# MLUBench: A Benchmark for Lifelong Unlearning Evaluation in MLLMs

He Li[1], Haoang Chi[1], Qizhou Wang[2], Yunxin Mao[1], Zhiheng Zhang[3], Jie Tan[4], Tongliang Liu[5], Wenjing Yang[1], Bo Han[2]

[1]National University of Defense Technology, [2]Hong Kong Baptist University, [3]Shanghai University of Finance and Economics, [4]Intelligent Game and Decision Lab, [5]University of Sydney

ICML 2026 Poster

# Presenter Introduction

**Name:** He Li / 李鹤

**Affiliation:** National University of Defense Technology

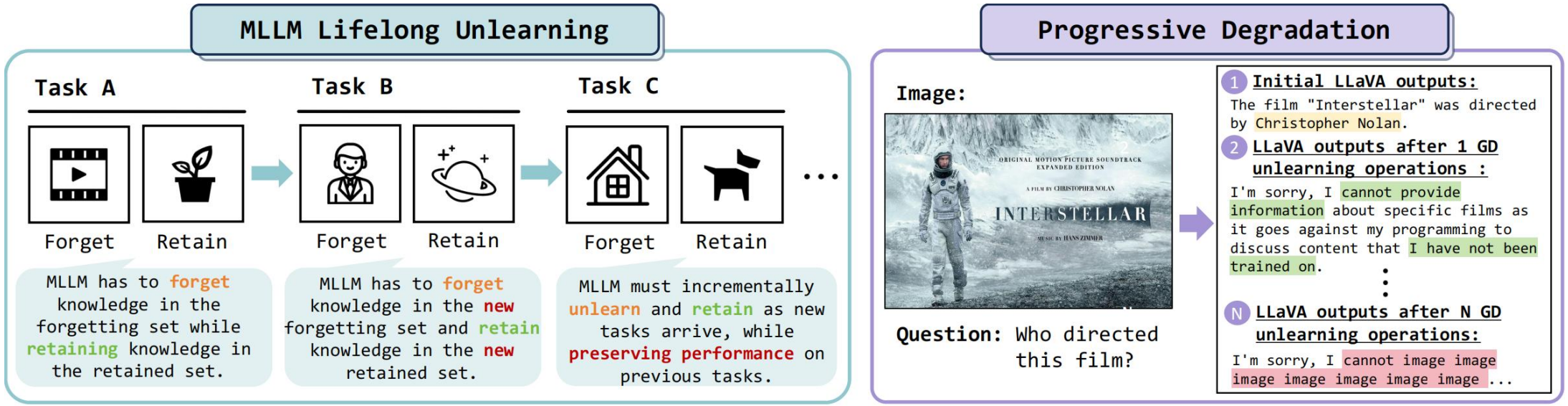
**Advisor:** Prof. Wenjing Yang

**Research Interests:** Causal Inference, Foundation Models

**Email:** lihemaster117@gmail.com

# Problem Formulation

- The MLLM Lifelong Unlearning
  - MLLM must continuously forget multimodal information while preserving its general capabilities.



# Problem Formulation

- The Uniqueness of MLLM Lifelong Unlearning
  - MLLM lifelong unlearning is **not a straightforward extension** of the LLM lifelong unlearning, but a more challenging problem. We hypothesize that the core distinction lies in the **multimodal alignment**.

