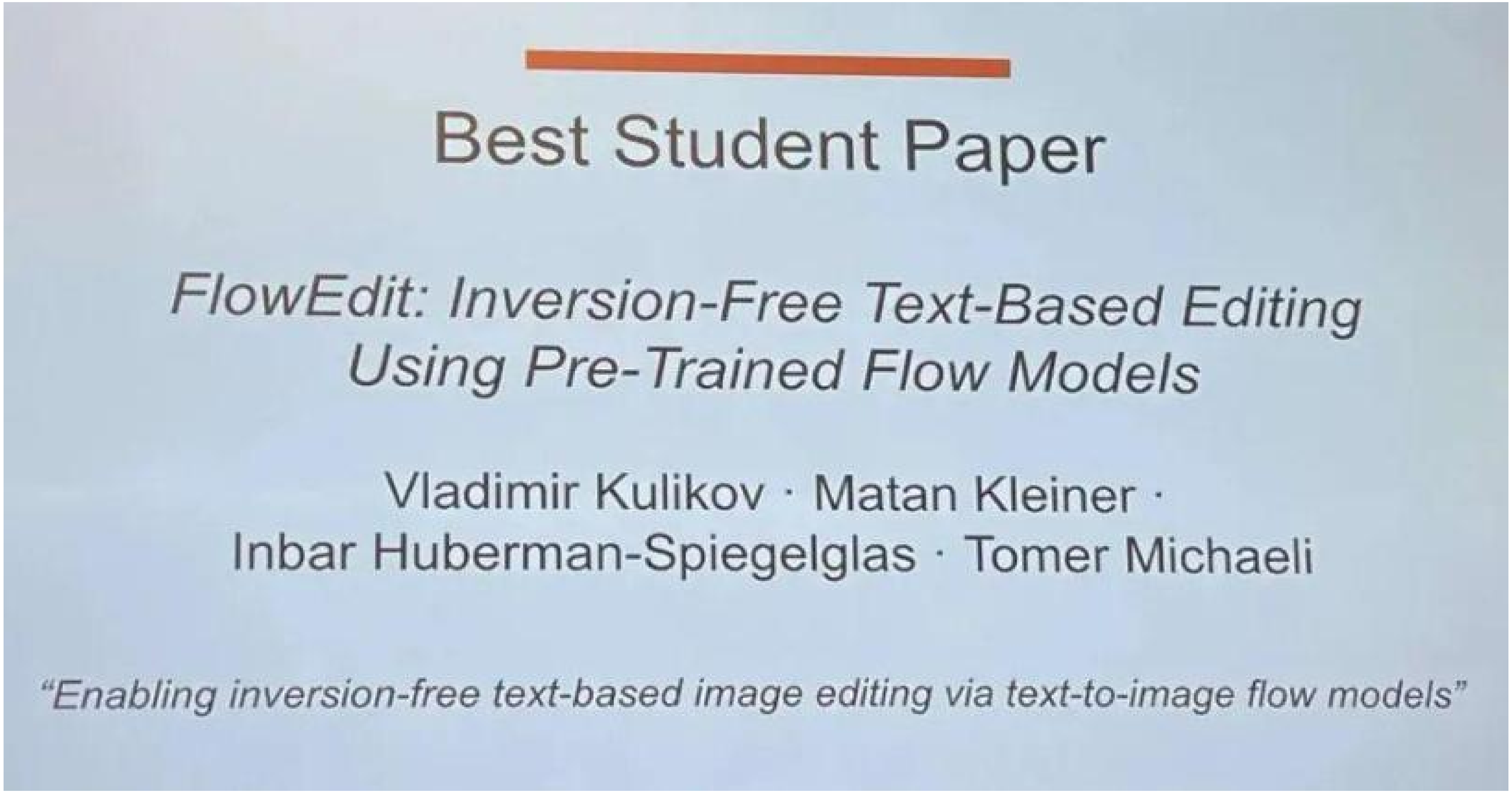
Seminar

2026.06.12
Lei Tu

ICCV 2025


Best Student Paper
FlowEdit: Inversion-Free Text-Based Editing
Using Pre-Trained Flow Models
Vladimir Kulikov · Matan Kleiner ·
Inbar Huberman-Spiegelglas · Tomer Michaeli
"Enabling inversion-free text-based image editing via text-to-image flow models"

Setting Clarification

Setting: image editing by a pre-trained text-to-image (T2I) model

Source Image	Source Prompt	Target Prompt
	A [dog] sitting on a wooden chair	A [cat] sitting on a wooden chair

Input & Output :

1: **Inputs:** source image x_{src}; prompts c_{src}, c_{tar}

2: **Output:** edited image x_{tar}.

Data(> 250 pairs): DIV2K dataset, pexels, pixabay, pxhere.

source prompt, which was obtained from LLaVA-1.5 and manually refined.

Evaluation Metric:

Semantic alignment:

CLIP score

Consistency：

DINO

DreamSim

MSE(latent space)

LPIPS(Learned Perceptual Image Patch Similarity)

FID(Frechet Inception Distance)

KID(Kernel Inception Distance)

Notation & Motivation

Pre-trained text-to-image model :

time: $t \in [0,1]$

image state: xt/Zt

text condition: c

$$\frac{dx_t}{dt} = v(x_t, t, c).$$

$$dZ_t = V(Z_t, t)\, dt.$$

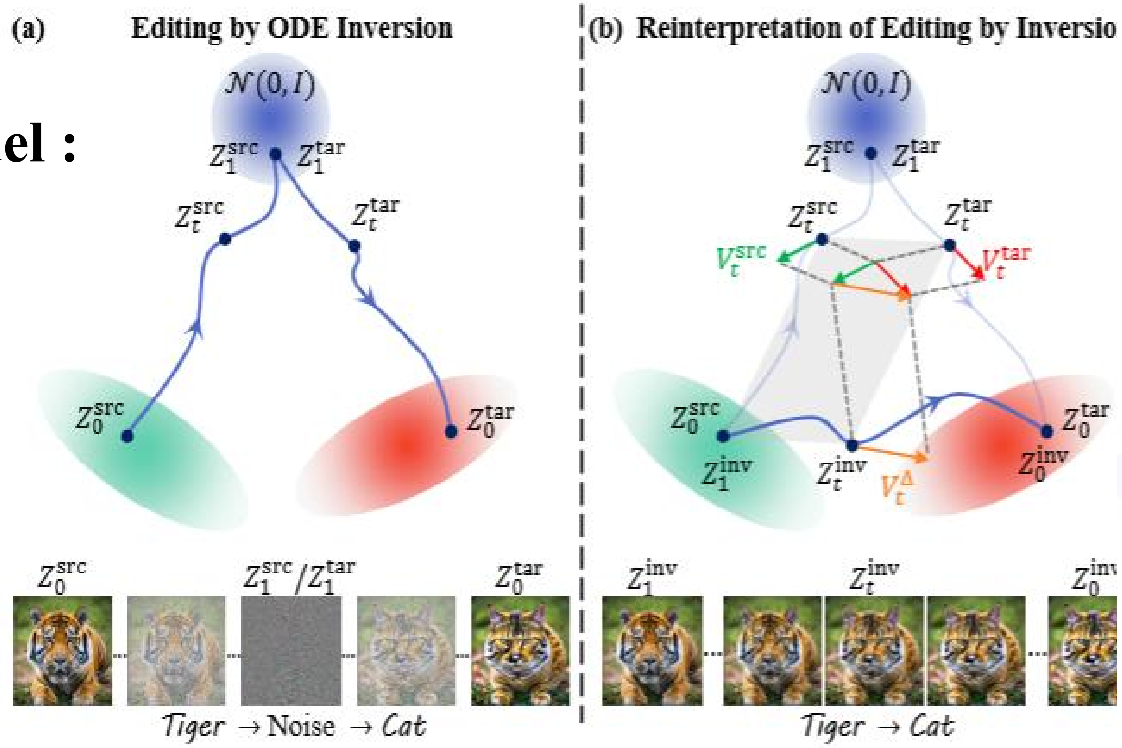
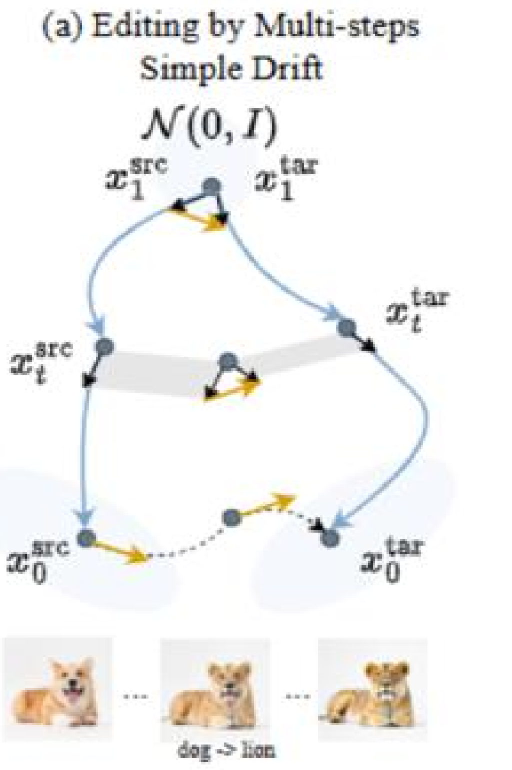


Image from ChordEdit



The Editing Problem:

$$V^{\mathrm{src}}(X_t, t) \triangleq V(X_t, t, c_{\mathrm{src}}) \text{ and } V^{\mathrm{tar}}(X_t, t) \triangleq V(X_t, t, c_{\mathrm{tar}})$$

$$V_t^{\Delta}(Z_t^{\mathrm{src}}, Z_t^{\mathrm{tar}}) = V^{\mathrm{tar}}(Z_t^{\mathrm{tar}}, t) - V^{\mathrm{src}}(Z_t^{\mathrm{src}}, t).$$

$$\boxed{Z_t^{\mathrm{inv}} = Z_0^{\mathrm{src}} + Z_t^{\mathrm{tar}} - Z_t^{\mathrm{src}}.} \qquad \boxed{dZ_t^{\mathrm{inv}} = V_t^{\Delta}(Z_t^{\mathrm{src}}, Z_t^{\mathrm{tar}})\, dt,}$$

$$dZ_t^{\mathrm{inv}} = V_t^{\Delta}(Z_t^{\mathrm{src}}, Z_t^{\mathrm{inv}} + Z_t^{\mathrm{src}} - Z_0^{\mathrm{src}})\, dt$$

Method

The Editing Problem:

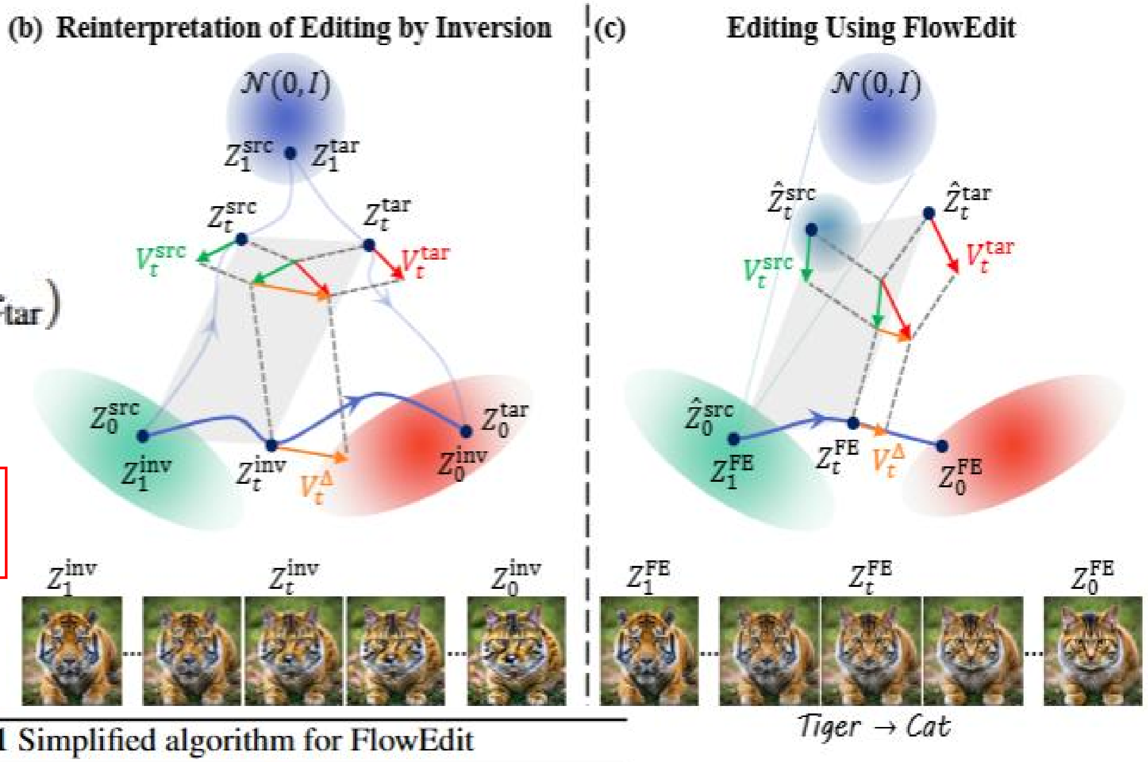
$$dZ_t = V(Z_t, t)\,dt.$$

$$V^{\text{src}}(X_t, t) \triangleq V(X_t, t, c_{\text{src}}) \text{ and } V^{\text{tar}}(X_t, t) \triangleq V(X_t, t, c_{\text{tar}})$$

$$V_t^{\Delta}(Z_t^{\text{src}}, Z_t^{\text{tar}}) = V^{\text{tar}}(Z_t^{\text{tar}}, t) - V^{\text{src}}(Z_t^{\text{src}}, t).$$

$$\boxed{Z_t^{\text{inv}} = Z_0^{\text{src}} + Z_t^{\text{tar}} - Z_t^{\text{src}}.} \qquad \boxed{dZ_t^{\text{inv}} = V_t^{\Delta}(Z_t^{\text{src}}, Z_t^{\text{tar}})\,dt,}$$

$$dZ_t^{\text{inv}} = V_t^{\Delta}(Z_t^{\text{src}}, Z_t^{\text{inv}} + Z_t^{\text{src}} - Z_0^{\text{src}})\,dt$$



