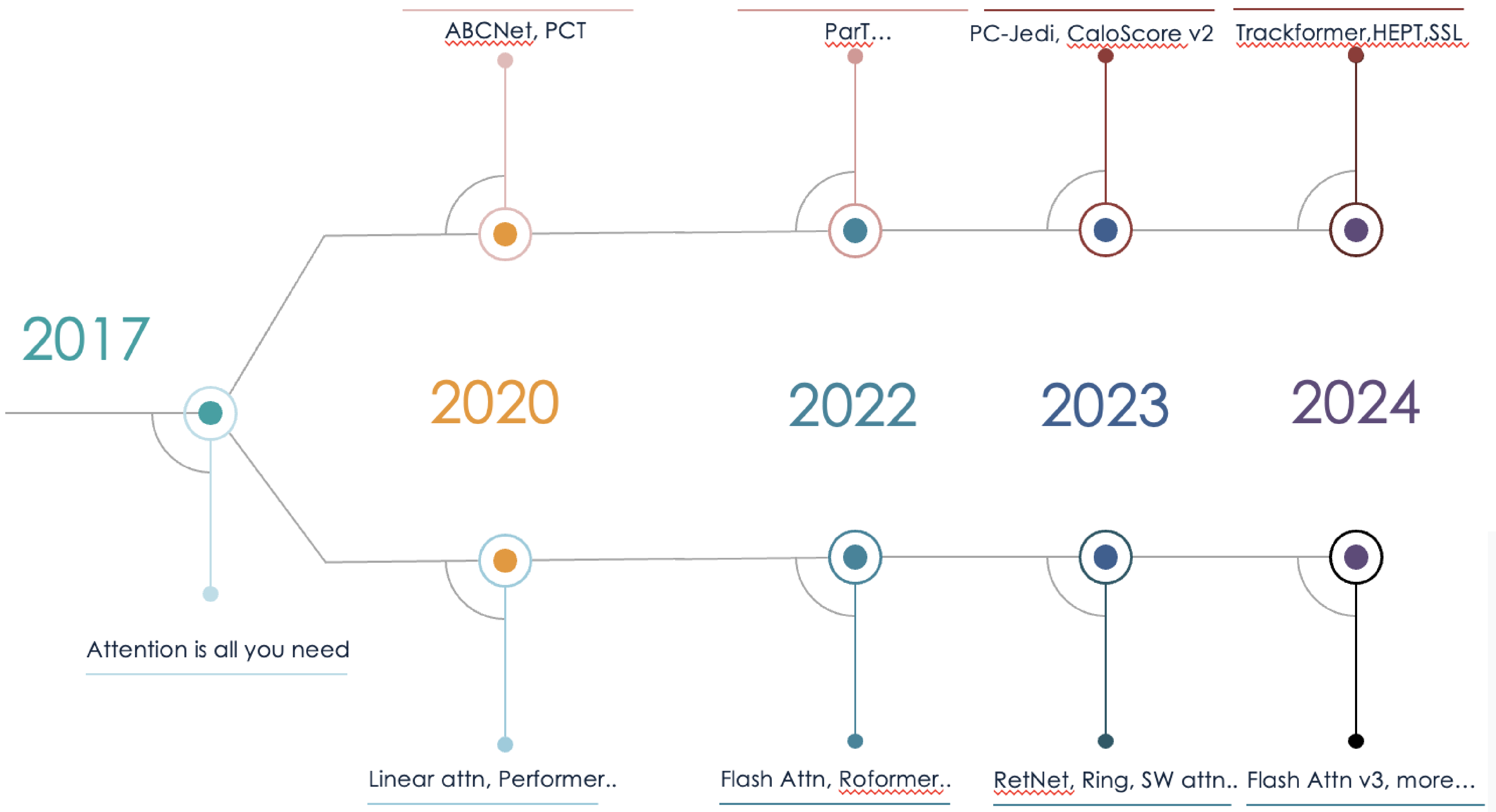
Efficient Particle Tracking and Pileup Mitigation with State space model

ML4Jets 2024, LPNHE, Paris

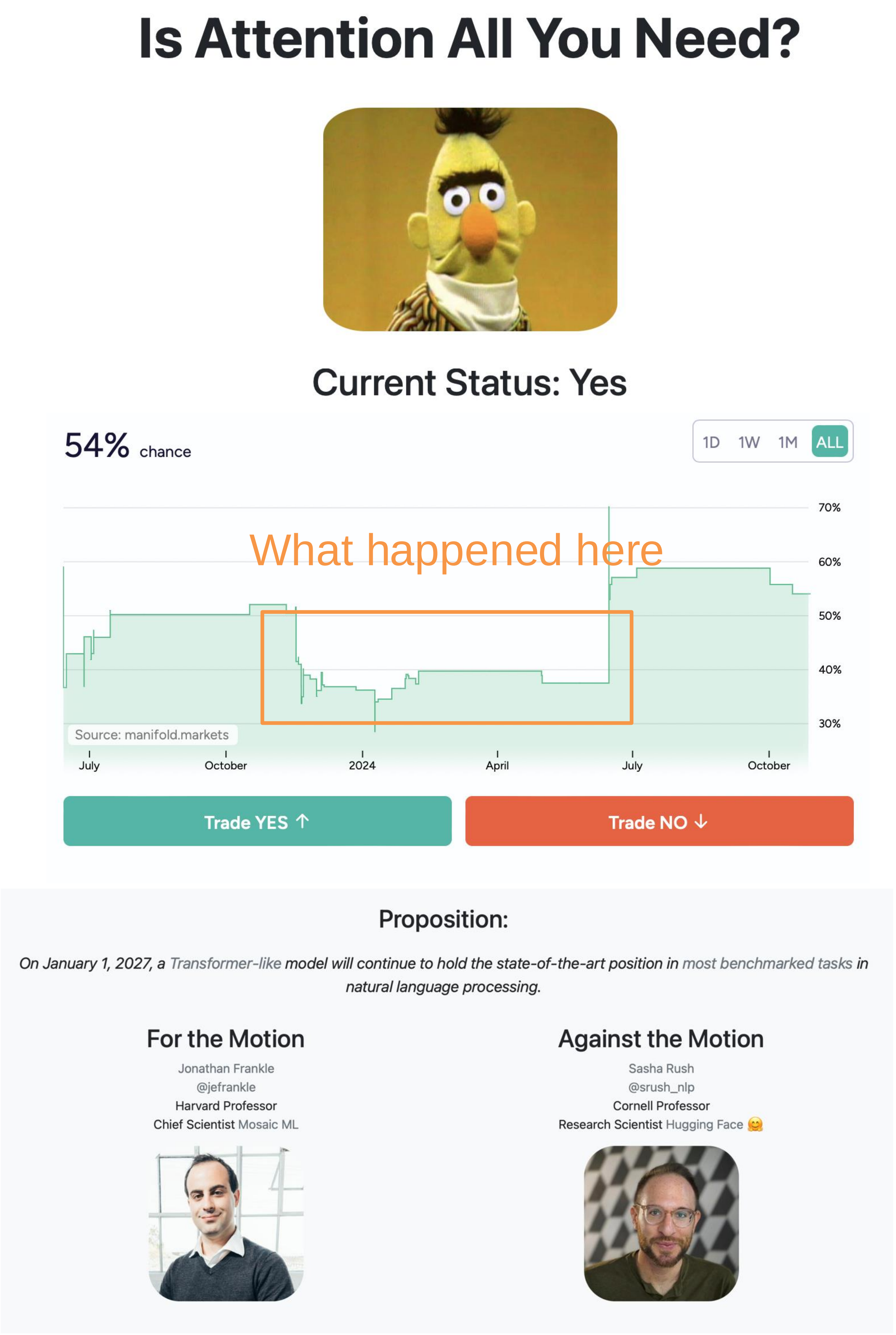
Image generated DALL•E 3 with prompt “Mamba with local attention on eyes”

Cheng Jiang
Sitian Qian
Huilin Qu
Yihui Lai
Yongbing Feng

How the attention starts



https://www.isattentionallyouneed.com



Is Attention All You Need?

Current Status: Yes

Proposition:

On January 1, 2027, a Transformer-like model will continue to hold the state-of-the-art position in most benchmarked tasks in natural language processing.

For the Motion

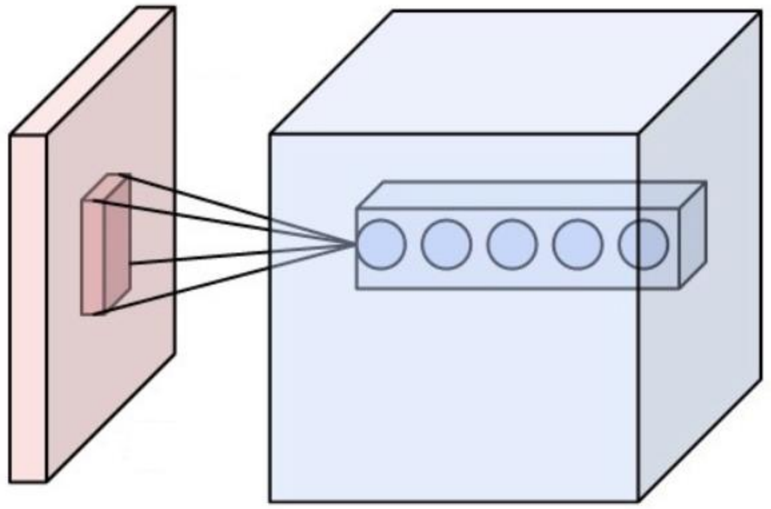
Jonathan Frankle
@jefrankle
Harvard Professor
Chief Scientist Mosaic ML

Against the Motion

Sasha Rush
@srush_nlp
Cornell Professor
Research Scientist Hugging Face 🤗

Different types

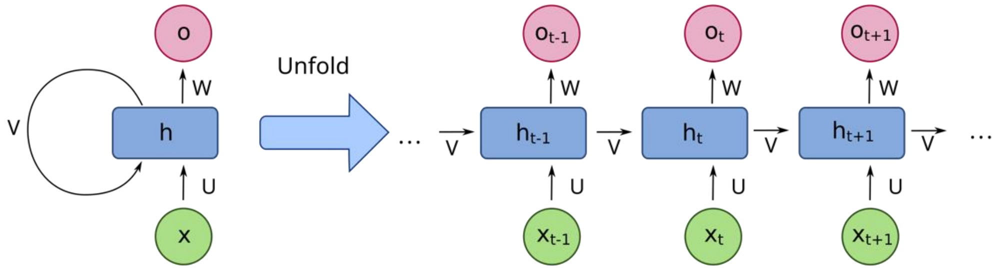
CNN



- Window size/speed depends on kernel.
- Easily parallelizable.

