

# Exploring Scaling Laws of CTR Model for Online Performance Improvement

## 探索点击率预测模型的缩放定律提高在线性能

赖伟江, 金蓓弘

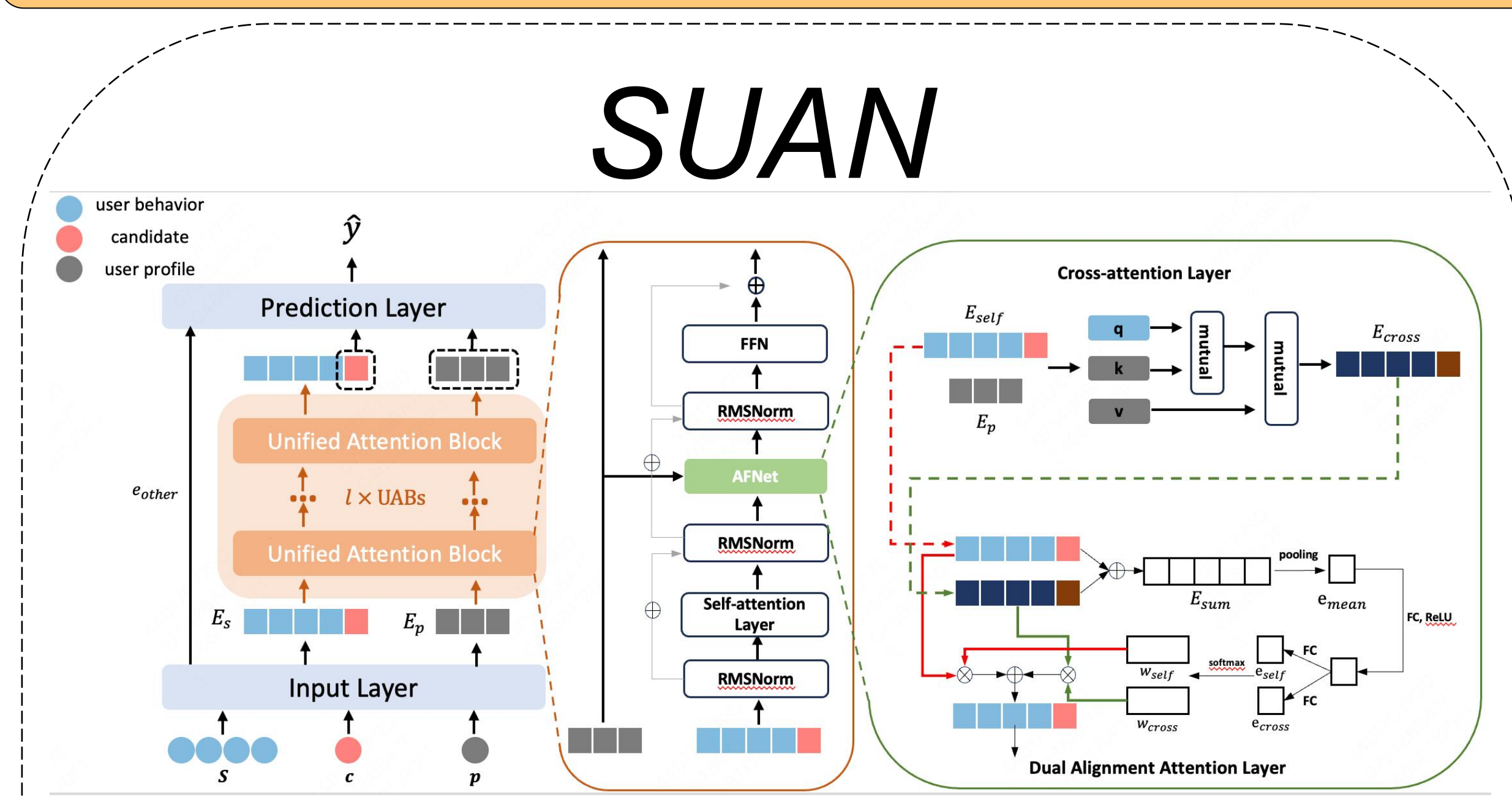
19th ACM Conference on Recommender Systems (accepted)

