

HUST

ĐẠI HỌC BÁCH KHOA HÀ NỘI
HANOI UNIVERSITY OF SCIENCE AND TECHNOLOGY

ONE LOVE. ONE FUTURE.



**ĐẠI HỌC
BÁCH KHOA HÀ NỘI**
HANOI UNIVERSITY
OF SCIENCE AND TECHNOLOGY

FlowEdit: Inversion-Free Text-Based Editing Using Pre-Trained Flow Models

ONE LOVE. ONE FUTURE.

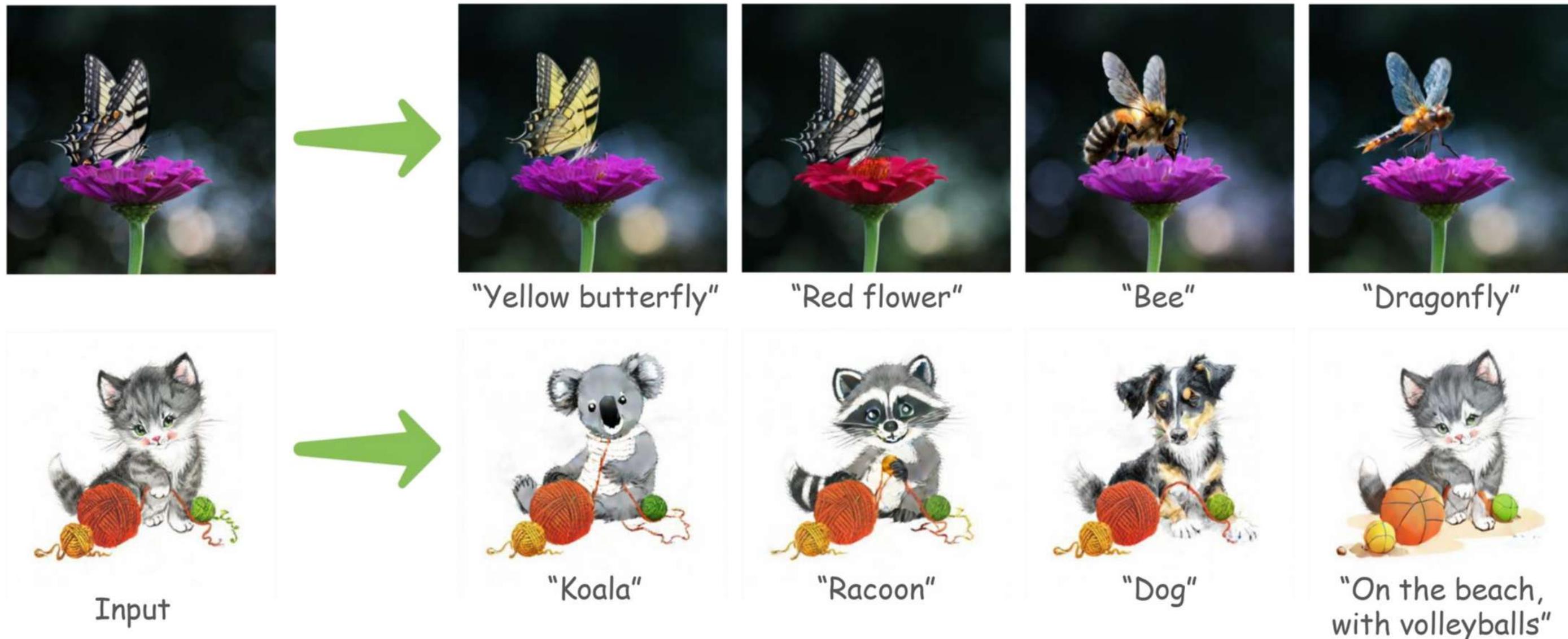
Nội dung chính

- Text-Based Image Editing
- ODE Inversion: Reinterpretation
- FlowEdit
- InstaFlow
- Ứng dụng FlowEdit vào InstaFlow



Text-Based Image Editing

- Text-Based Image Editing là một bài toán biến đổi hình ảnh có điều kiện can thiệp vào không gian ngữ nghĩa của bức ảnh



- ODE Inversion chính là phương pháp cơ bản nhất gồm 2 giai đoạn Inversion và Generation
- Quay trở lại với công thức cập nhật của Diffusion/Flow model:
 - Với mô hình dự đoán nhiễu:

$$x_{s_1} = \sqrt{\frac{\alpha_{s_1}}{\alpha_{s_2}}} (x_{s_2} - \sqrt{1 - \alpha_{s_2}} \varepsilon(x_{s_2}, s_2, C)) + \sqrt{1 - \alpha_{s_1}} \varepsilon(x_{s_2}, s_2, C)$$

- Với mô hình dự đoán vận tốc:

$$x_{s_1} = x_{s_2} + (s_1 - s_2) v(x_{s_2}, s_2, C)$$

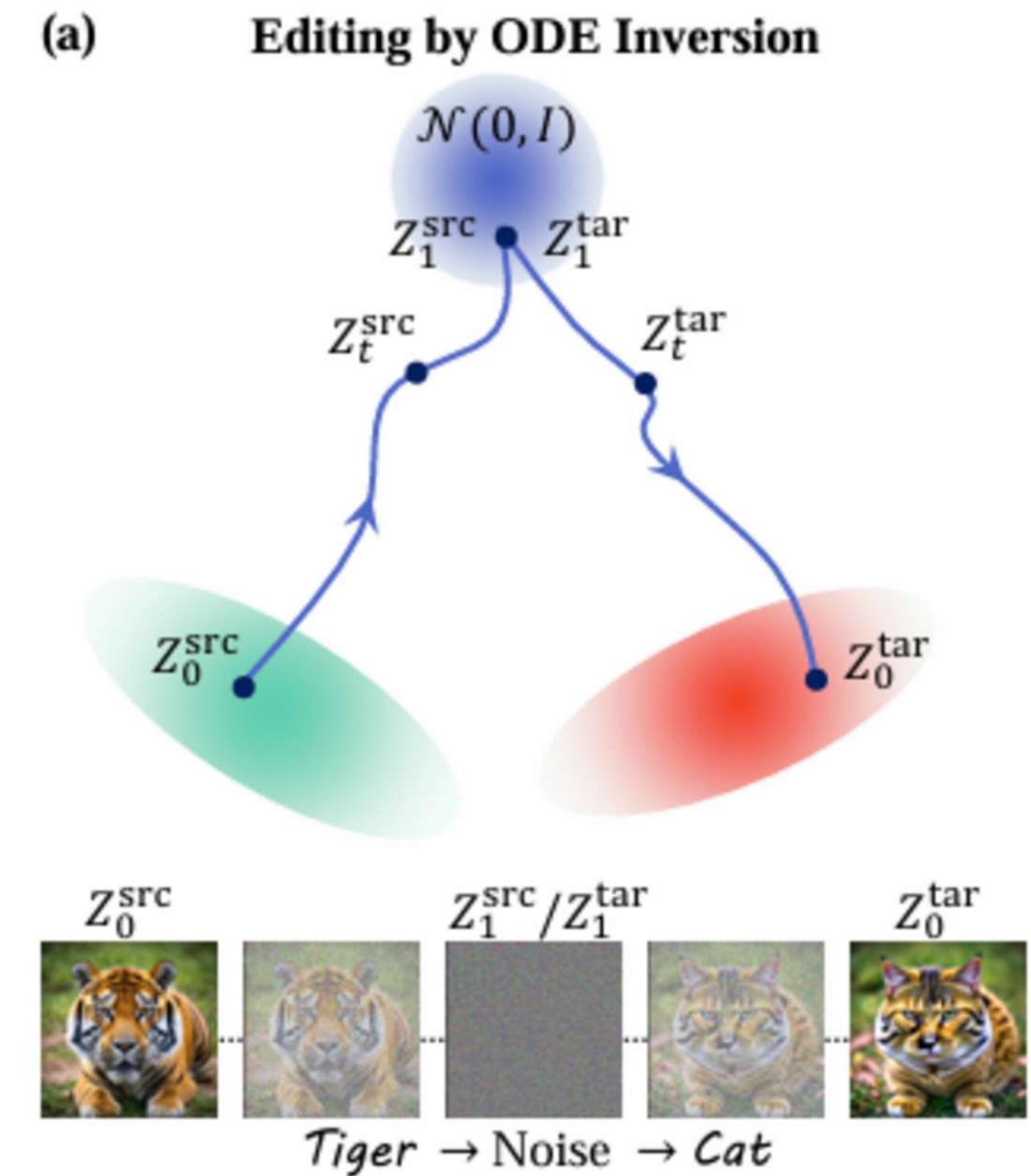
- Để đi ngược từ ảnh thật về nhiễu chỉ cần cho $s_1 > s_2$

