

# Differential Transformer

Tianzhu Ye\*, Li Dong\*, Yuqing Xia\*, Yutao Sun\*, Yi Zhu, Gao Huang, Furu Wei

Microsoft Research, Tsinghua University

ICLR 2025 (Oral)

Presenter: Jinyi Luo  
2025.02.23

## Different Modalities in Data Processing:

Image



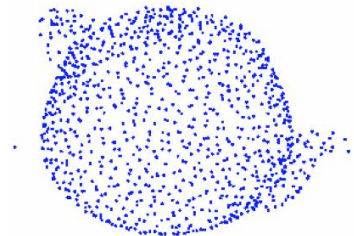
0	10	20	30	30	30	20	10
0	20	40	60	60	60	40	20
0	30	60	90	90	90	60	30
0	30	50	80	80	90	60	30
0	30	50	80	80	90	60	30
0	20	30	50	50	60	40	20
10	20	30	30	30	30	20	10
10	10	10	0	0	0	0	0

Text

When did you fall asleep?  
- I stayed up until 4 o'clock.



3-D Scene



Language: sequence processing — autoregressive models

## Challenges in Language Processing:



Long term dependency:

我的书柜太宽了，而且非常重，我没有办法把**它**搬出书房

Attention



Sophisticated meaning:

我**原以为**这部电影挺**无聊**的，**没想到**还不错



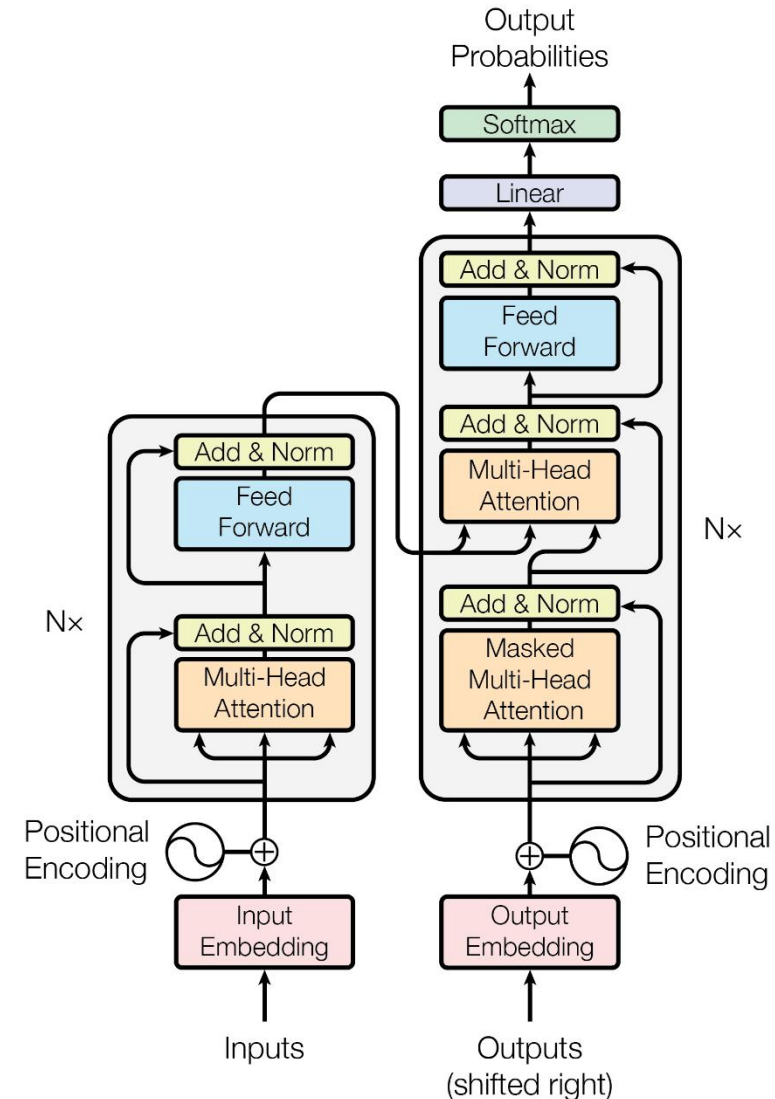
Word order matters a lot:

屡战屡败？屡败屡战！

Positional  
Encoding

## Vanilla Transformer:

- Encoder-Decoder architecture
- Multi-head attention
- Feed-Forward Network
- Residual connection  
+ Layer normalization
- Positional encoding



## Attention Mechanism:

### Hash Table:

- consists of key-value pairs
- a query arrives, search through all the keys, find the one that matches
- return the corresponding value

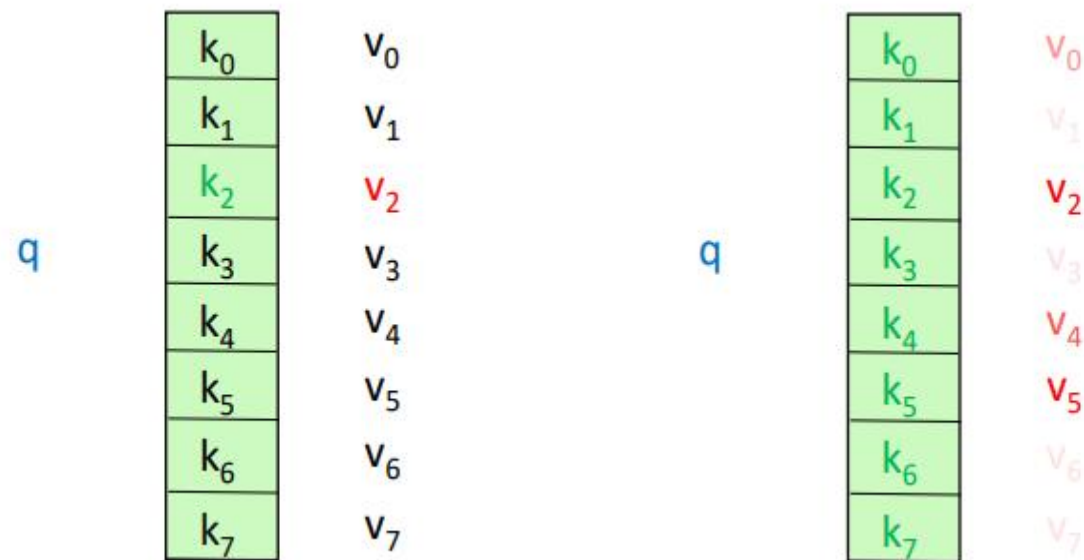
q

$k_0$	$v_0$
$k_1$	$v_1$
$k_2$	$v_2$
$k_3$	$v_3$
$k_4$	$v_4$
$k_5$	$v_5$
$k_6$	$v_6$
$k_7$	$v_7$

## Attention Mechanism:

A soft version of Hash matching:

- consists of key-value pairs
- a query arrives, calculate the correlation between the query and every key
- return the linear combination of the values



## Attention Mechanism:

Each word is represented as a vector  $x_i$

- $q_i = W^Q x_i$ ,  $k_i = W^K x_i$ ,  $v_i = W^V x_i$

word	Q	K	V	Score	Softmax
Natural	$q_1$	$k_1$	$v_1$	$\mathbf{q_1 \cdot k_1 / \sqrt{d_k}}$	$x_{11}$
Language		$k_2$	$v_2$	$\mathbf{q_1 \cdot k_2 / \sqrt{d_k}}$	$x_{12}$
Processing		$k_3$	$v_3$	$\mathbf{q_1 \cdot k_3 / \sqrt{d_k}}$	$x_{13}$

## Attention Mechanism:

Each word is represented as a vector  $x_i$

•  $q_i = W^Q x_i, k_i = W^K x_i, v_i = W^V x_i$

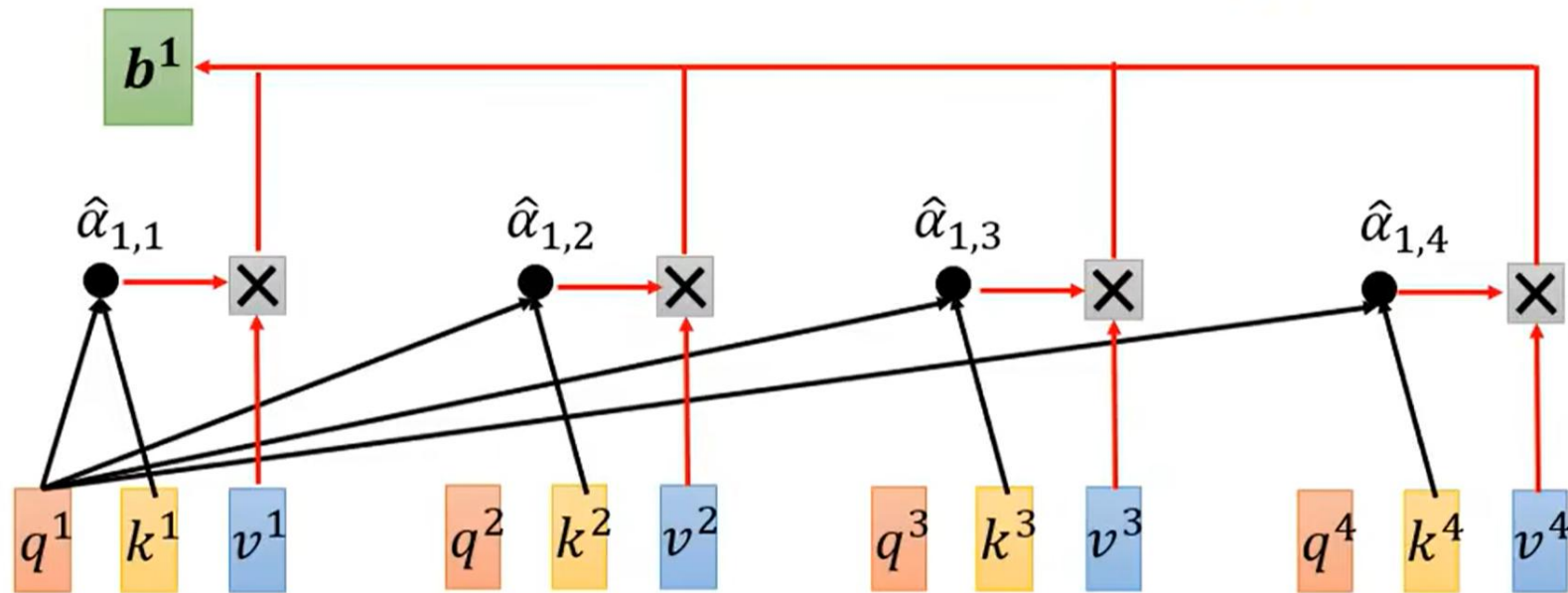
word	Q	K	V	Score	Softmax
Natural	$q_1$	$k_1$	$v_1$	$q_1 \cdot k_1 / \sqrt{d_k}$	$x_{11}$
Language		$k_2$	$v_2$	$q_1 \cdot k_2 / \sqrt{d_k}$	$x_{12}$
Processing		$k_3$	$v_3$	$q_1 \cdot k_3 / \sqrt{d_k}$	$x_{13}$

word	Q	K	V	Score	Softmax	Softmax*v	Sum
Natural	$q_1$	$k_1$	$v_1$	$q_1 \cdot k_1 / \sqrt{d_k}$	$x_{11}$	$x_{11} v_1$	$z_1$
Language		$k_2$	$v_2$	$q_1 \cdot k_2 / \sqrt{d_k}$	$x_{12}$	$x_{12} v_2$	
Processing		$k_3$	$v_3$	$q_1 \cdot k_3 / \sqrt{d_k}$	$x_{13}$	$x_{13} v_3$	