Algorithm 1 Simplified algorithm for FlowEdit

Input: real image X^{src}, $\{t_i\}_{i=0}^{T}$, $n_{\max}$, n_{avg}
Output: edited image X^{tar}
Init: $Z_{t_{\max}}^{\text{FE}} = X_0^{\text{src}}$
for $i = n_{\max}$ **to** 1 **do**

$$\left.\begin{aligned} &N_{t_i} \sim \mathcal{N}(0, 1)\\ &Z_{t_i}^{\text{src}} \leftarrow (1 - t_i)X^{\text{src}} + t_i N_{t_i}\\ &Z_{t_i}^{\text{tar}} \leftarrow Z_{t_i}^{\text{FE}} + Z_{t_i}^{\text{src}} - X^{\text{src}}\\ &V_{t_i}^{\Delta} \leftarrow V^{\text{tar}}(Z_{t_i}^{\text{tar}}, t_i) - V^{\text{src}}(Z_{t_i}^{\text{src}}, t_i) \end{aligned}\right\} \text{Optionally average } n_{\text{avg}} \text{ samples}$$

$$Z_{t_{i-1}}^{\text{FE}} \leftarrow Z_{t_i}^{\text{FE}} + (t_{i-1} - t_i)V_{t_i}^{\Delta}$$

end for
Return: $Z_0^{\text{FE}} = X_0^{\text{tar}}$

FlowEdit:

$$\hat{Z}_t^{\text{src}} = (1 - t)Z_0^{\text{src}} + tN_t, \qquad N_t \sim \mathcal{N}(0, 1)$$

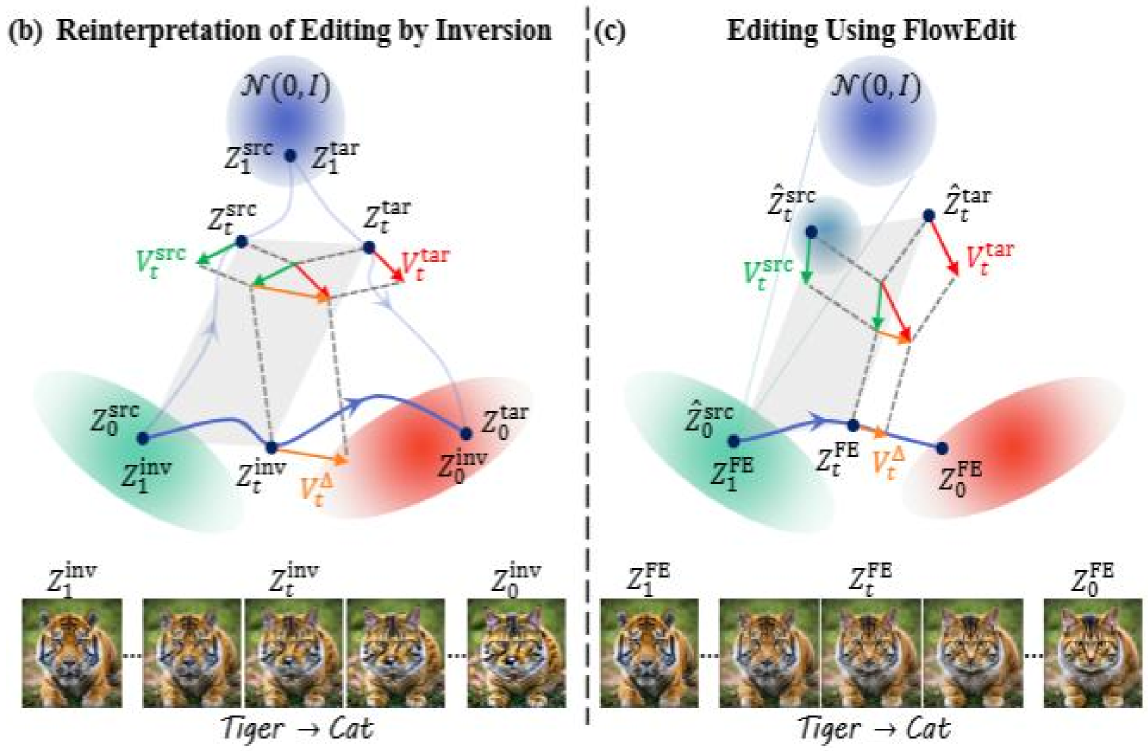
$$dZ_t^{\text{FE}} = \mathbb{E}\left[V_t^{\Delta}(\hat{Z}_t^{\text{src}}, Z_t^{\text{FE}} + \hat{Z}_t^{\text{src}} - Z_0^{\text{src}}) \middle| Z_0^{\text{src}}\right] dt$$

FlowEdit

Algorithm 1 Simplified algorithm for FlowEdit

Input: real image X^{src}, $\{t_i\}_{i=0}^{T}$, $n_{\max}$, n_{avg}
Output: edited image X^{tar}
Init: $Z_{t_{\max}}^{\text{FE}} = X_0^{\text{src}}$
for $i = n_{\max}$ **to** 1 **do**
  $N_{t_i} \sim \mathcal{N}(0, 1)$
  $Z_{t_i}^{\text{src}} \leftarrow (1 - t_i)X^{\text{src}} + t_i N_{t_i}$
  $Z_{t_i}^{\text{tar}} \leftarrow Z_{t_i}^{\text{FE}} + Z_{t_i}^{\text{src}} - X^{\text{src}}$
  $V_{t_i}^{\Delta} \leftarrow V^{\text{tar}}(Z_{t_i}^{\text{tar}}, t_i) - V^{\text{src}}(Z_{t_i}^{\text{src}}, t_i)$
  $Z_{t_{i-1}}^{\text{FE}} \leftarrow Z_{t_i}^{\text{FE}} + (t_{i-1} - t_i)V_{t_i}^{\Delta}$
end for
Return: $Z_0^{\text{FE}} = X_0^{\text{tar}}$

Optionally average n_{avg} samples



Practical Considerations:

$\mathbb{E}[N_t N_s] = 0$ for every $|t - s| > \delta$, where δ is the ODE discretization step, so that the noise becomes independent across timesteps.

$$dZ_t^{\text{FE}} = \mathbb{E}\left[V_t^{\Delta}(\hat{Z}_t^{\text{src}}, Z_t^{\text{FE}} + \hat{Z}_t^{\text{src}} - Z_0^{\text{src}}) \middle| Z_0^{\text{src}}\right] dt \longleftrightarrow \mathbb{E}[V^{\text{tar}}(\hat{Z}_t^{\text{tar}}, t)|Z_0^{\text{src}}] - \mathbb{E}[V^{\text{src}}(\hat{Z}_t^{\text{src}}, t)|Z_0^{\text{src}}],$$

When $n_{\max} < T$, the first $(T - n_{\max})$ timesteps are skipped

Cats-Dogs Experiment

“a photo of cat”

Llama3 8B instruction

“a photo of XX cat” x 1000

SD3

“a photo of XX dog” x 1000

SD3

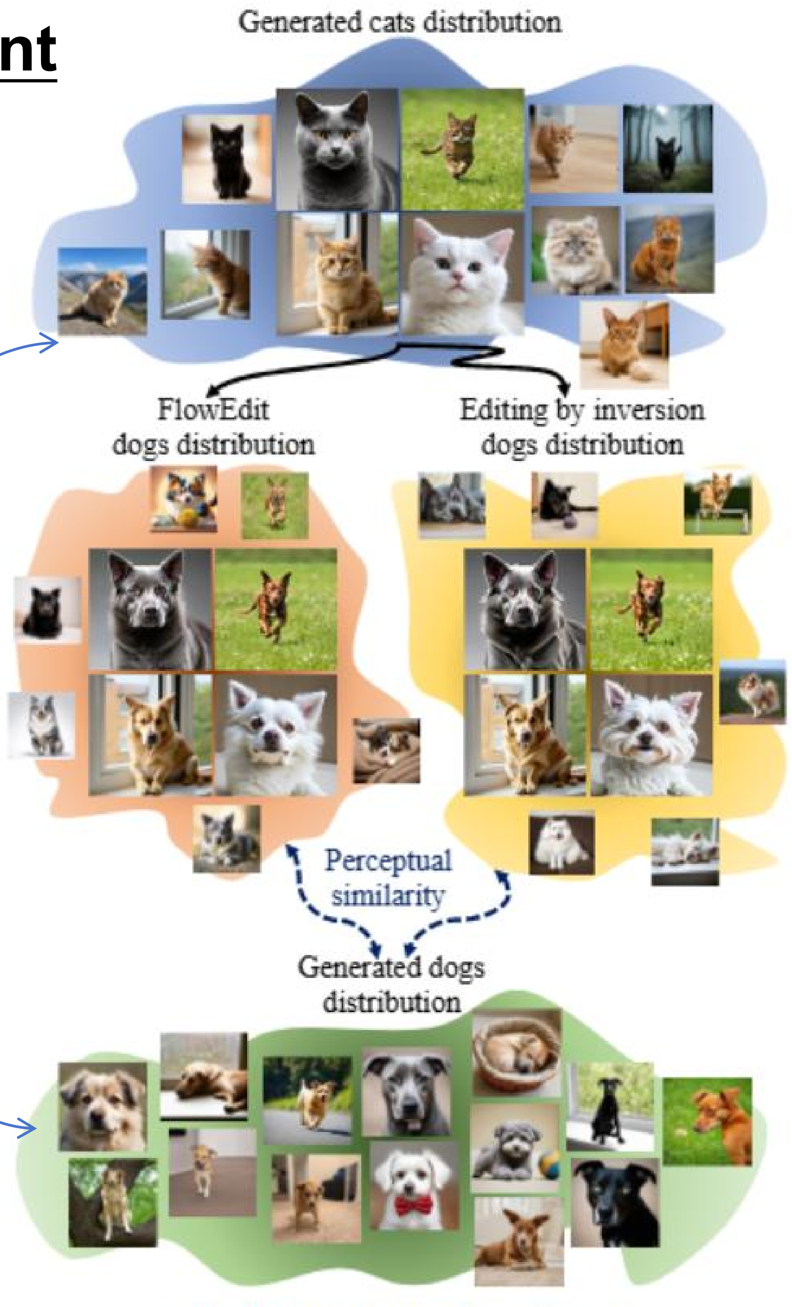


Figure S17. Illustration of the Cats-Dogs experiment.

FlowEdit achieves a lower transport cost:

	MSE(in latent space) ↓	LPIPS ↓
FlowEdit	1376	0.15
inversion	2239	0.25

FlowEdit is able to produce images from the target distribution(similarly to inversion):

	FID ↓	KID ↓
FlowEdit	51.14	0.017
inversion	55.88	0.023

Experiments

Table S5. **SD3 metrics.** The first, second and third best scores are highlighted for each metric. CLIP-T measures adherence to text, while the other four scores measure structure preservation.

	CLIP-T ↑	CLIP-I ↑	LPIPS ↓	DINO ↑	DreamSim ↓
SDEdit 0.2	0.33	0.885	0.251	0.634	0.213
SDEdit 0.4	0.34	0.854	0.316	0.564	0.273
ODE Inv.	0.337	0.813	0.318	0.549	0.326
iRFDS	0.335	0.822	0.376	0.534	0.327
FlowEdit	0.344	0.872	0.181	0.719	0.253

Table S6. **FLUX metrics.** The first, second and third best scores are highlighted for each metric. CLIP-T measures adherence to text, while the other four scores measure structure preservation.

	CLIP-T ↑	CLIP-I ↑	LPIPS ↓	DINO ↑	DreamSim ↓
SDEdit 0.5	0.316	0.902	0.264	0.637	0.18
SDEdit 0.75	0.331	0.862	0.348	0.557	0.26
ODE Inv.	0.341	0.822	0.374	0.505	0.328
RF Inv.	0.334	0.856	0.34	0.558	0.266
RF Edit 2	0.344	0.833	0.335	0.53	0.32
RF Edit 5	0.332	0.876	0.22	0.65	0.22
FlowEdit	0.337	0.875	0.223	0.682	0.252

Experiments

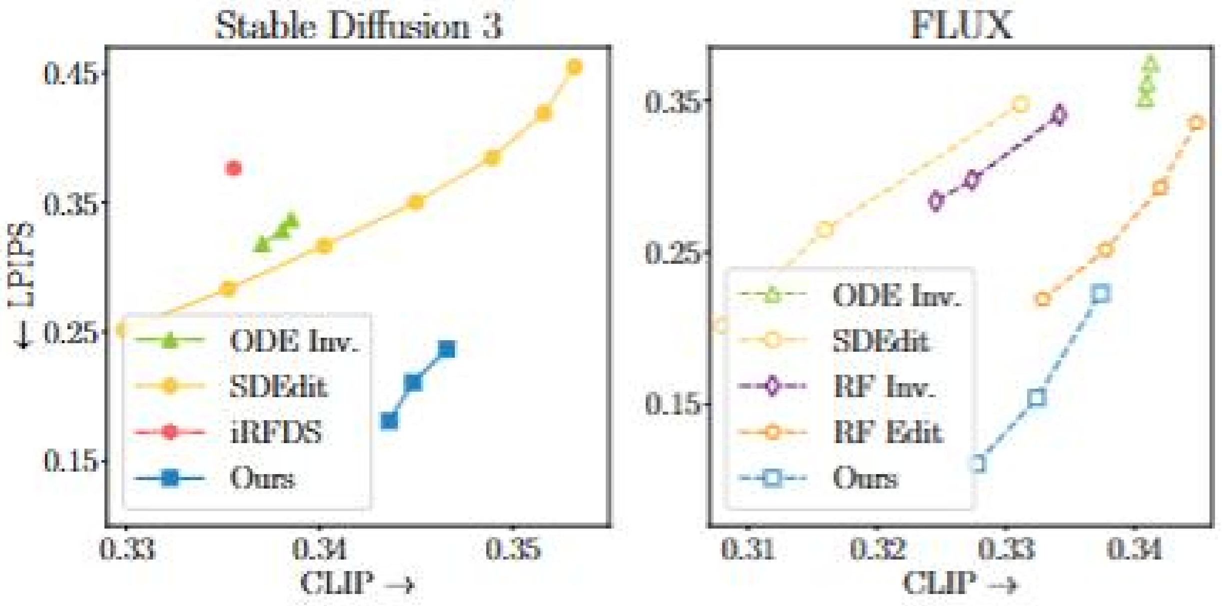
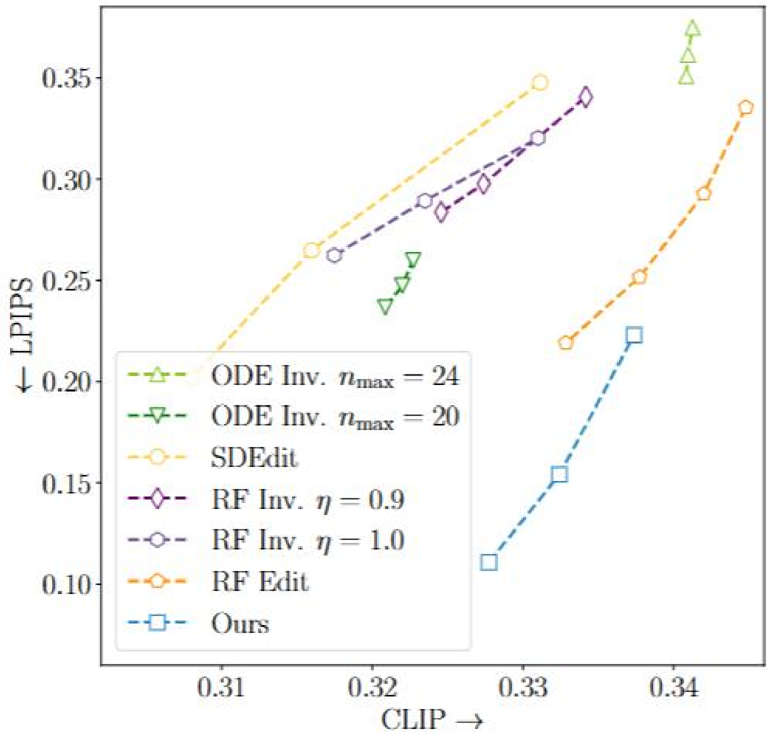


Figure 7. **Quantitative comparisons.** FlowEdit achieves a favorable balance between text adherence (CLIP) and structure preservation (LPIPS) compared to other methods. Connected markers represent different hyperparameters (see App. B).



