Eta Inversion: Designing an Optimal Eta Function for Diffusion-based Real Image Editing

Wonjun Kang*, Kevin Galim*, Hyung Il Koo

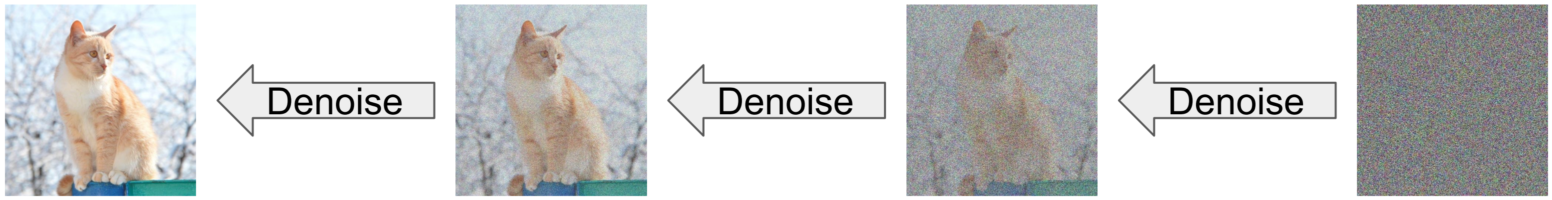
Code: github.com/furiosa-ai/eta-inversion

* indicates equal contribution

Diffusion

1. Start with random Gaussian noise
2. Repeat for n timesteps:
    - Predict the noise using the diffusion model with the input prompt
    - Remove the predicted noise from the image

“an orange cat sitting on top of a fence”

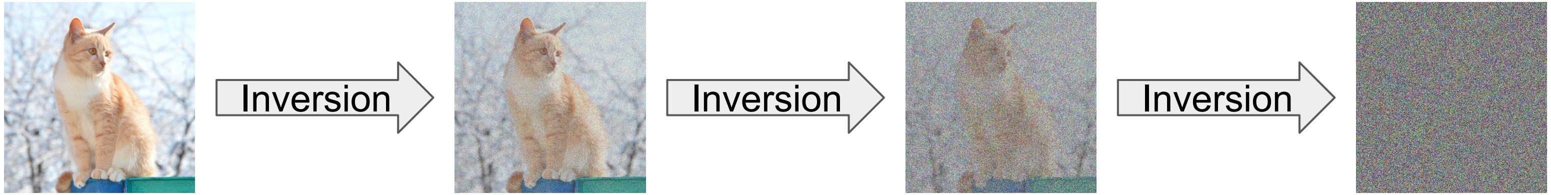


Prompt-based Real Image Editing with Diffusion

- Input
    - Source image
    - Source prompt, describing the source image
    - Target prompt, describing the desired output image
- Steps
    1. Invert the diffusion process for the source image and prompt to retrieve latent
    2. Denoise inverse latent with source prompt
    3. Denoise inverse latent with target prompt, forwarding information from the source denoising process to the target