## Attention Mechanism:

$$\text{Attention}(Q, K, V) = \text{softmax} \left( \frac{QK^T}{\sqrt{d_k}} \right) V$$

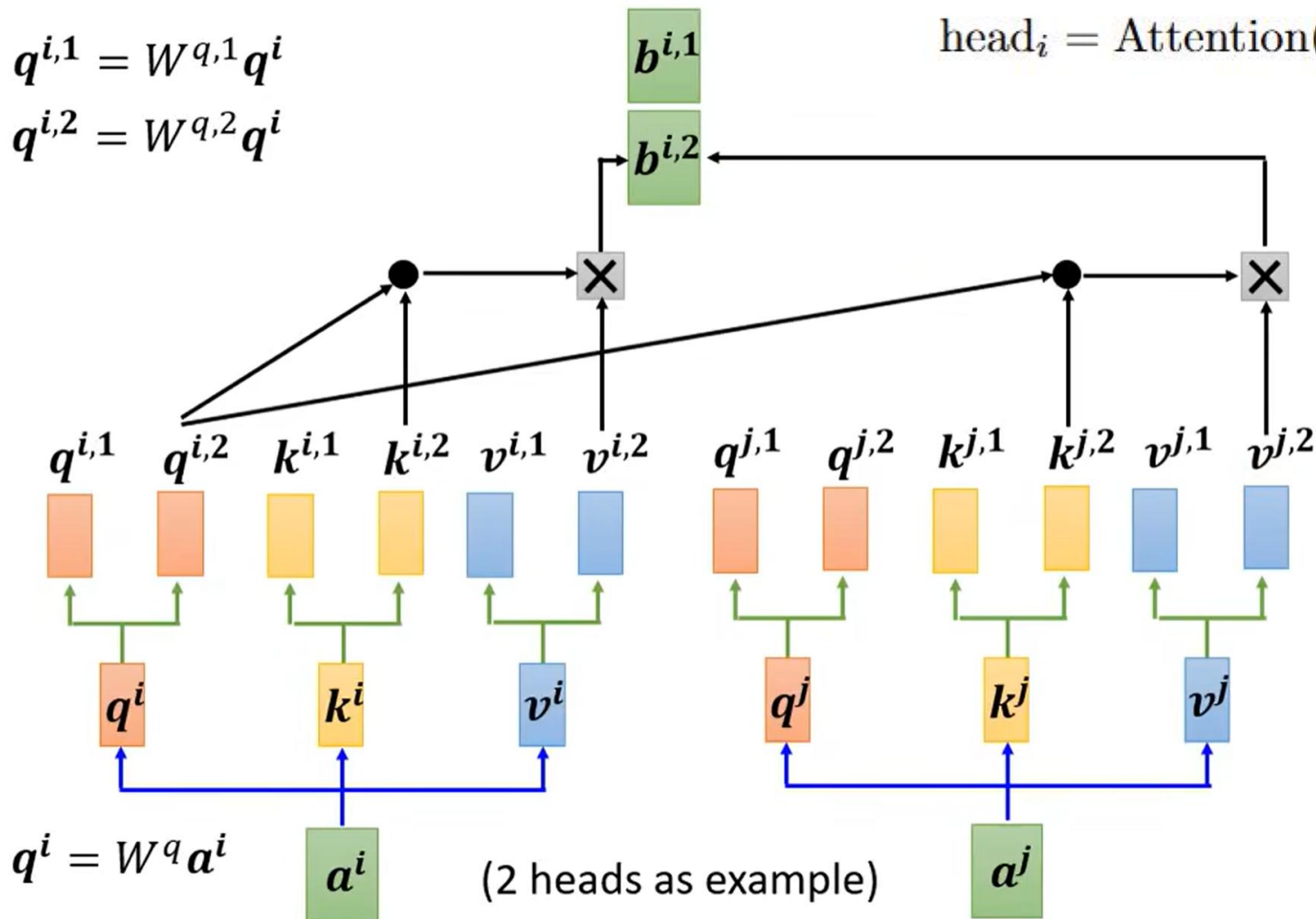


$$\begin{matrix}
 b^1 & b^2 & b^3 & b^4 \\
 \hline
 O
 \end{matrix}
 =
 \begin{matrix}
 v^1 & v^2 & v^3 & v^4 \\
 \hline
 V
 \end{matrix}
 \begin{matrix}
 \alpha'_{1,1} & \alpha'_{2,1} & \alpha'_{3,1} & \alpha'_{4,1} \\
 \alpha'_{1,2} & \alpha'_{2,2} & \alpha'_{3,2} & \alpha'_{4,2} \\
 \alpha'_{1,3} & \alpha'_{2,3} & \alpha'_{3,3} & \alpha'_{4,3} \\
 \alpha'_{1,4} & \alpha'_{2,4} & \alpha'_{3,4} & \alpha'_{4,4} \\
 \hline
 A'
 \end{matrix}$$

## Multi-head Attention

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \text{head}_2, \dots, \text{head}_h) W^O$$

$$\text{head}_i = \text{Attention}(XW_i^Q, XW_i^K, XW_i^V)$$



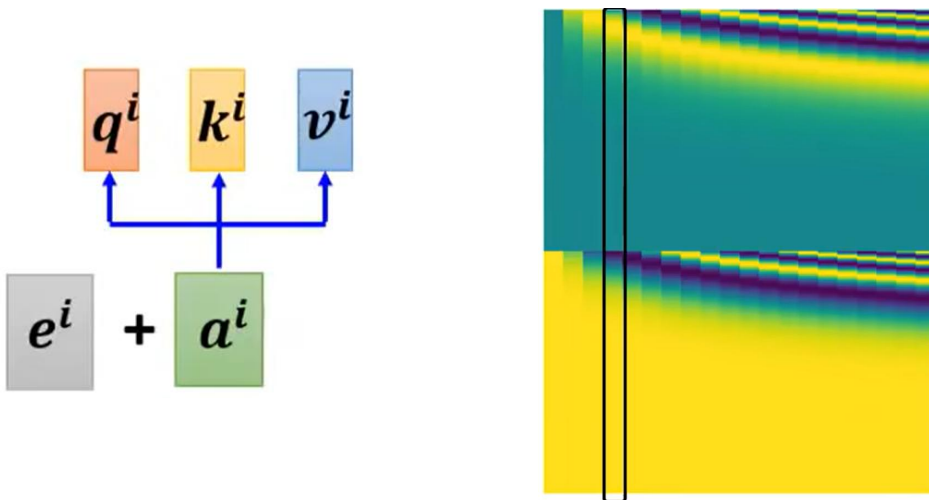
## Feed-Forward Network

$$\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2$$

## Layer Normalization

$$\mu_{ij} = \frac{1}{k} \sum_{l=1}^k X_{ijl}$$
$$\sigma_{ij}^2 = \frac{1}{k} \sum_{l=1}^k (X_{ijl} - \mu_{ij})^2$$
$$\hat{X}_{ijl} = \frac{X_{ijl} - \mu_{ij}}{\sqrt{\sigma_{ij}^2 + \epsilon}}$$

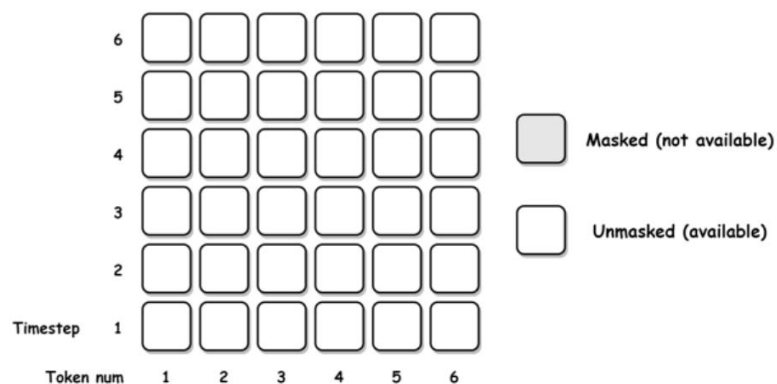
## Positional Encoding



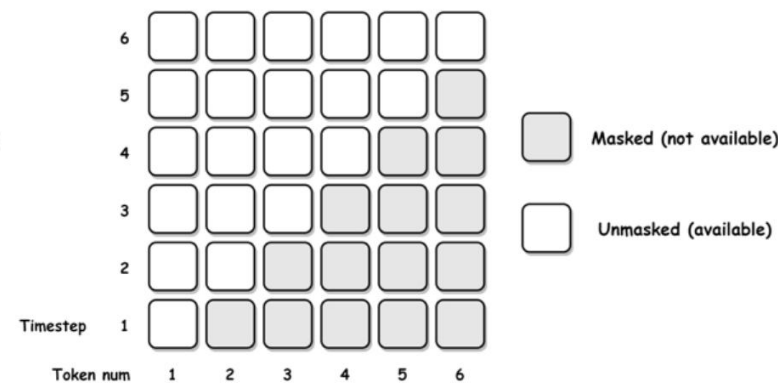
## Transformer Architectures:

- Encoder-Decoder
  - Transformer
  - T5
- Encoder-Only
  - BERT
  - RoBERTa
- Decoder-Only
  - GPTs

