# FSI-Edit: Frequency and Stochasticity Injection for Flexible Diffusion-Based Image Editing

Kaixiang Yang[1], Xin Li[1], Yuxi Li[1], Qiang Li, Zhiwei Wang[*]

Wuhan National Laboratory for Optoelectronics, Huazhong University of Science and Technology

[1]: Co-first authors, [*]: Corresponding author.

2025-12-4

CONTENTS

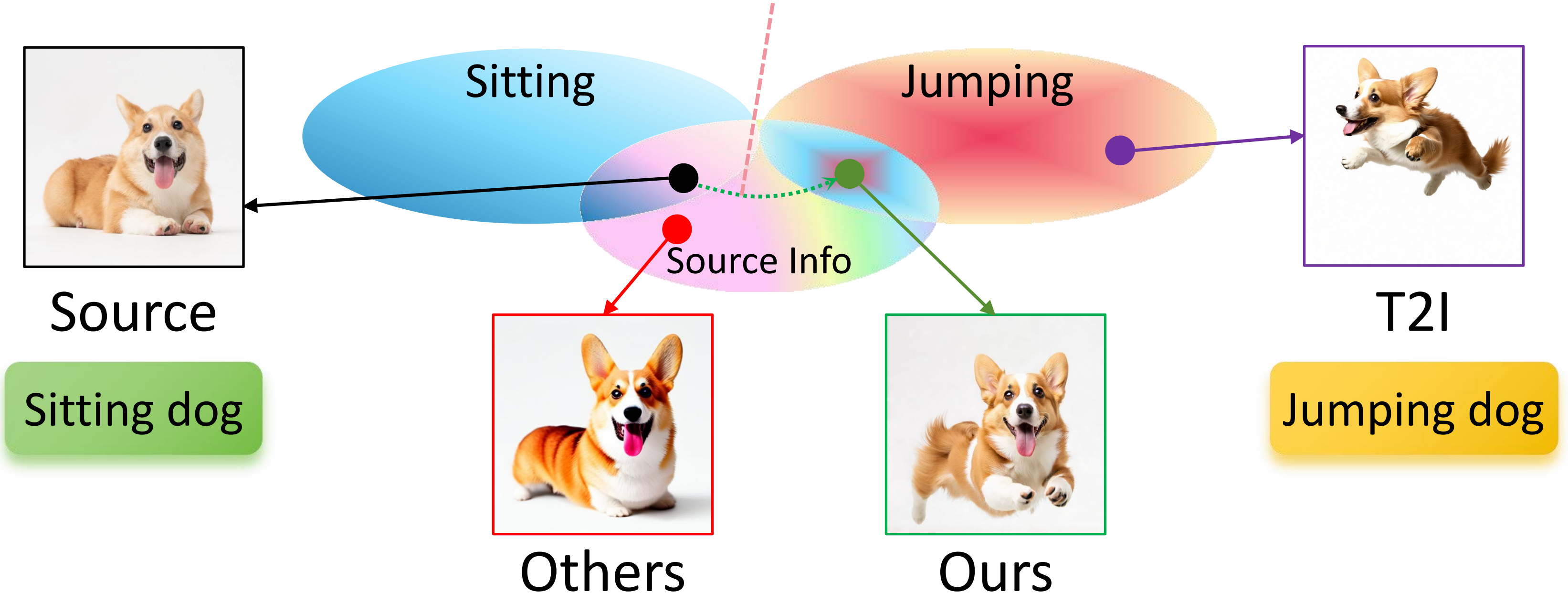
## Image editing

modify *source image* according to *target textual prompt* while preserving the unedited regions

### Non-rigid editing

substantial modifications including object addition, removal, or significant pose alteration

