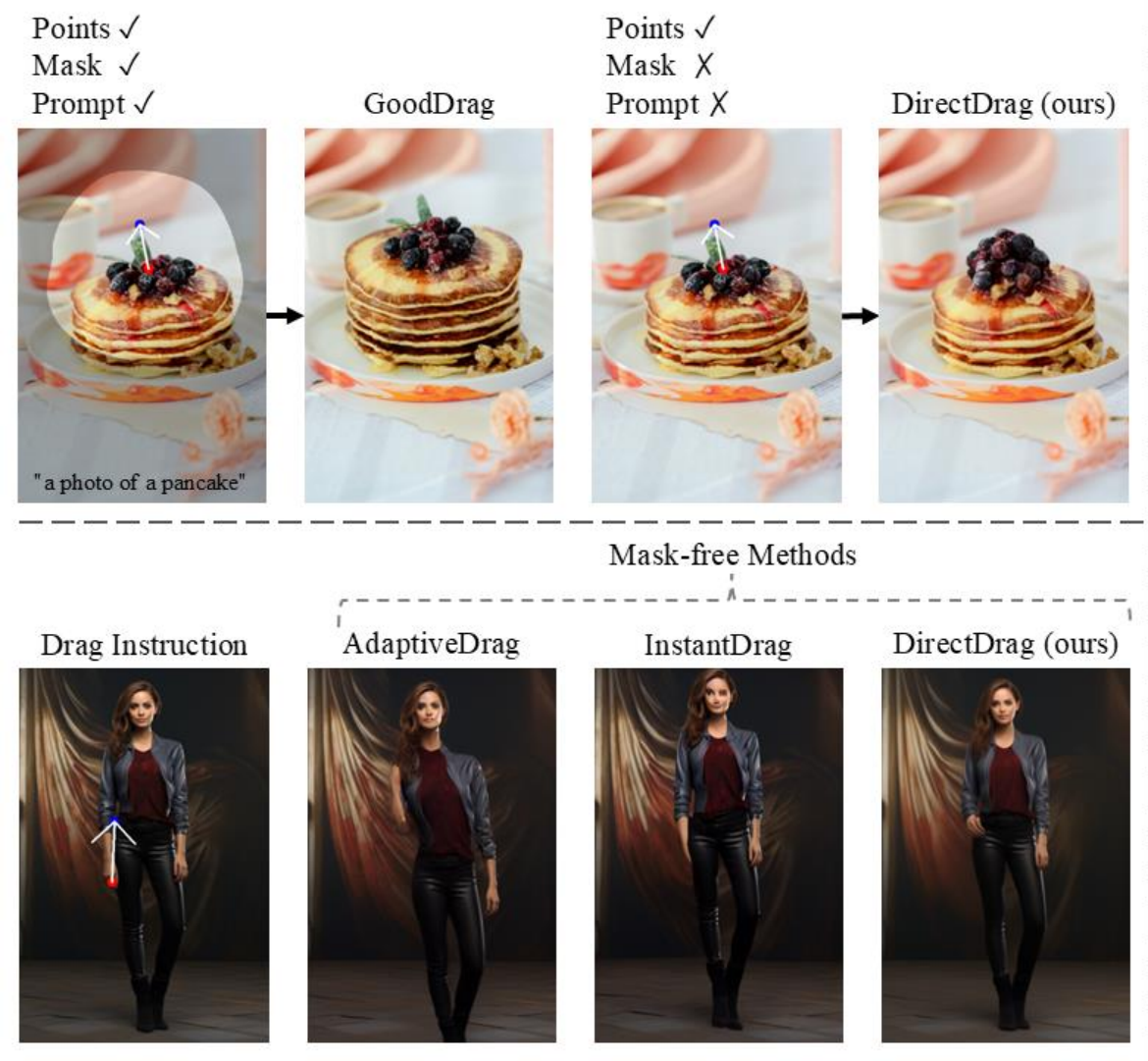
WACV 2026
TUCSON, AZ
3/6 - 3/10

DirectDrag: High-Fidelity, Mask-Free, Prompt-Free Drag-based Image Editing via Readout-Guided Feature Alignment

Sheng-Hao Liao[1], Shang-Fu Chen[2], Tai-Ming Huang[2], Wen-Huang Cheng[2], Kai-Lung Hua[1,3]
[1]National Taiwan University of Science and Technology, [2]National Taiwan University,
[3]Microsoft Taiwan

國立臺灣科技大學
TAIWAN TECH NATIONAL TAIWAN UNIVERSITY OF SCIENCE AND TECHNOLOGY

NTUST MVC LAB
MULTIMEDIA AND VISUAL COMPUTING LAB


Points ✓
Mask ✓
Prompt ✓
GoodDrag
Points ✓
Mask ✗
Prompt ✗
DirectDrag (ours)
"a photo of a pancake"
Mask-free Methods
Drag Instruction
AdaptiveDrag
InstantDrag
DirectDrag (ours)

Outline

國立臺灣科技大學
NATIONAL TAIWAN UNIVERSITY OF SCIENCE AND TECHNOLOGY

01
Introduction

Image Editing in Deep Learning

國立臺灣科技大學
NATIONAL TAIWAN UNIVERSITY OF SCIENCE AND TECHNOLOGY

Traditional Image Editing

- Less user control
- Hard to make precise control

HyperStyle

Interactive Image Editing

- Precise control
- More creative

InstructPix2Pix

Alaluf et al., HyperStyle: StyleGAN Inversion with HyperNetworks for Real Image Editing. CVPR 2022
Brooks et al., InstructPix2Pix: Learning to Follow Image Editing Instructions. CVPR 2023

Interactive Image Editing

國立臺灣科技大學
NATIONAL TAIWAN UNIVERSITY OF SCIENCE AND TECHNOLOGY

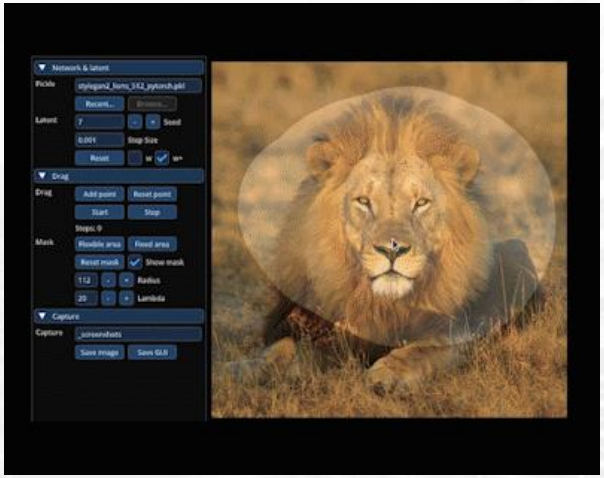
Text-based Image Editing

- Natural language instructions
- Lack of pixel level control

Prompt-to-prompt

Drag-based Image Editing

- Enable precise spatial control
- But often causing unintended modification



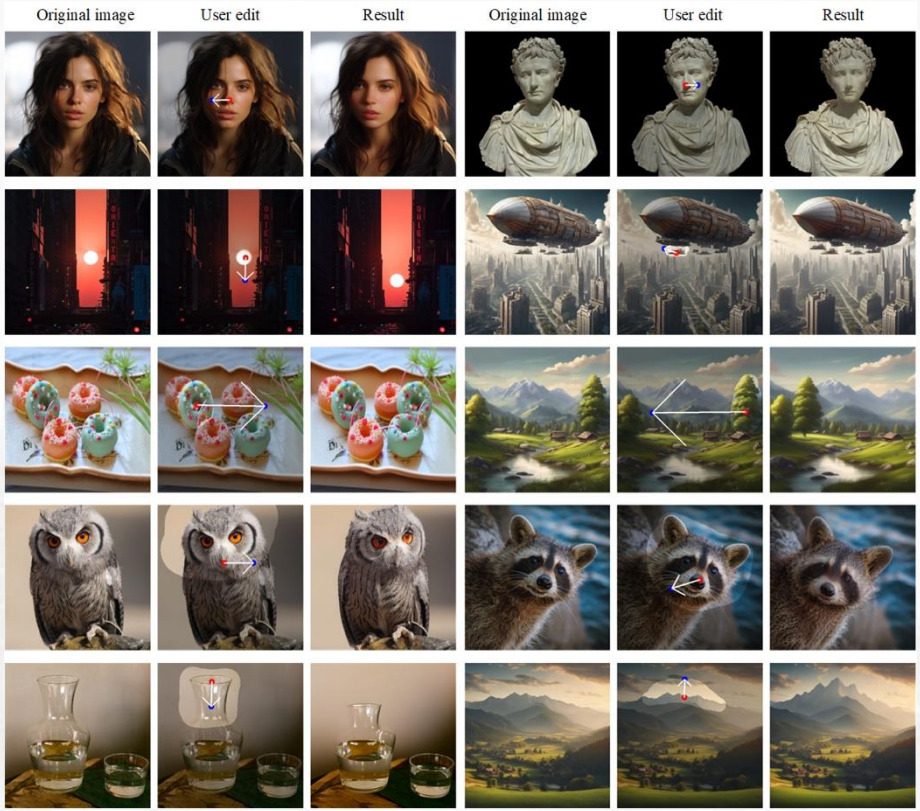
DragGAN

Hertz et al., Prompt-to-Prompt Image Editing with Cross-Attention Control. ICLR 2023
Pan et al., Drag Your GAN: Interactive Point-based Manipulation on the Generative Image Manifold. SIGGRAPH 2023