## Corresponding Attention Masks:

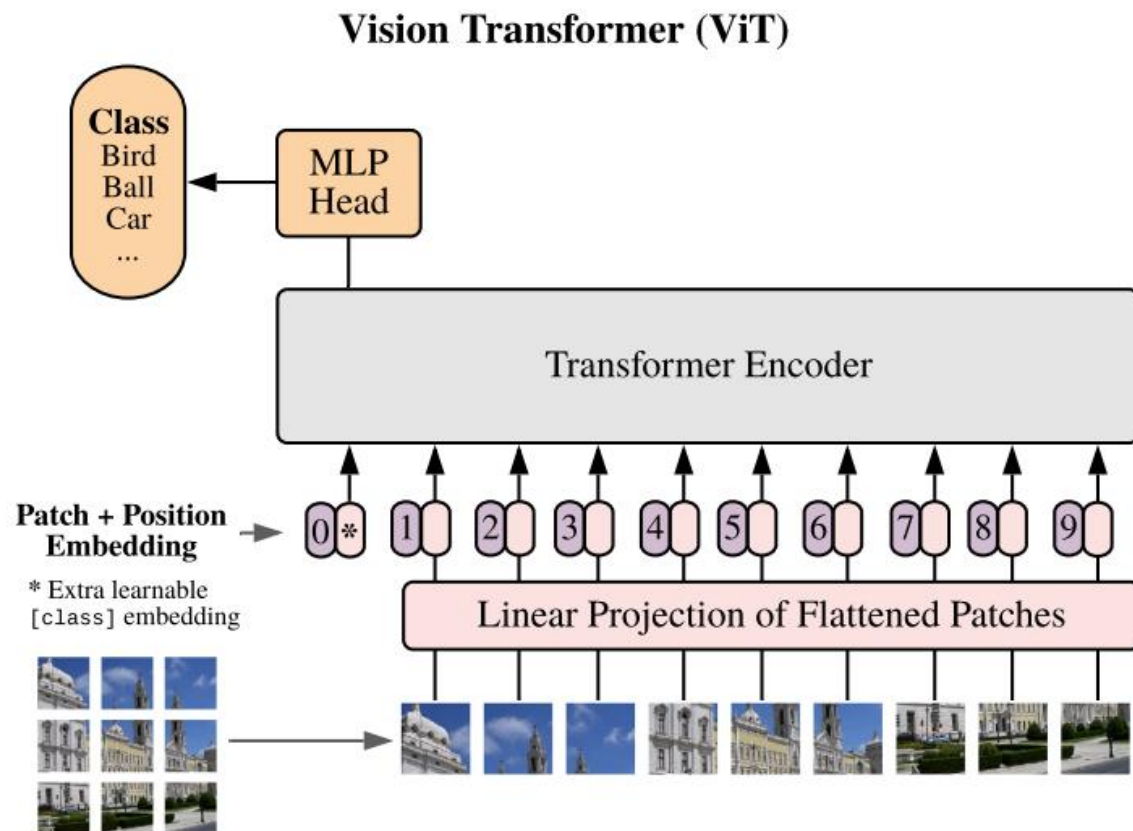


Encoder

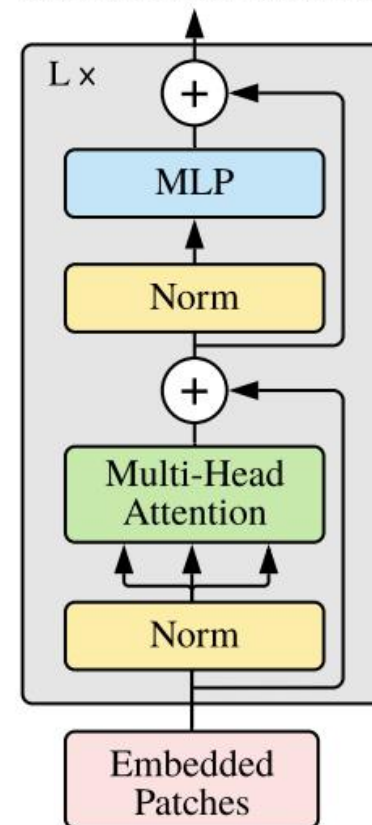


Decoder

## Vision Transformers

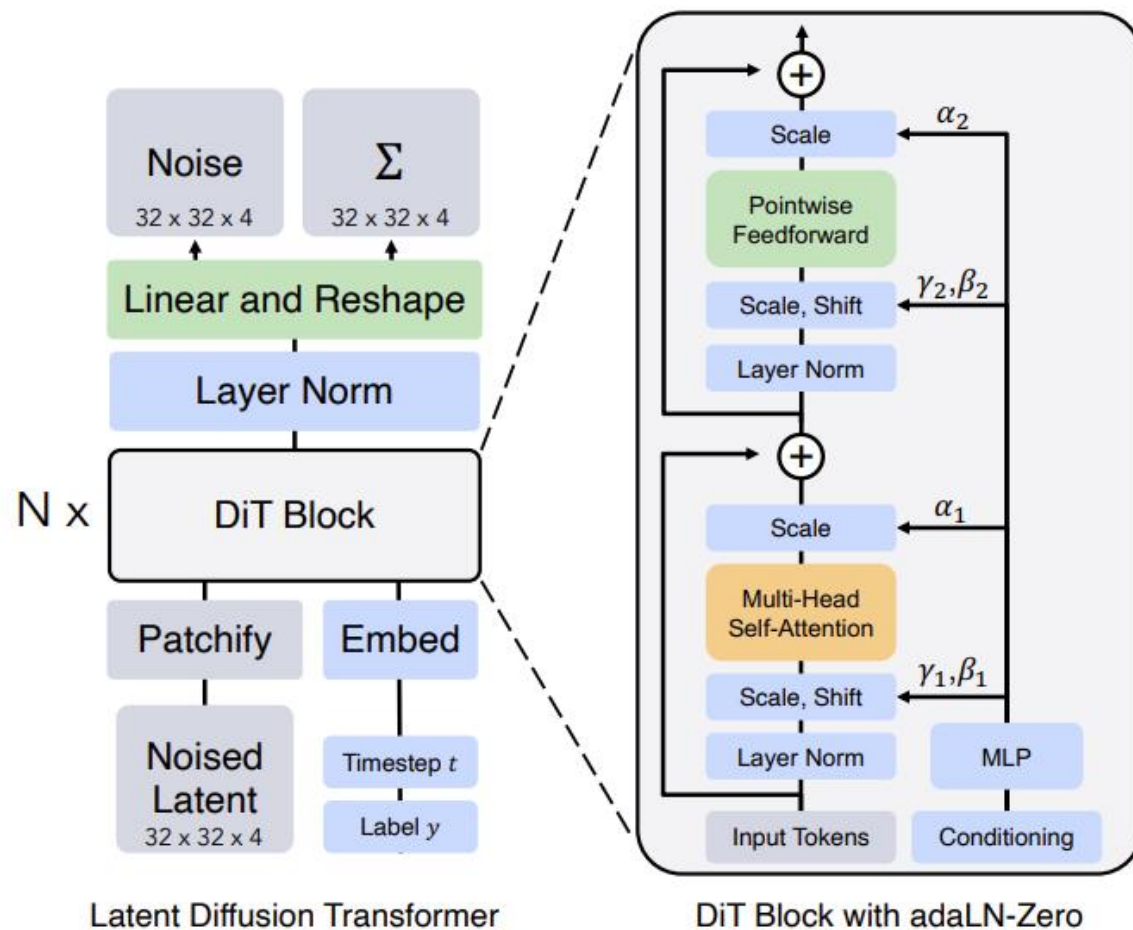


### Transformer Encoder



- Patchify images as token sequences
- Transformer encoder for classification
- Broader spatial correlation

## Diffusion Transformers



- Replace U-Net with Transformer blocks
- Long-range correlation in latent denoising



## Attention Setbacks: Irrelevant Attention

### Modified Problem

Maria buys a large bar of French soap that lasts her for 2 months. She spends \$8.00 per bar of soap. *Every 10 months, Maria's neighbor buys a new shampoo and moisturizer for Maria's neighbor.* If Maria wants to stock up for the entire year, how much will she spend on soap?

### Correct CoT Output to the Original Problem

Maria spends \$8.00 per bar of soap. She buys a bar of soap every 2 months. There are 12 months in a year. Maria will spend  $\$8.00 \times 12 / 2 = \$48.00$  on soap for the year. **The answer is \$48.00.**

---

---

### Output to the Modified Problem

#### CoT

Maria buys a large bar of French soap that lasts her for 2 months. She spends \$8.00 per bar of soap. Every 10 months, Maria's neighbor buys a new shampoo and moisturizer for Maria's neighbor. If Maria wants to stock up for the entire year, she will spend  $\$8.00 \times 12 = \$96.00$  on soap. **The answer is \$96.00.**

---

---

#### LTM

Let's break down this problem: 1. How much does Maria spend on soap every 10 months? 2. How much does Maria spend on soap every year?  
1. Maria spends \$8.00 per bar of soap. She buys a new bar of soap every 10 months. So Maria spends  $\$8.00 \times 10 = \$80.00$  every 10 months.  
2. Maria spends \$80.00 every 10 months. So Maria spends  $\$80.00 \times 12 = \$960.00$  every year.  
**The answer is \$960.00.**

## Attention Setbacks: Disperse Attention



Possible Reasons:

1. Softmax can't produce zero scores
2. Producing near-zero scores needs a wide input range which harms backpropagation of softmax



## Attention Setbacks: *Lost-in-Middle*

### Original Prompt / Model's Response

Write a high-quality answer for the given question using only the provided search results (some of which might be irrelevant).

Question: what is mercy mercy me by marvin gaye about

Search results:

Document [1] ...The follow-up single, "Mercy Mercy Me (The Ecology)", peaked at number-four on the Hot 100...

Document [10] Mercy Mercy Me (The Ecology)... became regarded as one of popular music's most poignant anthems of sorrow regarding the environment

Document [20]...

Question: what is mercy mercy me by marvin gaye about

Answer: ...The song was released as the follow-up single to the album's title track and peaked at number-four on the Hot 100...

### Shuffled Prompt / Model's Response

Write a high-quality answer for the given question using only the provided search results (some of which might be irrelevant).

Question: what is mercy mercy me by marvin gaye about

Search results:

Document [1] ...The album became Gaye's first million-selling album launching two more top ten singles

Document [13] Mercy Mercy Me (The Ecology)... became regarded as one of popular music's most poignant anthems of sorrow regarding the environment

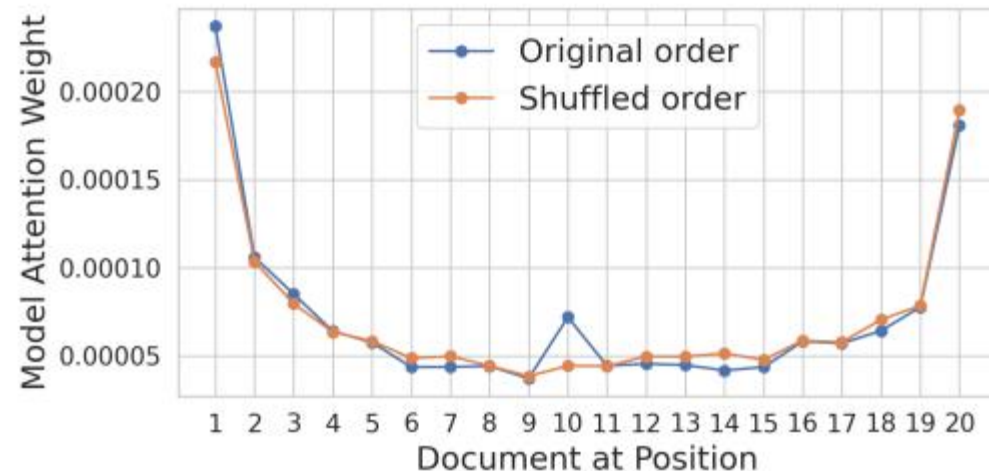
Document [20]...

Question: what is mercy mercy me by marvin gaye about

Answer: ...The song was released as a single in August 1971 and became Gaye's first million-selling single...

## Attention Setbacks: *Lost-in-Middle*

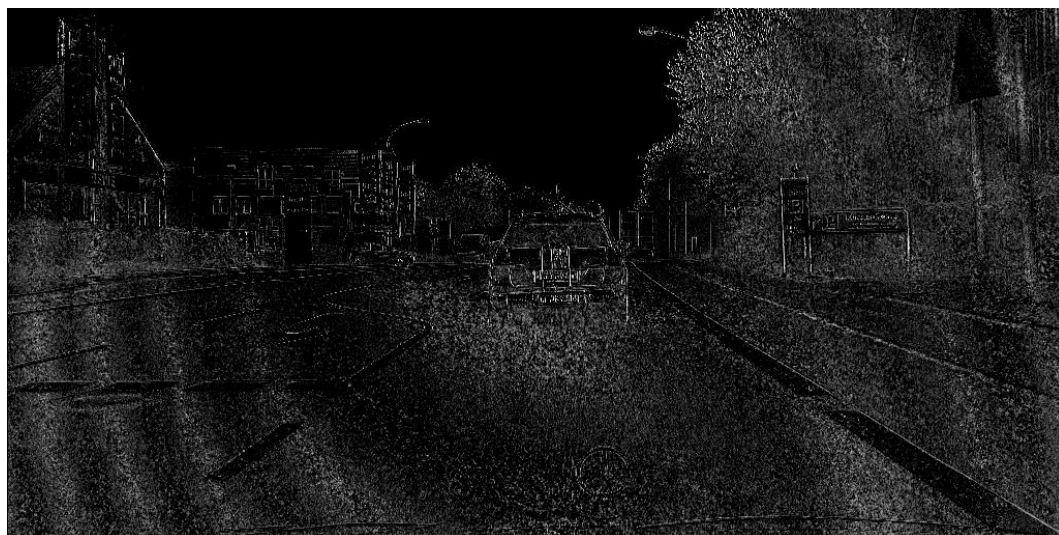
Original Prompt / Model's Response	Shuffled Prompt / Model's Response
<p>Write a high-quality answer for the given question using only the provided search results (some of which might be irrelevant).</p> <p>Question: what is mercy mercy me by marvin gaye about</p> <p>Search results:</p> <p>Document [1] ...The follow-up single, "Mercy Mercy Me (The Ecology)", peaked at number-four on the Hot 100...</p> <p>Document [10] Mercy Mercy Me (The Ecology)... became regarded as one of popular music's most poignant anthems of sorrow regarding the environment</p> <p>Document [20]...</p> <p>Question: what is mercy mercy me by marvin gaye about</p> <p>Answer: ...The song was released as the follow-up single to the album's title track and peaked at number-four on the Hot 100...</p>	<p>Write a high-quality answer for the given question using only the provided search results (some of which might be irrelevant).</p> <p>Question: what is mercy mercy me by marvin gaye about</p> <p>Search results:</p> <p>Document [1] ...The album became Gaye's first million-selling album launching two more top ten singles</p> <p>Document [13] Mercy Mercy Me (The Ecology)... became regarded as one of popular music's most poignant anthems of sorrow regarding the environment</p> <p>Document [20]...</p> <p>Question: what is mercy mercy me by marvin gaye about</p> <p>Answer: ...The song was released as a single in August 1971 and became Gaye's first million-selling single...</p>



