STRUCT Group Seminar

# FlowDirector: Training-Free Flow Steering for Precise Text-to-Video Editing

Guangzhao Li  Yanming Yang  Chenxi Song  Chi Zhang
AGI Lab, Westlake University  Central South University

arXiv 2506

Presenter: Wenshuo Gao

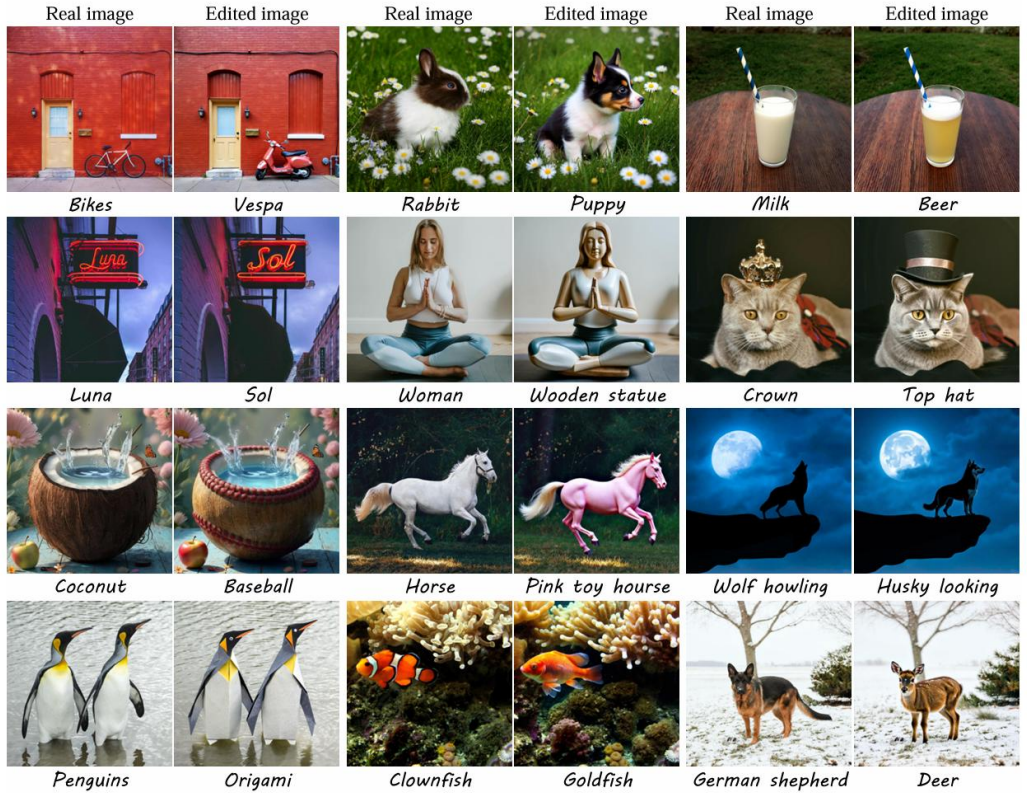
2025.07

# Content

- Author
- **Background**
- Method
- Experiments
- Conclusion

# Background: FlowEdit

## Training-Free Image Editing




Vladimir Kulikov, Matan Kleiner, Inbar Huberman-Spiegelglas, Tomer Michaeli, "FlowEdit: Inversion-Free Text-Based Editing Using Pre-Trained Flow Models", ICCV 2025

# Background: FlowEdit

## Flow models

Generation process defined by ODE:

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

Usually, $t \in [0,1]$ , $Z_1 \sim \mathcal{N}(0, \boldsymbol{I})$ and $Z_0$ distribute like real images

## Rectified Flow

A linear interpolation: $Z_t \sim (1-t)\, Z_0 + t\, Z_1\,.$

# Background: FlowEdit

## Inversion-based editing in flow models

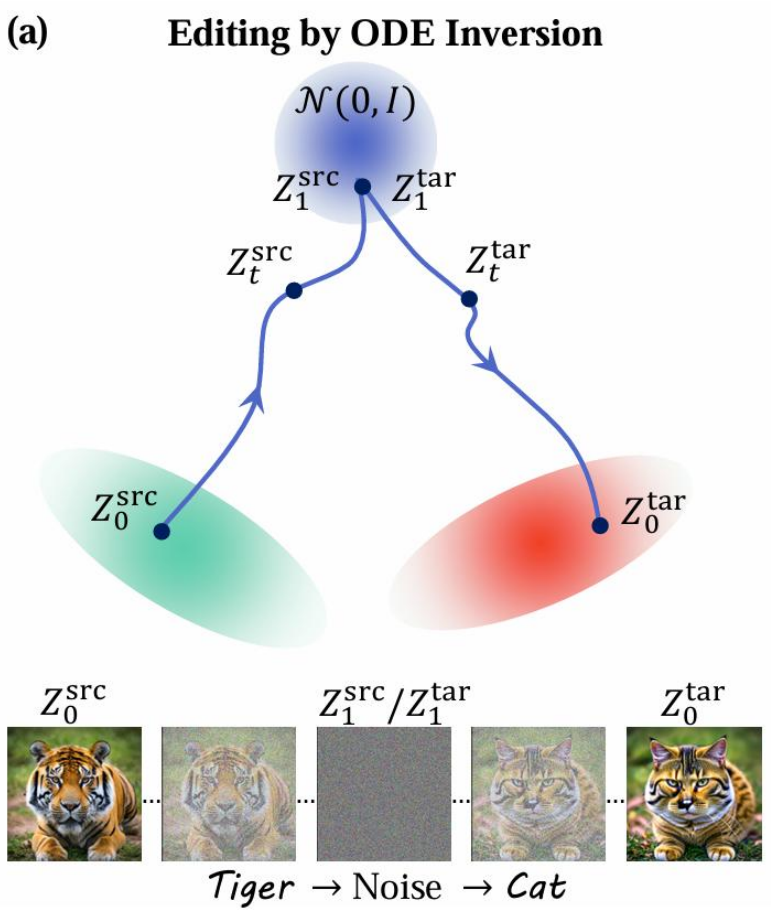
Traversing the forward process defined by ODE:

$$dZ_t^{\text{src}} = V^{\text{src}}(Z_t^{\text{src}}, t)\, dt,$$

And then, the reverse ODE:

$$dZ_t^{\text{tar}} = V^{\text{tar}}(Z_t^{\text{tar}}, t)\, dt,$$

is solved backward in time

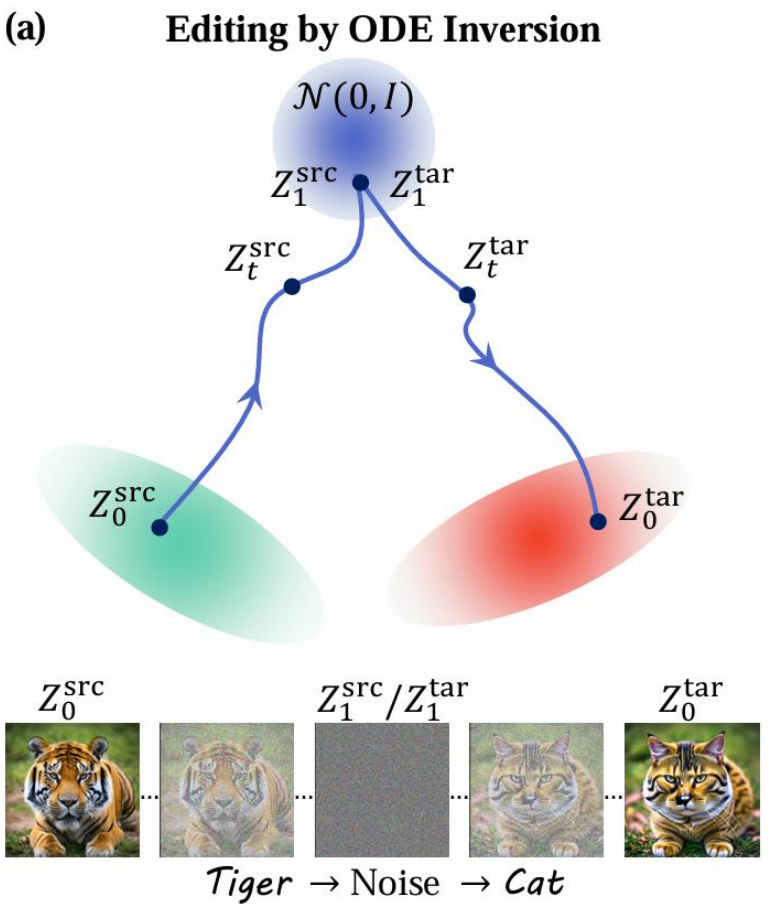


# Background: FlowEdit

## Why does inversion perform badly?

Drawbacks of Inversion:

- Inaccurate discretization
- Inaccurate even on perfect reconstruction
  - Prompt replacement affected generation process
  - Requires intervention (e.g. injecting feature maps)



# Background: FlowEdit

## Reinterpretation

Definition of a new path:

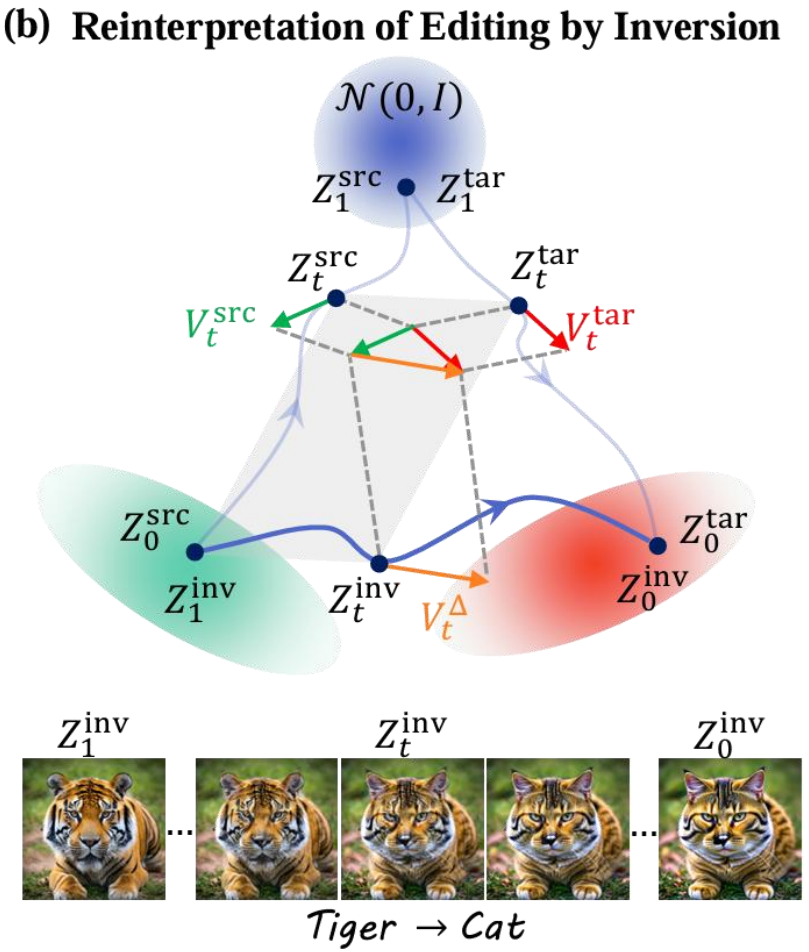
$$Z_t^{\text{inv}} = Z_0^{\text{src}} + Z_t^{\text{tar}} - Z_t^{\text{src}}.$$

