ACM MOBIHOC 2025

OCT 27-30, 2025, Houston, USA

Local-Cloud Inference Offloading for LLMs in Multi-Modal, Multi-Task, Multi-Dialogue Settings

Liangqi Yuan[1], Dong-Jun Han[2], Shiqiang Wang[3], and Christopher G. Brinton[1]

[1] Purdue University, [2] Yonsei University, [3] IBM T. J. Watson Research Center

Oct 28, 2025

Local LLM vs. Cloud LLM

✓ **Lightweight local LLM** can process simple tasks efficiently with low latency and cost.

✗ **Lightweight local LLM** struggle to handle complex or multi-modal tasks.

✓ **Large-scale cloud LLM** can manage complex reasoning and multi-modal data.

✗ **Large-scale cloud LLM** suffer from high inference latency and usage costs.

Objective: Develop a system that **combines the strengths of both approaches** — achieving high response quality from cloud LLMs while maintaining low latency and usage cost through local inference.

Simple Tasks

Complex Tasks

Apple Intelligence

Distributed Inference

Goal: Perform inference using most suitable models, on local device or potentially offloading to the cloud

Advantages: Performance-resource trade-offs, capability to address complex tasks, etc.

Apple Intelligence

Distributed
Multi-Modal Inference

Goal: Perform inference using most suitable models and modalities, on local device or potentially offloading to the cloud

Advantages: Performance-resource trade-offs, capability to address complex tasks, etc.

Challenges: Lack of common frameworks and datasets, difficulty in decision-making, etc.

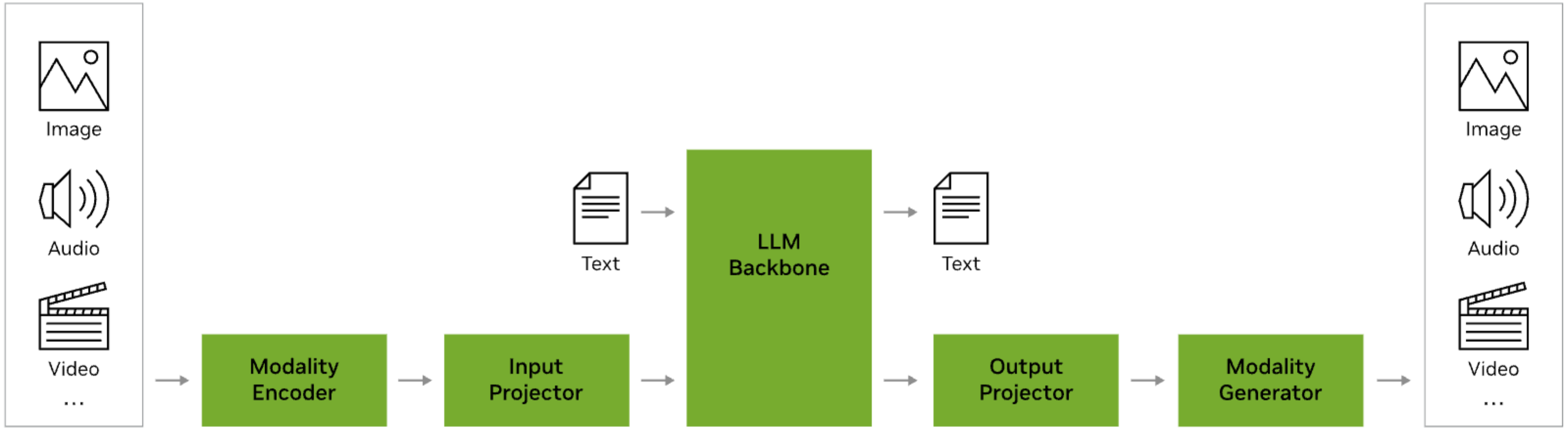
Existing works only consider single-modality and single-round scenarios

Apple iPhone 16 Pro Max

Multi-Modal LLMs over Networks

- Multimodal LLMs have been widely deployed on edge devices via network access, such as the ChatGPT application on smartphones, direct web access, or API calls.

- A typical scenario is that **users manually select the service and the relevant data**, for example, using natural language to query objects within an image.



