CVPR JUNE 3-7, 2026 DENVER COLORADO

ϕ-DPO: Fairness Direct Preference Optimization Approach to Continual Learning in Large Multimodal Models

Thanh-Dat Truong[1,2], Huu Thien Tran[1,2], Jackson Cothren[3], Bhiksha Raj[3], Khoa Luu[1,2]

[1]CVIU Lab, University of Arkansas [2]Dep. of EECS, University of Arkansas
[3]Dep. of GEOS, University of Arkansas [4]Carnegie Mellon University
https://uark-cviu.github.io/projects/Fai-DPO

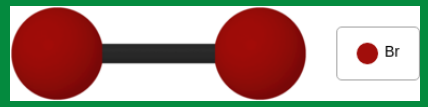
Arkansas High Performance Computing Center

Project, Code, and Data

Motivation

Time

ScienceQA



Question: The model below represents a molecule of bromine. Liquid bromine is made in chemical factories. It can be used to make couches and mattresses that are fire-resistant. Bromine is [..]
A. an elementary substance B. a compound
Answer: A

ImageNet

Question: What is the object in the image?
Answer: African crocodile

Grounding

Question : Please provide the bounding box coordinate of the region this sentence describes: lady in the middle
Answer: [0.3,0.19,0.62,0.64]