Ablation:

Source Prompt：

Table S9. **Effect of source prompt on the results.** CLIP-T measures adherence to text, while the other four scores measure structure preservation.

	CLIP-T ↑	CLIP-I ↑	LPIPS ↓	DINO ↑	DreamSim ↓
original source prompt	0.344	0.872	0.181	0.719	0.253
different source prompt	0.343	0.872	0.181	0.713	0.254
w/o source prompt	0.343	0.872	0.192	0.709	0.254

Noise：

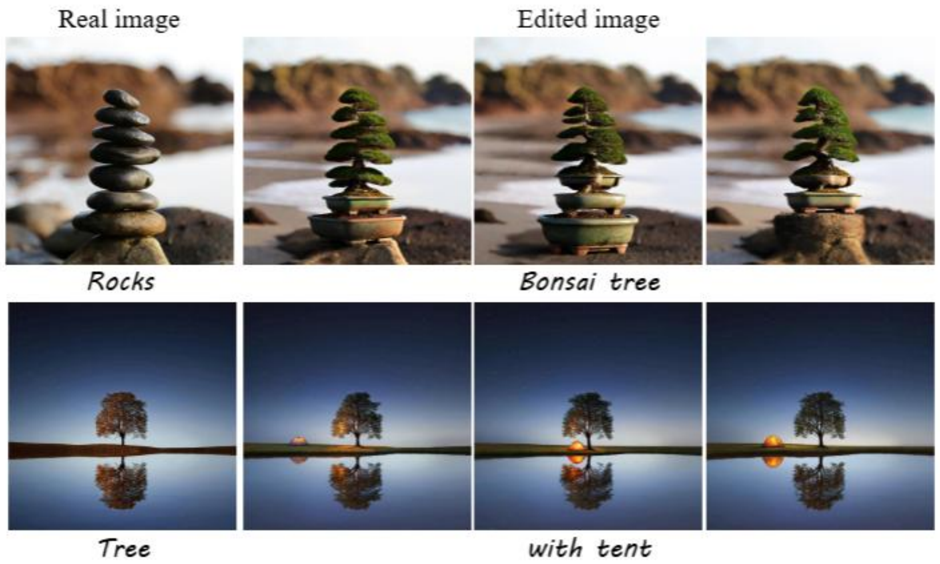


Figure S7. **FlowEdit diverse results due to different added noise.**

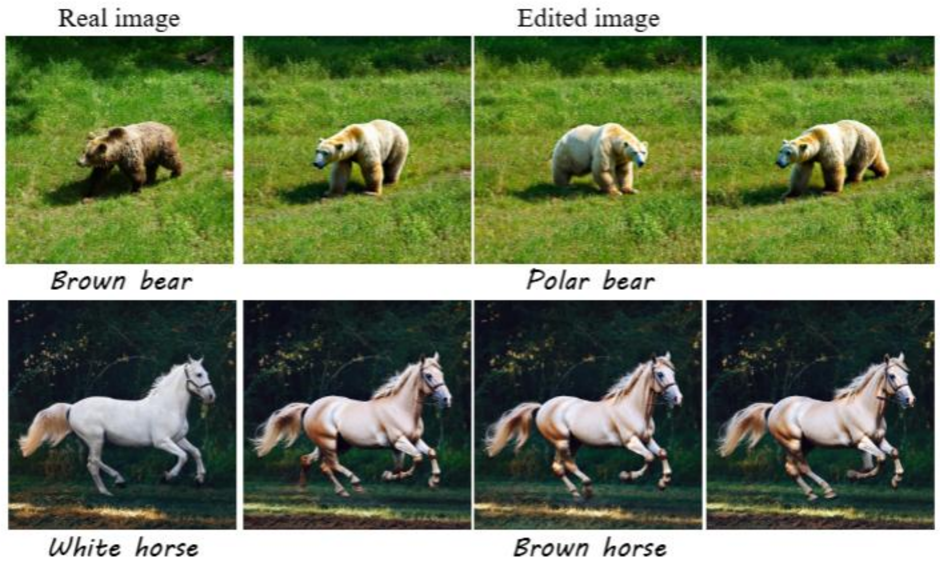
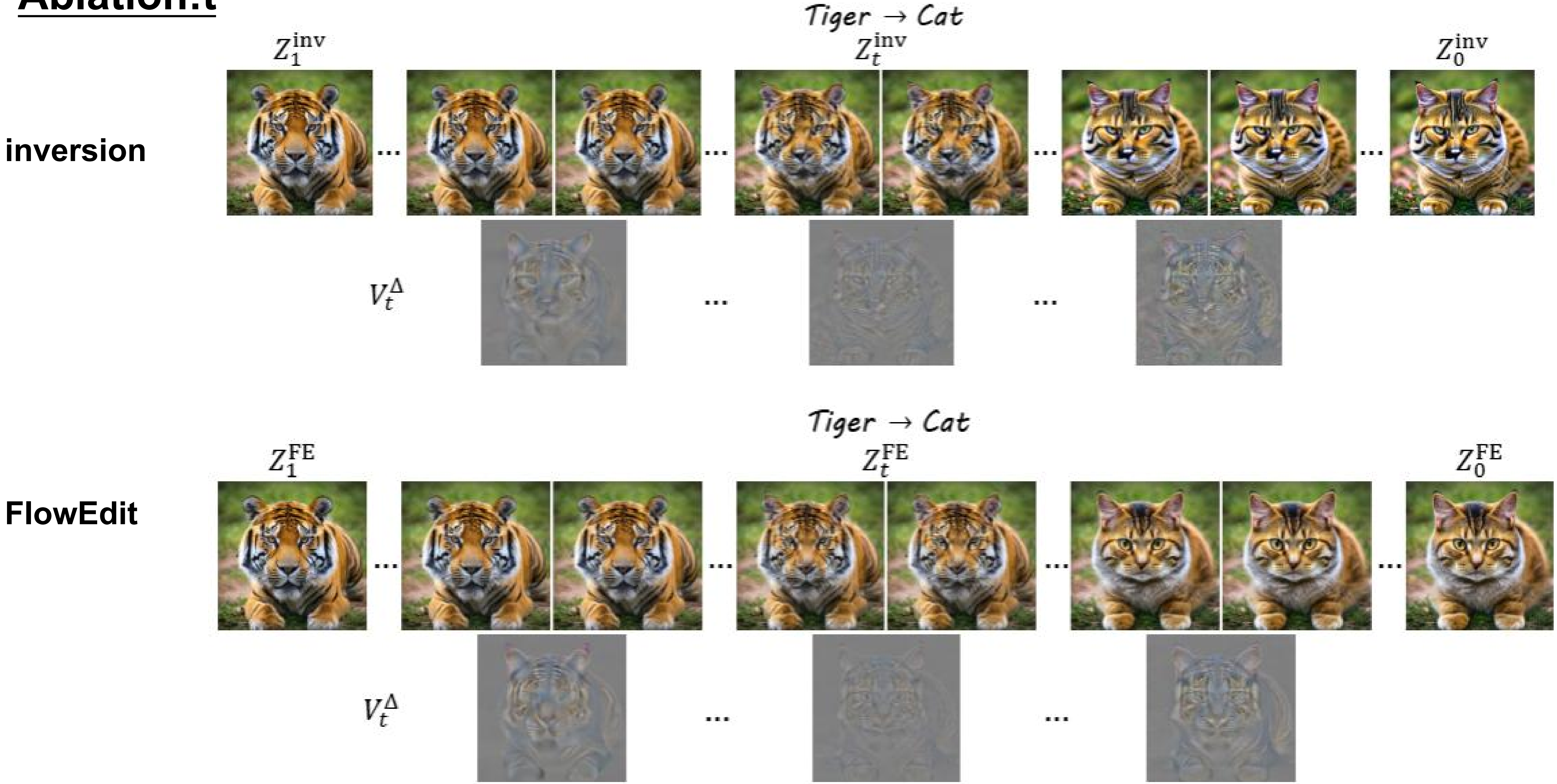
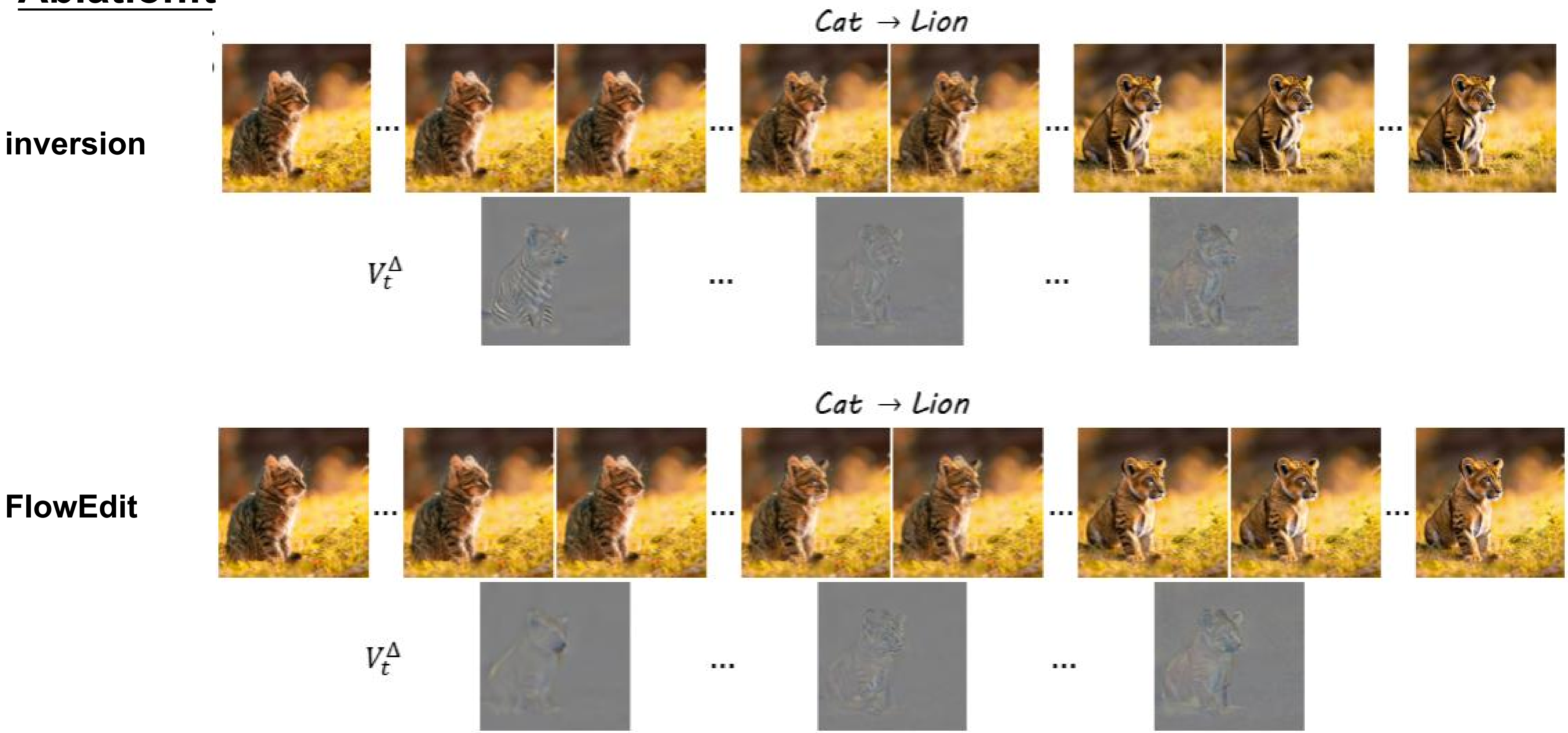


Figure S8. **Failure cases due to different added noise.**

Ablation:t


Tiger → Cat
Z_1^{inv}
Z_t^{inv}
Z_0^{inv}
inversion
V_t^{Δ}
Tiger → Cat
Z_1^{FE}
Z_t^{FE}
Z_0^{FE}
FlowEdit
V_t^{Δ}

Ablation:t



When $n_{\max} < T$, the first $(T - n_{\max})$ timesteps are skipped

Ablation: n_{max}

Algorithm 1 Simplified algorithm for Fl

Input: real image X^{src}, $\{t_i\}_{i=0}^{T}$, $n_{\max}$,
Output: edited image X^{tar}
Init: $Z_{t_{\max}}^{\text{FE}} = X_0^{\text{src}}$
for $i = n_{\max}$ **to** 1 **do**
$N_{t_i} \sim \mathcal{N}(0, 1)$
$Z_{t_i}^{\text{src}} \leftarrow (1 - t_i)X^{\text{src}} + t_i N_{t_i}$
$Z_{t_i}^{\text{tar}} \leftarrow Z_{t_i}^{\text{FE}} + Z_{t_i}^{\text{src}} - X^{\text{src}}$
$V_{t_i}^{\Delta} \leftarrow V^{\text{tar}}(Z_{t_i}^{\text{tar}}, t_i) - V^{\text{src}}(Z_{t_i}^{\text{src}}, t_i)$
$Z_{t_{i-1}}^{\text{FE}} \leftarrow Z_{t_i}^{\text{FE}} + (t_{i-1} - t_i)V_{t_i}^{\Delta}$
end for
Return: $Z_0^{\text{FE}} = X_0^{\text{tar}}$