New ODE becomes:

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

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

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

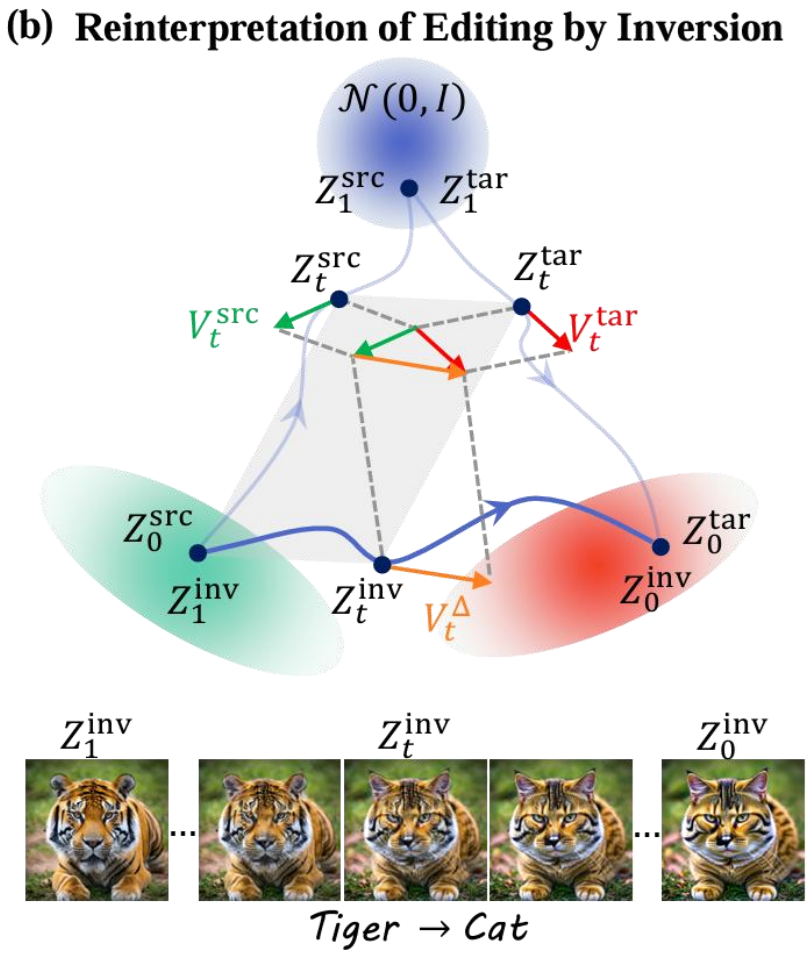


# Background: FlowEdit

## Reinterpretation

Properties of the new path:

- Noise-free path
- Coarse-to-fine evolution (why?)



# Background: FlowEdit

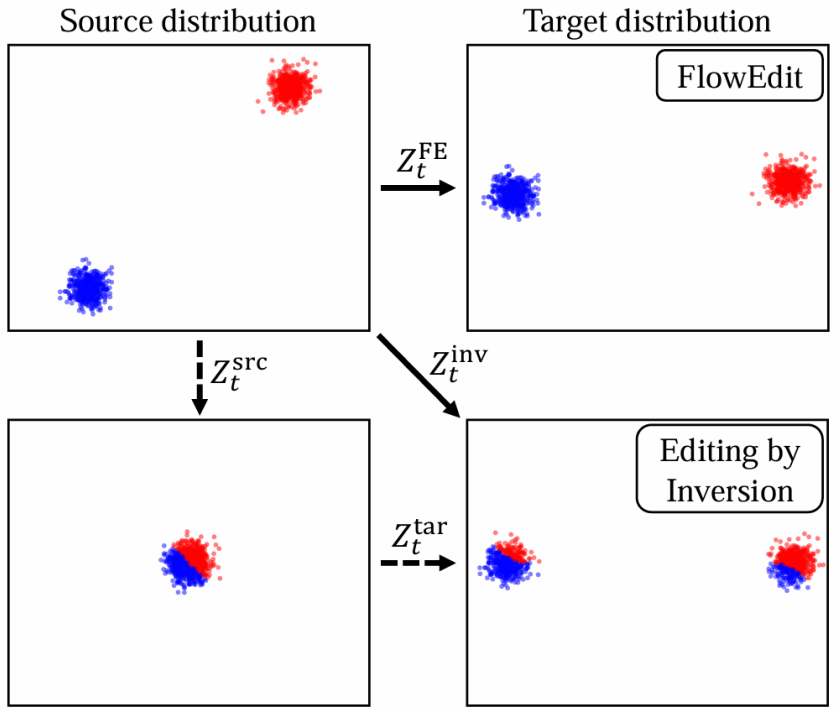
## Why does inversion perform badly?

An visualized explanation:

Two modes each in blue and red

**Expected:**
Map each mode in the source distribution
to its closest mode in the target distribution

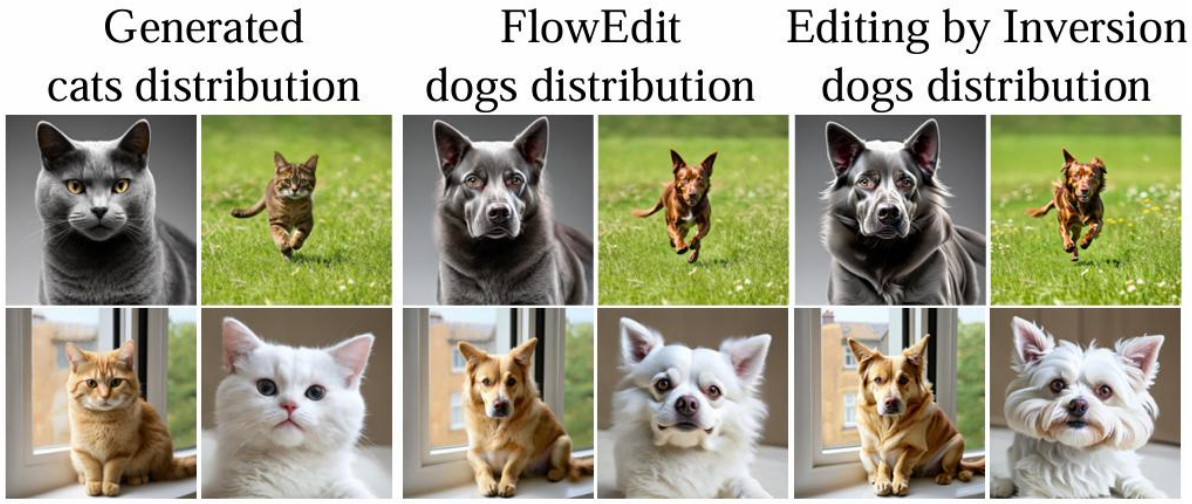
**Inversion:**
Mixed, each mapped to initial noise,
not the closest solution

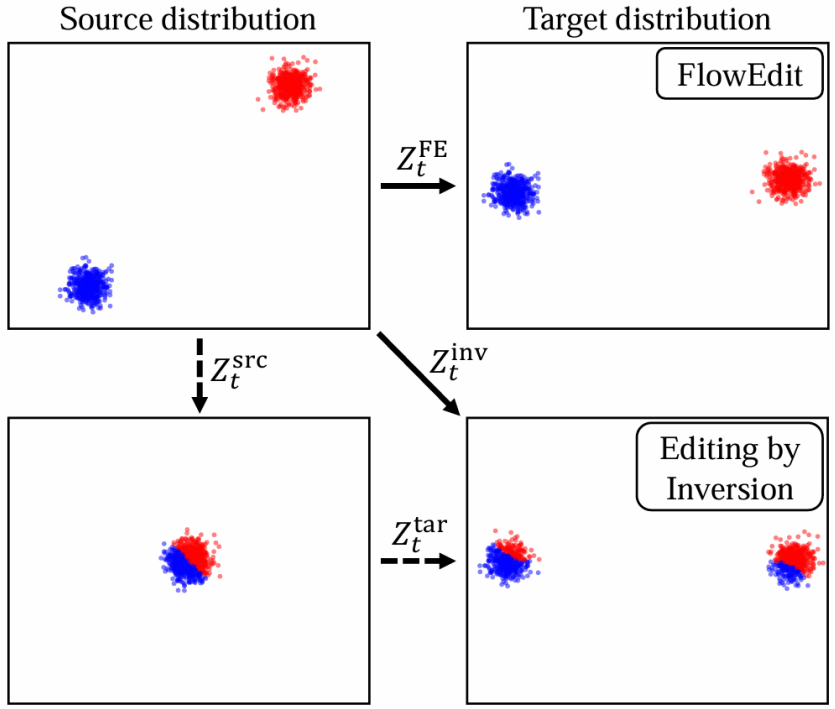


# Background: FlowEdit

## FlowEdit

How to construct the closest path?
(Least MSE of images)



