Subject-driven Text-to-Image Generation via Preference-based Reinforcement Learning

Yanting Miao, William Loh, Suraj Kothawade, Pascal Poupart, Abdullah Rashwan, Yeqing Li

The Challenge of Existing Diffusion Models

Hard to keep identity

Reference Images

A corgi dog wearing a green sunglasses

Goal: Design an efficient way to fine-tune diffusion models for subject-driven tasks.

Method

Intuitively, the objective function can be the image similarity loss, $\mathscr{L}_{\text{sim}}$, for the reference dataset.

Overfit!

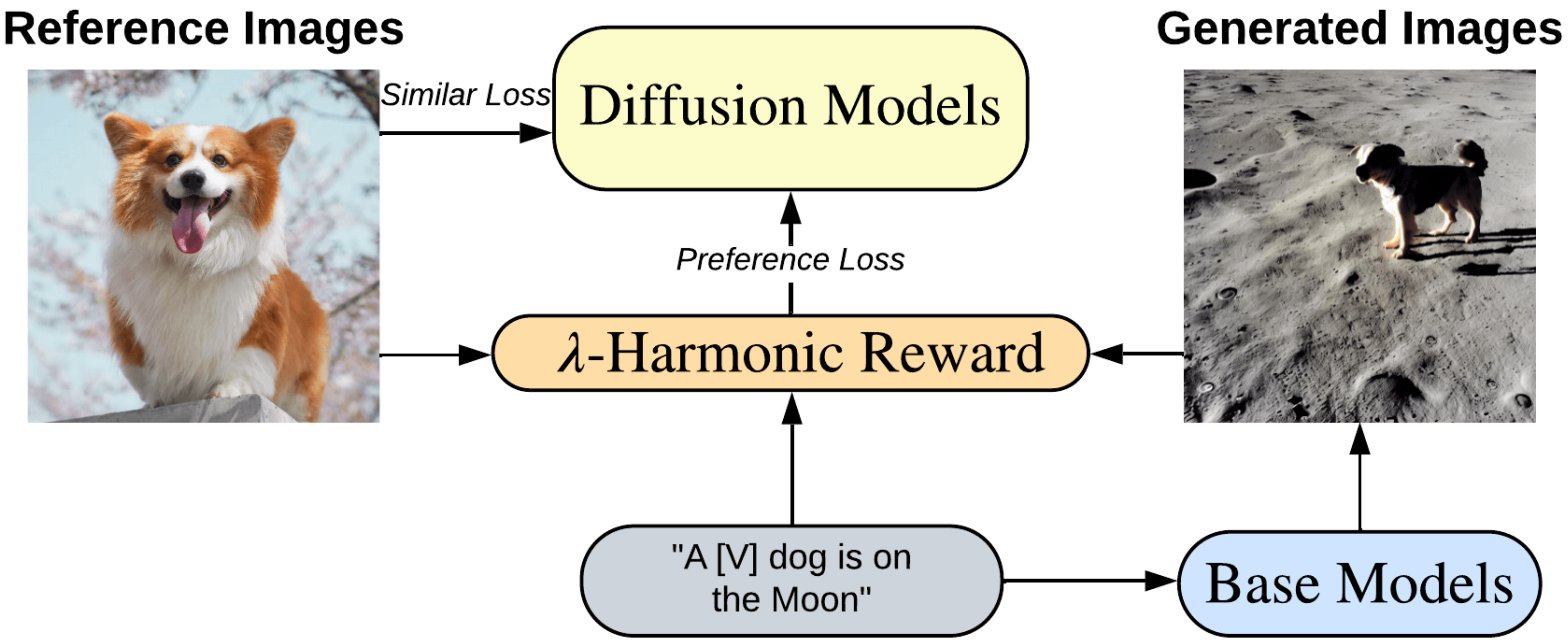
Reference Image

pink sunglasses on the beach

Method

Solution: Using preference-based RL for text-to-image alignment.

RPO: Reward Preference Optimization



$$\mathscr{L} = \mathscr{L}_{\text{sim}} + \mathscr{L}_{\text{pref}}$$

Method

$$\lambda\text{-Harmonic reward} := \frac{1}{\frac{\lambda}{\text{ALIGN-I}} + \frac{1-\lambda}{\text{ALIGN-T}}},$$

where ALIGN-I and ALIGN-T are the image alignment and text alignment scores, respectively.