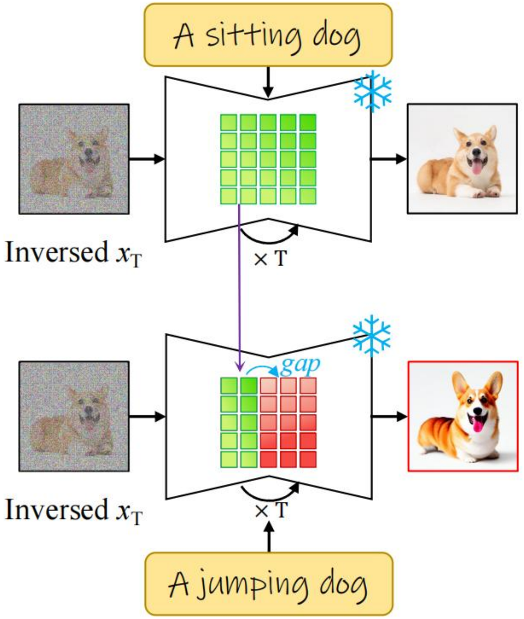
Large-scale non-rigid editing remains highly challenging



CONTENTS

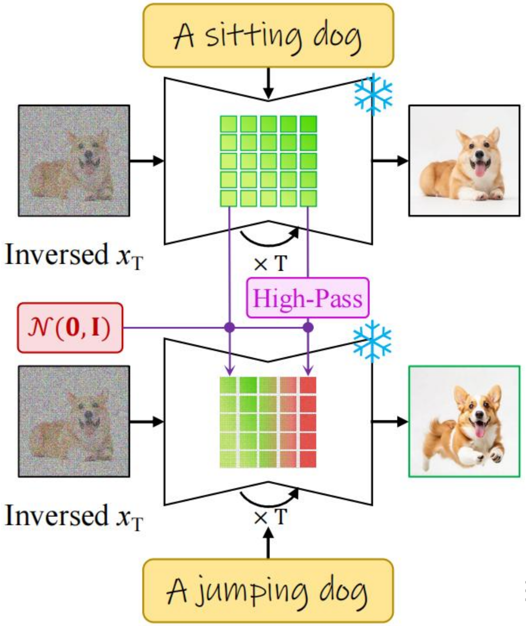
# 2.1 Limitation in previous methods

## Editing Paradigm Comparison




- ✗ **Semantic gap: directly inject** attention features from the reconstruction branch to target branch
- ✗ **Constrained generative capacity**: excessive **reliance on the source** limits the model's ability to fully unleash its generative potential on non-rigid editing

- ✓ **Frequency residual fusion:** selectively **inject high-frequency** component from source, avoiding interfe-rence from structural infor-mation
- ✓ **Stochastic noise injection: controlled perturbation** enriches the latent space and empowers the model to perform substantial structural changes