ODE Inversion

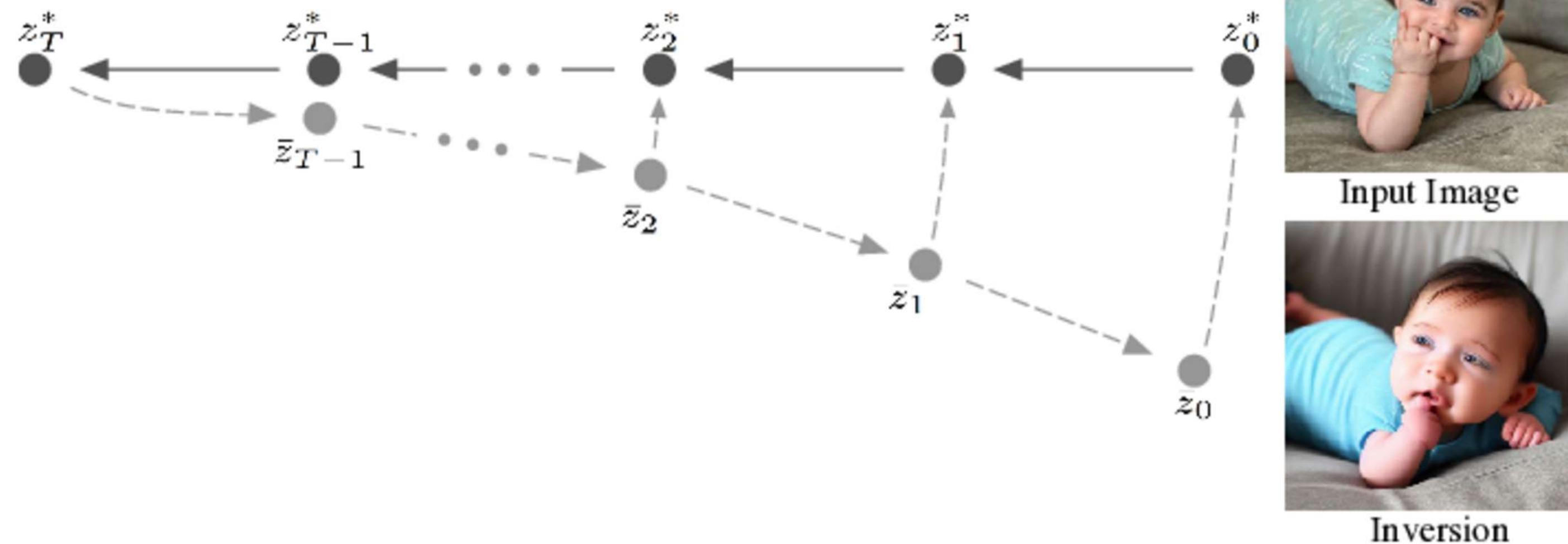
- Ta có một bức ảnh X_{src} và câu mô tả c_{src} , ta muốn dùng câu mô tả c_{tar} để biến thành ảnh X_{tar}
- Ta thực hiện hai giai đoạn: đưa ảnh gốc về nhiễu rồi đưa nhiễu thành ảnh ta cần

$$X_{src} = Z_0^{src} \rightarrow \dots \rightarrow Z_1^{src} \sim N(0, 1)$$

$$Z_1^{src} = Z_1^{tar} \rightarrow \dots \rightarrow Z_0^{tar} = X_{tar}$$



- ODE Inversion còn có nhiều hạn chế:
 - Sai số tích lũy: Mô hình chỉ có thể xấp xỉ giá trị thật chứ không chính xác tuyệt đối
 - Xung đột Guidance:



Reinterpretation of ODE Inversion

- Nhóm tác giả FlowEdit định nghĩa lại cách đi của ODE Inversion, tạo ra một đường mới đi từ ảnh nguồn đến ảnh đích:

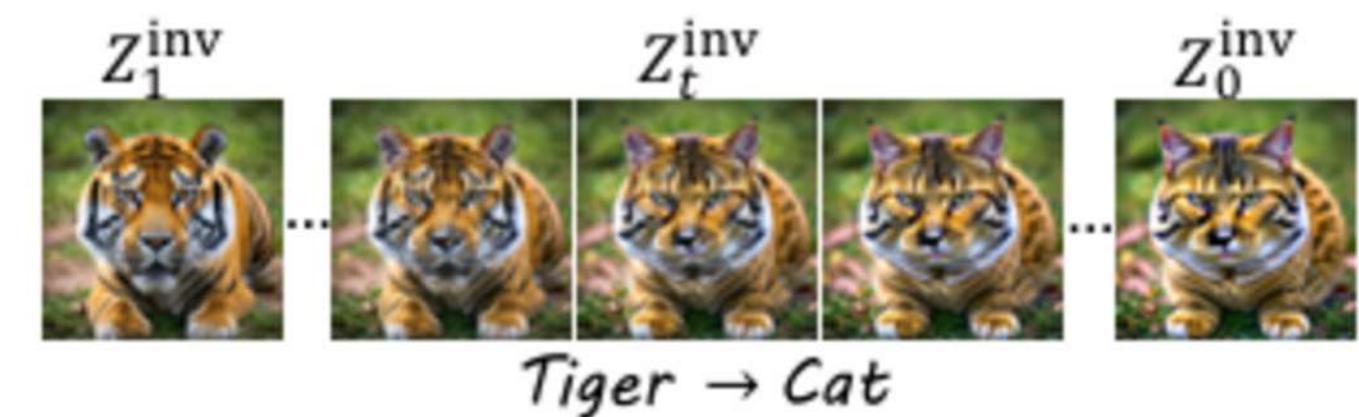
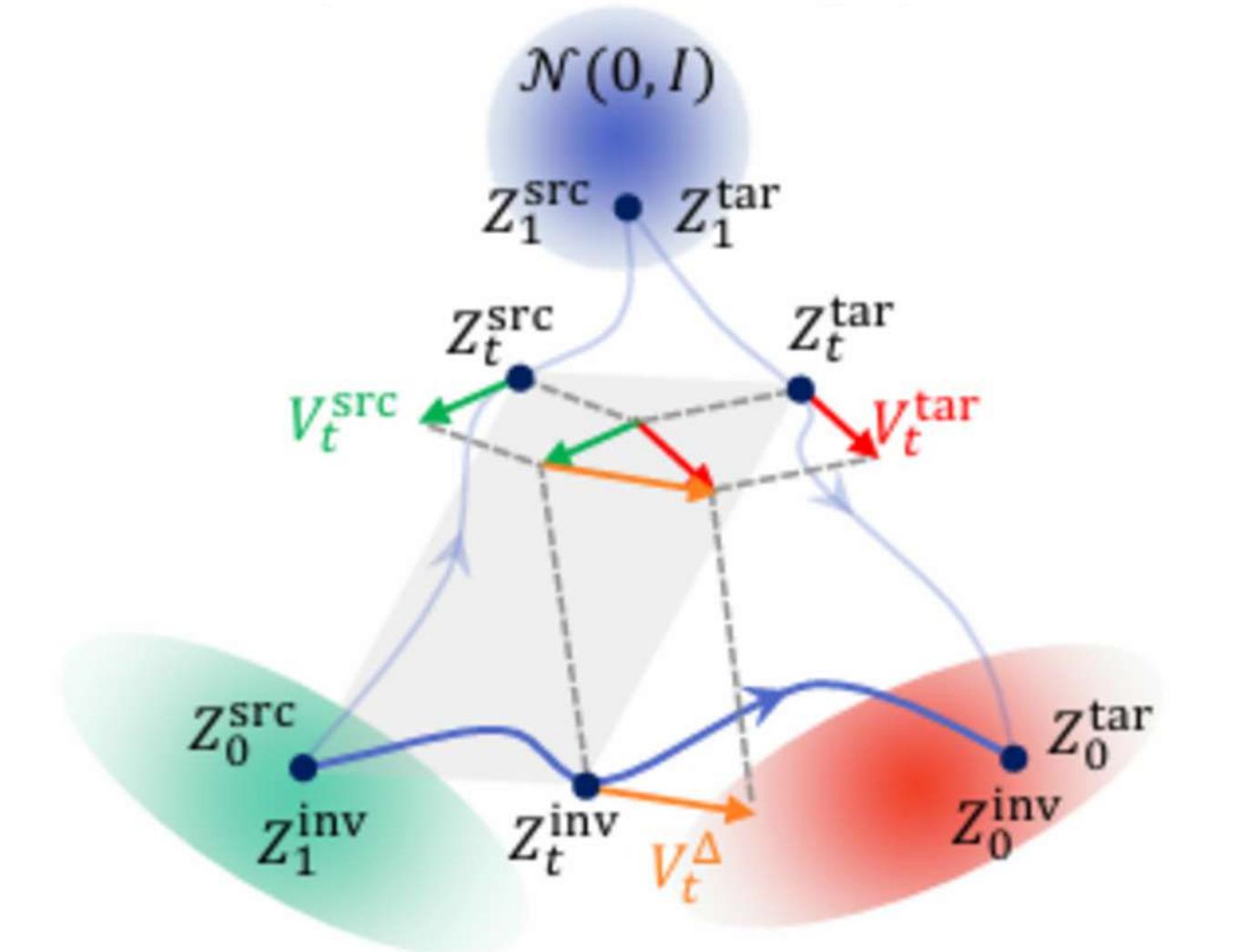
$$Z_t^{inv} = Z_0^{src} + Z_t^{tar} - Z_t^{src}$$

