SOUTHEAST UNIVERSITY

NetLLM: Adapting Large Language Models for Networking

Duo Wu[1], Xianda Wang[1], Yaqi Qiao[1], Zhi Wang[2], Junchen Jiang[3], Shuguang Cui[1], Fangxin Wang[1]*

The Chinese University of Hong Kong, Shenzhen
Tsinghua University，The University of Chicago

SIGCOMM 2024

香港中文大學（深圳）
The Chinese University of Hong Kong, Shenzhen

汇报人：董兵

2024 年 11月 10日

提纲

1 Background

2 Method

3 Evaluation

4 Conclusion

提纲

1 Background

2 Method

3 Evaluation

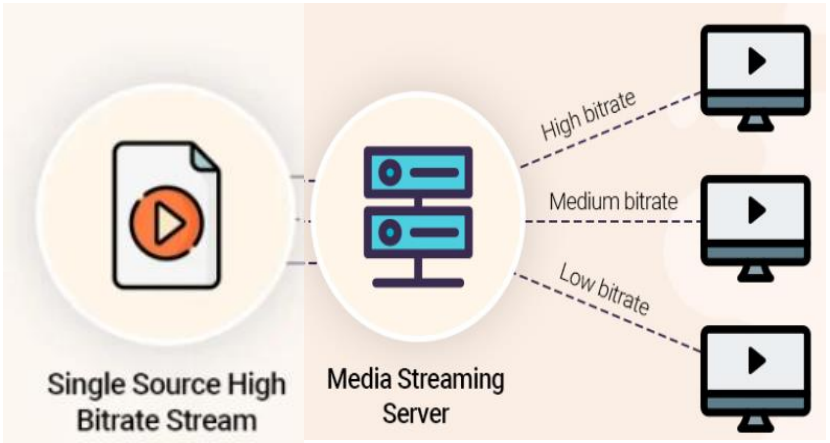
4 Conclusion

Background

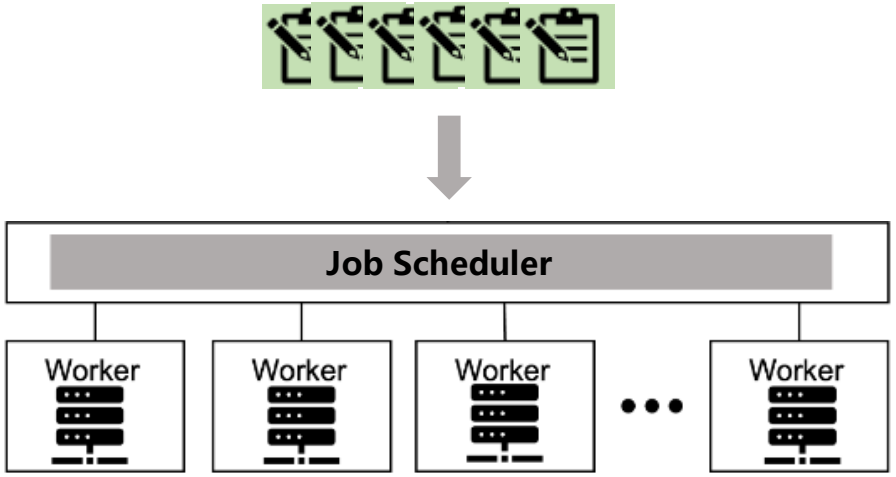
Network optimization tasks：

Viewport Prediction (VP)

Adaptive Bitrate Streaming (ABR)



Cluster Job Scheduling （CJS）



Prediction tasks

Decision-making tasks

Background: Deep Learning in Networking

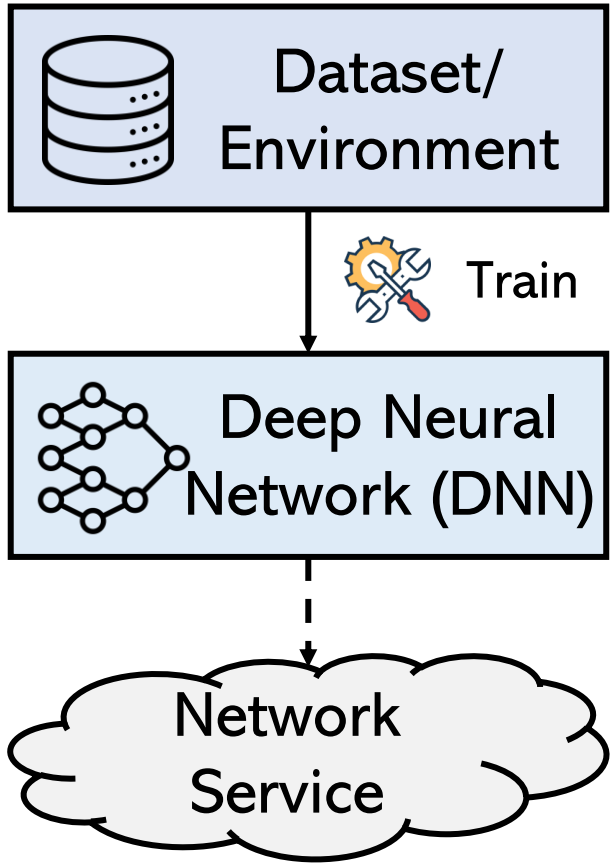
Deep learning (DL) has been used to solve complex prediction and decision-making tasks in networking.

Prediction tasks

- Traffic classification (WWW '22)
- Bandwidth prediction (NSDI '20)
- **Viewport prediction** (TPAMI '22)

Decision-making tasks

- Congestion control (SIGCOMM '20, '23)
- **Adaptive bitrate streaming** (SIGCOMM '17, '22)
- **Cloud cluster job scheduling** (SIGCOMM '19)

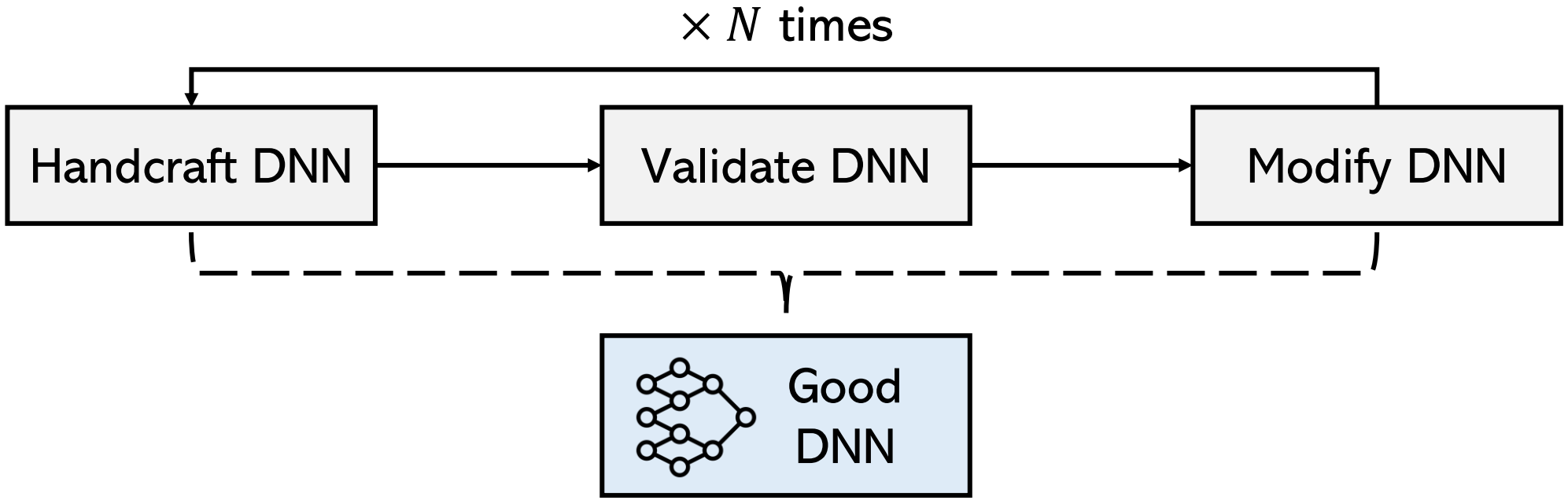


DL in networking suffers from limits

Limitation：Labor-intensive model engineering

The success of DL-based algorithms heavily relies on the manual design of black-box DNNs [1].

- Trial-and-error design manner.



