Attend to the Active: Structure-Aware Dynamic Attention in LLMs for Compositional Instruction Following

Fangrui Lv, Yulei Qin, Ruixin Hong, Jian Liang, Jinyang Wu,
Ke Li, Xing Sun, Changshui Zhang

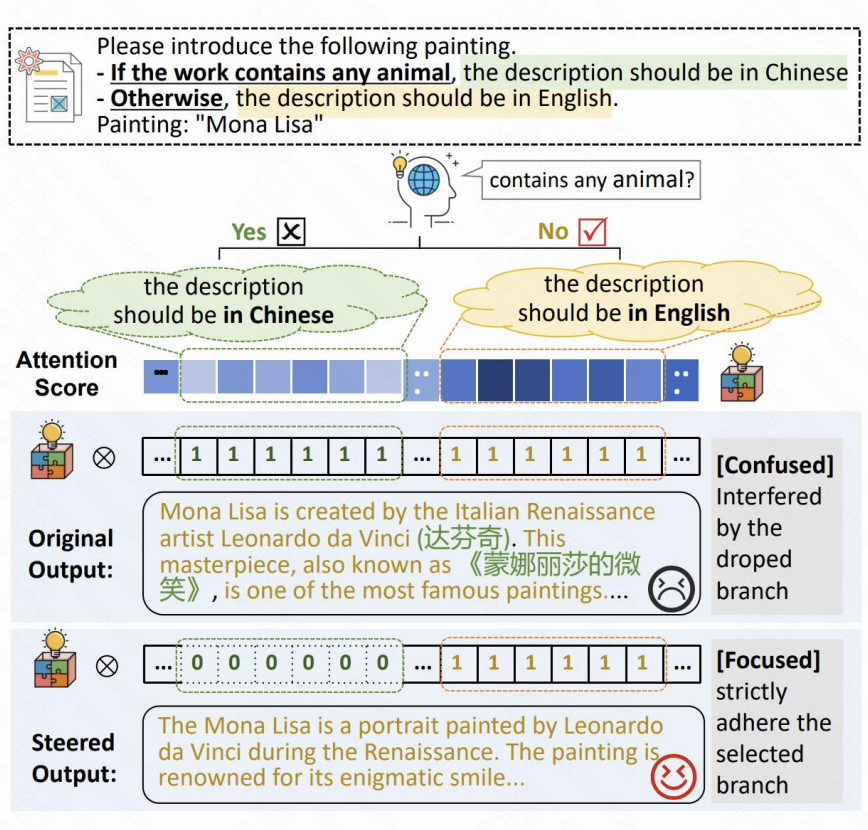
Email: lvfr23@mails.tsinghua.edu.cn

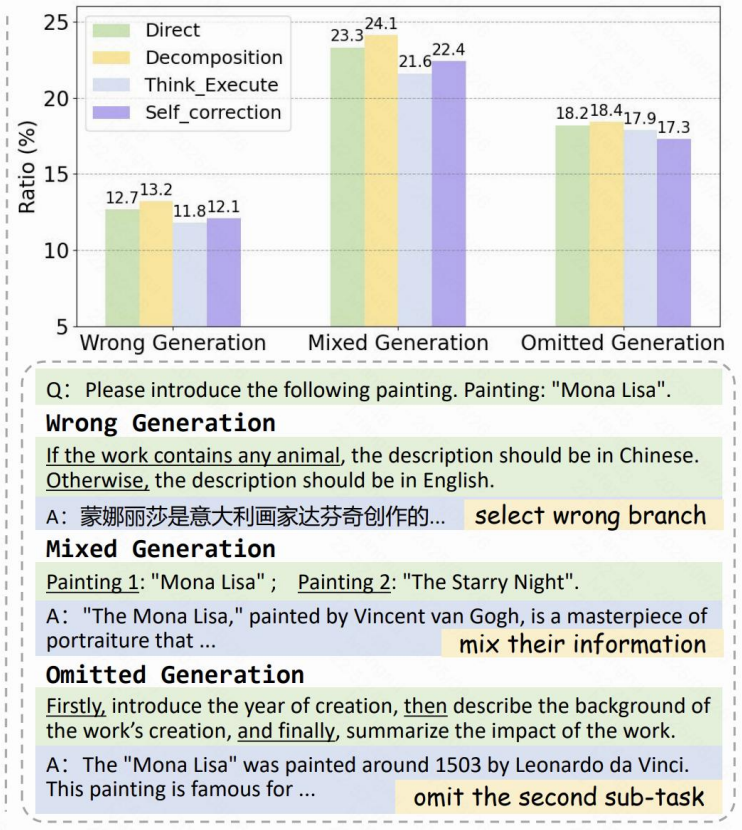
Background & Challenges

- **Background：** **Attention Distraction** in Compositional Instruction Following

  LLMs often struggle with **compositional instructions** that involve **multiple interleaved yet logically independent** sub-tasks.