# Background: FlowEdit

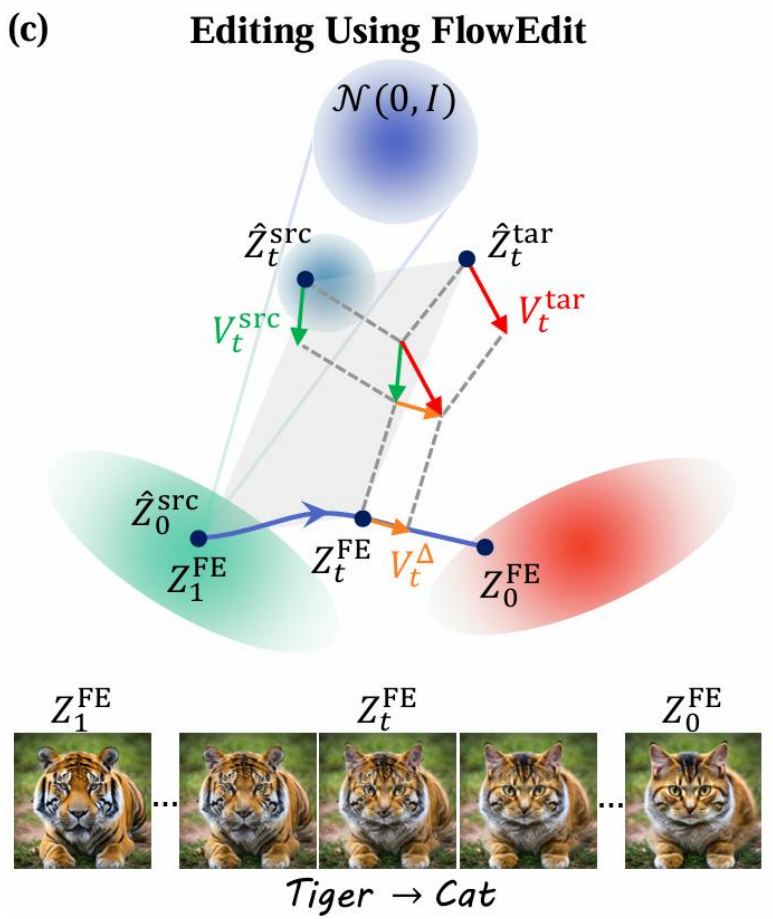
## FlowEdit

How to construct the closest path?
**Key idea:** calculating the expectation of evolution
Consider an alterative forward process:

$$\hat{Z}_t^{\text{src}} = (1-t)Z_0^{\text{src}} + tN_t,$$

where $N_t \sim \mathcal{N}(0,1)$ independently,
and construct a new path:

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



# Background: FlowEdit

## FlowEdit

Analysis:

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

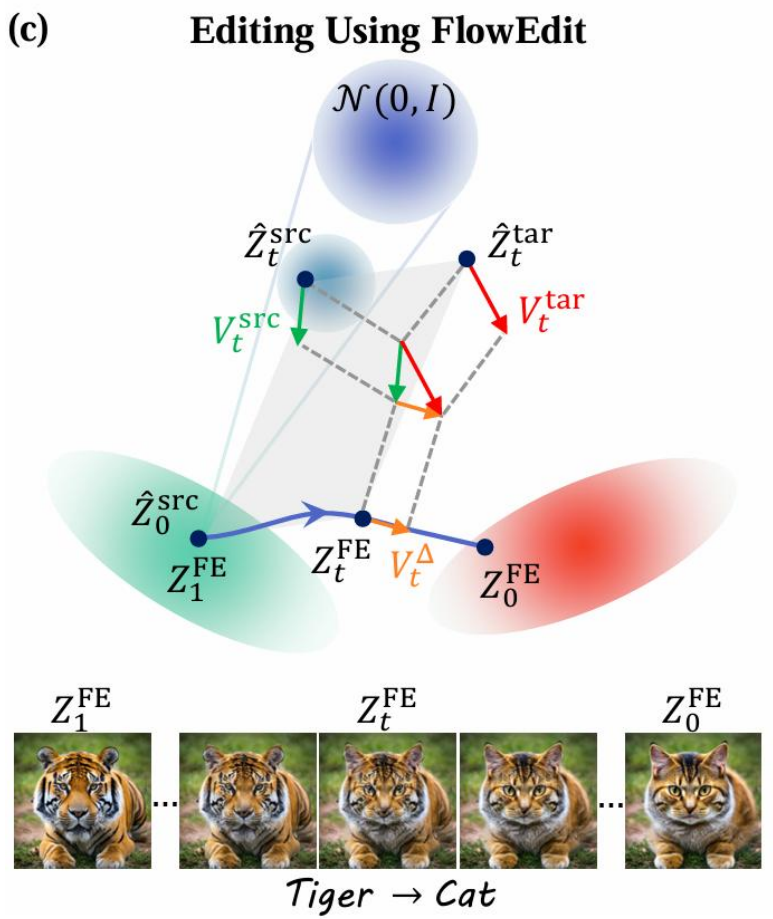
$$\hat{Z}_t^{\mathrm{tar}} = Z_t^{\mathrm{FE}} + \hat{Z}_t^{\mathrm{src}} - Z_0^{\mathrm{src}},$$

$$V_t^{\Delta}(\hat{Z}_t^{\mathrm{src}}, \hat{Z}_t^{\mathrm{tar}}) = V^{\mathrm{tar}}(\hat{Z}_t^{\mathrm{tar}}, t) - V^{\mathrm{src}}(\hat{Z}_t^{\mathrm{src}}, t)$$

Note that $\hat{Z}_t^{\mathrm{tar}}$ and $\hat{Z}_t^{\mathrm{src}}$ share the same noise $N_t$

When calculating difference:

- responce to this **common noise** will **cancel out**
- **semantic changes** will **remain**



# Background: FlowEdit

## FlowEdit

Analysis:

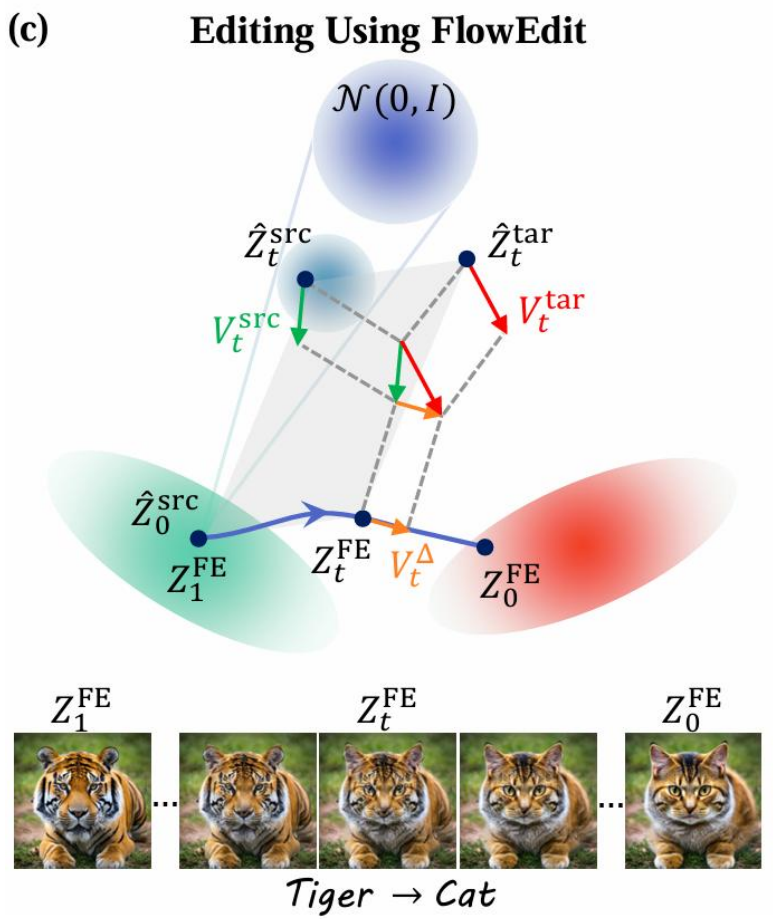
$$dZ_t^{\mathrm{FE}} = \mathbb{E}\left[V_t^{\Delta}(\hat{Z}_t^{\mathrm{src}}, Z_t^{\mathrm{FE}} + \hat{Z}_t^{\mathrm{src}} - Z_0^{\mathrm{src}}) \middle| Z_0^{\mathrm{src}}\right] dt$$

$$\hat{Z}_t^{\mathrm{tar}} = Z_t^{\mathrm{FE}} + \hat{Z}_t^{\mathrm{src}} - Z_0^{\mathrm{src}},$$

$$V_t^{\Delta}(\hat{Z}_t^{\mathrm{src}}, \hat{Z}_t^{\mathrm{tar}}) = V^{\mathrm{tar}}(\hat{Z}_t^{\mathrm{tar}}, t) - V^{\mathrm{src}}(\hat{Z}_t^{\mathrm{src}}, t)$$

Focusing on differences:

- FlowEdit can process images above generation ablitity
- Inversion method can not