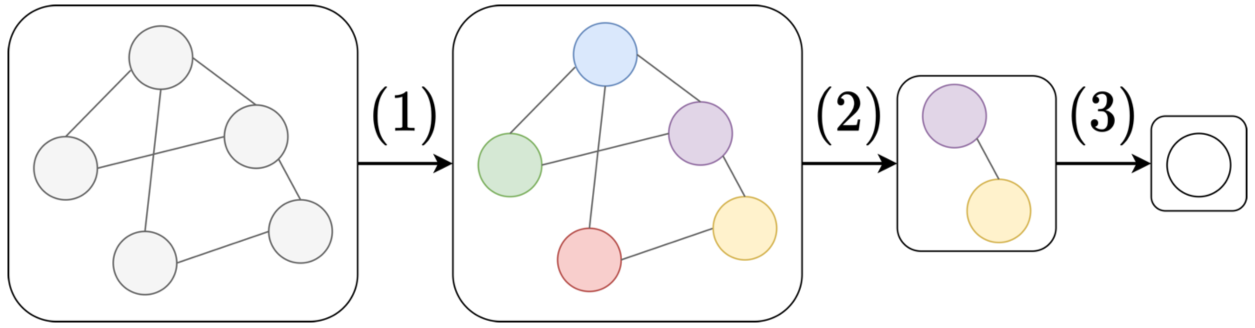
Source: Wikipedia

RNN



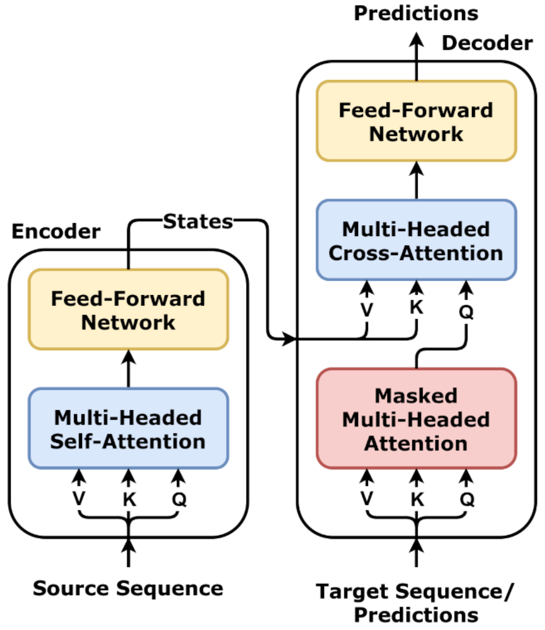
- Infinite window but compressed state (poor long-range dep).
- Fast inference.
- Not parallelizable (slow training).

GNN

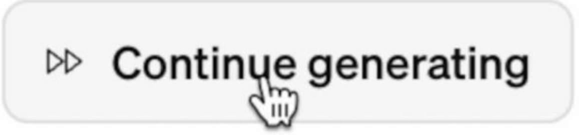


- Good performance (IFF?) finding the best data reps/feature connections.
- Slow inference/training depends on how the graph constructed.

Transformer


Predictions
Decoder
Feed-Forward Network
Multi-Headed Cross-Attention
V K Q
Masked Multi-Headed Attention
V K Q
Target Sequence/ Predictions
Encoder
States
Feed-Forward Network
Multi-Headed Self-Attention
V K Q
Source Sequence

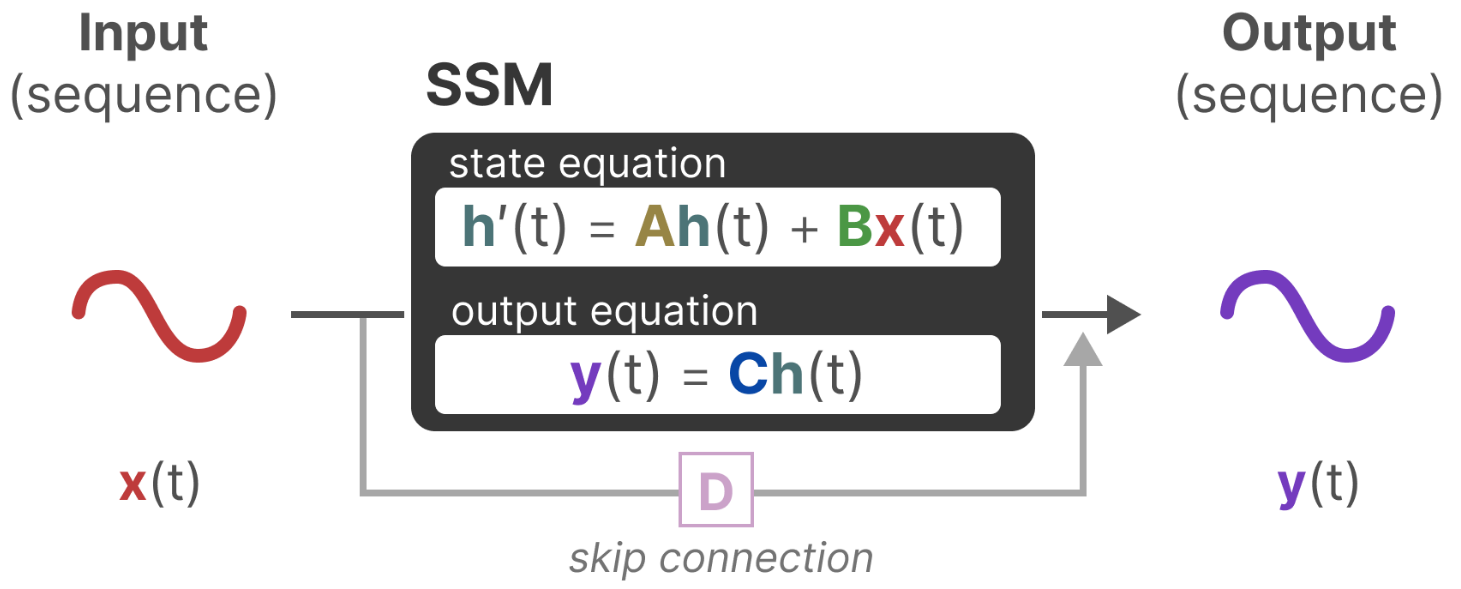
- Parallelizable training.
- Good scalability.
- Slow inference.
- Finite context window.

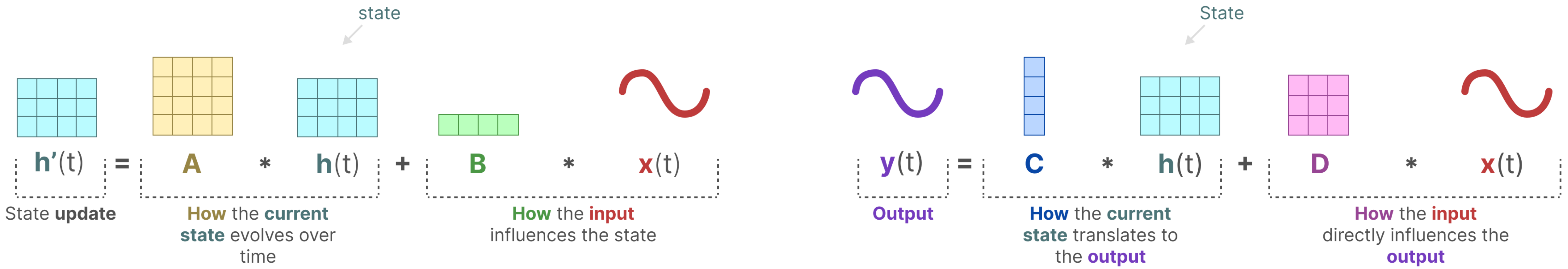


State space model

- The concept proposed by Kalman (the Kalman filter we know) in 60s.

