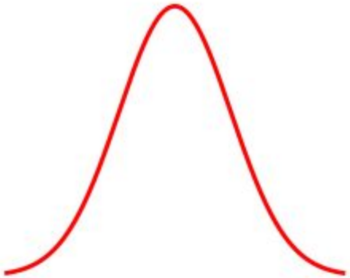
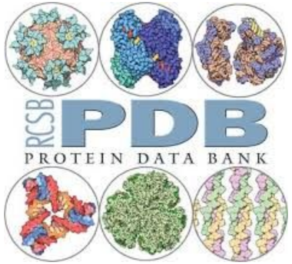
Test-time Alignment of Diffusion Models without Reward Over-optimization

Sunwoo Kim, Minkyu Kim, Dongmin Park

Why Alignment of Diffusion Models?



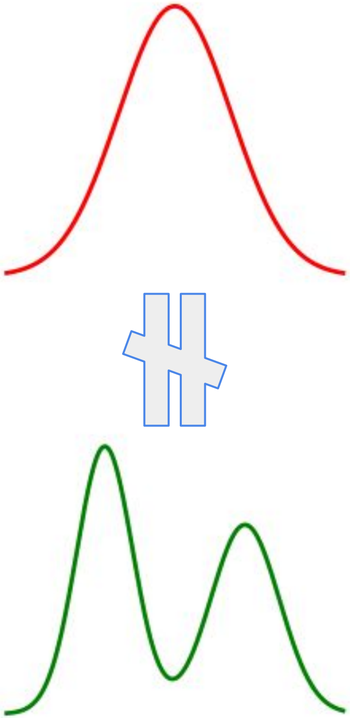
Pre-trained Distribution



$\neq$

Desired Distribution

Why Alignment of Diffusion Models?



Pre-trained Distribution

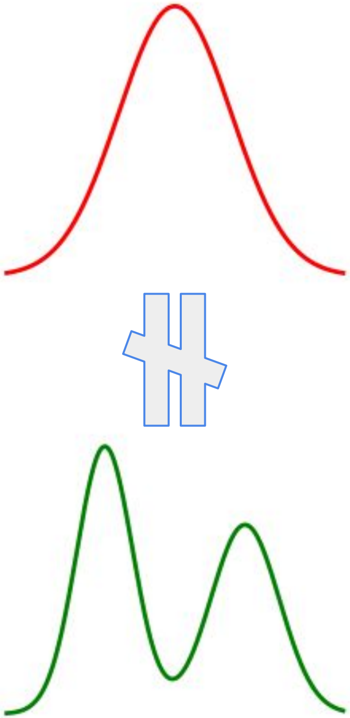
– Low **aesthetic quality**

crocodile

Desired Distribution

+ High **aesthetic quality**

Why Alignment of Diffusion Models?



Pre-trained Distribution

– Low aesthetic quality

– Low **text-image alignment**

Desired Distribution

+ High aesthetic quality

+ High **text-image alignment**

crocodile

cat and a dog

Alignment Without Over-optimization

Alignment Without Over-optimization

Pre-trained

Alignment Without Over-optimization

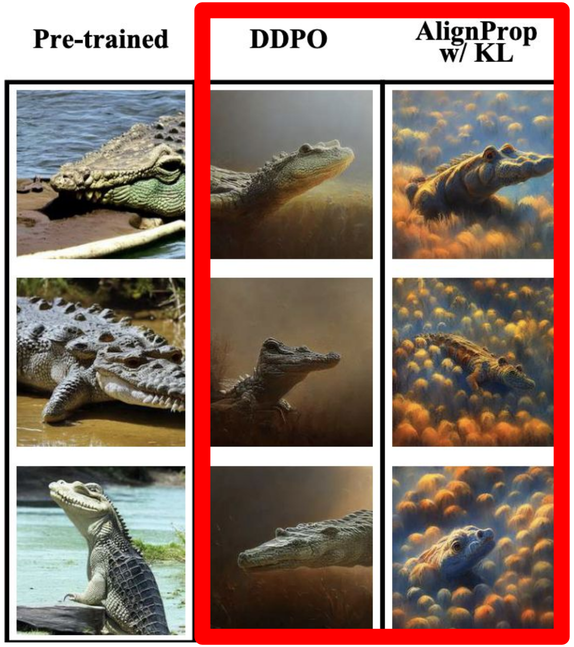
Target Reward:

Aesthetic

Alignment Without Over-optimization

Target Reward:

Aesthetic



Reward **Over-optimization**

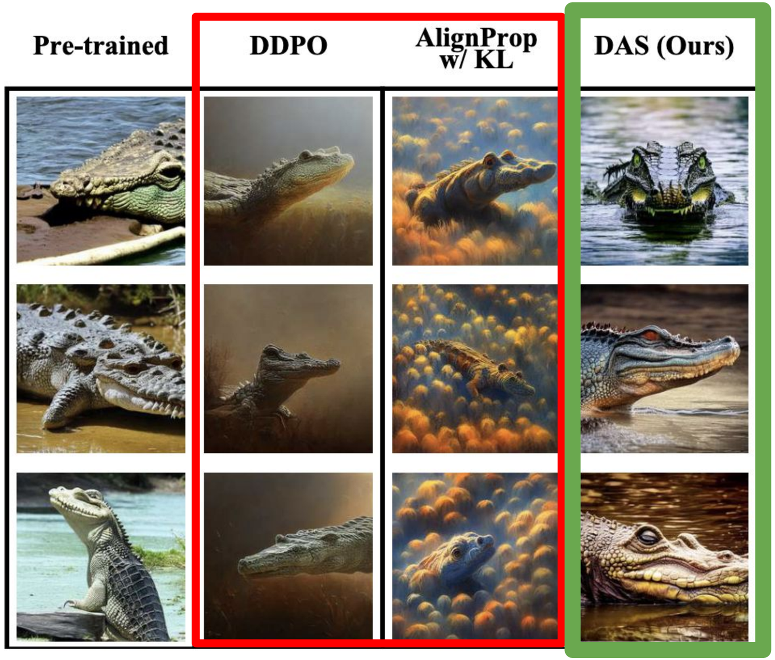
– **Low** Diversity

– **Low** Unseen Reward

Alignment Without Over-optimization

Target Reward:

Aesthetic



Reward **Over-optimization**

– Low Diversity

– Low Unseen Reward

+ **High** Target Reward

+ **High** Diversity

+ **High** Unseen Reward

Problem Formulation

$$p_{\text{tar}} = \arg\max_{p} \mathbb{E}_{x \sim p}\left[r\left(x\right)\right] - \alpha D_{\text{KL}}\left(p \,||p_{\text{pre}}\right)$$

Problem Formulation

$$p_{\text{tar}} = \arg\max_{p} \mathbb{E}_{x\sim p}\left[r\left(x\right)\right] - \alpha D_{\text{KL}}\left(p\,||p_{\text{pre}}\right)$$

maximize
expected reward

Problem Formulation

$$p_{\text{tar}} = \arg\max_{p} \mathbb{E}_{x \sim p}\left[r\left(x\right)\right] - \alpha D_{\text{KL}}\left(p \,||p_{\text{pre}}\right)$$

maximize
expected reward

stay close to
pre-trained distribution

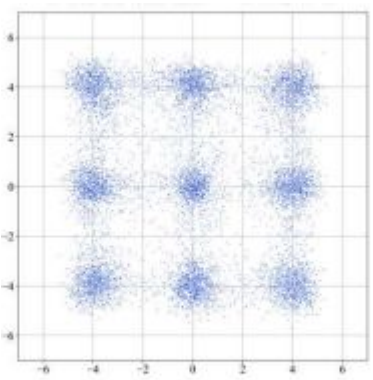
Problem Formulation

$$p_{\mathrm{tar}} = \arg\max_{p} \mathbb{E}_{x \sim p}\left[r\left(x\right)\right] - \alpha D_{\mathrm{KL}}\left(p \,||p_{\mathrm{pre}}\right)$$

$$p_{\mathrm{tar}} = \frac{1}{\mathcal{Z}} p_{\mathrm{pre}}\left(x\right) \exp\left(\frac{r\left(x\right)}{\alpha}\right)$$

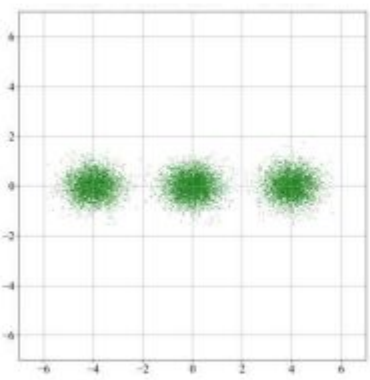
*Optimization ≈ **Inference***

Reward Aligned Target Distribution



$$p_{\text{pre}}$$

Higher reward if closer to x-axis

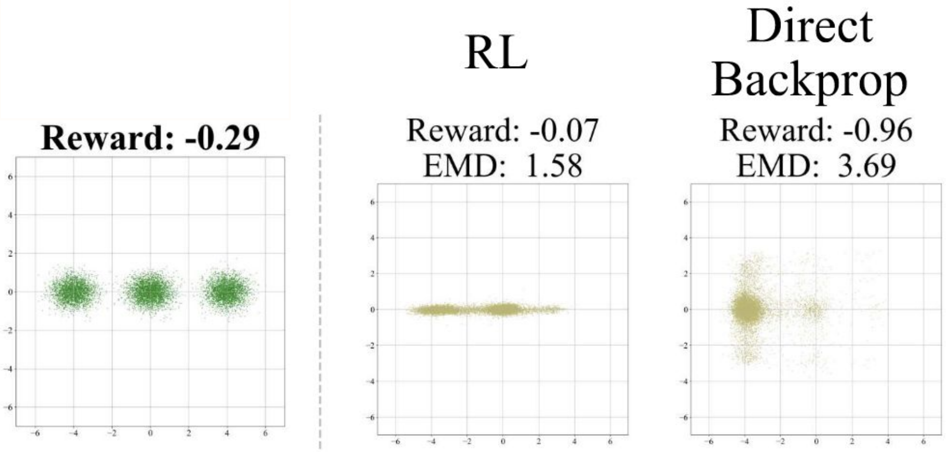


$$p_{\text{tar}}$$

Limitations in Prior Works - **Fine-tuning**

Solve '**Optimization**' Problem

$$\underset{\theta}{\text{minimize}} \mathcal{D}_{\mathrm{KL}}(p_\theta \| p_{\mathrm{tar}})$$



p_{tar}

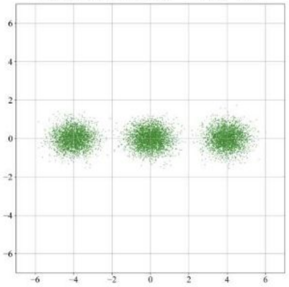
mode seeking
= ***over-optimize***

Limitations in Prior Works - **Guidance**

Solve ‘**Inference**’ Problem

$$\nabla_{x_t} \log p_{\mathrm{tar},\,t}(x_t) \approx \nabla_{x_t} \log p_{\mathrm{pre},\,t}(x_t) + \frac{1}{\alpha}\nabla_{x_t} r(\hat{x}_0(x_t))$$