Contact: Beihong@iscas.ac.cn

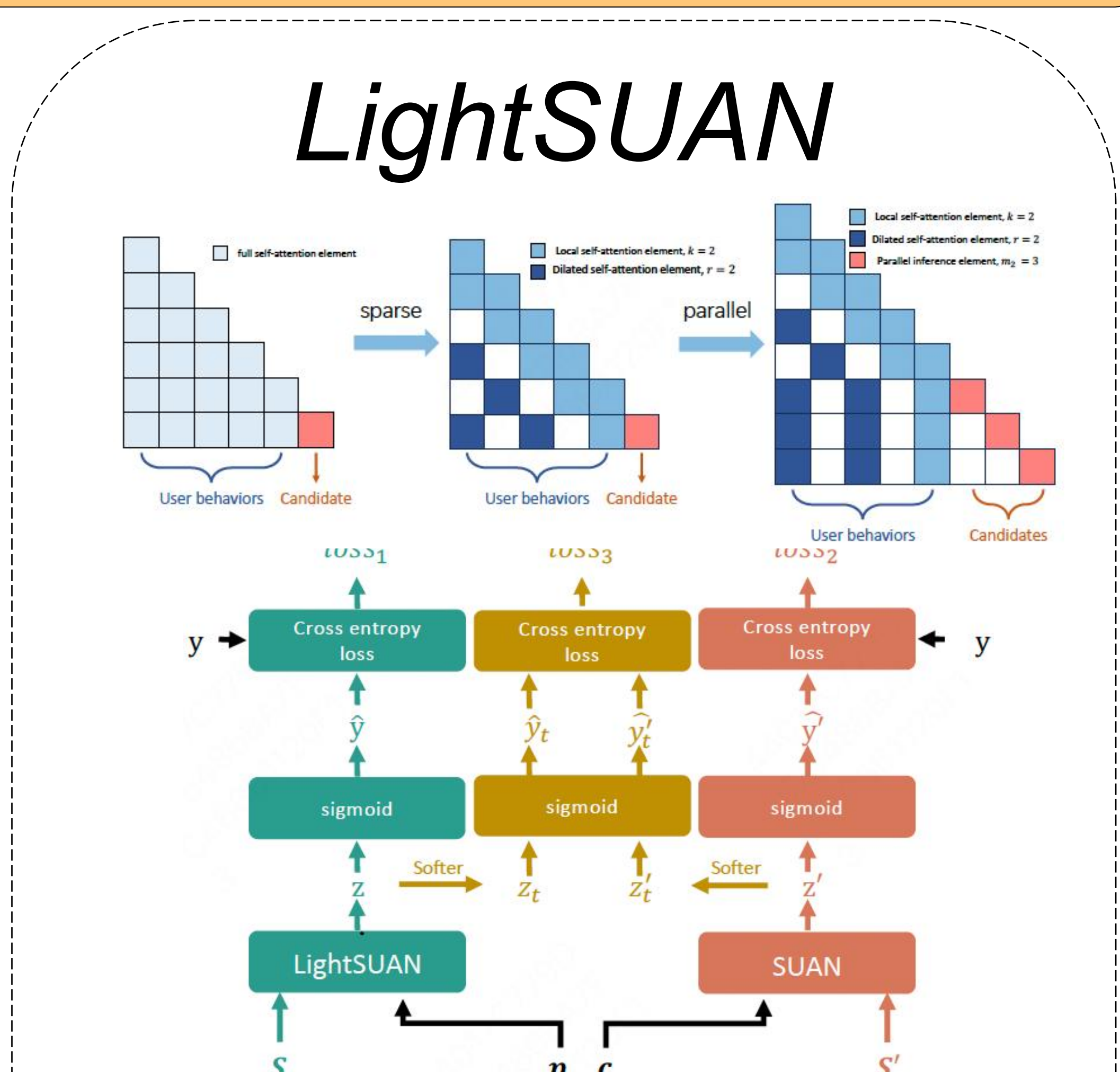
### Introduction

- Click-Through Rate (CTR) prediction is a critical task in online personalized services, directly influencing user experience and business revenue. Existing CTR models face performance bottlenecks, and even minor improvements are highly valued.
- Inspired by scaling laws from large language models, this work proposes a two-stage paradigm: first, train a large CTR model whose accuracy scales with model size and data size; then, distill its knowledge into a lightweight model for online deployment.
- Following this line of thought, we propose a CTR model named SUAN (Stacked Unified Attention Network) to offer scalable, high-accuracy CTR predictions

### Methodology



- We propose SUAN (Stacked Unified Attention Network) with stacked Unified Attention Blocks (UABs) as the core component.
- Each UAB integrates multiple attention mechanisms:
  - Self-attention for capturing dependencies in user behavior sequences.
  - Cross-attention for modeling the importance of behavior features from the perspective of user profile features.
  - Dual alignment attention for adaptively aligning and highlighting informative features while suppressing less relevant ones.
- Incorporates RMSNorm and SwiGLU to enhance training stability and expressiveness.



- We constructs LightSUAN, a lightweight SUAN variant with sparse self-attention and parallel inference for efficient online deployment.
- Employs online distillation, training LightSUAN (student) with guidance from a high-grade SUAN (teacher), combining high performance with low inference latency.

### Experiments

We conduct extensive offline experiments on three datasets, and the experimental results show that SUAN not only has excellent AUCs compared to multiple competitors but also holds the AUCs that scale with model grade and data size spanning three orders of magnitude.

Dataset	Metric	Group I		Group II				Group III		SUAN	SUAN(L)
		DIN	CAN	SoftSIM	HardSIM	ETA	TWIN	BST	HSTU		
Industry	AUC	0.7002	0.7004	0.7025	0.7020	0.7024	0.7028	0.7028	0.7036	<b>0.7098±0.00004</b>	<b>0.7135±0.00048</b>
	RelaImpr	0.00%	0.10%	1.15%	0.90%	1.10%	1.30%	1.30%	1.70%	<b>4.80%±0.02%</b>	<b>6.64%±0.24%</b>
Eleme	AUC	0.6363	0.6378	0.6399	0.6389	0.6398	0.6410	0.6600	0.6631	<b>0.6669±0.00028</b>	<b>0.6690±0.00090</b>
	RelaImpr	0.00%	1.10%	2.64%	1.90%	2.56%	3.44%	17.38%	19.66%	<b>22.45%±0.21%</b>	<b>23.99%±0.66%</b>
Taobao	AUC	0.6198	0.6184	0.6212	0.6239	0.6220	0.6215	0.6370	0.6397	<b>0.6472±0.00011</b>	<b>0.6495±0.00088</b>
	RelaImpr	0.00%	-1.17%	1.17%	3.42%	1.84%	1.42%	14.36%	16.61%	<b>22.87%±0.09%</b>	<b>24.97%±0.73%</b>