Learning Step t

Instruction Data

- Catastrophic Forgetting
- Imbalanced Data

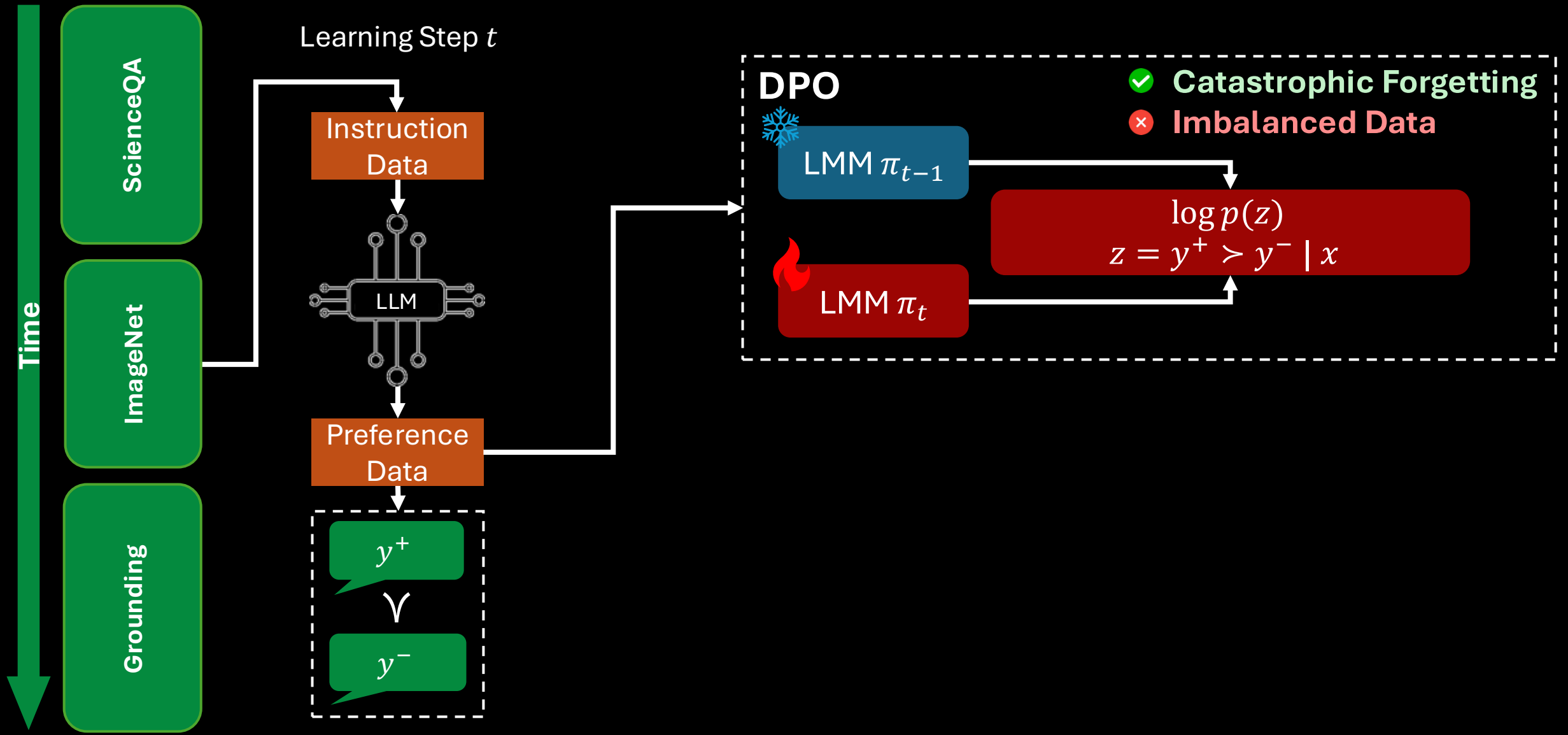
LoRA

LMM π_{t-1}

LoRA

LMM π_t

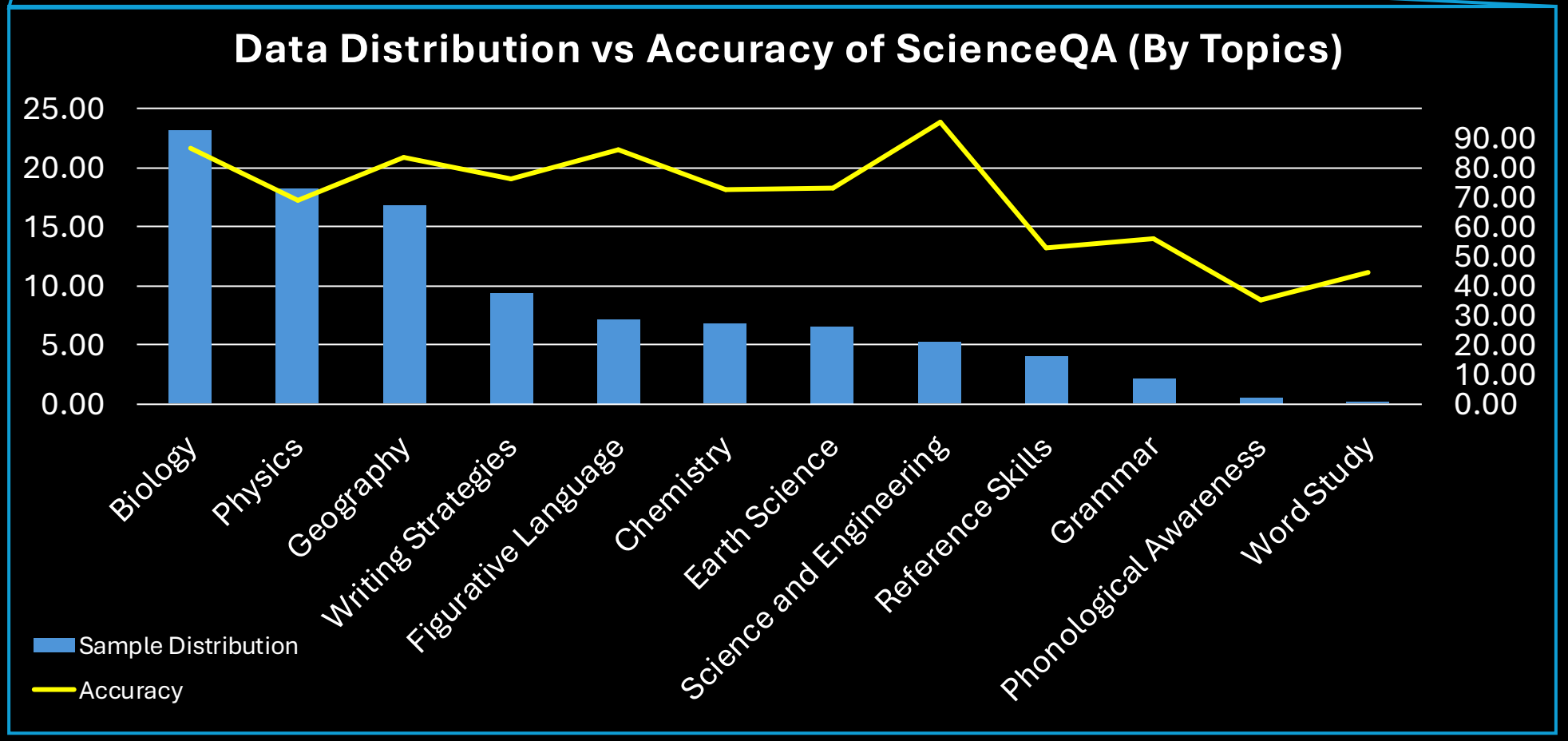
Motivation


Time
ScienceQA
ImageNet
Grounding
Learning Step t
Instruction Data
LLM
Preference Data
y^+
y^-
DPO
Catastrophic Forgetting
Imbalanced Data
LMM π_{t-1}
LMM π_t
$\log p(z)$
$z = y^+ \succ y^- \mid x$