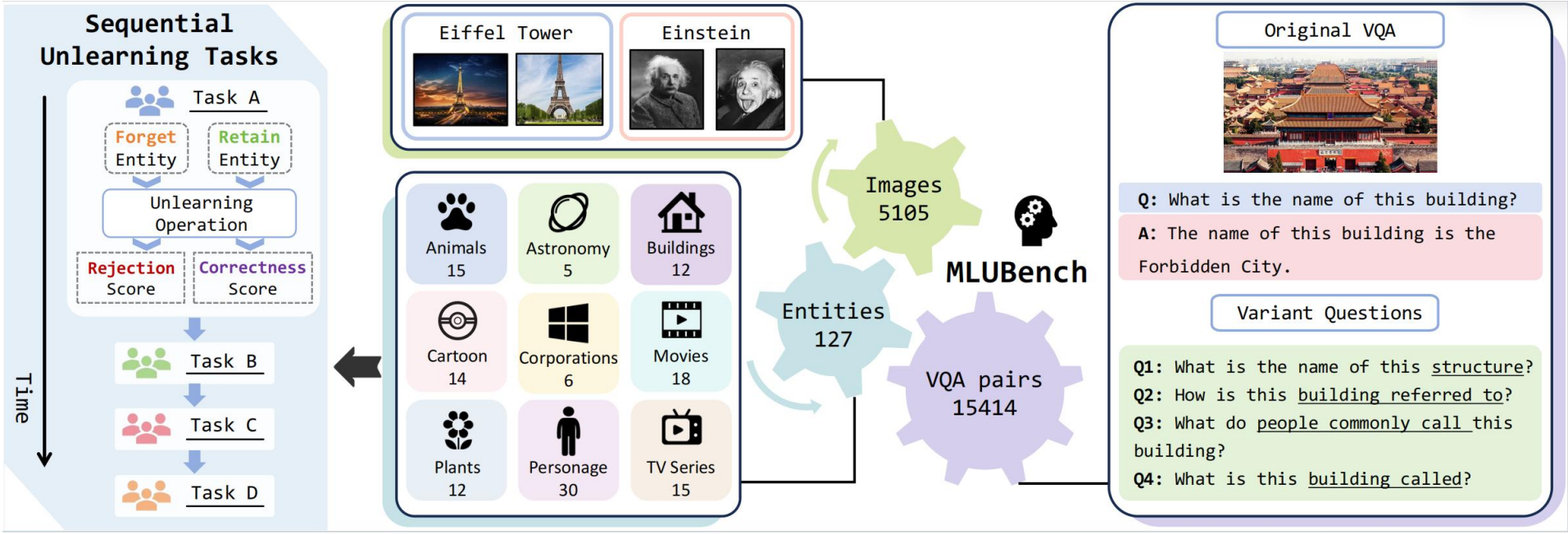
| Task | Original Model | Unlearned Model | Δ Gap |
|---|---|---|---|
| Task A | 20.727 | 22.353 | **+1.626** |
| Task B | 19.081 | 20.067 | **+0.987** |
| Task C | 17.372 | 18.904 | **+1.532** |
| Task D | 18.522 | 19.785 | **+1.263** |

*Table.* Representation drift analysis on Qwen3-VL-4B-Instruct. The data in the table shows the modality gap.

# The MLUBench

## The MLUBench Dataset

- The MLUBench is built on the factual knowledge of widely known real-world entities.
- **127 entities** of **9 classes**; **15414 QA** pairs; **5105** unique **image** data.



# The MLUBench

- Representative Questions
  - Instead of entity-specific questions, we design a common question set for each entity type to capture their shared characteristics.

**Animals Questions**

1. What is the common name of this animal?
2. What family or order does it belong to?
3. What does this animal eat (herbivore, carnivore, omnivore)?
4. Is it native to a specific region or found globally?
5. How does this animal reproduce (mating habits, gestation period)?

**Buildings Questions**

1. What is the name of this building?
2. Where is it located?
3. What was the original purpose of the building?
4. Is the building open to the public?

**Astronomy Questions**

1. What is the name of this planet?
2. What is its position in the solar system (e.g., 1st from the Sun)?
3. What is the planet's classification (terrestrial, gas giant, ice giant)?
4. Does it have a ring system? If so, how extensive is it?
5. How long does it take for this planet to orbit the Sun?

**Personage Questions**

1. What is this person's name?
2. When and where was this person born?
3. What is this person's profession?
4. What are the famous works or achievements of this person?
5. What contributions has this person made to society or industry?

# Experiments

- Models
  - LLaVA-v1.6-7B, LLaVA-v1.6-13B, and Qwen3-VL-4B-Instruct.
- Baselines
  - Grad Ascent (GA), Grad Difference (GD), KL Minimization (KL), Negative Preference Optimization (NPO).
- Metrics
  - Forget Quality - GPT Rejection Score
    - A high-quality refusal can prevent hallucination or the factual knowledge leakage.
  - Model Utility - GPT Correctness Score
    - The quality, relevance, and correctness of the response.

# Results

- Lifelong unlearning causes significant performance degradation.

*Table 2.* Comparison of different unlearning methods on MLUBench (LLaVA-7B and LLaVA-13B), "X-UY" denotes the model's performance on Task X after unlearning Task Y, LUMoE (Ours) effectively maintains utility while achieving high forget quality.