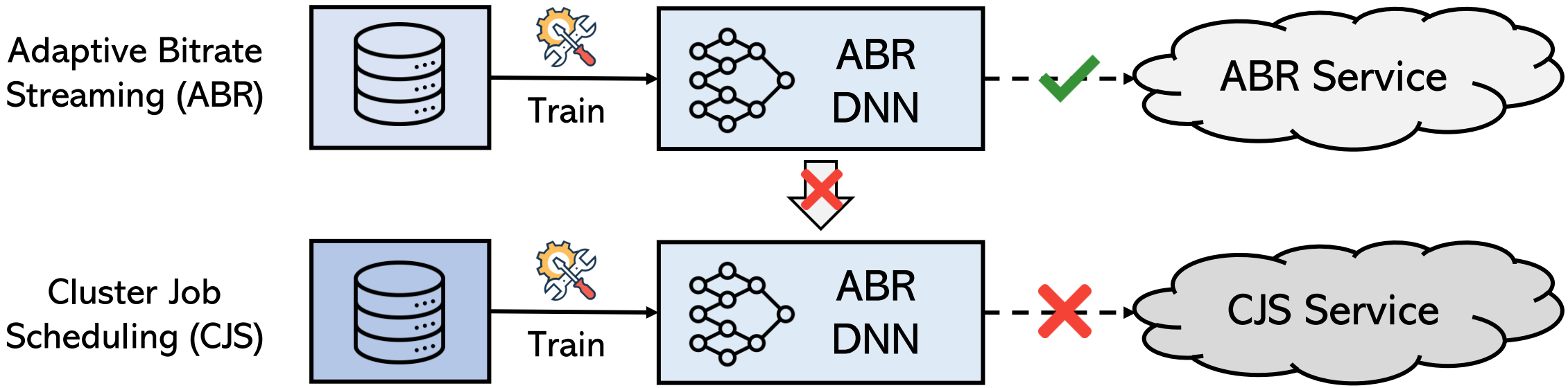
[1] Risto Miikkulainen, et al.. “Evolving deep neural networks”. In Artificial Intelligence in the Age of Neural Networks and Brain Computing (Second Edition). Academic Press, 269–287, 2024.

Limitation： Labor-intensive model engineering

The success of DL-based algorithms heavily relies on the manual design of black-box DNNs [1].

- Trial-and-error design manner
- DNN is not shareable across tasks

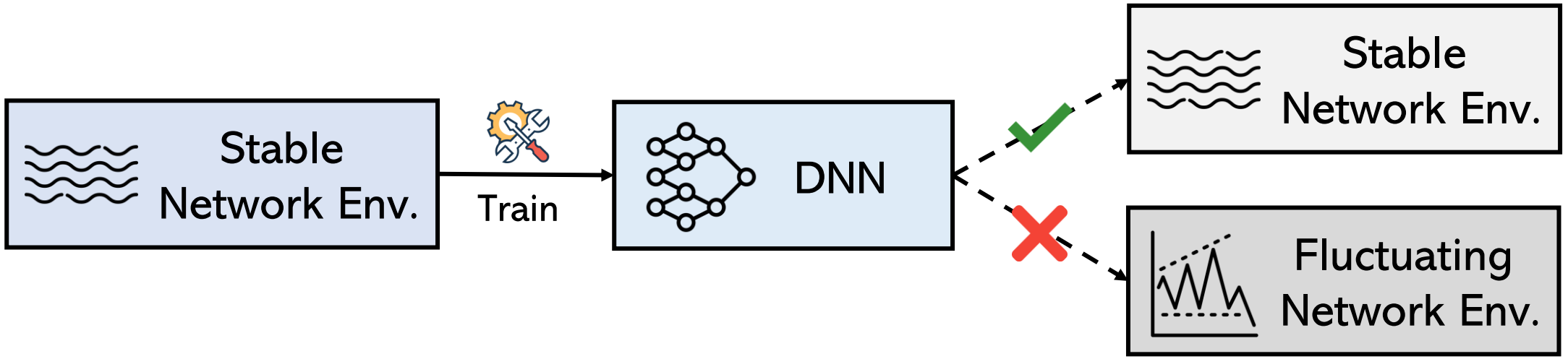
One model only for one task



[1] Risto Miikkulainen, et al.. “Evolving deep neural networks”. In Artificial Intelligence in the Age of Neural Networks and Brain Computing (Second Edition). Academic Press, 269–287, 2024.

Limitation: Poor generalization

DL-based algorithms tend to achieve poor generalization on unseen data distributions or environments.

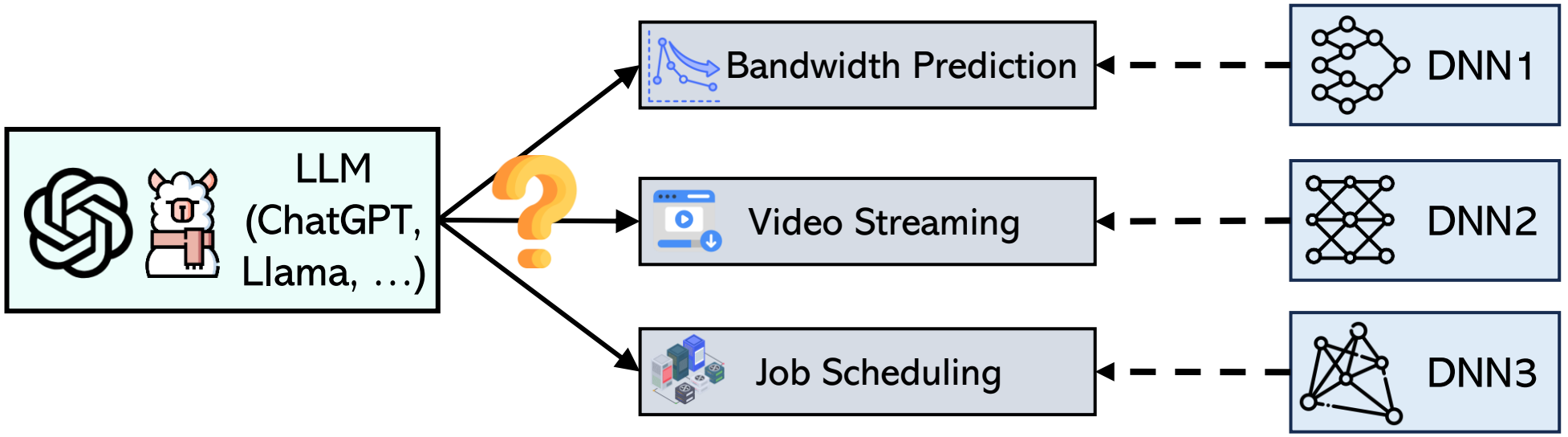


Can we use **one model for all networking tasks** with even **stronger generalization**?

Opportunity: Using LLM in Networking

Large language model (LLM), e.g., ChatGPT, Llama.

- LLM exhibits some emergent abilities.
- These abilities prove to be transferable across domains.



Key question: Can we efficiently use the LLM to achieve "one model for all networking tasks" with even better performance?

Challenge：Using LLM in Networking

Challenge 1: How to enable the LLM to understand networking information?

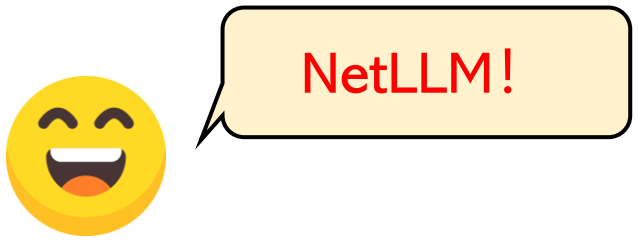
- Large input modality gap （Convert to text❌）

Challenge 2: How to enable the LLM to generate answers for networking efficiently?

- Inefficiency of token-based answer generation , Invalid answers

Challenge 3: How to fine-tune the LLM to learn networking knowledge efficiently?