Motivation - Fairness in Continual Learning

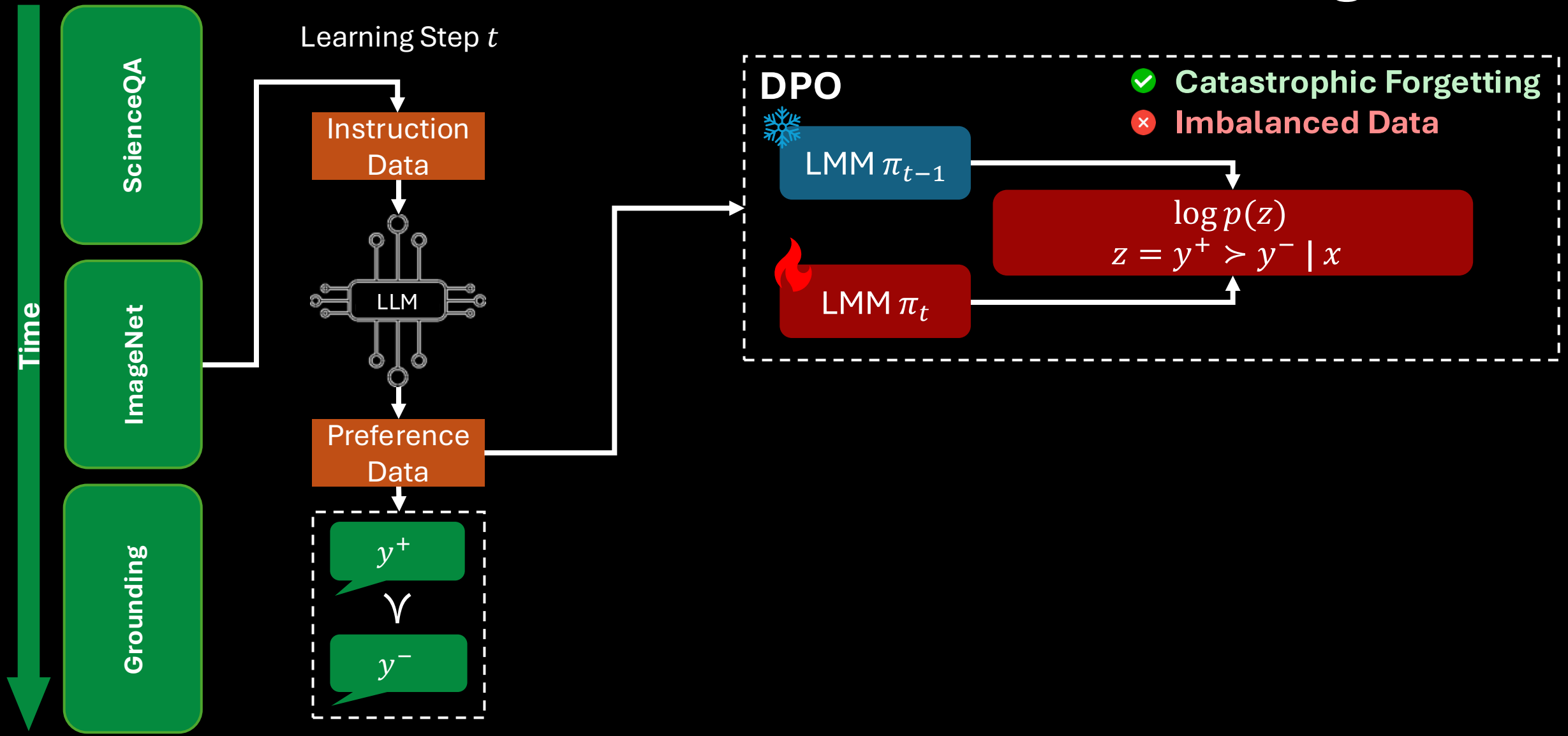
Learning Task Over Time

ScienceQA | ImageNet | VisWiz | Grounding | TextVQA | GQA | VQAv2 | OCR

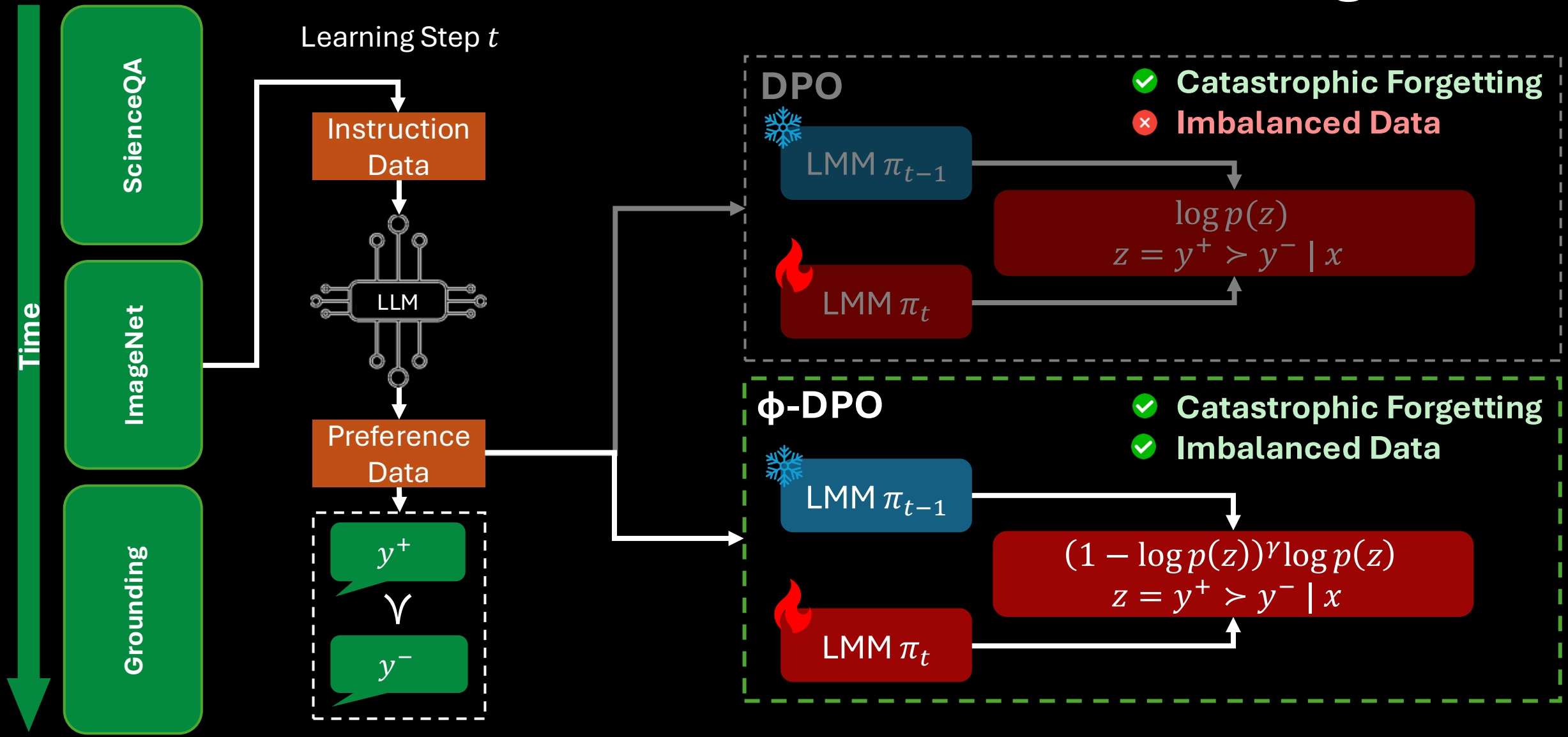


Significant Topic Imbalance

Motivation–Fairness in Continual Learning



Motivation–Fairness in Continual Learning



Contributions

- We propose a novel Fairness Direct Preference Optimization (**FaiDPO** or **ϕ-DPO**) approach to Continual Learning in LMMs.
- Our contributions can be summarized as follows.
  - ✓Introduce a new continual learning paradigm via DPO, which addresses the ***catastrophic forgetting*** problem.
  - ✓Present a novel ***Fairness DPO loss*** to address the fairness problem caused by ***imbalanced data***, by analyzing the limitations of traditional DPO.
  - ✓Provide a comprehensive theoretical analysis to show that our **ϕ-DPO** can address both ***catastrophic forgetting*** and ***imbalanced data***.
  - ✓Construct preference data labels for current continual learning benchmarks, to support the DPO learning in our framework.
  - ✓Achieve State-of-the-Art performance via intensive experiments and ablation studies.

Continual Learning via DPO

Traditional Continual Learning Approach

$$\pi_t^* = \underset{\pi_t}{\operatorname{argmax}}\, \mathbb{E}_{x,y\in\mathcal{D}_t}[\log p(y|x)] + D_{\text{forget}}(\pi_t \parallel \pi_{t-1})$$