Prompt-based Real Image Editing with Diffusion

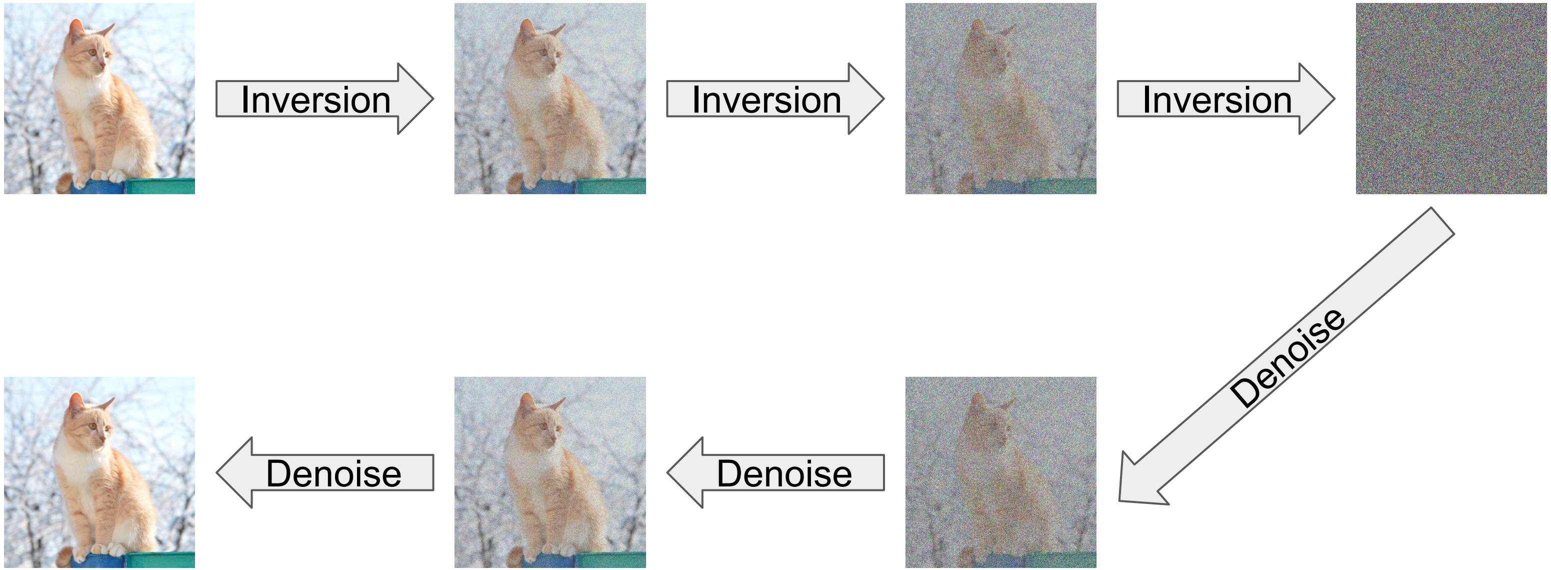
Source image
Invert with
source prompt



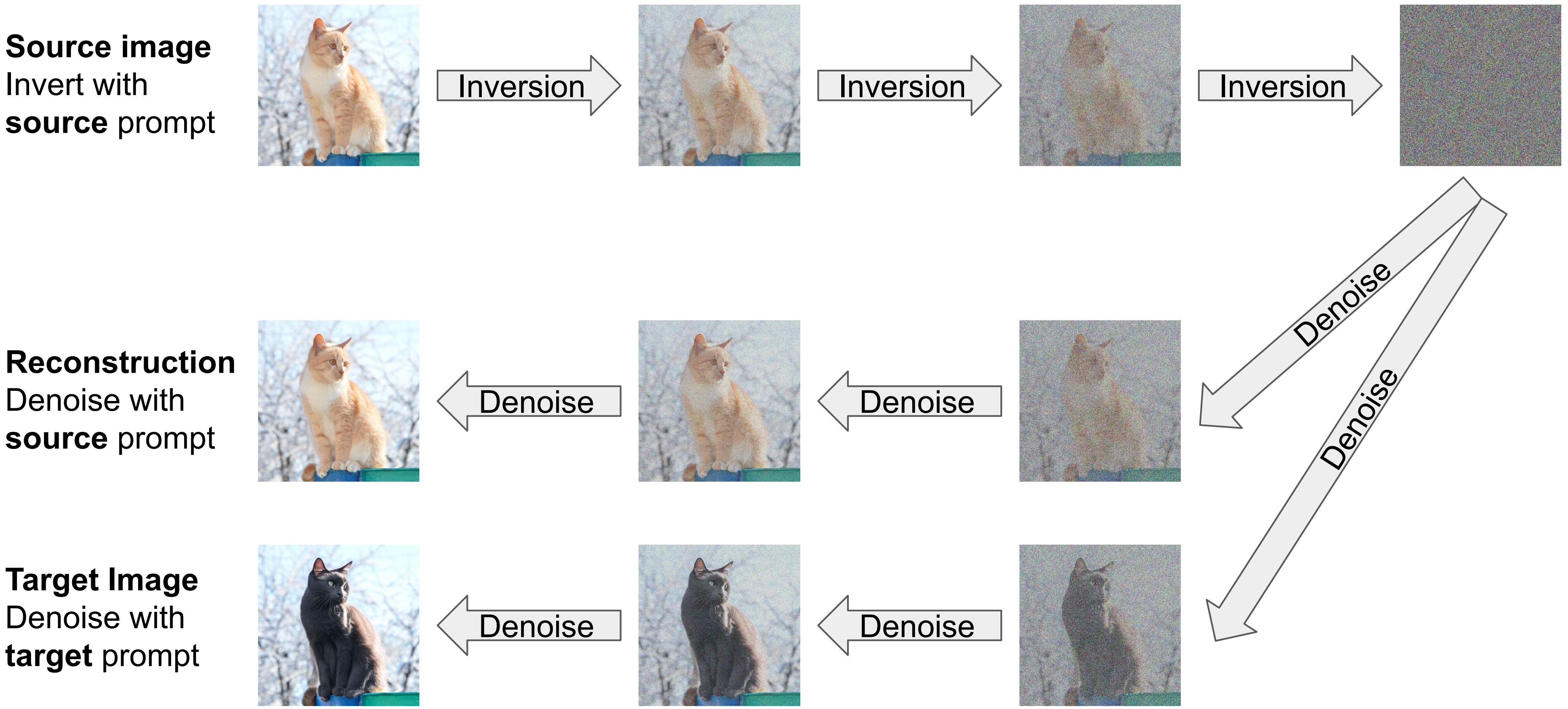
Prompt-based Real Image Editing with Diffusion