Traditional SSM:




state
h’(t) = A * h(t) + B * x(t)
State update
How the current state evolves over time
How the input influences the state
State
y(t) = C * h(t) + D * x(t)
Output
How the current state translates to the output
How the input directly influences the output

Source: https://newsletter.maartengrootendorst.com/p/a-visual-guide-to-mamba-and-state

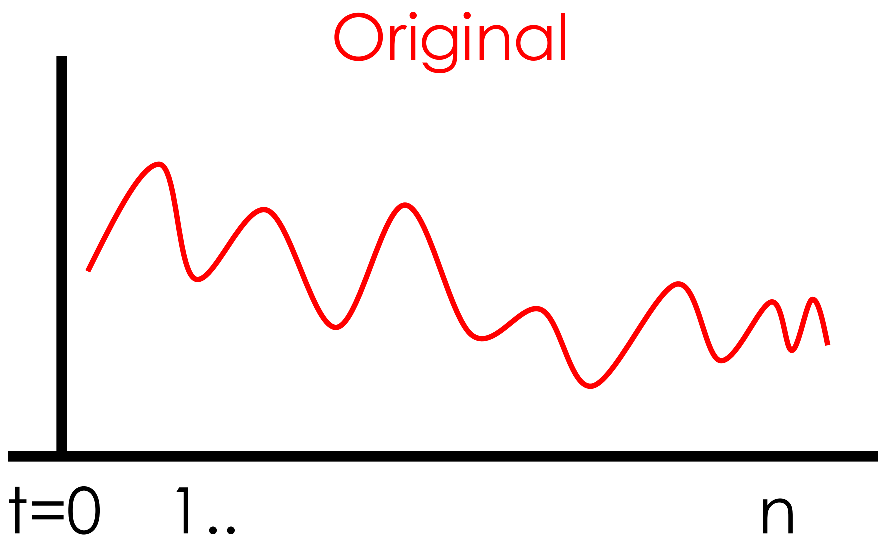
→ Structured SSM (S4)

- Combine the pros of continuous SSM (irregular sampling) + RNN (fast inference/ long window) + CNN (local info/ fast training)

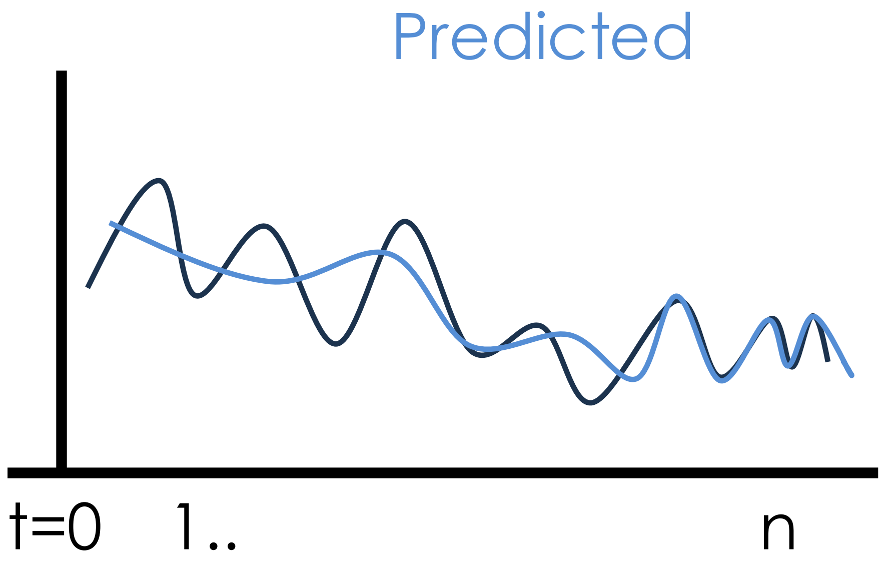
- Mathematical details can refer to Albert Gu's 300+ page thesis

- Main breakthrough:

1. Discretization for viewing model in either CNN/RNN mode,
2. **Hi**gh-order **P**olynomial **P**rojection **O**perator (HiPPO) for long-range dependency



If think exponential moving average as low-order approx, HiPPO doing it in high order by tracking the coeff of Legendre Polynomials



→ Structured SSM (S4)

- S4 gives surprisingly good performance in 2022, especially at very long sequence.

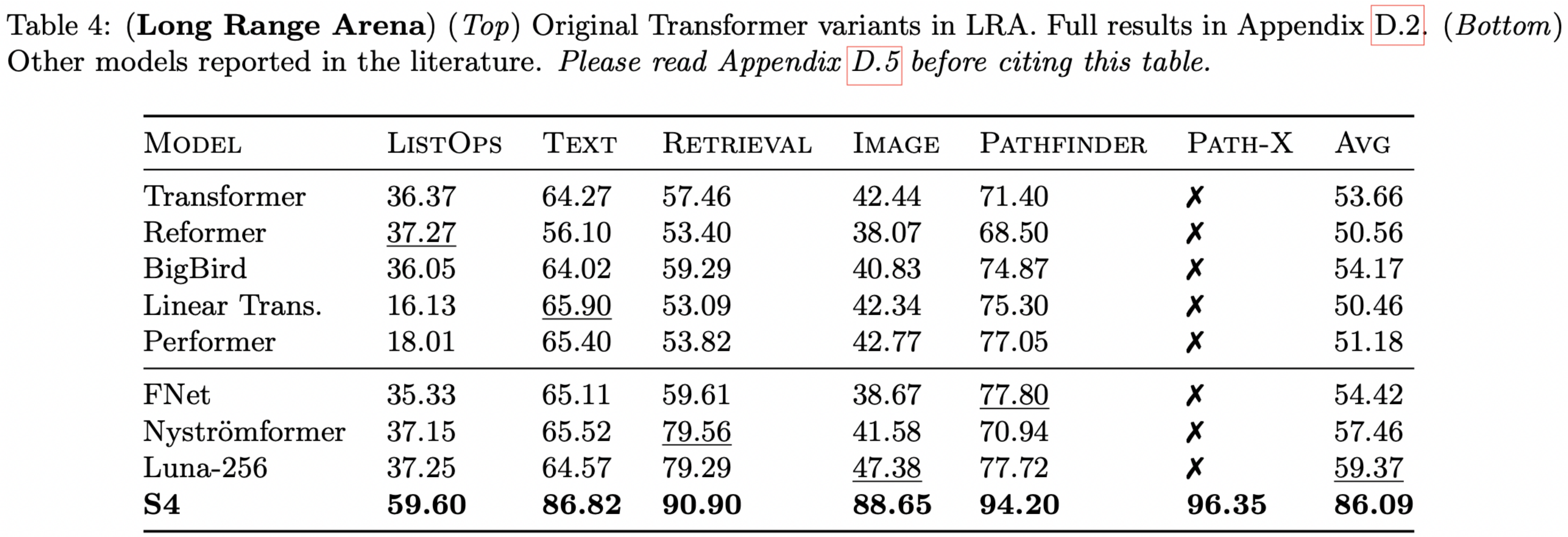


Table 4: (**Long Range Arena**) (*Top*) Original Transformer variants in LRA. Full results in Appendix D.2. (*Bottom*) Other models reported in the literature. *Please read Appendix D.5 before citing this table.*

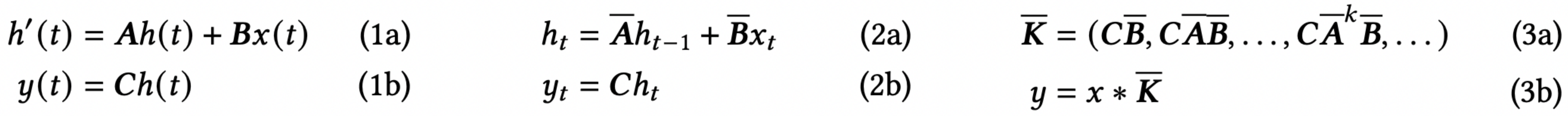
Model	ListOps	Text	Retrieval	Image	Pathfinder	Path-X	Avg
Transformer	36.37	64.27	57.46	42.44	71.40	✗	53.66
Reformer	37.27	56.10	53.40	38.07	68.50	✗	50.56
BigBird	36.05	64.02	59.29	40.83	74.87	✗	54.17
Linear Trans.	16.13	65.90	53.09	42.34	75.30	✗	50.46
Performer	18.01	65.40	53.82	42.77	77.05	✗	51.18
FNet	35.33	65.11	59.61	38.67	77.80	✗	54.42
Nyströmformer	37.15	65.52	79.56	41.58	70.94	✗	57.46
Luna-256	37.25	64.57	79.29	47.38	77.72	✗	59.37
S4	**59.60**	**86.82**	**90.90**	**88.65**	**94.20**	**96.35**	**86.09**

Source: s4 paper

- But.. In both Recurrent/Convolutional view, A,B,C is time invariant (input-free).