Supervised Fine-tuning

Forgetting Mitigation

Continual Learning via DPO

$$\pi_t^* = \underset{\pi_t}{\operatorname{argmax}}\, \mathbb{E}_{x,y\sim\pi_t}[r(x,y)] - \beta D_{\text{KL}}(\pi_t(\cdot|x) \parallel \pi_{t-1}(\cdot|x))$$

Reward Model

KL Divergence

Lagrangian Multiplier

Continual Learning via DPO – Objective

- $\mathcal{L}_{\mathrm{DPO}}(\pi_t, \pi_{t-1}) = -\mathbb{E}_{x,y^+,y^-}\left[\log\sigma\left(\beta\log\frac{\pi_t(y^+|x)}{\pi_{t-1}(y^+|x)} - \beta\log\frac{\pi_t(y^-|x)}{\pi_{t-1}(y^-|x)}\right)\right]$

- $\mathcal{L}_{\mathrm{DPO}}(\pi_t, \pi_{t-1}) = -\mathbb{E}_{x,y^+,y^-}[\log\sigma(\beta(\log\pi_t(y^+|x) - \log\pi_t(y^-|x)) - \beta(\log\pi_{t-1}(y^+|x) - \log\pi_{t-1}(y^-|x)))]$

Trainable Policy

Fixed Reference

DPO Objective prevents the policy from drifting away from the previous learning task

Mitigates Catastrophic Forgetting ✅

Continual Learning via DPO–Theoretical Analysis

(Lemma 1–The Lower Bound)

$$\frac{1}{C_{\text{lower}}}\left(\log 2 - \mathcal{L}_{\text{DPO}}(\pi_t, \pi_{t-1})\right)^2$$

(Lemma 2–The Upper Bound)

$$C_{\text{upper}}\mathcal{L}_{\text{DPO}}(\pi_t, \pi_{t-1})$$

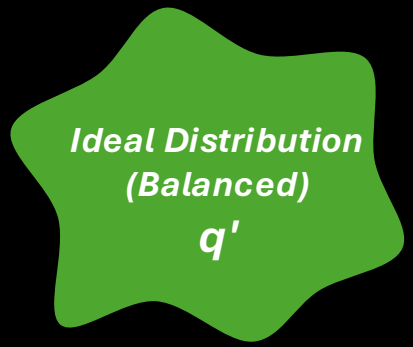
$$D_{\text{KL}}(\pi_t \parallel \pi_{t-1})$$

- ***The lower bound ensures π_t stays close to π_{t-1}, thereby mitigating catastrophic forgetting.***
- ***The upper bound regularizes updates, ensuring adaptability to a new learning task.***