Source image
Invert with
source prompt

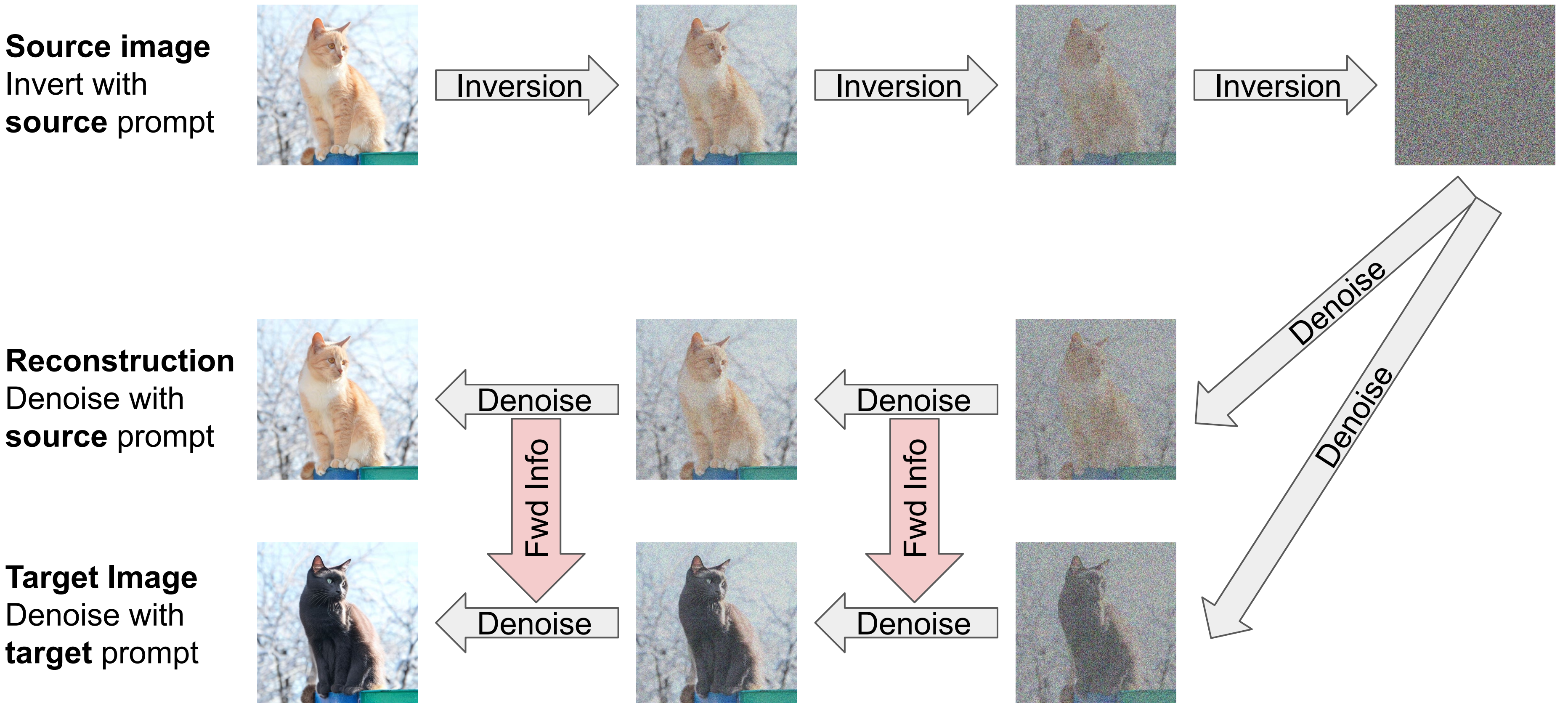
Reconstruction
Denoise with
source prompt