Drag-based Image Editing

Typically require following input:

- RGB Image
- Mask Area
- Text prompt
- Multiple drag instructions



FastDrag

Zhao et al., FastDrag: Manipulate Anything in One Step. NeurIPS 2024

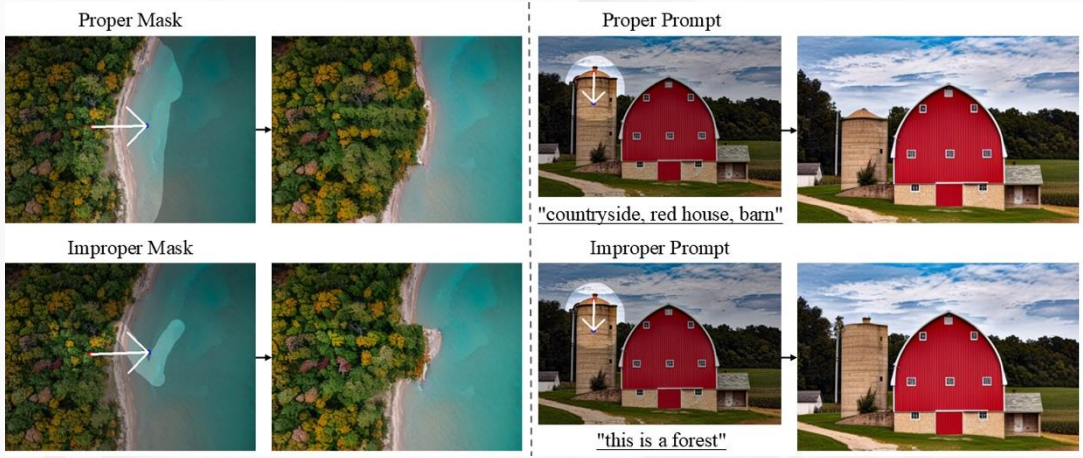
Drag-based Image Editing

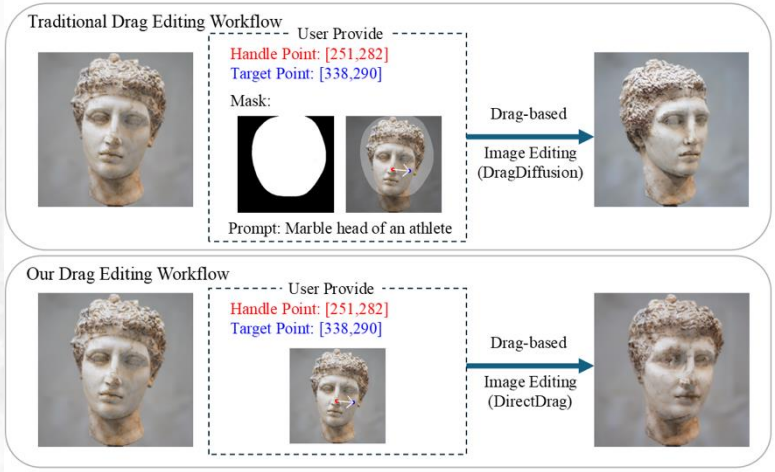
Three key challenges

- Precision — Make the dragged point reach the target
  - Metric: Mean Distance (MD)
- Visual Quality — Preserve realism and consistency
  - Metric: Image Fidelity (e.g., 1 - LPIPS)
- Efficiency — Fast and responsive editing
  - Metric: Average drag time per point

Motivation

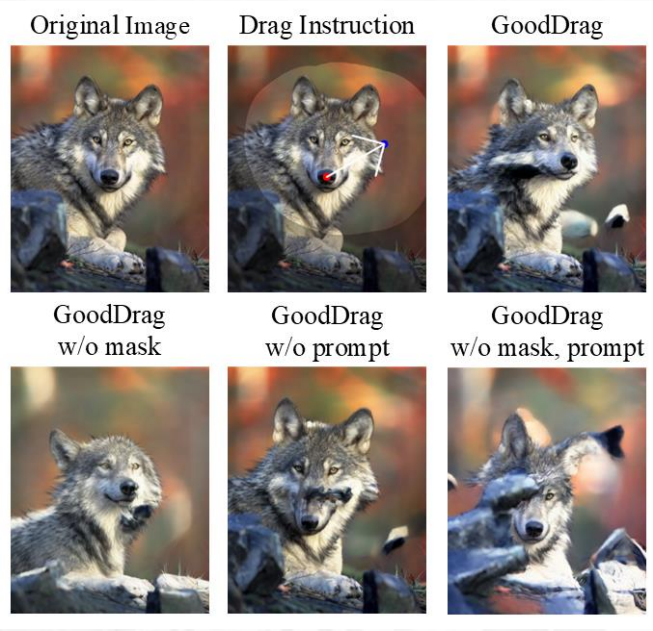
- Existing drag-based methods require manual mask and prompt
- Improper inputs lead to visual artifacts
- Workflow is complex and hard to control
- We simplify editing with only point inputs





Contribution

- Without mask and prompt, results are distorted and unreliable
- We propose **DirectDrag**, a fully mask-free and prompt-free drag-based image editing framework
- Our method includes three key contributions:
  - **Auto Soft Mask Generation**: Automatically infers editable regions from drag instruction
  - **Readout-Guided Feature Alignment**: Aligns intermediate features to maintain structural consistency
  - **Latent Warpage Function**: Provides geometry-aware initialization for more accurate edits



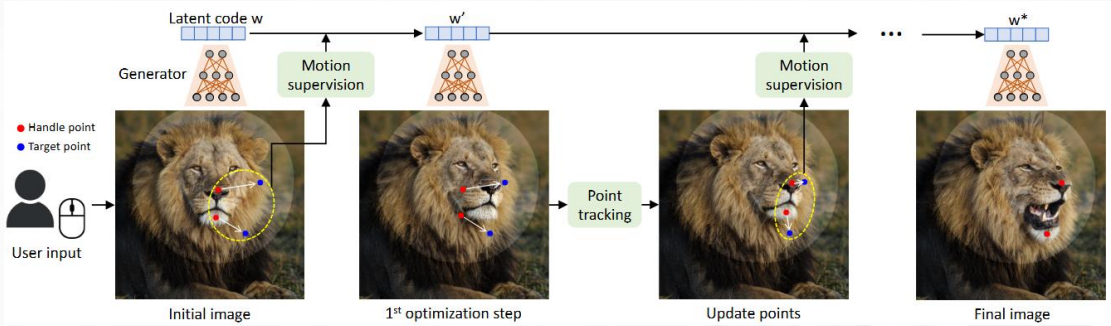
DirectDrag (ours)

02
Related Work

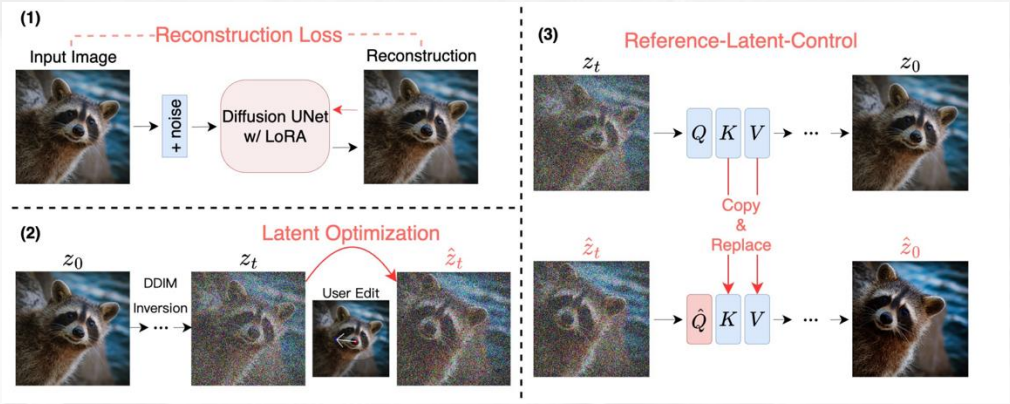
Seminal Papers

DragGAN

- Build on StyleGAN
- Proposed motion supervision and point tracking



Pan et al., Drag Your GAN: Interactive Point-based Manipulation on the Generative Image Manifold. SIGGRAPH 2023