- High adaptation costs



提纲

1 Background

2 Method

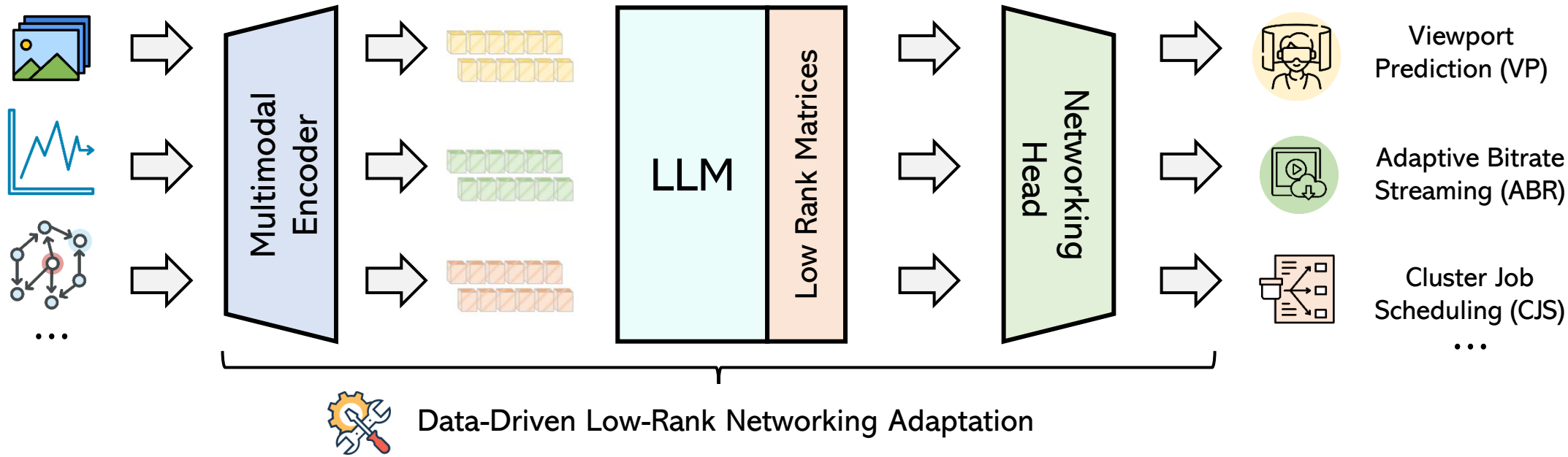
3 Evaluation

4 Conclusion

Method: Overview of NetLLM

NetLLM: the first framework to adapt the LLM to solve networking tasks with low efforts.

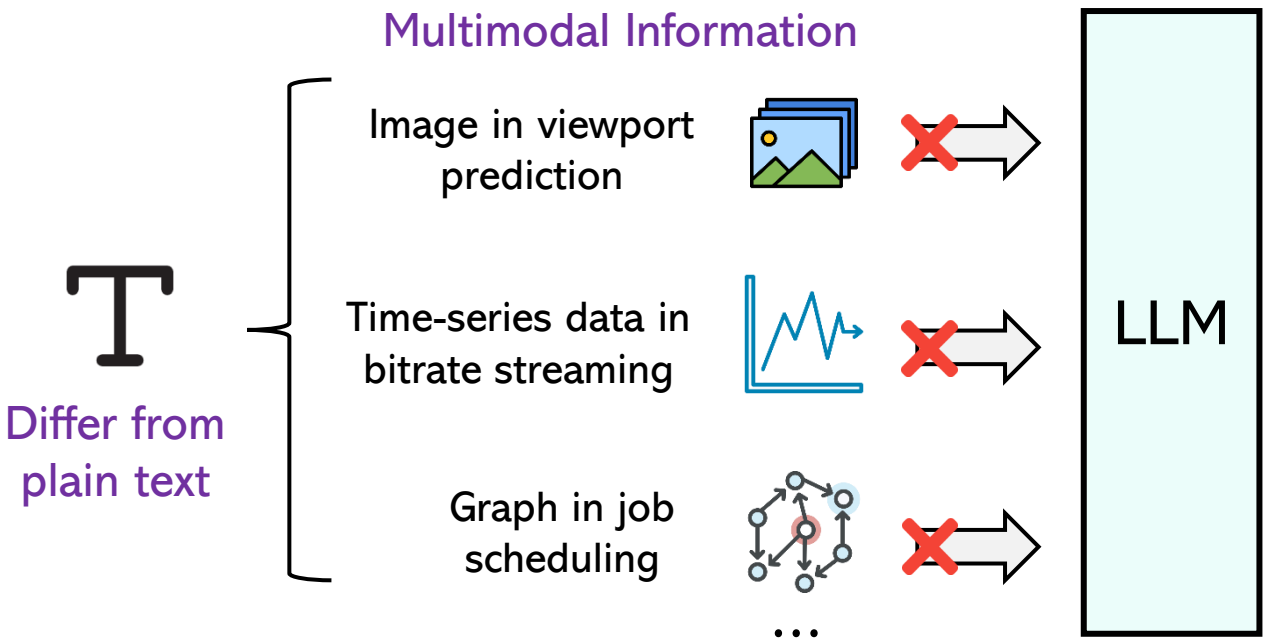
- Multimodal encoder (let LLM understand networking information effectively)
- Networking head (let LLM generate answers for networking efficiently)
- Data-driven low-rank networking adaptation (let LLM learn networking knowledge efficiently)



Method: Multimodal Encoder

Challenge 1: How to enable the LLM to understand networking information?

- Large input modality gap (Convert to text ❌)

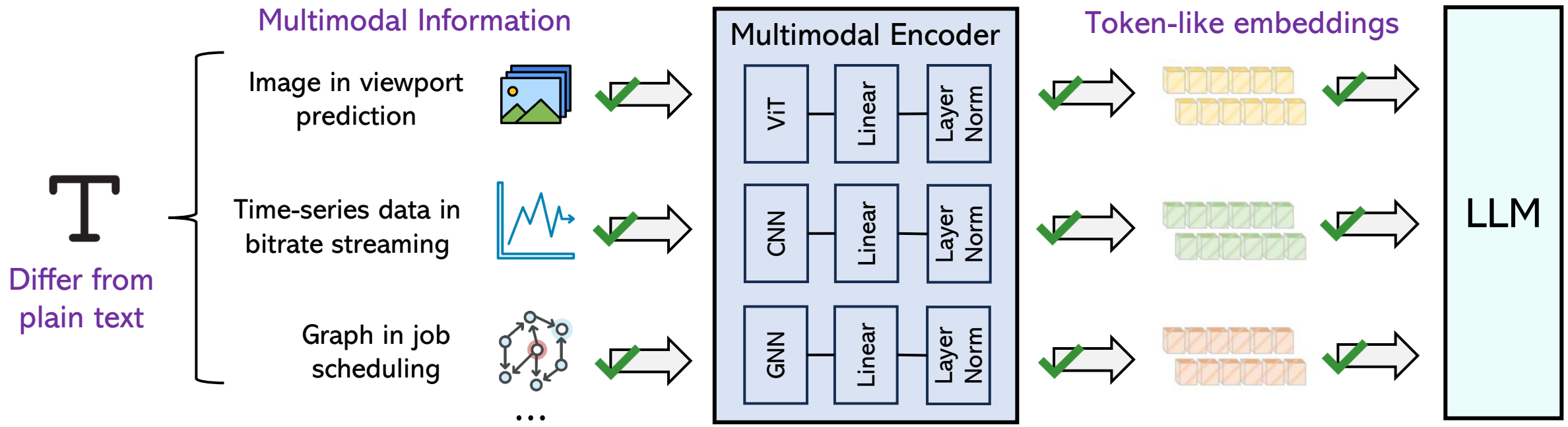


Method: Multimodal Encoder

Challenge 1: How to enable the LLM to understand networking information?

- Large input modality gap (Convert to text ❌)

Multimodal encoder: Project data into the same feature space as texts.

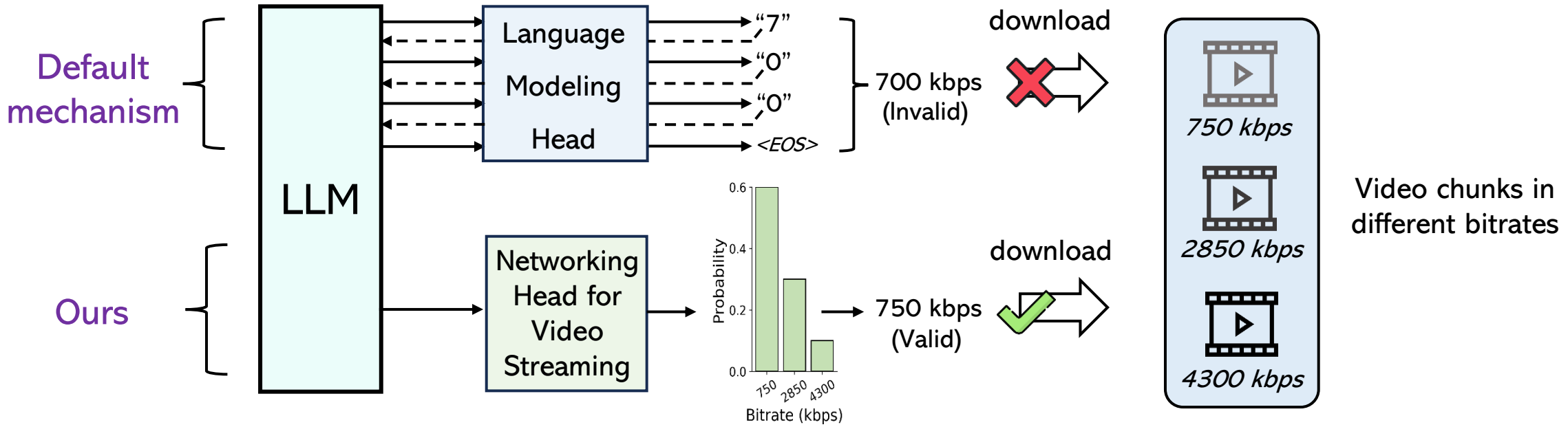


Method: Networking Head

Challenge 2: How to enable the LLM to generate answers for networking efficiently?