Prompt-based Real Image Editing with Diffusion



Prompt-based Real Image Editing with Diffusion



Preliminaries: DDIM Sampling

- Samplers, such as DDIM, remove a certain amount of predicted noise from the image
- DDIM is a common deterministic sampling technique for diffusion

$$\boldsymbol{x}_{t-1} = \sqrt{1/\alpha_t}(\boldsymbol{x}_t - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_t) + \sqrt{1-\bar{\alpha}_{t-1}-\sigma_t^2}\boldsymbol{\epsilon}_t + \sigma_t\boldsymbol{\epsilon}_{\text{add}}$$

Preliminaries: DDIM Sampling

(fixed) noise schedule

random Gaussian noise

$$\boxed{\boldsymbol{x}_{t-1}} = \sqrt{1/\boxed{\alpha_t}}\left(\boxed{\boldsymbol{x}_t} - \sqrt{1-\boxed{\bar{\alpha}_t}}\boxed{\boldsymbol{\epsilon}_t}\right) + \sqrt{1-\boxed{\bar{\alpha}_{t-1}}-\boxed{\sigma_t^2}}\boxed{\boldsymbol{\epsilon}_t} + \boxed{\sigma_t}\boxed{\boldsymbol{\epsilon}_{\text{add}}}$$

noisy image at timestep t

predicted noise (model)

denoised image at
timestep t-1

Controls random noise
injection amount
Set to 0 by default
(deterministic)

Preliminaries: DDIM Sampling (η)

- η ≥ 0 controls the amount of random noise being injected by DDIM
- DDIM uses η=0 by default (deterministic)

$$\boldsymbol{x}_{t-1} = \sqrt{1/\alpha_t}(\boldsymbol{x}_t - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_t) + \sqrt{1-\bar{\alpha}_{t-1} - \boxed{\sigma_t^2}}\boldsymbol{\epsilon}_t + \boxed{\sigma_t}\boldsymbol{\epsilon}_{\text{add}}$$

$$\boxed{\sigma_t} = \boxed{\eta_t}\sqrt{(1-\bar{\alpha}_{t-1})/(1-\bar{\alpha}_t)}\sqrt{1-\bar{\alpha}_t/\bar{\alpha}_{t-1}}$$

DDIM Inversion

- Foundation for most diffusion inversion methods
- Predicted noise of previous denoising step $\boldsymbol{\epsilon}_{t+1}$ unknown