DragDiffusion

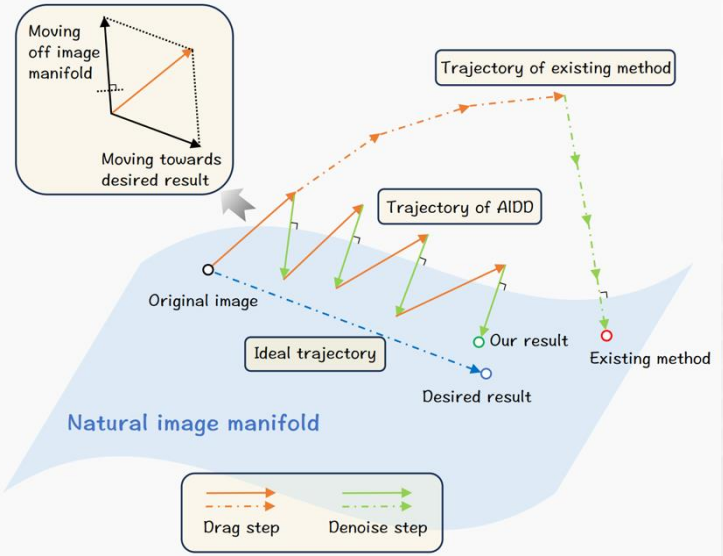
- Build on LDM
- Latent Optimization

Shi et al., DragDiffusion: Harnessing Diffusion Models for Interactive Point-based Image Editing. CVPR 2024

Editing Result Improvement

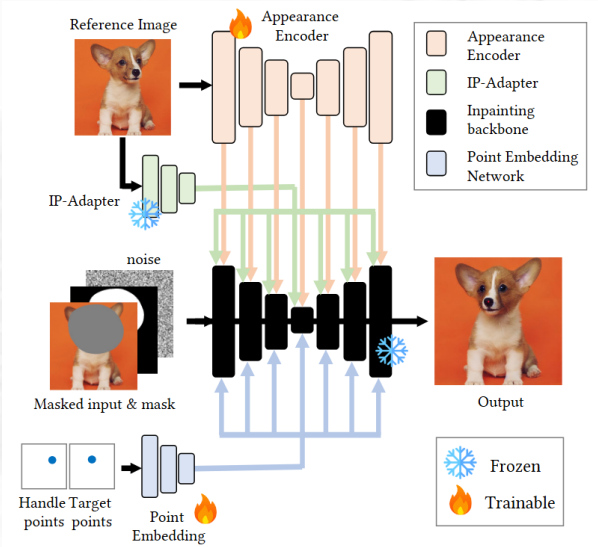
GoodDrag

- Drag and denoise strategy



LightningDrag

- Trained on video data
- Appearance Encoder
- Point embedding network



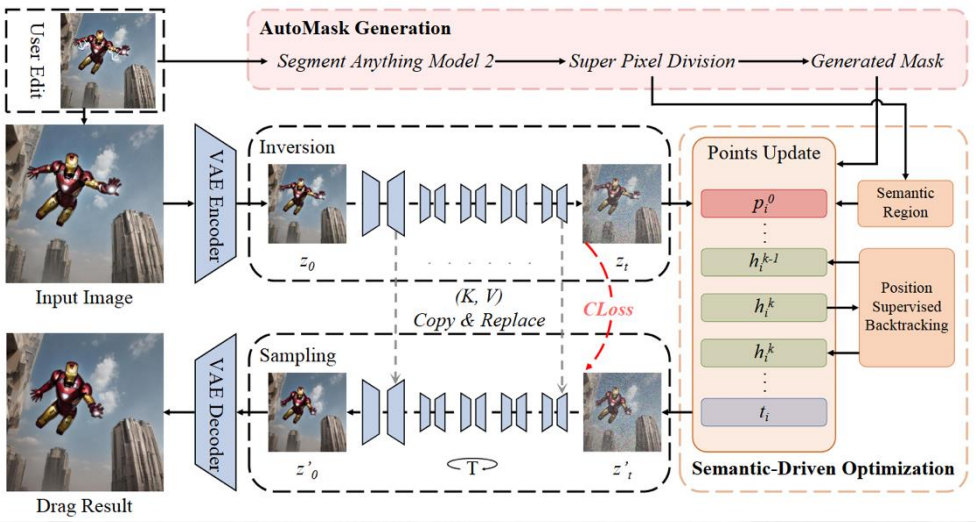
Zhang et al., GoodDrag: Towards Good Practices for Drag Editing with Diffusion Models. ICLR 2025
Shi et al., LightningDrag: Lightning Fast and Accurate Drag-based Image Editing Emerging from Videos. ICML 2025

Mask-free Drag-based Image Editing

國立臺灣科技大學
NATIONAL TAIWAN UNIVERSITY OF SCIENCE AND TECHNOLOGY

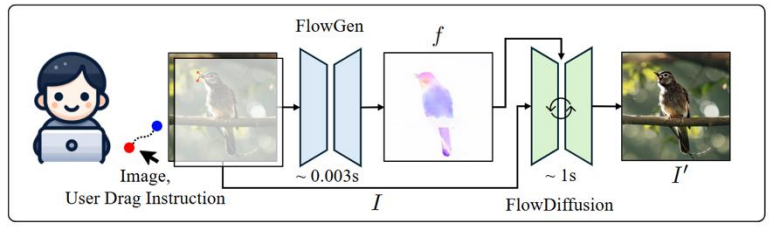
AdaptiveDrag

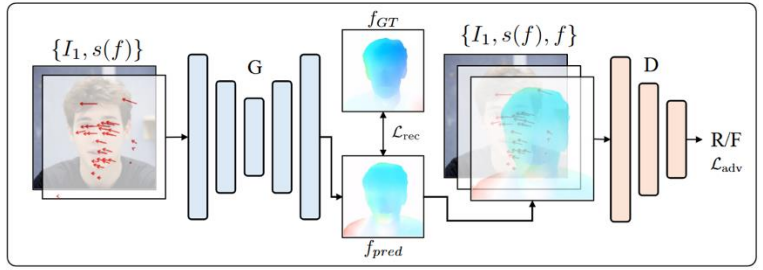
- Rely on segmentation model
- Still require prompt input