# Background: FlowEdit

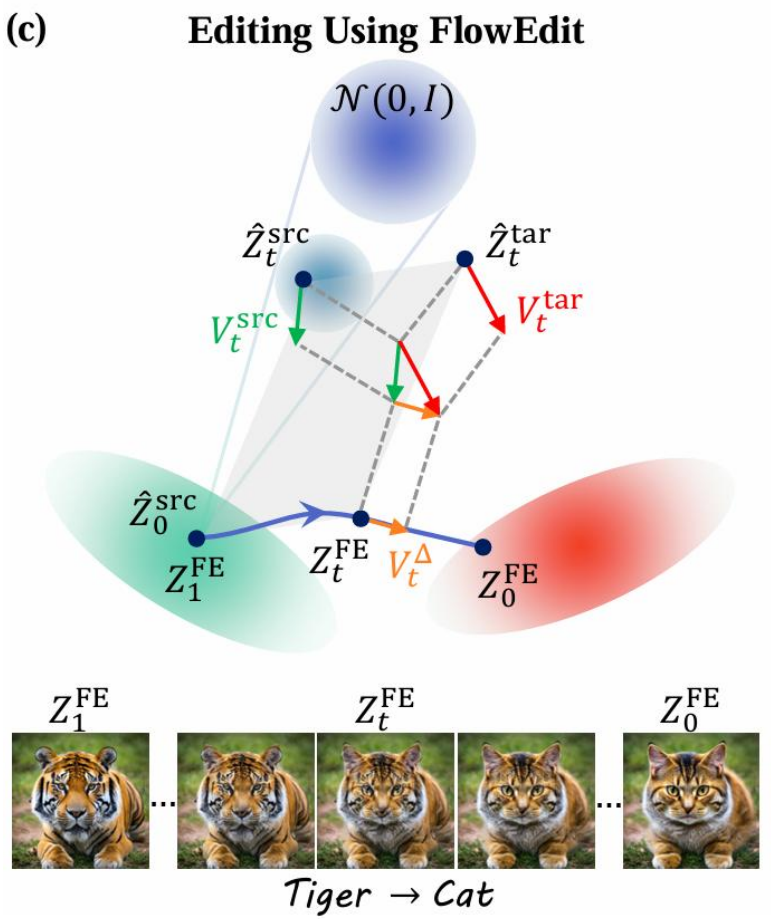
## FlowEdit

**Algorithm 1** Simplified algorithm for FlowEdit

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

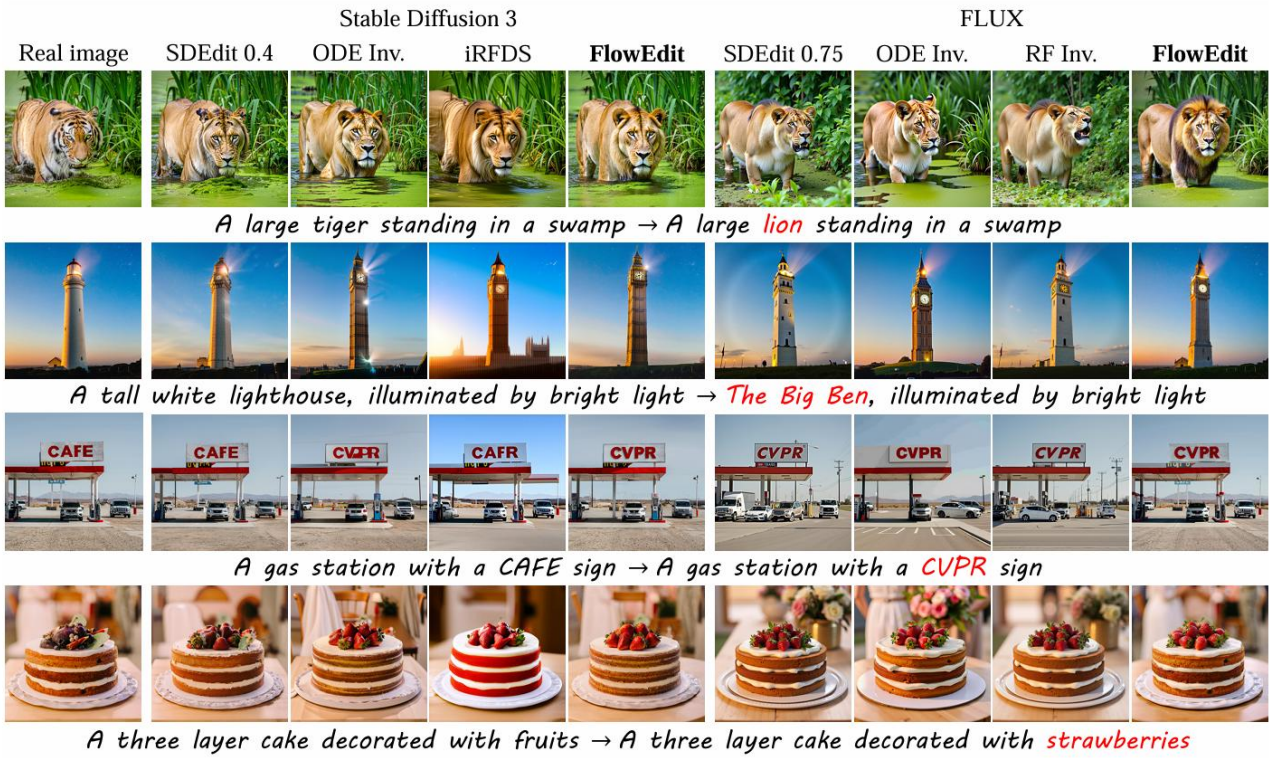
Optionally average $n_{\text{avg}}$ samples

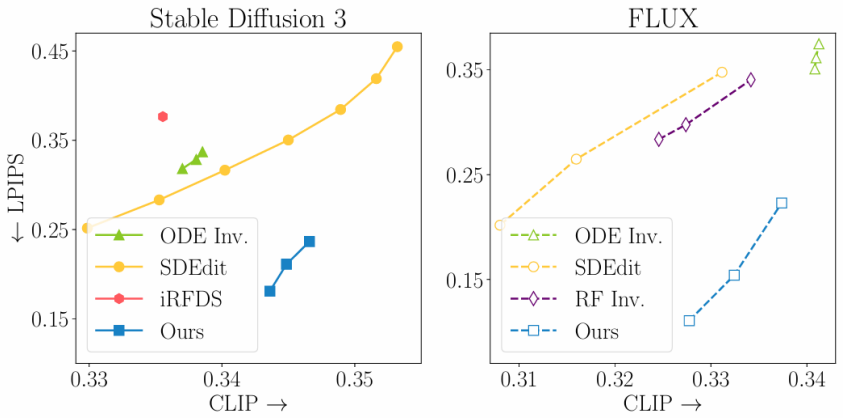
$n_{\text{max}} < T$, why?