Recurrent view

Convolutional view



$$h'(t) = \boldsymbol{A}h(t) + \boldsymbol{B}x(t) \quad (1a)$$
$$y(t) = \boldsymbol{C}h(t) \quad (1b)$$

$$h_t = \overline{\boldsymbol{A}}h_{t-1} + \overline{\boldsymbol{B}}x_t \quad (2a)$$
$$y_t = \boldsymbol{C}h_t \quad (2b)$$

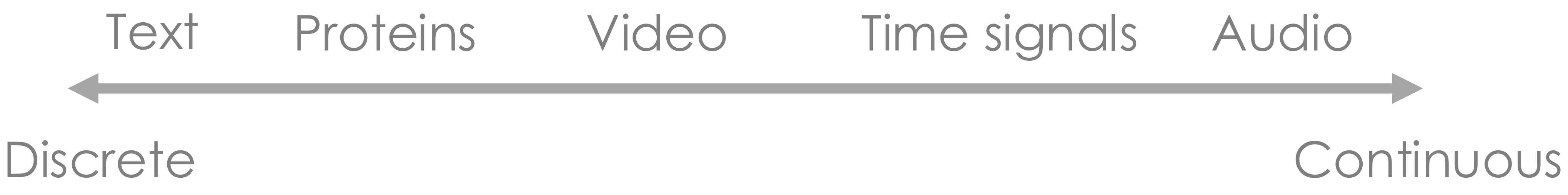
$$\overline{\boldsymbol{K}} = (\boldsymbol{C}\overline{\boldsymbol{B}}, \boldsymbol{C}\overline{\boldsymbol{A}\boldsymbol{B}}, \dots, \boldsymbol{C}\overline{\boldsymbol{A}}^k\overline{\boldsymbol{B}}, \dots) \quad (3a)$$
$$y = x * \overline{\boldsymbol{K}} \quad (3b)$$

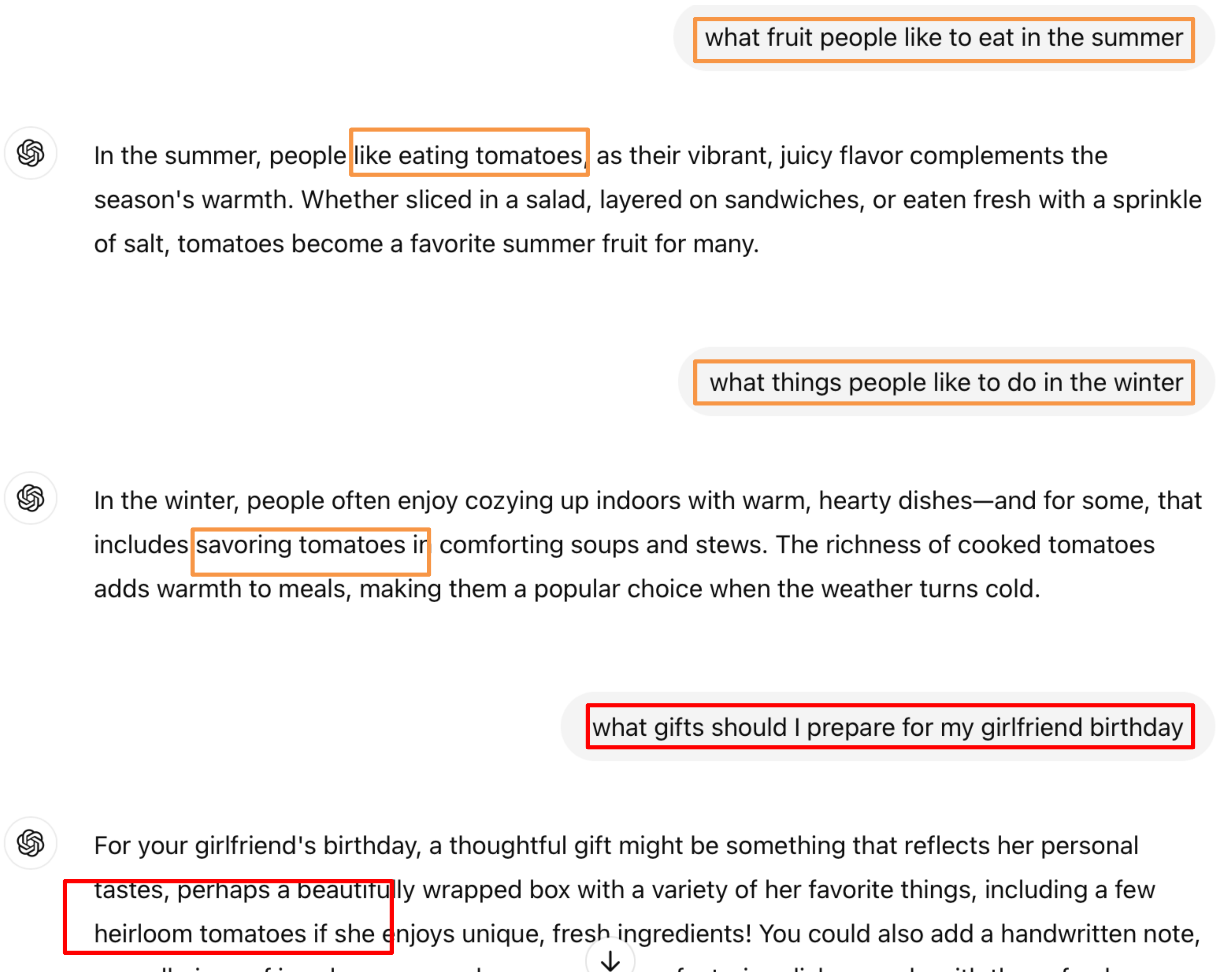
- Intrinsically would have inductive bias.

GP…S4?

- Now if we want to build a LLM with S4, what would it be like.

- Suppose my A,B,C really want the hidden state to ”eat tomatoes” (inductive bias).
- LLM need to be tokenized, quite “discrete” input, now what could improvement S4 make to generalized in less continuous data, and where is HEP data on the line below?





what fruit people like to eat in the summer

In the summer, people like eating tomatoes, as their vibrant, juicy flavor complements the season's warmth. Whether sliced in a salad, layered on sandwiches, or eaten fresh with a sprinkle of salt, tomatoes become a favorite summer fruit for many.

what things people like to do in the winter

In the winter, people often enjoy cozying up indoors with warm, hearty dishes—and for some, that includes savoring tomatoes in comforting soups and stews. The richness of cooked tomatoes adds warmth to meals, making them a popular choice when the weather turns cold.

what gifts should I prepare for my girlfriend birthday

For your girlfriend's birthday, a thoughtful gift might be something that reflects her personal tastes, perhaps a beautifully wrapped box with a variety of her favorite things, including a few heirloom tomatoes if she enjoys unique, fresh ingredients! You could also add a handwritten note,

BTW, she never liked this

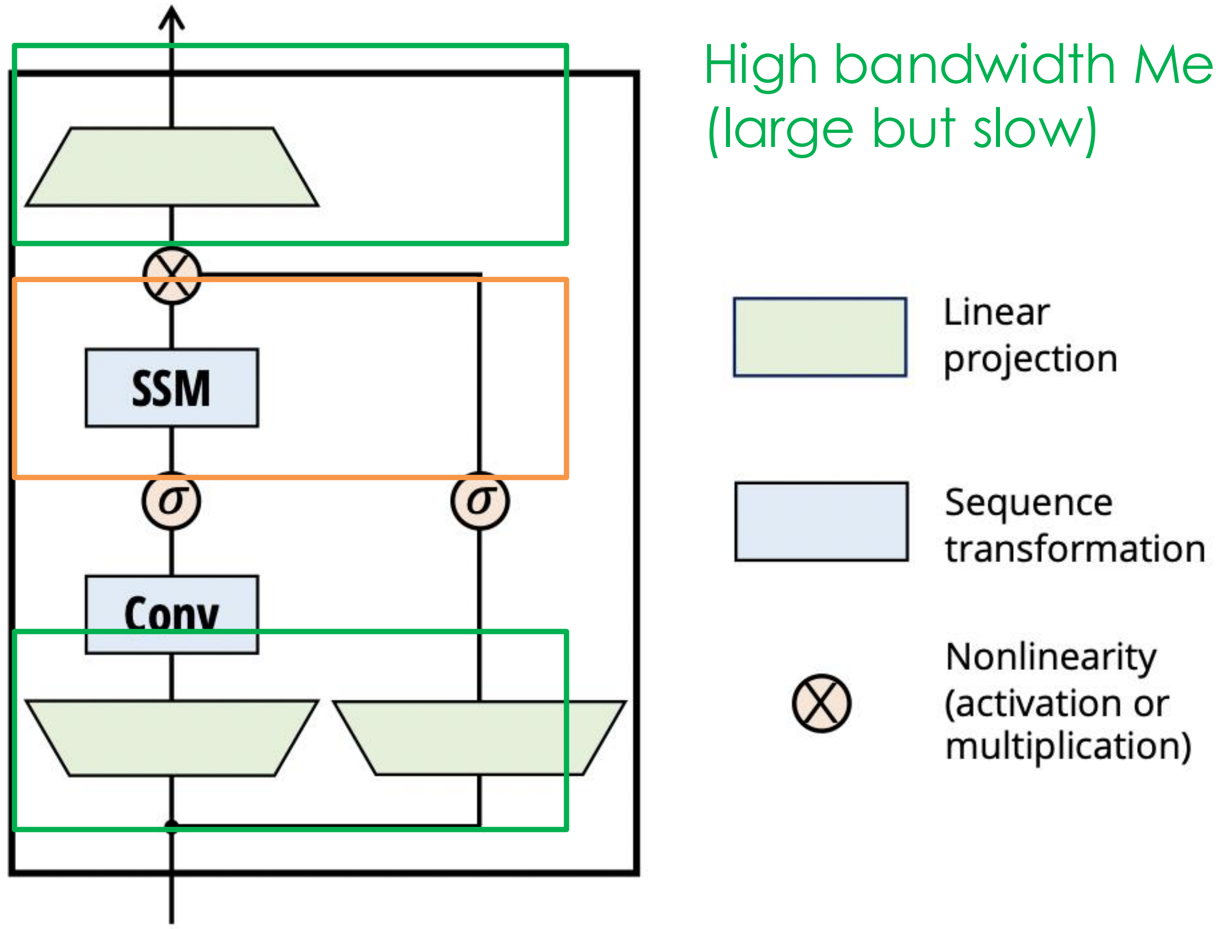
Mamba

- Or called Selective State Space Model with Hardware-aware State Expansion.. Albert Gu (SSM), Tri Dao (Flash Attn), paper

- A,B,C now depends on the input, but also means can not be precomputable (no CNN mode), need to find a way to speed up training.