*Hsieh et al*, Found in the Middle: Calibrating Positional Attention Bias Improves Long Context Utilization, ACL 2024



## U-shape Attention in Visual Tasks:



input feature

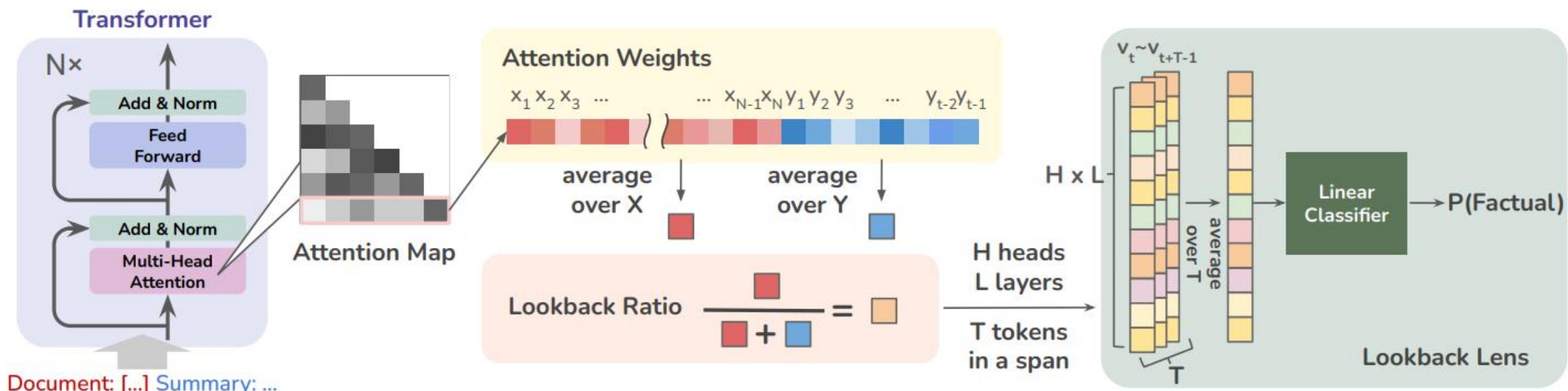
self-attn →



attention score

**The attention signal is noisy!**  
**How can we reduce the noise?**

## Reduce Attention Noise with Hallucination Detection - Selection



$$A_t^{l,h}(\text{context}) = \frac{1}{N} \sum_{i=1}^N \alpha_{h,i}^l,$$

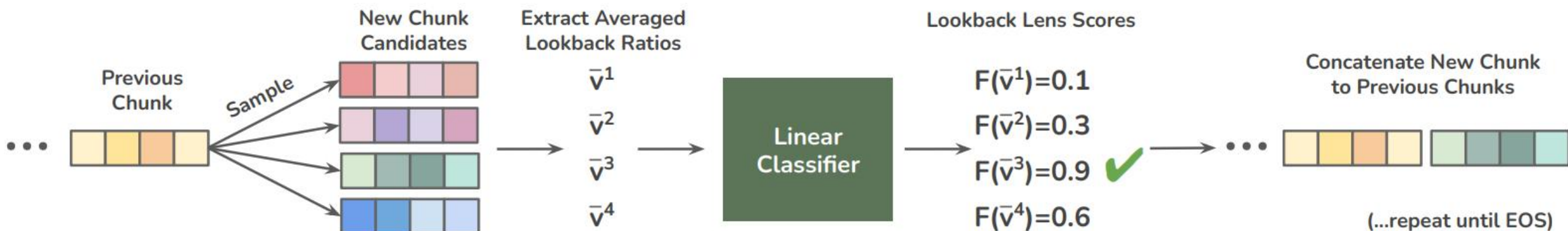
$$A_t^{l,h}(\text{new}) = \frac{1}{t-1} \sum_{j=N+1}^{N+t-1} \alpha_{h,j}^l,$$

$$\text{LR}_t^{l,h} = \frac{A_t^{l,h}(\text{context})}{A_t^{l,h}(\text{context}) + A_t^{l,h}(\text{new})}.$$

$$\mathbf{v}_t = [\text{LR}_t^{1,1}, \text{LR}_t^{1,2}, \dots, \text{LR}_t^{L,H}]$$

$$P(y = 1 | \mathbf{v}) = \mathcal{F}(\mathbf{v}) = \sigma(\mathbf{w}^T \mathbf{v} + b)$$

## Reduce Attention Noise with Hallucination Detection - Selection



## Selective Attention: Use Mask to Reduce Noise

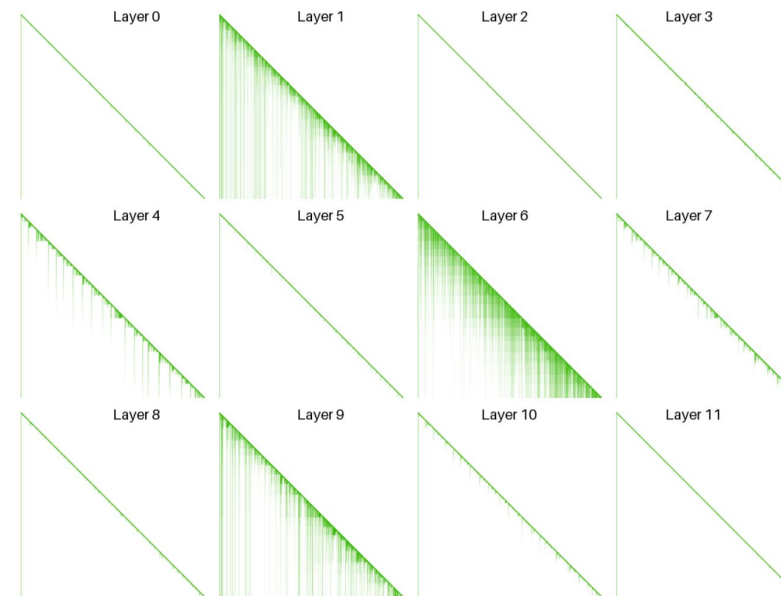
Construct a constraint matrix  $S$  :

1. Zero out negative values (i.e. applying ReLU), only reducing attention, never boosting it
2. Zero out the first column, so as not to mask the <BOS> token.
3. Zero out the diagonal, so as not to let a token mask itself.

$$F_{i,j} = \sum_{k \leq i-1} S_{k,j}$$

$$\text{SelectiveAttention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}} - F\right)V$$

$$\mathcal{L}_{mem} = \mathcal{L}_{ppl} + \epsilon \cdot \frac{\sum_{l=1}^L \max_i M_i^l}{L \cdot n_{\neq pad}}$$





## Calibrated Attention: Decouple Attention and Positional Bias

Input:  $x^{\text{prompt}} = [x^{\text{q}}, x_1^{\text{doc}}, \dots, x_k^{\text{doc}}, x^{\text{q}}]$

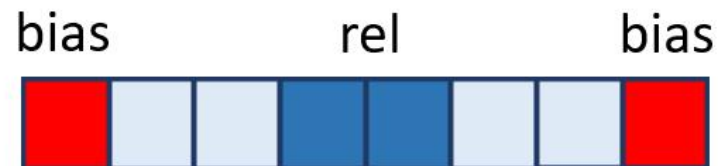
Position-related score:

$$\text{Attn}(x^{\text{prompt}}, k) = \frac{1}{N_k} \sum_{i=1}^{N_k} \text{attn}(x_k^{\text{doc}})$$

Modeling position bias:

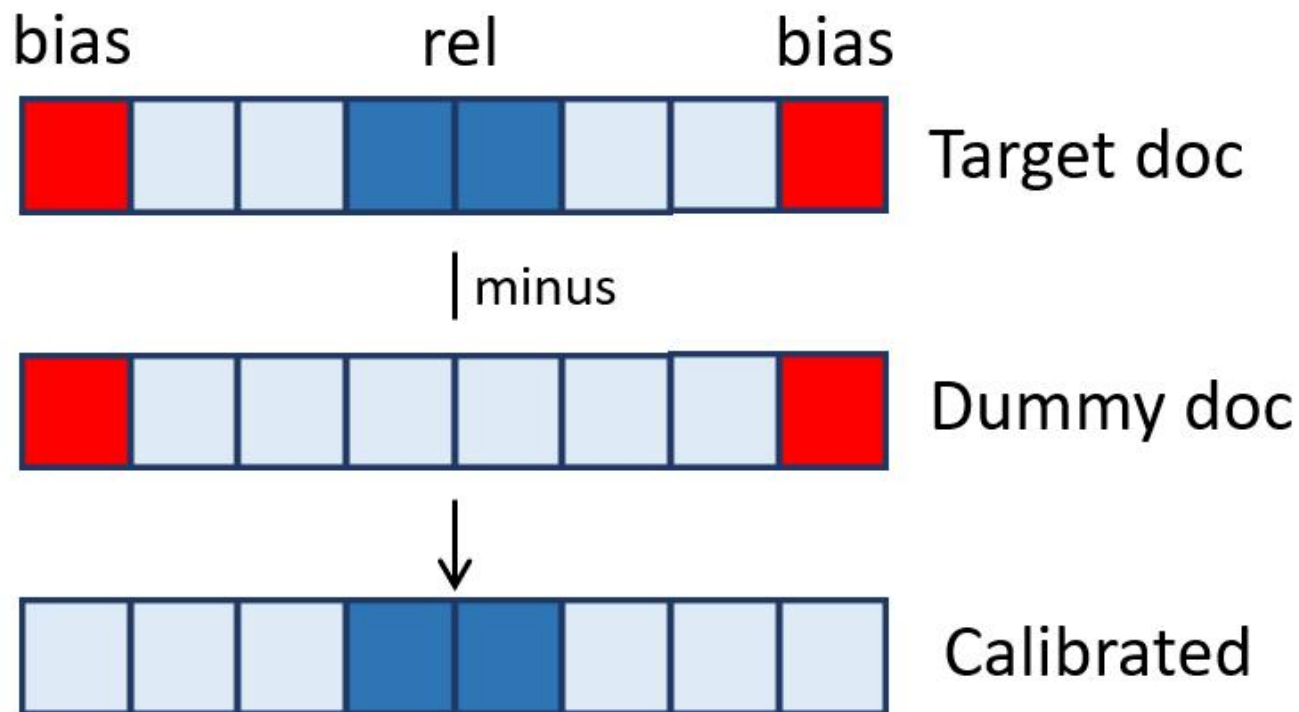
$$\text{Attn}(x^{\text{prompt}}, k) = f(\text{rel}(x_k^{\text{doc}}), \text{bias}(k))$$

$$\text{Attn}(x^{\text{doc}}, k) = \text{rel}(x^{\text{doc}}) + \text{bias}(k) + \epsilon$$