InstantDrag

- Optical flow model
- Require training data, and huge computational resources

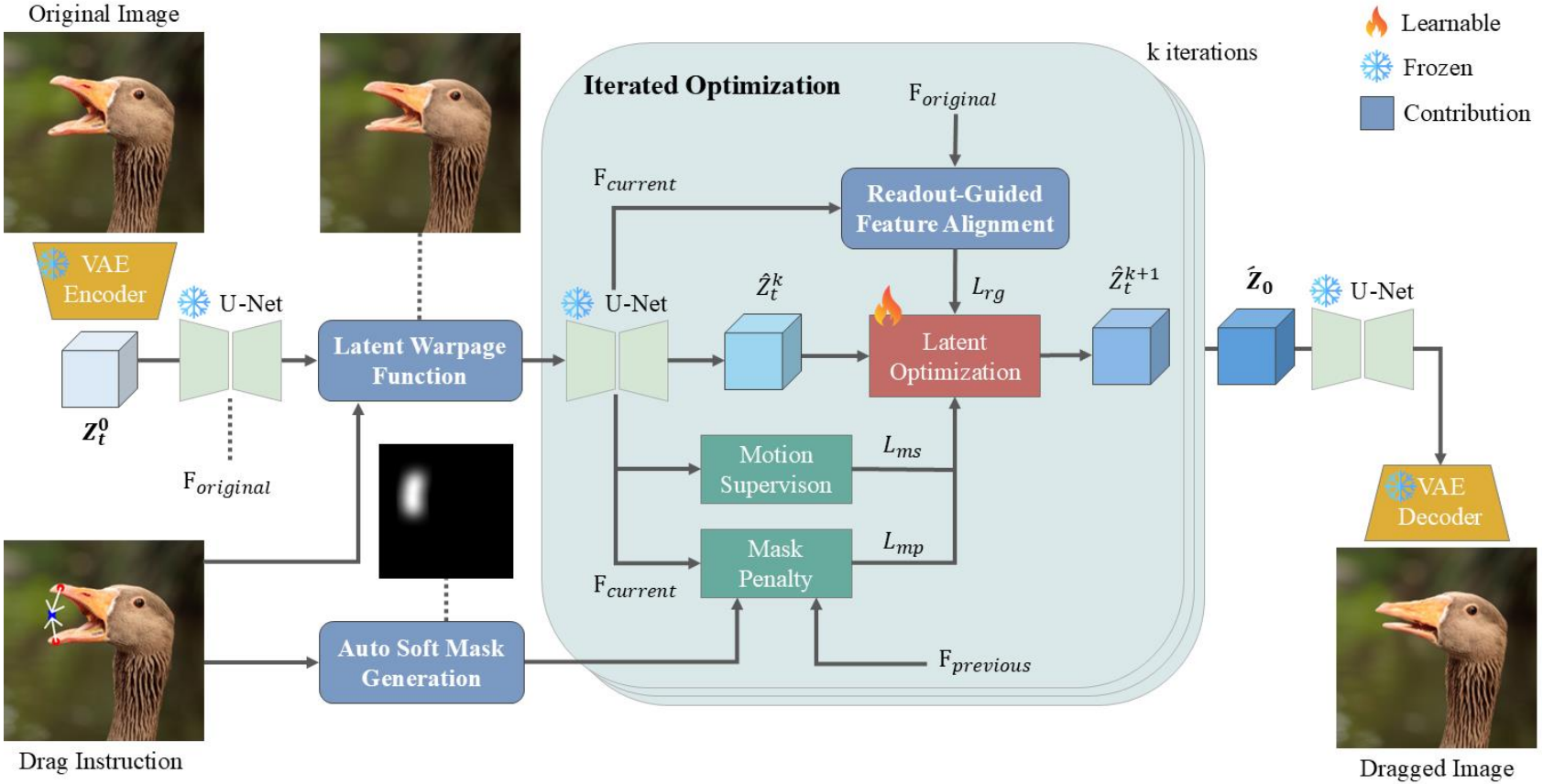




Chen et al., AdaptiveDrag: Semantic-Driven Dragging on Diffusion-Based Image Editing. ArXiv 2024
Shin et al., InstantDrag: Improving Interactivity in Drag-based Image Editing. SIGGRAPH Asia 2024

03
Method

Overview


Original Image
VAE Encoder
U-Net
z_t^0
$F_{original}$
Latent Warpage Function
Drag Instruction
Auto Soft Mask Generation
k iterations
Iterated Optimization
$F_{original}$
$F_{current}$
Readout-Guided Feature Alignment
U-Net
$\hat{Z}_t^k$
L_{rg}
Latent Optimization
$\hat{Z}_t^{k+1}$
Motion Supervison
L_{ms}
Mask Penalty
L_{mp}
$F_{current}$
$F_{previous}$
$\hat{Z}_0$
U-Net
VAE Decoder
Dragged Image
Learnable
Frozen
Contribution

Drag-based Image Editing

- Motion Supervision : Supervises feature movement from handle to target points
- Point Tracking : Keeps track of the handle point's updated location after each editing step

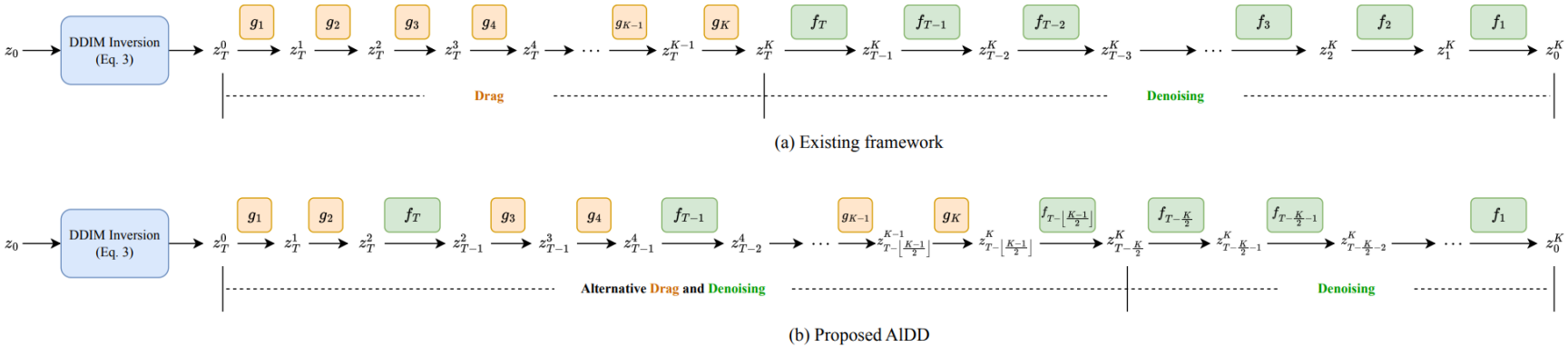
$$\mathcal{L}_{\text{ms}} = \sum_{i=1}^{n} \sum_{q} \left\| \mathcal{F}_{q+d_i}(\hat{\mathbf{z}}_t^k, \hat{\mathbf{c}}^k) - \text{sg}(\mathcal{F}_q(\hat{\mathbf{z}}_t^k, \hat{\mathbf{c}}^k)) \right\|_1$$

n : Number of dragging points
q : Spatial position in the feature map
d_i : Displacement vector
Z_t^k : Latent feature at timestep t during the k-th drag step
$\hat{c}^k$: Conditioning input (e.g., text, semantic map, etc.)
$F_q(z, \hat{c})$: Feature vector at position q given latent and condition
$sg(\cdot)$: Stop-gradient operation (no backpropagation)

Zhang et al., GoodDrag: Towards Good Practices for Drag Editing with Diffusion Models. ICLR 2025

AlDD Strategy

- Alternative Drag and Denoising strategy from GoodDrag
- Alternates between dragging and denoising at each timestep instead of separating them
- Prevents feature drift and quality degradation caused by editing everything before denoising



Zhang et al., GoodDrag: Towards Good Practices for Drag Editing with Diffusion Models. ICLR 2025

Auto Soft Mask Generation (1/2)

- Automatically generates a soft mask along the drag trajectory
- Places 1's along interpolated points from handle to target
- Applies **Gaussian blur** to smooth the mask and expand its region
- Normalizes the result to get a soft, differentiable mask
- Eliminates the need for manual mask annotation

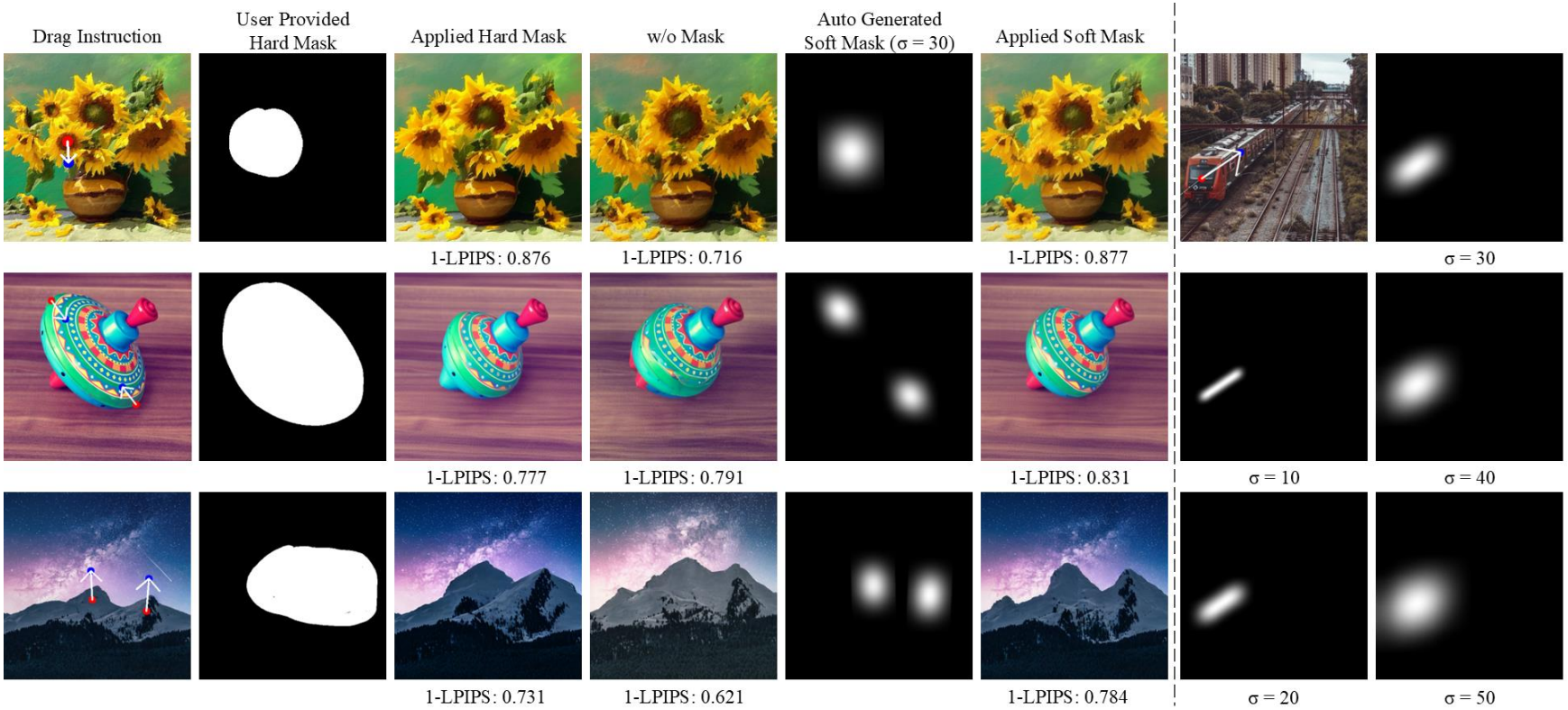
$$\tilde{M}(x_k, y_k) = 1, \quad \text{where}$$