# 2.2 Framework

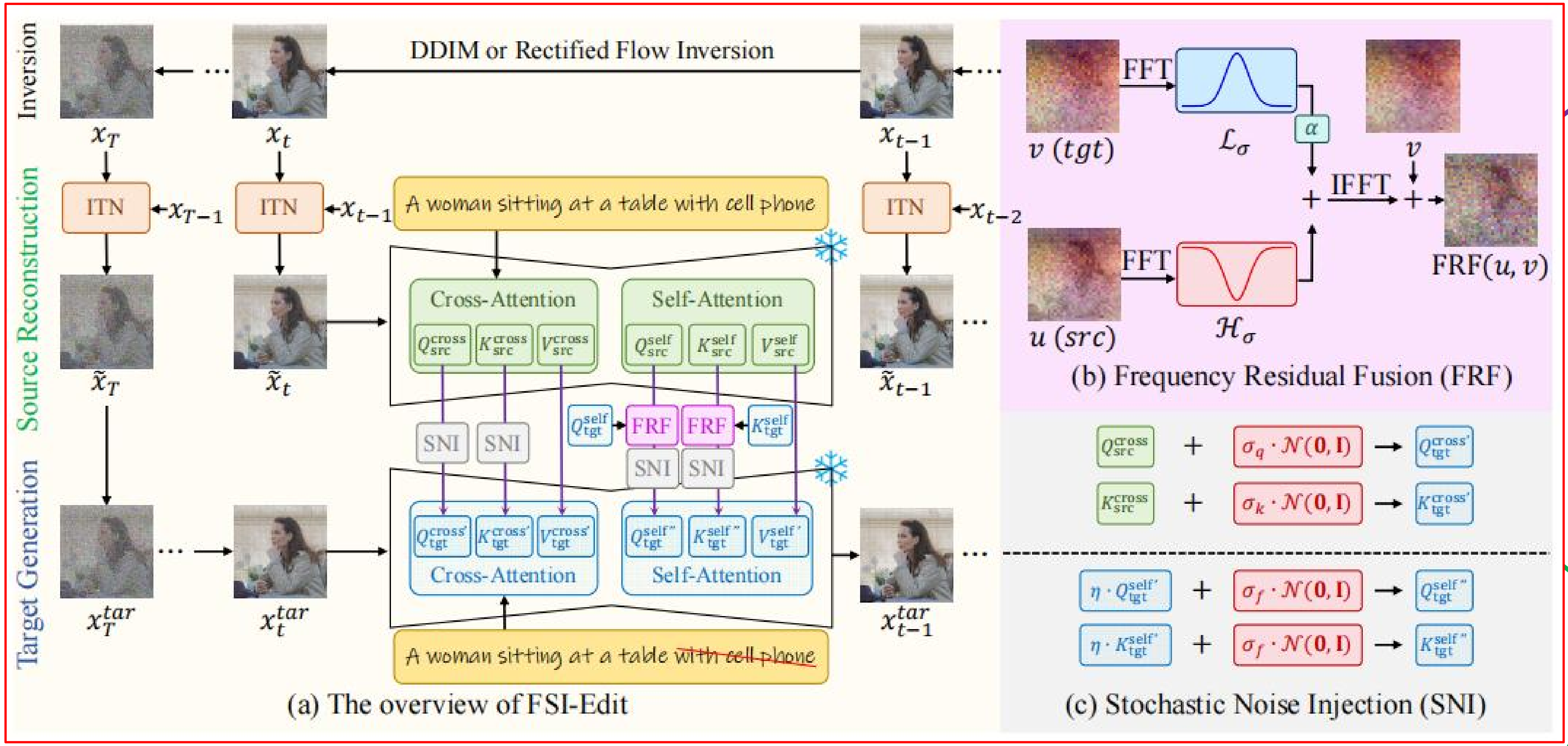
## FSI-Edit

✓ **FRF:** Residual injection of ***source high-frequency*** details into the target

$$\mathcal{F}_{\text{fuse}} = \mathcal{H}_\sigma \cdot \text{FFT}(u) + \alpha \cdot \mathcal{L}_\sigma \cdot \text{FFT}(v)$$

✓ **ITN:** Fusing ***low-frequency*** guidance from ***early timestep*** to enhance recons-truction

$$\tilde{x}_t = \text{IFFT}\big(\mathcal{H}_\sigma \cdot \text{FFT}(x_t) + \mathcal{L}_\sigma \cdot \text{FFT}(x_{t-1})\big) + \sigma_x \cdot \mathcal{N}(\mathbf{0}, \mathbf{I})$$




✓ **SNI:** Injecting ***bounded noise*** to unleash the generative cap-acity of the base model

$$K^{\text{cross}'}_{\text{tgt}} = K^{\text{cross}}_{\text{src}} + \sigma_k \cdot \mathcal{N}(\mathbf{0}, \mathbf{I}) \qquad Q^{\text{cross}'}_{\text{tgt}} = Q^{\text{cross}}_{\text{src}} + \sigma_q \cdot \mathcal{N}(\mathbf{0}, \mathbf{I})$$

CONTENTS

# 3.1 Datasets

## Dataset

1. PIE-Bench[1]
2. 700 image-prompt pairs across 10 diverse editing categories

## Metrics

- Structural Similarity

  **Structure Distance**

- Content Preservation

  **PSNR, LPIPS, MSE, SSIM**

- Text-Image Consistency

  **CLIP similarity**

- **original_prompt:**

  *a slanted mountain bicycle on the road in front of a building*

- **editing_prompt:**

  *a slanted [rusty] mountain bicycle on the road in front of a building*

[1] Ju X, Zeng A, Bian Y, et al. Direct inversion: Boosting diffusion-based editing with 3 lines of code[J]. arXiv preprint arXiv:2310.01506, 2023.