- Inefficiency of token-based answer generation , Invalid answers

Networking head: a linear output layer to generate task-specific answers directly.

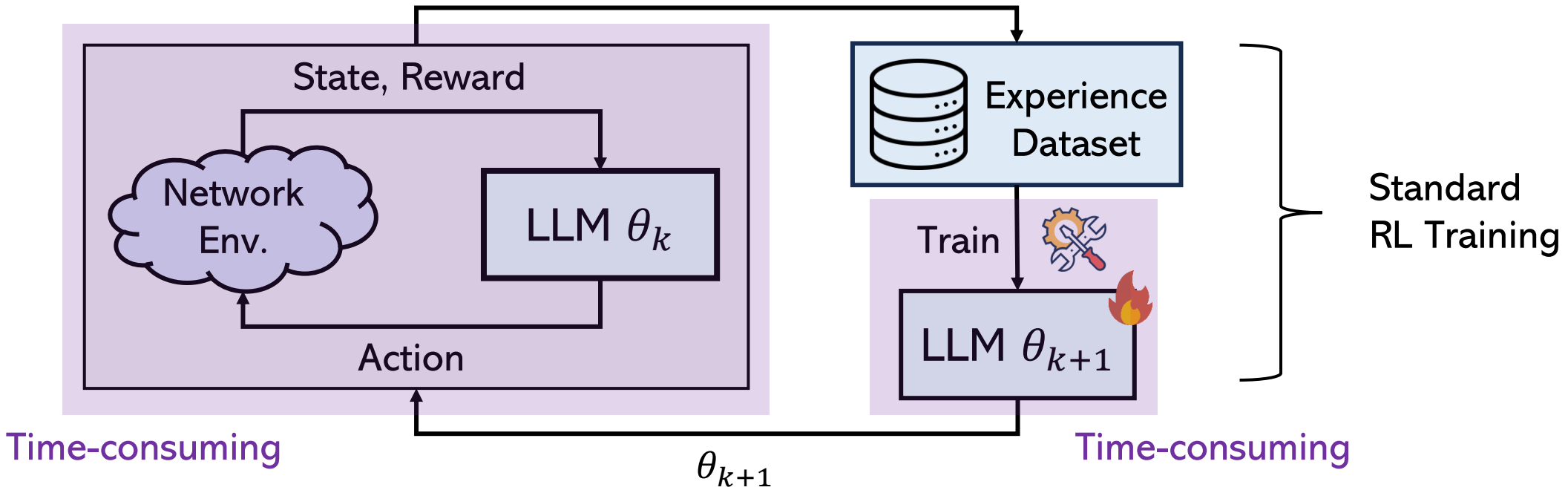


Method: Data-Driven Low-Rank Networking Adaptation

Challenge 3: How to fine-tune the LLM to learn networking knowledge efficiently?

The costs of fine-tuning the LLM are expensive because of the large parameter size.

- Standard reinforcement learning (RL) requires active environment interaction.
- Fully fine-tuning the LLM is expensive.



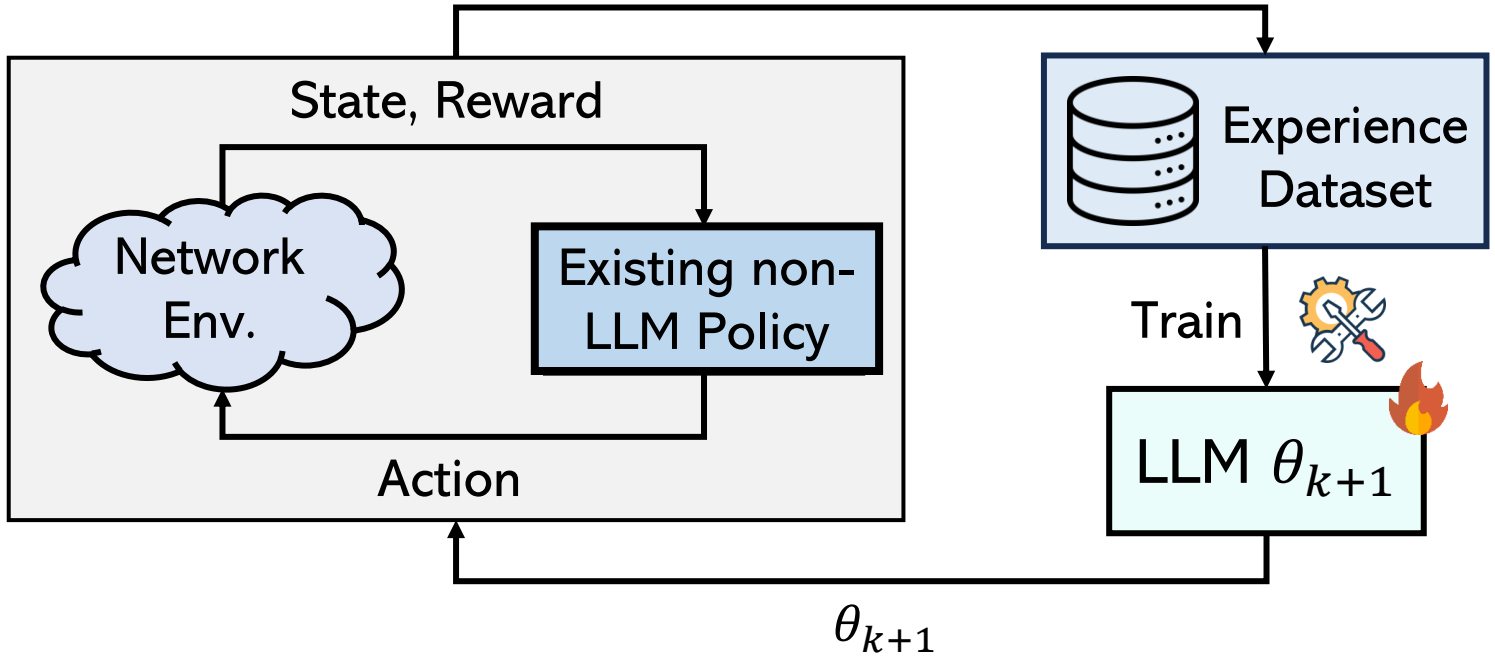
Method: Data-Driven Low-Rank Networking Adaptation

Challenge 3: How to fine-tune the LLM to learn networking knowledge efficiently?

DD-LRNA: Significantly reduce the fine-tuning costs of LLM.

- Remove interactions between LLM and environments based on data-driven RL.
- Insert low-rank matrices to reduce the number of trainable parameters.

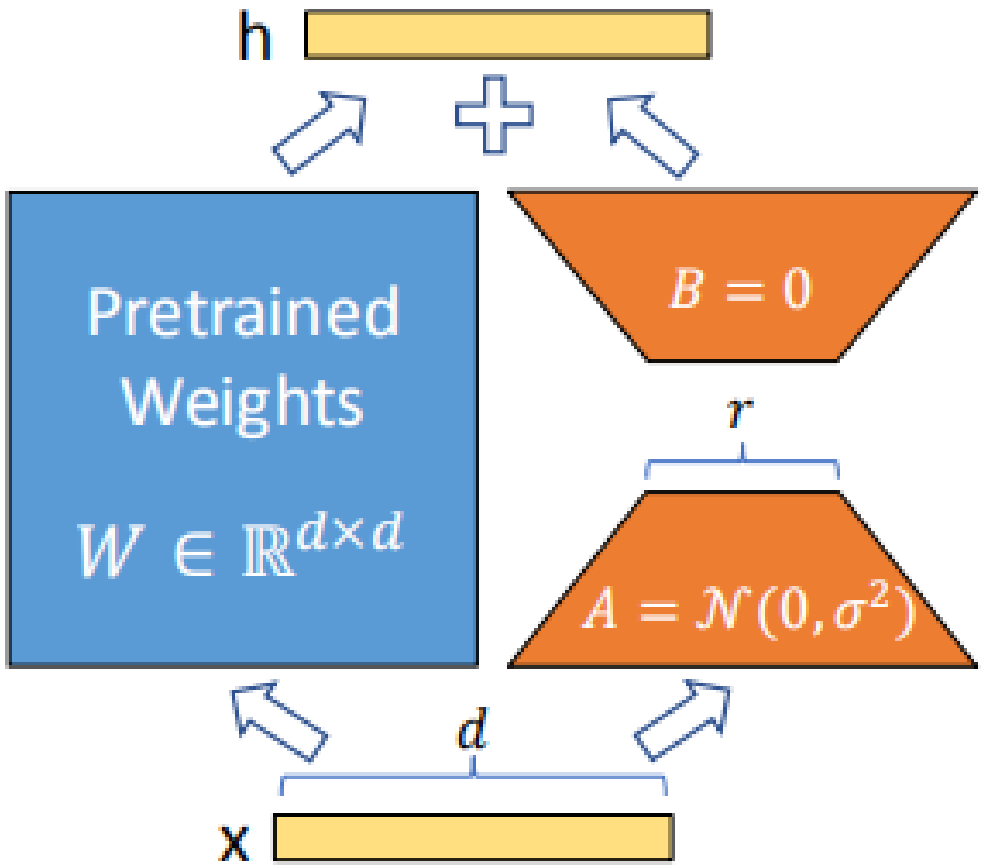
Reduce >30% training time



Method: Data-Driven Low-Rank Networking Adaptation

Challenge 3: How to fine-tune the LLM to learn networking knowledge efficiently?

