h-Edit: Effective and Flexible Diffusion-Based Editing via Doob’s h-Transform

Toan Nguyen*, Kien Do*, Duc Kieu, Thin Nguyen
A2I2, Deakin University, Australia. *Equal contribution

CVPR 2025

CVPR Nashville JUNE 11-15, 2025

DEAKIN UNIVERSITY | DEAKIN APPLIED ARTIFICIAL INTELLIGENCE INITIATIVE

Training-Free Image Editing with Diffusion Models (1)

Goal: Generate images that align with editing prompt (editing **accuracy**) while being **faithful** to the original image.

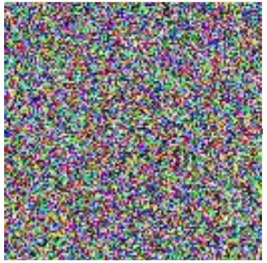
→ How to achieve and balance both targets? (**Problem 1**)

How diffusion models perform editing?

<u>Inversion</u>:

Source

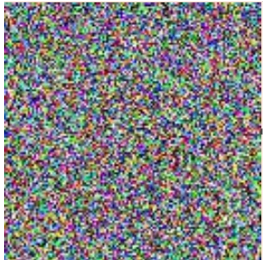
$$x_0^{\text{orig}} \longrightarrow x_T^{\text{orig}}$$



- Two popular types are **DDIM Inversion** and **DDPM Inversion** (slightly better).

<u>Editing</u>: Perform **step-by-step sampling** with editing conditions.

- Often requires further **attention control** for faithfulness.



$$x_T^{\text{orig}} \longrightarrow x_0^{\text{edit}}$$

Training-Free Image Editing with Diffusion Models (2)

Problem 2:

How to **combine** multiple types of editing effectively?

<u>For example</u>: Text-guided and Style Editing.

Training-Free Image Editing with Diffusion Models (3)

Problem 3:

- Current literature do not pay attention on the editing process of diffusion models, lacking of **theoretical foundation.**

→ Struggle with the first two problems!!

- Most of current text-guided methods focus on sampling from $p\left(x_{t-1} \mid x_t, c^{\text{edit}}\right)$

→ **No guarantee** to fall into the **target** distribution $p\left(x_0\right) p\left(c^{\text{edit}} \mid x_0\right)$

→ By this approach, there is **no control for the trade-off** between editing & faithfulness.

→ How to **combine** with other editing conditions?

Motivation of *h*-Edit (1)

- We know the abstract target distribution is $p\left(x_0\right) p_{\mathcal{Y}}\left(y \mid x_0\right)$

Can be **anything**: text, graph, reference images, audio.

- We know the starting distribution for editing is $p\left(x_T\right)$

- Why don't we **build a bridge** to connect them?

→ **Guarantee** to fall into the target distribution.