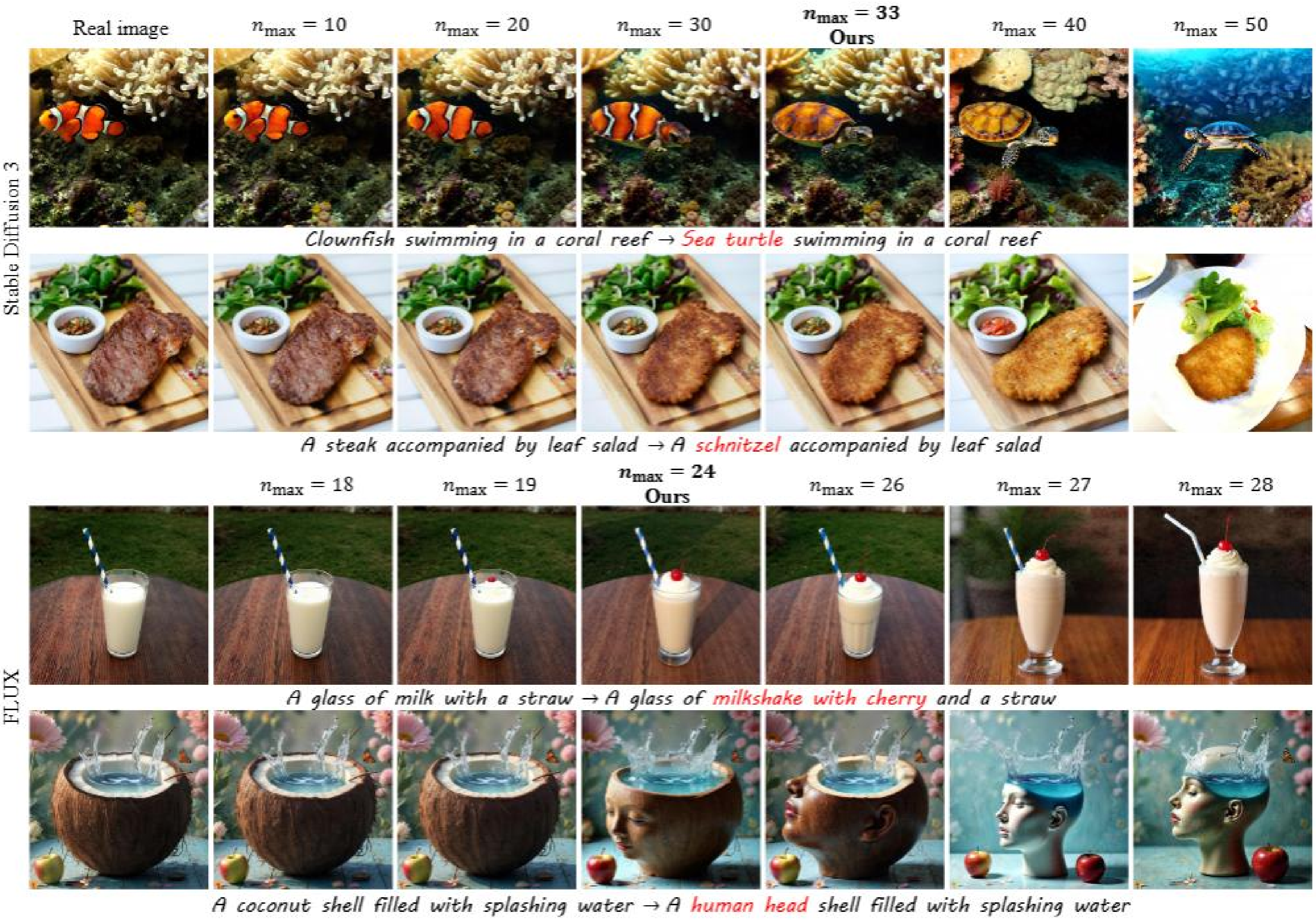


Figure S4. **Effect of $n_{\max}$.**

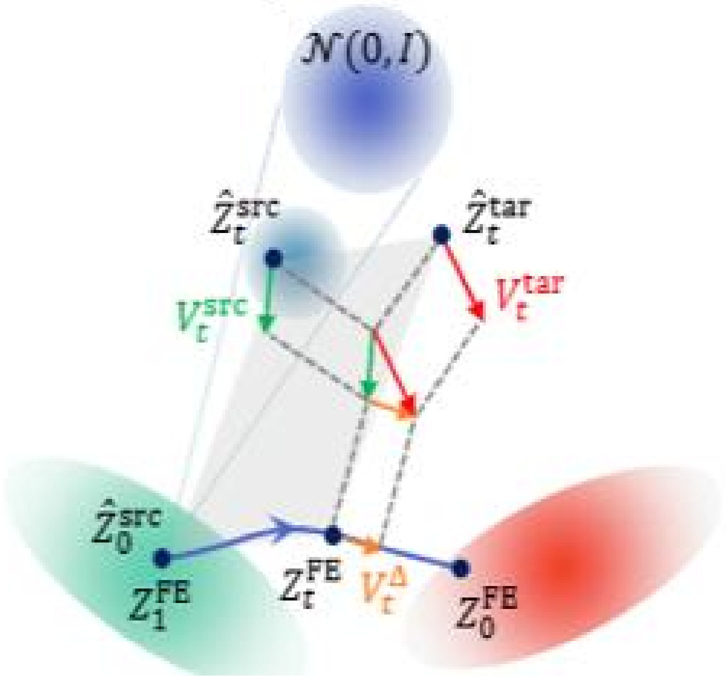
Ablation: n_{min}

A bear walking in a river in Pixar style

Algorithm S1 Full FlowEdit algorithm

Input: real image X^{src}, $\{t_i\}_{i=0}^{T}$, $n_{\max}$, $n_{\min}$, n_{avg}
Output: edited image X^{tar}
Init: $Z^{\text{FE}}_{t_{\max}} = X^{\text{src}}_0$
for $i = n_{\max}$ **to** $n_{\min+1}$ **do**
  $N_{t_i} \sim \mathcal{N}(0, 1)$
  $Z^{\text{src}}_{t_i} \leftarrow (1 - t_i)X^{\text{src}} + t_i N_{t_i}$
  $Z^{\text{tar}}_{t_i} \leftarrow Z^{\text{FE}}_{t_i} + Z^{\text{src}}_{t_i} - X^{\text{src}}$
  $V^{\Delta}_{t_i} \leftarrow V^{\text{tar}}(Z^{\text{tar}}_{t_i}, t_i) - V^{\text{src}}(Z^{\text{src}}_{t_i}, t_i)$
  Optionally average n_{avg} samples
  $Z^{\text{FE}}_{t_{i-1}} \leftarrow Z^{\text{FE}}_{t_i} + (t_{i-1} - t_i)V^{\Delta}_{t_i}$
end for
if $n_{\min} = 0$ **then**
  Return: $Z^{\text{FE}}_0 = X^{\text{tar}}_0$
else
  $N_{n_{\min}} \sim \mathcal{N}(0, 1)$
  $Z^{\text{src}}_{t_{n_{\min}}} \leftarrow (1 - t_{n_{\min}})X^{\text{src}} + t_{n_{\min}} N_{n_{\min}}$
  $Z^{\text{tar}}_{t_{n_{\min}}} \leftarrow Z^{\text{FE}}_{t_{n_{\min}}} + Z^{\text{src}}_{t_{n_{\min}}} - X^{\text{src}}$
  for $i = n_{\min}$ **to** 1 **do**
    $Z^{\text{tar}}_{t_{i-1}} \leftarrow Z^{\text{tar}}_{t_i} + (t_{i-1} - t_i)V^{\text{tar}}(Z^{\text{tar}}_{t_i}, t_i)$
  end for
  Return: $Z^{\text{tar}}_0 = X^{\text{tar}}_0$
end if

(c) Editing Using FlowEdit



Ablation: n_{avg}

Algorithm S1 Full FlowEdit algorithm

Input: real image X^{src}, $\{t_i\}_{i=0}^{T}$, $n_{\max}, n_{\min}, n_{\text{avg}}$
Output: edited image X^{tar}
Init: $Z^{\text{FE}}_{t_{\max}} = X^{\text{src}}_0$
for $i = n_{\max}$ **to** $n_{\min+1}$ **do**

$$\left.\begin{aligned}
&N_{t_i} \sim \mathcal{N}(0,1) \\
&Z^{\text{src}}_{t_i} \leftarrow (1-t_i)X^{\text{src}} + t_i N_{t_i} \\
&Z^{\text{tar}}_{t_i} \leftarrow Z^{\text{FE}}_{t_i} + Z^{\text{src}}_{t_i} - X^{\text{src}} \\
&V^{\Delta}_{t_i} \leftarrow V^{\text{tar}}(Z^{\text{tar}}_{t_i}, t_i) - V^{\text{src}}(Z^{\text{src}}_{t_i}, t_i)
\end{aligned}\right\}\ \text{Optionally average } n_{\text{avg}} \text{ samples}$$

$Z^{\text{FE}}_{t_{i-1}} \leftarrow Z^{\text{FE}}_{t_i} + (t_{i-1} - t_i)V^{\Delta}_{t_i}$

end for
if $n_{\min} = 0$ **then**
  Return: $Z^{\text{FE}}_0 = X^{\text{tar}}_0$
else
  $N_{n_{\min}} \sim \mathcal{N}(0,1)$
  $Z^{\text{src}}_{t_{n_{\min}}} \leftarrow (1 - t_{n_{\min}})X^{\text{src}} + t_{n_{\min}} N_{n_{\min}}$
  $Z^{\text{tar}}_{t_{n_{\min}}} \leftarrow Z^{\text{FE}}_{t_{n_{\min}}} + Z^{\text{src}}_{t_{n_{\min}}} - X^{\text{src}}$
  for $i = n_{\min}$ **to** 1 **do**
    $Z^{\text{tar}}_{t_{i-1}} \leftarrow Z^{\text{tar}}_{t_i} + (t_{i-1} - t_i)V^{\text{tar}}(Z^{\text{tar}}_{t_i}, t_i)$
  end for
  Return: $Z^{\text{tar}}_0 = X^{\text{tar}}_0$
end if

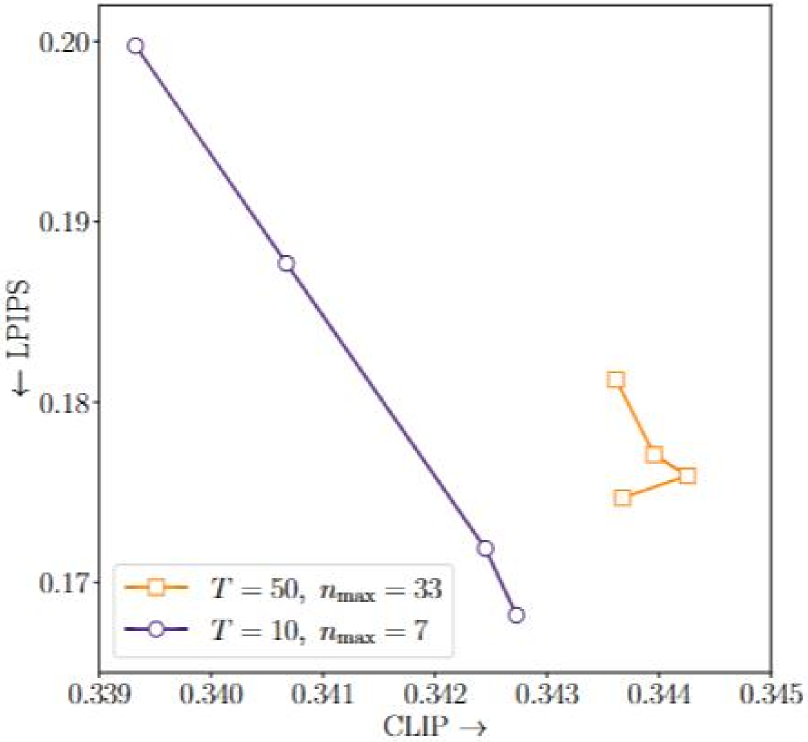


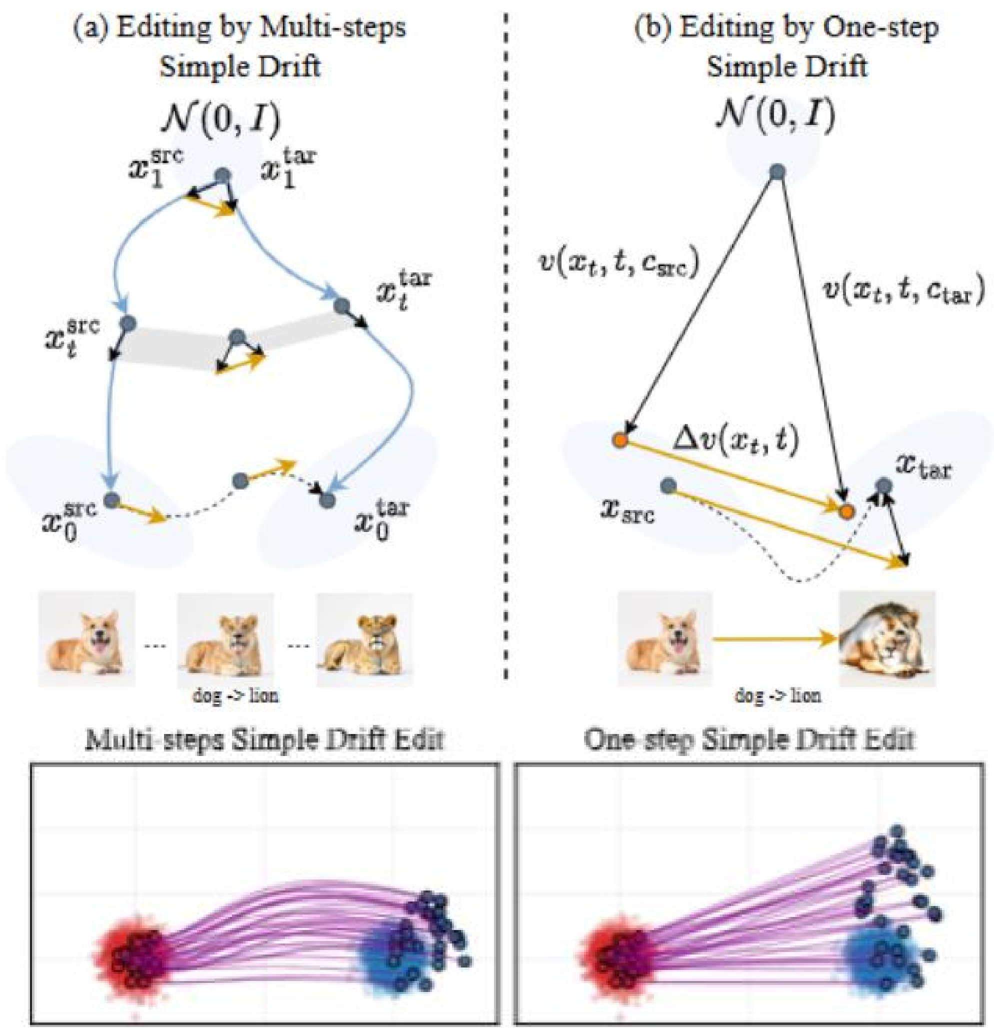
Figure S14. **CLIP vs. LPIPS for different values of n_{avg}.** From top to bottom, the markers correspond to n_{avg} of $1,\ 3,\ 5,\ 10$.