$$Z_t^{src} = (1-t)Z_0^{src} + tZ_1^{src}$$

$$Z_t^{tar} = (1-t)Z_0^{tar} + tZ_1^{tar}$$

$$\begin{aligned} Z_t^{inv} &= Z_0^{src} + (1-t)(Z_0^{tar} - Z_0^{src}) \\ &= tZ_0^{src} + (1-t)Z_0^{tar} \end{aligned}$$

(b) Reinterpretation of Editing by Inversion



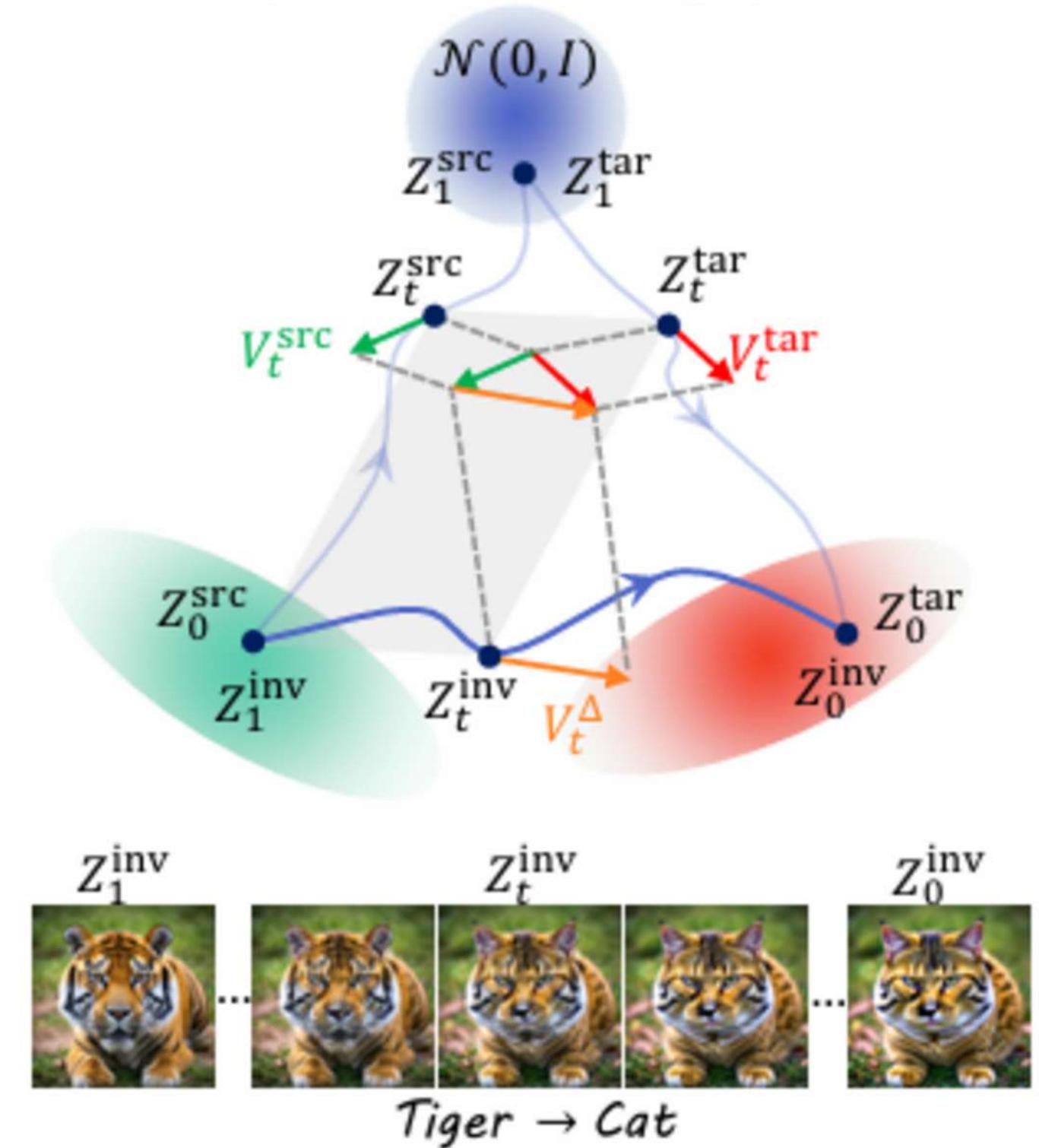
Reinterpretation of ODE Inversion

- Điểm đặc biệt của con đường mới:
Triệt tiêu nhiễu (Noise-free)

$$Z_t^{inv} = Z_0^{src} + (1 - t)(Z_0^{tar} - Z_0^{src})$$

$$\begin{aligned} V_t^{inv} &= \frac{dZ_t^{inv}}{dt} = -Z_0^{tar} + Z_0^{src} \\ &= V^{tar}(Z_t^{tar}, t) - V^{src}(Z_t^{src}, t) \end{aligned}$$