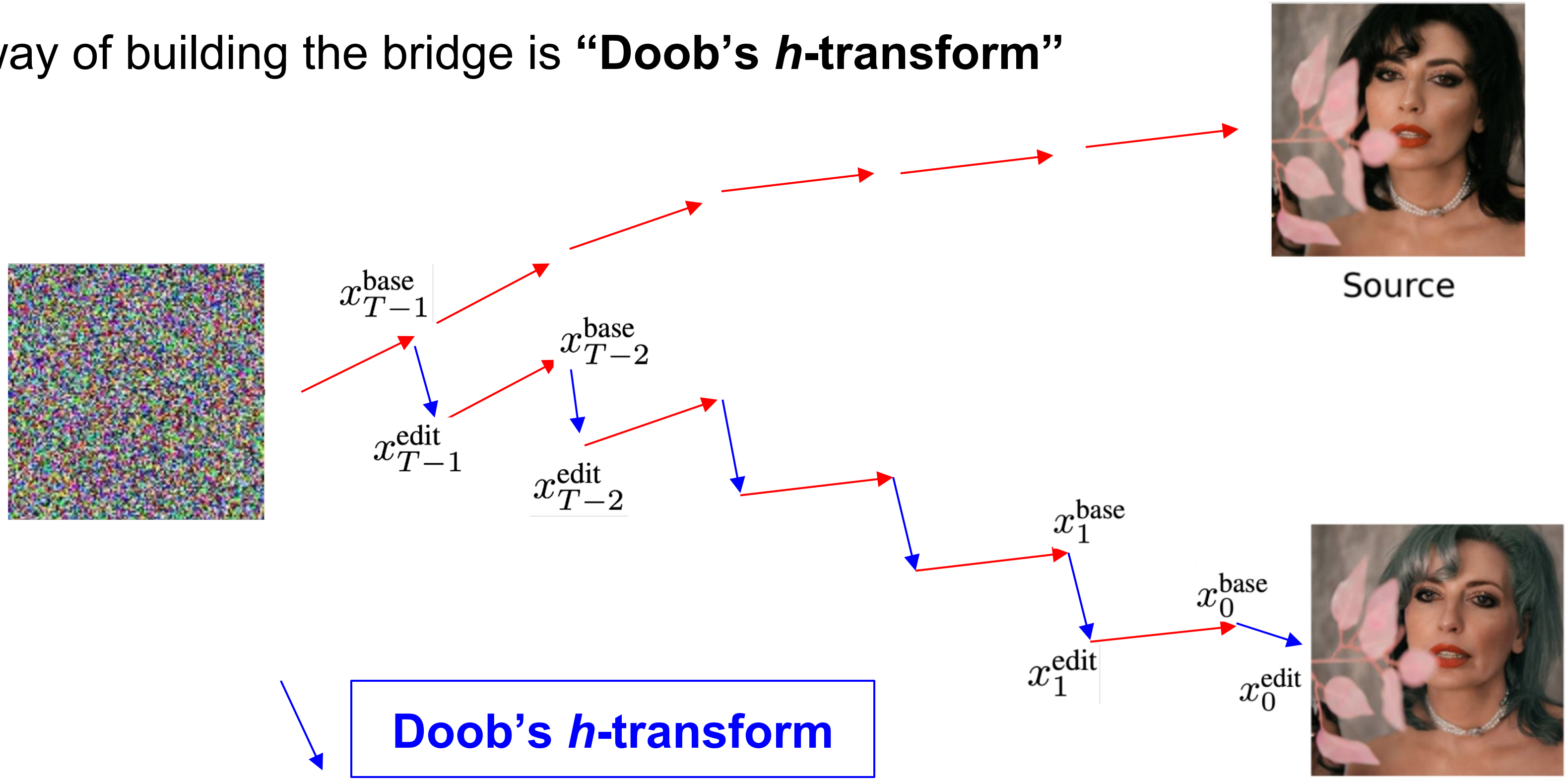
→ For **combining**, we adjust the target distribution to

$$p\left(x_0\right) p_{\mathcal{Y}}\left(y_1 \mid x_0\right) p\left(x_0\right) p_{\mathcal{Y}}\left(y_2 \mid x_0\right) \ldots p\left(x_0\right) p_{\mathcal{Y}}\left(y_n \mid x_0\right)$$

→ Naturally decompose the update into **"reconstruction" term and "editing" term**.

How *h*-Edit works?

A way of building the bridge is **“Doob’s *h*-transform”**



How *h*-Edit works? *h*-transform + LMC sampling!

Implicit form: $$p_\theta^h\left(x_{t-1}|x_t\right) = p_\theta\left(x_{t-1}|x_t\right)\frac{h\left(x_{t-1}, t-1\right)}{h\left(x_t, t\right)}$$ $p_{\mathcal{Y}}\left(y \mid x_{t-1}\right)$

Explicit form: $$p^h\left(x_t\right) = \frac{p(x_t)h(x_t,t)}{\mathbb{E}_{p(x_0)}[h(x_0,0)]}$$ $p_{\mathcal{Y}}\left(y \mid x_t\right)$

To sample x_{t-1}^{edit} , we perform Langevin Monte Carlo sampling to sample from $p_\theta^h\left(x_{t-1}|x_t\right)$ or with the score of $p^h\left(x_t\right)$