$$(x_k, y_k) = \lfloor (1 - \alpha_k)(x_0, y_0) + \alpha_k (x_1, y_1) \rceil ,$$

$$\alpha_k = \frac{k}{N-1} = \frac{k}{\max(|x_1 - x_0|, |y_1 - y_0|)}.$$

$$M = \frac{\text{GaussianBlur}(\tilde{M}, \sigma)}{\max\left(\text{GaussianBlur}(\tilde{M}, \sigma)\right)}.$$

Auto Soft Mask Generation (2/2)



Readout-Guided Feature Alignment (1/2)

國立臺灣科技大學
NATIONAL TAIWAN UNIVERSITY OF SCIENCE AND TECHNOLOGY

Readout Network

- Lightweight **model extracts appearance-preserving features** from frozen denoiser
- Trained via triplet loss to separate distorted vs. preserved appearances

$$\mathcal{L}_{\text{triplet}} = \max\big(0, D(F(I_a), F(I_p)) - D(F(I_a), F(I_n)) + \delta\big)$$

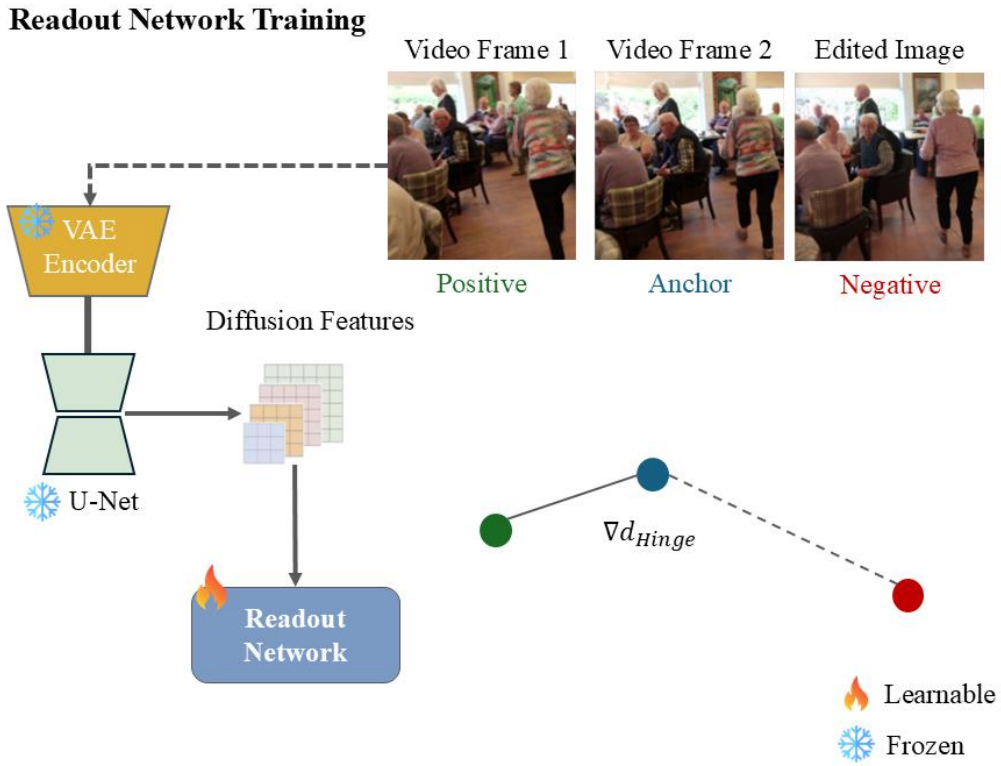
Inference-Time Guidance

- Compare edited latent features with original image features
- Apply L2 loss to align current appearance with initial image

$$\mathcal{L}_{\text{rg}} = \|F(\bar{\mathbf{z}}_t^{\text{k}}) - F(\mathbf{z}_t^0)\|_2^2,$$

Luo et al., Readout Guidance: Learning Control from Diffusion Features. CVPR 2024

Readout-Guided Feature Alignment (2/2)



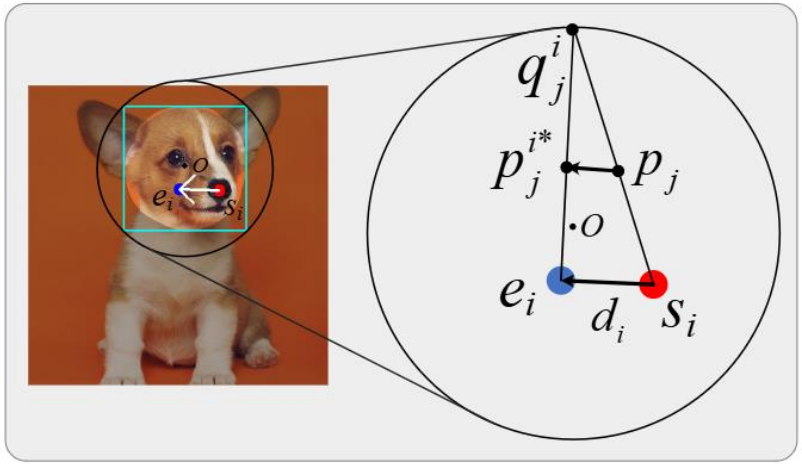
Latent Warpage Function

- Initializes latent features with **geometry-aware deformation**
- Computes pixel-wise displacement as a weighted sum of drag vectors
- Introduces a scaling ratio ρ to prevent over-deformation