[NVIDIA] Multimodal Large Language Models

Our Focus

Goal: To optimize trade-offs among

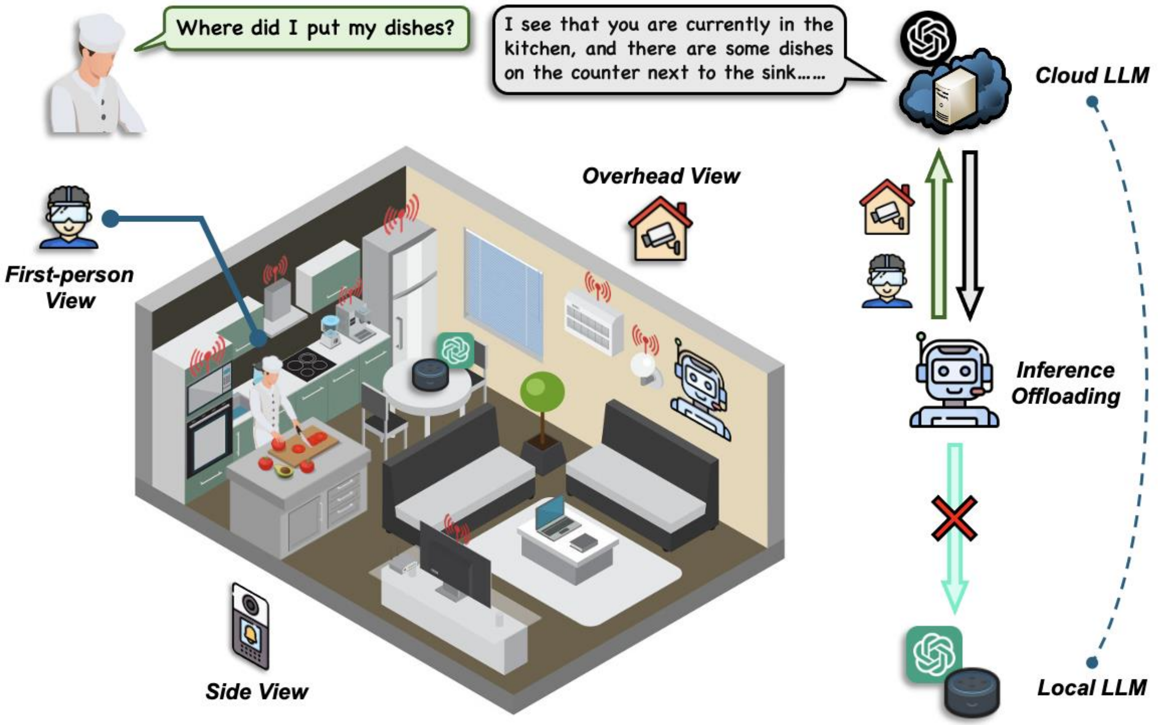
response quality, latency, usage cost

in local-cloud LLM collaborative inference across

multi-modal, multi-task, multi-dialogue scenarios

under resource constraints

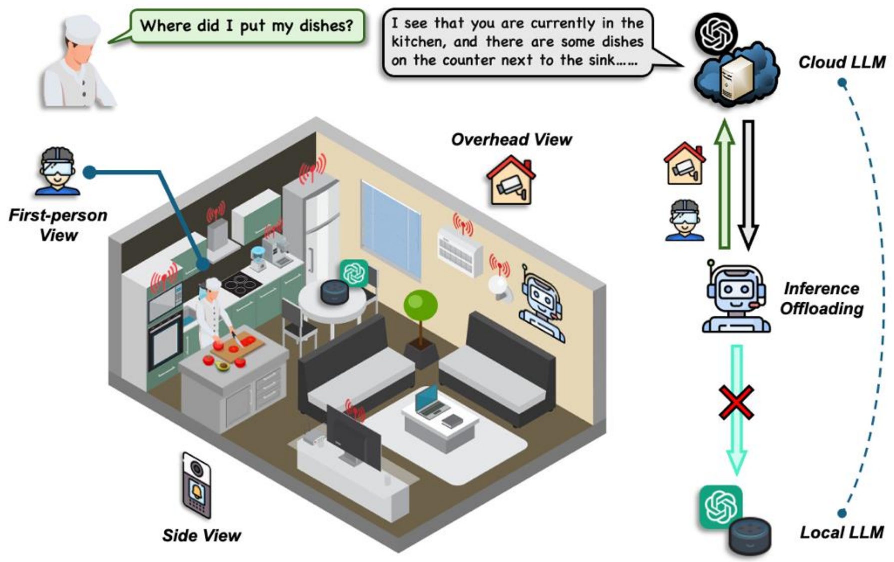
- **Lightweight local LLM** can process simple tasks at high speed.
- **Large-scale cloud LLM** can handle complex tasks and multi-modal data sources.



Challenges

New challenges arise!

- **Local LLM vs. Cloud LLM:** Difficult to determine inference locations due to trade-offs among response quality, latency, cost, and resource constraints.
- **Multi-Modal Data for Multi-Task in Multi-Dialogues:** Challenging to identify the most informative and relevant multi-modal data sources.
- **Inherent Uncertainty in LLM Inference:** LLM-generated responses are uncertain (e.g., hallucination), which makes offline training of other models difficult.



How to decide which LLMs?

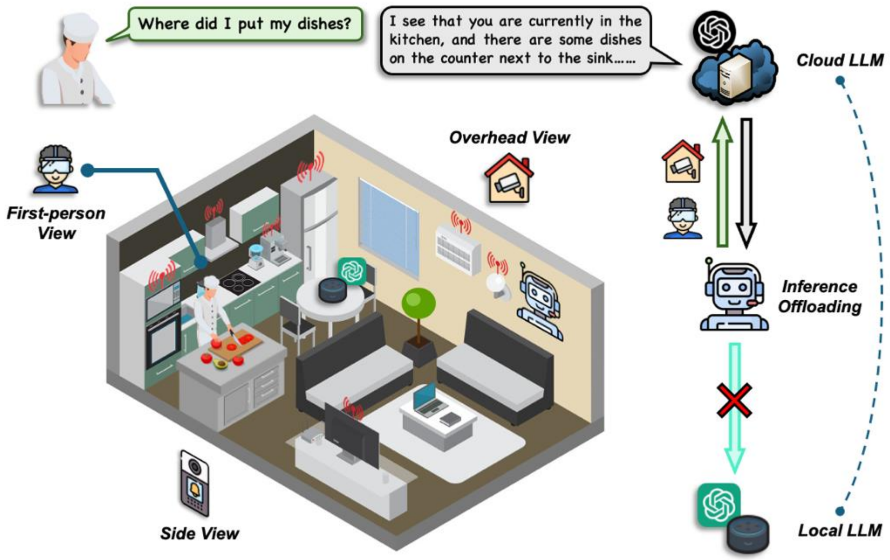
How to decide which modalities?

Uncertainty in LLM responses

Our Solution

TMO: Multi-Modal, Multi-Task, Multi-Dialogue Inference Offloading

- **Resource-Constrained RL:** Maximize cumulative reward across multi-dialogue interactions by selecting the **optimal LLMs** and **modalities** for inference.
- Handle **two types of uncertainty** in LLM responses:
  - **Non-Deterministic Evaluation (NDE):** Identical state-action pairs may yield different response quality scores.
  - **Out-Of-Distribution (OOD):** Estimating response quality is challenging for unseen state-action pairs.



How to decide which LLMs?