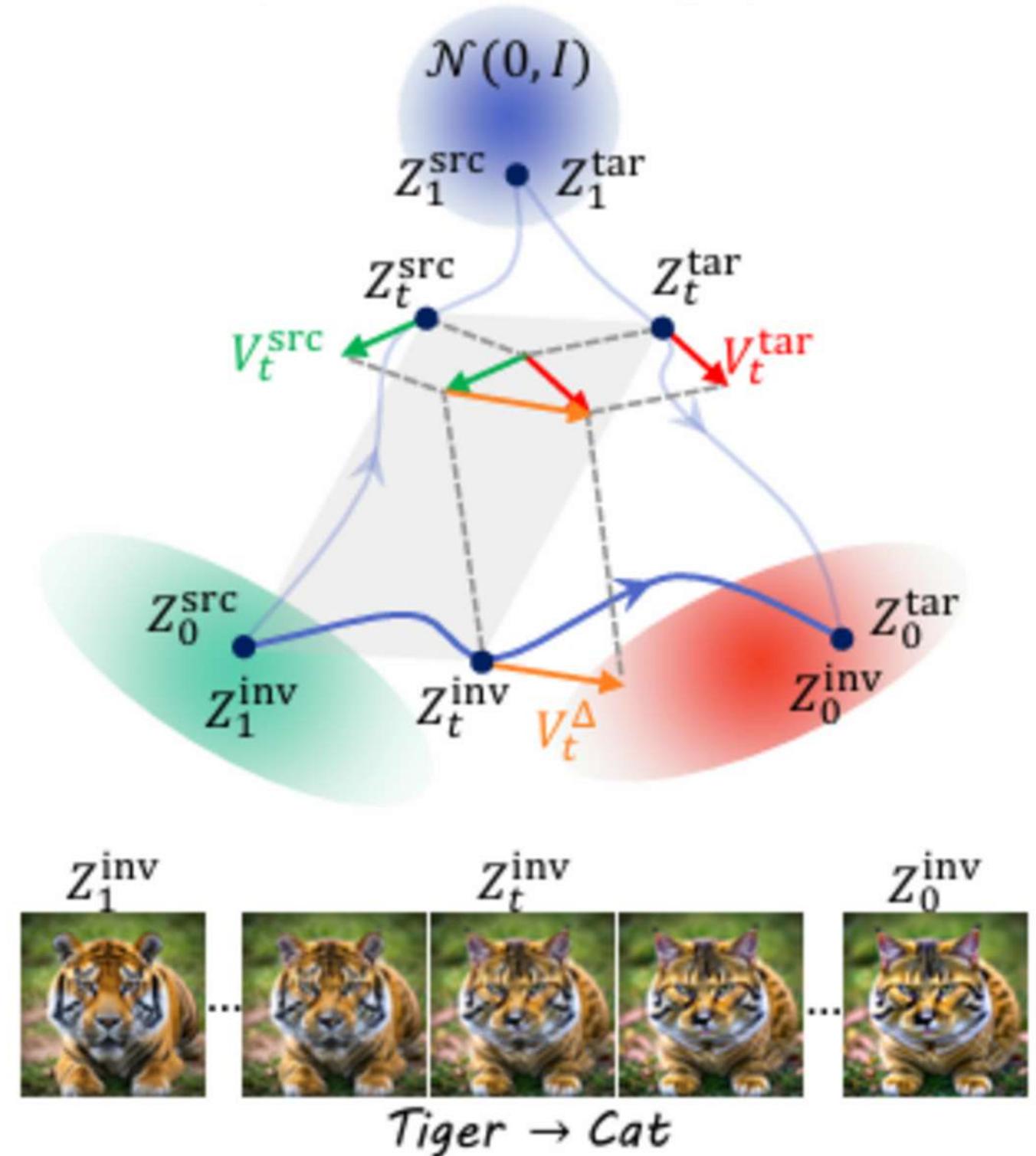
(b) Reinterpretation of Editing by Inversion



Reinterpretation of ODE Inversion

- Dù định nghĩa con đường mới nhưng trên thực tế khi tính toán ta vẫn thực hiện hai bước Inversion và Generation
- Con đường mới phụ thuộc hoàn toàn vào nhiễu ban đầu của ảnh nguồn

(b) Reinterpretation of Editing by Inversion



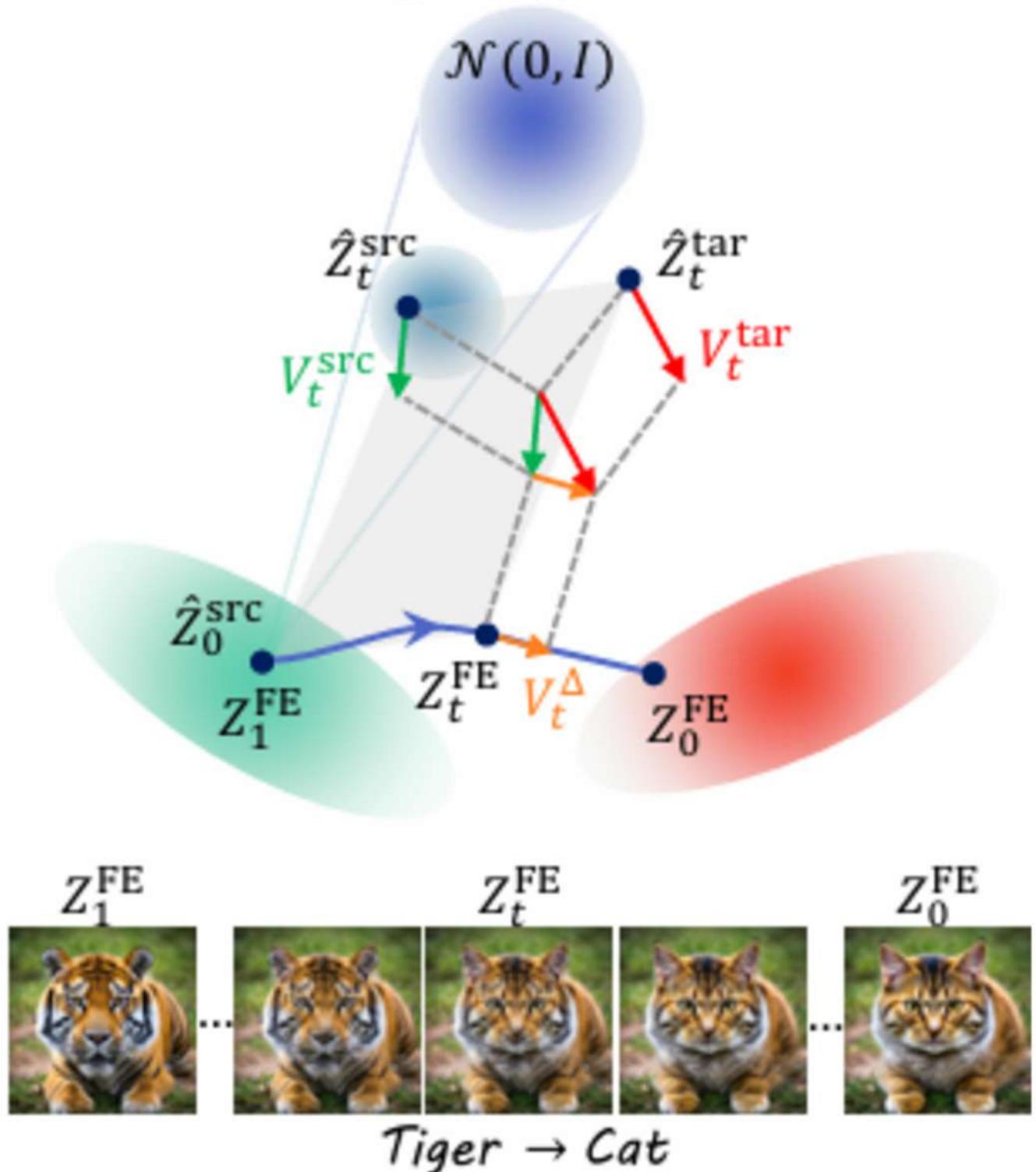
- Miễn là hai ảnh Z_t^{src}, Z_t^{tar} có cùng lượng nhiễu thì khi trừ đi sẽ chỉ còn lại thông tin thay đổi cấu trúc

$$Z_t^{inv} = Z_0^{src} + Z_t^{tar} - Z_t^{src}$$

$$V_t^{inv} = V^{tar}(Z_t^{tar}, t) - V^{src}(Z_t^{src}, t)$$

- Ở mỗi bước, FlowEdit sẽ lấy ngẫu nhiên nhiễu bất kỳ để thực hiện tính toán

(c) Editing Using FlowEdit



FlowEdit

- Khởi tạo: $Z_1^{FE} = Z_0^{src}$
- Ở mỗi bước:

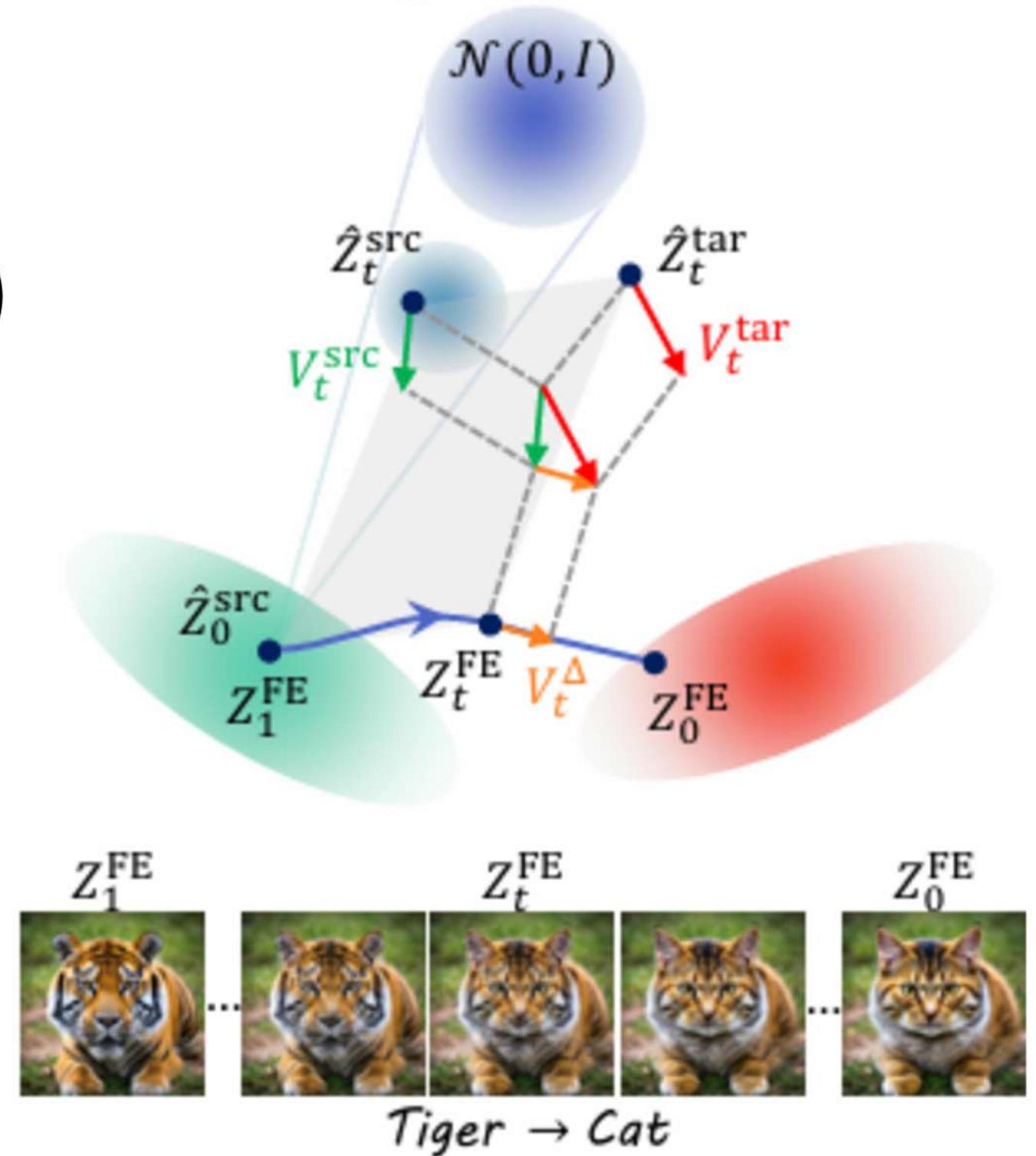
$$Z_t^{src} = (1 - t)Z_0^{src} + tN, N \sim N(0, 1)$$

$$Z_t^{tar} = Z_t^{FE} + Z_t^{src} - Z_0^{src}$$

$$V_t^{FE} = V^{tar}(Z_t^{tar}, t) - V^{src}(Z_t^{src}, t)$$

$$Z_s^{FE} = Z_t^{FE} + (s - t)V_t^{FE}$$

(c) Editing Using FlowEdit

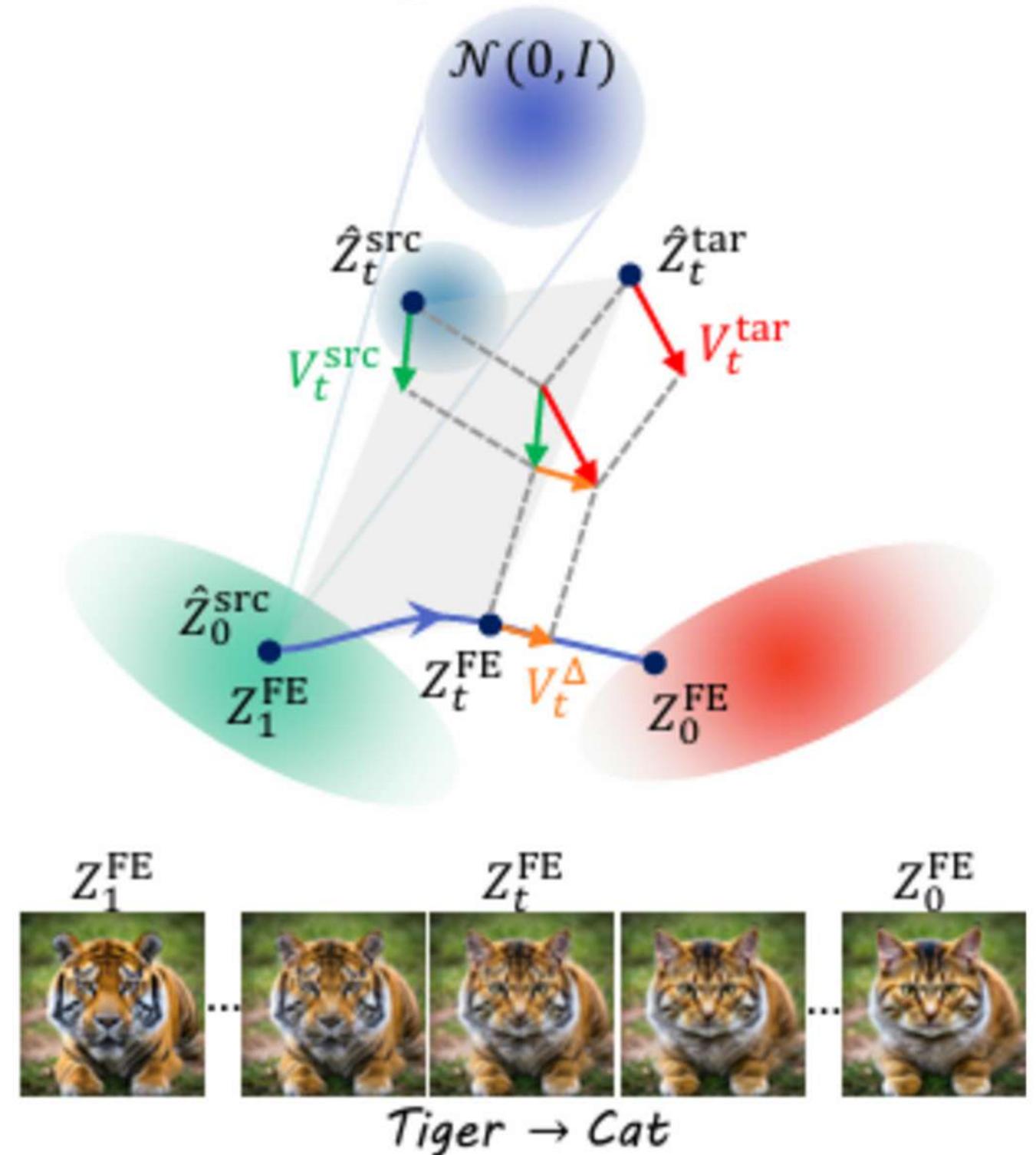


- Khi chạy thực tế, có một số tham số cần lưu ý
- n_{max} : Điểm bắt đầu can thiệp
- n_{min} : Điểm kết thúc can thiệp
- n_{avg} : Số lượng mẫu trung bình
- CFG Scale:

$$V^{src} = V_{uncond}^{src} + \alpha_{src} (V_{cond}^{src} - V_{uncond}^{src})$$

$$V^{tar} = V_{uncond}^{tar} + \alpha_{tar} (V_{cond}^{tar} - V_{uncond}^{tar})$$