How *h*-Edit works? *h*-transform + LMC sampling! (2)

Explicit form:

$$
\begin{aligned}
x_{t-1} &\approx x_t + \eta \nabla_{x_t} \log \left(p\left(x_t\right) h\left(x_t, t\right) \right) + \sqrt{2\eta} z \\
&= \left(x_t + \eta \nabla_{x_t} \log p\left(x_t\right) + \sqrt{2\eta} z \right) \\
&\quad + \eta \nabla_{x_t} \log h\left(x_t, t\right) \\
&= \underbrace{x_{t-1}^{\text{base}}}_{\text{rec.}} + \eta \underbrace{\nabla_{x_t} \log h\left(x_t, t\right)}_{\text{editing}}
\end{aligned}
$$

- Naturally decompose the update into **“reconstruction” term and “editing” term**.
- We can have **multiple** “editing” terms with any form.

How *h*-Edit works? *h*-transform + LMC sampling! (3)

Implicit form:

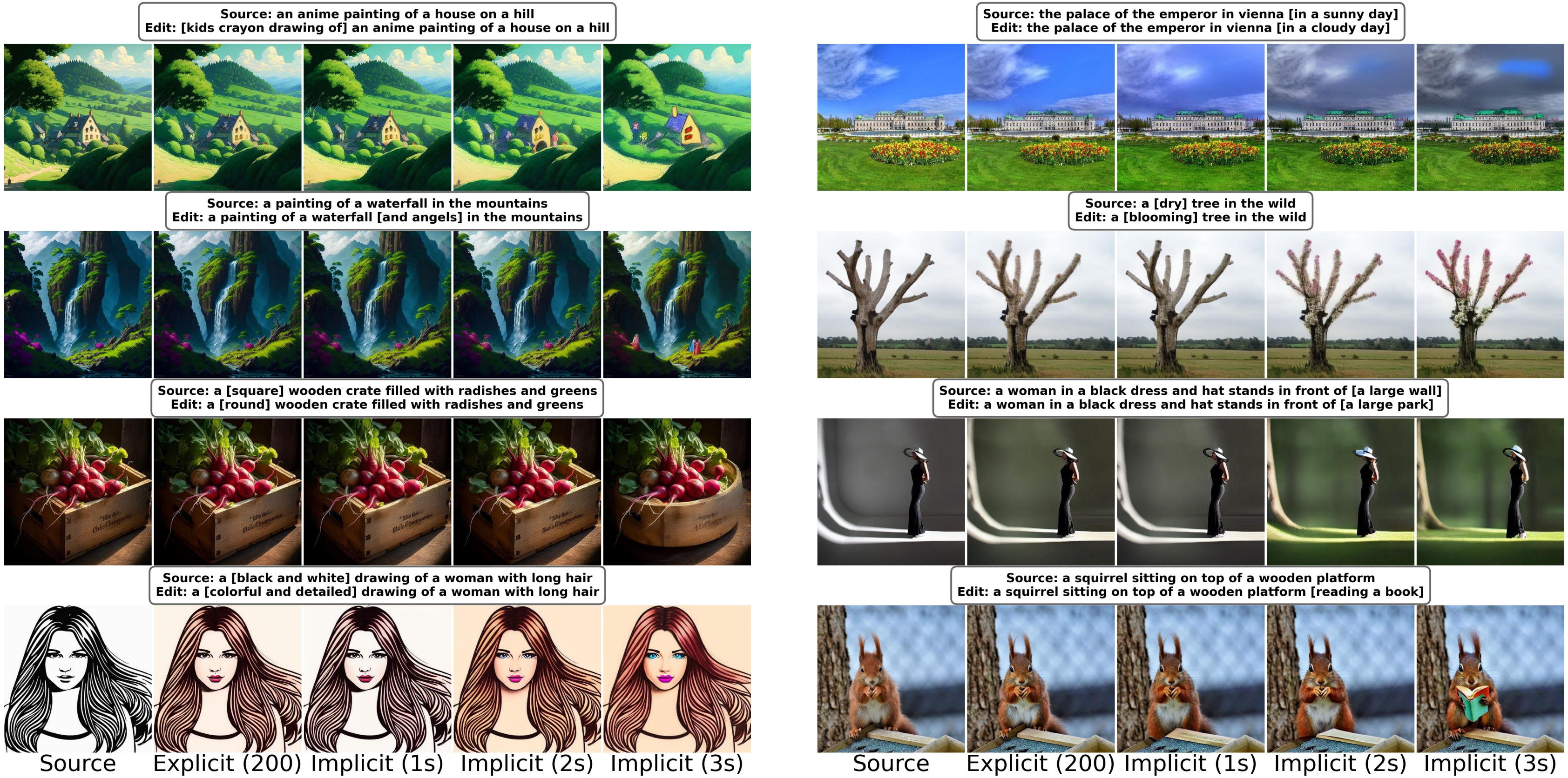
$$\begin{aligned} x_{t-1} &\approx x_{t-1}^{\text{init}} + \gamma \nabla_{x_{t-1}} \log p^h\left(x_{t-1}|x_t\right) + \sqrt{2\gamma} z \\ &= \left(x_{t-1}^{\text{init}} + \gamma \nabla_{x_{t-1}} \log p\left(x_{t-1}|x_t\right) + \sqrt{2\gamma} z\right) \\ &\quad + \gamma \nabla_{x_{t-1}} \log h\left(x_{t-1}, t-1\right) \\ &\approx \underbrace{x_{t-1}^{\text{base}}}_{\text{rec.}} + \gamma \underbrace{\nabla_{x_{t-1}} \log h\left(x_{t-1}^{\text{base}}, t-1\right)}_{\text{editing}} \end{aligned}$$