→ Approximate as $\boldsymbol{\epsilon}_{t+1} \approx \boldsymbol{\epsilon}_t$

DDIM sampling

$$\begin{aligned}\boldsymbol{x}_{t-1} &= \sqrt{1/\alpha_t}(\boldsymbol{x}_t - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_t) + \sqrt{1-\bar{\alpha}_{t-1}}\boldsymbol{\epsilon}_t \\ &= \sqrt{\bar{\alpha}_{t-1}/\bar{\alpha}_t}\boldsymbol{x}_t + \sqrt{\bar{\alpha}_{t-1}}(\sqrt{1/\bar{\alpha}_{t-1}-1} - \sqrt{1/\bar{\alpha}_t - 1})\boldsymbol{\epsilon}_t\end{aligned}$$

$$\Downarrow$$

DDIM inversion

$$\boldsymbol{x}_{t+1} = \sqrt{\bar{\alpha}_{t+1}/\bar{\alpha}_t}\boldsymbol{x}_t + \sqrt{\bar{\alpha}_{t+1}}(\sqrt{1/\bar{\alpha}_{t+1}-1} - \sqrt{1/\bar{\alpha}_t - 1})\boldsymbol{\epsilon}_t$$

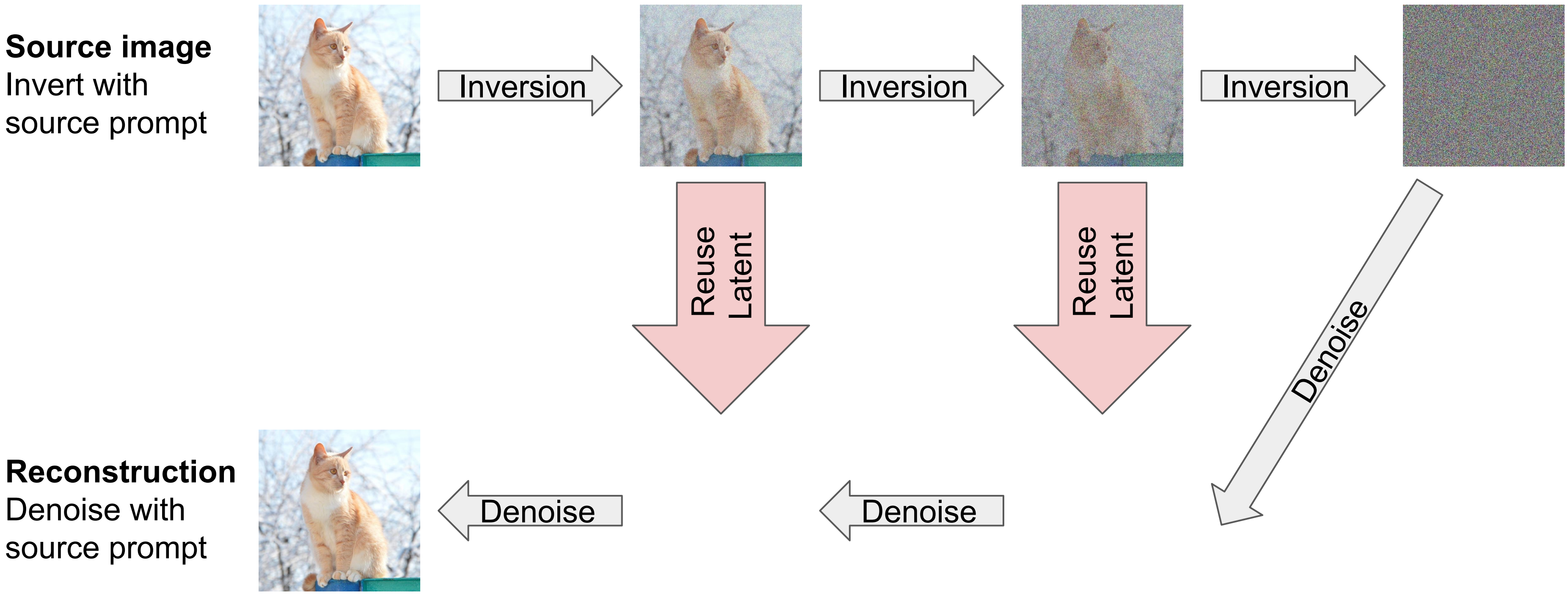
DDIM Inversion

- Approximation ($\boldsymbol{\epsilon}_{t+1} \approx \boldsymbol{\epsilon}_t$) leads to reconstruction errors
- Reconstruction error is non-neglectable (when using classifier-free guidance)
- Goal of diffusion inversion methods: Reduce reconstruction error