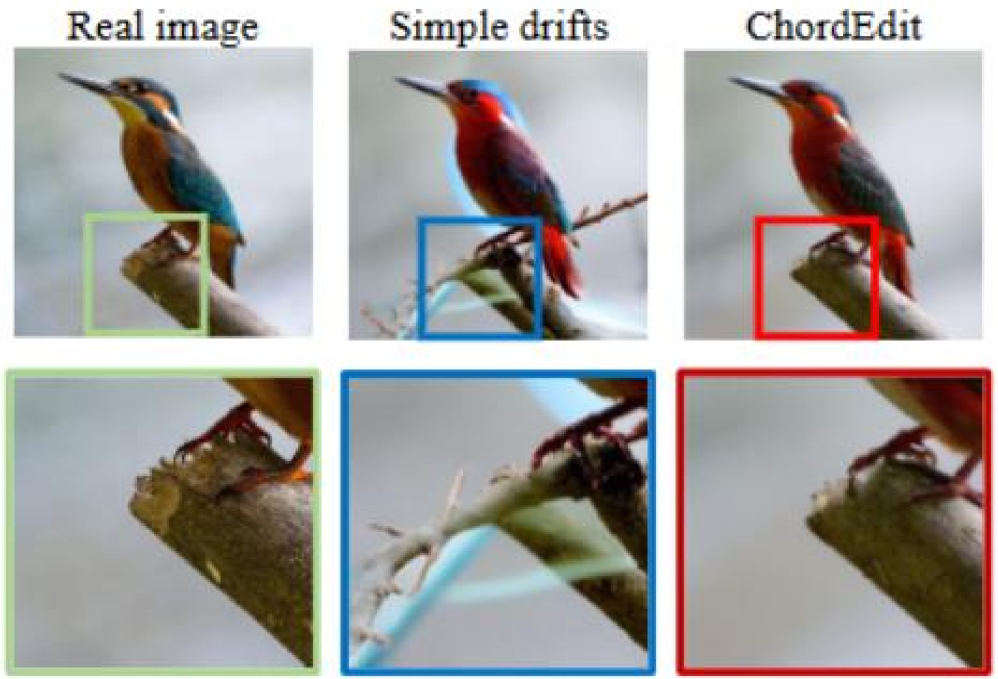
Limitations:

Figure S6. **Limitations.** FlowEdit is often fails at making substantial modifications to large region of the image, as in pose (left) and background (right) editing.

Multi-step vs. Single-step :

Multi-step editing reduces the impact of single-step variance.





ples and their paired target samples. We emphasize that our method is based on a heuristic, which is not guaranteed to precisely map to the target distribution. In particular, it does not necessarily coincide with the optimal transport mapping. Yet, as we will see in the context of images, it achieves state-of-the-art results in terms of structure preservation (*i.e.* small transportation cost) while adhering to the target text prompt (generating samples in the target distribution).

CVPR 2026

Best Student Paper Honorable Mention

ChordEdit: One-Step Low-Energy Transport for Image Editing

Liangsi Lu[1], Xuhang Chen[2], Minzhe Guo[1], Shichu Li[3], Jingchao Wang[4], Yang Shi[1†]
[1]Guangdong University of Technology [2]Huizhou University
[3]Shenzhen University [4]Peking University
[†] Corresponding author: sudo.shiyang@gmail.com

Project page: https://chordedit.github.io

Implementation details. Code is available in supplementary material. All experiments are conducted on a single NVIDIA Titan 24GB GPU. Our framework consists of the 1-NFE Chord transport step and an optional 1-NFE prox-

A **one-step**, training-free, **inversion-free** image editor that frames editing as low-energy optimal transport — **theoretically grounded** and **surprisingly effective**.

Setting Clarification

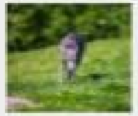
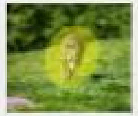
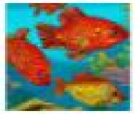
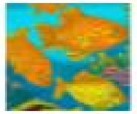
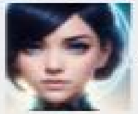
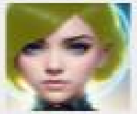
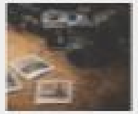
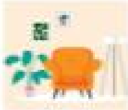
Setting: image editing by a pre-trained text-to-image (T2I) model

Input & Output :

1: **Inputs:** source image x_{src}; prompts $c_{\text{src}}, c_{\text{tar}}$; step time t; window δ; step scale λ; Proximal Refinement time t_c.
2: **Output:** edited image x_{tar}.

PIE bench:

4.3 BENCHMARK CONSTRUCTION

Editing Type	Source Image	Source Prompt	Target Prompt	Editing Instruction	Editing Mask	Editing Type	Source Image	Source Prompt	Target Prompt	Editing Instruction	Editing Mask
0 Random		A [dog] sitting on a wooden chair	A [cat] sitting on a wooden chair	Change the animal from a cat to a dog		5 Change Pose		A greyhound [walking] through the grass	A greyhound [jumping] through the grass	Change the greyhound from walking to jumping	
1 Change Object		[Fishes] in the ocean	[Sharks] in the ocean	Change the fishes to sharks		6 Change Color		A [white] shirt	A [yellow] shirt	Change the color of the shirt from white to yellow	
2 Add Object		Asian woman with black hair	Asian woman with [flowers on her] black hair	Add flowers to the woman's hair		7 Change Material		Camera, polaroid, and travel photography	[Wooden toy] camera, polaroid, and travel photography	Make the camera a wooden toy	
3 Delete Object		A painting of [an orange chair and] a lamp in a living room	A painting of a lamp in a living room	Remove the chair		8 Change Backrgound		A girl sitting in the [ruins]	A girl sitting in the [beach]	Change the background from ruins to a beach	
4 Change Content		A cave with trees and [sparkling] river	A cave with trees and [blood] river	Change the river from sparkling to blood		9 Change Style		[Painting of] a bird standing on tree branch	[A real photo of] a bird standing on tree branch	Change the image from illustration to photo	

Evaluation Metric

Semantic alignment:

CLIP-Whole score：

$$\text{CLIP-Whole} = \text{cs}\big(\text{CLIP}_I(I_{\text{edit}}),\ \text{CLIP}_T(T_{\text{trg}})\big)$$

CLIP-Edited score：

$$\Delta I = \text{CLIP}_I(I_{\text{edit}}) - \text{CLIP}_I(I_{\text{src}})$$

$$\Delta T = \text{CLIP}_T(T_{\text{trg}}) - \text{CLIP}_T(T_{\text{src}})$$

$$\text{CLIP-Edited} = \text{cs}(\Delta I, \Delta T)$$

Background consistency：

MSE

PSNR(computed exclusively on the nonedited regions):

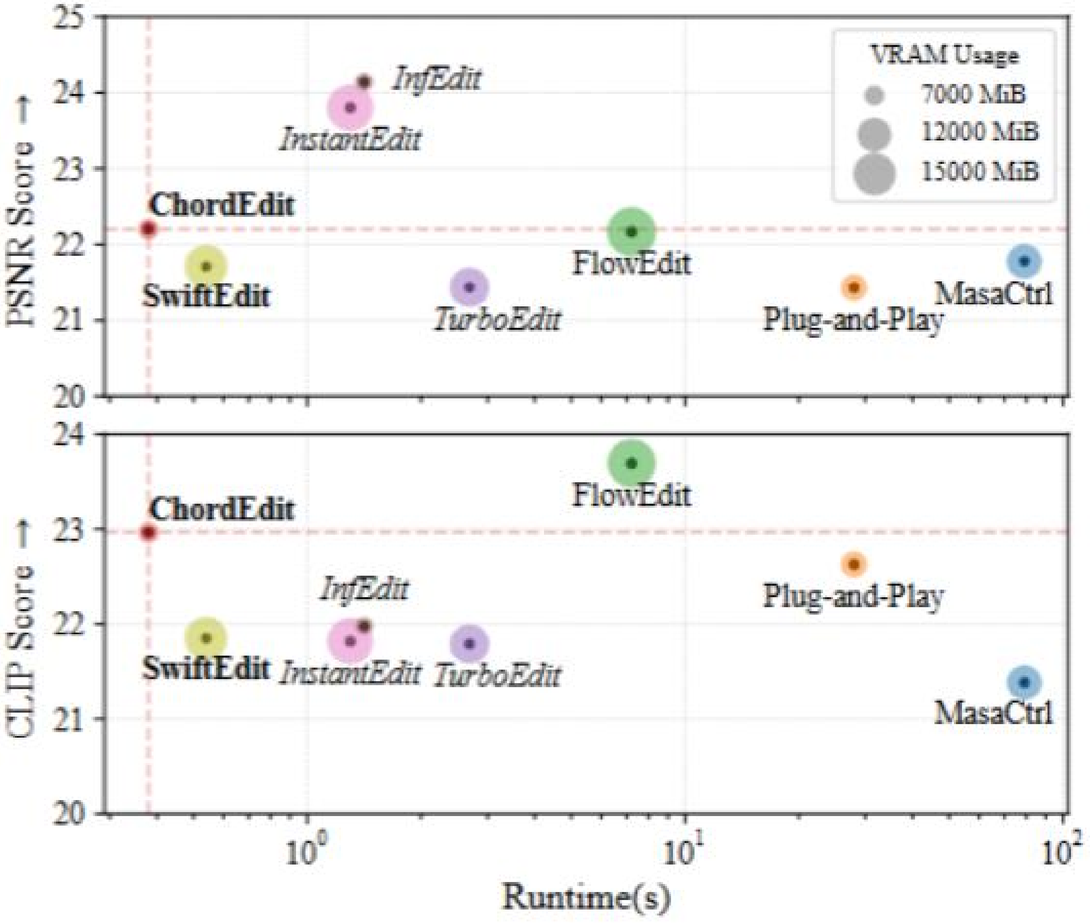
$$PSNR = 10 \cdot log_{10}(\frac{MAX_I^2}{MSE})$$

LPIPS(Learned Perceptual Image Patch Similarity)

SSIM(Structural Similarity Index Measure)

Structure Distance

Runtime



Notation & Motivation

Pre-trained text-to-image model :

time: $t \in [0,1]$

image state: xt

text condition: c

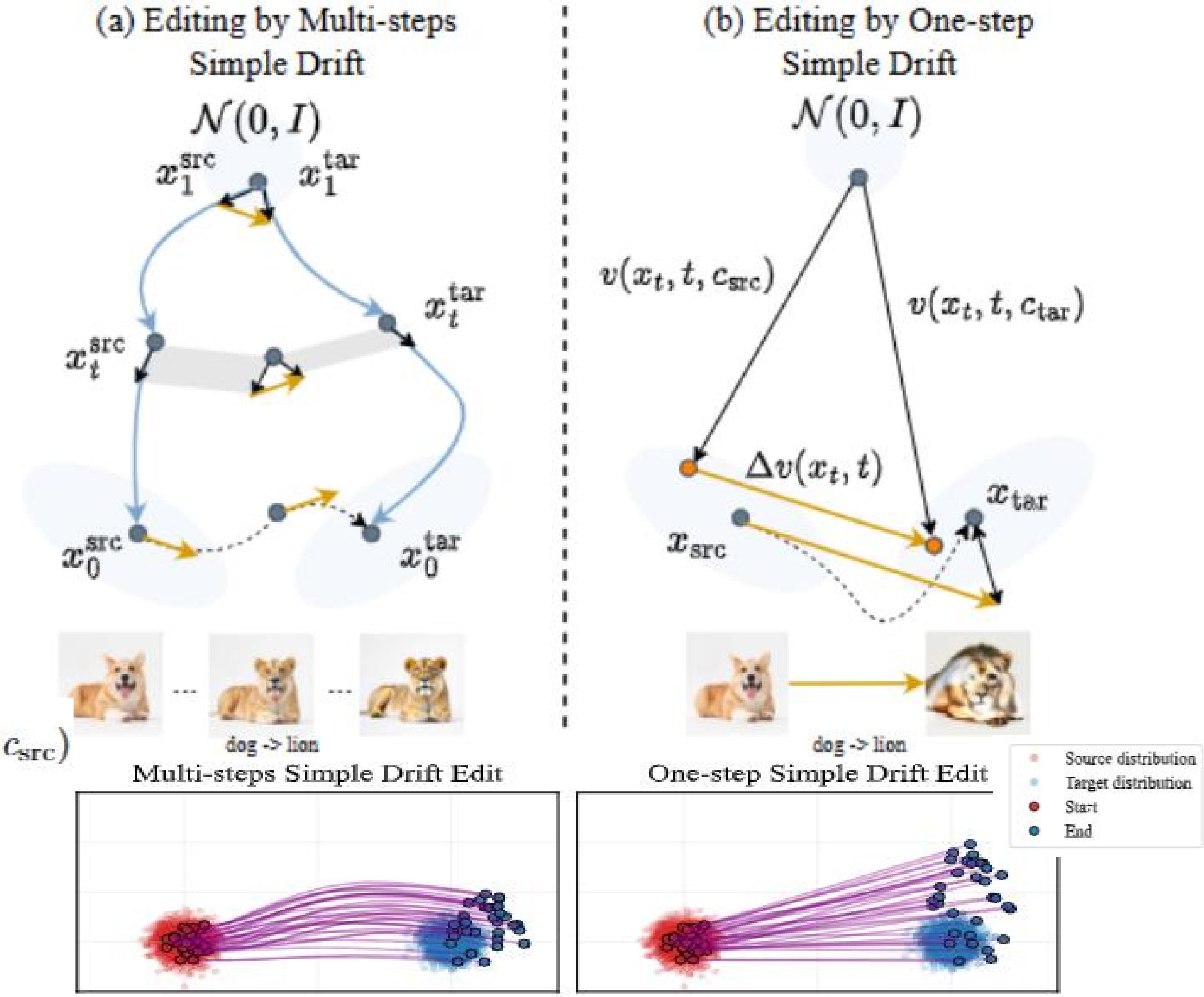
$$\frac{dx_t}{dt} = v(x_t, t, c).$$

The Editing Problem:

$$x_{\text{src}} := x_1 \sim p_1(x \mid c_{\text{src}})$$

$$x_{\text{tar}} := x_0 \sim p_0(x \mid c_{\text{tar}})$$

$$\Delta v(x_t, t) = v(x_t, t, c_{\text{tar}}) - v(x_t, t, c_{\text{src}})$$



Notation

The Editing Problem:

$$x_{\mathrm{src}} := x_1 \sim p_1(x \mid c_{\mathrm{src}})$$

$$x_{\mathrm{tar}} := x_0 \sim p_0(x \mid c_{\mathrm{tar}})$$

$$\Delta v(x_t, t) \; = \; v(x_t, t, c_{\mathrm{tar}}) - v(x_t, t, c_{\mathrm{src}})$$

