**What the paper is doing:**

Diffusion models are great at generating high-quality images. However, they don’t align well with human preferences.

Goal: Integrate human preferences into diffusion model training using RL.

The paper introduces DDPO (Denoising Diffusion Policy Optimization), a reinforcement learning-based method that treats the denoising process as a multi-step decision process (an MDP).

Questions:

1. Why DDPO is only setting awards for the final image? Is it possible or will it be better to set some other rewards in the middle or somewhere during the whole process?