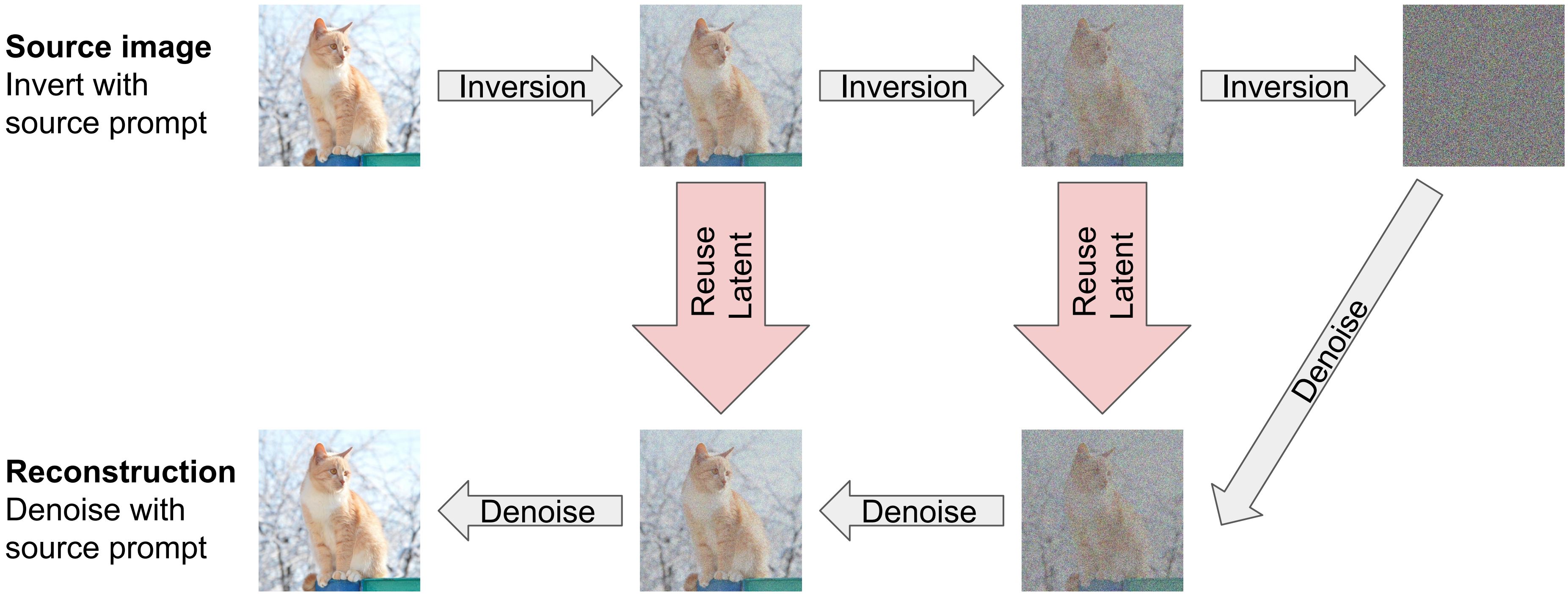
Perfect Inversion Methods

- Store all intermediate latents during the inversion process
- Replace latent with stored latent in the denoising process
  - → Compensates for the error between the inversion and reconstruction branches
  - → Ensures perfect reconstruction
- Employed by existing methods such as CycleDiffusion, DDPM Inversion, and Direct Inversion
- Also employed by Eta Inversion

Perfect Inversion Methods



Perfect Inversion Methods



Issues with Existing Perfect Inversion Methods

- Issues of latent replacement/compensation
  - Compensation is not normally distributed
  - May limit editability
- Consequences
  - Edited image is too close to the source image and does not follow the target prompt

“a woman in a **jacket** standing in the rain”

Source

“a woman in a **blouse** standing in the rain”