- What's special about implicit form? → **Optimization** over the space of x_{t-1} as the editing term acts on x_{t-1}

$$\begin{aligned} x_{t-1}^{(0)} &= x_{t-1}^{\text{base}} \\ x_{t-1}^{(k+1)} &= x_{t-1}^{(k)} + \gamma \nabla_{x_{t-1}} \log h\left(x_{t-1}^{(k)}, t-1\right) \end{aligned}$$

Multiple optimization steps for hard editing cases

Results of Implicit h-Edit using multiple optimization steps



How *h*-Edit works? Design *h*-function effectively!

- Text-guided editing

We introduce $\hat{w}^{\text{orig}}$ at editing time (different to w^{orig} at inversion time)

$$
\begin{aligned}
&\nabla \log h\left(x_{t-1}, t-1\right) \\
&\quad = \nabla \log p\left(y \middle| x_{t-1}\right) \\
&\quad = \nabla \log p\left(x_{t-1} \middle| y\right) - \nabla \log p\left(x_{t-1}\right)
\end{aligned}
$$

$$\frac{-\boxed{\tilde{\epsilon}_\theta\left(x_{t-1}, t-1, c^{\text{edit}}\right)}}{\sigma_{t-1}} \qquad \frac{-\boxed{\tilde{\epsilon}_\theta\left(x_{t-1}, t-1, c^{\text{orig}}\right)}}{\sigma_{t-1}}$$

$$\boxed{\left(1-w^{\text{edit}}\right) \epsilon_\theta\left(x_{t-1}, t-1, \varnothing\right) + w^{\text{edit}} \epsilon_\theta\left(x_{t-1}, t-1, c^{\text{edit}}\right)} \qquad \boxed{\left(1-\hat{w}^{\text{orig}}\right) \epsilon_\theta\left(x_{t-1}, t-1, \varnothing\right) + \hat{w}^{\text{orig}} \epsilon_\theta\left(x_{t-1}, t-1, c^{\text{orig}}\right)}$$

$\hat{w}^{\mathrm{orig}}$ plays the role of removing "negative" information!