(c) Editing Using FlowEdit



FlowEdit

Input: real image X^{src} , $\{t_i\}_{i=0}^T$, n_{\max} , n_{\min} , n_{avg}
Output: edited image X^{tar}

Init: $Z_{t_{\max}}^{\text{FE}} = X_0^{\text{src}}$

for $i = n_{\max}$ **to** $n_{\min+1}$ **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

if $n_{\min} = 0$ **then**

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

else

$N_{n_{\min}} \sim \mathcal{N}(0, 1)$

$Z_{t_{n_{\min}}}^{\text{src}} \leftarrow (1 - t_{n_{\min}})X^{\text{src}} + t_{n_{\min}} N_{n_{\min}}$

$Z_{t_{n_{\min}}}^{\text{tar}} \leftarrow Z_{t_{n_{\min}}}^{\text{FE}} + Z_{t_{n_{\min}}}^{\text{src}} - X^{\text{src}}$

for $i = n_{\min}$ **to** 1 **do**

$Z_{t_{i-1}}^{\text{tar}} \leftarrow Z_{t_i}^{\text{tar}} + (t_{i-1} - t_i)V^{\text{tar}}(Z_{t_i}^{\text{tar}}, t_i)$

end for

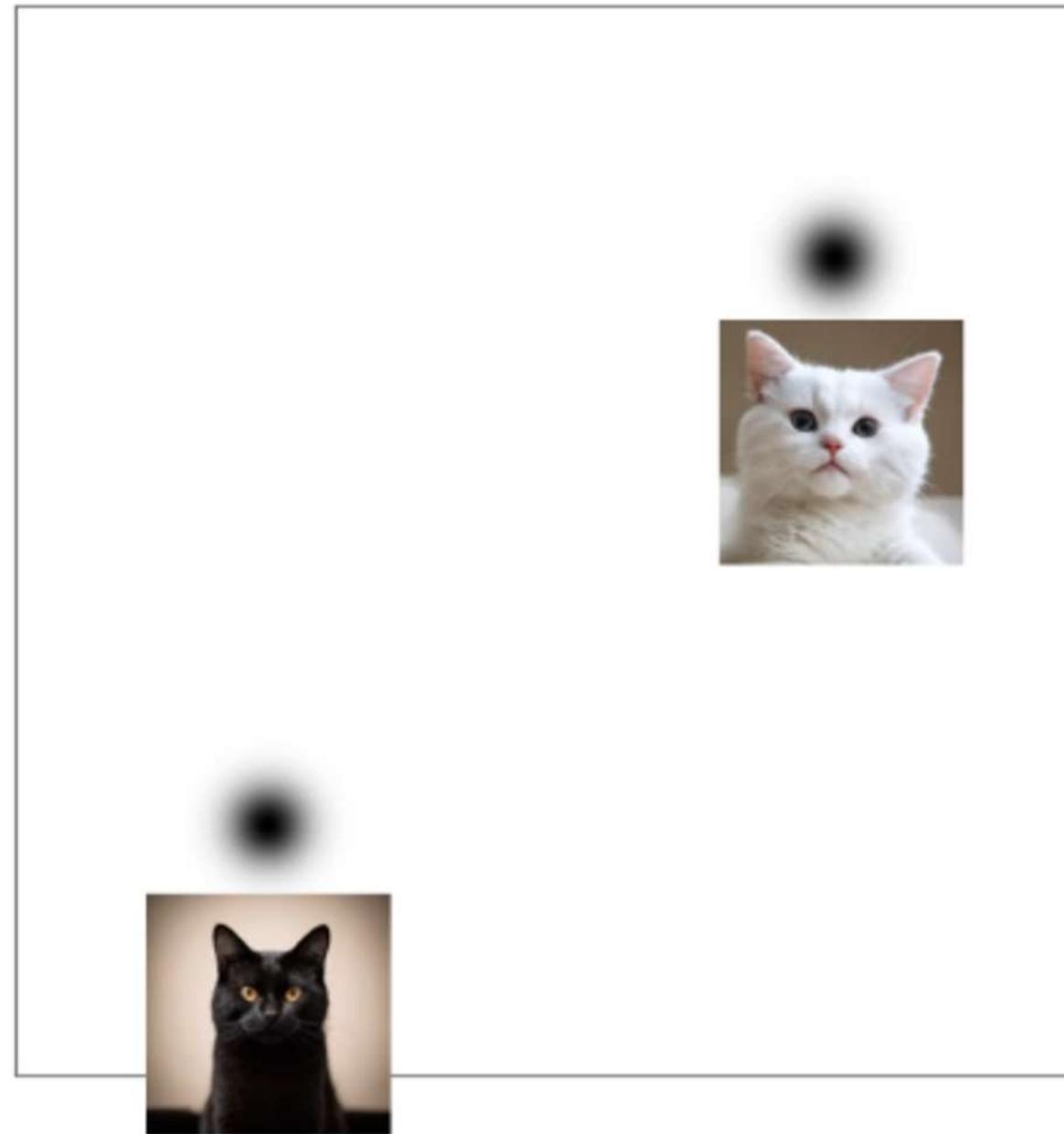
Return: $Z_0^{\text{tar}} = X_0^{\text{tar}}$