## Calibrated Attention: Decouple Attention and Positional Bias



*Hsieh et al*, Found in the Middle: Calibrating Positional Attention Bias Improves Long Context Utilization, ACL 2024

## Calibrated Attention: Decouple Attention and Positional Bias

Extracting content relevance:

$$\text{Attn}(x^{\text{dum}}, k) = \text{rel}(x^{\text{dum}}) + \text{bias}(k) + \epsilon$$

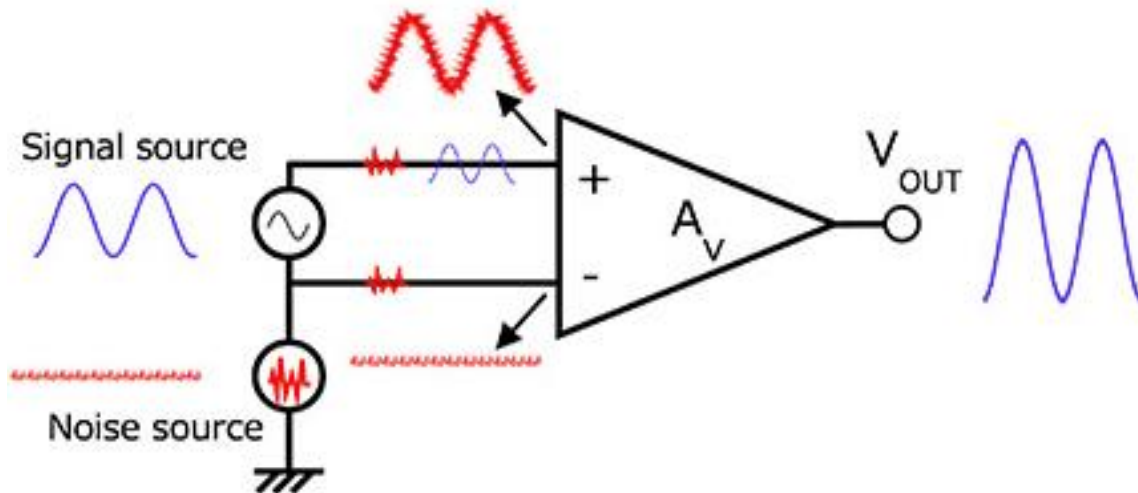
$$\text{rel}(x^{\text{doc}}) = \text{Attn}(x^{\text{doc}}, k) - \text{Attn}(x^{\text{dum}}, k) + \text{rel}(x^{\text{dum}})$$

Calibrate the attention:

$$\text{attn}_{\text{calibrated}}(x_{k,i}^{\text{doc}}) = \frac{\alpha_k}{\text{Attn}_{\text{original}}(x_k^{\text{doc}})} \cdot \text{attn}_{\text{original}}(x_{k,i}^{\text{doc}}) \cdot C$$

$$\alpha_k = \text{Softmax}(\text{rel}(x_k^{\text{doc}}), t)$$

## Noise Reduction with Differential Amplifier:



$$V_{OUT} = A_V \cdot (V_{\text{signal}} - V_{\text{noise}})$$

## Differential Attention:

$$Q = XW^Q, \quad K = XW^K, \quad V = XW^V$$

$$\text{Attn}(X) = \left( \text{softmax} \left( \frac{QK^T}{\sqrt{d}} \right) \right) V$$

## Differential Attention:

$$Q = XW^Q, \quad K = XW^K, \quad V = XW^V$$

$$\text{Attn}(X) = \left( \text{softmax} \left( \frac{QK^T}{\sqrt{d}} \right) \right) V$$

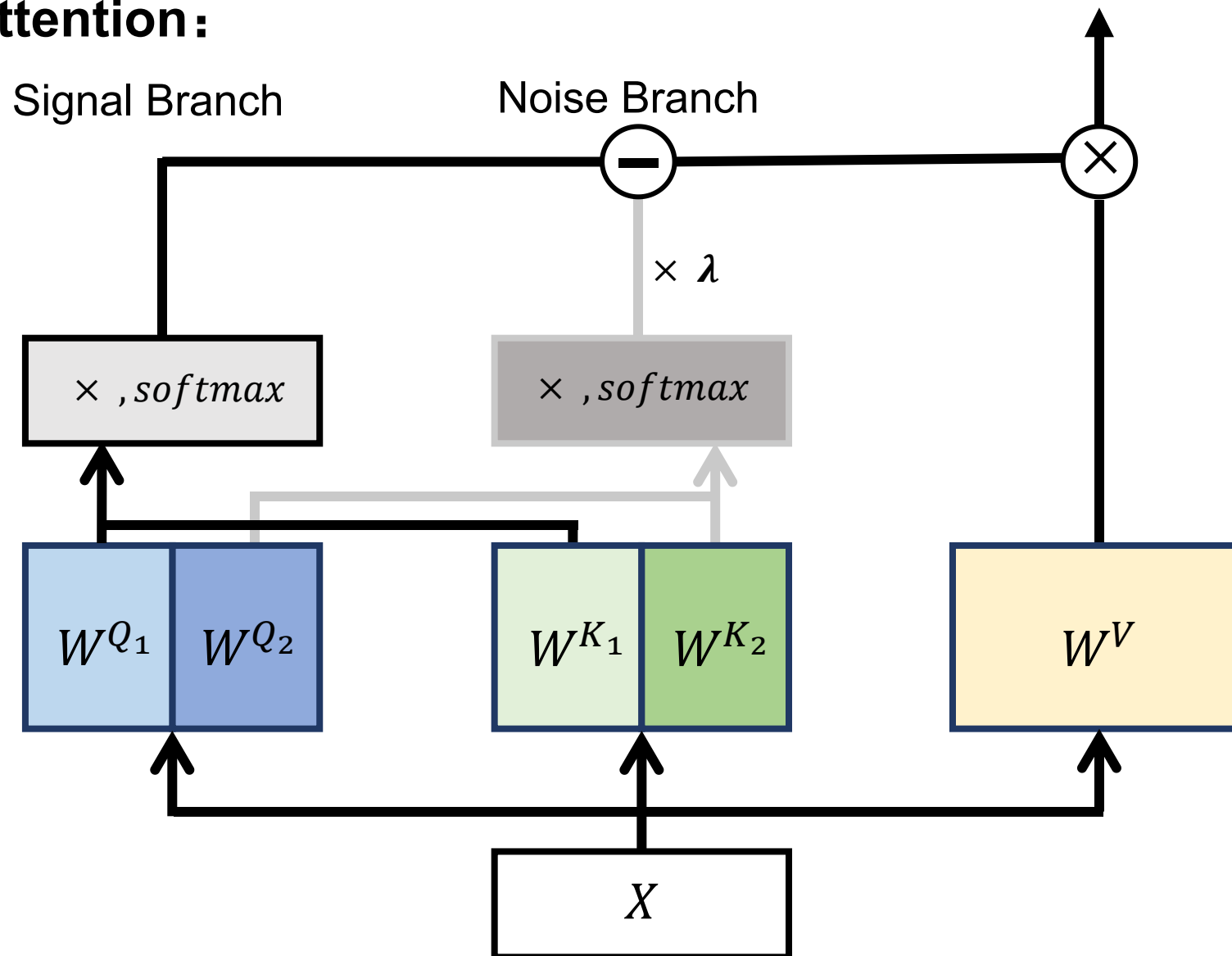


$$[Q_1; Q_2] = XW^Q, \quad [K_1; K_2] = XW^K, \quad V = XW^V$$

$$\text{DiffAttn}(X) = \left( \text{softmax} \left( \frac{Q_1 K_1^T}{\sqrt{d}} \right) - \lambda \text{softmax} \left( \frac{Q_2 K_2^T}{\sqrt{d}} \right) \right) V$$

---

## Differential Attention:



### Re-parameterizing $\lambda$ :

To align the learning rate of the parameters:

Original: 
$$\lambda \cdot \text{softmax}(qK^T) = \lambda \cdot \frac{\exp\{qK^T\}}{\sum_i \exp\{(qK^T)_i\}}$$

Exponential form: 
$$\lambda = \exp\{\lambda_q \cdot \lambda_k\}, \quad \lambda \cdot \text{softmax}(qK^T) = \frac{\exp\{qK^T + \lambda_q \cdot \lambda_k\}}{\sum_i \exp\{(qK^T)_i\}}$$

Initialization: 
$$\lambda = \exp\{\lambda_q \cdot \lambda_k\} + \lambda_{init}$$

Enable lower values: 
$$\lambda = \exp\{\lambda_{q_1} \cdot \lambda_{k_1}\} - \exp\{\lambda_{q_2} \cdot \lambda_{k_2}\} + \lambda_{init}$$

## Overall Framework:

Similar multi-head attention:

$$\text{head}_i = \text{DiffAttn}(X; W_i^Q, W_i^K, W_i^V, \lambda)$$

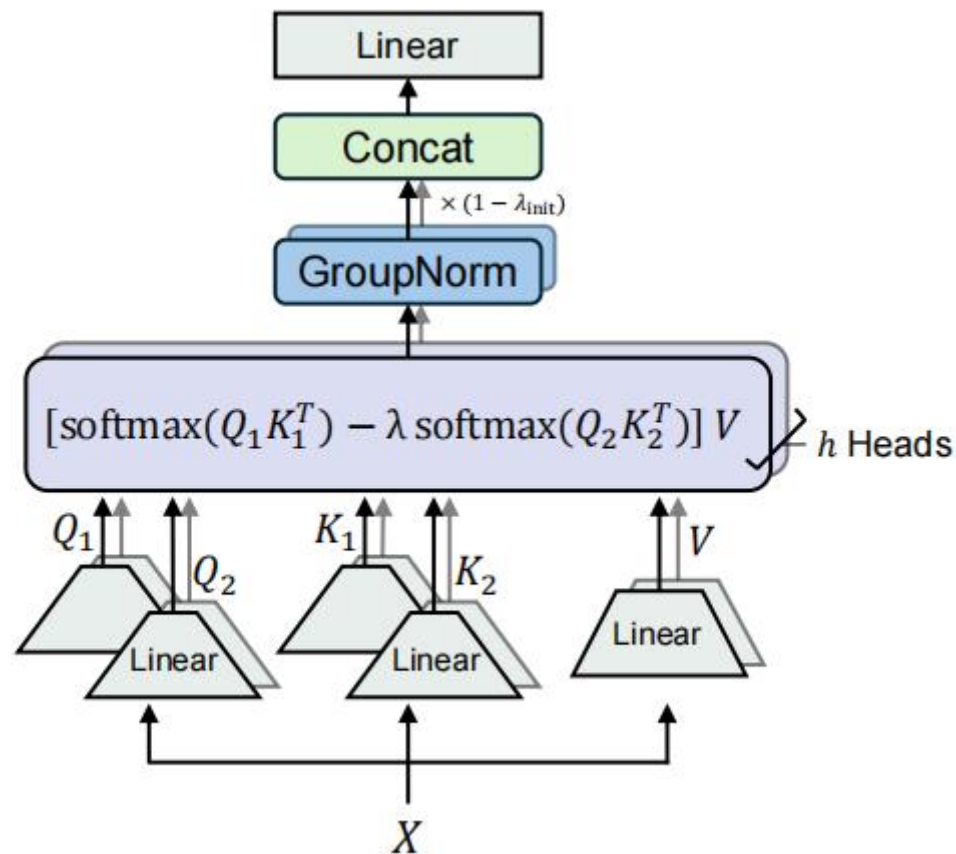
$$\overline{\text{head}}_i = (1 - \lambda_{\text{init}}) \cdot \text{LN}(\text{head}_i)$$

$$\text{MultiHead}(X) = \text{Concat}(\overline{\text{head}}_1, \dots, \overline{\text{head}}_h) W^O$$