$$\mathbf{v}_j = \sum_{i=1}^{k} w_j^i \cdot \lambda_j^i \cdot \mathbf{d}_i,$$

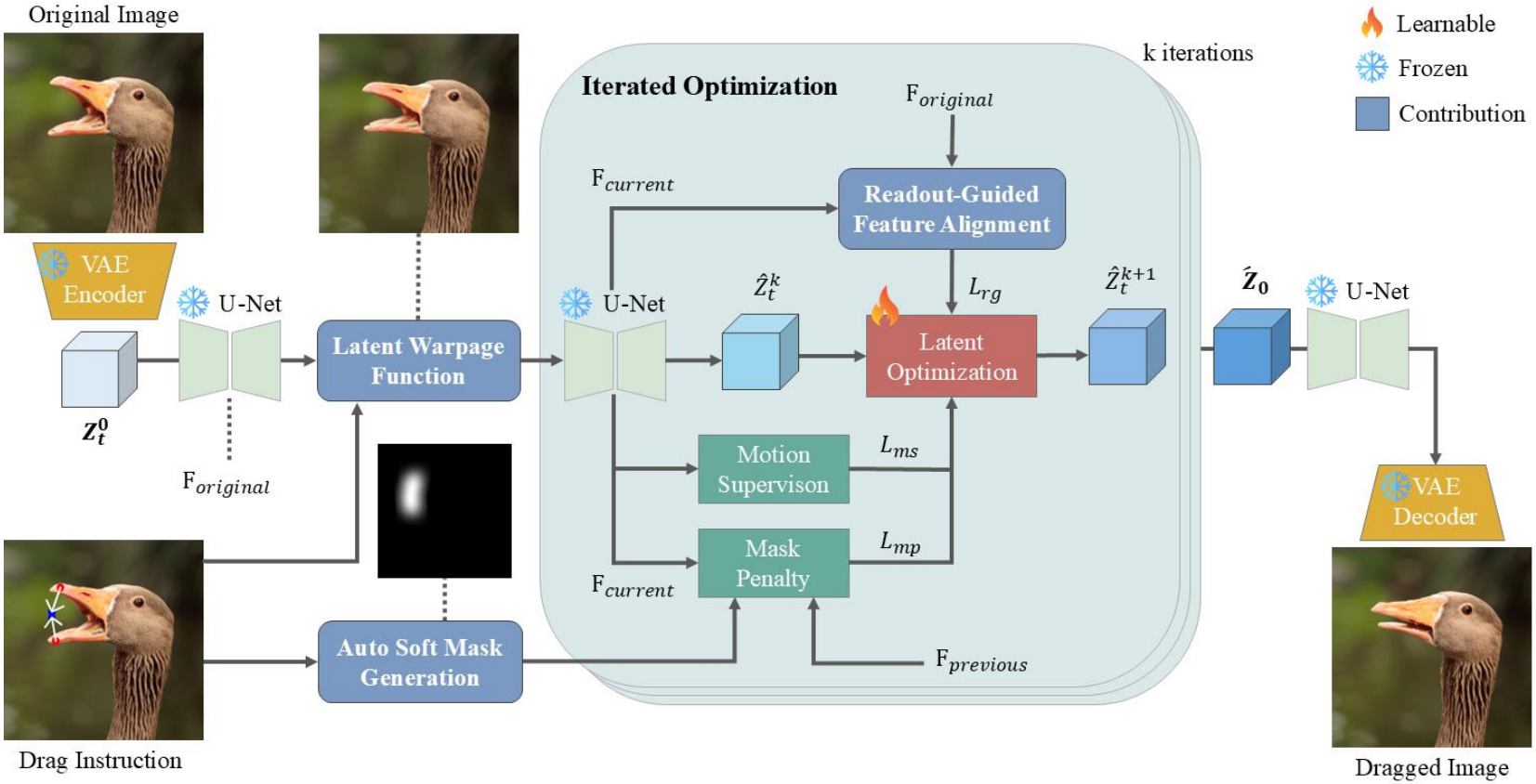
$$\mathbf{d}_i' = \rho \cdot (e_i - s_i),$$

w_j^i : inverse distance weight to each handle point
λ_j^i : geometric stretch factor based on path intersection



Zhao et al., FastDrag: Manipulate Anything in One Step. NeurIPS 2024

Method Summary



04
Experiment

Implementation Details

國立臺灣科技大學
NATIONAL TAIWAN UNIVERSITY OF SCIENCE AND TECHNOLOGY

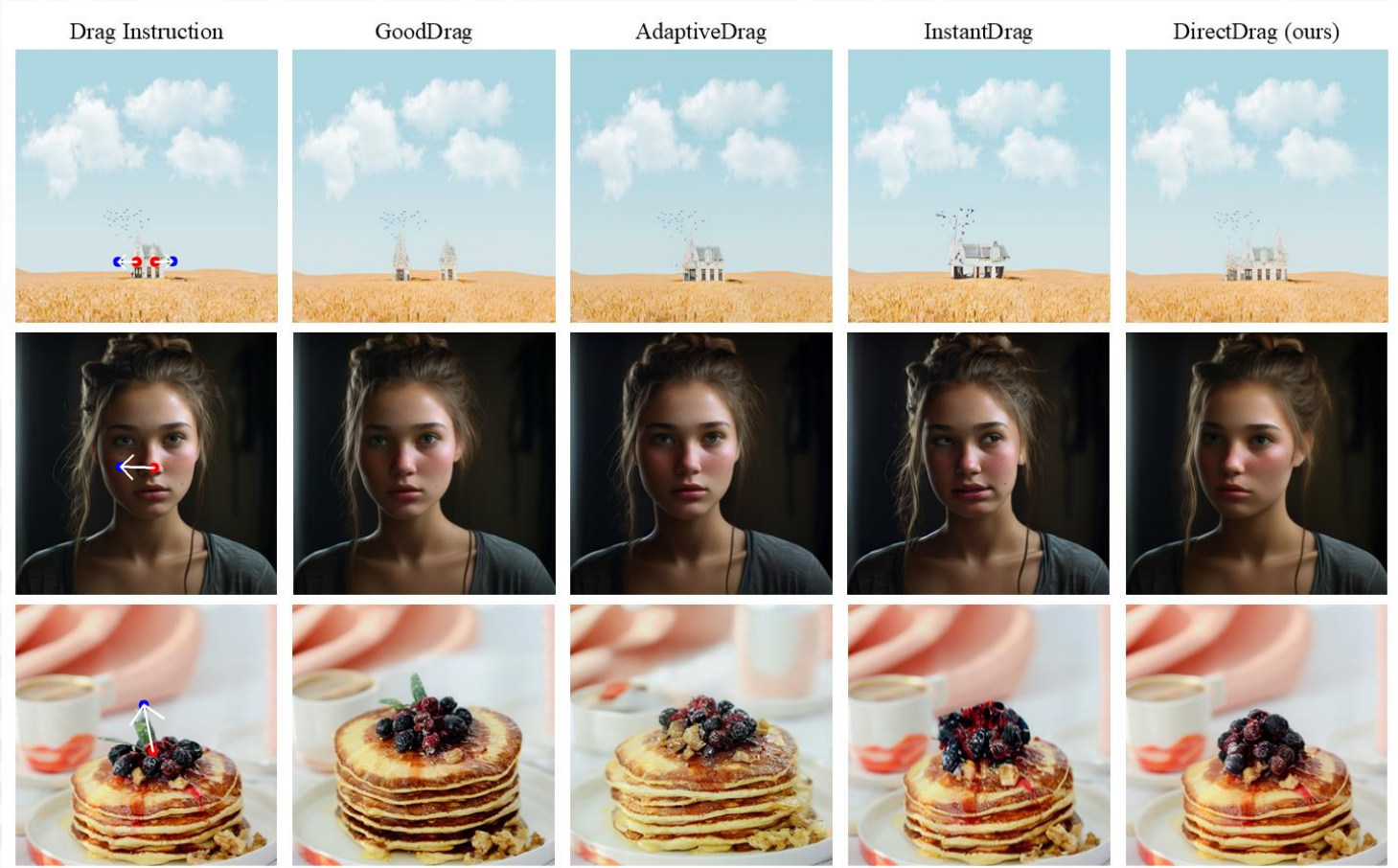
- Built on Stable Diffusion v1.5, using DDIM inversion with 50 steps and guidance scale 1.0
- All experiments run on **NVIDIA RTX 4090** GPU
- Dataset : DragBench (Proposed by DragDiffusion)
- Key configurations:
  - Soft Mask: Gaussian blur with σ=30 for smooth transitions
  - Readout-Guided Weight: Loss scaled by 350× to emphasize appearance preservation
  - Latent Warpage Function: Apply only 15% of displacement to prevent over-dragging
- All other settings follow GoodDrag baseline
- Readout Network
  - Trained on Pascale VOC (Image edit by SDEdit)
  - NVIDIA A40 40G for 3 hours training

Quantitative Evaluation

Method	Venue	Mask	Prompt	IF↑	CLIP SIM↑	MD↓	Model Params	Tuning Params
DragDiffusion [29]	CVPR'24	✓	✓	0.883	0.977	32.87	865M	0.07M
FreeDrag [12]	CVPR'24	✓	✓	0.897	0.977	33.82	865M	0.07M
DiffEditor [17]	CVPR'24	✓	✓	0.877	0.966	31.70	865M	0.07M
DragNoise [13]	CVPR'24	✓	✓	0.899	0.972	37.92	865M	0.33M
FastDrag [38]	NeurIPS'24	✓	✓	0.859	0.963	32.66	865M	0
GoodDrag [37]	ICLR'25	✓	✓	0.869	0.977	25.28	865M	0.07M
DragText [3]	WACV'25	✓	✓	0.870	0.971	34.25	865M	0.12M
LightningDrag [28]	ICML'25	✓	✓	0.881	0.970	29.95	933M	933M
Mask-free methods								
EasyDrag* [8]	CVPR'24	✗	✓	0.882	–	34.44	1770M	**0.07M**
Readout Guidance [14]	CVPR'24	✗	✗	0.867	0.951	55.12	**871M**	5.97M
AdaptiveDrag [2]	ArXiv'24	✗	✓	0.867	0.975	33.94	1168M	**0.07M**
InstantDrag [30]	SIGGRAPH Asia'24	✗	✗	0.878	0.968	30.41	914M	914M
DirectDrag (ours)$_{\text{w/o LWF}}$	–	✗	✓	**0.918**	**0.982**	31.91	**871M**	5.97M
DirectDrag (ours)	–	✗	✗	0.891	0.976	**29.65**	**871M**	5.97M

Table 1. **Quantitative evaluation** on the DragBench [29] dataset. IF = 1 - LPIPS. CLIP SIM = CLIP [22] Similarity. MD = Mean Distance. ✓: Required, ✗: Not Required. LWF: Latent Warpage Function. Model Params: Total parameters used in model. Tunning Params: Parameters require to training in correspond method. * means scores are taken from the another publication.