Fairness DPO in Continual Learning

Gradient Difference

$$B(\theta) = \nabla_\theta \mathcal{L}_{\text{DPO}}(\theta; q) - \nabla_\theta \mathcal{L}_{\text{DPO}}(\theta; q')$$



$$= \sum_{k=1}^{K} (q_k - q'_k) m_k(\theta)$$

Group Mean Gradient

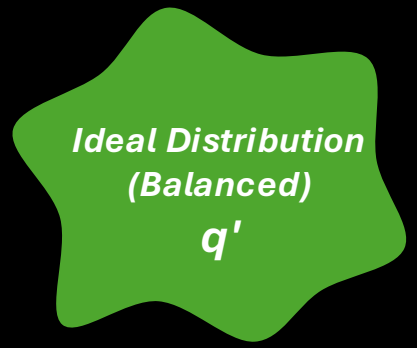
Fairness DPO in Continual Learning

$$B(\theta) = \nabla_\theta \mathcal{L}_{\text{DPO}}(\theta; q) - \nabla_\theta \mathcal{L}_{\text{DPO}}(\theta; q')$$

$$= \underbrace{(q_j - q'_j) m_j(\theta)}_{\neq 0} - \sum_{\substack{k=1 \\ k \neq j}}^{K} (q_k - q'_k) m_k(\theta)$$

$$\Rightarrow \|B(\theta)\| \neq 0$$

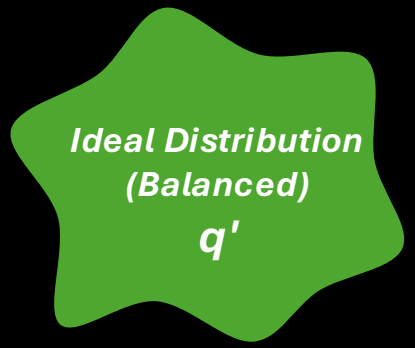
Fairness DPO in Continual Learning



$$B(\theta) = \nabla_\theta \mathcal{L}_{\mathrm{DPO}}(\theta; q) - \nabla_\theta \mathcal{L}_{\mathrm{DPO}}(\theta; q')$$

$$\mathcal{L}_{\mathrm{DPO}}(\theta; \mu) = -\mathbb{E}_\mu[\log p(y^+ \succ y^- \mid x)]$$

Fairness DPO in Continual Learning



$$B(\theta) = \nabla_\theta \mathcal{L}_{\mathrm{DPO}}(\theta; q) - \nabla_\theta \mathcal{L}_{\mathrm{DPO}}(\theta; q')$$

$$\mathcal{L}_{\mathrm{DPO}}(\theta; \mu) = -\mathbb{E}_\mu[\log p(z)]$$

Introducing new **Fair DPO Loss:**

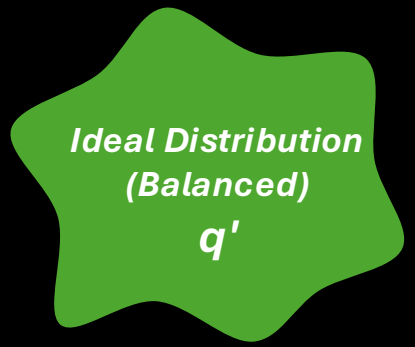
$$\mathcal{L}^{\gamma}_{\mathrm{DPO}}(\theta; \mu) = -\mathbb{E}_\mu\left[(1 - p(z))^{\gamma} \log p(z)\right]$$

γ : Focusing Parameter, balancing the gradients of each group.

Fairness DPO in Continual Learning

$$B(\theta) = \nabla_\theta \mathcal{L}_{\mathrm{DPO}}(\theta; q) - \nabla_\theta \mathcal{L}_{\mathrm{DPO}}(\theta; q')$$

$$\mathcal{L}_{\mathrm{DPO}}(\theta; \mu) = -\mathbb{E}_\mu[\log p(z)]$$



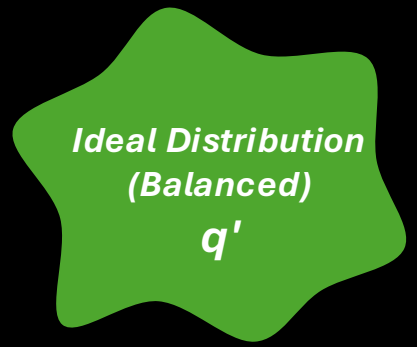
$$B^\gamma(\theta) = \nabla_\theta \mathcal{L}^\gamma_{\mathrm{DPO}}(\theta; q) - \nabla_\theta \mathcal{L}^\gamma_{\mathrm{DPO}}(\theta; q')$$

$$\mathcal{L}^\gamma_{\mathrm{DPO}}(\theta; \mu) = -\mathbb{E}_\mu\left[\left(1 - p(z)\right)^\gamma \log p(z)\right]$$

Fairness DPO in Continual Learning

$$B(\theta) = \nabla_\theta \mathcal{L}_{\text{DPO}}(\theta; q) - \nabla_\theta \mathcal{L}_{\text{DPO}}(\theta; q'), \|B(\theta)\| \neq 0$$

$$B^\gamma(\theta) = \nabla_\theta \mathcal{L}^\gamma_{\text{DPO}}(\theta; q) - \nabla_\theta \mathcal{L}^\gamma_{\text{DPO}}(\theta; q')$$



(Lemma 3)

$$\lim_{\gamma \to \infty} \|B^\gamma(\theta)\| = 0$$

Fair DPO Loss can yield fairer gradient updates as γ increases.

The Framework of Φ-DPO

Time

ScienceQA

Question: A bulldozer clears a path for a new road. A force from the bulldozer pushes loose dirt out of the way. What is the direction of this push?
A. away from the bulldozer
B. toward the bulldozer
Answer: A

...

Grounding

Question: Please provide the bounding box coordinate of the region this sentence describes: the top right slice of bread on a sandwich

Answer: [0.46,0.18,0.9,0.45]

...