Observable Model at Noisy States:

the clean source state $x_\tau := x_1$.

$$z \sim K_t(\cdot \mid x_\tau)$$

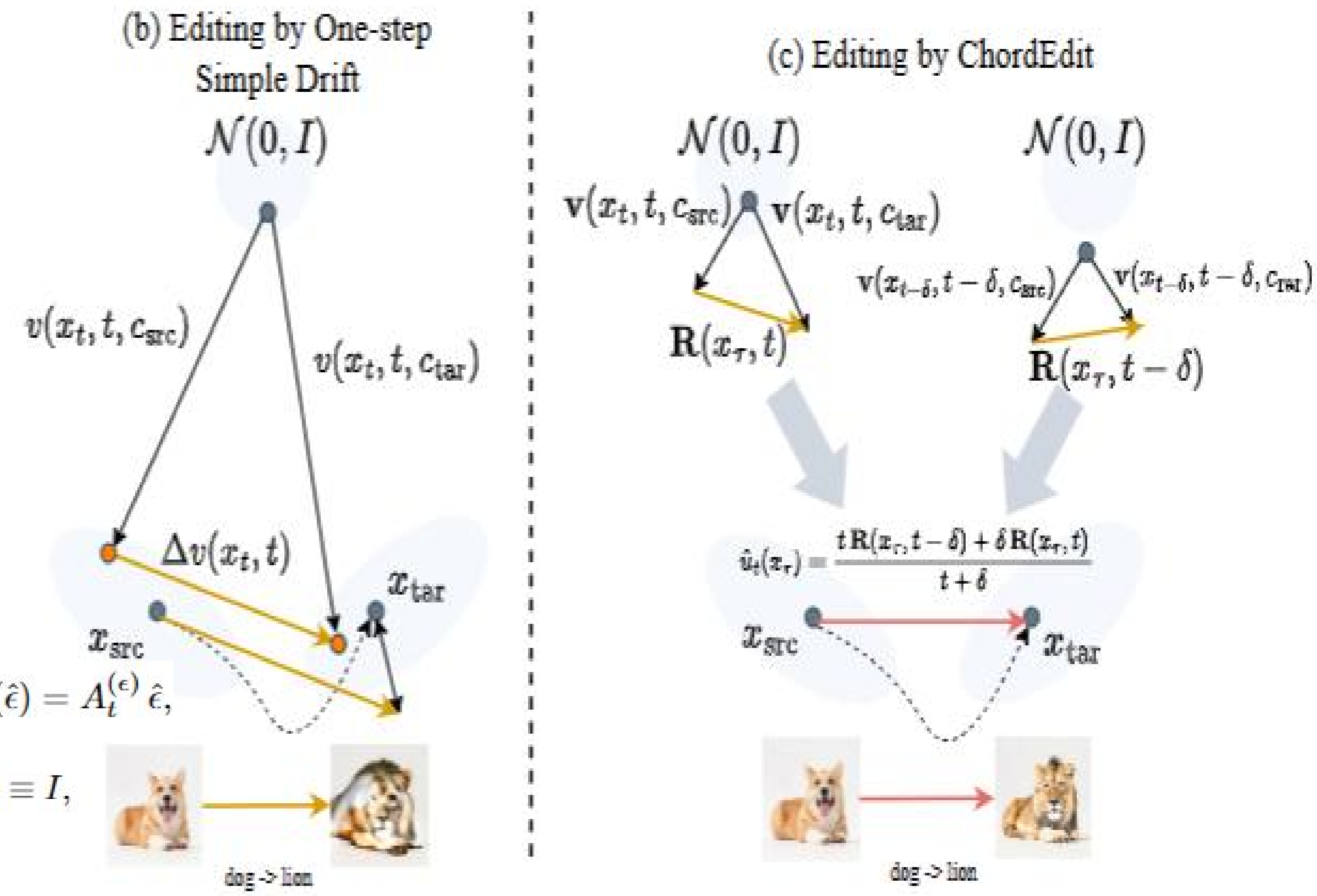
model's observable output: $Q(z, t, c)$

diffusion: $Q(z,t,c) = \hat{\epsilon}_\theta(z,t,c)$, $\quad \mathcal{B}_t(\hat{\epsilon}) = A_t^{(\epsilon)}\,\hat{\epsilon}$,

flow model: $Q(z,t,c) = \mathbf{v}_\theta(z,t,c)$, $\quad \mathcal{B}_t \equiv I$,

$$\mathbf{R}(x_\tau, t) \; = \; \mathbb{E}_{\boxed{z\sim K_t(\cdot|x_\tau)}}\big[\mathcal{B}_t\, \Delta Q(z,t)\big]$$

$$\Delta Q(z,t) \; = \; Q(z,t,c_{\mathrm{tar}}) - Q(z,t,c_{\mathrm{src}})$$



DDIM: $x_{t-1} = \sqrt{\alpha_{t-1}}\underbrace{\left(\frac{x_t - \sqrt{1-\alpha_t}\,\epsilon_\theta^{(t)}(x_t)}{\sqrt{\alpha_t}}\right)}_{\text{predicted } x_0} + \sqrt{1-\alpha_{t-1}-\sigma_t^2}\cdot\underbrace{\epsilon_\theta^{(t)}(x_t)}_{\text{predicted noise}} + \underbrace{\sigma_t\epsilon_t}_{\text{random noise}}$

unstable

$\boxed{\text{use } u_{\text{nai}} = \mathrm{R}}$ → **ChordEdit**

Optimal Transport (OT) View

Observable Model at Noisy States:

the clean source state $x_\tau := x_1$.

$z \sim K_t(\cdot \mid x_\tau)$

model's observable output: $Q(z,t,c)$

$$\mathrm{R}(x_\tau,t) = \mathbb{E}_{z\sim K_t(\cdot|x_\tau)}\big[\mathcal{B}_t\,\Delta Q(z,t)\big]$$

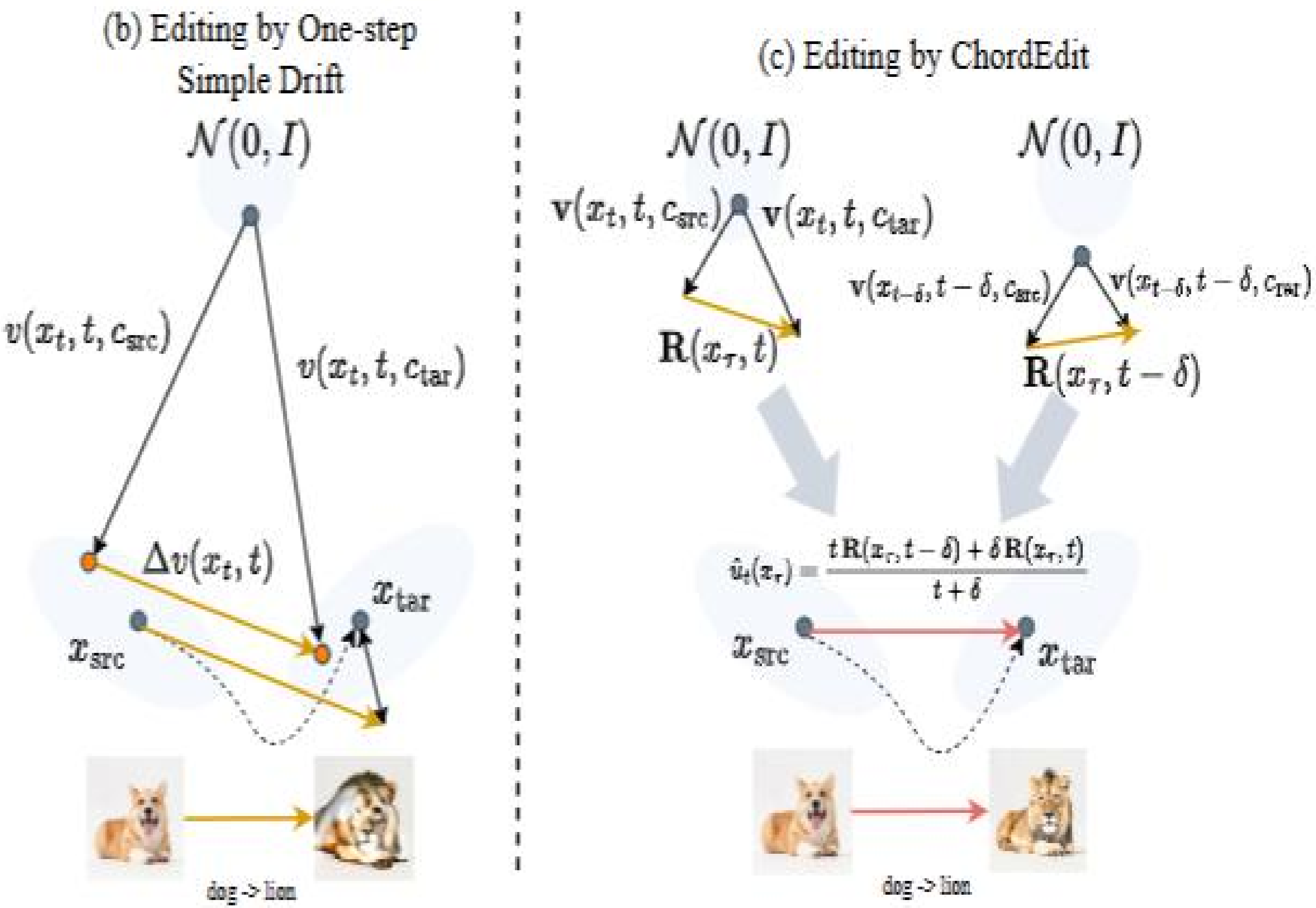
$$\Delta Q(z,t) = Q(z,t,c_{\text{tar}}) - Q(z,t,c_{\text{src}})$$

OT View: Benamou–Brenier dynamic OT problem

$$\rho_0 = p_0(\cdot \mid c_{\text{tar}}) \qquad \rho_1 = p_1(\cdot \mid c_{\text{src}})$$

$$\min_{\rho,u} \int_0^1\!\!\int \frac{1}{2}\|u_t(x)\|^2\,\rho_t(x)\,dx\,dt$$

$$\text{s.t.}\quad \partial_t\rho_t(x) + \nabla_x\cdot\big(\rho_t(x)\,u_t(x)\big) = 0.$$



The ideal u_t is the one that solves the Benamou–Brenier dynamic OT problem

But: $\mathrm{R}(x_\tau,t) = u_t(x_\tau) + \varepsilon_t, \quad \mathbb{E}[\varepsilon_t] = 0.$

Chord Control: A Low-Energy Local Estimator

OT View:Benamou–Brenier dynamic OT problem

the clean source state $x_\tau := x_1$.

$$\rho_1 = p_1(\cdot \mid c_{\text{src}}) \qquad \rho_0 = p_0(\cdot \mid c_{\text{tar}})$$

$$\min_{\rho, u} \int_0^1 \int \frac{1}{2} \|u_t(x)\|^2 \rho_t(x)\, dx\, dt$$

$$\text{s.t.} \quad \partial_t \rho_t(x) + \nabla_x \cdot (\rho_t(x)\, u_t(x)) = 0.$$

$$\mathbf{R}(x_\tau, t) = u_t(x_\tau) + \varepsilon_t, \quad \mathbb{E}[\varepsilon_t] = 0.$$

Where ut is unavailable, use estimation:

a strictly convex quadratic surrogate：

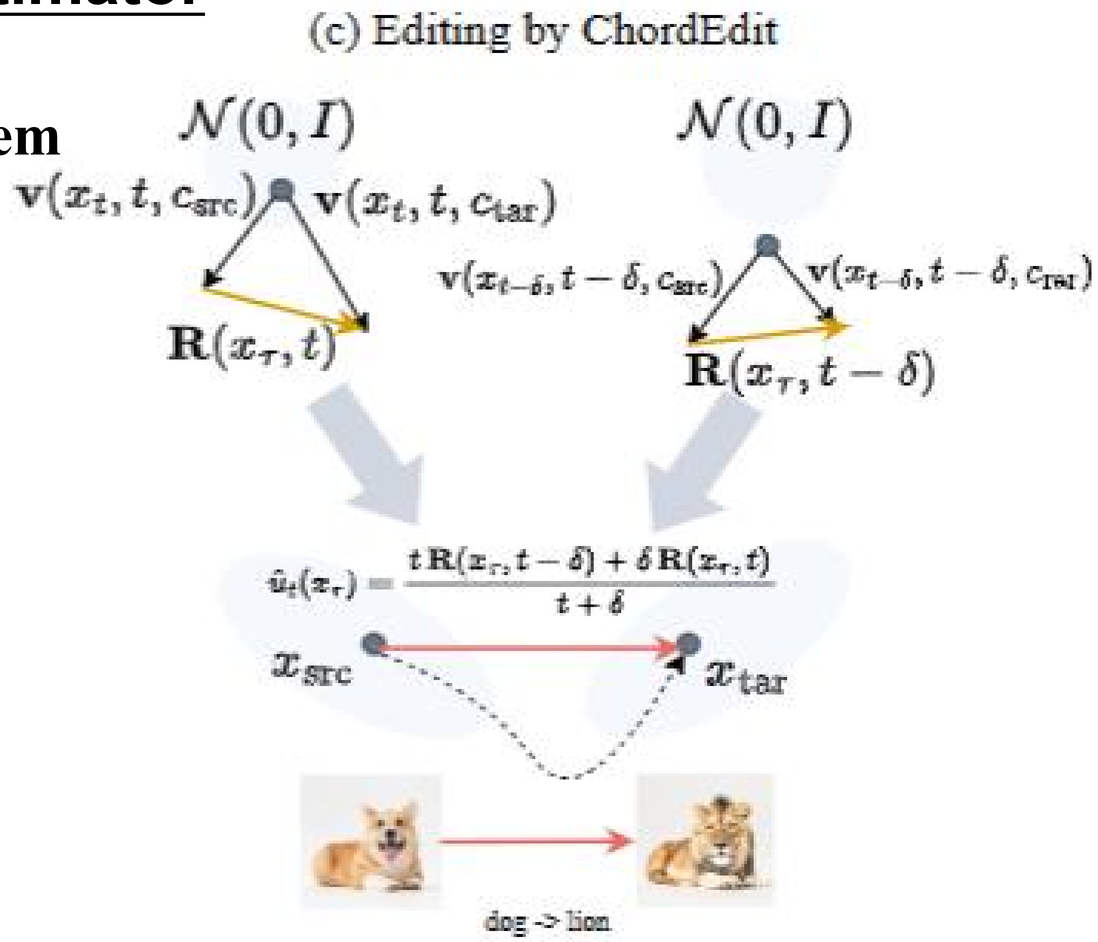
$$\Phi_t(u; x_\tau) = t\,\|u - \hat{u}_{t-\delta}(x_\tau)\|^2 + \int_{t-\delta}^{t} \|u - \mathbf{R}(x_\tau, \xi)\|^2\, d\xi.$$