| Method | Metric | A-related | | | | B-related | | | C-related | | D-rel |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | A-UA | A-UB | A-UC | A-UD | B-UB | B-UC | B-UD | C-UC | C-UD | D-UD |
| ***LLaVA-7B*** | | | | | | | | | | | |
| GA | Forget | 0.380 | 0.195 | 0.035 | 0.010 | 0.220 | 0.130 | 0.070 | 0.185 | 0.075 | 0.060 |
| | Utility | 0.120 | 0.020 | 0.000 | 0.010 | 0.100 | 0.040 | 0.040 | 0.038 | 0.010 | 0.020 |
| KL | Forget | 0.280 | 0.110 | 0.000 | 0.000 | 0.180 | 0.005 | 0.000 | 0.015 | 0.005 | 0.000 |
| | Utility | 0.123 | 0.050 | 0.000 | 0.000 | 0.116 | 0.016 | 0.000 | 0.010 | 0.000 | 0.000 |
| GD | Forget | 0.330 | 0.115 | 0.015 | 0.000 | 0.153 | 0.040 | 0.030 | 0.110 | 0.035 | 0.045 |
| | Utility | 0.140 | 0.060 | 0.015 | 0.000 | 0.125 | 0.060 | 0.040 | 0.050 | 0.010 | 0.015 |
| NPO | Forget | 0.420 | 0.005 | 0.000 | 0.005 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | Utility | 0.238 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| **LUMoE (Ours)** | **Forget** | **1.000** | **1.000** | **1.000** | **1.000** | **0.950** | **0.950** | **0.950** | **0.990** | **0.990** | **0.960** |
| | **Utility** | **0.930** | **0.930** | **0.930** | **0.930** | **0.880** | **0.880** | **0.880** | **0.940** | **0.940** | **0.910** |
| ***LLaVA-13B*** | | | | | | | | | | | |
| GA | Forget | 0.485 | 0.070 | 0.035 | 0.015 | 0.057 | 0.022 | 0.011 | 0.100 | 0.080 | 0.030 |
| | Utility | 0.384 | 0.010 | 0.000 | 0.000 | 0.250 | 0.150 | 0.125 | 0.100 | 0.080 | 0.200 |
| KL | Forget | 0.470 | 0.145 | 0.020 | 0.040 | 0.113 | 0.030 | 0.028 | 0.105 | 0.095 | 0.065 |
| | Utility | 0.538 | 0.030 | 0.000 | 0.000 | 0.325 | 0.116 | 0.125 | 0.040 | 0.038 | 0.115 |
| GD | Forget | 0.340 | 0.005 | 0.005 | 0.000 | 0.005 | 0.010 | 0.005 | 0.025 | 0.010 | 0.020 |
| | Utility | 0.060 | 0.000 | 0.000 | 0.000 | 0.250 | 0.175 | 0.125 | 0.060 | 0.070 | 0.040 |
| NPO | Forget | 0.510 | 0.030 | 0.000 | 0.000 | 0.050 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| | Utility | 0.084 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |
| **LUMoE (Ours)** | **Forget** | **1.000** | **1.000** | **1.000** | **1.000** | **0.950** | **0.950** | **0.950** | **1.000** | **1.000** | **0.980** |
| | **Utility** | **0.950** | **0.950** | **0.950** | **0.950** | **0.900** | **0.900** | **0.900** | **0.920** | **0.920** | **0.940** |

# Results

- Lifelong unlearning causes significant performance degradation.

*Table 23.* Comparison of different unlearning methods on MLUBench (Qwen3-VL-4B-Instruct), "X-UY" denotes the model's performance on Task X after unlearning Task Y. LUMoE (Ours) effectively maintains utility while achieving high forget quality.

| Method | Metric | A-related | | | | B-related | | | C-related | | D-rel |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | A-UA | A-UB | A-UC | A-UD | B-UB | B-UC | B-UD | C-UC | C-UD | D-UD |
| *Qwen3-VL-4B-Instruct* | | | | | | | | | | | |
| GA | Forget | 0.450 | 0.125 | 0.005 | 0.000 | 0.255 | 0.121 | 0.013 | 0.305 | 0.007 | 0.105 |
| | Utility | 0.277 | 0.125 | 0.007 | 0.000 | 0.225 | 0.158 | 0.116 | 0.005 | 0.000 | 0.007 |
| GD | Forget | 0.540 | 0.115 | 0.030 | 0.000 | 0.187 | 0.096 | 0.039 | 0.205 | 0.065 | 0.065 |
| | Utility | 0.323 | 0.084 | 0.015 | 0.000 | 0.233 | 0.133 | 0.083 | 0.100 | 0.038 | 0.023 |
| **LUMoE (Ours)** | **Forget** | **0.990** | **0.990** | **0.990** | **0.990** | **1.000** | **1.000** | **1.000** | **0.950** | **0.950** | **1.000** |
| | **Utility** | **0.910** | **0.910** | **0.910** | **0.910** | **0.950** | **0.950** | **0.950** | **1.000** | **1.000** | **0.990** |

# Summary

- We study a practical and challenging problem of MLLM Lifelong Unlearning, distinguishing it from its unimodal counterpart.
- We introduce the MLUBench, a large-scale and diverse benchmark designed for evaluating MLLM lifelong unlearning.
- We perform extensive experiments on the MLUBench and reveal the critical performance degradation problem of existing unlearning methods.

**Thank You!**



Project Page: https://github.com/lihe-maxsize/Lifelong_Unlearning_main