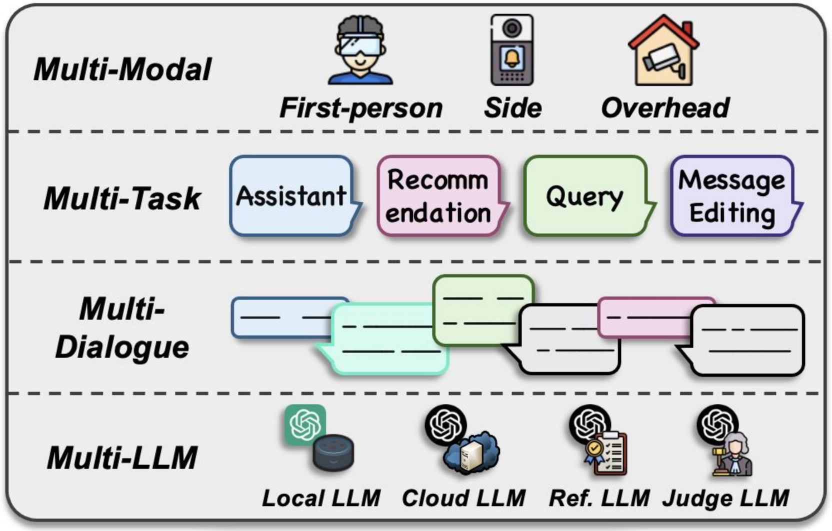
How to decide which modalities?

Uncertainty in LLM responses

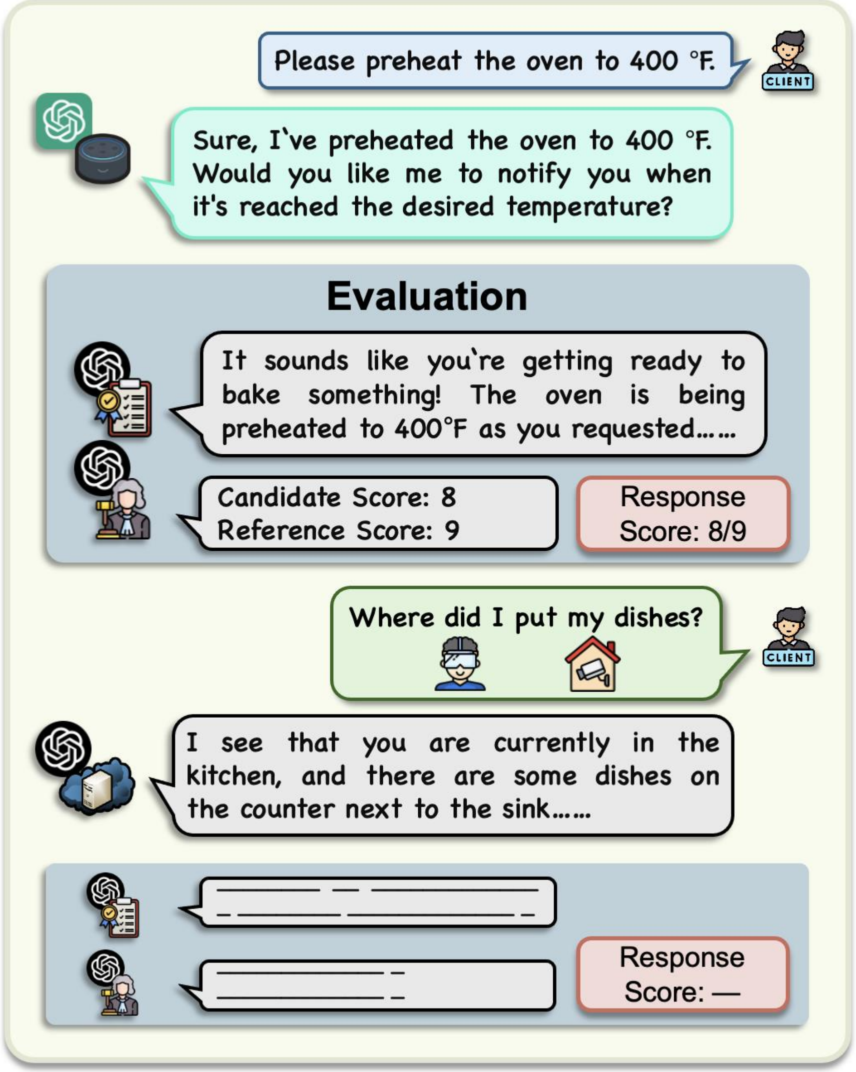
Use Case

Kitchen Assistant with LLMs: The scenario simulates 2-5 dialogues between users and LLMs across different kitchen activities.