Same macro layout:

$$Y^l = \text{MultiHead}(\text{LN}(X^l)) + X^l$$

$$X^{l+1} = \text{SviGLU}(\text{LN}(Y^l)) + Y^l$$





### Fixed Multiplier ( $1 - \lambda_{\text{init}}$ ) for Multi-Head:

$$\text{head}_i = \text{DiffAttn}(X; W_i^Q, W_i^K, W_i^V, \lambda)$$

$$\overline{\text{head}_i} = \underline{(1 - \lambda_{\text{init}})} \cdot \text{LN}(\text{head}_i)$$

$$\text{MultiHead}(X) = \text{Concat}(\overline{\text{head}_1}, \dots, \overline{\text{head}_h}) W^O$$

### Fixed Multiplier $(1 - \lambda_{\text{init}})$ for Multi-Head:

$$\begin{aligned}\text{head}_i &= \text{DiffAttn}(X; W_i^Q, W_i^K, W_i^V, \lambda) \\ \overline{\text{head}}_i &= (1 - \lambda_{\text{init}}) \cdot \text{LN}(\text{head}_i) \\ \text{MultiHead}(X) &= \text{Concat}(\overline{\text{head}}_1, \dots, \overline{\text{head}}_h) W^O\end{aligned}$$

Controlling Gradient Flow: We expect the magnitude of

$$\frac{\partial L}{\partial W^O}, \frac{\partial L}{\partial W^V}, \frac{\partial L}{\partial W^{Q_1}}, \frac{\partial L}{\partial W^{Q_2}}, \frac{\partial L}{\partial W^{K_1}}, \frac{\partial L}{\partial W^{K_2}}$$

to remain the same as conventional Transformers

## Fixed Multiplier $(1 - \lambda_{\text{init}})$ for Multi-Head:

$$\begin{aligned}
 & \frac{\partial L}{\partial W_{Q_1}} \\
 &= \frac{\partial L}{\partial O} \frac{\partial O}{\partial \text{head}} \frac{\partial \overline{\text{head}}}{\partial \text{head}} \frac{\partial \text{head}}{\partial A_1} \frac{\partial A_1}{\partial Q_1} \frac{\partial Q_1}{\partial W_{Q_1}} \\
 &= \frac{1}{\sqrt{d}} X^T \left[ \underline{A_1} \odot \left( \frac{\partial L}{\partial O} (W^O)^T V^T - (A_1 \odot \left( \frac{\partial L}{\partial O} (W^O)^T V^T \right)) J \right) \right] K_1 \underline{\frac{\partial \overline{\text{head}}}{\partial \text{head}}}
 \end{aligned}$$

## Fixed Multiplier ( $1 - \lambda_{\text{init}}$ ) for Multi-Head:

$$\begin{aligned}
 & \frac{\partial L}{\partial W_{Q_1}} \\
 &= \frac{\partial L}{\partial O} \frac{\partial O}{\partial \text{head}} \frac{\partial \overline{\text{head}}}{\partial \text{head}} \frac{\partial \text{head}}{\partial A_1} \frac{\partial A_1}{\partial Q_1} \frac{\partial Q_1}{\partial W_{Q_1}} \\
 &= \frac{1}{\sqrt{d}} X^T \left[ \underline{A_1} \odot \left( \frac{\partial L}{\partial O} (W^O)^T V^T - (A_1 \odot \left( \frac{\partial L}{\partial O} (W^O)^T V^T \right)) J \right) \right] K_1 \underline{\frac{\partial \overline{\text{head}}}{\partial \text{head}}}
 \end{aligned}$$

$$\overline{\text{head}} = GN(\text{head})$$

$$= \frac{\text{head} - \mathbb{E}(\text{head})}{\text{Std}(\text{head})}$$

$$\mathbb{E}(\text{head}) = 0, \text{Std}(\text{head}) = \mathbb{E}(\|\text{head}\|_2^2)$$

$$\frac{\partial \overline{\text{head}}}{\partial \text{head}} = \frac{1}{\sqrt{\mathbb{E}(\|\text{head}\|_2^2)}}$$

## Fixed Multiplier (1 - $\lambda_{\text{init}}$ ) for Multi-Head:

In conventional Transformers:

$$\begin{aligned}\text{head} &= \text{Attn}(Q, K) \\ &= \text{softmax} \left( \frac{QK^T}{\sqrt{d}} \right) V \\ \mathbb{E} \left( \frac{QK^T}{\sqrt{d}} \right) &= 0 \\ \mathbb{E} (\|\text{head}\|_2^2) &= 0.5 \mathbb{E} (\|V\|_2^2) \\ &= \Theta(1)\end{aligned}$$

## Fixed Multiplier (1 - $\lambda_{\text{init}}$ ) for Multi-Head:

In conventional Transformers:

$$\begin{aligned}\text{head} &= \text{Attn}(Q, K) \\ &= \text{softmax} \left( \frac{Q_1 K_1^T}{\sqrt{d}} \right) V \\ \mathbb{E} \left( \frac{QK^T}{\sqrt{d}} \right) &= 0 \\ \mathbb{E} (\|\text{head}\|_2^2) &= 0.5 \mathbb{E} (\|V\|_2^2) \\ &= \Theta(1)\end{aligned}$$

In Diff-Transformers:

$$\begin{aligned}\text{head} &= \text{DiffAttn}(Q, K) \\ &= \left( \text{softmax} \left( \frac{Q_1 K_1^T}{\sqrt{d}} \right) - \lambda \text{softmax} \left( \frac{Q_2 K_2^T}{\sqrt{d}} \right) \right) V \\ \mathbb{E} \left( \frac{QK^T}{\sqrt{d}} \right) &= 0, \quad \mathbb{E}(\lambda) = \lambda_{\text{init}} \\ \mathbb{E} (\|\text{head}\|_2^2) &= 0.5(1 - \lambda_{\text{init}})^2 \mathbb{E} (\|V\|_2^2) \\ &= \Theta((1 - \lambda_{\text{init}})^2)\end{aligned}$$

## Settings:

- Decoder-only setup
- Compared with augmented Transformer architecture as in LLaMA

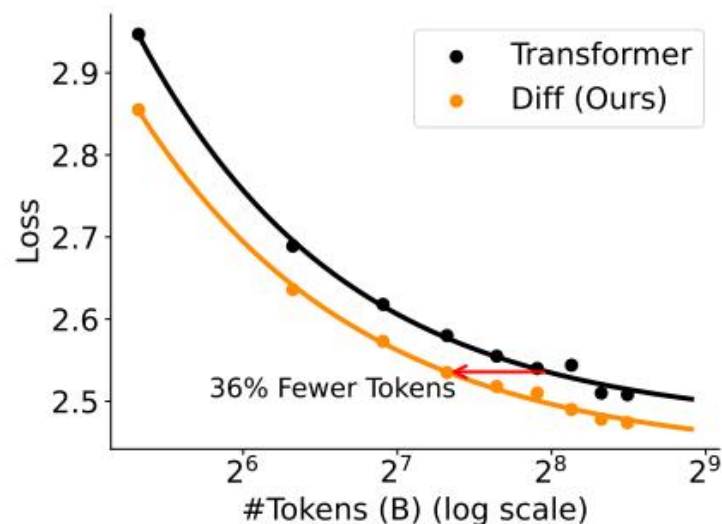
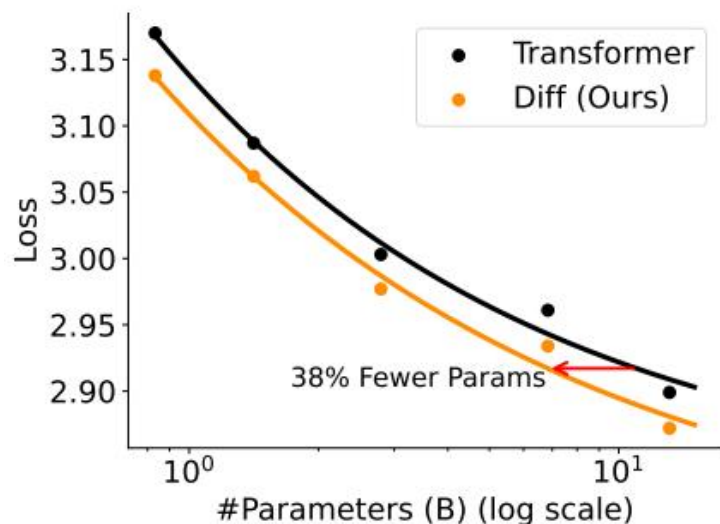
## Evaluation:

- Various downstream tasks
- Long-sequence modeling
- Key information retrieval, contextual hallucination evaluation, and in-context learning
- Activation outliers

## Better Performance & Scalability:

Model	ARC-C	ARC-E	BoolQ	HellaSwag	OBQA	PIQA	WinoGrande	Avg
<i>Zero-Shot</i>								
Transformer-3B	32.2	66.8	<b>62.9</b>	63.4	26.2	74.5	61.6	55.4
DIFF-3B	<b>33.0</b>	<b>68.3</b>	60.1	<b>66.2</b>	<b>27.6</b>	<b>75.5</b>	<b>62.7</b>	<b>56.2</b>
<i>5-Shot</i>								
Transformer-3B	34.0	<b>69.5</b>	65.3	63.4	25.0	75.2	62.6	56.4
DIFF-3B	<b>35.0</b>	<b>69.5</b>	<b>67.2</b>	<b>66.9</b>	<b>27.6</b>	<b>76.1</b>	<b>63.8</b>	<b>58.0</b>