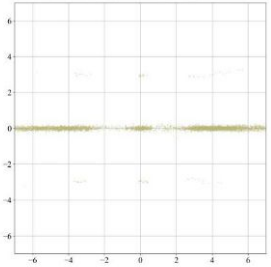
Reward: -0.29



p_{tar}

low reward
= ***under-optimize***

Approx.
Guidance
Reward: -338
EMD: 15.49

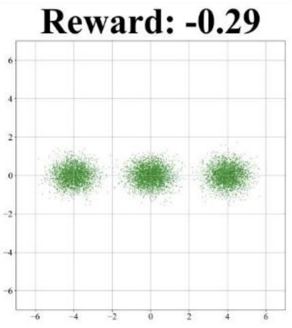


Solution - DAS (Diffusion Alignment as Sampling)

Solve ‘**Inference**’ Problem

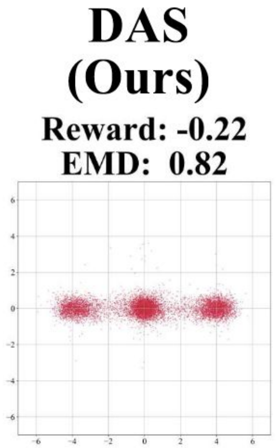
\+ **High target reward** via test-time search based on **tempered SMC**

\+ **Overcome over-optimization** via direct sampling



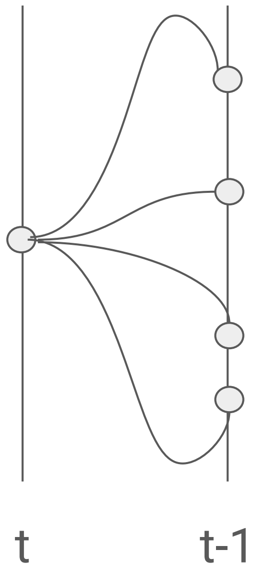
p_{tar}

High target reward
w/o mode seeking



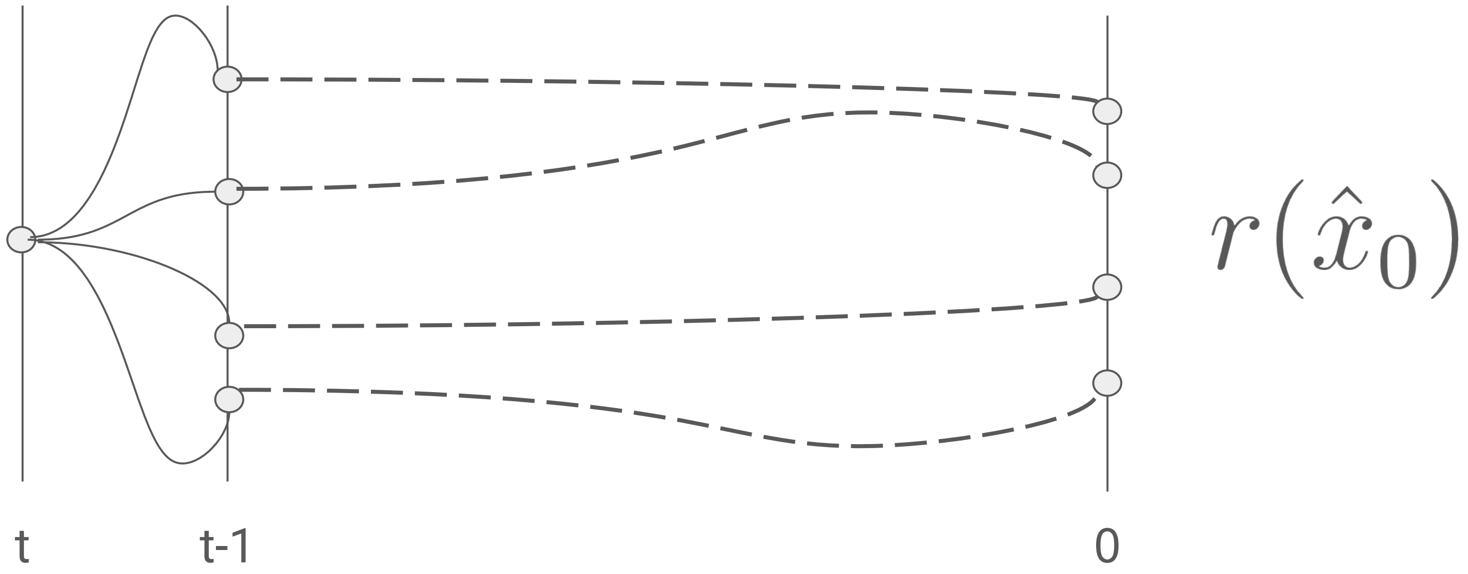
DAS: Method Overview

1. **Propose** multiple samples at current denoising step using **reward guidance**