# Background: FlowEdit

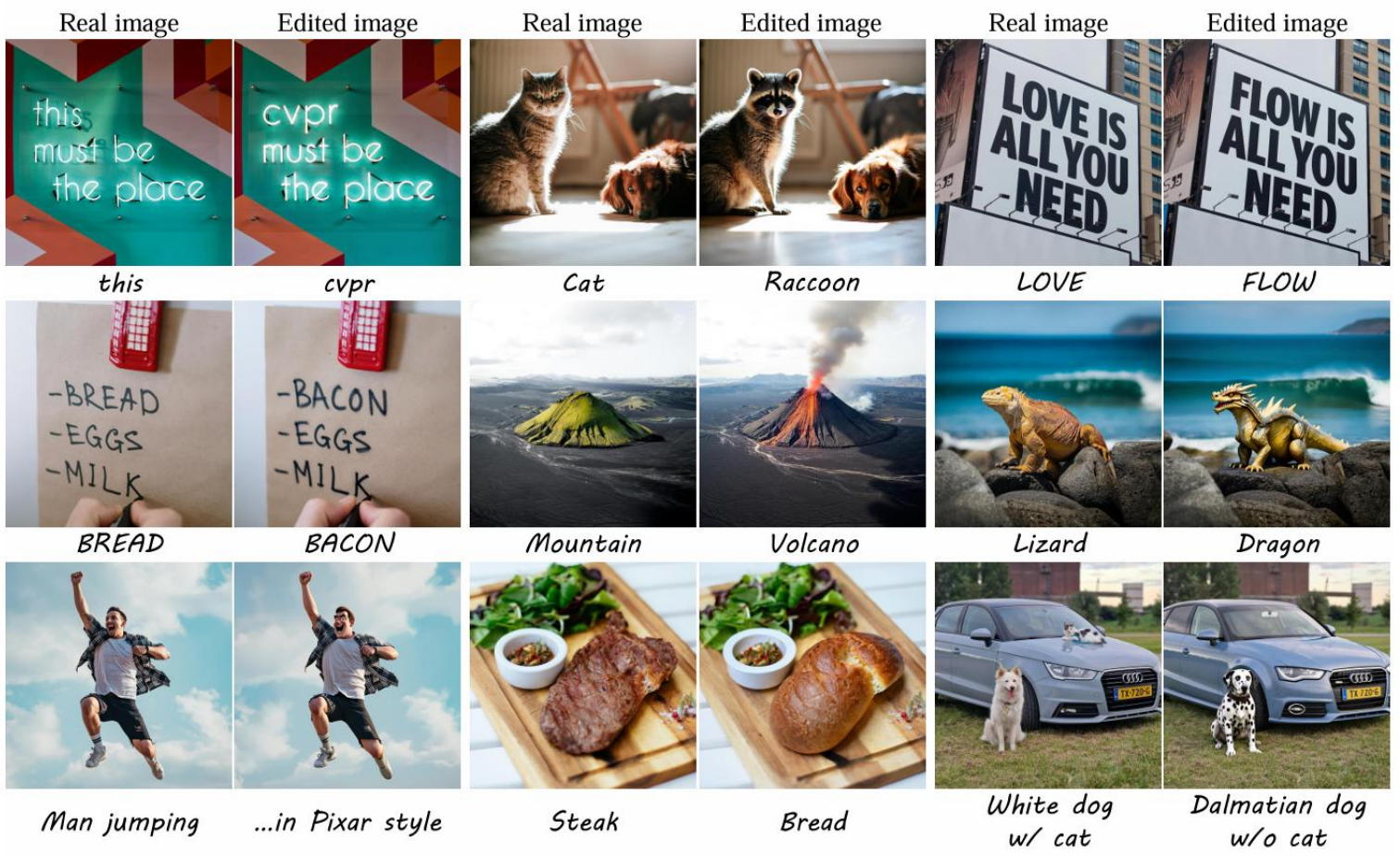
## FlowEdit



# Background: FlowEdit

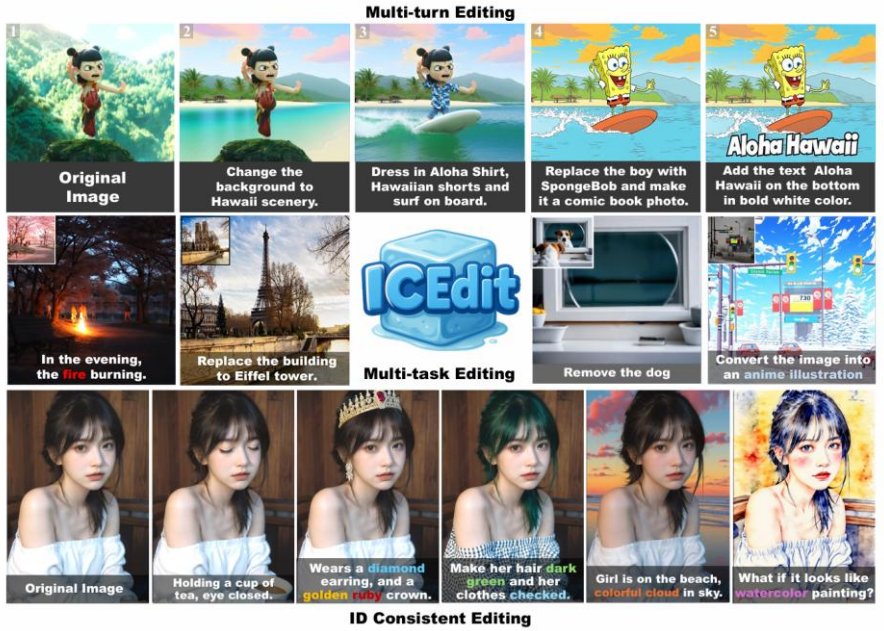
## FlowEdit



# Background: ICEdit

## ICEdit

Training-free / LoRA finetuning method for image editing

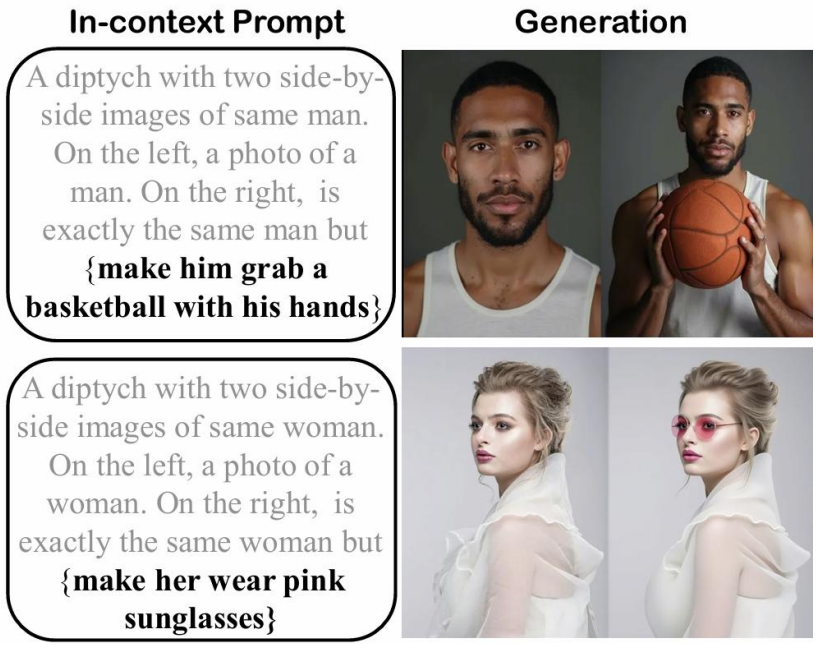



Zechuan Zhang, Ji Xie, Yu Lu, Zongxin Yang, Yi Yang, “In-Context Edit: Enabling Instructional Image Editing with In-Context Generation in Large Scale Diffusion Transformer”, arXiv 2504

# Background: ICEdit

## Insight: in-context prompt

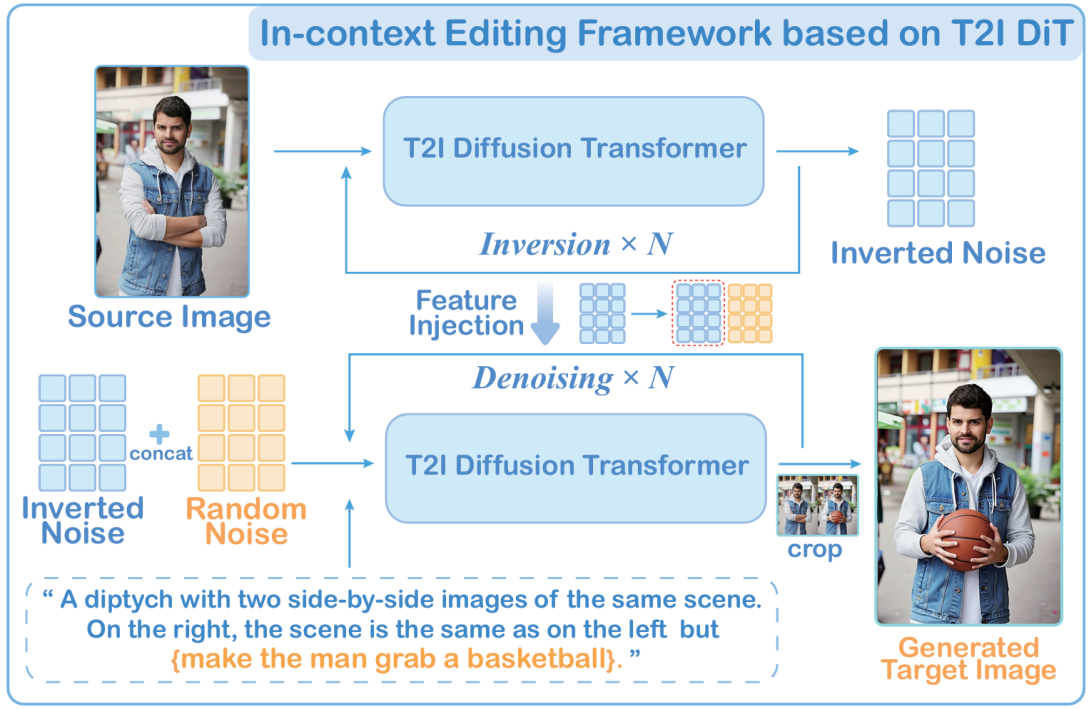
FLUX model is capable of generating ***side-by-side*** images



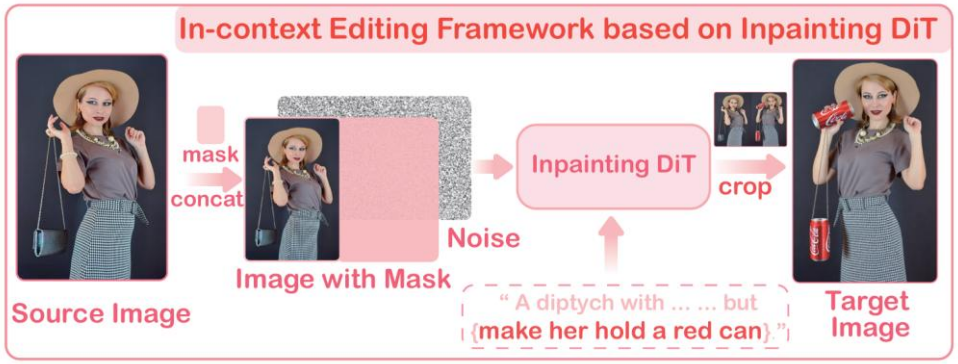
# Background: ICEdit

## Insight: in-context prompt

Editing using ***side-by-side*** images generation




inversion-based

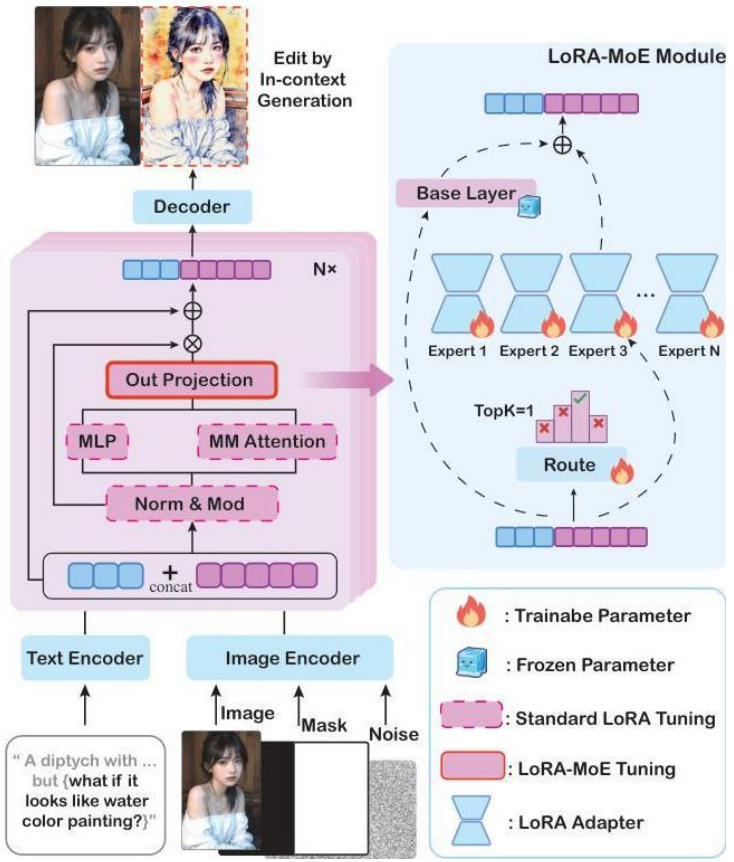



inpainting-based

# Background: ICEdit

## LoRA finetuning on inpainting method

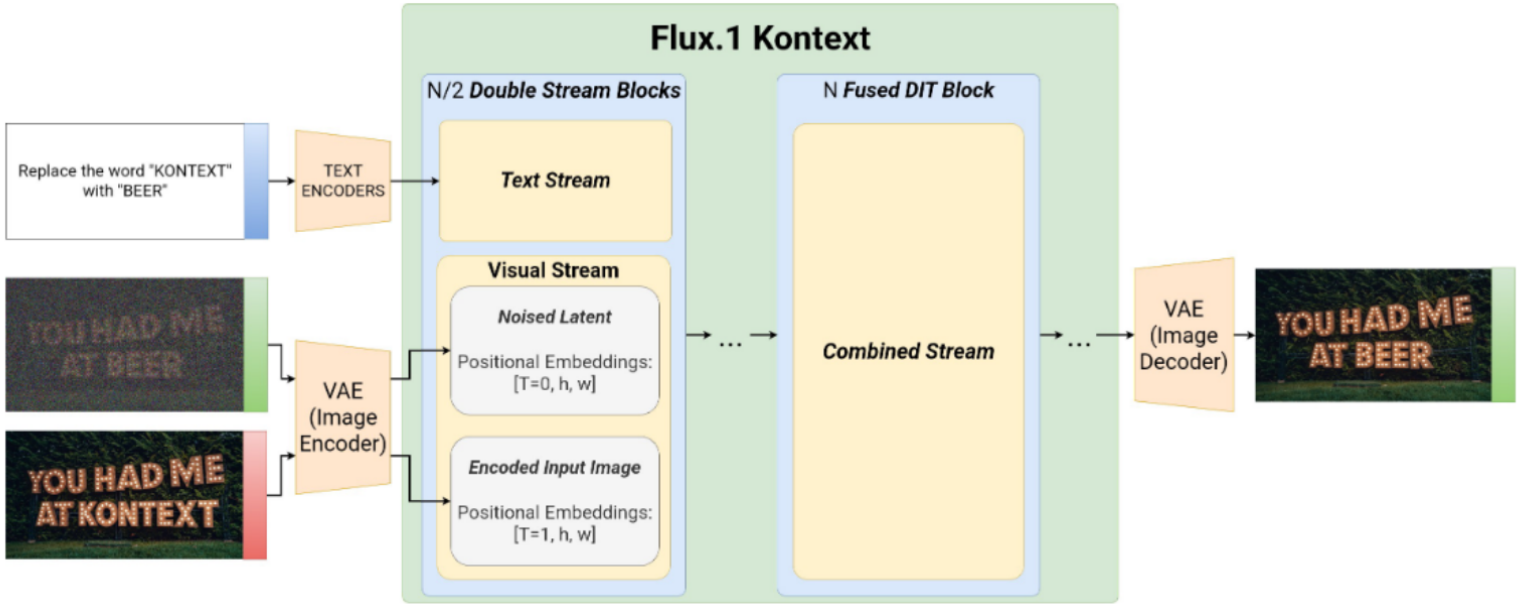
MoE: each expert for one kind of editing task



# Background: FLUX.1 Kontext

## FLUX.1 Kontext

SOTA diffusion-based image editing method (not training-free)




Black Forest Labs, "FLUX.1 Kontext: Flow Matching for In-Context Image Generation and Editing in Latent Space", arXiv 2506

## FLUX.1 Kontext

Character remain unchanged, the cars in the background remain unchanged, change the car‘s color to red.