Scenario based on ActionSense, involving 20 kitchen activities

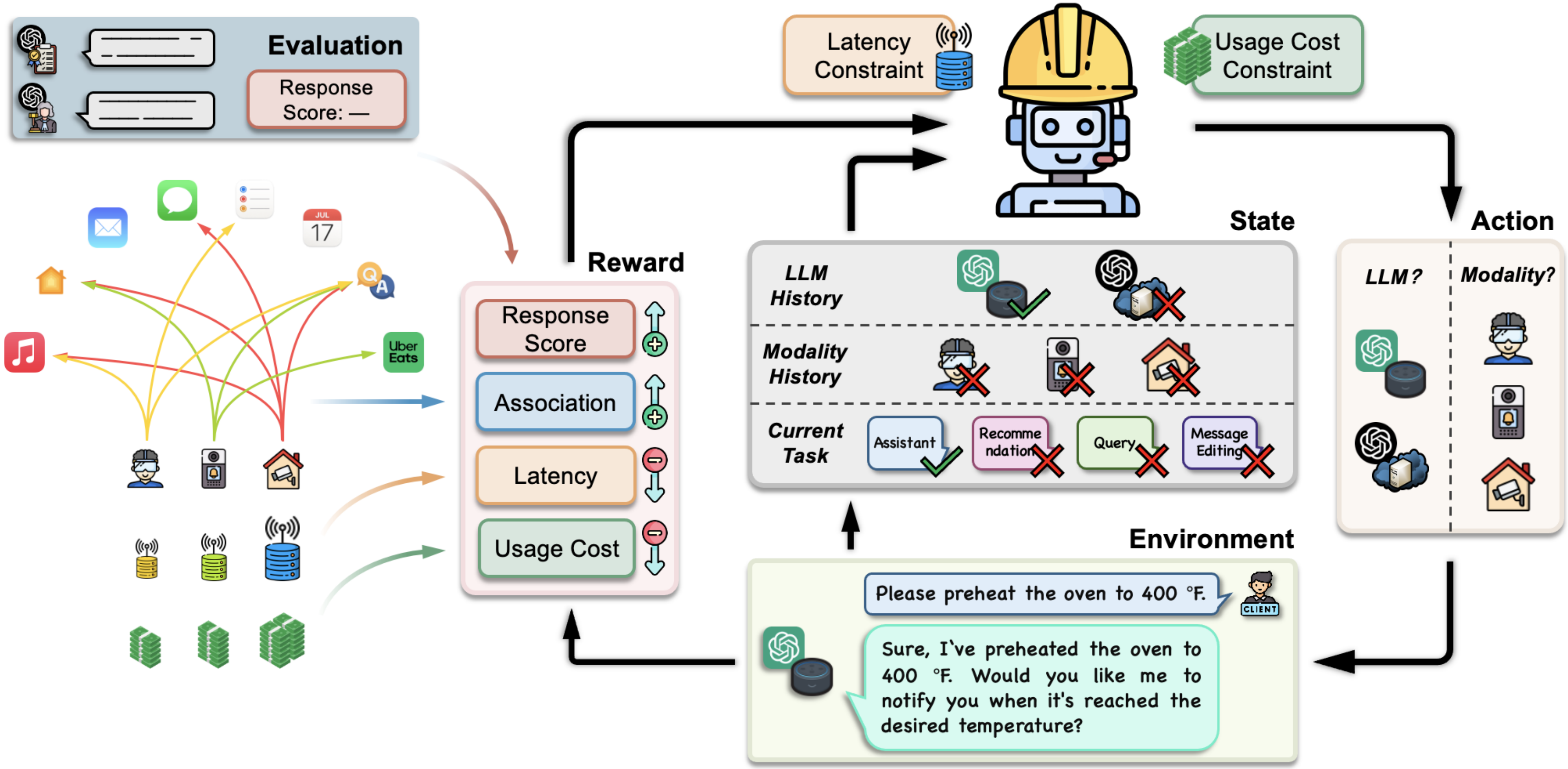


ActionSense, NeurIPS 2022



Response scores evaluated within our M4A1 dataset using LLM-as-Judge

Resource-Constrained RL



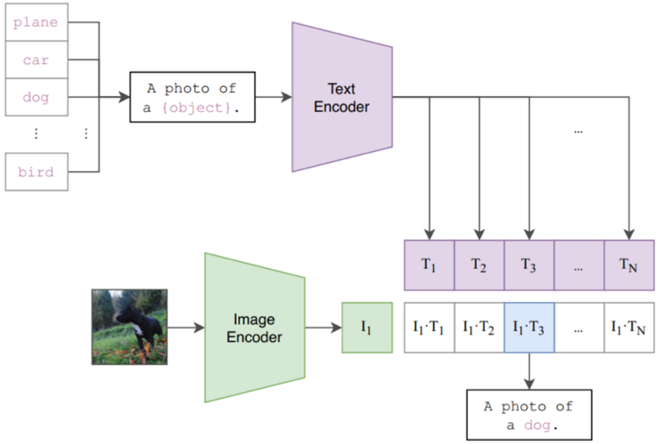
Reward Function Composition

- **Response Score:**

  Step 1. LLM-as-Judge evaluates candidate responses and assigns a relative response score.

  Step 2. Handles uncertainty through weighted averaging of similar state-action pairs.

- **Task-Modality Association:** Measures how well the text and target modality (e.g., image) align, based on cosine similarity between their representations from the CLIP model.



CLIP, OpenAI

- **Local LLM:**

  - **Latency:** Estimated using device FLOPS and model efficiency.
  - **Usage Cost:** Calculated from electricity consumption.

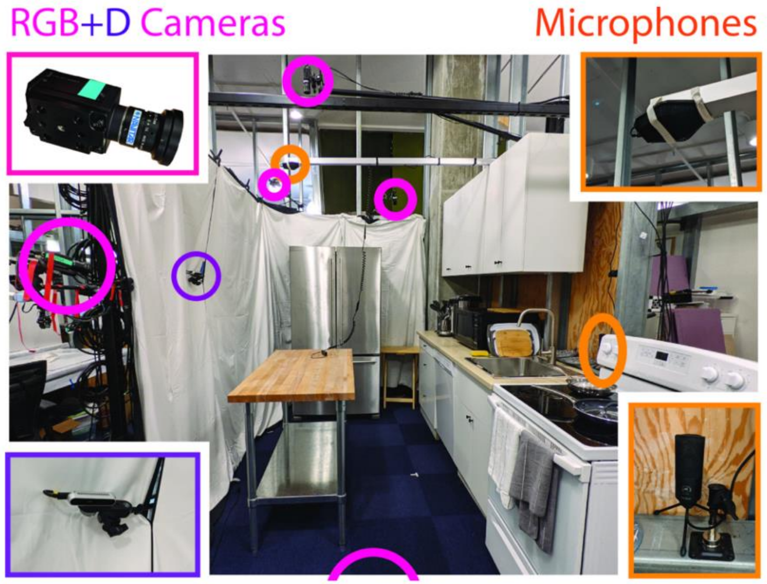
- **Cloud LLM:**

  - **Latency:** Measured real inference time (upload → inference → download).
  - **Usage Cost:** Calculated based on provider pricing (e.g., OpenAI API rates).

Dataset Curation – Image Data

20 kitchen activities, including Peel a Cucumber, Peel a Potato, Clear Cutting Board, etc.