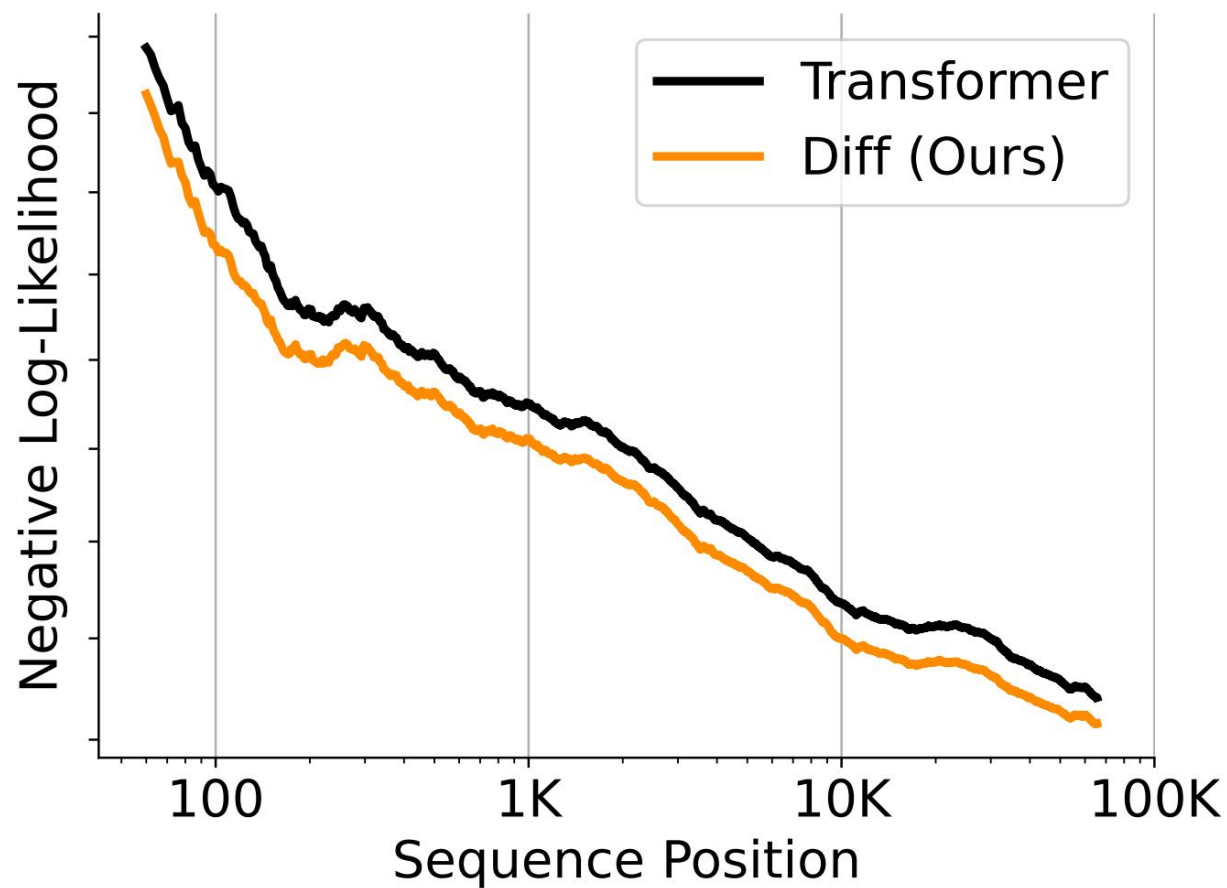
Table 1: Comparison of DIFF Transformer with Transformer on LM Eval Harness (Gao et al., 2023). DIFF Transformer achieves better accuracy in the zero- and few-shot settings.





## Long Context Modeling:

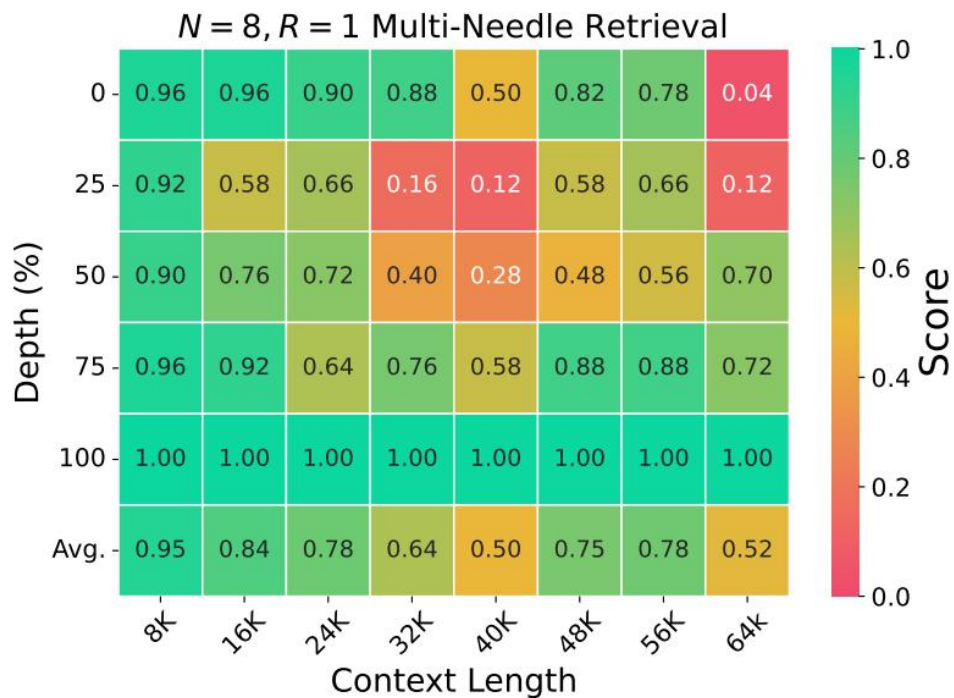
Culminative NLL on text-to-text:



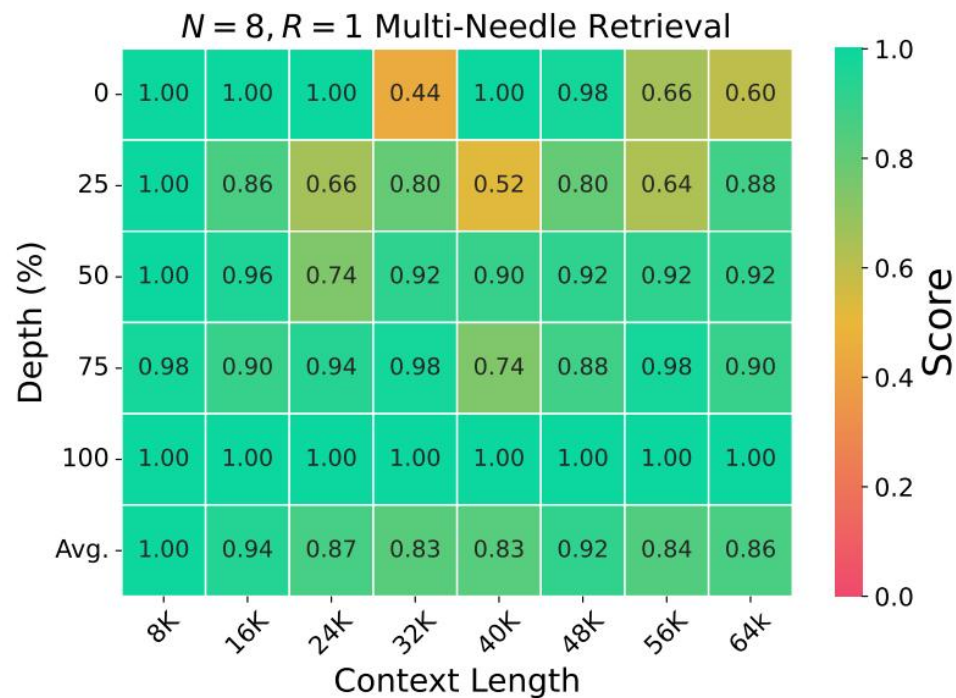
## Information Retrieval:

4K retrieval:

Model	$N = 1$	$N = 2$	$N = 4$	$N = 6$
	$R = 1$	$R = 2$	$R = 2$	$R = 2$
Transformer	<b>1.00</b>	0.85	0.62	0.55
DIFF	<b>1.00</b>	<b>0.92</b>	<b>0.84</b>	<b>0.85</b>

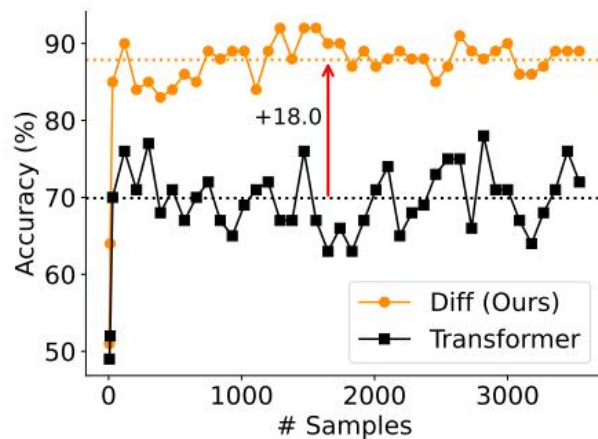


(a) Transformer.

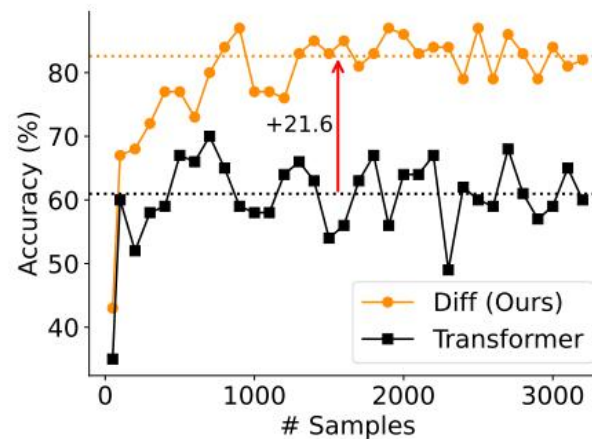


(b) DIFF Transformer.

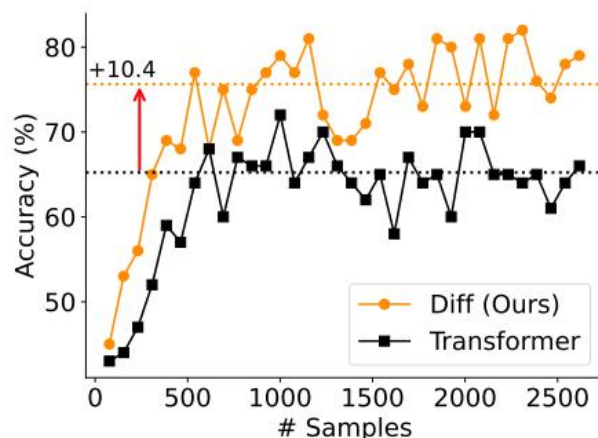
## In-Context-Learning:



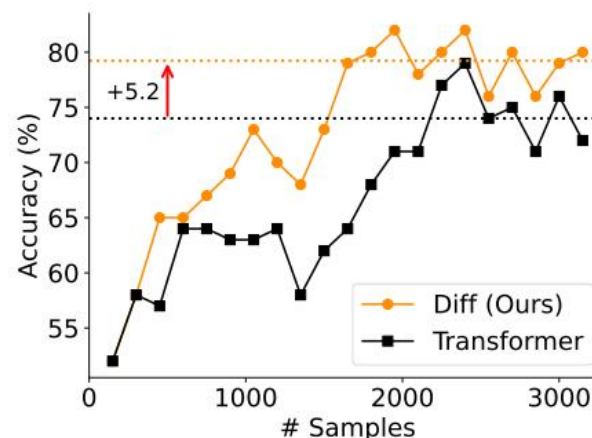
(a) TREC with 6 classes.



(b) TREC-fine with 50 classes.



(c) Banking-77 with 77 classes.



(d) Clinic-150 with 150 classes.