GQA

Question: What is on the street?
Answer: Car

Instruction Data

x_{img} x_{ins} y

Learning Step t

LLM

Preference Data

x_{img} x_{txt} $y^+ \succ y^-$

Continual Learning via RLHF

Reward Model $r(x, y)$

LMM at Learning Step t: π_t

Evaluate Forgetting and Adaptability Level

From RLHF to DPO via Pairwise Preference

Continual Learning via Fairness Φ-DPO

LMM at Learning Step $t-1$: π_{t-1}

Fairness DPO Loss

LMM at Learning Step t: π_t

Preference Data Curation

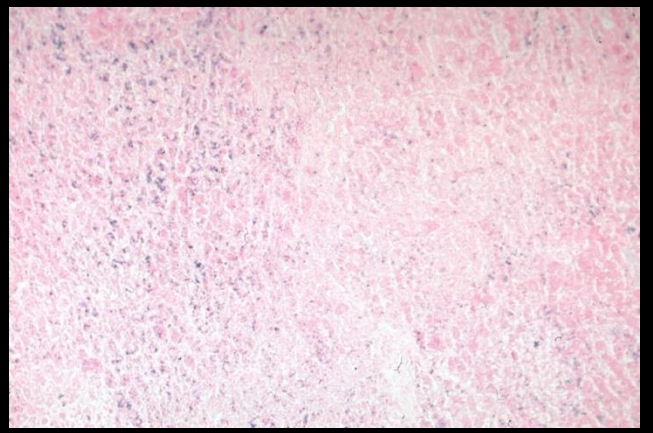
Prior Benchmarks:

Instruction-following data suitable for LLMs,

but ***lacked preference annotations required for DPO!***

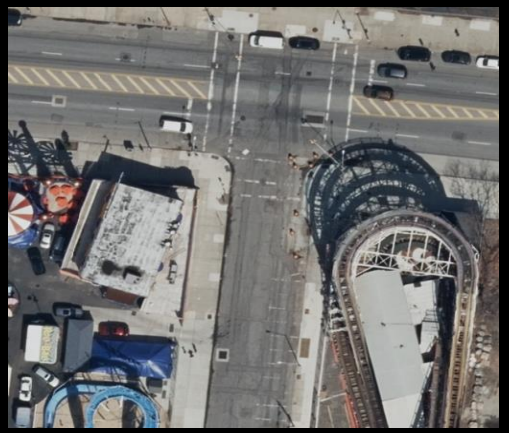
Query: What **actions** could the ego vehicle take based on **<c1,CAM_BACK,[507, 804]>**? **Why** take this action and what's the **probability**? Objects are encoded using <c,CAM,[cx,cy] [...]

Answer: The action is to **keep going at the same speed**, the reason is that the object has **little reference value** for the ego vehicle, with a **high probability**.



Query: **where** is this part in the figure?

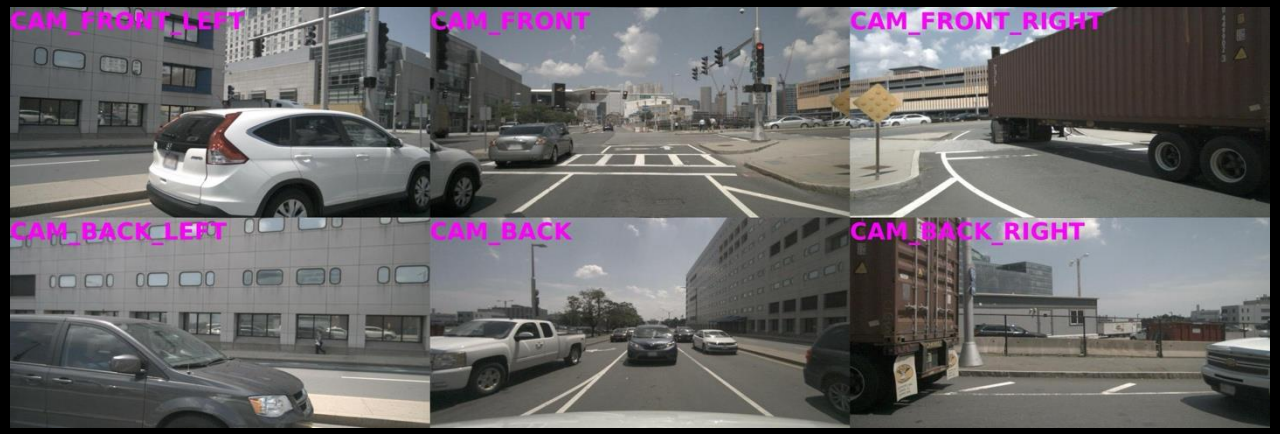
Answer: **endocrine system**



Query: What is **the area** covered by parkings? [...] single word or phrase.