end if

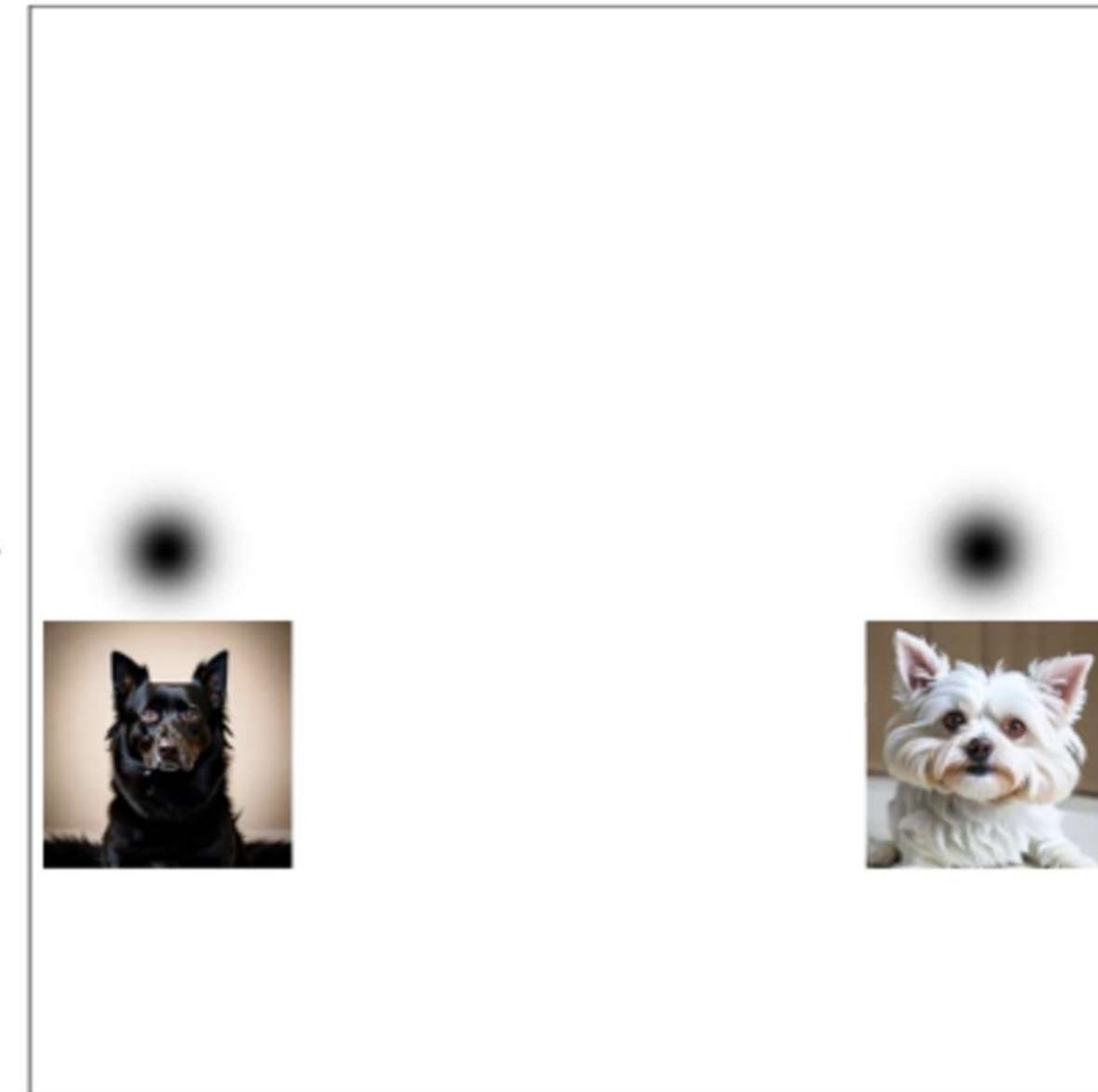
} Optionally average n_{avg} samples



Source Distribution

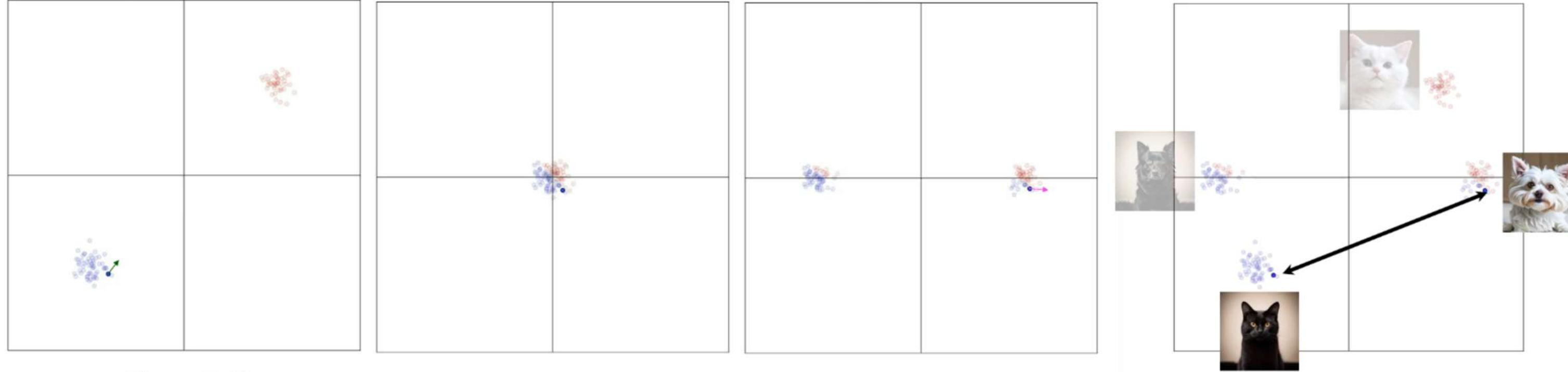


Target Distribution



FlowEdit

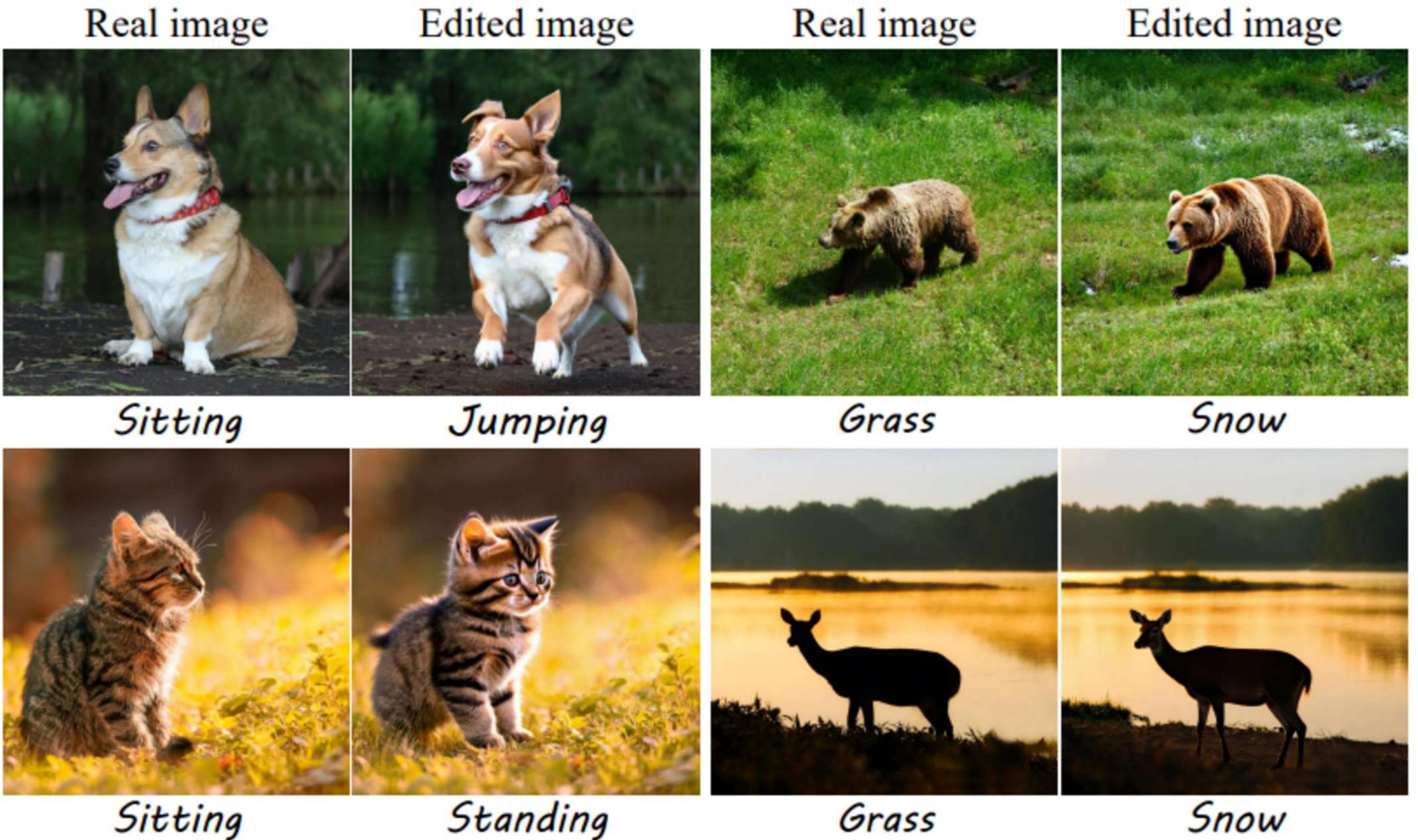
Editing by Inversion



FlowEdit



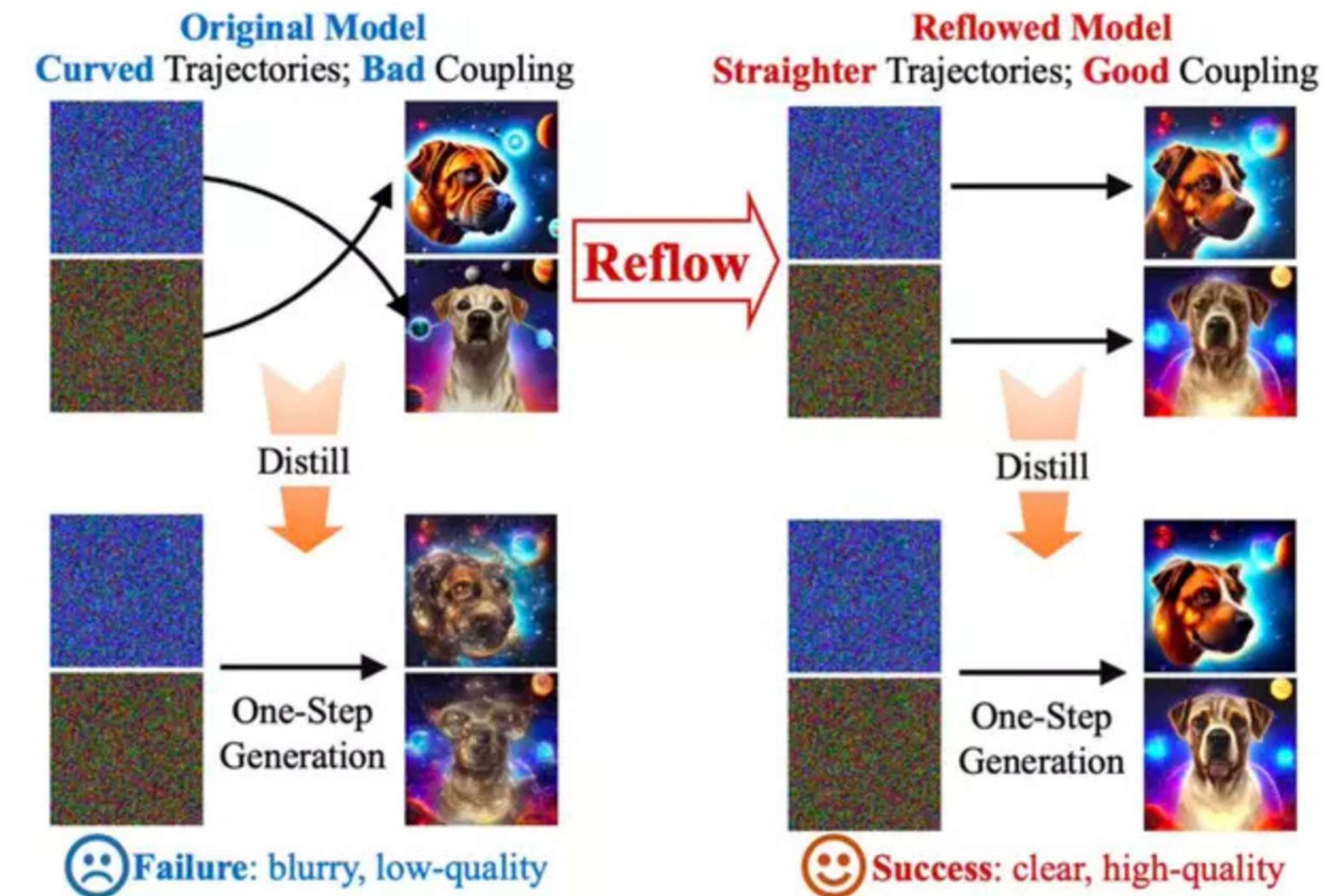
FlowEdit



- FlowEdit được thiết kế đặc biệt cho Rectified Flow Models chứ không phải các mô hình Diffusion cổ điển
- Nguyên nhân chính đến từ vector vận tốc:
 - Với Rectified Flow Models: đường đi nằm trên siêu phẳng, vector vận tốc là đường thẳng và giá trị là hằng
 - Với Diffusion Models cổ điển: đường đi nằm trên siêu cầu, vector vận tốc là vector tiếp tuyến của siêu cầu và giá trị thay đổi liên tục

InstaFlow

- InstaFlow áp dụng lý thuyết Rectified Flow để "ép" mô hình SD1.5 thành quỹ đạo thẳng



- InstaFlow gồm hai loại:
 - Few-step: “XCLiu/2_rectified_flow_from_sd_1_5”
 - Mô hình SD1.5 được nắn thẳng hai lần
 - Giai đoạn sampling chưa đến 20 bước
 - One-step: “XCLiu/instaflow_0_9B_from_sd_1_5”
 - Mô hình học ánh xạ trực tiếp nhiễu sang ảnh
 - Giai đoạn sampling đúng 1 bước

Ứng dụng FlowEdit vào InstaFlow

- Bộ dữ liệu nhóm tác giả FlowEdit:
 - 77 ảnh kích thước 1024x1024 lấy từ DIV2K và các nguồn miễn phí trực tuyến
 - Mỗi bức ảnh có 1 prompt nguồn và 1 hoặc nhiều prompt đích. Tổng cộng có hơn 200 cặp text-image



Ứng dụng FlowEdit vào InstaFlow

- Một vài lưu ý:

- Đầu ra mô hình: InstaFlow finetune từ SD1.5 nên đầu ra vẫn là nhiễu nên phải thêm bước chuyển đổi:

$$V_t = \sqrt{\alpha_t} Z_1 - \sqrt{1 - \alpha_t} Z_0$$

- Quy ước về thời gian và vận tốc của InstaFlow