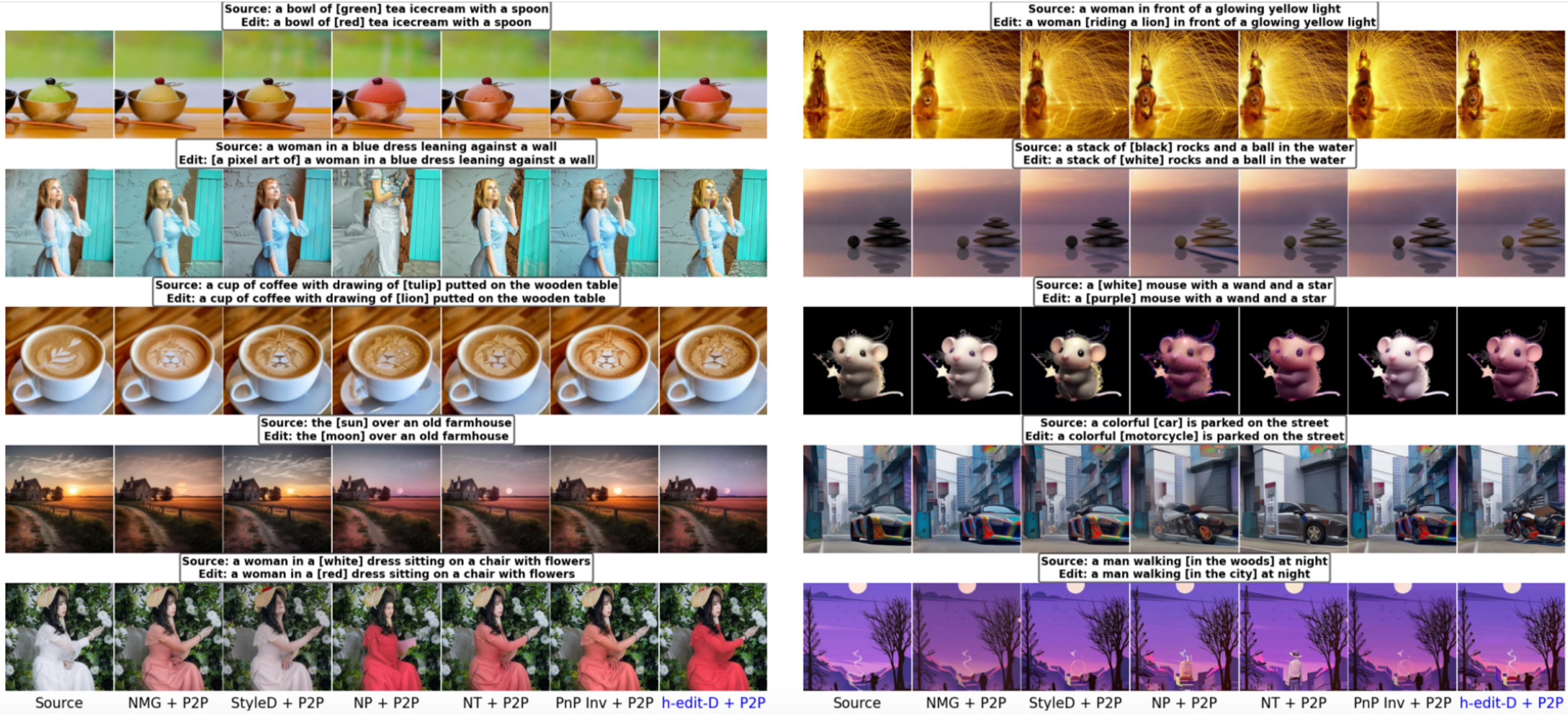
Quantitative Results on PIE-Bench

Inv.	Attn.	Method	CLIP Sim.↑	Local CLIP↑	DINO Dist.$_{\times 10^2}$↓	LPIPS$_{\times 10^2}$↓	SSIM$_{\times 10}$↑	PSNR↑
Deter.	P2P	NP	0.246	0.140	1.62	6.90	8.34	26.21
		NT	0.248	0.130	1.34	6.07	8.41	27.03
		StyleD	0.248	0.085	**1.17**	6.61	8.34	26.05
		NMG	0.249	0.087	1.32	5.59	8.47	27.05
		PnP Inv	0.250	0.095	**1.17**	5.46	8.48	27.22
		h-Edit-D	**0.253**	**0.147**	**1.17**	**4.85**	**8.54**	**27.87**
Random	None	EF	0.254	0.122	1.29	6.09	8.37	25.87
		LEDITS++	0.254	0.113	2.34	8.88	8.11	23.36
		h-Edit-R	**0.255**	**0.148**	**1.28**	**5.55**	**8.46**	**26.43**
	P2P	EF	0.255	0.126	1.51	5.70	8.40	26.30
		h-Edit-R	**0.256**	**0.159**	**1.45**	**5.08**	**8.50**	**26.97**