During training, $\lambda_{\text{train}} = 0.$

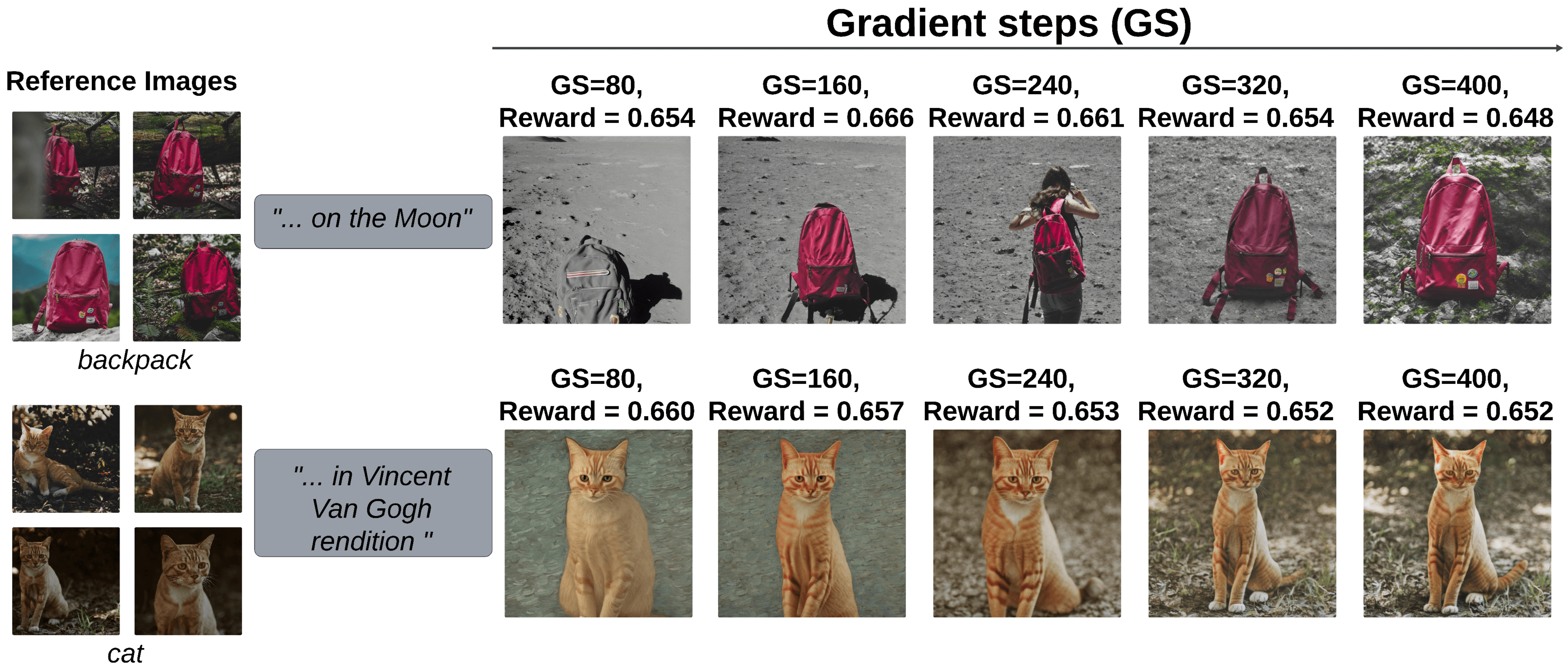
Method

$$\lambda\text{-Harmonic reward} := \frac{1}{\frac{\lambda}{\text{ALIGN-I}} + \frac{1-\lambda}{\text{ALIGN-T}}},$$

During validation, $\lambda_{\text{val}} > 0$.

Observation

Reward changes in the $\lambda_{\text{val}} = 0.3$ during training process.



Early-stopping also alleviate overfitting.