- **Previous Work**:
  - ➢ Finetuning Models: require **large-scale training data** and substantial **computational overhead**;
  - ➢ High-level Planning & Iterative Self-reflection: **fail to fully suppress attention distraction** from irrelevant sub-tasks.





Research Question & Contributions

How can we effectively and efficiently eliminate attention inteference from irrelevant sub-tasks?

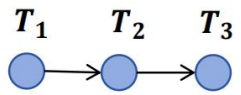
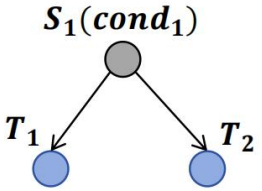
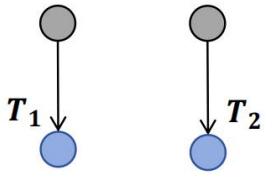
Contribution:

- **Identify three prototypical types** of composition structure;
- **First to systematically analyse** the reasons behind the performance degradation of LLMs in tackling compositional instruction;
- **Propose a Structure-Aware Dynamic Attention framework**, which dynamically and precisely steers the model's focus to attend to the active sub-task at each generation step, while suppressing attention to inactive sub-tasks.
- Comprehensive experiments across diverse benchmarks demonstrate our **effectiveness, efficiency, and generalization ability.**

Formulation

- **Structure Types of Sub-tasks**

These sub-tasks, being **structurally mutually exclusive**, dynamically alternate between **active** roles that govern the output and **dormant** roles that exert interference throughout the generation process.

Structure Type	Definition	Example	Illustration
Chain	The generation is required to complete multiple sub-tasks sequentially.	Please introduce "Mona Lisa" briefly. **Firstly,** introduce the year of creation, **then** describe the background of the work's creation, **and finally**, summarize the impact of the work.	
Branch	The generation is required to select different branches according to certain conditions.	Please introduce the following painting. - **If the work contains any animal**, the description should be in English - **Otherwise**, the description should be in Chinese. Painting: "Mona Lisa"	
Parallel	The generation is required to complete multiple independent tasks in parallel .	Please introduce the following paintings. - **Painting 1**: "Mona Lisa" - **Painting 2**: "The Starry Night"	

Methods

- **Structure Analysis and Identification**

  - identify the structure type S;
  - extract the sub-tasks $\mathcal{T} = \{T_1, \ldots T_m\}$

$$S, T_1, T_2, \cdots, T_m = \mathcal{LLM}(\mathcal{T}|\mathcal{P}),$$

  - ✓ serves as a reliable guidance to conduct **precise** and **safe** attention steering:
    - attention steering is confined to the identified mutually exclusive sub-tasks;
    - attention masking is performed at the granularity of independent and complete sub-task

Prompt template $\mathcal{P}$ of structure identification

You are an excellent logic expert. Given a compositional task and the definition of structure types, please determine the structure label of the task and identify the corresponding sub-tasks.

Structure Type:
- Chain: The task is required to complete multiple sub-tasks sequentially.
- Branch: The task is required to select different branches according to certain conditions.
- Parallel: The task is required to complete multiple independent sub-tasks in parallel.