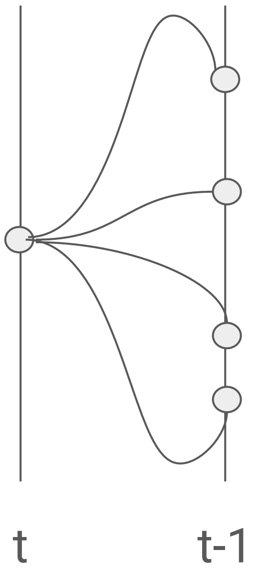
DAS: Method Overview

2. Estimate **expected rewards** of each samples using one-step denoising



DAS: Method Overview

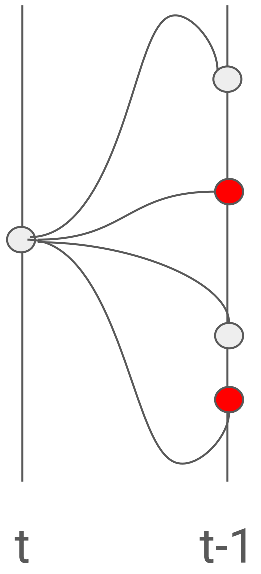
3. Calculate the **probability** of each samples **respect to pre-trained model**



$$p_{\mathrm{pre},\, t}(x_{t-1})$$

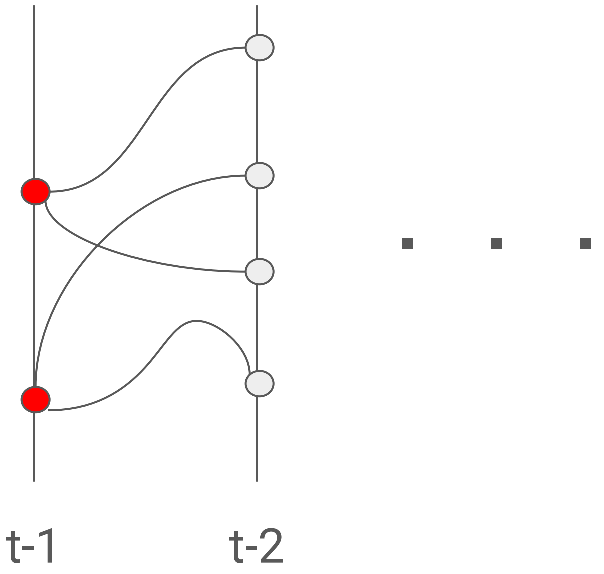
DAS: Method Overview

4. Combine two criteria and **select top samples**

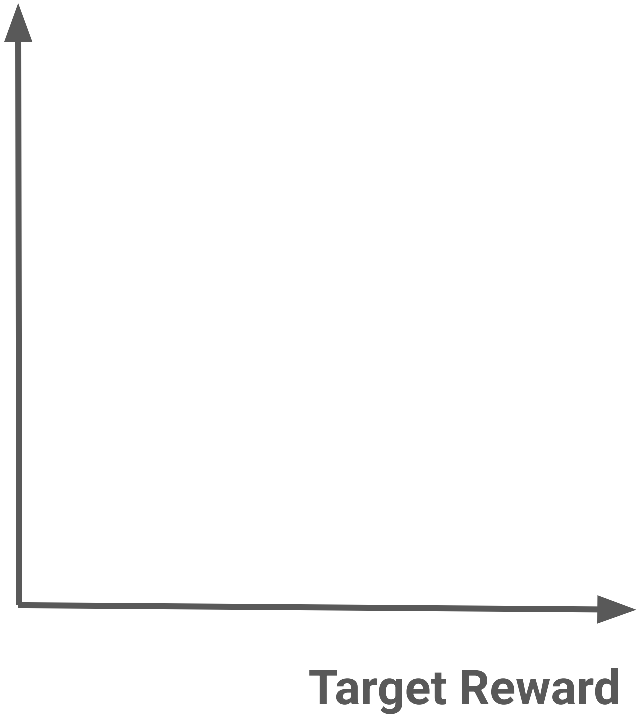


DAS: Method Overview

5. Repeat the previous with **tempering**



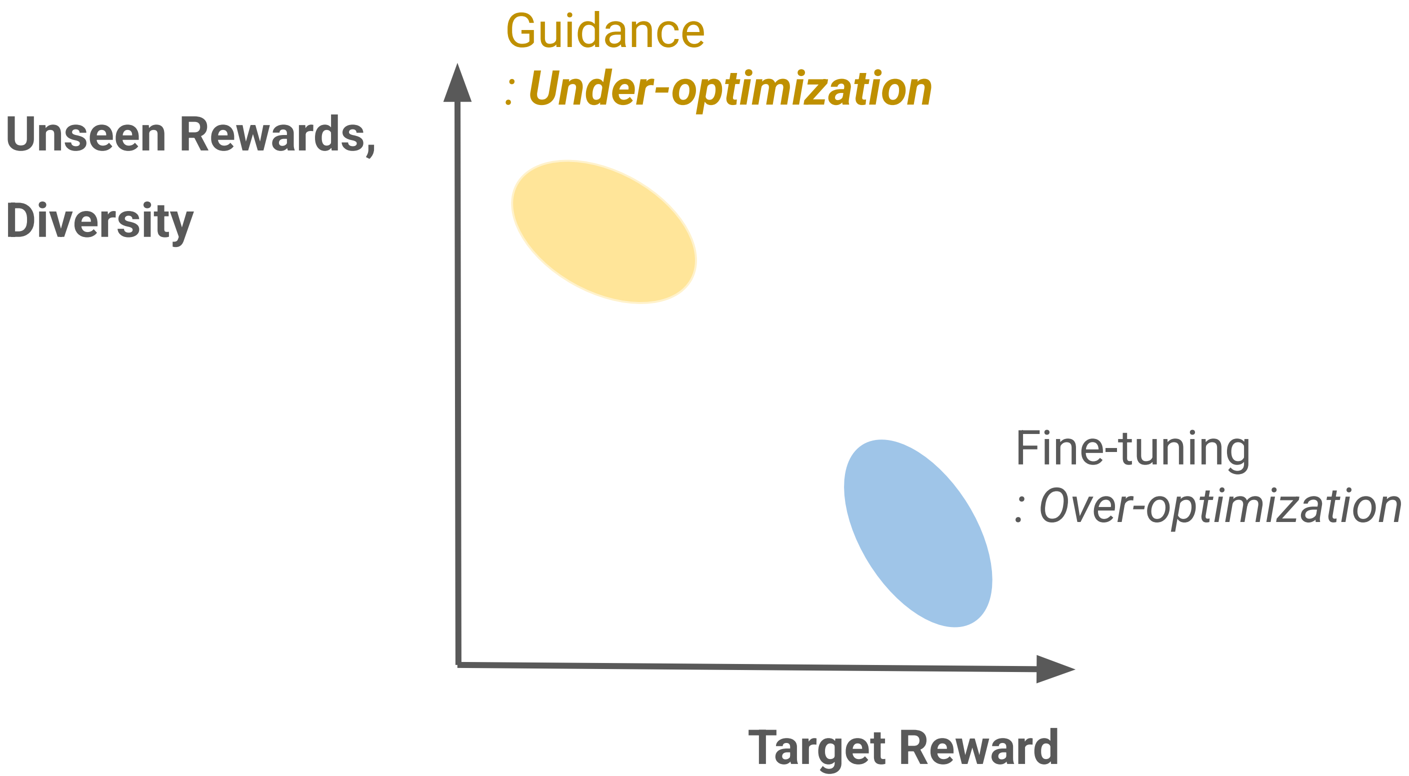