ActionSense, NeurIPS 2022



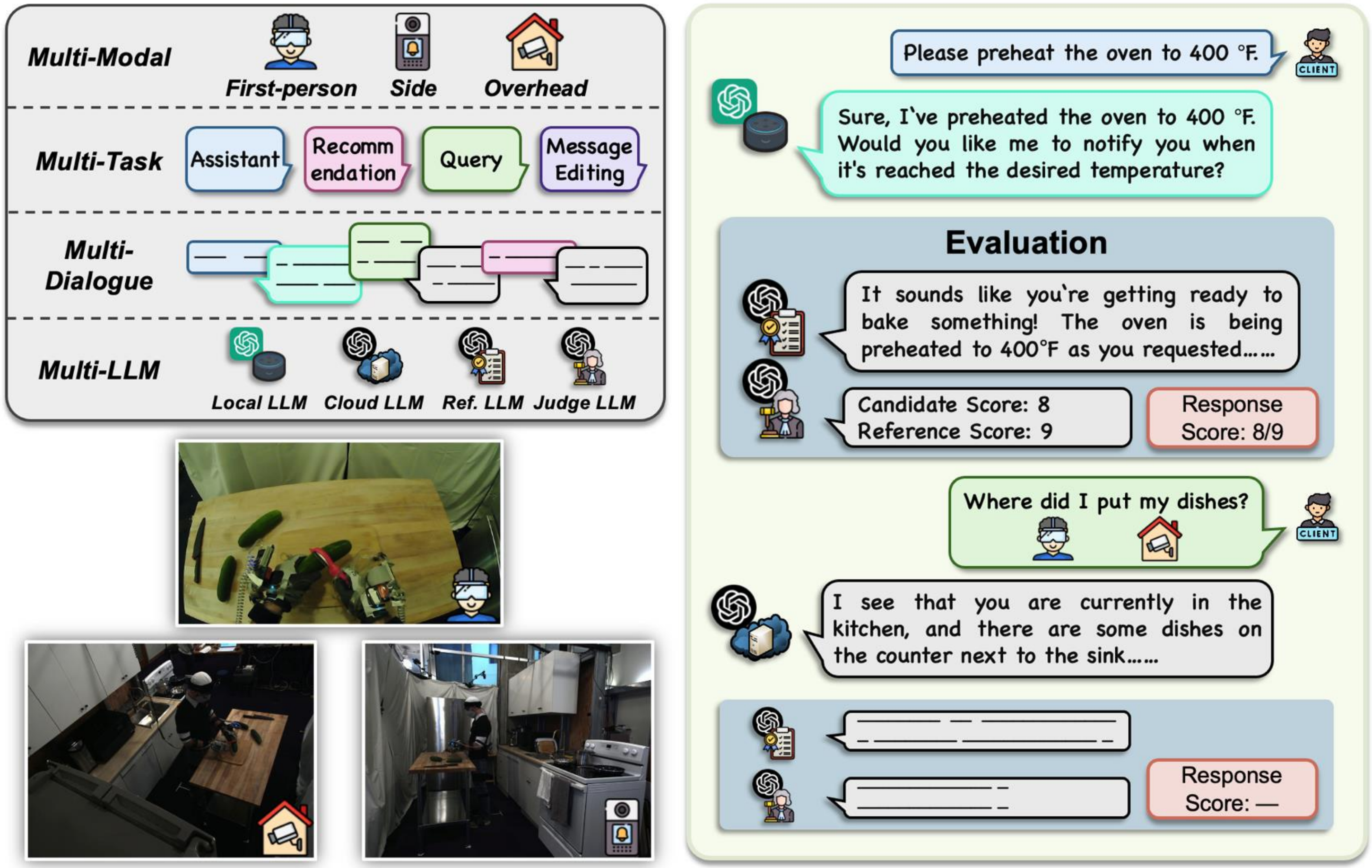
We use image data from three RGB cameras to construct our M4A1 dataset.

Dataset Curation – M4A1

Four types of task instructions, such as:

- Please preheat the oven to 400 degrees Fahrenheit.
- I'm feeling a little hot right now, please turn the temperature down a bit.
- Where did I put my dishes?
- Please draft a text message asking my friend to bring some ketchup on their way over.

ActionSense, NeurIPS 2022



Main Results

- ✓ **TMO outperforms all baselines:** naive, LLM-as-Agent, and SOTA methods.