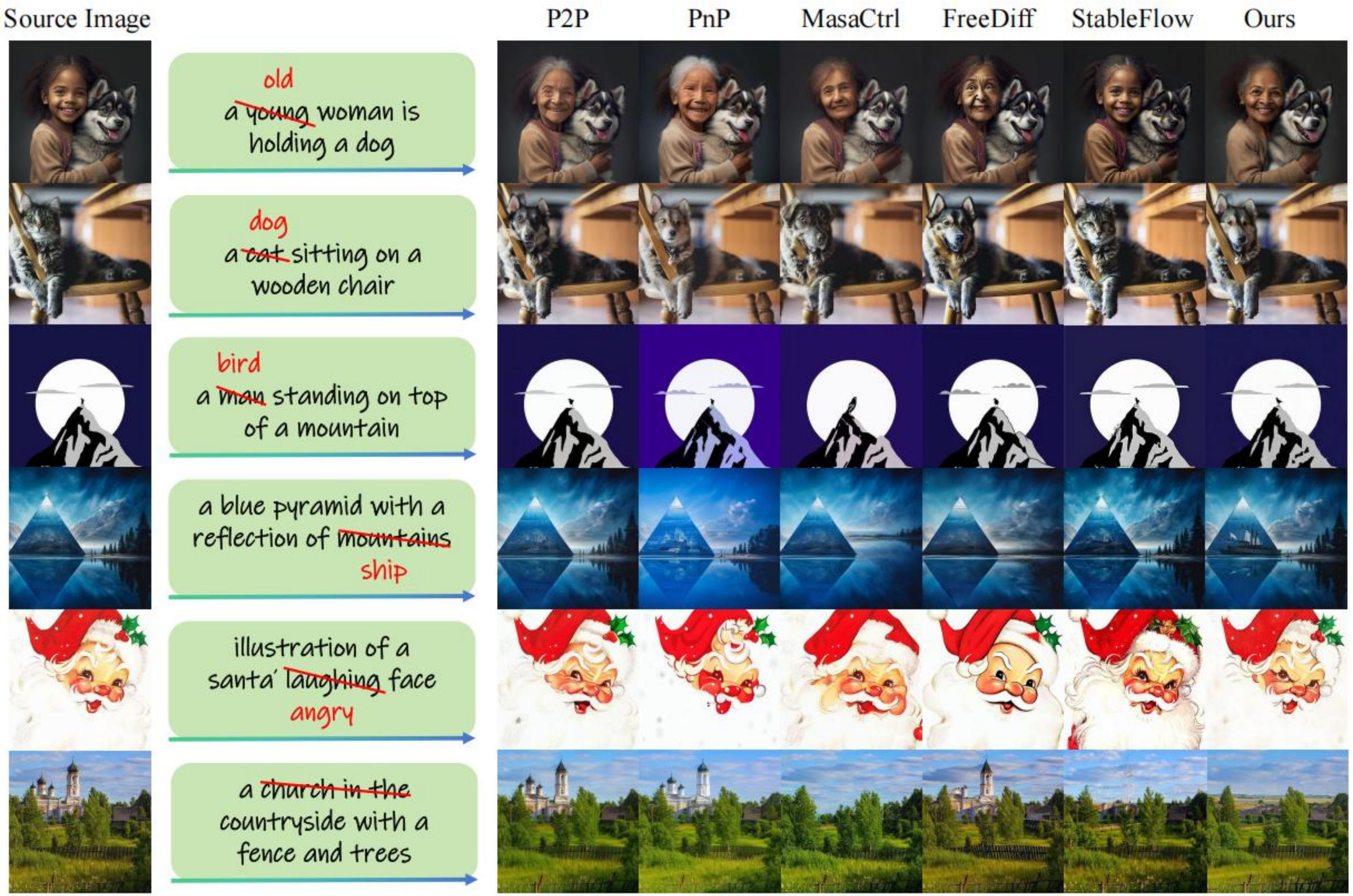
# 3.2 Comparison with SOTA

| Method | Model | Structure | Background Preservation | | | | CLIP Similarity | |
|---|---|---|---|---|---|---|---|---|
| | | $Distance_{\times 10^3} \downarrow$ | $PSNR \uparrow$ | $LPIPS_{\times 10^3} \downarrow$ | $MSE_{\times 10^4} \downarrow$ | $SSIM_{\times 10^2} \uparrow$ | $Whole \uparrow$ | $Edited \uparrow$ |
| P2P [8] | UNet | **11.65** | **27.22** | **54.55** | **32.86** | **84.76** | 25.02 | 22.10 |
| PnP [9] | UNet | 24.29 | 22.46 | 106.06 | 80.45 | 79.68 | 25.41 | 22.62 |
| MasaCtrl [19] | UNet | 24.70 | 22.64 | 87.94 | 81.09 | 81.33 | 24.38 | 21.35 |
| FlexiEdit [16] | UNet | 22.13 | 25.74 | 80.45 | 58.45 | 82.62 | 25.15 | **22.87** |
| FreeDiff [34] | UNet | 18.70 | 24.73 | 89.76 | 55.32 | 81.68 | 25.03 | 22.12 |
| Ours-LDM | UNet | 15.84 | 24.69 | 88.42 | 52.21 | 81.93 | **25.46** | 22.30 |
| RF-Inv [39] | Transformer | 48.76 | 19.51 | 195.85 | 155.74 | 68.95 | 25.11 | 22.50 |
| StableFlow [40] | Transformer | 19.24 | 23.04 | **76.94** | 84.85 | **87.22** | 24.30 | 21.28 |
| RF-Edit [31] | Transformer | 24.45 | 24.41 | 113.44 | 56.46 | 83.84 | 25.03 | 22.28 |
| DCEdit [33] | Transformer | 22.36 | 25.41 | 94.17 | 48.09 | 85.60 | 25.47 | **22.71** |
| Ours-DiT* | Transformer | **13.71** | **26.61** | 85.44 | **36.50** | 86.25 | **25.69** | 22.50 |

*: Ours runs on an RTX 4090 GPU (24 GB memory) and completes within 20 seconds.

