Aligning Compound AI Systems via System-level DPO

Xiangwen Wang[1,2*] Yibo Jacky Zhang[1*] Zhoujie Ding[1] Katherine Tsai[1] Sanmi Koyejo[1]
[1]Stanford University [2]University of Science and Technology of China

Motivation & Introduction

Compound AI systems consist of multiple interacting AI components.
Examples: LLM + image generator; multi-agent systems.

- The example below shows GPT-4's inconsistent collaboration with DALL-E. User prompt: "Generate three separate images of a cat being progressively angrier."

(a) Calm Cat (b) Slightly Irritated Cat (c) Very Angry Cat
(d) Slightly Annoyed Cat (e) Angry Cat (f) Furious Cat

Open problems: aligning compound AI systems, due to

1. **Non-differentiability:** prevents end-to-end gradient optimization such as vanilla DPO and RLHF.
2. **Credit assignment**: the system's preference not easily decompose into individual component's preference.
3. **Datasets**: alignment datasets may exist for the system's overall task, but not for the sub-tasks of components.

Question

How to align compound AI systems in a principled way?

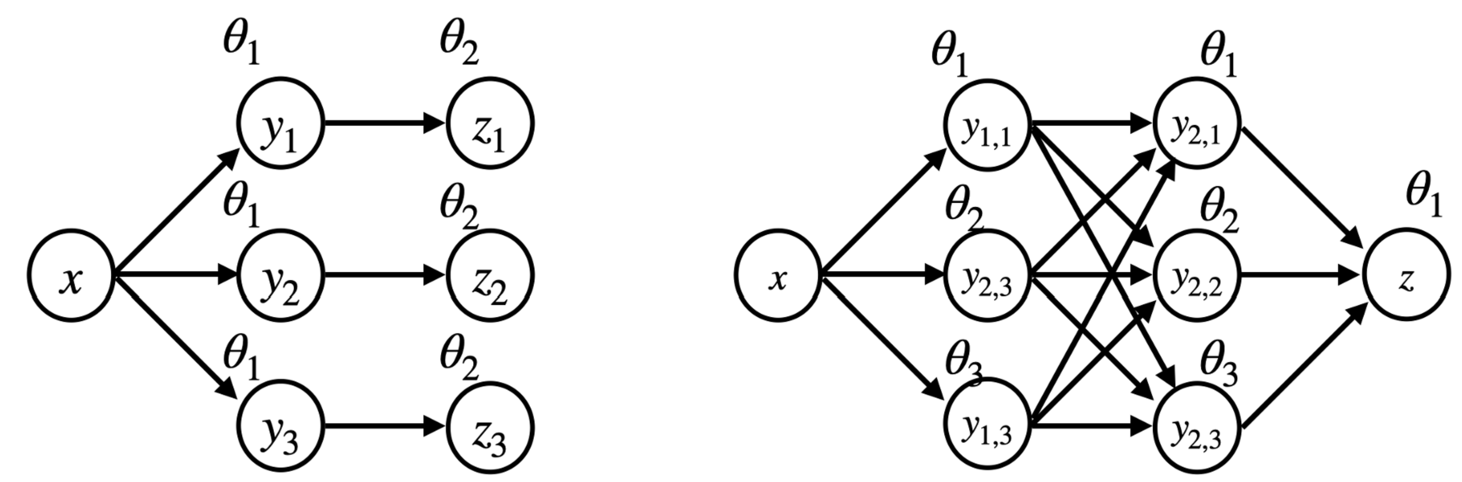
Contributions

- Define the problem of alignment of compound AI system; propose the SysDPO Framework for solving it;
- Apply SysDPO to align a system of an LLM agent and a text-to-image diffusion model;
- Demonstrate that aligning compound AI systems increases the performance complex tasks.

Our work represents an initial step in forming a foundation for aligning compound AI systems as cohesive entities.

The SysDPO Framework

1. **System Representation.** We represent the compound AI system as a Directed Acyclic Graph (DAG). Node x is the input; y_i are intermediate outputs; z_j are final outputs.



(a) LLM + Diffusion Models (b) Mixture-of-agents[3]

2. **Probability Factorization.** The DAG structure encodes the conditional independence of the generated data. Denote $s = \{y_i, z_j\}_{i\in I, j\in J}$ as the set of all generated outputs.

$$p_\theta(s|x) = \prod_{i,j} \; p_{\theta_i}(y_i|parent(y_i)) \cdot p_{\theta_j}(z_j|parent(z_j))$$

3. **Preference Dataset Construction.** Given a query x, the system generates two versions of the responses: s^w, s^l.
4. **Loss Function Design.** Given a dataset of (x, s^w, s^l), an AI system formulated as a DAG, we can apply DPO:

$$L(\theta) = -\mathbb{E}\left[\log\sigma\left(\beta\log\frac{p_\theta(s^w|x)}{p_{\bar\theta}(s^w|x)} - \beta\log\frac{p_\theta(s^l|x)}{p_{\bar\theta}(s^l|x)}\right)\right],$$

where $\bar\theta$ is the reference model.

Application: LLM + Diffusion Model

Goal: apply SysDPO to a group-image-generation application (Figure (a)): an LLM ψ and a Diffusion Model ϕ.
Issue: the diffusion model does not directly provide the likelihood p_ϕ.
Method: to obtain a tractable loss function in this application, we prove the following theorem.