Method	Response Score (↑)	Latency (s) (↓)	Usage Cost (1e-3 USD) (↓)	Overall Reward (↑)	Local	Cloud - Num. Selected Modalities				Constraint Violation (↓)	
						0 (text-only)	1	2	3	Latency	Usage Cost
Random	0.86 ±0.02	10.00 ±0.25	12.32 ±0.23	0.71 ±0.01	2097	261	775	771	255	0.95 ±0.13	0.89 ±0.83
Local	0.74 ±0.04	0.04 ±0.00	0.00 ±0.00	0.74 ±0.04	4203	0	0	0	0	0.00 ±0.00	0.00 ±0.00
Cloud	1.04 ±0.01	20.24 ±0.18	25.14 ±0.32	0.75 ±0.01	0	515	1583	1540	544	1.11 ±0.05	2.69 ±0.34
LLM-as-Agent [14, 25] for Inference Offloading (Ignored LLM Agent's Latency and Usage Cost)											
Phi-3-mini	0.81 ±0.04	5.25 ±0.31	7.54 ±0.44	0.74 ±0.04	3089	0	613	207	295	0.00 ±0.00	0.00 ±0.00
Phi-3.5-mini	0.83 ±0.04	5.63 ±0.38	8.39 ±0.57	0.76 ±0.04	3118	0	42	1044	0	0.00 ±0.00	0.00 ±0.00
LLaMA-3.2-3B	1.05 ±0.02	22.04 ±0.27	35.08 ±0.39	0.74 ±0.02	78	101	58	2955	1011	2.11 ±0.04	4.02 ±0.16
LLaMA-3.1-8B	0.89 ±0.02	10.84 ±0.18	12.41 ±0.31	0.74 ±0.02	1933	381	1017	545	326	1.54 ±0.22	3.79 ±1.02
Mistral-7B-v0.3	0.83 ±0.03	14.46 ±0.28	1.10 ±0.02	0.64 ±0.03	1519	2684	0	0	0	0.00 ±0.00	0.00 ±0.00
FLAN-T5-large	1.01 ±0.04	24.45 ±0.28	47.91 ±0.55	0.68 ±0.04	0	0	0	0	4203	4.90 ±0.01	11.14 ±0.36
FLAN-T5-xl	0.85 ±0.05	12.95 ±0.15	9.59 ±0.27	0.65 ±0.05	1452	1053	1408	0	289	0.00 ±0.00	0.00 ±0.00
Gemma-2-2b	0.74 ±0.04	0.04 ±0.00	0.00 ±0.00	0.74 ±0.04	4203	0	0	0	0	0.00 ±0.00	0.00 ±0.00
Gemma-2-9b	0.74 ±0.04	0.04 ±0.00	0.00 ±0.00	0.74 ±0.04	4203	0	0	0	0	0.00 ±0.00	0.00 ±0.00
GPT-3.5-turbo	0.99 ±0.02	19.96 ±0.51	33.75 ±0.66	0.73 ±0.01	527	190	504	692	2292	2.97 ±0.15	5.86 ±0.39
GPT-4o-mini	0.74 ±0.04	0.04 ±0.00	0.00 ±0.00	0.74 ±0.04	4203	0	0	0	0	0.00 ±0.00	0.00 ±0.00
GPT-4o	0.85 ±0.05	12.92 ±0.29	0.98 ±0.02	0.67 ±0.05	1807	2396	0	0	0	2.26 ±0.01	0.00 ±0.00
OpenAI o1-mini *	0.77 ±0.03	8.21 ±0.43	0.62 ±0.03	0.66 ±0.03	222	127	0	0	0	1.13 ±0.81	0.00 ±0.00
OpenAI o1 *	0.76 ±0.02	5.18 ±0.16	0.39 ±0.01	0.69 ±0.02	269	80	0	0	0	0.38 ±0.54	0.00 ±0.00
OpenAI o3-mini *	0.86 ±0.04	16.91 ±1.18	1.28 ±0.09	0.63 ±0.03	87	262	0	0	0	2.09 ±0.06	0.00 ±0.00
Comparison with SOTA Exploration-Decision Baselines											
AIwRG [8]	1.02 ±0.05	23.73 ±0.98	44.01 ±3.50	0.69 ±0.06	113	244	0	0	3852	4.68 ±0.18	9.26 ±1.83
PerLLM [26]	0.97 ±0.06	21.12 ±0.08	1.60 ±0.01	0.66 ±0.06	281	3922	0	0	0	0.00 ±0.00	0.00 ±0.00
TMO (PPO)	1.02 ±0.00	17.89 ±0.82	22.48 ±1.48	0.76 ±0.01	142	40	2499	1355	121	0.70 ±0.77	1.66 ±2.35
TMO (DQN)	1.02 ±0.08	19.57 ±2.18	26.23 ±6.36	0.75 ±0.05	0	5	1820	2367	0	0.81 ±0.57	0.00 ±0.00
TMO (A2C)	1.05 ±0.03	18.81 ±3.44	26.44 ±13.05	0.79 ±0.07	98	0	2753	210	1168	1.48 ±2.09	2.38 ±3.36
TMO (RC-PPO)	1.03 ±0.03	18.33 ±1.87	22.43 ±3.57	0.77 ±0.04	90	105	2599	1309	110	0.45 ±0.64	0.00 ±0.00
TMO (RC-DQN)	1.05 ±0.05	18.87 ±0.81	23.29 ±3.59	0.78 ±0.06	46	174	2249	1636	85	0.53 ±0.33	0.42 ±0.60
TMO (RC-A2C)	1.09 ±0.08	16.63 ±0.67	17.97 ±1.18	**0.85 ±0.07**	74	12	3871	225	6	0.00 ±0.00	0.00 ±0.00
Our TMO System - Ablation Study (Using RC-A2C as Backbone)											
w/o LLM Sel.	0.85 ±0.00	9.25 ±1.02	11.67 ±3.14	0.72 ±0.02	2128	6	1228	830	1	0.00 ±0.00	0.00 ±0.00
w/o Modality Sel.	1.04 ±0.02	20.13 ±0.07	24.82 ±0.21	0.75 ±0.02	0	528	1591	1519	528	1.17 ±0.07	2.74 ±0.87
w/o Score Est.	1.01 ±0.01	16.71 ±0.28	17.76 ±0.59	0.76 ±0.01	0	0	4095	89	0	0.00 ±0.00	0.00 ±0.00

Main Results

✓ **TMO optimally selects** LLMs and modalities, outperforming in response quality, latency, and usage cost without constraint violations.

Method	Response Score (↑)	Latency (s) (↓)	Usage Cost (1e-3 USD) (↓)	Overall Reward (↑)	Local	Cloud - Num. Selected Modalities 0 (text-only)	1	2	3	Constraint Violation (↓) Latency	Usage Cost
Comparison with SOTA Exploration-Decision Baselines											
AIwRG [8]	1.02 ±0.05	23.73 ±0.98	44.01 ±3.50	0.69 ±0.06	113	244	0	0	3852	4.68 ±0.18	9.26 ±1.83
PerLLM [26]	0.97 ±0.06	21.12 ±0.08	1.60 ±0.01	0.66 ±0.06	281	3922	0	0	0	0.00 ±0.00	0.00 ±0.00
TMO (PPO)	1.02 ±0.00	17.89 ±0.82	22.48 ±1.48	0.76 ±0.01	142	40	2499	1355	121	0.70 ±0.77	1.66 ±2.35
TMO (DQN)	1.02 ±0.08	19.57 ±2.18	26.23 ±6.36	0.75 ±0.05	0	5	1820	2367	0	0.81 ±0.57	0.00 ±0.00
TMO (A2C)	1.05 ±0.03	18.81 ±3.44	26.44 ±13.05	0.79 ±0.07	98	0	2753	210	1168	1.48 ±2.09	2.38 ±3.36
TMO (RC-PPO)	1.03 ±0.03	18.33 ±1.87	22.43 ±3.57	0.77 ±0.04	90	105	2599	1309	110	0.45 ±0.64	0.00 ±0.00
TMO (RC-DQN)	1.05 ±0.05	18.87 ±0.81	23.29 ±3.59	0.78 ±0.06	46	174	2249	1636	85	0.53 ±0.33	0.42 ±0.60
TMO (RC-A2C)	1.09 ±0.08	16.63 ±0.67	17.97 ±1.18	**0.85 ±0.07**	74	12	3871	225	6	0.00 ±0.00	0.00 ±0.00