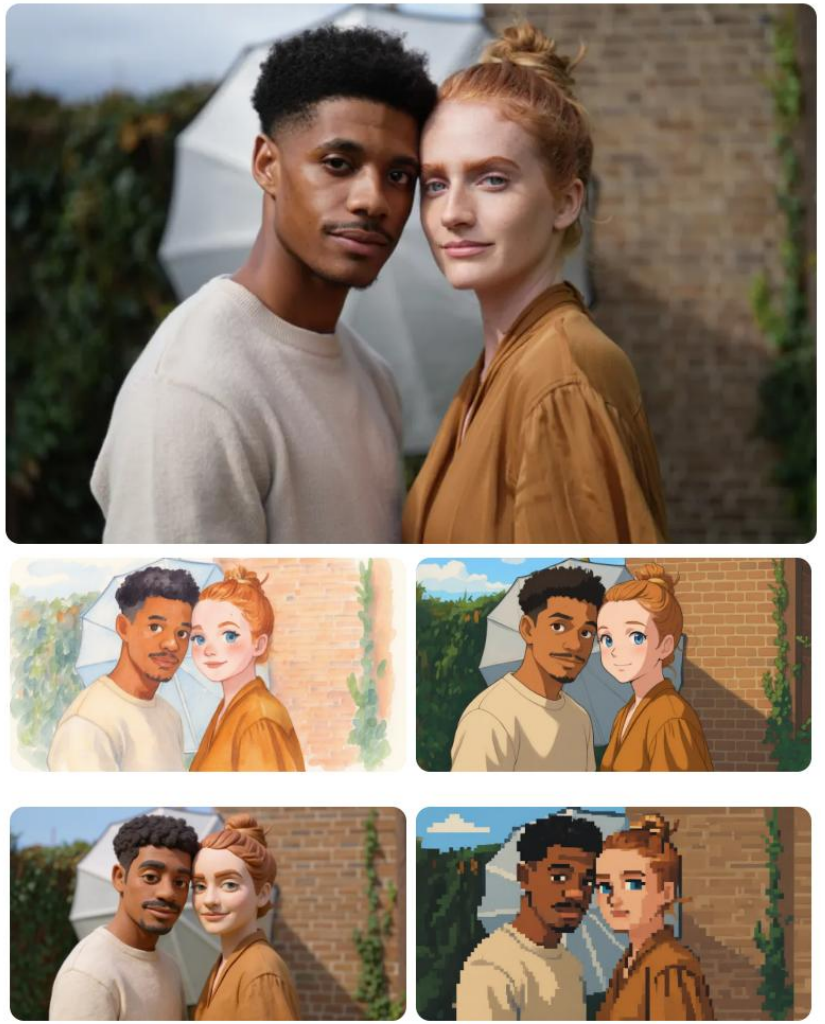
# Background: FLUX.1 Kontext

## FLUX.1 Kontext


Replace the toothbrush in the hand of the main character on the left with a microphone, making the character look like they are singing. Change the text content to "Singing, louder, more noise."

# Background: FLUX.1 Kontext

## FLUX.1 Kontext



# Background: Conclusion

## Comparisons of image editing methods

| Method | Training need | Core idea |
|---|---|---|
| FlowEdit | Training-free | Editing flow calculation |
| ICEdit | Training-free / LoRA | Side-by-side generation |
| FLUX.1 Kontext | Required training | Simple training |

## FlowDirector

Can **FlowEdit** transferred to video editing task?
**FlowEdit:** training-free, model-independent
Backbone model: FLUX -> Wan2.1

bear

pandas

dinosaur

## Editing Flow Generation

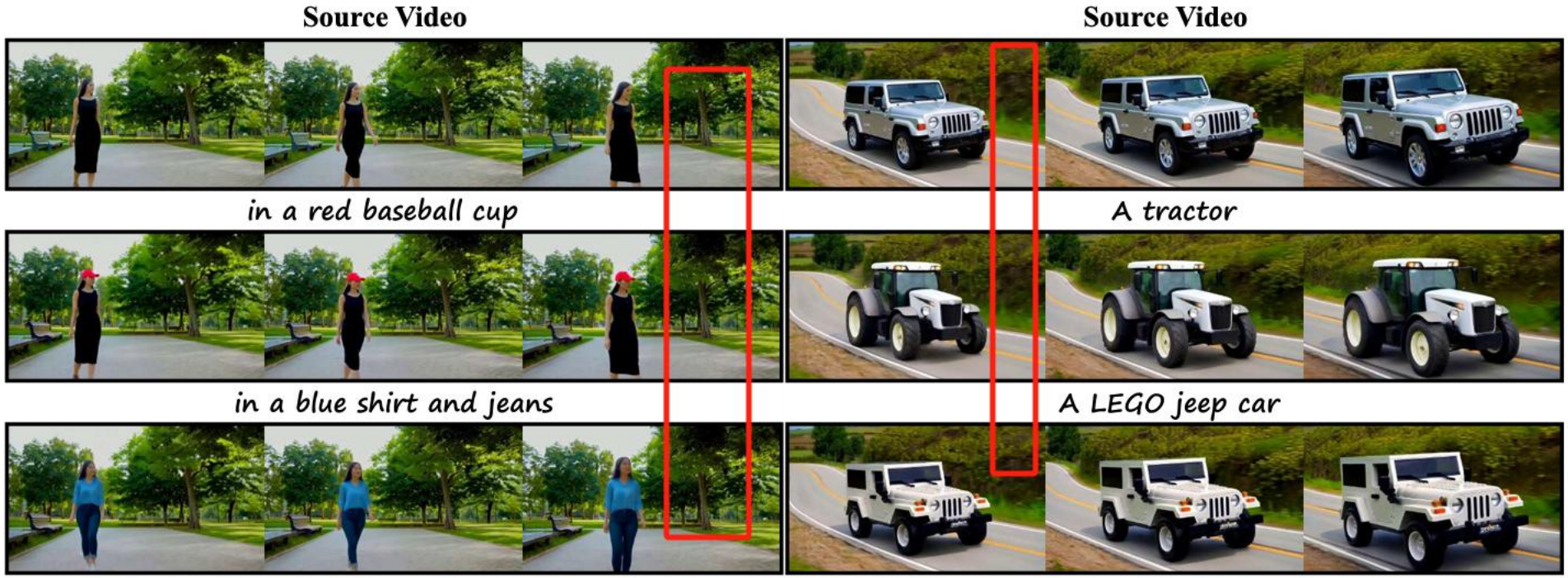
Identical equations to FlowEdit:

$$\frac{dZ_t^{\text{edit}}}{dt} = V_{\text{edit}}(t) = v_\theta(Z_t^{\text{tar}}, t, c_{\text{tar}}) - v_\theta(Z_t^{\text{src}}, t, c_{\text{src}}),$$

$$Z_t^{\text{edit}} = X_{\text{src}} - Z_t^{\text{src}} + Z_t^{\text{tar}}.$$

## Spatially Attentive Flow Correction

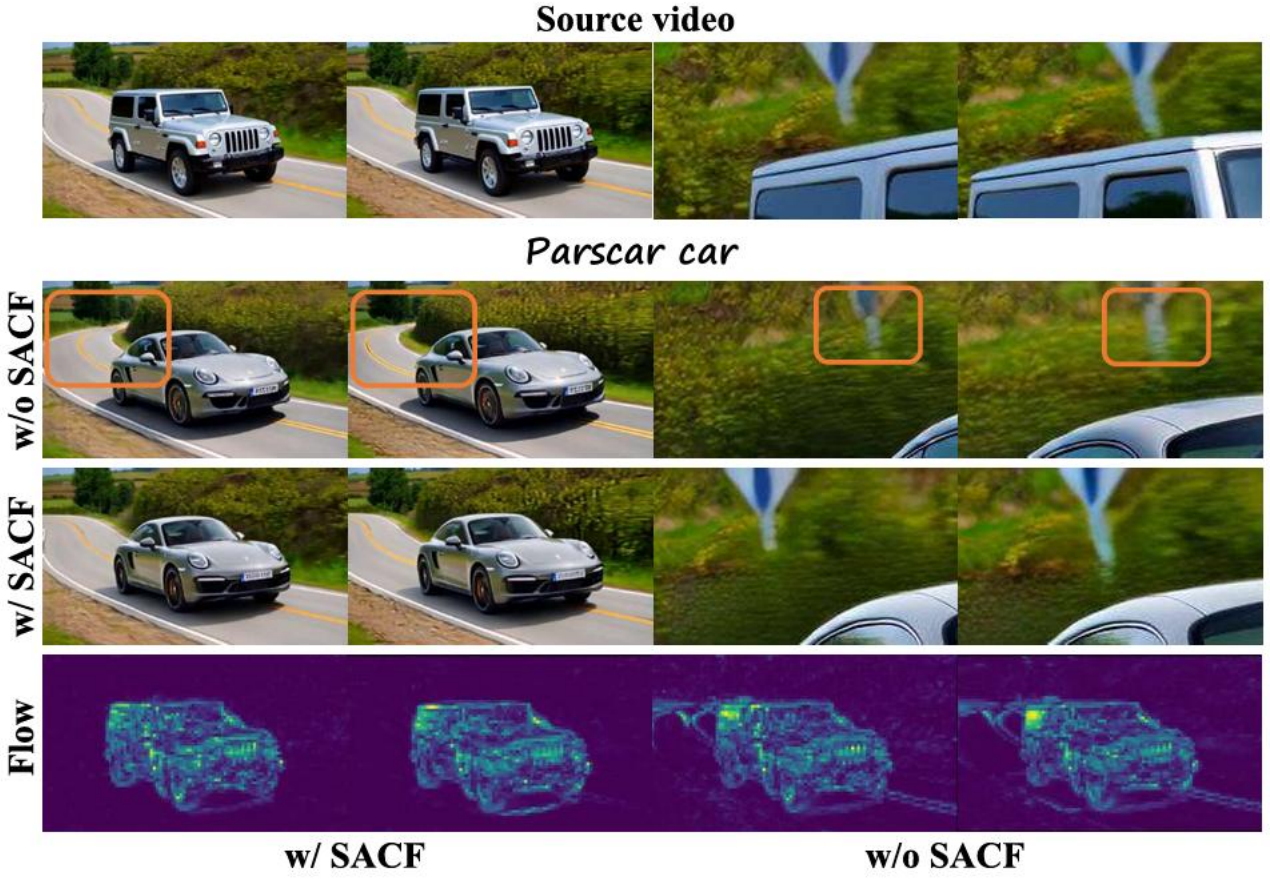
How to maintain non-edited regions?