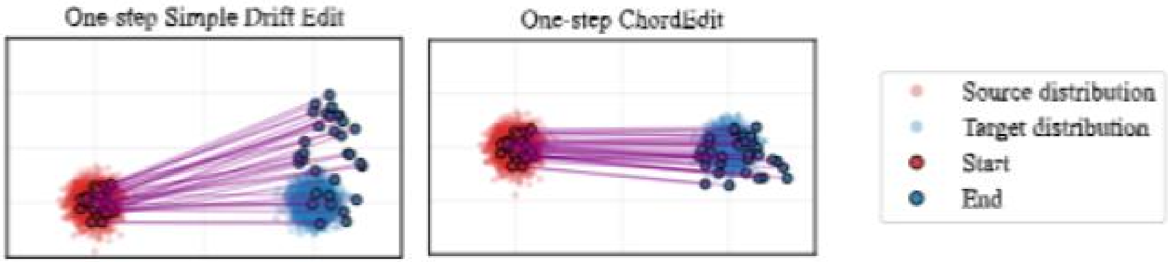
$$u_t^\star(x_\tau) = \frac{t}{t+\delta}\,\hat{u}_{t-\delta}(x_\tau) + \frac{1}{t+\delta} \int_{t-\delta}^{t} \mathbf{R}(x_\tau, \xi)\, d\xi.$$

$$\hat{u}_{t-\delta} \approx \mathbf{R}(x_\tau, t-\delta) \text{ and } \int_{t-\delta}^{t} \mathbf{R}(x_\tau, \xi) d\xi \approx \delta \mathbf{R}(x_\tau, t),$$

Chord Control Field(CCF):

$$\hat{u}_t(x_\tau) = \frac{t\,\mathbf{R}(x_\tau, t-\delta) + \delta\,\mathbf{R}(x_\tau, t)}{t+\delta}.$$





Proximal Refinement

Notation:

model's observable output: $Q(z,t,c)$

$$\mathbf{R}(x_\tau, t) = \mathbb{E}_{z\sim K_t(\cdot|x_\tau)}\left[\mathcal{B}_t\, \Delta Q(z,t)\right]$$

$$\Delta Q(z,t) = Q(z,t,c_{\text{tar}}) - Q(z,t,c_{\text{src}})$$

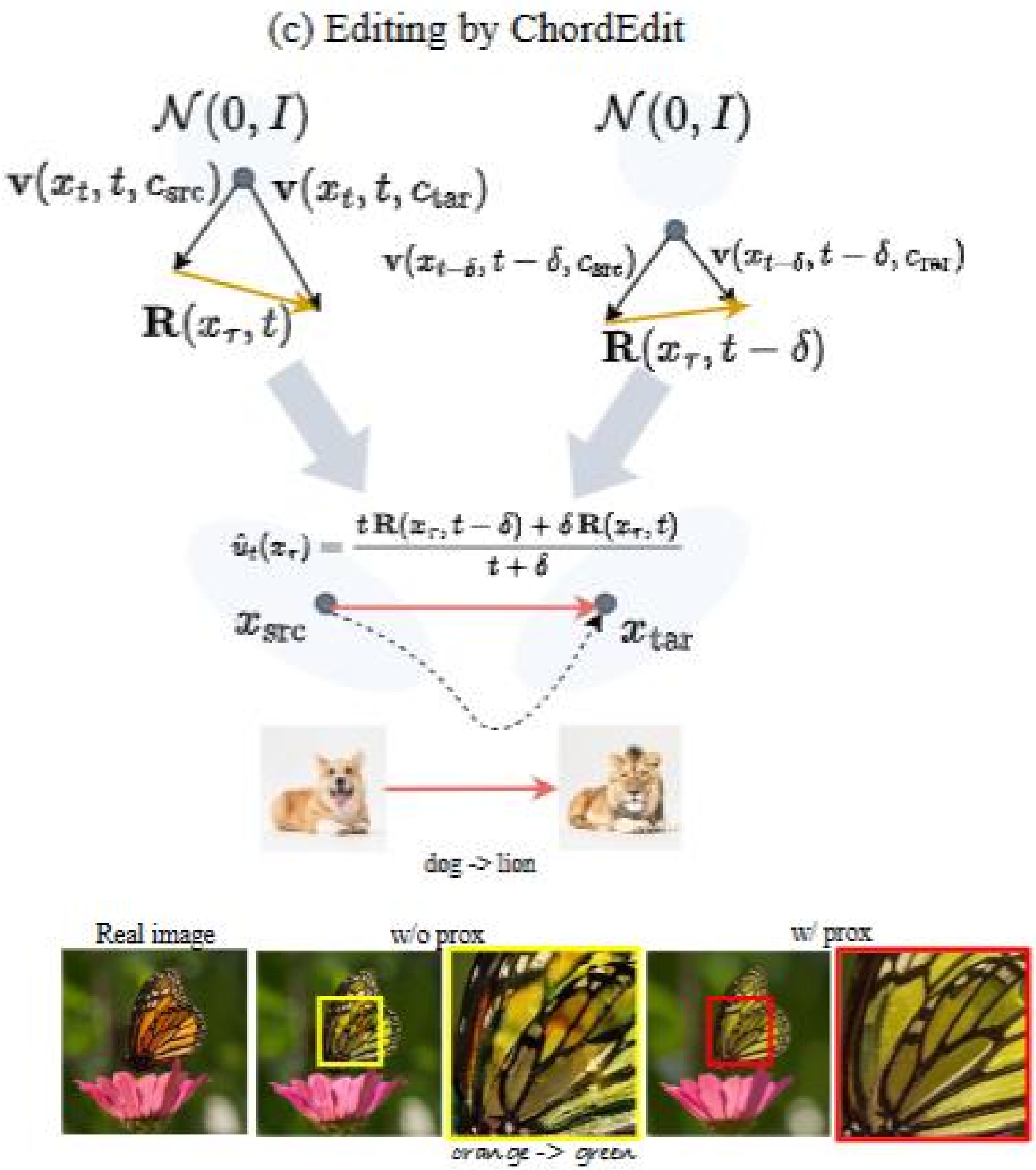
Proximal Refinement:

$$\operatorname{prox}\left(x^{\text{pred}}, t_c, c_{\text{tar}}\right) = \mathcal{B}_{t_c}\, Q\left(x^{\text{pred}}, t_c, c_{\text{tar}}\right)$$

Algorithm 1 Simplified algorithm for ChordEdit

Inputs: source image x_{src}; prompts $c_{\text{src}}, c_{\text{tar}}$; step time t; window δ; step scale λ; Proximal Refinement time t_c.
Output: edited image x_{tar}.
Init: $x_{\text{in}} \leftarrow x_{\text{src}}$
$\hat{u} \leftarrow \frac{t\,\mathbf{R}(x_{\text{in}}, t-\delta)+\delta\,\mathbf{R}(x_{\text{in}}, t)}{t+\delta}$
$x^{\text{pred}} \leftarrow x_{\text{in}} + \lambda\,\hat{u}$
$x_{\text{tar}} \leftarrow \operatorname{prox}\left(x^{\text{pred}}, t_c, c_{\text{tar}}\right)$ ▷ Optional
Return x_{tar}

$t = 0.90, \delta = 0.15, \lambda = 1.00, t_c = 0.30$



Experiments

Table 1. **Quantitative comparison on PIE-bench** [11]. **T-free**: Training-free. **I-free**: Inversion-free. The best/second/third results in each numeric column are highlighted with yellow / orange / blue backgrounds, respectively. A comprehensive table with extended metrics (e.g., SSIM, Structure Distance) is available in Appendix.

Type	Method	Consistency			CLIP Semantics		Properties		Efficiency			
		PSNR↑	MSE_{10^3} ↓	$LPIPS_{10^3}$ ↓	Whole↑	Edited↑	T-free	I-free	Step↓	NFE↓	Runtime(s)↓	VRAM(MiB)↓
Multi-step (≥ 30 steps)	DDIM + MasaCtrl [3, 30]	21.25	8.58	106.59	24.13	21.13	✓	✗	50	100	55.20	12272
	Direct Inversion + MasaCtrl [3, 11]	21.78	7.99	87.38	24.42	21.38	✓	✗	50	100	79.10	12272
	DDIM + PnP [30, 33]	21.26	8.42	113.58	25.45	22.54	✓	✗	50	100	28.01	9262
	Direct Inversion + PnP [11, 33]	21.43	8.10	106.26	25.48	22.63	✓	✗	50	100	28.03	9262
	FlowEdit (SD3) [14]	22.17	7.69	104.81	26.64	23.69	✓	✓	33	33	7.22	17140
Few-step (4 steps)	TurboEdit (SDXL-Turbo) [6]	21.44	9.49	108.60	24.66	21.79	✓	✓	4	4	2.69	13826
	InfEdit (SD1.4) [36]	24.14	6.82	55.69	24.89	21.88	✓	✓	4	4	1.41	6502
	InstantEdit (PeRFlow-SD1.5) [7]	23.80	4.21	60.92	24.97	21.82	✓	✗	4	8	1.30	16270
One-step	SwiftEdit (SwiftBrush-v2) [24]	21.71	8.22	91.22	24.93	21.85	✗	✗	1	2	0.54	15060
	ChordEdit (SwiftBrush-v2)	22.04	7.13	111.22	25.12	22.58	✓	✓	1	2	**0.38**	6988
	ChordEdit (w/o prox, SD-Turbo)	23.89	5.05	88.36	24.97	21.87	✓	✓	1	1	**0.20**	6988
	ChordEdit (SD-Turbo)	22.20	6.84	128.25	25.58	22.96	✓	✓	1	2	**0.38**	6988

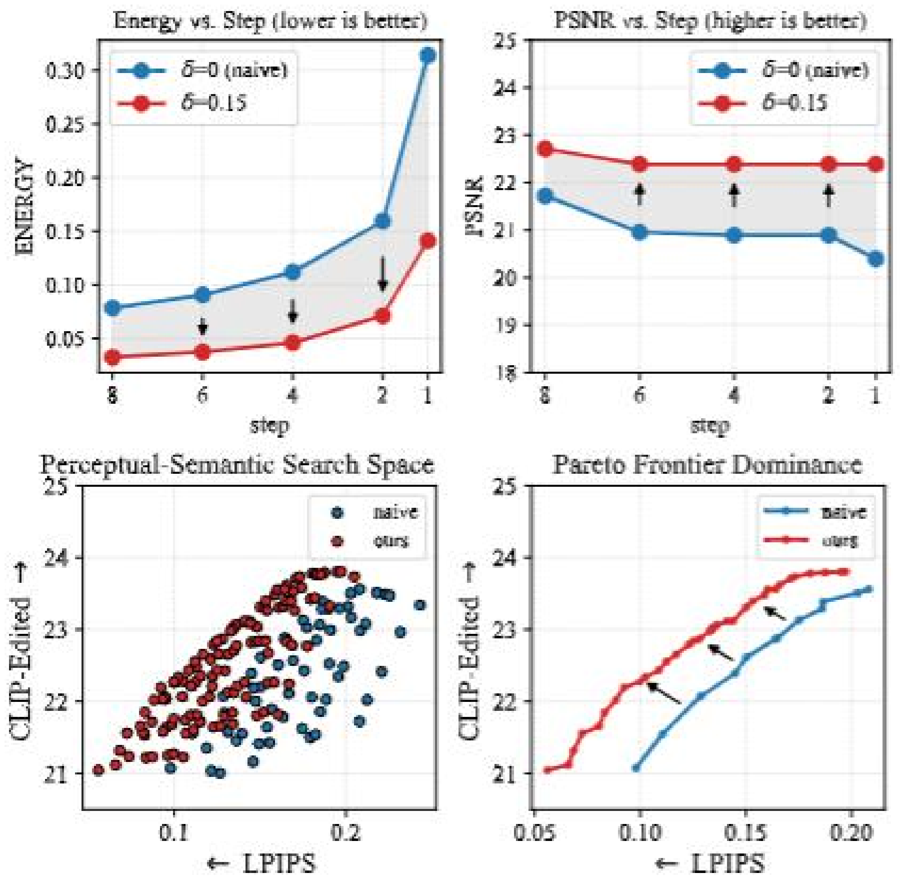
Table 4. Quantitative comparison of our method against other editing methods on PIE Bench.