DAS is Pareto-optimal



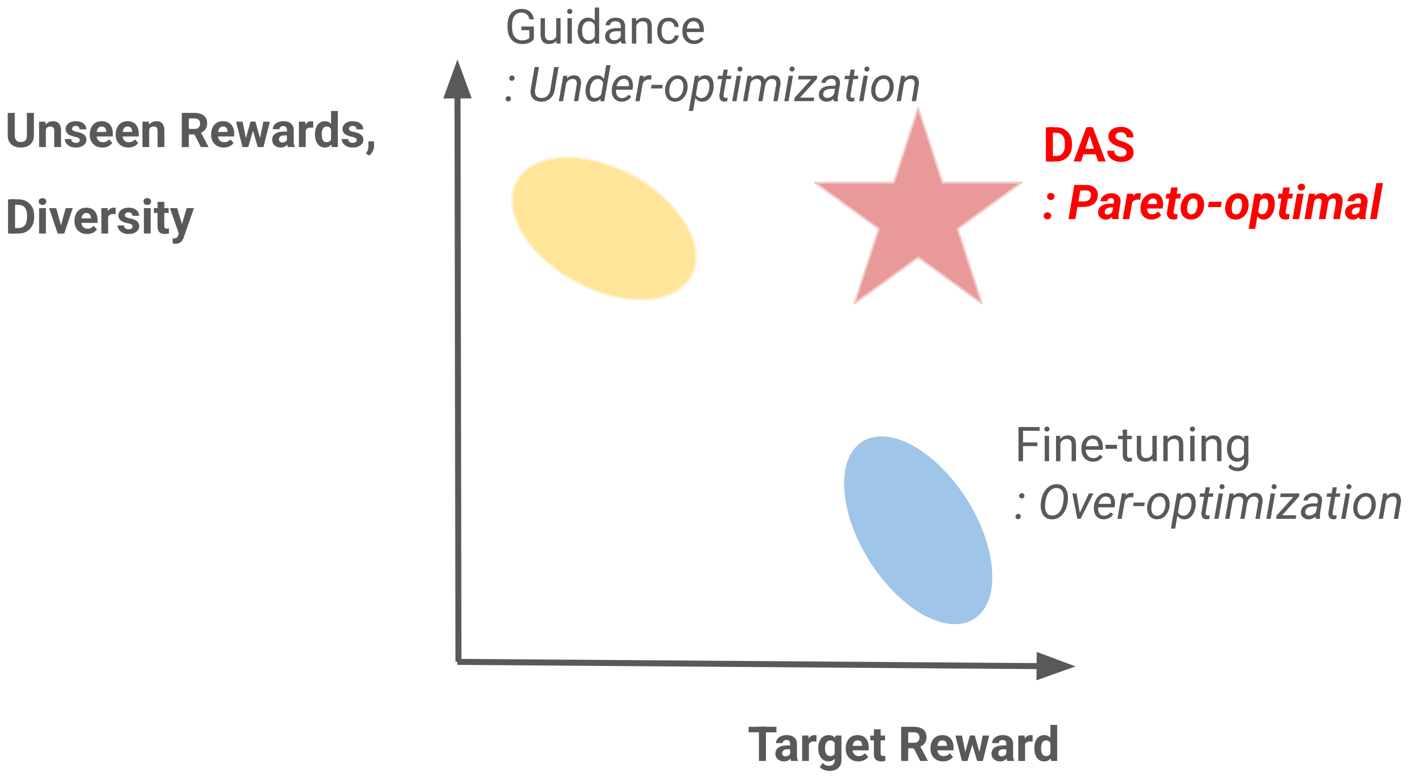
DAS is Pareto-optimal

DAS is Pareto-optimal

DAS is Pareto-optimal

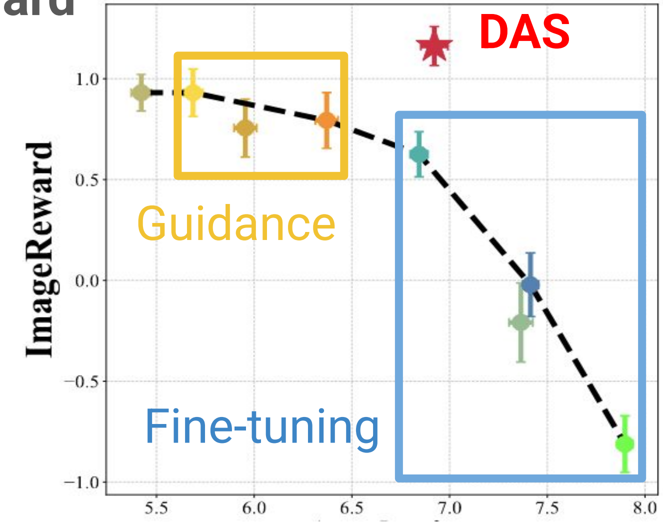


DAS is Pareto-optimal

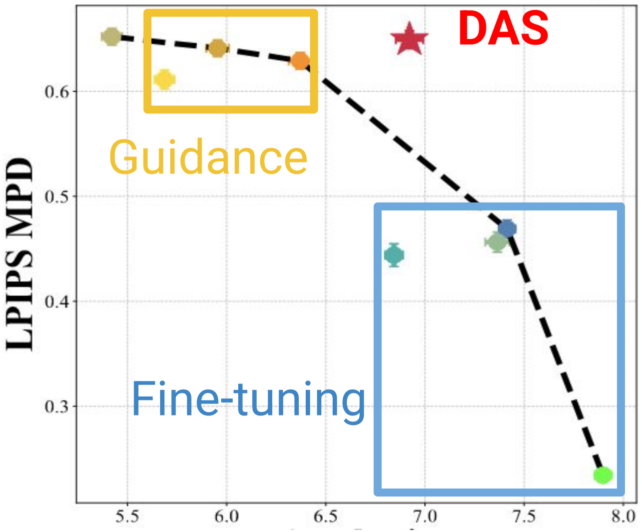


DAS is Pareto-optimal