Direct Inversion

Eta Inversion - Motivation

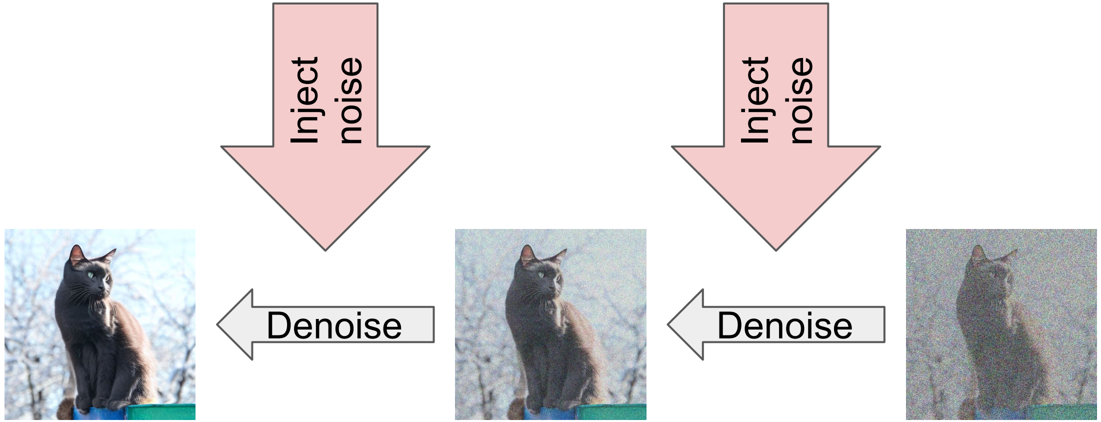
- Previous methods employ a fixed η value of either 0 or 1
- However, based on theoretical analysis
  - Score networks (diffusion models) are imperfect in practice
  - If score networks are imperfect, η impacts the marginal distribution
  - Thus, it could be beneficial to optimize η for practical scenarios
- Our question: Can we improve editability by using a dynamic η function?
- Answer
  - Yes!

$$\boldsymbol{x}_{t-1} = \sqrt{1/\alpha_t}(\boldsymbol{x}_t - \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_t) + \sqrt{1-\bar{\alpha}_{t-1} - \sigma(\eta_t)^2}\boldsymbol{\epsilon}_t + \sigma(\eta_t)\boldsymbol{\epsilon}_{\text{add}}$$

Improving Editability with Eta Inversion

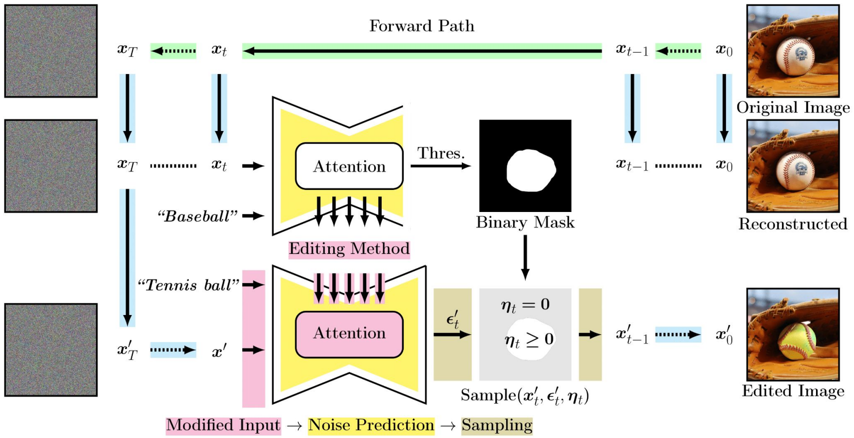
- Use time-dependent η to control noise intensity per timestep
- Maintain the similarity of the background by injecting noise only into the object to be edited

Target Image
Denoise with
target prompt



Our Method - Eta Inversion

- Employ a time- and region-dependent η_t based on theoretical analysis
  - Use larger η_t at earlier denoising steps to increase the editing effect
  - Use smaller η_t at later denoising steps to maintain structural details
  - Only injects (real Gaussian) noise in the foreground part of the image to avoid editing the background
  - Retrieve the foreground-background map from cross-attention maps



SOTA Quantitative Results (PIE-Bench)