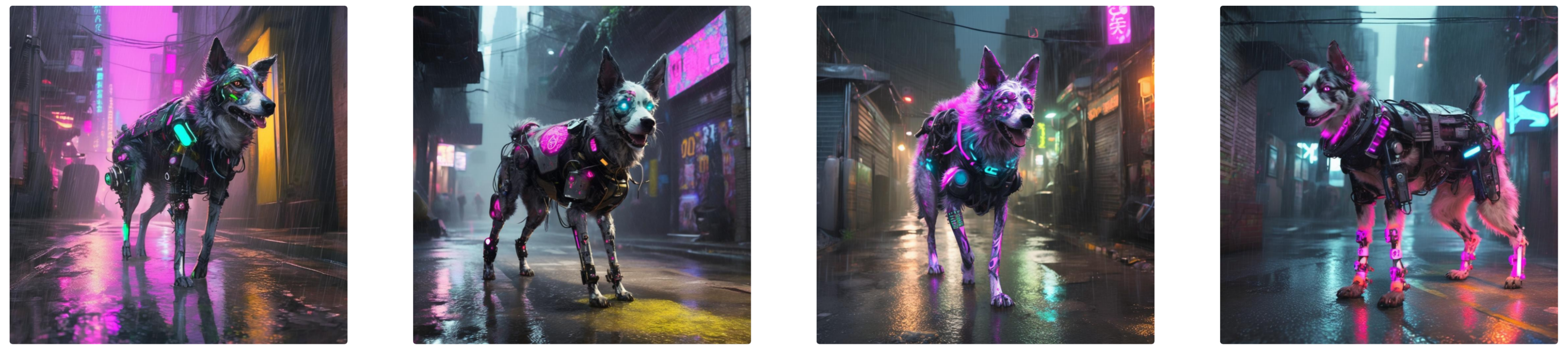
Quantitative Results

Method	Backbone	Iterations ↓	DINO ↑	CLIP-I ↑	CLIP-T ↑
Reference Images	N/A	N/A	0.774	0.885	N/A
DreamBooth [23]	Imagen [24]	1000	0.696	0.812	0.306
DreamBooth [23]	SD [22]	1000	0.668	0.803	0.305
Textual inversion [12]	SD [22]	5000	0.569	0.780	0.255
SuTI [7]	Imagen [24]	1.5×10^5	**0.741**	0.819	0.304
Re-Imagen [8]	Imagen [24]	2×10^5	0.600	0.740	0.270
DisenBooth [6]	SD[22]	3000	0.574	0.755	0.255
Custom Diffusion [16]	SD[22]	500	0.695	0.801	0.245
ELETE [31]	SD[22]	3000	0.652	0.765	0.255
IP-Adapter [32]	SD[22]	10^6	0.608	0.809	0.274
SSR-Encoder [33]	SD[22]	10^6	0.612	0.821	**0.314**
Ours: RPO	SD [22]	**400**	0.652	**0.833**	**0.314**

We report results for $\lambda_{\text{val}} = 0.5$.

Qualitative Results

Reference Images



" a cyberpunk [V] dog with neon fur patterns and robotic limbs walking down a rain-soaked alleyway in a bustling city"

"a superhero [V] dog soaring through the sky over a city skyline, with a cape billowing in the wind"

"an astronaut [V] dog in a sleek space suit floating weightlessly inside a spacecraft."