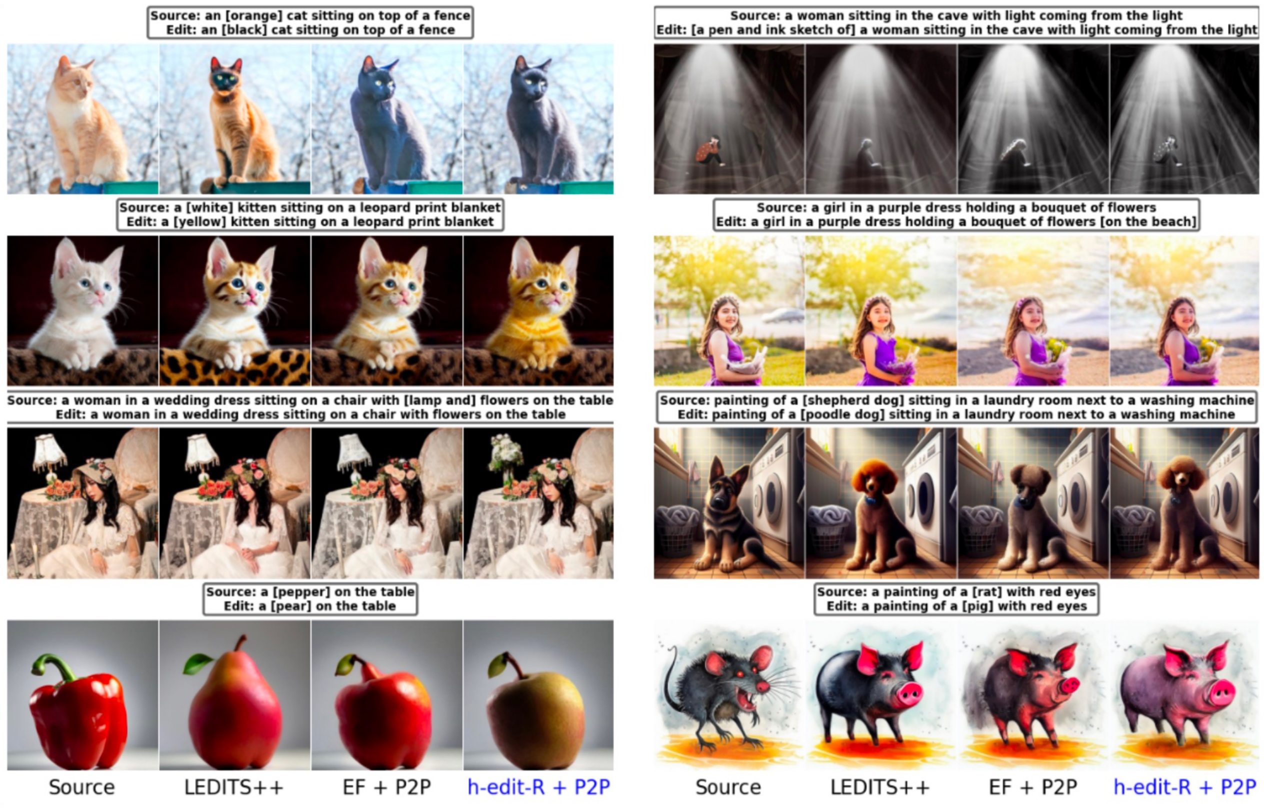
Better editing fidelity!

Better faithfulness!