 - SD3, FLUX:

$$Z_t = (1 - t)Z_0 + tZ_1 \Rightarrow V_t = Z_1 - Z_0$$

- InstaFlow:

$$Z_t = tZ_0 + (1 - t)Z_1 \Rightarrow V_t = Z_0 - Z_1$$



Ứng dụng FlowEdit vào InstaFlow

- Các chỉ số sử dụng để đánh giá kết quả:
 - CLIP-T (Càng cao càng tốt): Đo xem ảnh có khớp với văn bản mô tả không
 - CLIP-I (Càng cao càng tốt): Đo độ tương đồng giữa ảnh nguồn và ảnh đích
 - DINO (Càng cao càng tốt): Đo độ tương đồng về cấu trúc và bố cục
 - LPIPS (Càng thấp càng tốt): Đo sự khác biệt về thị giác như màu sắc, độ mờ, nhiễu
 - DreamSim (Càng thấp càng tốt): Đo độ tương đồng giữa hai ảnh giống mắt người nhìn nhất



Ứng dụng FlowEdit vào InstaFlow

- Các tham số:
 - $T_{steps} = 25$
 - $n_{avg} = 1$
 - $n_{min} = 1$
 - $n_{max} \in \{19, 21, 23, 25\}$
 - $src_{guidance} = 1.5$
 - $tar_{guidance} \in \{16.5, 18.5, 20.5\}$



Ứng dụng FlowEdit vào InstaFlow

A large brown bear → A large polar bear



Two golden retriever puppies → Two husky puppies



A clownfish swimming → A small sea turtle swimming



Ứng dụng FlowEdit vào InstaFlow

Target Guidance	CLIP-T	CLIP_I	DINO	LPIPS	DreamSim
16.5	0.207	0.878	0.750	0.159	0.252
18.5	0.208	0.867	0723	0.181	0.276
20.5	0.207	0.857	0.701	0.203	0.298



Ứng dụng FlowEdit vào InstaFlow

Model	CLIP-T	CLIP_I	DINO	LPIPS	DreamSim
InstaFlow	0.207	0.878	0.750	0.159	0.252
SD3.5	0.344	0.872	0.713	0.181	0.253
FLUX	0.337	0.875	0.682	0.223	0.252



Ứng dụng FlowEdit vào InstaFlow

CAFE → CVPR



Luna → Sol



LOVE IS ALL YOU NEED → ECCV IS ALL YOU NEED



A large, faint watermark of the HUST logo is visible across the entire background of the slide.

HUST

THANK YOU !