Main Results

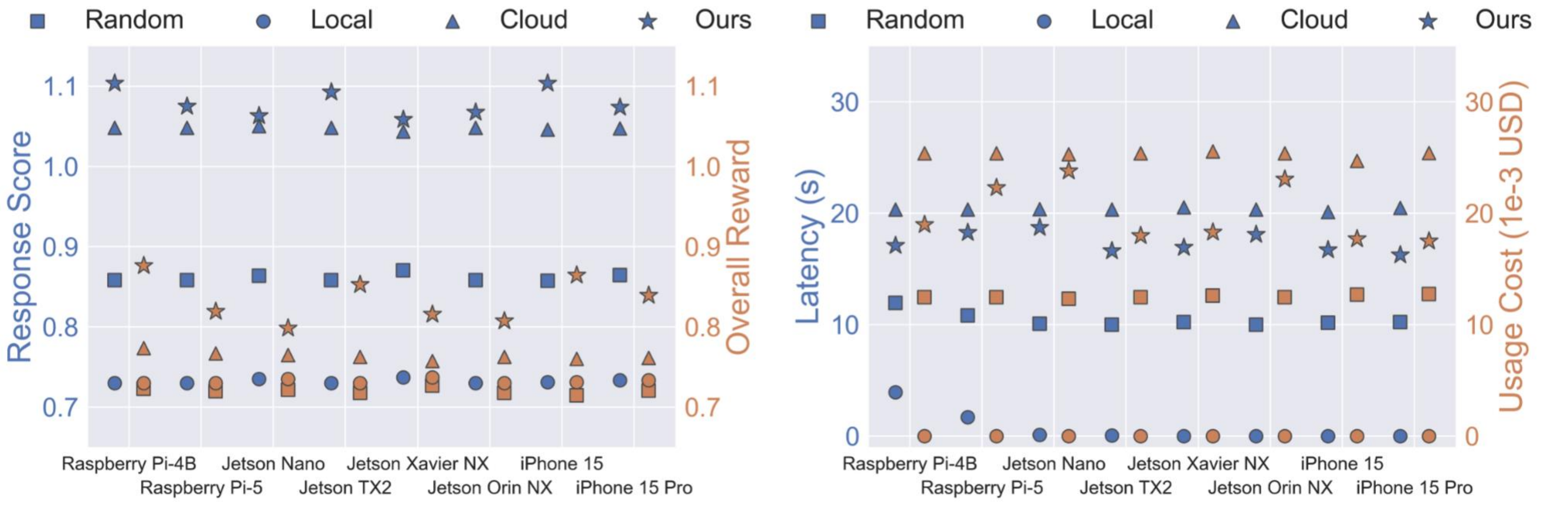
✓ **Ablation studies confirm** that LLM selection, modality selection, and score estimation are all crucial to performance.

Method	Response Score (↑)	Latency (s) (↓)	Usage Cost (1e-3 USD) (↓)	Overall Reward (↑)	Local	Cloud - Num. Selected Modalities 0 (text-only)	1	2	3	Constraint Violation (↓) Latency	Usage Cost
TMO (RC-A2C)	1.09 ±0.08	16.63 ±0.67	17.97 ±1.18	**0.85 ±0.07**	74	12	3871	225	6	0.00 ±0.00	0.00 ±0.00
Our TMO System - Ablation Study (Using RC-A2C as Backbone)											
w/o LLM Sel.	0.85 ±0.00	9.25 ±1.02	11.67 ±3.14	0.72 ±0.02	2128	6	1228	830	1	0.00 ±0.00	0.00 ±0.00
w/o Modality Sel.	1.04 ±0.02	20.13 ±0.07	24.82 ±0.21	0.75 ±0.02	0	528	1591	1519	528	1.17 ±0.07	2.74 ±0.87
w/o Score Est.	1.01 ±0.01	16.71 ±0.28	17.76 ±0.59	0.76 ±0.01	0	0	4095	89	0	0.00 ±0.00	0.00 ±0.00

Different Local Devices

Local Device	Performance (TFLOPS)	Power (Watts)	Latency (s)	Usage Cost (1e-3 USD)
Raspberry Pi-4B	0.0135	8	1.12593	4.17e-4
Raspberry Pi-5	0.0314	12	0.48408	2.69e-4
Jetson Nano	0.472	10	0.03220	1.49e-5
Jetson TX2	1.33	15	0.01143	7.94e-6
Jetson Xavier NX	21	20	0.00072	6.71e-7
Jetson Orin NX	100	25	0.00015	1.76e-7
iPhone 15 (A16)	15.8	15	0.00096	6.69e-7
iPhone 15 Pro (A17 Pro)	35	15	0.00043	3.02e-7

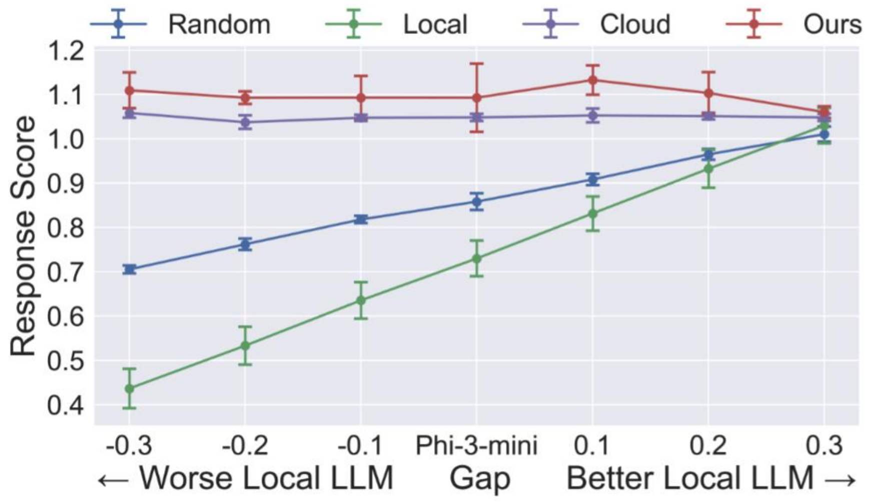
✓ **TMO operates robustly across diverse devices**, maintaining top performance while balancing latency and cost.