## Contextual Hallucination:

Model	XSum	CNN/DM	MultiNews
Transformer	0.44	0.32	0.42
DIFF	<b>0.53</b>	<b>0.41</b>	<b>0.61</b>

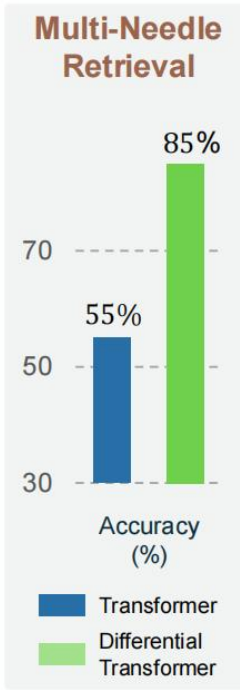
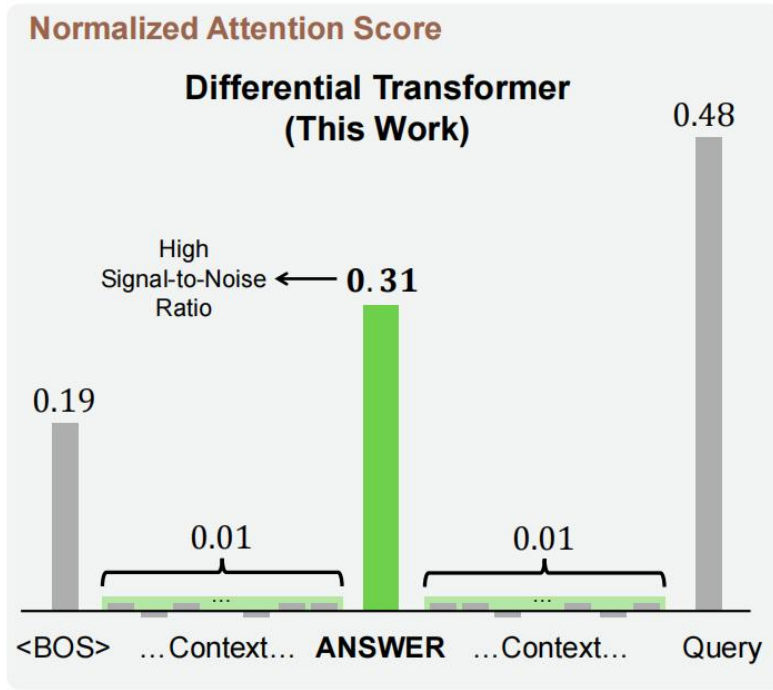
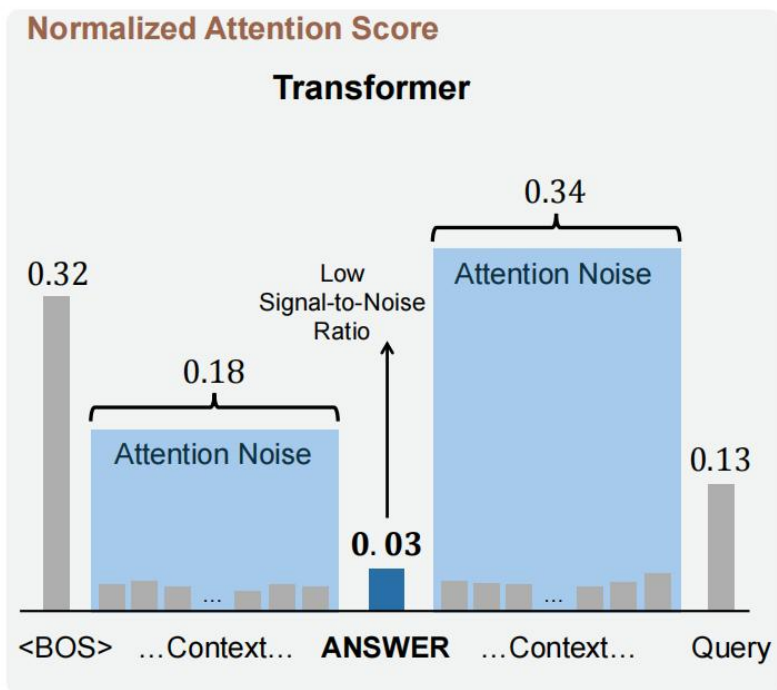
(a) Accuracy (i.e., free of hallucinations) on text summarization datasets.

Model	Qasper	HotpotQA	2WikiMQA
Transformer	0.28	0.36	0.29
DIFF	<b>0.39</b>	<b>0.46</b>	<b>0.36</b>

(b) Accuracy (i.e., free of hallucinations) on question answering datasets.

## Attention Allocation:

Model	Attention to Answer $\uparrow$					Attention Noise $\downarrow$				
	0%	25%	50%	75%	100%	0%	25%	50%	75%	100%
Transformer	0.03	0.03	0.03	0.07	0.09	0.51	0.54	0.52	0.49	0.49
DIFF	<b>0.27</b>	<b>0.30</b>	<b>0.31</b>	<b>0.32</b>	<b>0.40</b>	<b>0.01</b>	<b>0.02</b>	<b>0.02</b>	<b>0.02</b>	<b>0.01</b>






## Activation Outliers:

Model	Activation Type	Top-1	Top-2	Top-3	Top-10	Top-100	Median
Transformer	Attention Logits	318.0	308.2	304.9	284.7	251.5	5.4
DIFF	Attention Logits	38.8	38.8	37.3	32.0	27.4	3.3
Transformer	Hidden States	3608.6	3607.4	3603.6	3552.1	2448.2	0.6
DIFF	Hidden States	1688.2	1672.5	1672.1	1624.3	740.9	1.2

## Ablation Studies:

Model	#heads	$d$	GN	Valid. Set↓	Fine-Grained Slices	
					AR-Hit↓	Others↓
Transformer	16	128	✗	3.087	0.898	3.272
Transformer	8	256	✗	3.088	0.899	3.273
+ GroupNorm	8	256	✓	3.086	0.899	3.271
DIFF Transformer	8	128	✓	<b>3.062</b>	<b>0.880</b>	<b>3.247</b>
– GroupNorm	8	128	✗	3.122	0.911	3.309
with $\lambda_{\text{init}} = 0.8$	8	128	✓	3.065	0.883	3.250
with $\lambda_{\text{init}} = 0.5$	8	128	✓	3.066	0.882	3.251

## Further Questions

 tmp1234 Oct 9, 2024



Great work! Just wonder do you have any idea why two learned attentions tend to cancel noise, rather than canceling signals? For instance, if attention 1 learns  $S + N_1$ , and attention 2 learns  $S + N_2$  (where  $S$  is signal,  $N_1, N_2$  are different noises), by subtracting these two, the signal  $S$  gets canceled while noise becomes  $N_1 - N_2$  which could be more complicated. Is there any reason why the model would not do this instead?

 1 reply



1



 ytz20 Paper author Oct 9, 2024 • edited Oct 9, 2024



It's a good question. Our observation is that the model knows what signal is and what noise is. Notice that attention\_1 and attention\_2 are both calculated with learnable parameters, they can "perceive" each other in the training process. Then they can adjust themselves according to each other, to achieve lower loss. The result is that the model chooses to preserve signal and cancel out noise as long as we give it the chance to do so. And for a single softmax, it's difficult for it to learn the same solution, due to its formulation and gradient properties.

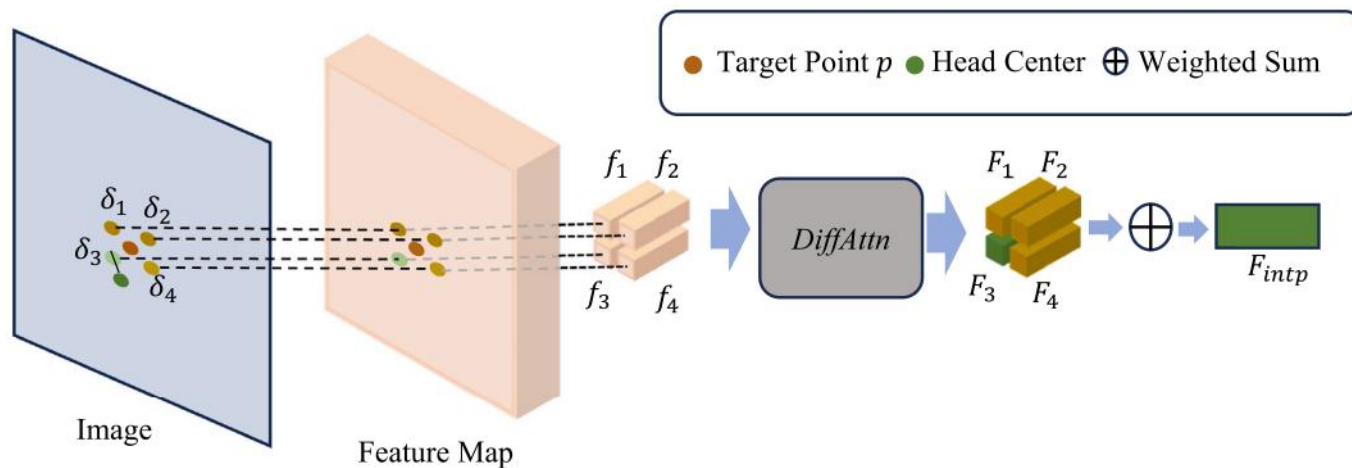


8





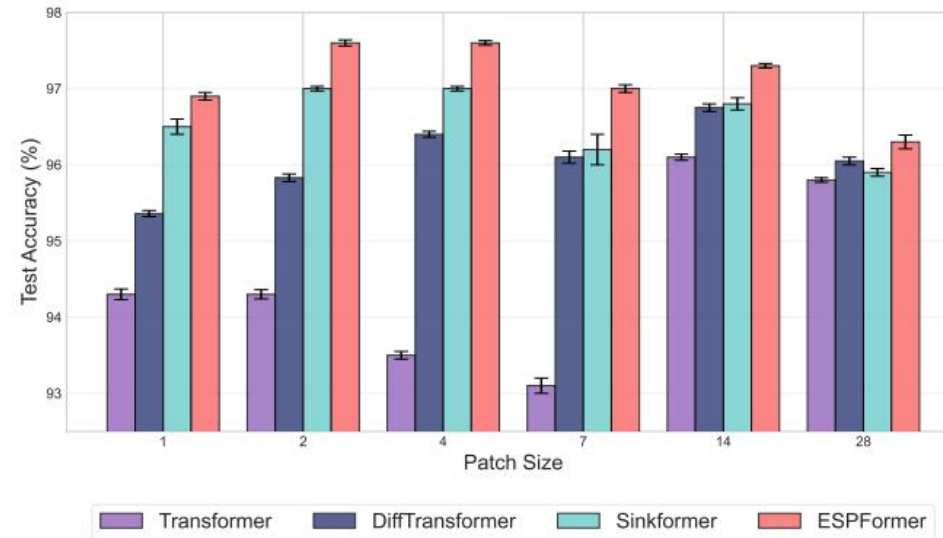
## Crowd Localization:



*Zhang et al*, DiffusionLoc: A diffusion model-based framework for crowd localization, Image and Vision Computing, 2025

## Image Classification:

Data Fraction	ESPFormer	Sinkformer	DiffTransformer	Transformer
1%	<b><math>55.66 \pm 3.95</math></b>	$55.07 \pm 3.34$	$53.78 \pm 0.28$	$49.71 \pm 0.31$
10%	<b><math>71.49 \pm 0.43</math></b>	$69.56 \pm 0.32$	$67.34 \pm 0.11$	$57.25 \pm 0.22$
25%	<b><math>75.40 \pm 0.38</math></b>	$74.56 \pm 0.58$	$74.86 \pm 0.17$	$72.25 \pm 0.16$
100%	<b><math>79.47 \pm 0.12</math></b>	$79.12 \pm 0.17$	$78.85 \pm 0.11$	$78.49 \pm 0.09$



*Shahbazi et al*, ESPFormer: Doubly-Stochastic Attention with Expected Sliced Transport Plans, arXiv 2025

## Highlights

- Simple approach, plug-in to all modern transformer-based models
- Extensive Experiment showing good performance on all tasks
- Makes a significant improvement for long-text tasks

## Current Limitations

- Only implemented on decoder-only architectures
- More modalities —— upcoming in the revised version

.....

- **Attention is noisy**
- **Differential Attention learns to decouple signal & noise**
- **Excels on long-context processing**
- **Simple design, plug-and-use**
- **Vision-to-be-explored**

# Thanks for listening!

Presenter: Jinyi Luo  
2025.02.23