## Spatially Attentive Flow Correction

Mask: cross-attention maps

$$\tilde{V}_{\text{edit}} = V_{\text{edit}} \odot (M_{src} \cup M_{tar}).$$



# Method

## Differential Averaging Guidance

Time cost: n_avg is expensive

Reducing time cost while enhancing quality:

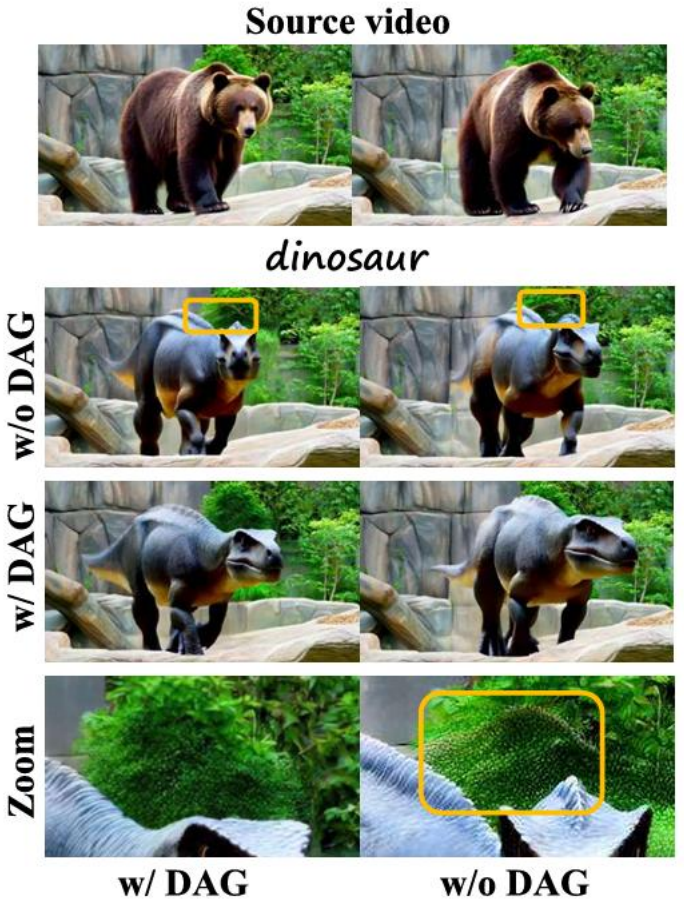
$$V_{\text{HQ}}(Z_t^{\text{edit}}, t) = \mathbb{E}_{N_t \sim \mathcal{N}(0,I);\, L_{\text{HQ}} \text{ samples}}[\tilde{V}_{\text{edit}}(Z_t^{\text{tar}}, Z_t^{\text{src}}, t, c^{\text{tar}}, c^{\text{src}})],$$

$$V_{\text{BL},i}(Z_t^{\text{edit}}, t) = \mathbb{E}_{N_t \sim \mathcal{N}(0,I);\, L_{\text{BL}} \text{ samples}}[\tilde{V}_{\text{edit}}(Z_t^{\text{tar}}, Z_t^{\text{src}}, t, c^{\text{tar}}, c^{\text{src}})]_i.$$

$$D_i = V_{\text{HQ}} - V_{\text{BL},i},$$

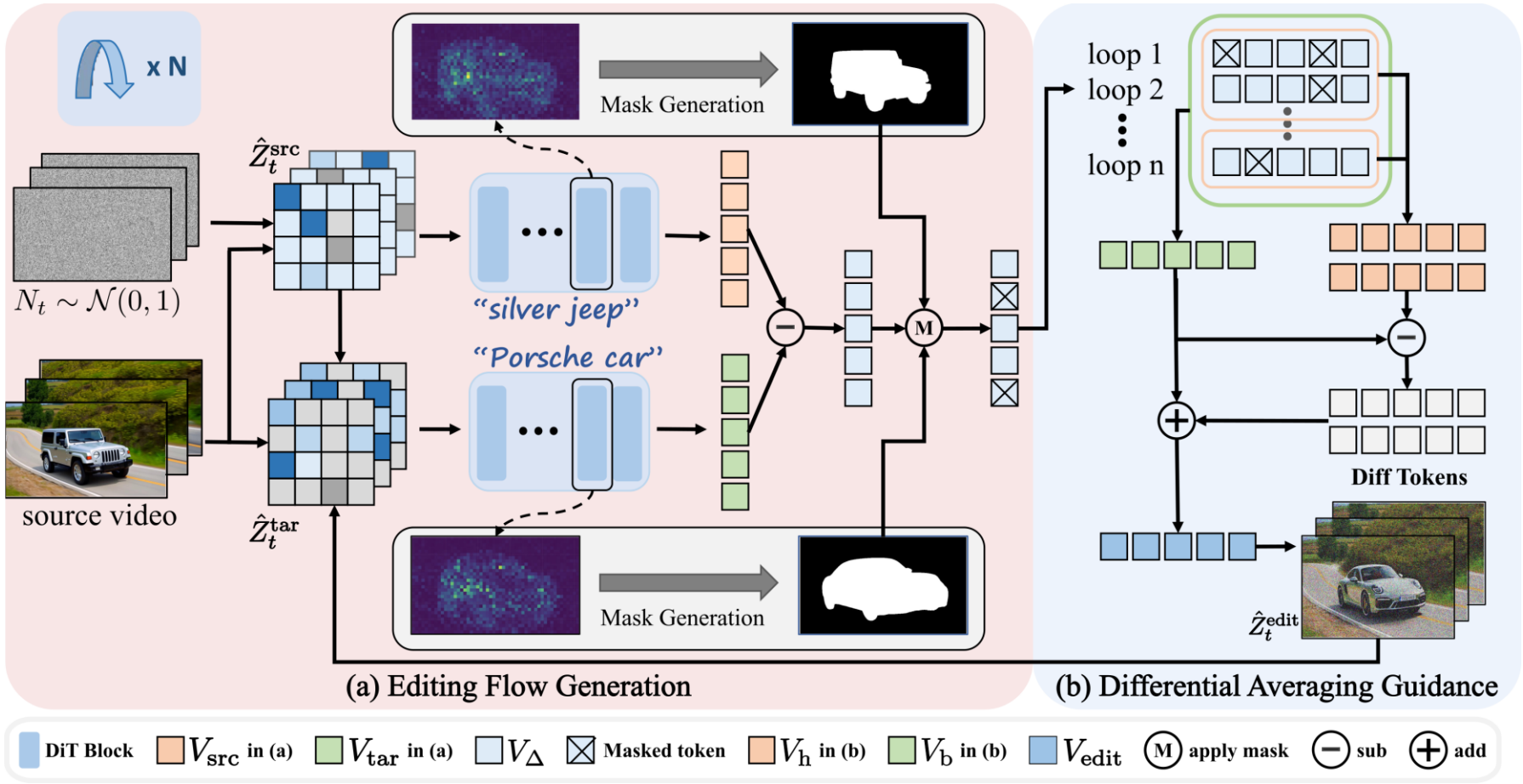
$$V_{\text{DAG}} = V_{\text{HQ}} + w \cdot \bar{D},$$

$$\bar{D} = \tfrac{1}{K} \textstyle\sum_{i=1}^{K} D_i.$$



# Method

## Full Framework



# Experiment

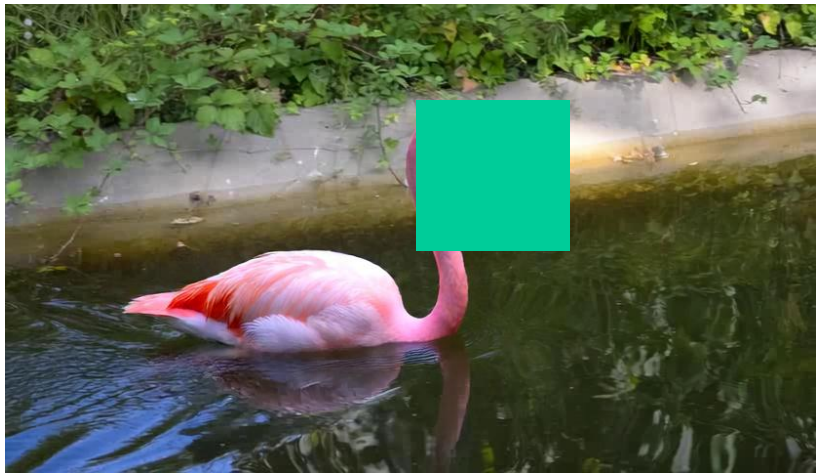
## Metrics

CLIP-T & Pick Score & Frame-Acc: Prompt Alignment
CLIP-F: Frame Consistency
WarpSSIM: Structural Preservation

| Method | Frame-Acc (%) ↑ | | | Pick Score (%) ↑ | | | CLIP-T ($\times 10^{-2}$) ↑ | | | CLIP-F ($\times 10^{-2}$) ↑ | | | WarpSSIM ($\times 10^{-2}$) ↑ | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| FateZero [33] | 45.83 | 46.67 | - | 21.01 | 20.98 | - | 32.13 | 32.06 | - | 96.83 | 96.61 | - | **78.88** | **78.50** | - |
| FLATTEN [6] | 75.00 | 74.29 | 73.96 | 20.93 | 20.88 | 20.88 | 33.45 | 33.45 | 33.50 | 96.54 | 96.54 | 96.20 | 77.47 | 77.84 | **77.60** |
| TokenFlow [8] | 74.22 | 71.88 | 72.83 | 21.42 | 21.42 | 21.40 | 32.96 | 32.96 | 32.81 | 97.04 | 97.03 | 96.85 | 75.62 | 75.42 | 75.19 |
| RAVE [17] | 73.39 | 72.58 | 72.67 | 20.99 | 20.98 | 20.99 | 33.33 | 33.25 | 33.25 | 97.13 | 97.14 | 96.91 | 77.01 | 77.04 | 76.61 |
| VideoDirector [48] | 77.82 | 72.87 | - | 20.76 | 20.60 | - | 32.93 | 32.57 | - | 95.40 | 91.19 | - | 76.67 | 75.88 | - |
| **FlowDirector (ours)** | **84.38** | **84.09** | **84.55** | **21.83** | **21.82** | **21.80** | **33.72** | **33.79** | **33.84** | **98.37** | **98.07** | **97.85** | 77.21 | 77.14 | 77.00 |