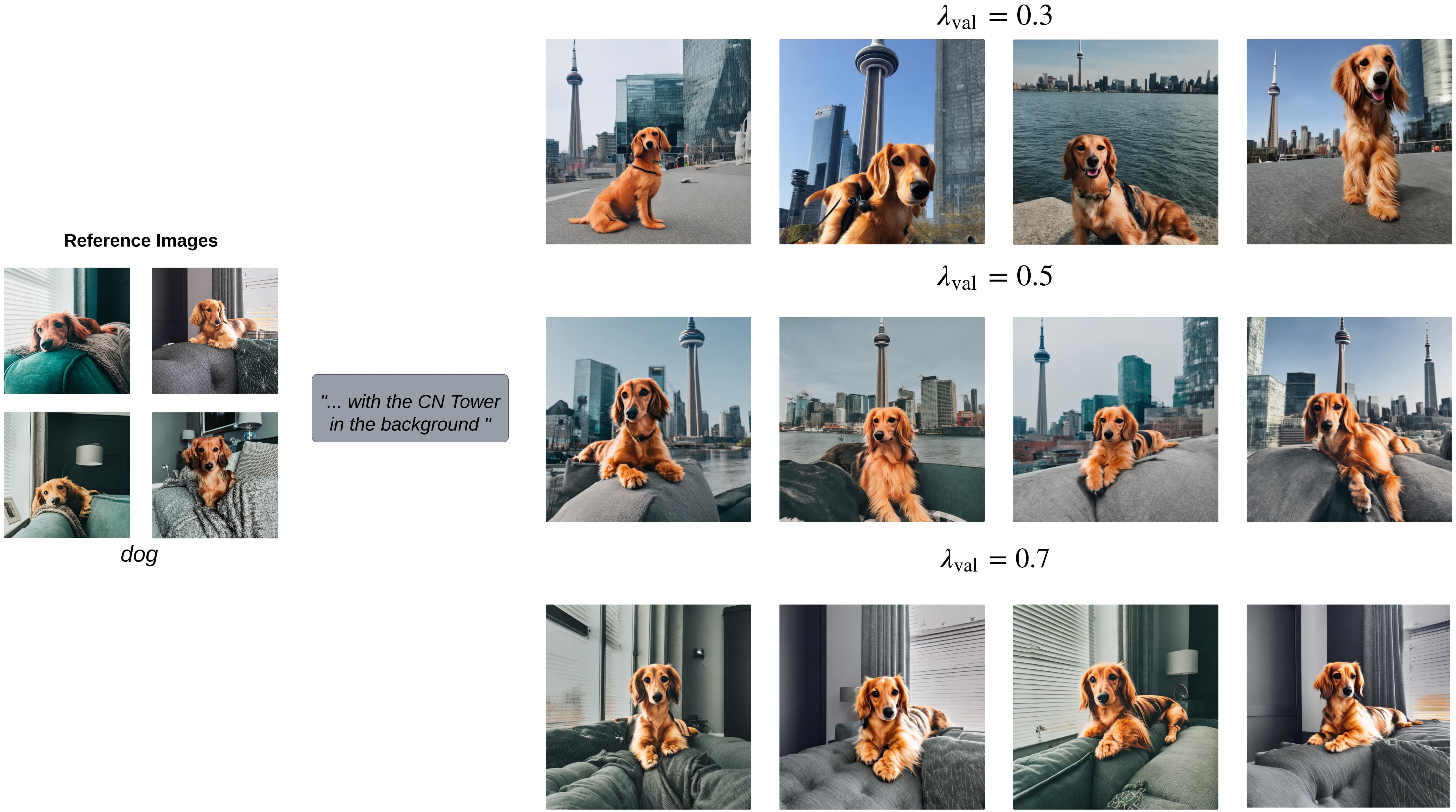
"an underwater [V] dog wearing a diving suit swimming through a vibrant coral reef surrounded by exotic marine life"

Ablation Study

Method	DINO ↑	CLIP-I ↑	CLIP-T ↑
Pure $\mathcal{L}_{\text{sim}}$	$\mathbf{0.695 \pm 0.077}$	$\mathbf{0.852 \pm 0.043}$	0.285 ± 0.027
$\mathcal{L}_{\text{pref}}$ w/o early-stopping	0.688 ± 0.082	0.845 ± 0.042	0.296 ± 0.027
Early-stopping w/o $\mathcal{L}_{\text{pref}}$	0.575 ± 0.124	0.799 ± 0.052	0.323 ± 0.025
RPO ($\lambda_{\text{val}} = 0.3$)	0.581 ± 0.113	0.798 ± 0.039	$\mathbf{0.329 \pm 0.021}$

Configuration	DINO ↑	CLIP-I ↑	CLIP-T ↑
$\lambda_{\text{val}} = 0.3$	0.581 ± 0.113	0.798 ± 0.039	$\mathbf{0.329 \pm 0.021}$
$\lambda_{\text{val}} = 0.5$	0.652 ± 0.082	0.833 ± 0.041	0.314 ± 0.022
$\lambda_{\text{val}} = 0.7$	$\mathbf{0.679 \pm 0.085}$	$\mathbf{0.850 \pm 0.045}$	0.304 ± 0.023

Ablation Study



Takeaway

- Introduce lambda-Harmonic reward function for the subject-driven tasks.
- Employ RPO to finetune the diffusion models.
- Lambda-Harmonic reward function serves as a model selection method.

SCAN ME