Answer: **451m2**

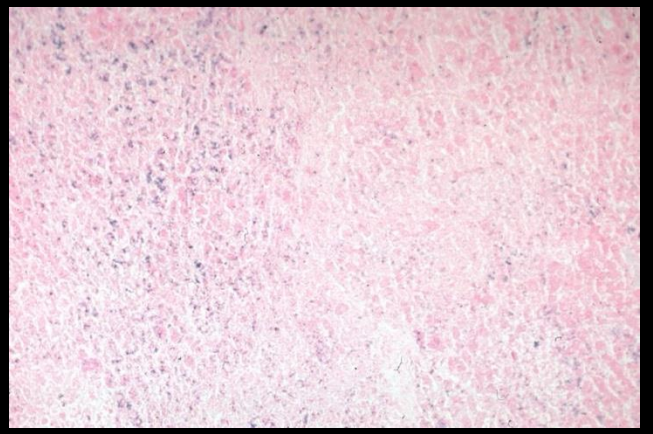
Preference Data Curation



Query: What **actions** could the ego vehicle take based on **<c1,CAM_BACK,[507, 804]>**? **Why** take this action and what's the **probability**? Objects are encoded using <c,CAM,[cx,cy] [...]

Preferred answer: The action is to **keep going at the same speed** ✓, the reason is that the object has **little reference value** ✓ for the ego vehicle, with a **high probability** ✓.

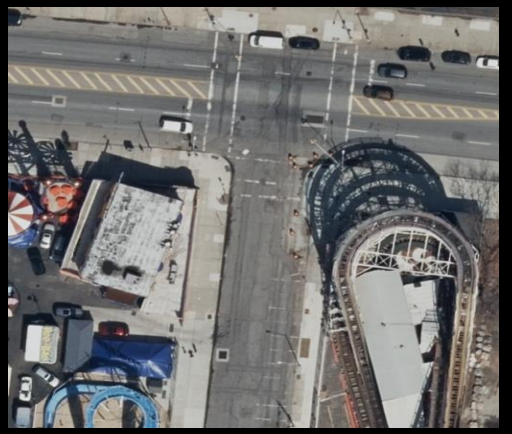
Dispreferred answer: The action is to **slow down** ×, the reason is that the object has **significant reference value** × for the ego vehicle, with a **low probability** ×.



Query: **where** is this part in the figure?

Preferred answer: **endocrine system** ✓

Dispreferred answer: **digestive system** ×



Query: What is **the area** covered by parkings? [...] single word or phrase.