1. Parallel prefix scan
2. Flash attention fashion

3. Selective Copying
(how is different with attention)



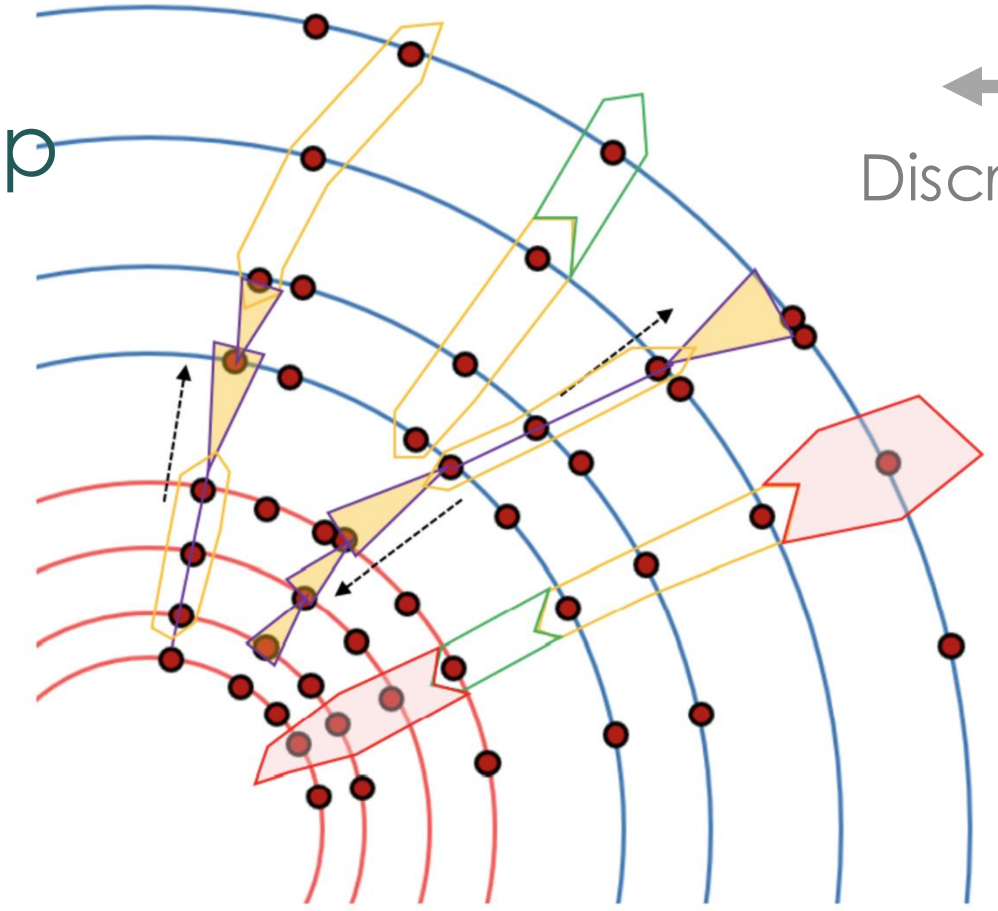
High bandwidth Memory (large but slow)

Static random access Memory (small but fast)

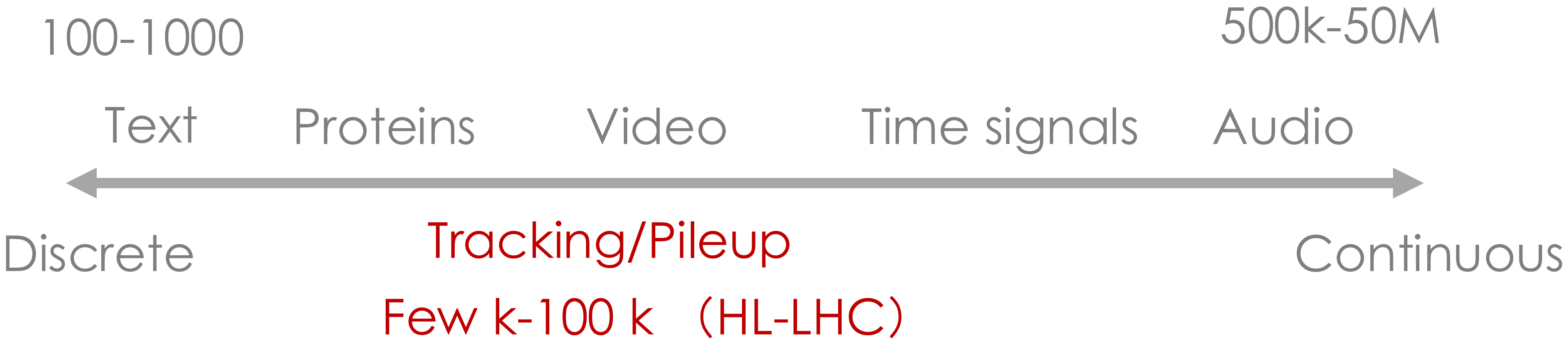
What Mamba suits

- Fast inference, high throughputs at long sequence
- Comparable performance with Transformer
- Tasks could be benefits from kinds of inductive bias

- Tracking
- Validation with Pileup Mitigation



Source: ATLAS software tutorial (Track seeding with Kalman Filter)



100-1000

500k-50M

Text Proteins Video Time signals Audio

Discrete

Tracking/Pileup

Few k-100 k （HL-LHC）

Continuous

Reference Work:

HEPT (Main)

EggNet

HGNN

GNN-OC

Training & Metrics (Tracking)

- The data-preprocessing follow the same TrackML input features and similar workflows as GNN-OC and HEPT. The loss function (contrastive predictive coding) use the same kinds as HEPT.

$$\mathcal{L}_{\text{InfoNCE}} = -\log \frac{\exp(\text{sim}(h_u, h_v^+))}{\exp(\text{sim}(h_u, h_v^+)) + \sum_{h_v^- \in \mathcal{N}} \exp(\text{sim}(h_u, h_v^-))}$$

$$\text{sim}_{L^2_{\text{RBF}}}(h_u, h_v) = \exp\left(-\frac{d_{uv}^2}{2\sigma^2}\right), \quad d_{uv} = \|h_u - h_v\|_2 = \sqrt{\sum_{k=1}^{\mathcal{N}} (h_{u,k} - h_{v,k})^2}$$

- Mainly changed the backbone.

Metrics:

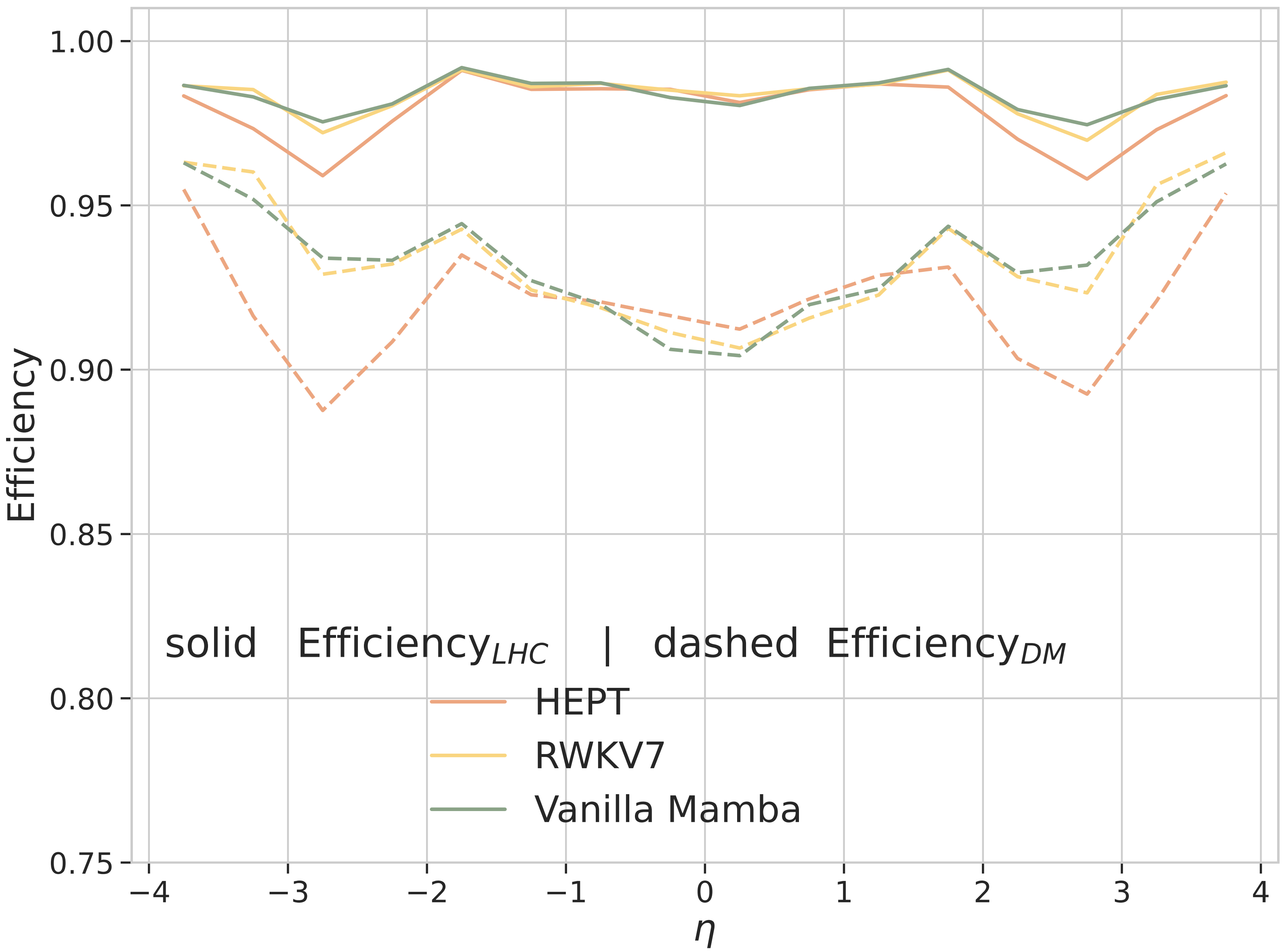
- Flops and throughputs during the inference.
- Double-Majority (DM) and LHC-style efficiency:

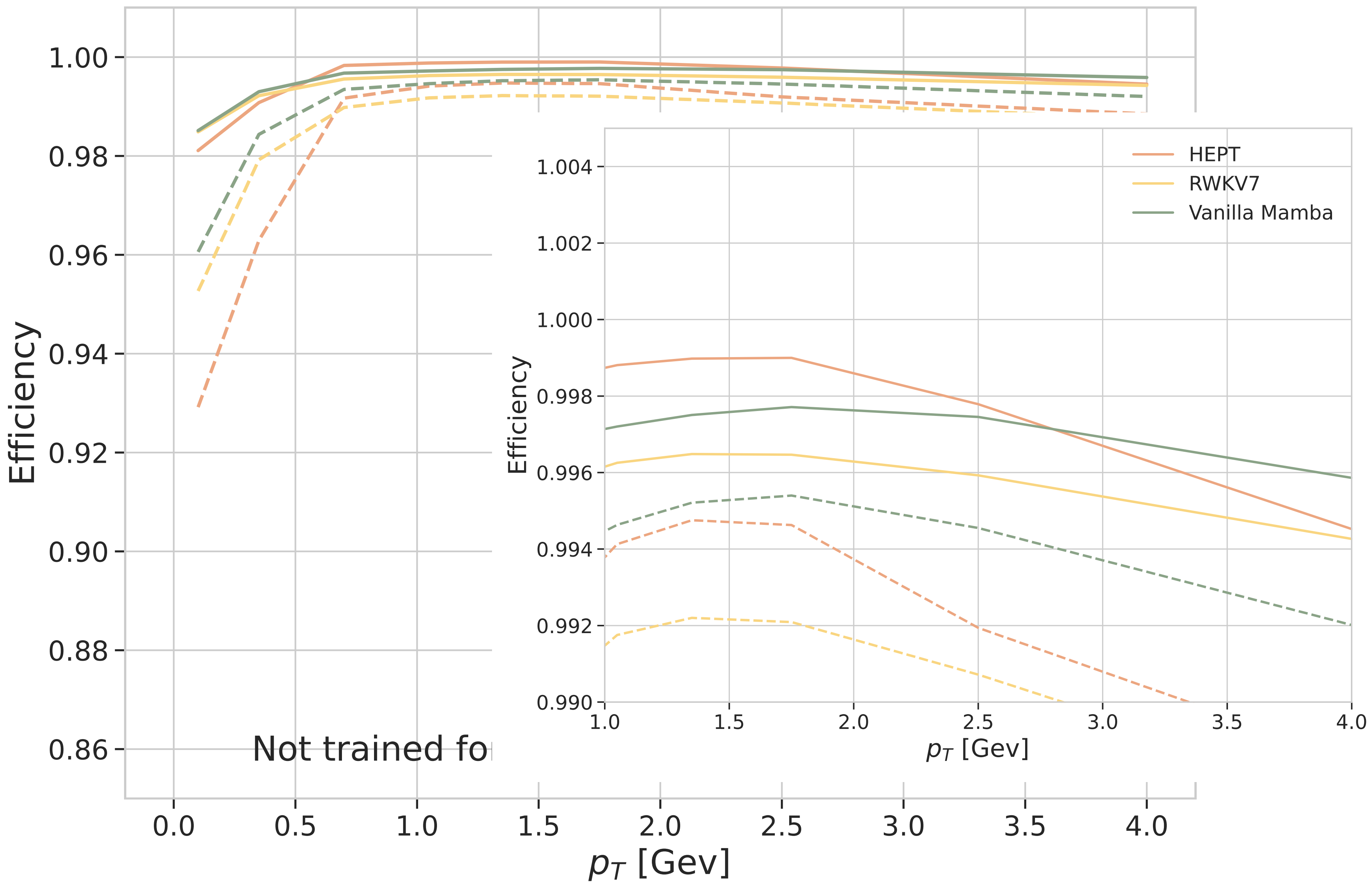
$$\#\,matched\ of\ reco\ particles \Big/ \#\,reco\ particles$$

DM: 50% hits belong to the particle, less than 50% of hits outside reco tracks

LHC: 75% hits belong to the particle

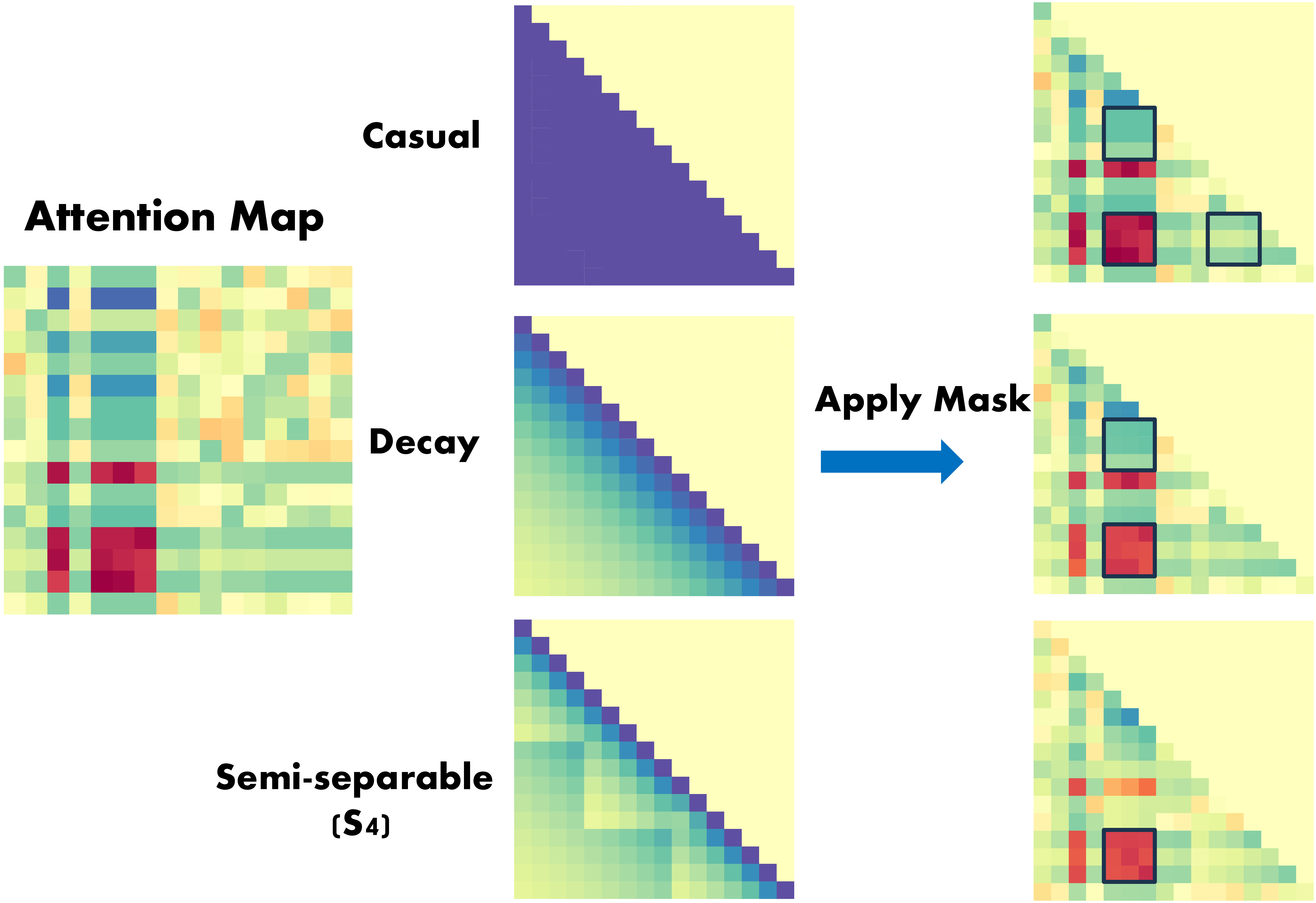
Performance (Physics)





- First trying to compare with some linear model like pure RWKV and Mamba, performed bit worse in double majority-Efficiency.
- Only limited statistics, not very smooth across all η region.

SSM ~ Attention



- Mamba 2 generalized linear attention class with structured SSM as the lower triangle mask.
- Current hybrid model perform overall better than either pure attention or pure Mamba/RWKV (SSM)
- Keys to efficient model: how small size of model can perform the same or even better with full size of full attention.

Good performance while high throughputs and low inference time

Architecture

- Want to explore how both pure Mamba model and hybrid Mamba+Transformer perform on the TrackML dataset.