# Result

“black swan”



“flamingo”

# Result

“wolf”

“fox”

# Result

“seaturtle”

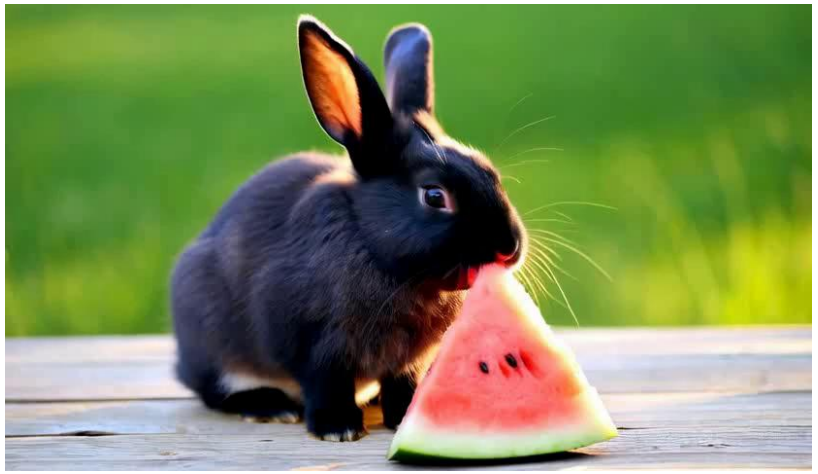
“dolphin”

# Result

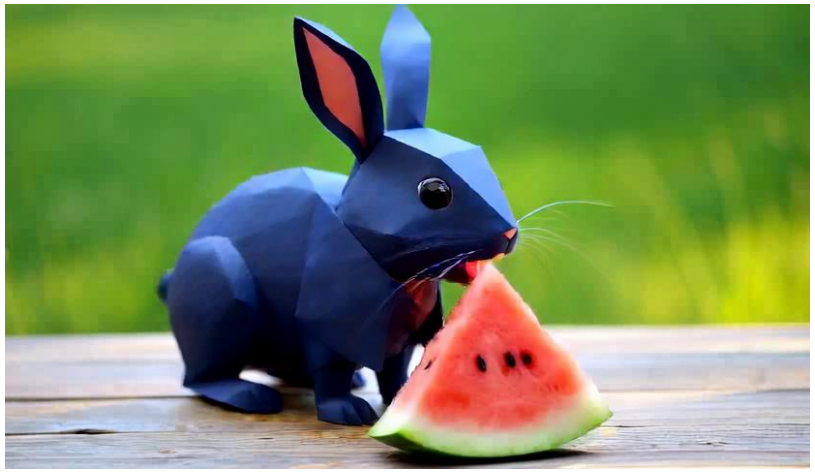
“dog with flower”

“dog”

# Result



“rabbit”



“origami rabbit”

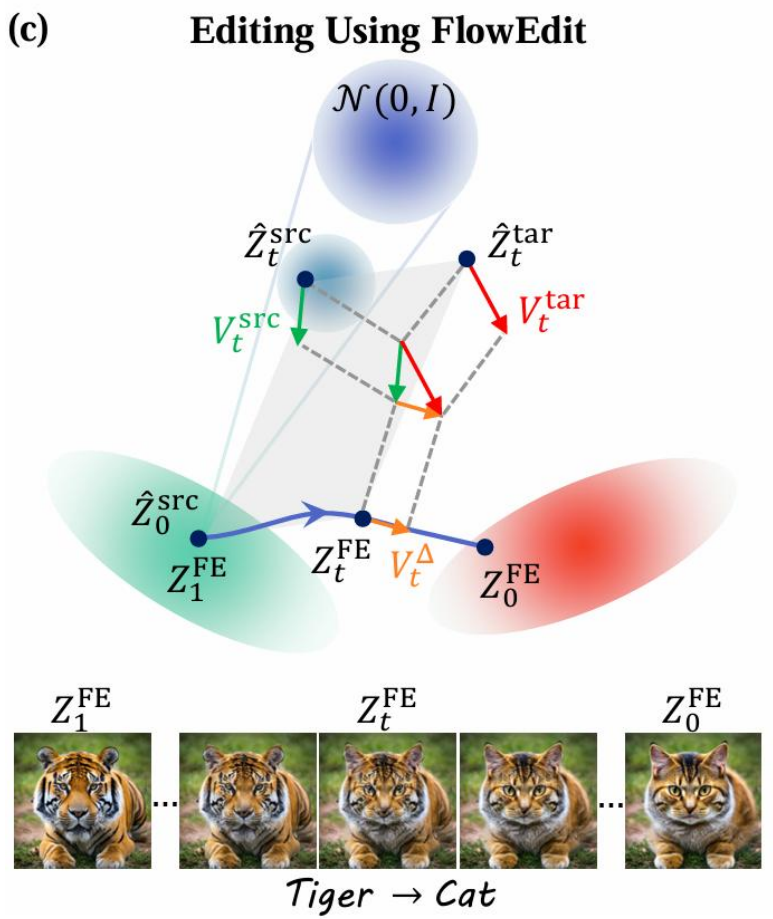
# Conclusion

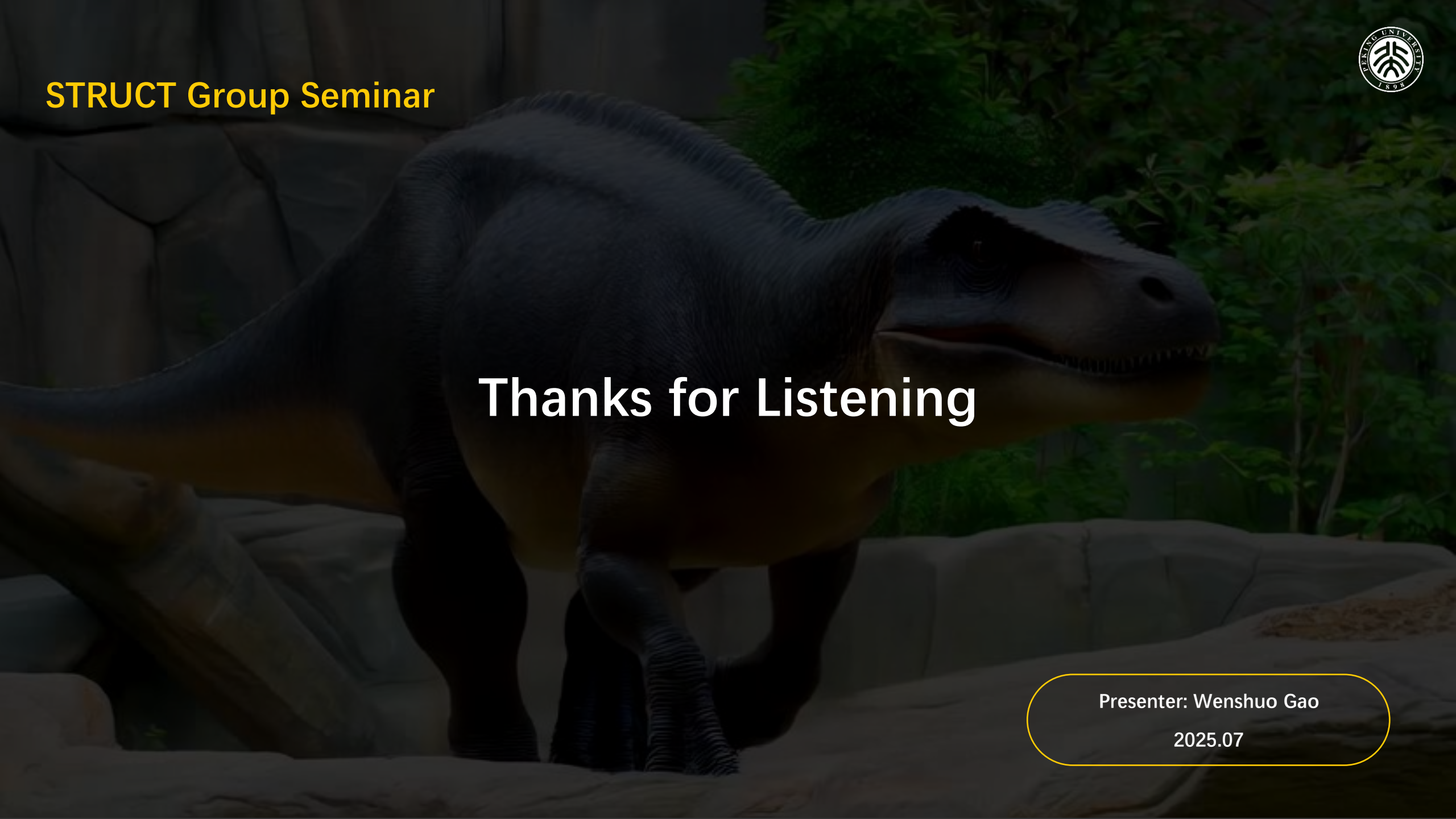
## FlowEdit

A useful training-free model-independent method for image/video editing

Other applications?

- ControlNet
- Image-to-Video models
- …





# Thanks for Listening

Presenter: Wenshuo Gao

2025.07