Metric (×10²)	CLIP similarity ↑			CLIP accuracy ↑			DINO ↓			LPIPS ↓			BG-LPIPS ↓		
Method	PtP	PnP	Masa	PtP	PnP	Masa	PtP	PnP	Masa	PtP	PnP	Masa	PtP	PnP	Masa
DDIM Inv. [39]	30.99	29.38	30.74	94.57	85.57	95.00	6.94	6.11	7.55	46.65	40.84	47.68	24.97	20.84	25.37
Null-text Inv. [24]	30.73	30.75	30.07	92.57	90.43	93.00	1.24	3.27	4.49	15.13	30.51	25.02	5.69	14.17	11.92
NPI [23]	30.49	30.73	29.54	92.71	91.29	87.29	2.03	2.67	4.51	19.28	26.18	26.03	8.24	11.57	12.41
ProxNPI [12]	30.31	30.54	29.49	92.43	90.71	88.14	1.92	2.29	3.92	17.69	21.76	22.99	7.76	9.57	10.99
EDICT [42]	29.28	24.69	29.68	92.71	63.43	93.29	0.41	4.26	0.79	6.65	30.22	8.59	3.10	14.96	4.20
DDPM Inv. [16]	29.43	30.26	29.57	92.71	94.86	93.00	0.42	1.04	0.75	6.87	12.50	8.65	3.27	5.84	4.12
Direct Inv. [17]	30.92	31.32	30.37	94.71	95.14	94.57	1.28	2.27	4.32	15.79	25.59	26.91	6.33	12.98	13.76
Eta Inversion (1)	31.01	31.33	30.39	95.00	94.86	93.14	1.34	2.34	3.66	16.58	27.33	23.12	6.57	14.05	11.57
Eta Inversion (1) w/o mask	31.00	31.34	30.37	95.29	95.00	92.71	1.37	2.37	3.69	16.85	27.68	23.40	6.74	14.33	11.79
Eta Inversion (2)	31.25	31.63	30.62	95.43	95.29	93.86	1.70	3.40	5.24	21.14	36.59	33.07	8.00	18.72	16.64
Eta Inversion (2) w/o mask	31.27	31.62	30.62	95.43	95.86	94.14	1.85	3.58	5.46	22.77	38.43	34.81	9.03	20.19	18.03

SOTA Quantitative Results (PIE-Bench) - Style Transfer

Metric ($\times 10^2$)	CLIP similarity ↑			CLIP accuracy ↑			DINO ↓			LPIPS ↓		
Method	PtP	PnP	Masa	PtP	PnP	Masa	PtP	PnP	Masa	PtP	PnP	Masa
DDIM Inv. [39]	31.00	30.21	30.67	83.75	73.75	86.25	6.47	6.09	6.90	46.76	42.63	47.42
Null-text Inv. [24]	32.06	32.79	29.97	88.75	91.25	86.25	1.60	3.98	4.07	19.60	37.26	25.81
NPI [23]	31.44	32.37	29.60	92.50	90.00	75.00	2.22	3.30	4.04	22.18	32.26	27.37
ProxNPI [12]	30.88	31.66	29.38	86.25	85.00	80.00	2.02	2.52	3.36	19.18	25.47	22.68
EDICT [42]	29.45	25.32	29.93	91.25	58.75	90.00	0.41	4.22	0.73	6.68	31.11	8.57
DDPM Inv. [16]	29.78	30.64	29.78	90.00	90.00	90.00	0.43	0.97	0.66	6.99	12.53	8.50
Direct Inv. [17]	31.71	32.51	30.37	91.25	93.75	85.00	1.64	2.47	3.79	19.87	27.22	26.56
Eta Inversion (3)	32.85	33.12	30.82	90.00	86.25	86.25	4.19	5.16	6.69	47.76	52.66	46.17

Injecting Noise Generates Various Plausible Results

Qualitative Results


"... jacket ..." → "... blouse ..."
"... laughing ..." → "... angry ..."
"... collie dog ..." → "... garfield cat ..."
Source
DDIMInv
NTI
DirInv
EtaInv