Comparison with baselines: Deterministic Inversion



Comparison with baselines: Random Inversion



Editing with external reward models

$$\nabla \log h\left(x_{t-1}, t-1\right) = \nabla \log \left[\sum_{x_0} p\left(x_0 \mid x_{t-1}\right) h\left(x_0, 0\right)\right]$$
$$= \nabla \log \left[\mathbb{E}_{p(x_0 \mid x_{t-1})}\left[h\left(x_0, 0\right)\right]\right]$$
$$\approx \nabla \log \left[h\left(x_{0|t-1}, 0\right)\right]$$

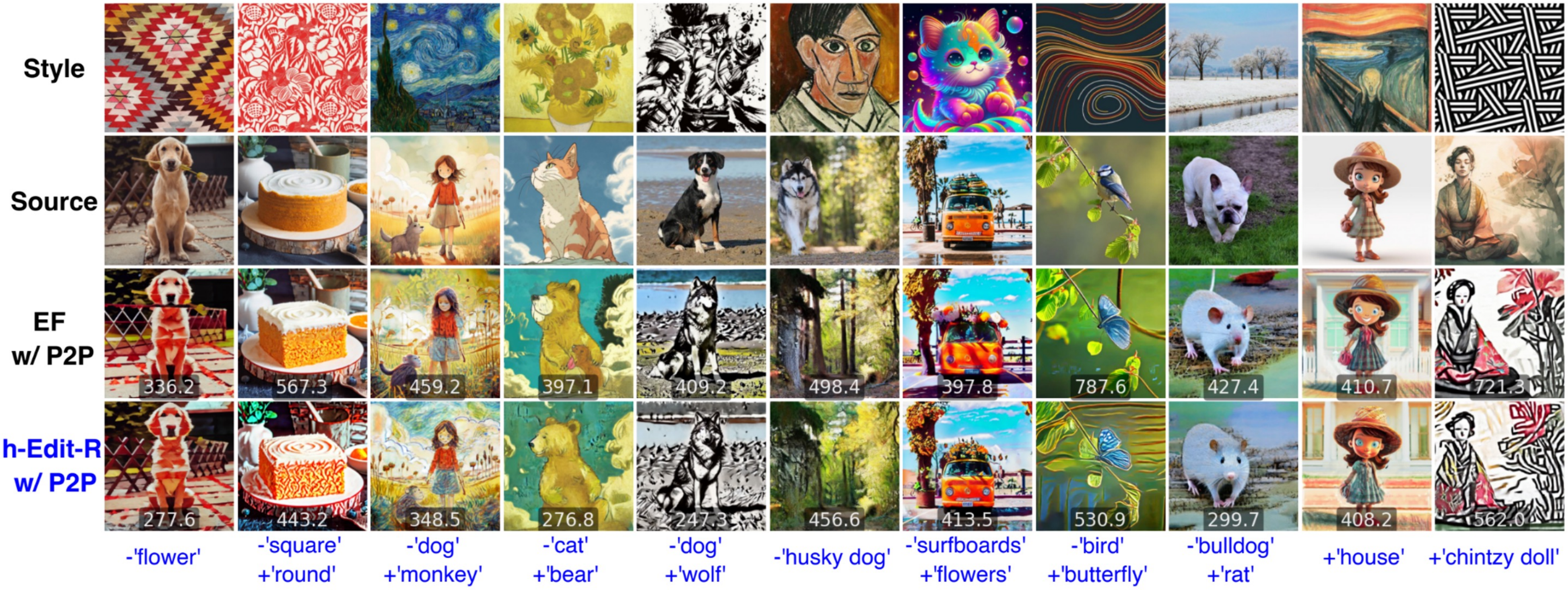
$x_{0|t-1}$ can be computed with **Tweedie's formula**

We can perform editing with gradients from external reward models trained on x_0

Editing with reward models: Face Swapping with ArcFace

Editing with conditional score and reward models:

Text-guided and Style Editing with CLIP



THANK YOU FOR LISTENING

Our paper: https://arxiv.org/abs/2503.02187
Source code: https://github.com/nktoan/h-edit