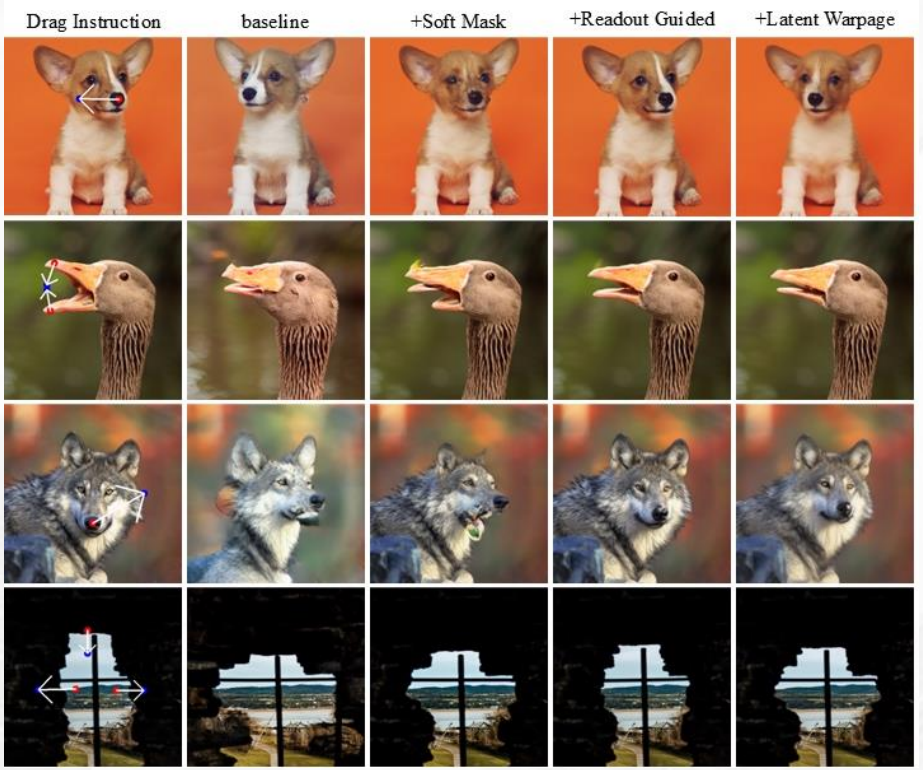
Qualitative Comparison



Ablation Study

國立臺灣科技大學
NATIONAL TAIWAN UNIVERSITY OF SCIENCE AND TECHNOLOGY

Method	SM	RG	LWF	IF↑	CLIP SIM↑	MD↓
Baseline				0.789	0.963	24.74
+ Soft Mask	✓			0.895	0.979	31.35
+ Readout Guided	✓	✓		**0.918**	**0.982**	33.75
+ Readout Guided $_{+prompt}$	✓	✓		**0.918**	**0.982**	31.91
+ Latent Warpage	✓	✓	✓	0.891	0.976	29.65
+ Latent Warpage $_{+prompt}$	✓	✓	✓	0.891	0.975	**29.18**



05

Conclusion

Conclusion

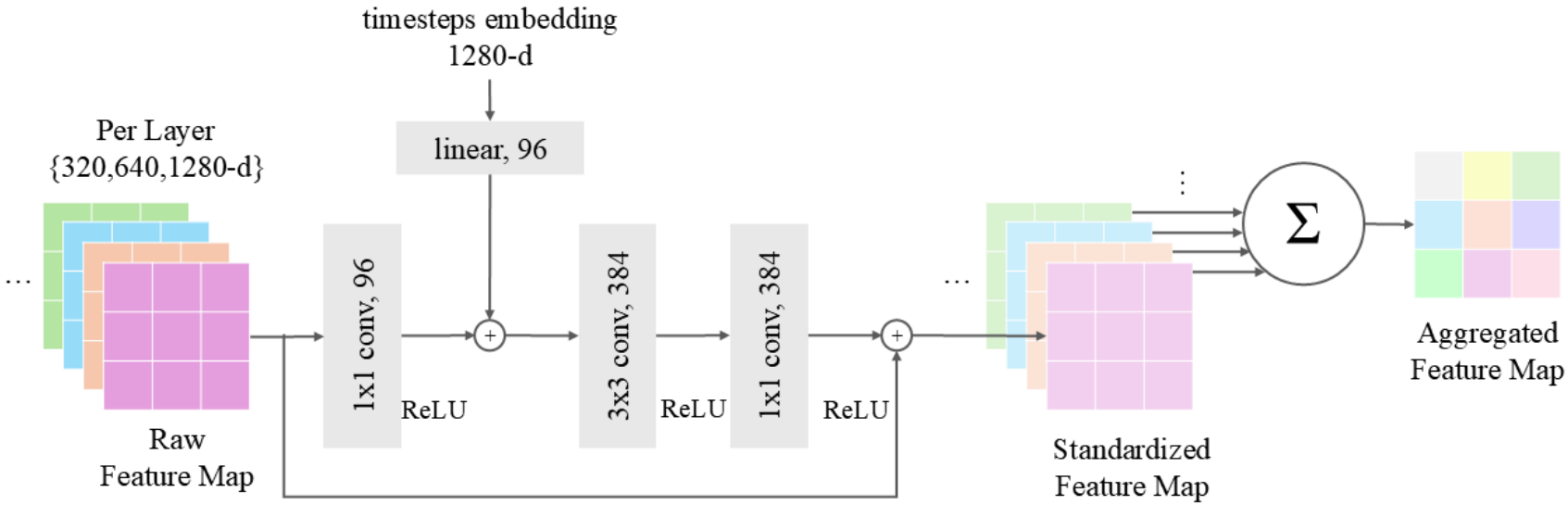
- Contributions
  - Auto Soft Mask Generation
  - Readout-Guided Feature Alignment
  - Latent Warpage Function
- Achieves
  - **SOTA on mask-free methods**
  - **SOTA on IF scores**
- Limitation
  - Over fidelity preserving
  - Lower drag accuracy