Unseen Reward



Diversity

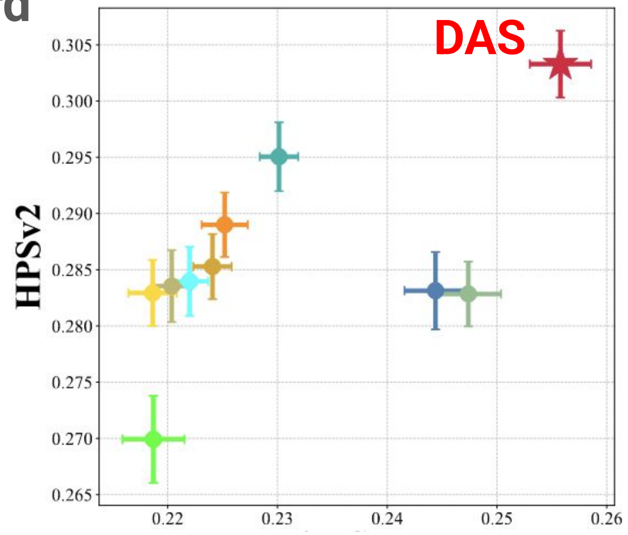


Aesthetic

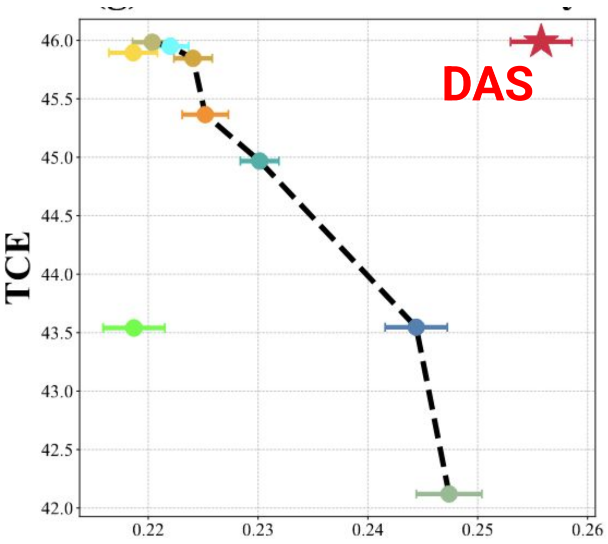
DAS is Pareto-optimal

Unseen

Reward



Diversity

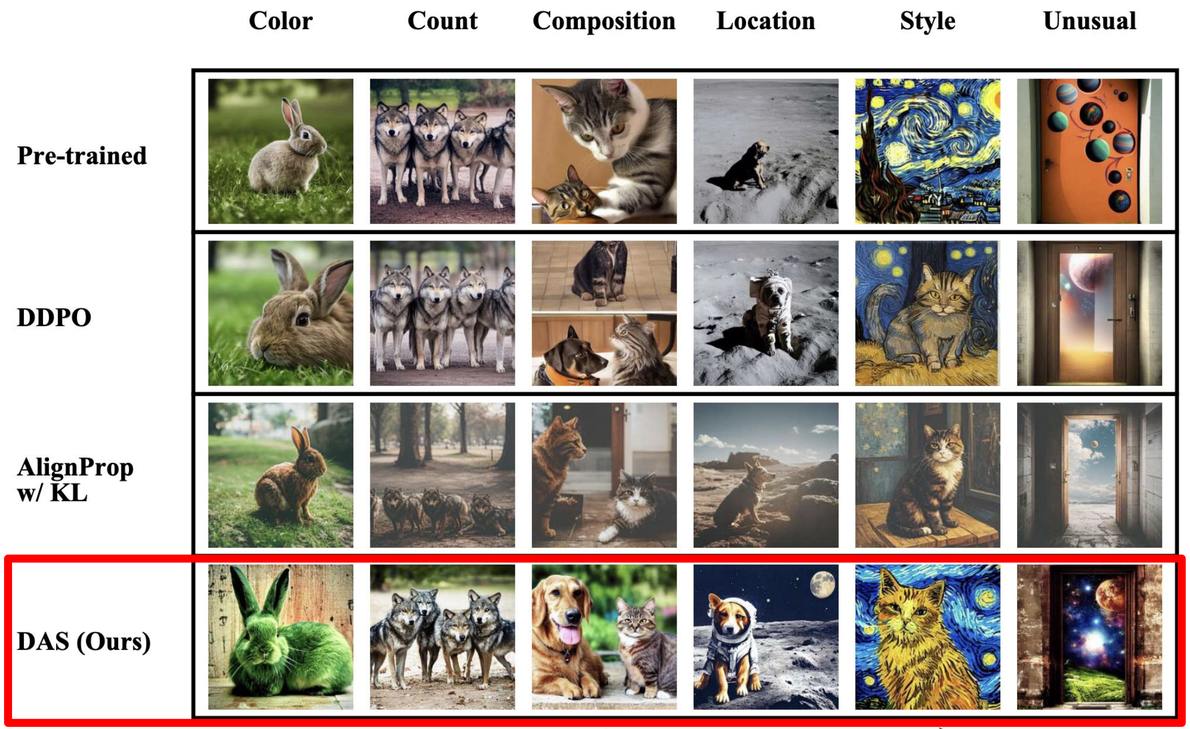


Human Preference

DAS effectively Optimizes Rewards

Target Reward:

Human Preference



green colored rabbit

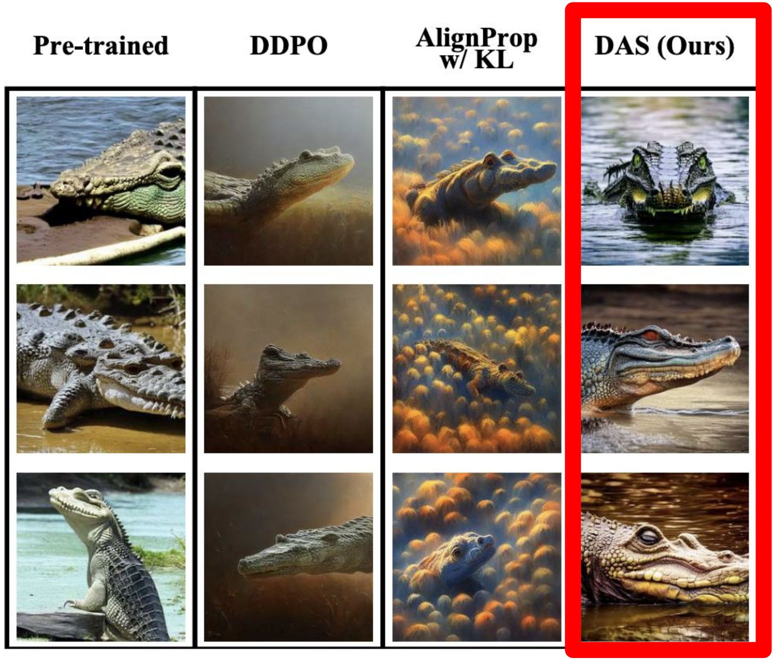
cat and a dog

cat in the style of Van Gogh

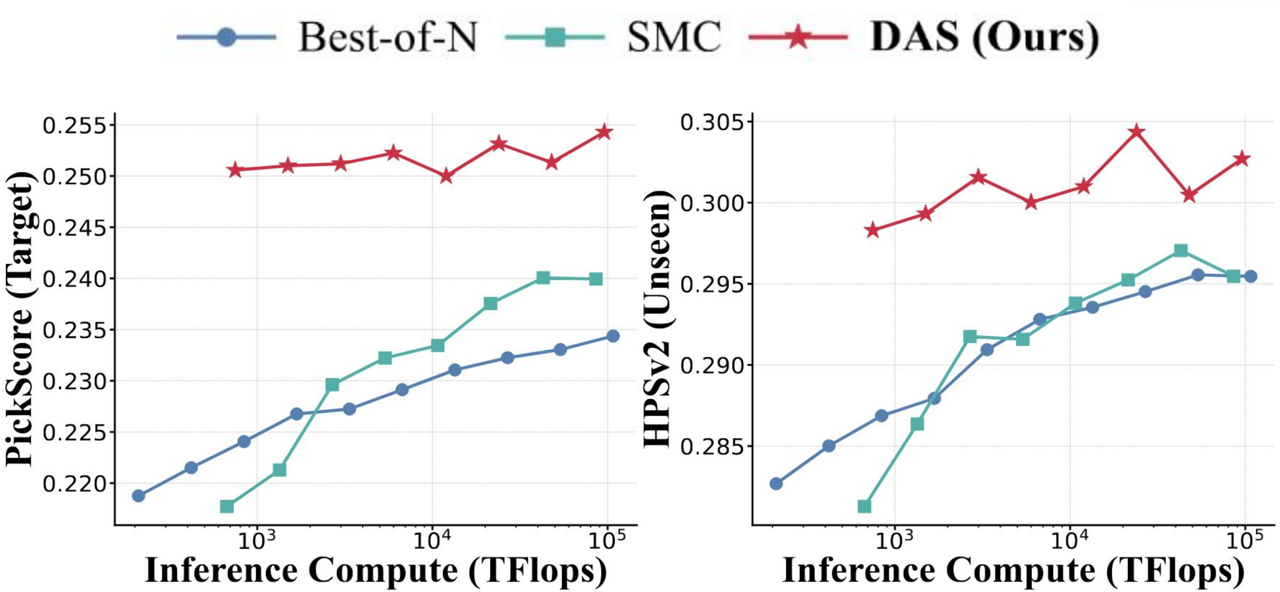
DAS effectively Mitigates Over-optimization

Target Reward:

Aesthetic



Compute-efficient Test-time Scaling



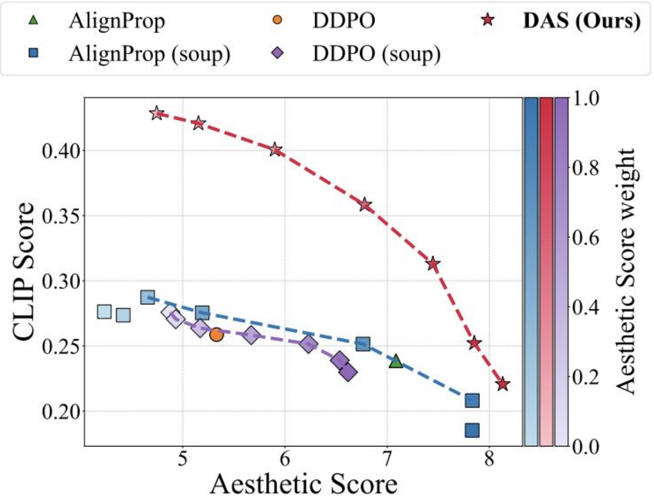
HOW?

- **Proposal** Distribution
- **Tempering** Technique

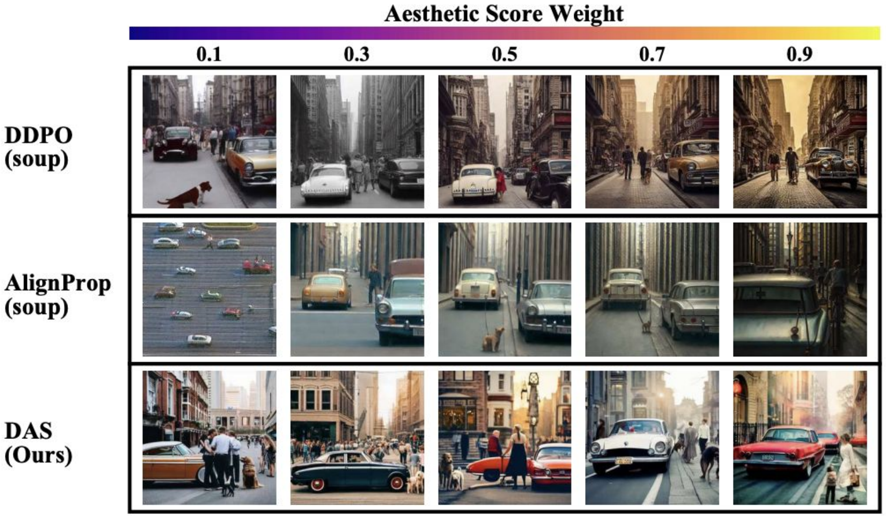
=> For theoretical properties, check out the paper!

Multi-reward Alignment

$$w \cdot \text{Aesthetic Score} + (1 - w) \cdot 20 \cdot \text{CLIPScore}$$



(a) Trade-off in multi-objective optimization.



(b) Generated samples according to reward weights

Thank You!

For more:

📄 paper: https://openreview.net/forum?id=vi3DjUhFVm

💻 code: https://github.com/krafton-ai/DAS