LoRA: Low-Rank Adaptation



LLM: W_0

Fine-tuning: $W_0 + \Delta W$

LoRA: $W_0 + \Delta W = W_0 + BA$

$B \in R^{d \times r}$, $A \in R^{r \times k}$, $r \ll \min(d, k)$

Inference: $h = W_0 x + \Delta W x = W_0 x + BAx$

Method: Data-Driven Low-Rank Networking Adaptation

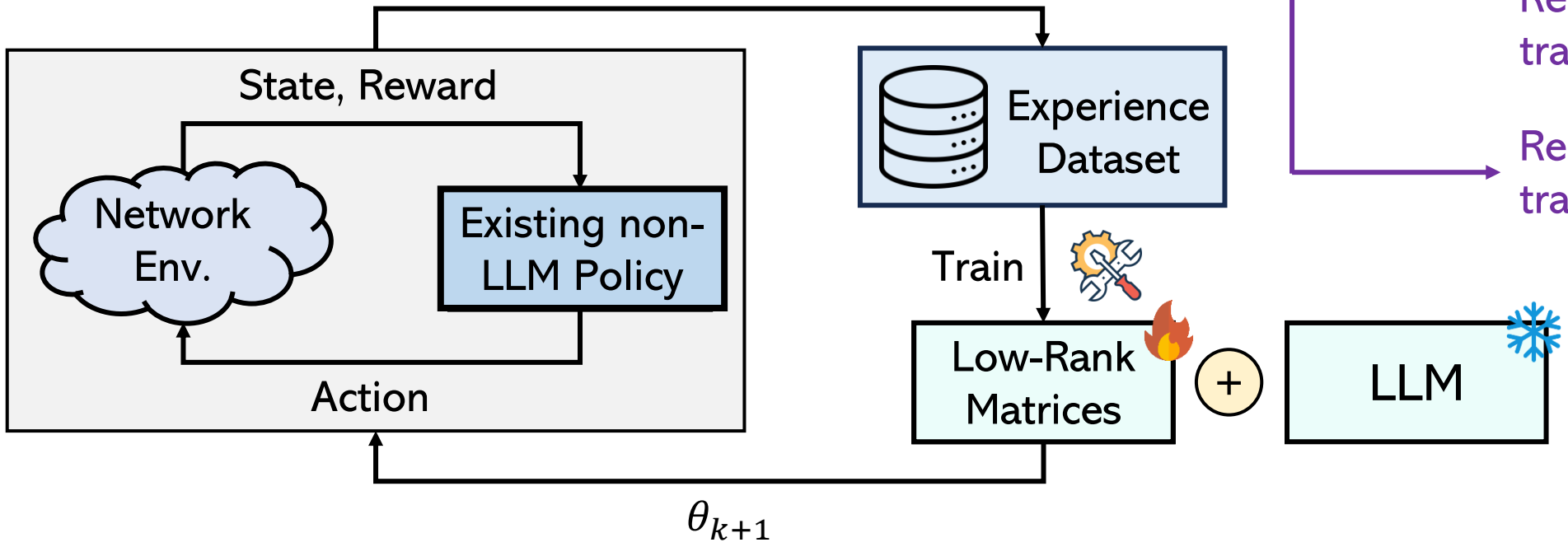
Challenge 3: How to fine-tune the LLM to learn networking knowledge efficiently?

DD-LRNA: Significantly reduce the fine-tuning costs of LLM.

- Remove interactions between LLM and environments based on data-driven RL.
- Insert low-rank matrices to reduce the number of trainable parameters.

Reduce >30% training time

Reduce >15% training time



提纲

1 Background

2 Method

3 Evaluation

4 Conclusion

Evaluation：Setup

Tasks: Viewport Prediction (VP), Adaptive Bitrate Streaming (ABR), Cluster Job Scheduling (CJS).

Metrics: Mean absolute error (MAE) for VP, Quality of Experience (QoE) for ABR, job completion time (JCT) for CJS.

LLM: We use Llama2-7B as the default LLM.

Baselines:

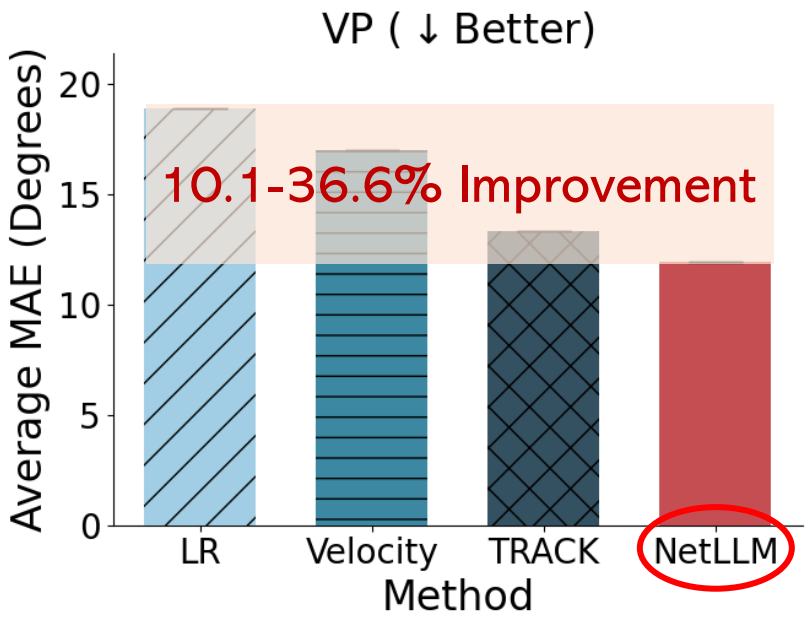
- DNN-based: TRACK(TPAMI'21) | GENET(SIGCOMM'22) | Decima(SIGCOMM'19)
- Rule-based: LR(Mobicom'18)、Velocity(TVCG'21) | BBA(SIGCOMM'14)、MPC(SIGCOMM'15) | FIFO、Fair

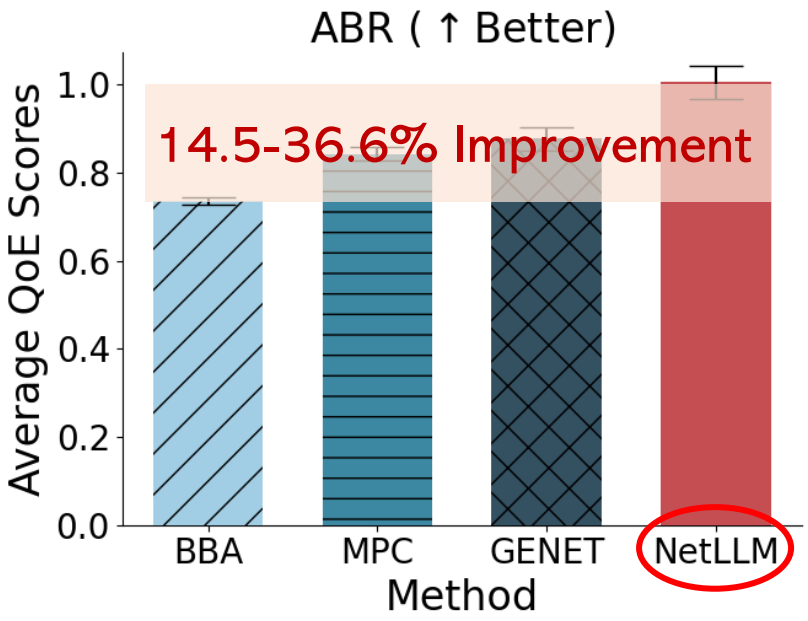
Datasets： Jin2022(SIGMM)、Wu2017(MMSys) | Envivio-Dash3、FCC(bandwidth trace) | TPC-H(ISPASS'16)