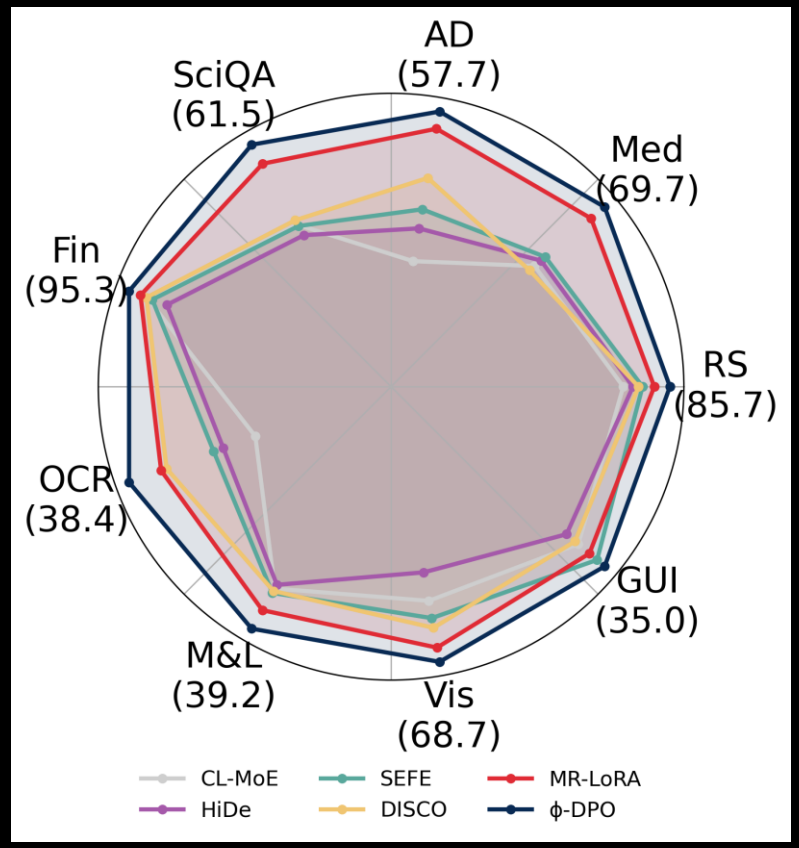
Preferred answer: **451m2** ✓

Dispreferred answer: **450cm2** ×

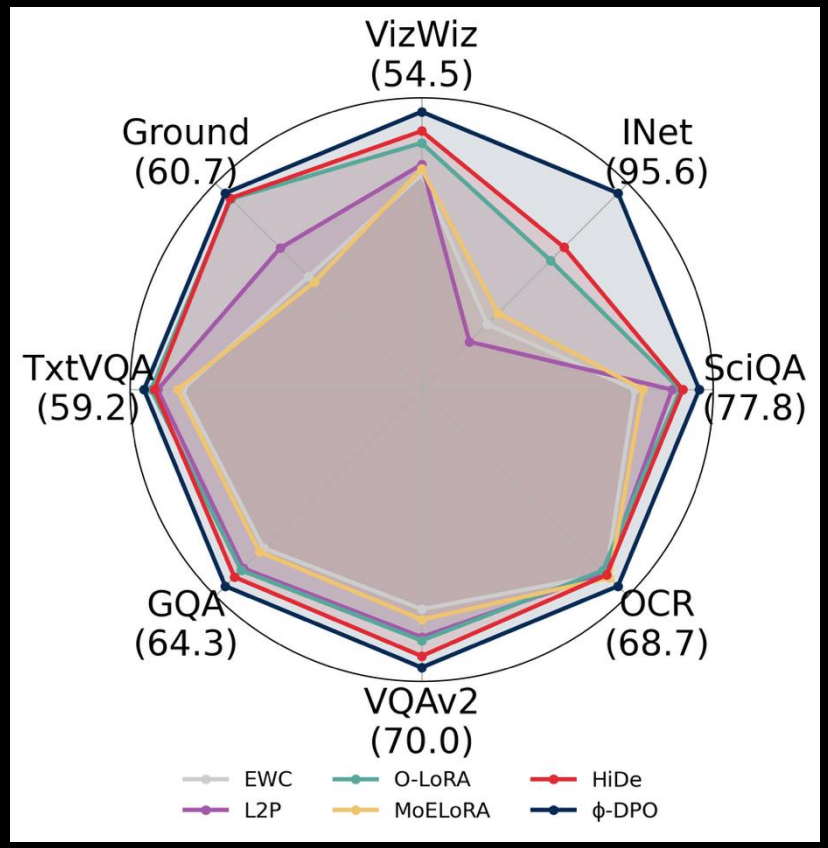
We construct pairwise preference data for each benchmark ➡ ***Enable continual learning in DPO*** ✅

Experimental Results

Results on MLLM-CL Benchmark

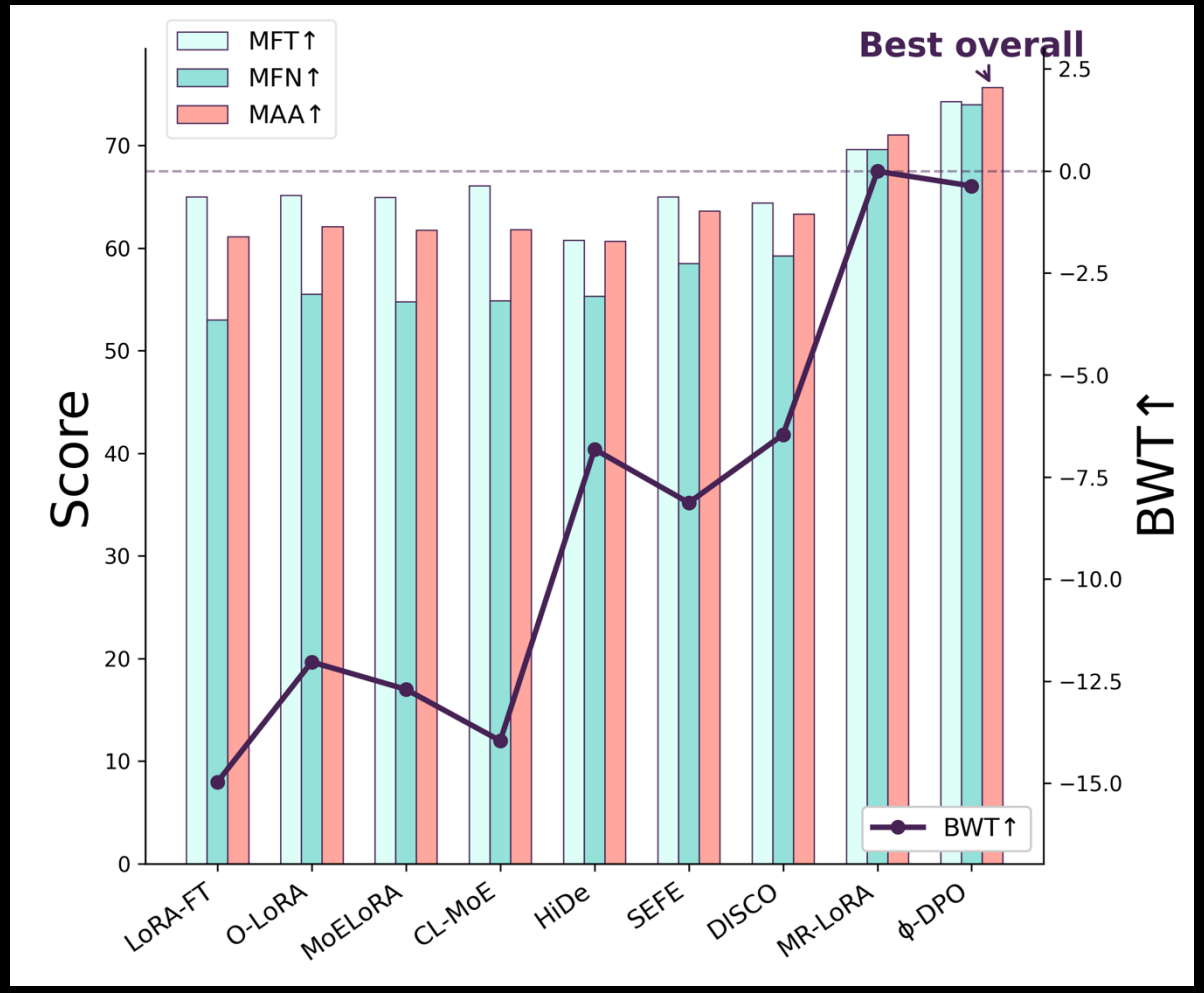


Results on CoIN Benchmark



Experimental Results

Results on MLLM-CL Domain Benchmark



Thank You For Watching!