Type	Method	Struct.	Background Preservation				CLIP Semantics		Efficiency			
		$Dist._{10^3}$ ↓	PSNR↑	MSE_{10^3} ↓	$SSIM_{10^2}$ ↑	$LPIPS_{10^3}$ ↓	Whole↑	Edited↑	Runtime(s)↓	Step↓	NFE↓	VRAM(MiB)↓
Multi-step (≥ 20 steps)	DDIM + MasaCtrl	28.79	21.25	8.58	80.11	106.59	24.13	21.13	55.19	50	100	12272
	Direct Inversion + MasaCtrl	24.46	21.78	7.99	81.74	87.38	24.42	21.38	79.10	50	100	12272
	DDIM + PnP	28.20	21.26	8.42	78.90	113.58	25.45	22.54	28.01	50	100	9262
	Direct Inversion + PnP	24.27	21.43	8.10	79.52	106.26	25.48	22.63	28.03	50	100	9262
	FlowEdit (SD3)	12.34	22.17	7.69	83.54	104.81	**26.64**	**23.69**	7.22	33	33	17140
Few-step (4 steps)	TurboEdit (SDXL-Turbo)	13.80	21.44	9.49	80.08	108.60	24.66	21.79	2.69	<u>4</u>	4	13826
	InfEdit (SD1.4)	17.06	**24.14**	6.82	**85.02**	<u>55.69</u>	24.89	21.88	1.41	<u>4</u>	4	<u>6502</u>
	InstantEdit (PeRFlow-SD1.5)	<u>7.14</u>	<u>23.80</u>	**4.21**	<u>84.84</u>	60.92	24.97	21.82	1.30	<u>4</u>	8	16270
One-step	SwiftEdit (SwiftBrush-v2)	12.96	21.71	8.22	74.84	91.22	24.93	21.85	<u>0.54</u>	**1**	<u>2</u>	15060
	ChordEdit (Naive, InstaFlow)	13.32	22.05	10.45	73.49	103.33	22.97	20.19	**0.38**	**1**	<u>2</u>	**6198**
	ChordEdit (Our, InstaFlow)	**6.33**	23.05	<u>5.45</u>	82.49	**53.33**	24.17	21.39	**0.38**	**1**	<u>2</u>	**6198**
	ChordEdit (Naive, SwiftBrush-v2)	16.33	20.52	17.17	73.42	127.43	23.78	21.06	**0.38**	**1**	<u>2</u>	6988
	ChordEdit (Our, SwiftBrush-v2)	12.96	22.04	7.13	75.84	111.22	25.12	22.58	**0.38**	**1**	<u>2</u>	6988
	ChordEdit (Naive, SD-Turbo)	25.44	21.38	9.73	74.39	131.30	25.11	21.96	**0.38**	**1**	<u>2</u>	6988
	ChordEdit (Naive w/o prox, SD-Turbo)	19.18	21.89	10.84	77.24	105.27	23.68	20.83	**0.20**	**1**	**1**	6988
	ChordEdit (Ours, SD-Turbo)	16.58	22.20	6.84	75.91	128.25	<u>25.58</u>	<u>22.96</u>	**0.38**	**1**	<u>2</u>	6988
	ChordEdit (Ours w/o prox, SD-Turbo)	<u>10.37</u>	23.89	5.05	81.24	88.36	24.97	21.87	**0.20**	**1**	**1**	6988

Ablation Study：Low-Energy

Benamou–Brenier kinetic energy：S is the total step count

Figure 8. **Qualitative Comparison and Energy Visualization.** We compare ChordEdit (Ours) against the naive baseline ($\delta = 0$, Naive). The naive method's high-energy field leads to artifacts and background corruption. Our ChordEdit derives a stable, low-energy field, resulting in high-fidelity edits that preserve object identity and non-edited regions. Results shown used SwiftBrush-v2 (first column) and SD-Turbo (second and third columns). Energy plots are computed as $E = \frac{1}{SC}\sum_{s=1}^{S}\sum_{\text{channel}}(\hat{u}_{t_s})^2$.



Ablation Study：Analysis of Noise

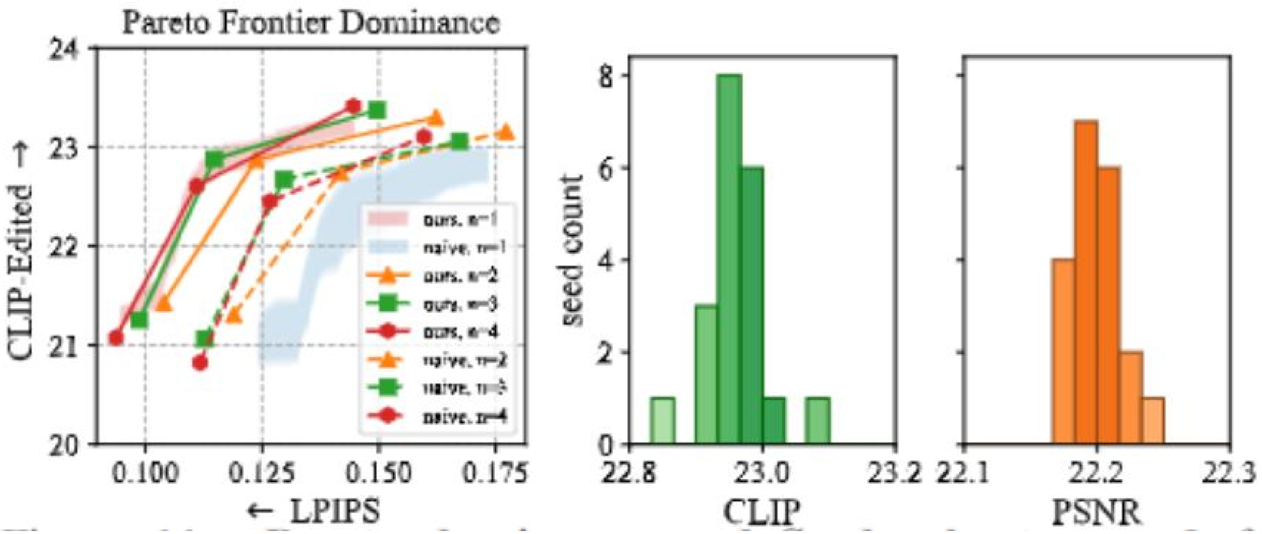
$$\mathbf{R}(x_\tau, t) \;=\; \mathbb{E}_{\boxed{z\sim K_t(\cdot|x_\tau)}}\big[\mathcal{B}_t\,\Delta Q(z,t)\big]$$

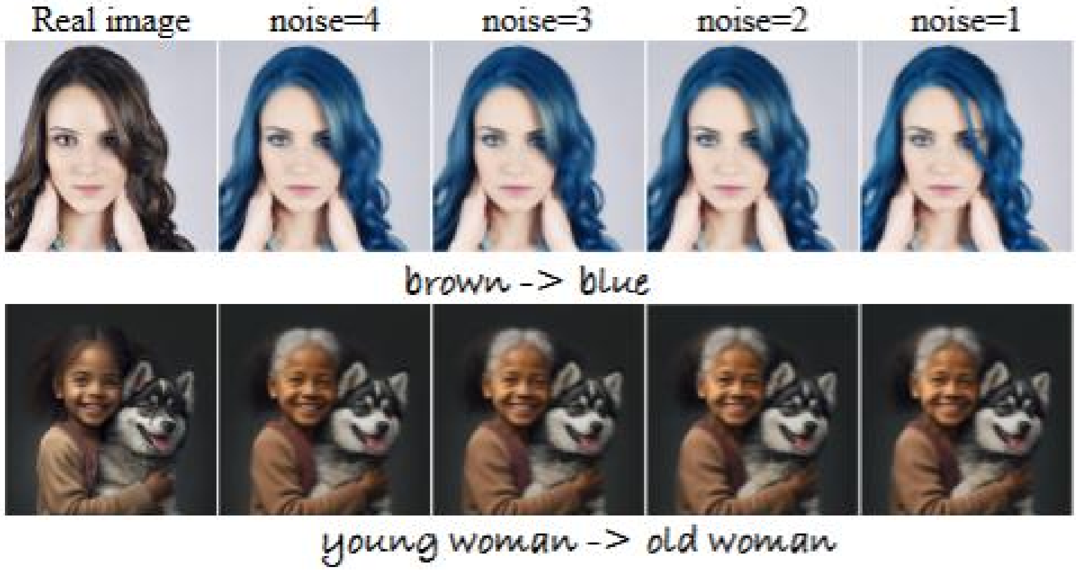
Algorithm 1 Simplified algorithm for ChordEdit

Inputs: source image x_{src}; prompts $c_{\text{src}}, c_{\text{tar}}$; step time t; window δ; step scale λ; Proximal Refinement time t_c.
Output: edited image x_{tar}.
Init: $x_{\text{in}} \leftarrow x_{\text{src}}$
$\hat{u} \leftarrow \frac{t\,\mathbf{R}(x_{\text{in}},\,t-\delta)+\delta\,\mathbf{R}(x_{\text{in}},\,t)}{t+\delta}$
$x^{\text{pred}} \leftarrow x_{\text{in}} + \lambda\,\hat{u}$
$x_{\text{tar}} \leftarrow \text{prox}\left(x^{\text{pred}},\, t_c,\, c_{\text{tar}}\right)$ ▷ Optional
Return x_{tar}

Algorithm 2 ChordEdit (Multi-noise $n > 1$ version)

Inputs: source image x_{src}; prompts $c_{\text{src}}, c_{\text{tar}}$; step time t; window δ; step scale λ; Proximal Refinement time t_c; number of noise samples n.
Output: edited image x_{tar}.
Init: $x_{\text{in}} \leftarrow x_{\text{src}}$
$\boxed{\hat{u}_{\text{sum}} \leftarrow 0}$
for $i = 1$ **to** n **do**
    $\mathbf{R}_{t-\delta} \leftarrow \mathbf{R}(x_{\text{in}}, t-\delta)$
    $R_t \leftarrow R(x_{\text{in}}, t)$
    $\hat{u}_i \leftarrow \frac{t\,\mathbf{R}_{t-\delta}+\delta\,\mathbf{R}_t}{t+\delta}$
    $\boxed{\hat{u}_{\text{sum}} \leftarrow \hat{u}_{\text{sum}} + \hat{u}_i}$
end for
$\boxed{\hat{u}_{\text{avg}} \leftarrow \hat{u}_{\text{sum}}/n}$
$x^{\text{pred}} \leftarrow x_{\text{in}} + \lambda\,\hat{u}_{\text{avg}}$
$x_{\text{tar}} \leftarrow \text{prox}\left(x^{\text{pred}},\, t_c,\, c_{\text{tar}}\right)$
Return x_{tar}





Ablation Study：Transport and Refinement

Table 2. **Ablation of Chord transport field and refinement.** Our Chord field drives consistency (PSNR), while the prox step boosts semantics (CLIP-Edited). Full metrics are in the Appendix.

Method	Naive ($\delta = 0$)		Ours ($\delta = 0.15$)		NFE
	PSNR ↑	CLIP-Edited ↑	PSNR ↑	CLIP-Edited ↑	
w/o prox	21.89	20.83	**23.89**	21.87	1
w/ prox	21.38	21.96	22.20	**22.96**	2

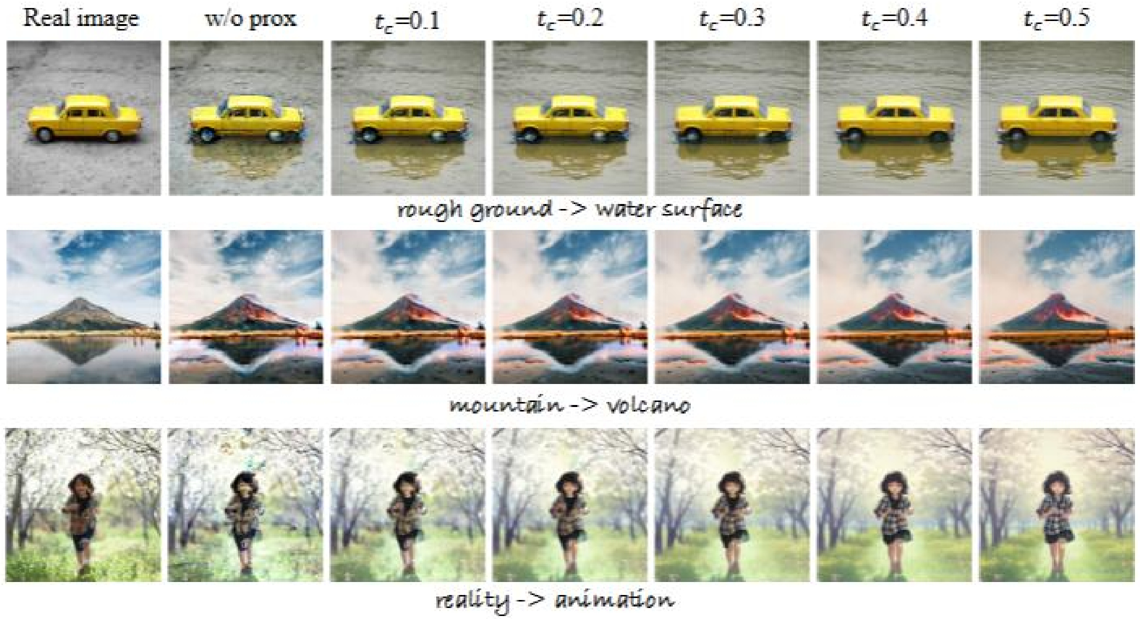
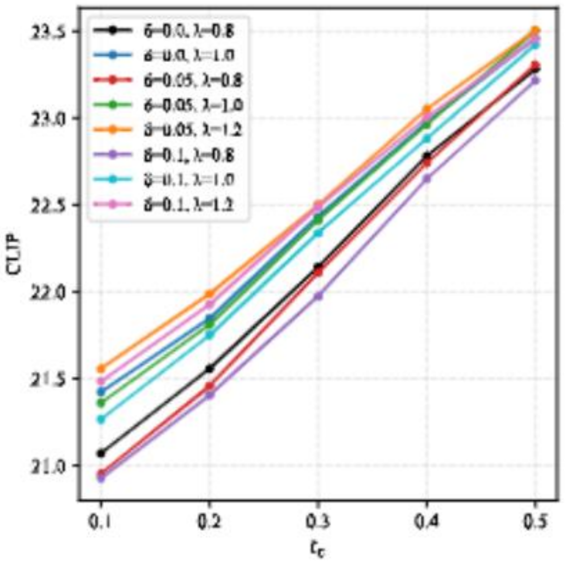
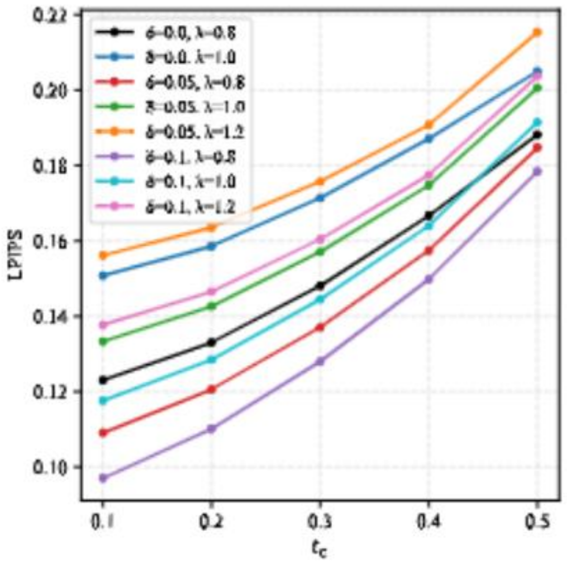


Figure 17. **Qualitative analysis of the proximal refinement time t_c.**





Thanks

2026.06.12
Lei Tu