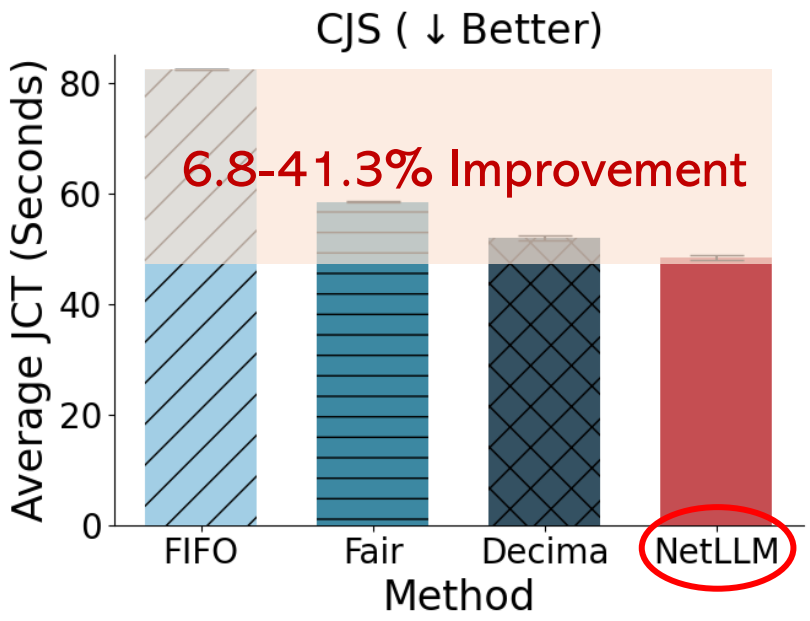
Hardware: A Linux server equipped with eight Intel(R) Xeon(R) Gold 5318Y CPUs and two NVIDIA 40GB A100 GPUs.

Evaluation: General Evaluation

Testing setting = Training setting

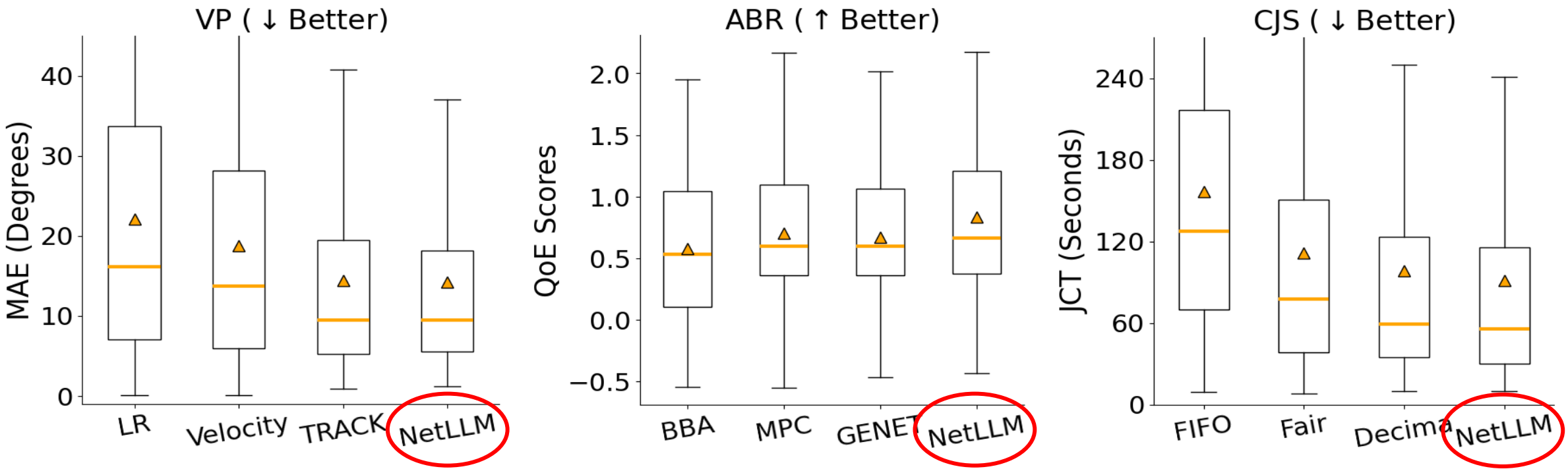






- Demonstrate the effectiveness of NetLLM in adapting the LLM for networking.
- Showcase the potential of "one model for all tasks with better performance"!

Evaluation: Generalization Evaluation

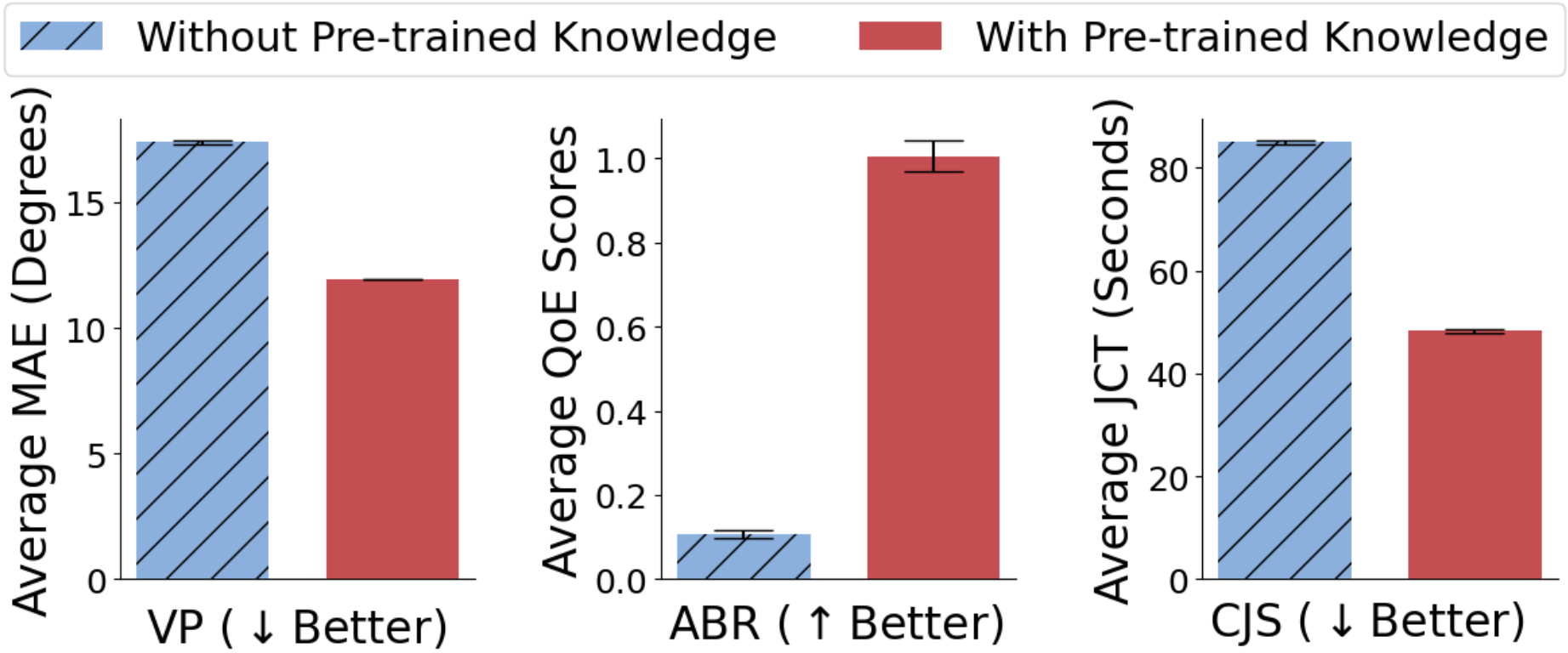


NetLLM-adapted Llama2 significantly outperforms baselines across all cases!

With NetLLM, we can indeed efficiently utilize the extensive knowledge of the LLM to achieve stronger generalization.