Output the structure type of input task (chosen from "Chain", "Branch", "Parallel") and the exactly sub-tasks (without modifying, deleting or adding any original text) in the following json format:
{ "Structure Type": "", "Sub-tasks": [sub-task1, sub-task2, ...] }

Task: {input task}

Methods

- **Structure-aware Dynamic Attention**

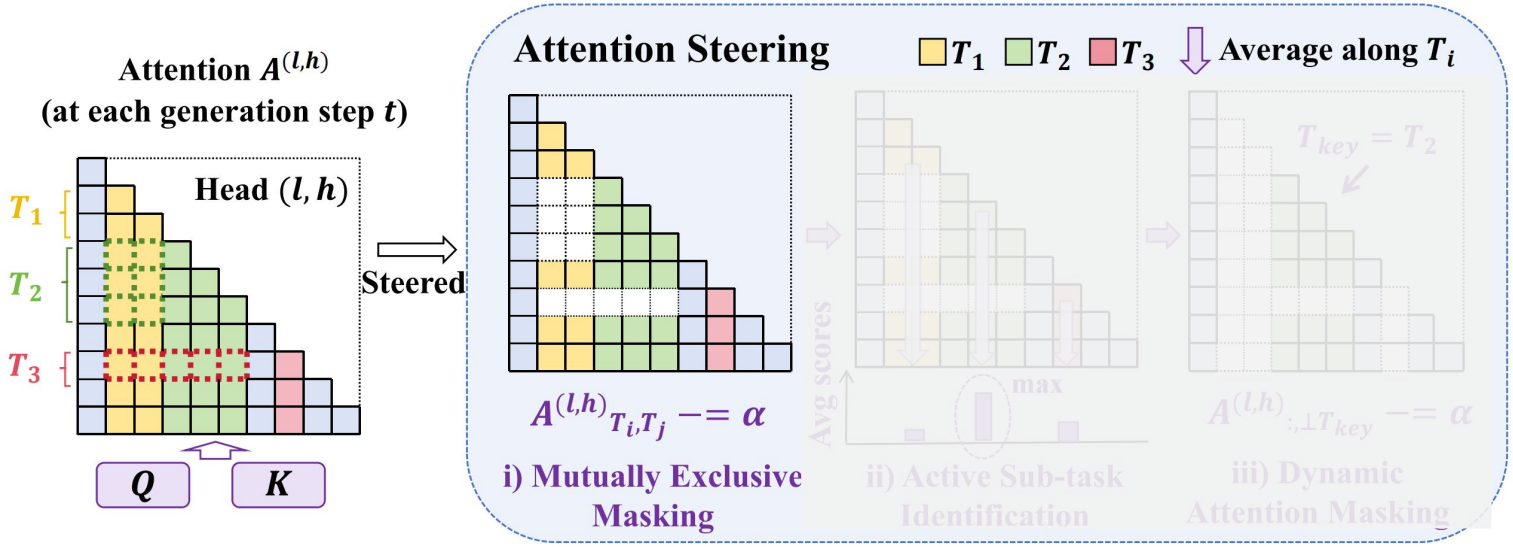
  ➢ Step1: Mutually Exclusive Masking

  - information interaction between the mutually exclusive sub-tasks may introduce blending comprehension;
  - **mask attention between exclusive sub-task pairs**, i.e., $T_i, T_j, i \neq j$, to mitigate potential mutual interference

  by **adopting a bias matrix M** on the multi-head attention mechanism:

$$\boldsymbol{H}^{(l,h)} = \text{Softmax}\left(\boldsymbol{A}^{(l,h)} + \boldsymbol{M}^{(l,h)}\right)\boldsymbol{V},$$

$$\boldsymbol{M}^{(l,h)}(T_i, T_j) = \begin{cases} -\alpha, & T_i \perp T_j \\ 0, & \text{otherwise} \end{cases}.$$



Methods

- **Structure-aware Dynamic Attention**

  ➢ Step2: Active Sub-task Identification

  - identify the active sub-task T_{key} at each generation step that **dominates next token prediction**:

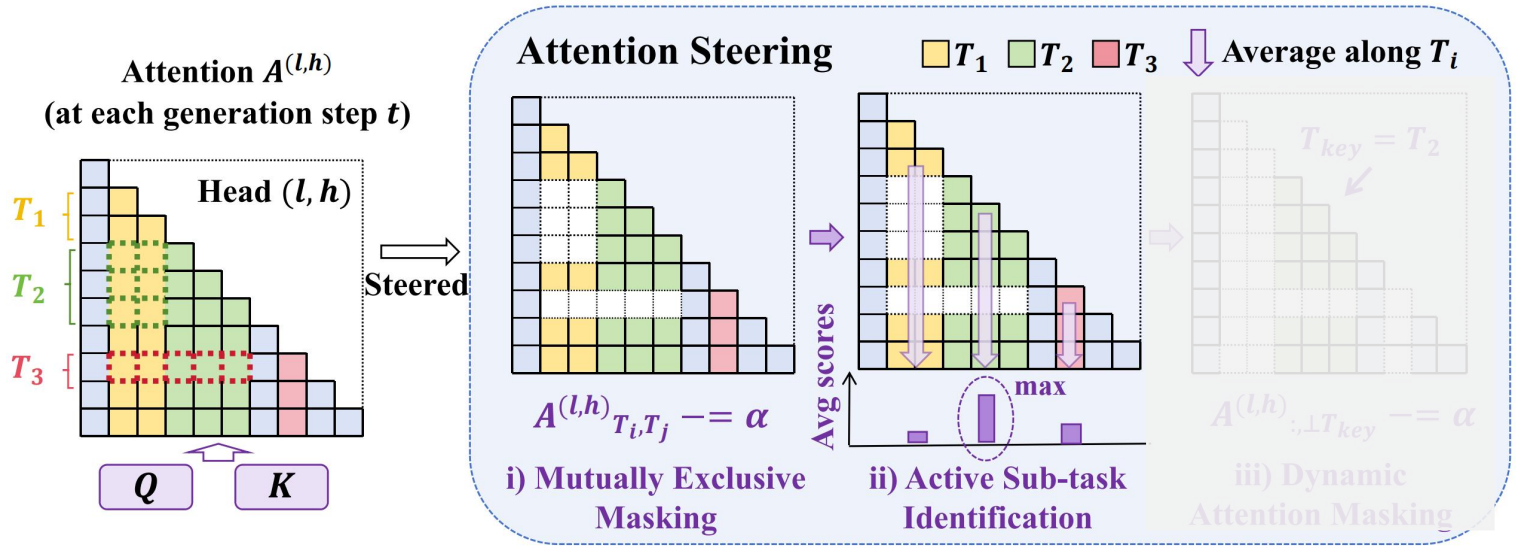
    ① **highest score**: attracts most attention from subsequent tokens.

    ② **low entropy**: deterministic and reliable identification.

    ③ **safe progression**: only permits active sub-task transitions that aligns the global structure logic.

$$score(T_i, t) = \frac{1}{|T_i|} \sum_{k \in T_i} \frac{1}{t-k} \sum_{k \le q \le t} A^{(l,h)}(q,k),$$

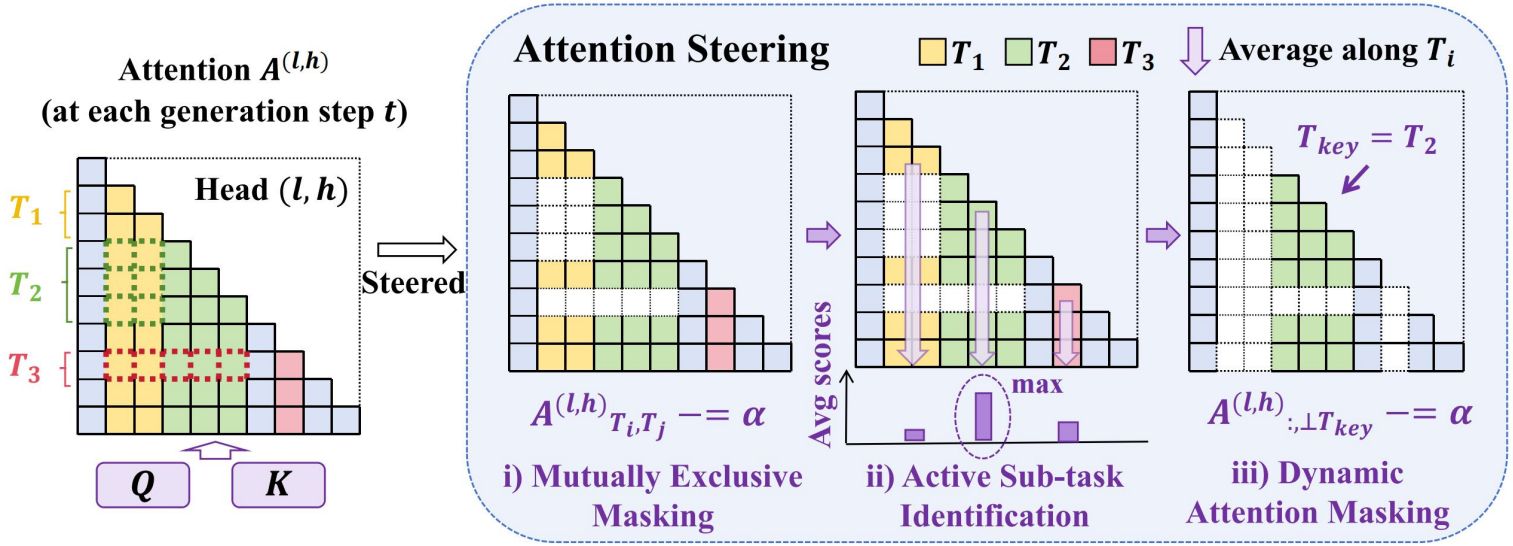
$$T_{key} = argmax_{T_i}(score(T_i, t)), \quad s.t. H([score(T_1, t), score(T_2, t), ...]) < \epsilon,$$