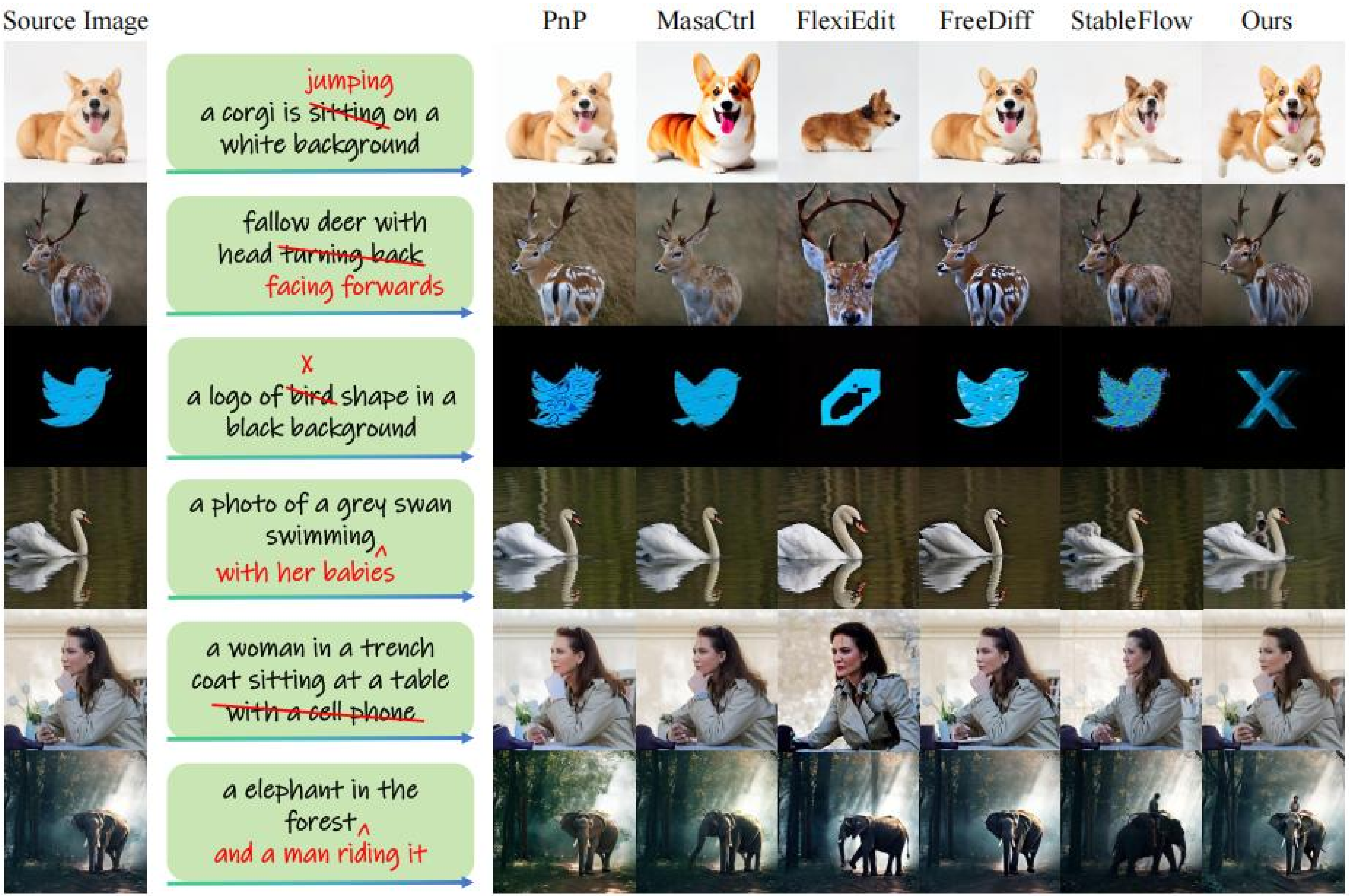
✦ The results show that our method better balances background fidelity in unedited areas and semantic consistency in edited regions.

# 3.3 Visualization




✦ Our edited images visually align better with the given editing instructions, producing more satisfactory visual results.

## 3.3 Visualization




✦ Our method achieves superior visual quality in non-rigid editing tasks.

CONTENTS

# 4 Conclusion

- We identify and address two core limitations of most current image editing methods for non-rigid edits, e.g., semantic inconsistency between reconstruction and generation features, and insufficient generative flexibility due to strong image priors.

- We propose FSI-Edit, featuring a new frequency residual fusion module that selectively transfers high-frequency details for more accurate feature alignment, and a stochastic noise injection strategy that expands the generation space to enable more precise and flexible structural transformations.

- We conduct extensive experiments on PIE-Bench benchmark, and the comparison results demonstrate that FSI-Edit significantly outperforms existing methods on both rigid and non-rigid editing tasks, confirming its effectiveness and generalizability across diverse editing scenarios.

# Thank you

FSI-Edit: Frequency and Stochasticity Injection for Flexible Diffusion-Based Image Editing

Kaixiang Yang[1], Xin Li[1], Yuxi Li[1], Qiang Li, Zhiwei Wang*

Wuhan National Laboratory for Optoelectronics, Huazhong University of Science and Technology

2025-12-4