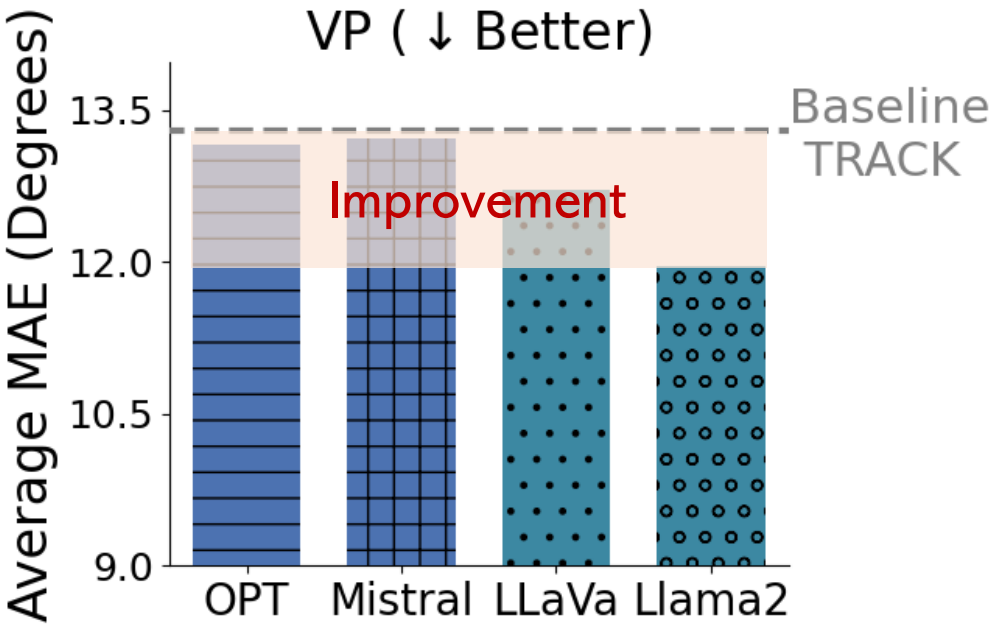
Evaluation: Importance of Pre-trained Knowledge

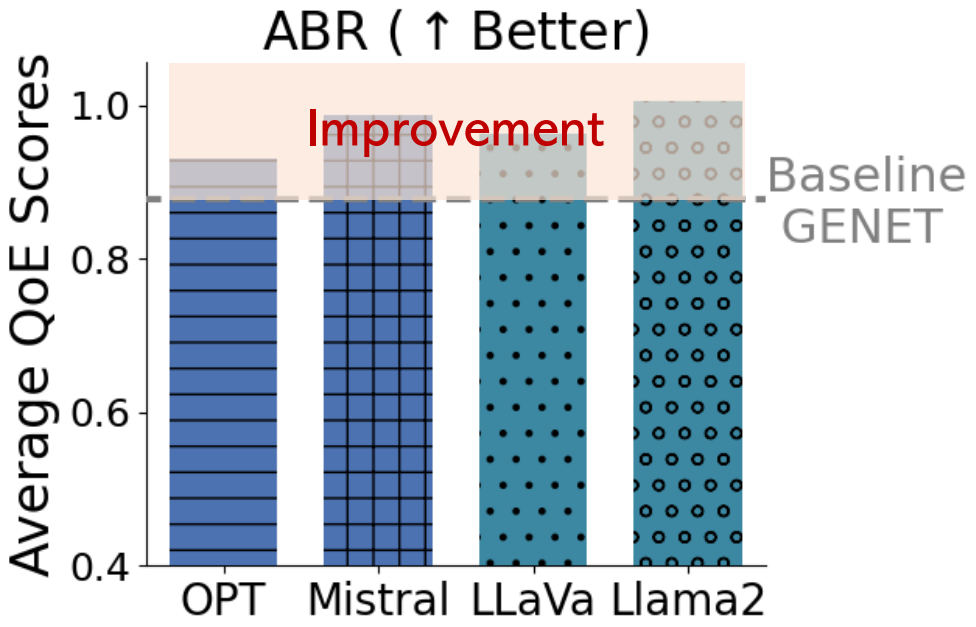
Without Pre-trained Knowledge: Random initialize LLM instead of using its pre-trained parameters



Pre-trained knowledge is important, indicating that LLM indeed contains some common knowledge useful for networking.

Evaluation: Impacts of Different Types of LLMs



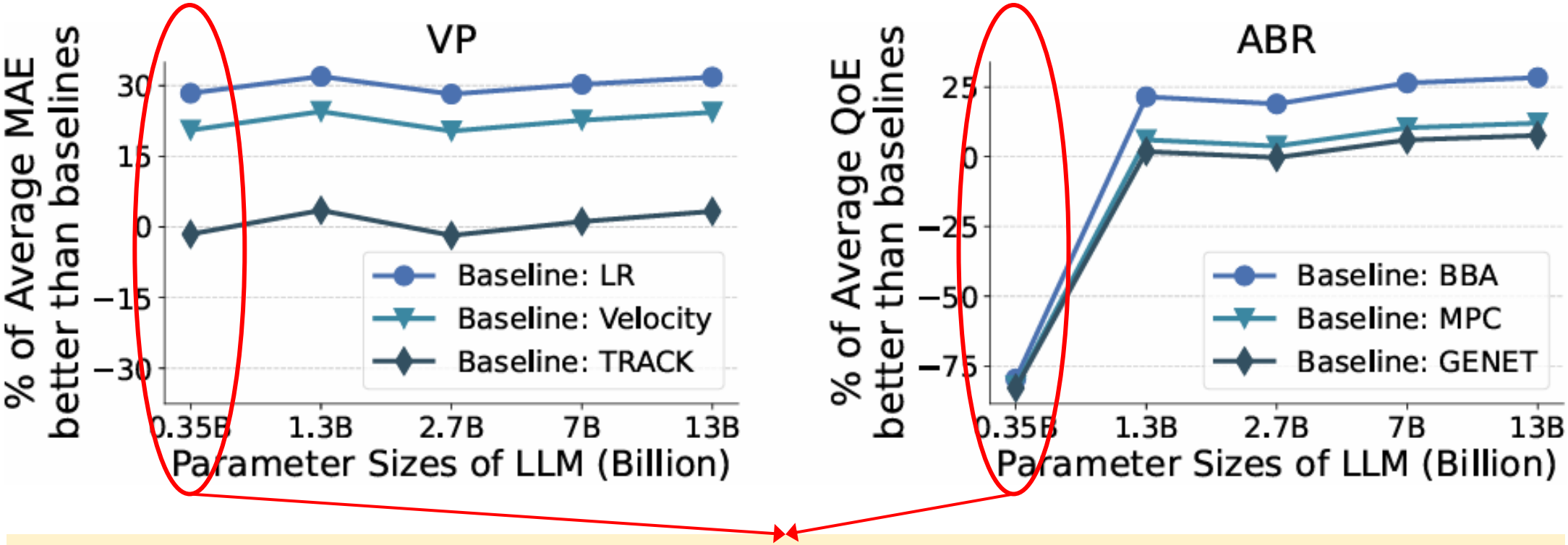


All the adapted LLMs outperform the state of the arts on both tasks.

Different LLMs can be adapted to solve networking tasks with our NetLLM framework

Evaluation: Impacts of LLM Parameter Size

We use OPT as the LLM for this investigation, which offers different versions with varying parameter sizes.



The LLM of 0.35 billion parameters performs well on VP but poorly on ABR.

Indicating that 1B could be the boarder line of using LLM for networking.

提纲

NetLLM

- **Problem: NetLLM利用单个LLM在多个网络任务中实现有效泛化。**
  - **Input:网络任务Viewport Prediction(VP)、Adaptive Bitrate Streaming (ABR)、Cluster Job Scheduling（CJS）；**
  - **Output:VP预测/，ABR、 CJS决策；**
  - **Significance: 将LLM作为网络的基础模型，利用LLM强大的泛化能力，单个模型解决不同网络任务；**
- **SoA & Limitations:**
  - **L1： High model engineering costs：从传统规则工程转向模型工程，网络任务的多样性无法实现不同任务共享DNN模型；**
  - **L2： Low generalization：模型表现能力易受训练数据影响；**
- **Opportunity:**
  - **O1: LLM具有广泛的知识，强大的泛化能力，已被证明可以成功地应用于其他领域；**
  - **O2: LLM可以作为网络的基础模型，规划能力可以用于更好的决策；**
- **Challenges:**
  - **C1 : Large input modality gap：不同的网络任务具有不同的模态输入信息，转换为文本输入在网络领域不可行；**
  - **C2：Inefficiency of token-based answer generation：自回归方式token生成时延高，生成的答案可能无效的（幻觉）；**
  - **C3：High adaptation costs：传统强化学习方式需要LLM与环境主动交互收集经验，这种微调方式成本过高。**
- **Design:**
  - **D1: Multimodal Encoder：使用已有特征编码器提取不同模态的信息，再训练一组线性层映射到token空间，可以被LLM有效利用；**
  - **D2: Networking Head：训练一组线性层，根据LLM的输出特征预测特定任务的答案，加速生成并将答案限制在有效范围内；**
  - **D3：Data-Driven Low-Rank Networking Adaptation：使用任何现有的（非 LLM）网络算法收集经验数据集，再利用LoRA微调；**

Conclusion

We use the LLM to achieve “one model for all networking tasks” with even better performance.