- We found the HEPT (LSH-based) transformer interacts well with Mamba.

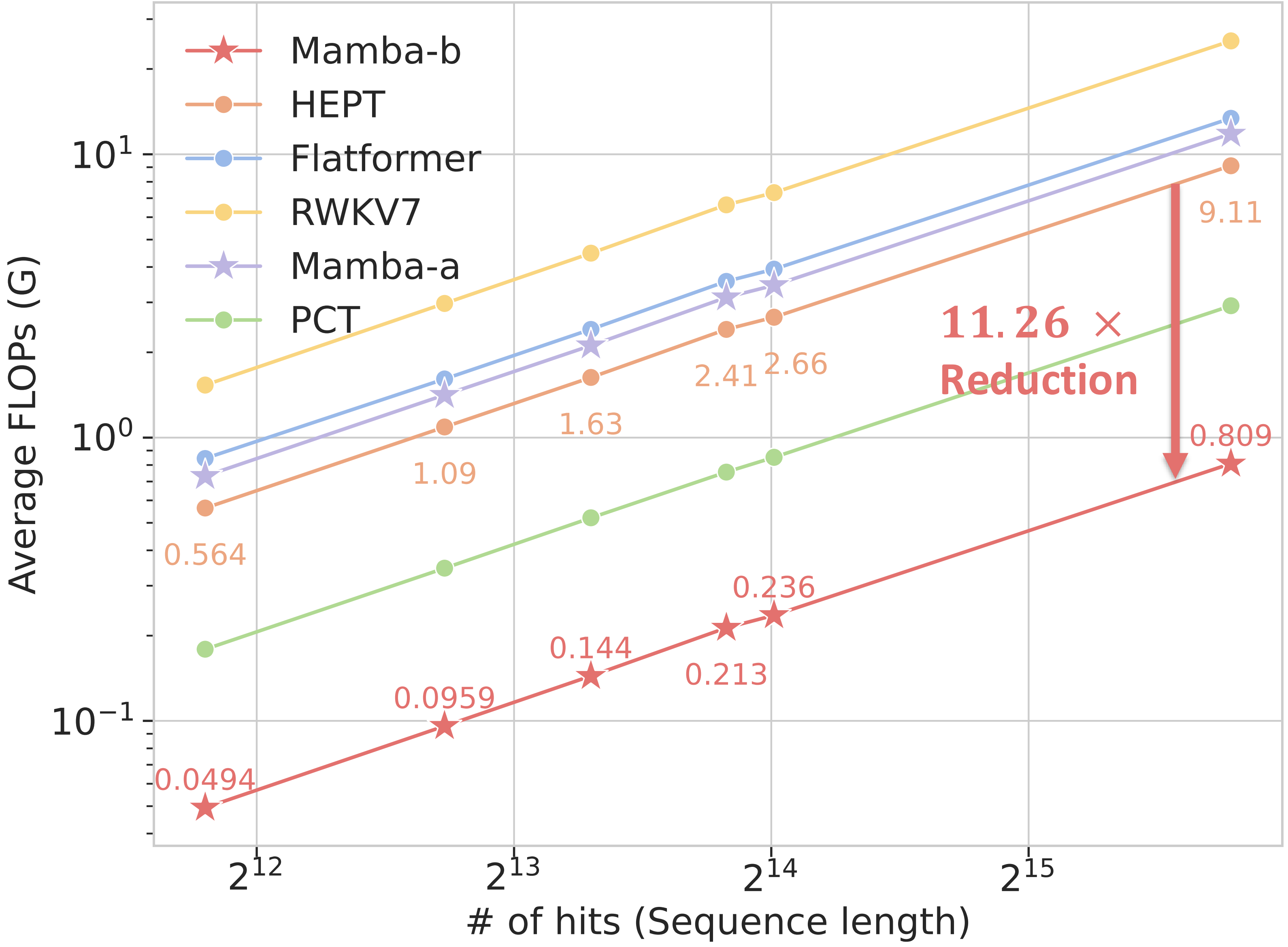
Mamba-a: Hybrid Mamba+HEPT (shared SSM+attention per blocks like Jamba)

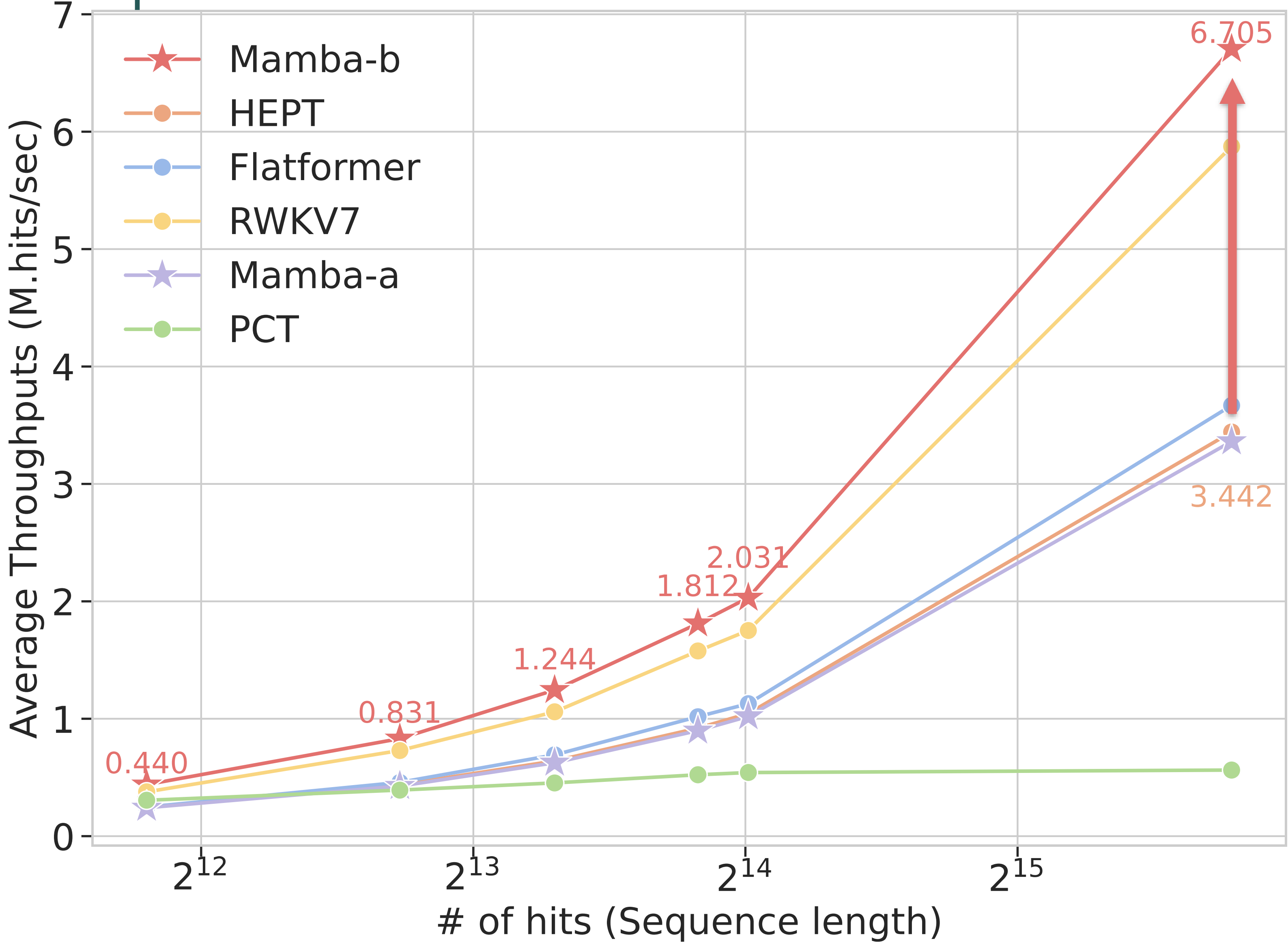
Mamba-b: Pure Mamba blocks with LSH fused before the selection mechanism.

- Since HEPT already set the benchmark for SOTA accuracy and hundreds of time speedup than traditional GNN, the performance comparison will directly compare with those effective and efficient transformer.