Theorem (Sketched)

$$L(\psi,\phi) \le -\mathbb{E}\left[\log\sigma\left(\beta\left(A^w - A^l\right)\right)\right], \text{ where}$$

$$A^w = \log\frac{p_\psi(y^w|x)}{p_{\bar\psi}(y^w|x)} + T\sum_i\left(-\ell_\epsilon(\phi; z_i^w, y_i^w) + \ell_\epsilon(\bar\phi; z_i^w, y_i^w)\right)$$

similarly for A^l; T is the num. of steps of the diffusion.

In the above, $\ell_\epsilon(\bar\phi; z_i^w, y_i^w)$ is the denoising loss function of the diffusion model.

Experiments

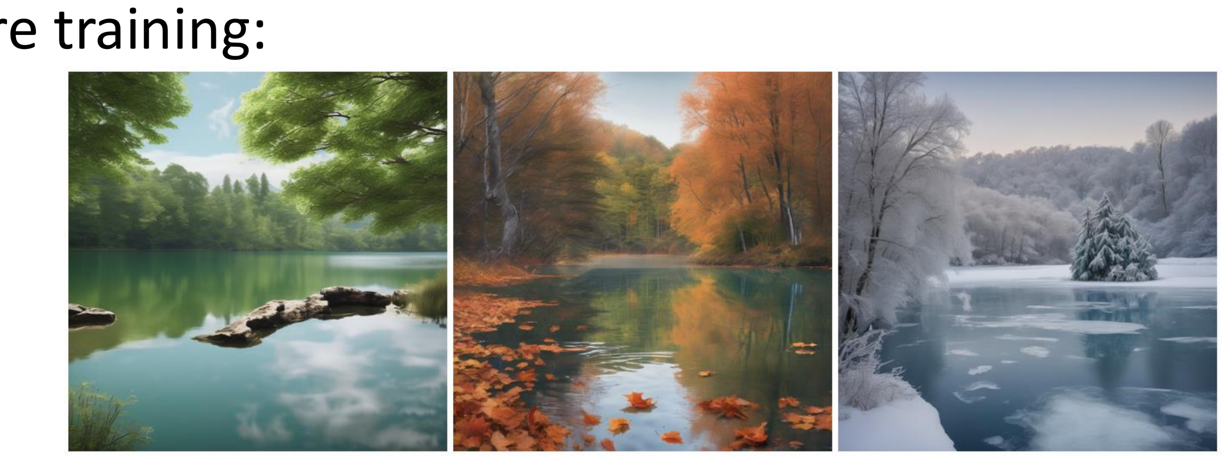
Task: multi-modal progression, where the system generates image sequences with a **progressively changing attribute**.
Dataset Construction:
1. 40 scene-related attributes (e.g., brightness, fog density) scored by a regressor.
2. GPT-4 generates 250 prompts for each attribute.
3. 6000 comparison pairs created by ranking generated sequences with the Preference Score (q).

Evaluation Metrics:
Preference Score (q): Measures **ordering consistency and evenness** of generated sequences.
Order Consistency Ratio: Evaluates how often sequences **maintain the correct order.**

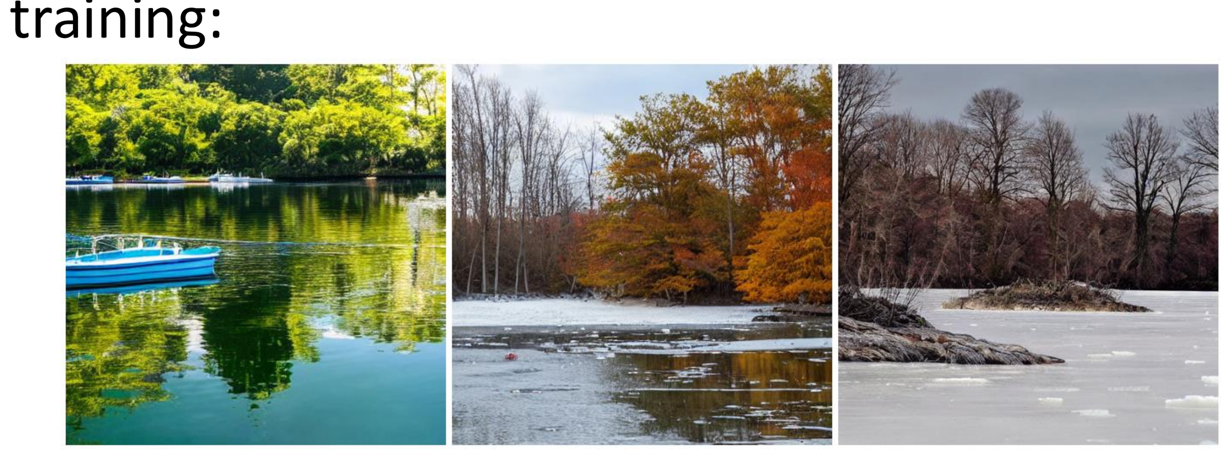
Results:

Method	Preference Score	Order Consistency Ratio
SysDPO (Proposed)	**0.25**	**70%**
System Before Alignment	-0.20	32%
Best-of-4 Sampling	0.16	67%
Only Train Language Model	0.23	65%
Only Train Diffusion Model	-0.03	35%

Visual examples: Prompt — "I want to see a series of images of a lake as the ice increases."
Before training:



After training:



The SysDPO approach significantly outperforms baselines, achieving the highest Preference Score (0.25) and Order Consistency Ratio (70%), demonstrating its ability to align compound AI systems effectively.

* Equal contribution

[3]Mixture-of-Agents Enhances Large Language Model Capabilities. Wang et al.