Methods

- **Structure-aware Dynamic Attention**
  - ➢ Step3: Dynamic Attention Masking
    - At each generation step, **suppressing the attention to inactive sub-tasks** to eliminate their interference
    - Ensuring **faithful and consistent execution** of current active sub-task
    - effectively mitigates attention interference and enhances model adherence to compositional instructions.

$$M^{(l,h)}(:, T_j) = \begin{cases} -\alpha, & T_j \perp T_{key} \\ 0, & \text{otherwise} \end{cases},$$



Experimental Results

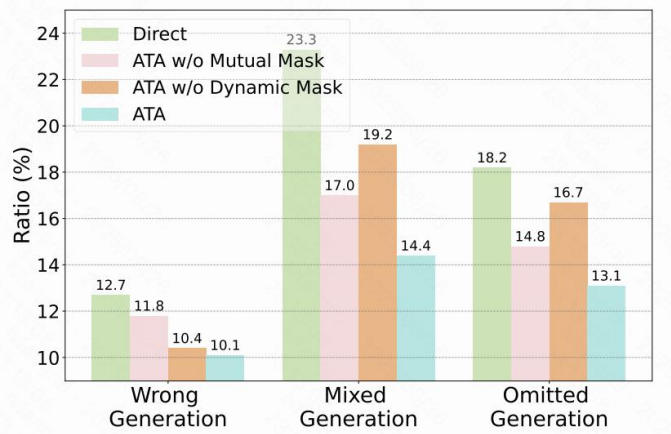
- Consistent Improvement across Composition Structure Types

Table 1: Performance of ATA and baselines on compositional instructions. The second best performance is underlined, while the best performance is **bold**.

Structure Type	**Chain**			**Branch**				**Parallel**				**All**
Complexity	2	3	Avg.	2	3	≥ 4	Avg.	2	3	≥ 4	Avg.	Avg.
Llama-3-8B-Instruct												
Direct I/O	66.93	56.82	59.21	59.74	55.26	48.32	54.63	61.63	60.17	55.83	59.80	57.88
CoT Prompting	64.05	55.73	57.69	56.62	55.22	53.39	55.02	64.17	67.05	62.08	64.97	59.23
Decomposition	61.19	52.61	54.63	53.88	51.49	45.66	50.04	62.98	66.74	56.26	62.58	55.75
Think-Execute	65.94	59.28	60.85	57.49	52.26	50.25	52.92	63.4	68.33	57.09	63.14	58.97
Self-correction	66.31	60.33	61.74	61.27	56.38	50.21	55.24	63.37	67.92	51.39	61.25	59.41
ATA	**69.26**	**61.04**	**62.98**	**64.38**	**58.42**	**53.79**	**58.74**	**72.50**	**71.03**	**65.42**	**69.91**	**63.88**
Mistral-7B-Instruct												
Direct I/O	57.41	55.68	56.31	55.40	49.37	41.19	48.83	37.87	35.24	31.50	34.89	46.67
CoT Prompting	55.78	53.64	54.19	57.14	51.24	42.98	50.45	40.45	39.07	33.89	37.62	47.42
Decomposition	52.50	51.17	51.62	48.74	44.26	37.59	43.77	40.26	38.44	32.15	37.41	44.27
Think-Execute	56.48	55.32	55.89	57.33	50.15	41.83	49.26	39.86	37.95	32.61	36.72	47.28
Self-correction	59.62	55.94	56.94	56.26	47.19	40.64	47.65	38.62	36.43	31.88	35.22	46.60
ATA	**61.73**	**57.28**	**58.37**	**60.44**	**52.93**	**43.76**	**52.16**	**44.68**	**41.77**	**37.59**	**41.34**	**50.62**