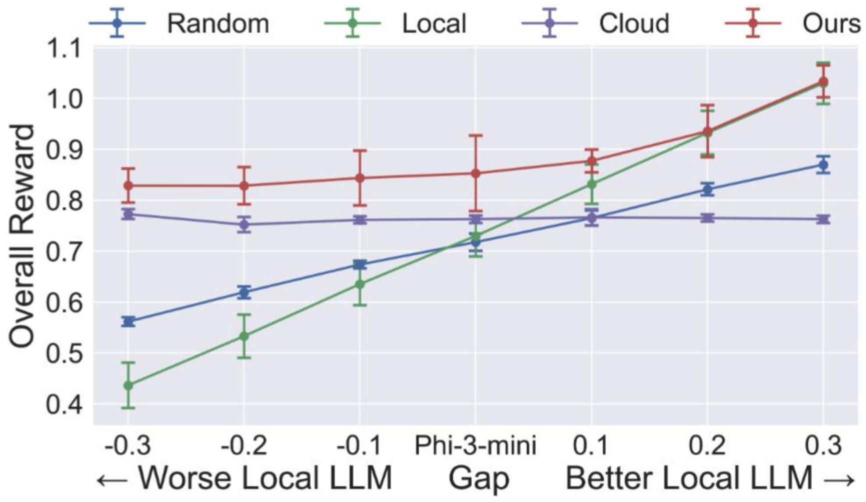
(a) Response Score & Overall Reward (b) Latency & Usage Cost

Different Local LLMs

✓ **TMO adapts to any local-cloud LLM combination**, consistently achieving strong performance.



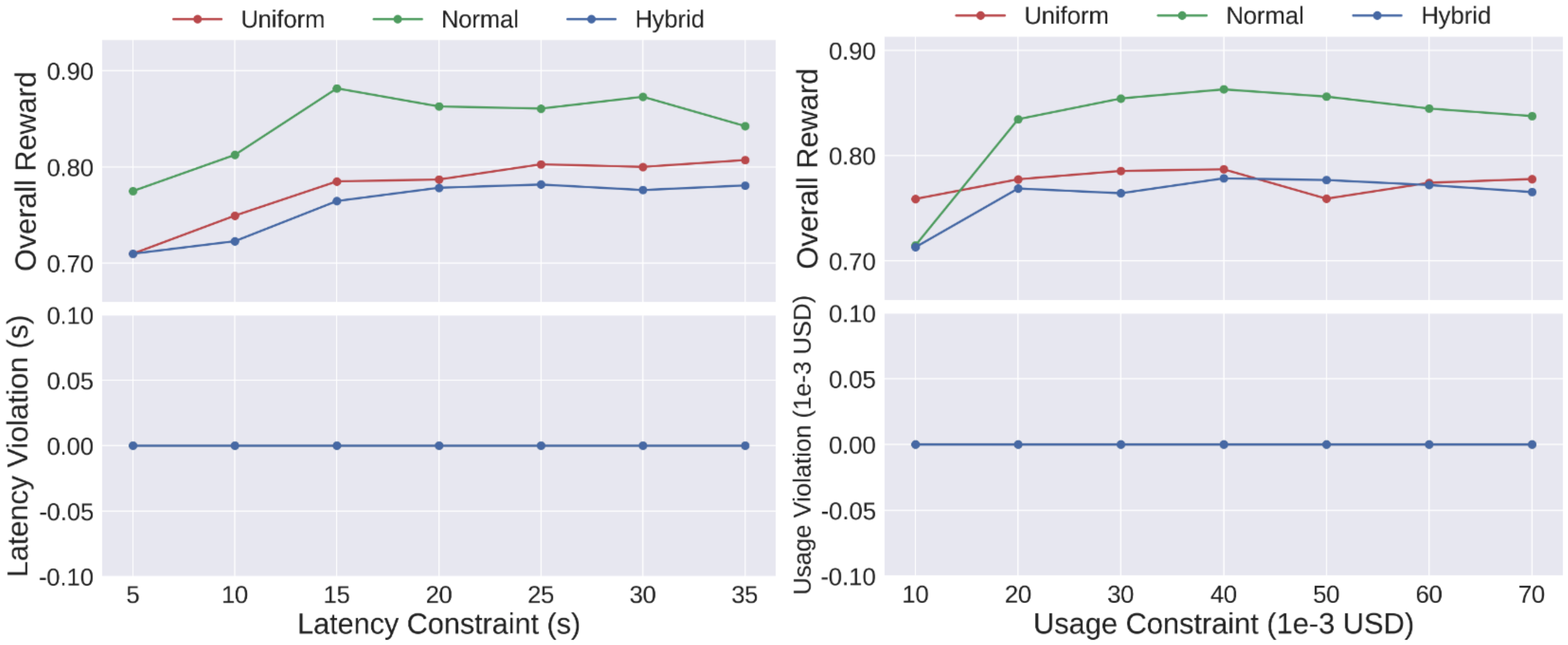
(a) Response Score



(b) Overall Reward

More Recently –
Better Resource Constraints

✓ **TMO adapts to arbitrarily user-specified resource constraints post-training,** consistently achieving strong performance without violating them.



➢ **Resource-Aware and Generalization:** Random resource budgets are assigned during training, following Uniform, Normal, or a Hybrid (Uniform + Normal) distribution.

Thank you!

Liangqi Yuan

Dong-Jun Han

Shiqiang Wang

Christopher G. Brinton

Scan for Full Paper