Performance (Inference)

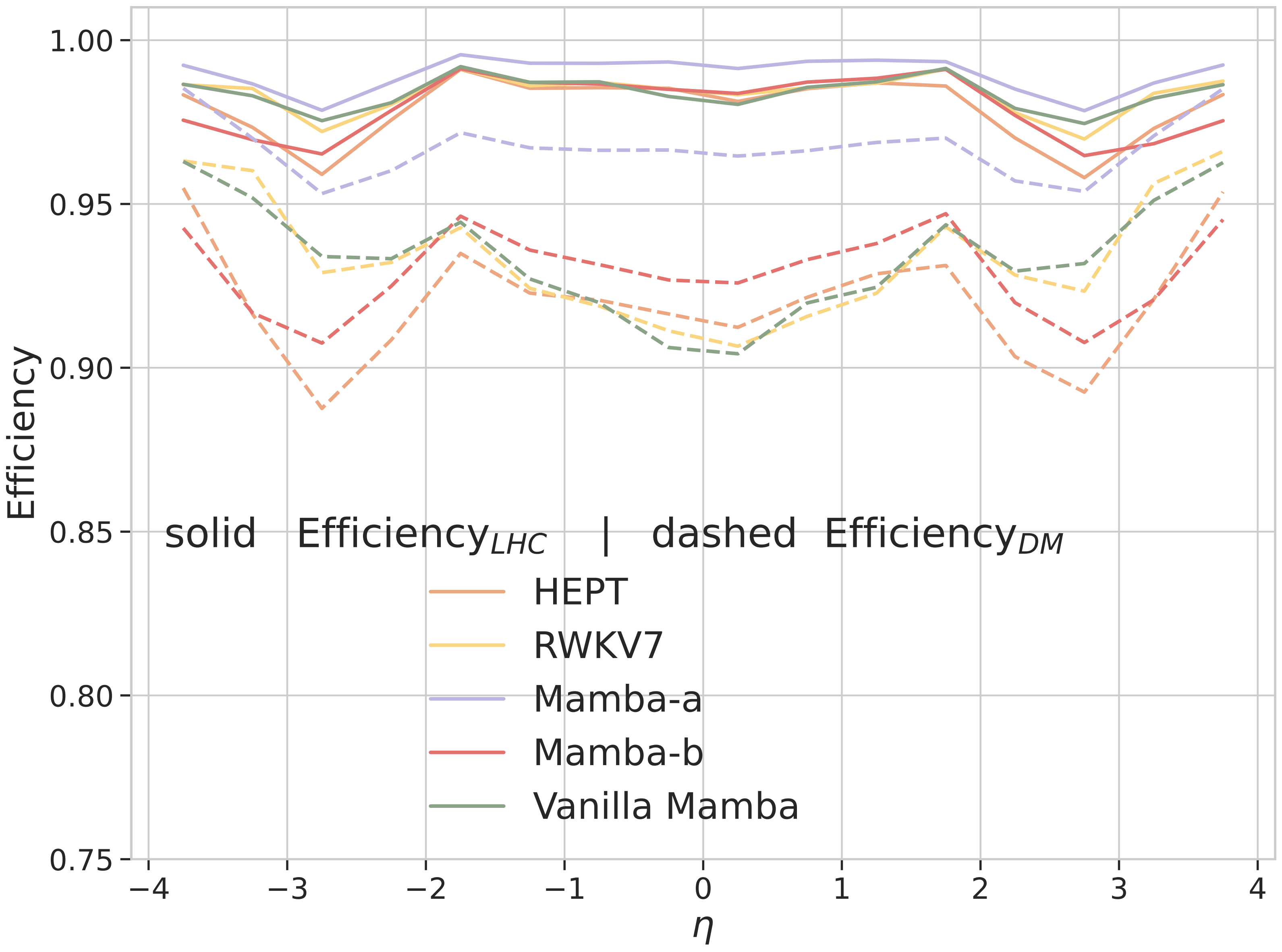
All model has ~ 0.3 M parameters

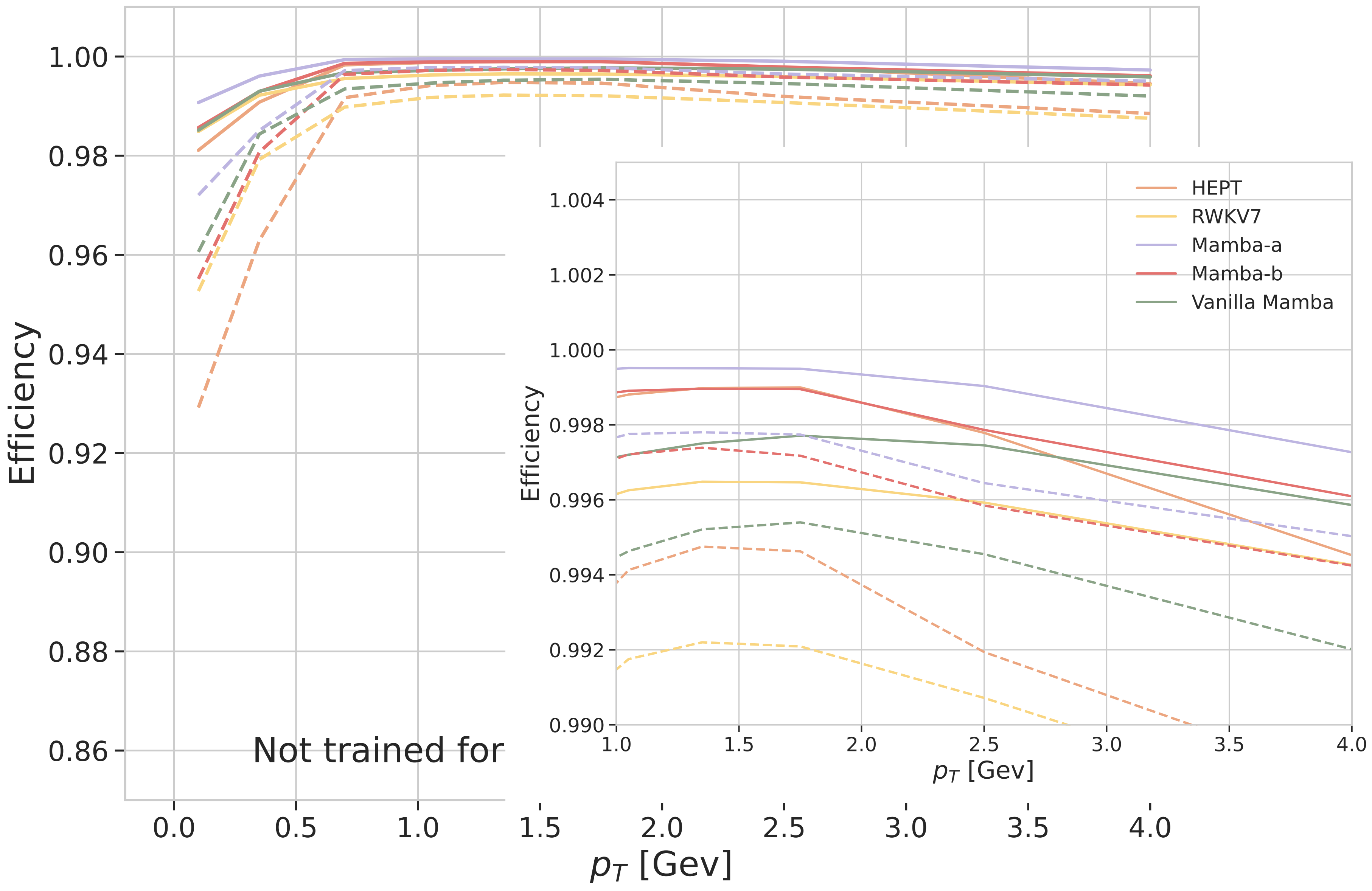




- Mamba-a has comparable performance as most RFF/LSH $O(nlogn)$ transformer, Mamba-b has almost the same scale of pure Mamba, **more than 10 times reduction in FLOPs than HEPT**. While the **raw inference time reduced by half**. Set sector to be 1,2,3,6,10,20 to test the performance under different number of tracking hits.

Performance (Physics)





- Hybrid mechanism works better over almost all trained region, especially Mamba-a, recover the efficiency in double majority.

Training & Metrics (Pileup)

- Total number of 25k (train/val/test: 20k/2k/3k) events with point cloud size 7k-10k each.
- Use Focal Loss to do binary classification among masked neutral particles.

$$\mathcal{L}_{\text{F.L.}} = -\alpha_t (1 - p_t)^{\lambda} \log(p_t)$$

Metrics:

- ROC/AUC values
- Resolution and bias for Jet pT distribution $^{(\text{reco pT}-\text{truth leading pT})}/_{\text{truth leading pT}}$

Performance (Table)

Tracking-60 k with different size of models

Binary Metrics	Accuracy	Recall
HEPT (1.01M)	**95.1**	97.5
FlatFormer (0.97M)	91.0	97.4
PCT (1.03M)	74.7	85.0
RWKV7 (0.65M)	80.1	97.3
Mamba (0.63M)	80.9	98.1
Mamba-a (0.35M)	**94.8**	97.4
Mamba-b (0.61M)	93.2	**99.4**

- HEPT has the SOTA accuracy, but recall rank is low, Most of Mamba models can have comparable performance or even better recall in smaller model size

Pileup-10k

Metrics	ROC	Bias	Resolution
HEPT (0.33M)	78.8	0.012	0.040
PCT (0.32M)	78.9	0.014	0.040
Mamba (0.33M)	78.5	0.014	0.042
Mamba-a (0.16M)	**79.4**	0.009	**0.035**
Mamba-b (0.32M)	79.1	0.009	0.039
PUPPI	70.4	0.021	0.078

- Pileup task as a validation: Most of the models have comparable performance in the similar model size, smaller hybrid Mamba-a model has the best performance with lowest bias and resolution.

Conclusion

- First trails to apply SSM on the tracking, results seems promising. Advantages especially in throughputs/FLOPs.

- The hybrid model, though not directly hold the highest throughputs, still outperform the pure Transformer and Mamba in many cases. (this hybridization is inherently superior or if there are alternative mechanisms beyond both)

- S4 was firstly introduced to tackle with long range arena, now becomes a challenger to transformer. (will the proposition in the beginning still hold yes in 2027?)

Cocktail Party

- How to hear one person’s words while everyone are talking on the party.

Image generated Leonardo AI with prompt “People on cocktail party”

- Whenever the scenario changed (new people comes, conversation changed)

RNN: we update the predictions in a compressed state (fixed scan)
GNN: we update the predictions based on the connections of each people with new one
Transformer: we update our focus at the recall time (look back at every previous with new one)
Mamba: we choose what focus and what to filter based on the previous compressed state (selectively compress).