LLM is powerful

Multimodal Encoder

Differ from plain text

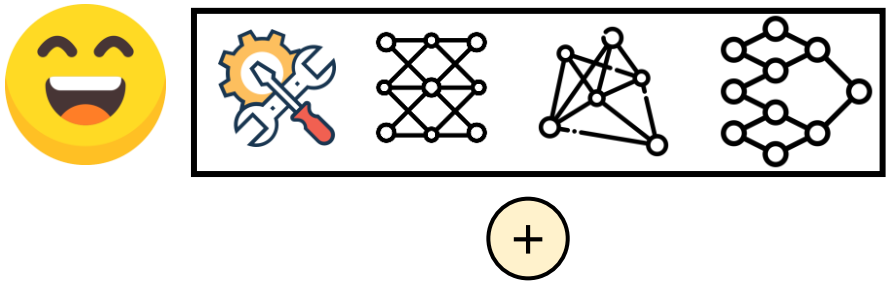
Multimodal Information

Image in viewport prediction

Time-series data in bitrate streaming

Graph in job scheduling

...



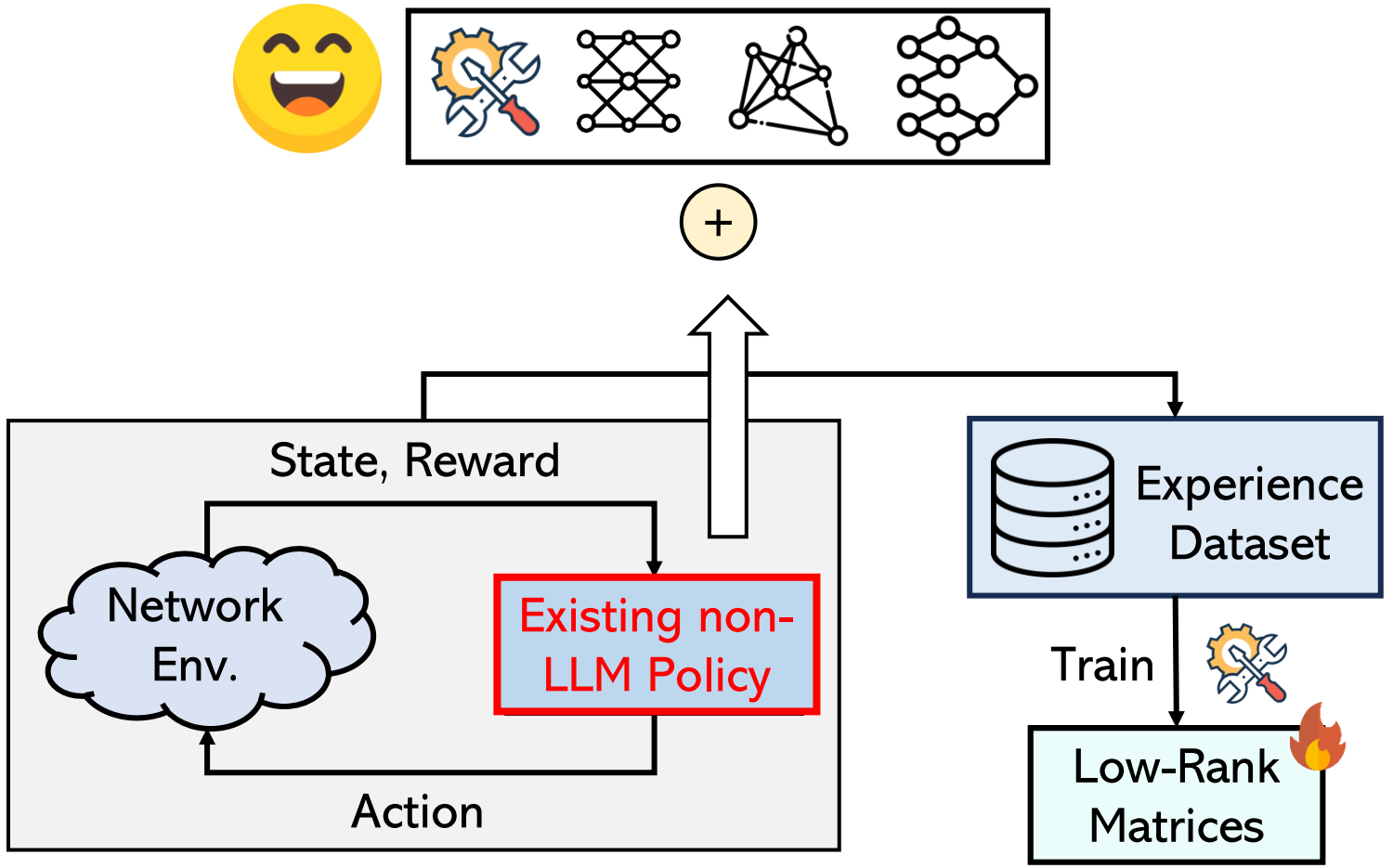
Conclusion

We use the LLM to achieve “one model for all networking tasks” with even better performance.

LLM is powerful

Multimodal Encoder

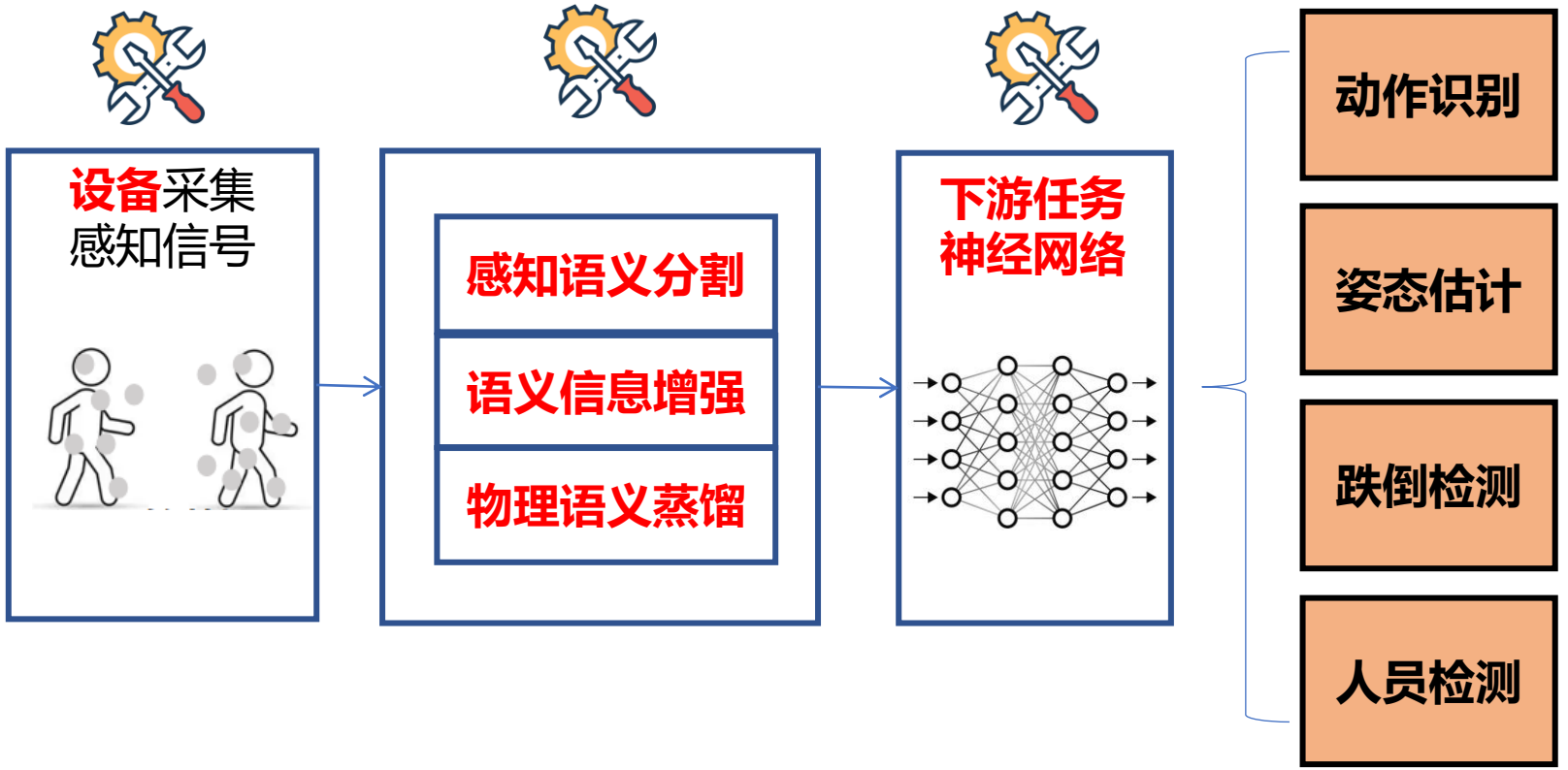
DD-LRNA



Conclusion

感知场景

LLM as Controller



東南大學
SOUTHEAST UNIVERSITY

恳请各位老师与同学批评指正！