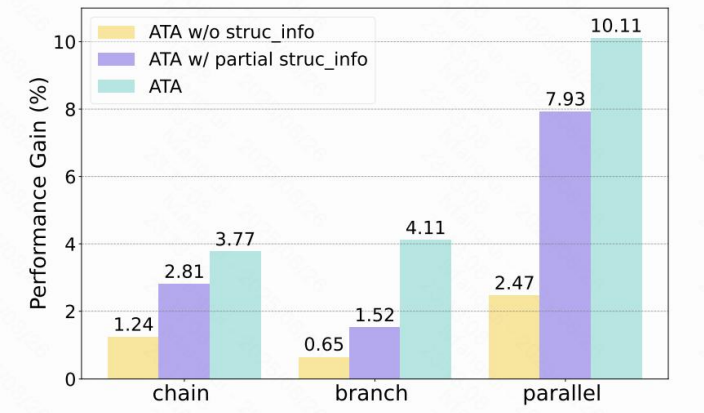
Ablation & Analysis

- ATA effectively **reduces all types of generation errors**



(a) Reduction of Generation Error (↓).

- ATA's performance is **robust to imperfect structure identification**



(b) Performance Impact of Structure Information (↑).

Ablation & Analysis

- **Structure information** serves as a safe guidance for attention steering;
- **Both attention masking module** of ATA is effective in improving model adherence;
- **The active control strategy** mitigates errors from occasional attention drift and ensure valid attention steering

Method	Chain	Branch	Parallel
Direct	59.21	54.63	57.88
ATA	**62.98**	**58.74**	**69.91**
- w/o Structure Info	60.45	55.28	60.35
- w/o Mutual Mask	61.82	57.03	67.42
- w/o Dynamic Mask	60.74	56.65	64.31
- w/o Active Control	61.14	57.23	66.82

Table 3: Effectiveness of Attention Steering.

Steering (St.) Strategy	Chain	Branch	Parallel
Direct (No St.)	59.21	54.63	57.88
SampleAttention	60.47	51.92	57.71
PASTA	61.24	54.16	60.43
ATA	**62.98**	**58.74**	**69.91**
- w/ Misguided St.	56.43	47.86	54.12
- w/ Random St.	57.38	51.04	56.81

Ablation & Analysis

- **Generalization** on Nested Structures & Larger LLMs

	Structure Type	Direct	ATA
Llama3-8B	Chain & Branch	48.42	**51.95**
	Chain & Parallel	50.17	**54.72**
	Branch & Parallel	46.83	**48.41**
Llama3-13B	Chain & Branch	52.26	**56.64**
	Chain & Parallel	56.48	**59.30**
	Branch & Parallel	51.65	**53.88**

- **Time Efficiency** of ATA

Complexity	2	3	≥ 4
Direct	9.69 (s)	14.91 (s)	18.66 (s)
ATA	10.14 (s)	15.83 (s)	19.95 (s)
Time overhead gain (relative)	4.6%	6.2%	6.9%

Conclusion

- **First work** to sysmetically analyse the reason behind the performance degradation of LLMs in tackling compositional instruction and propose to tackle this challenge by dynamic attention steering :
  - **Superior performance**, effectively mitigates attention interference and enhances model adherence to compositional instructions.
  - **Higher efficiency**, operates within a single forward pass without requiring parameter updates.
  - **More flexible**, plug-and-play compatibility with various off-the-shelf LLMs and existing planning-based techniques.

- **Future works**: to further broaden ATA's applicability
  - Incorporate with other structure-discovery techniques to tackle challenging scenarios where structure is implicit;
  - How to conduct structure-agnostic attention steering.

ICLR

Thank You!

Fangrui Lv, Yulei Qin, Ruixin Hong, Jian Liang, Jinyang Wu,
Ke Li, Xing Sun, Changshui Zhang

Email: lvfr23@mails.tsinghua.edu.cn