Limitation Examples

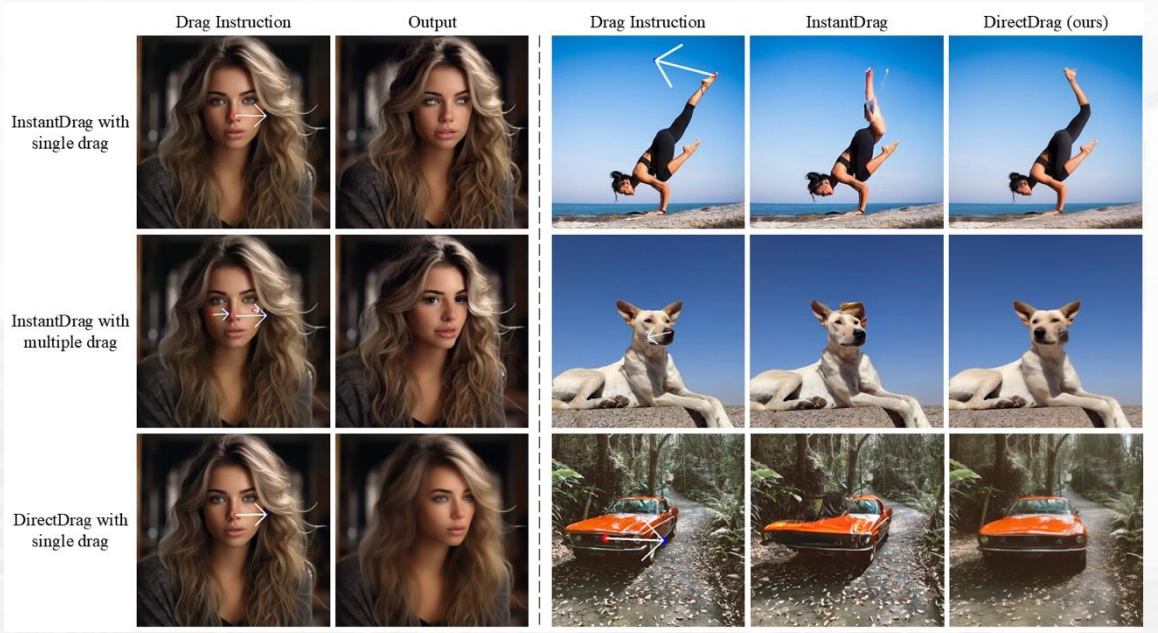
06

Supplementary Material

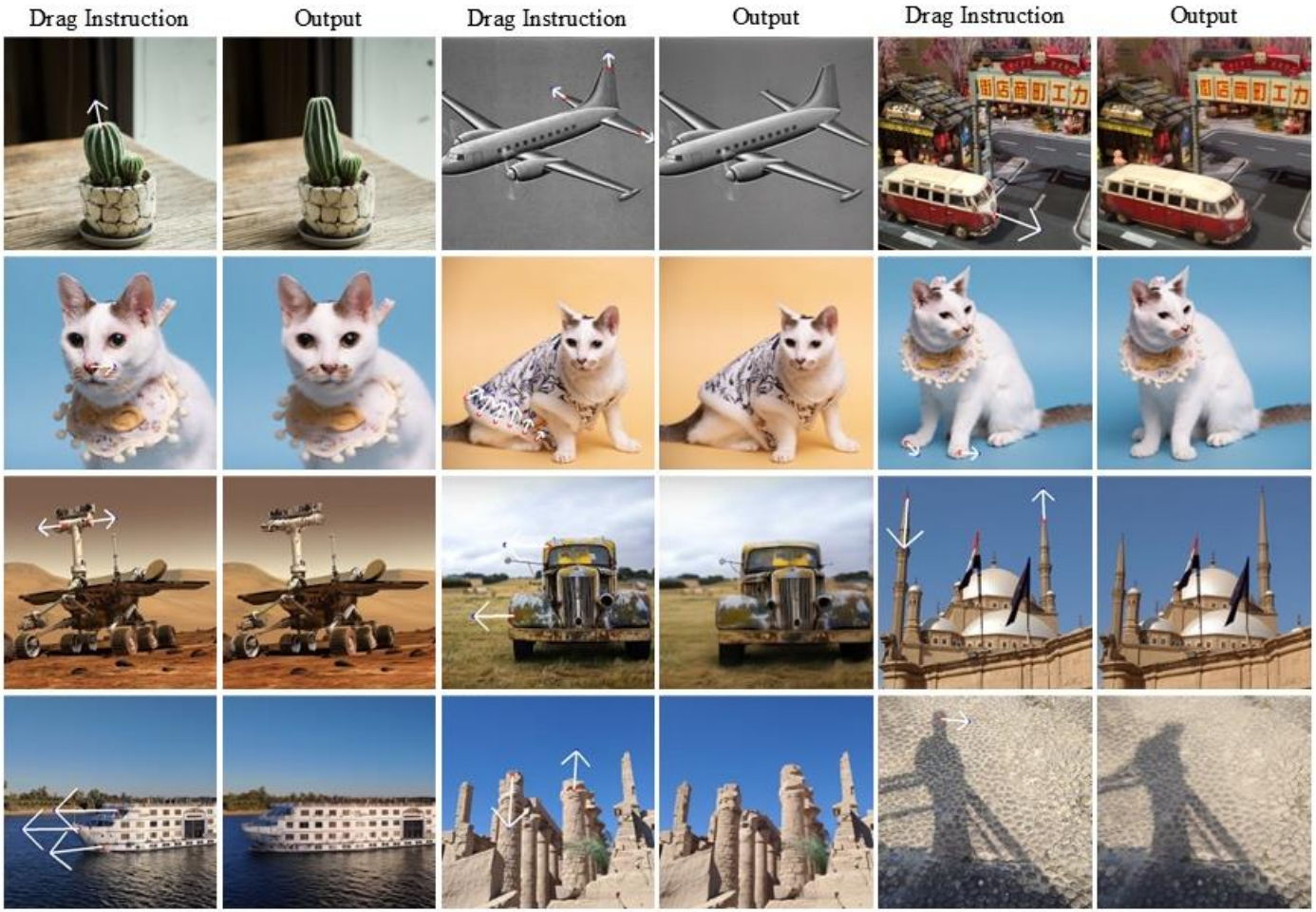
Readout Network Architecture



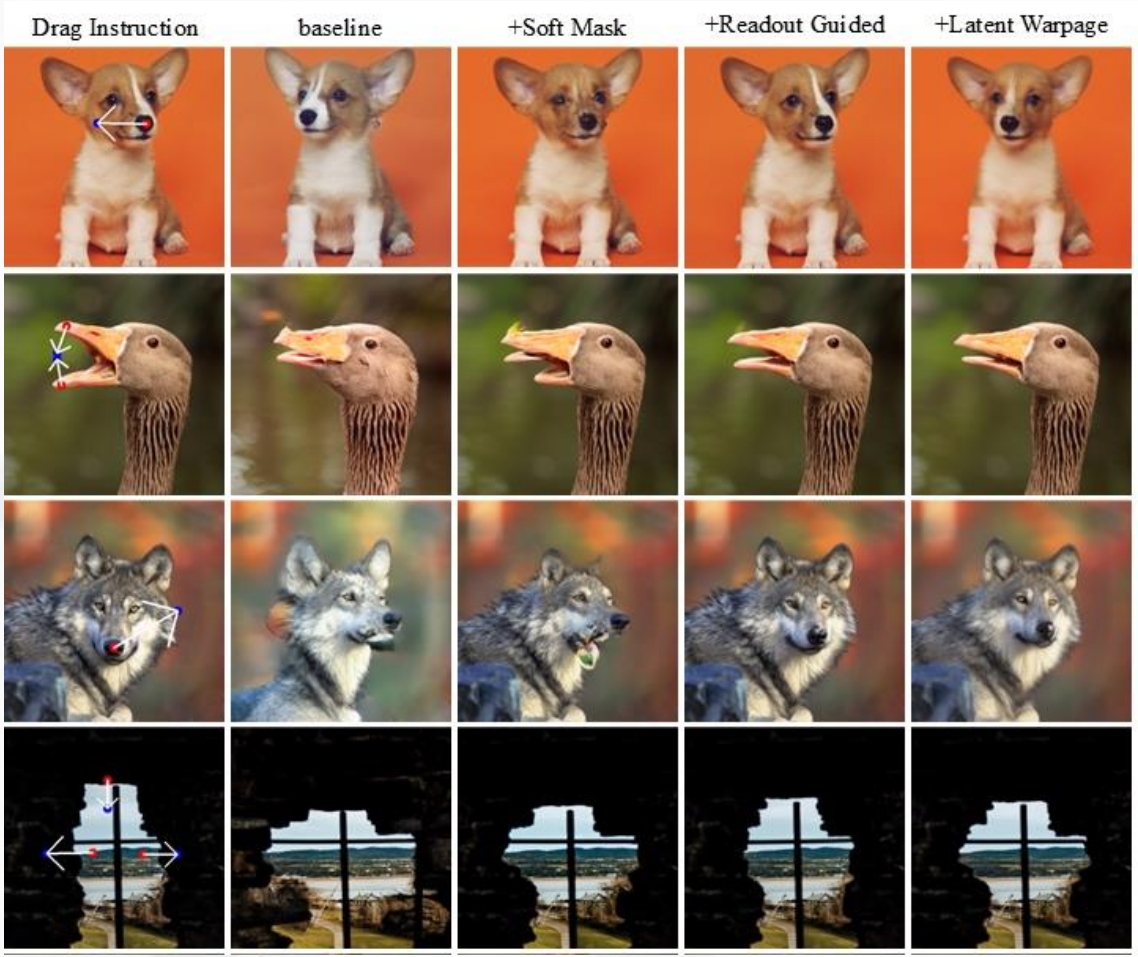

Drag Instruction
GoodDrag
GoodDrag
w/o mask,prompt

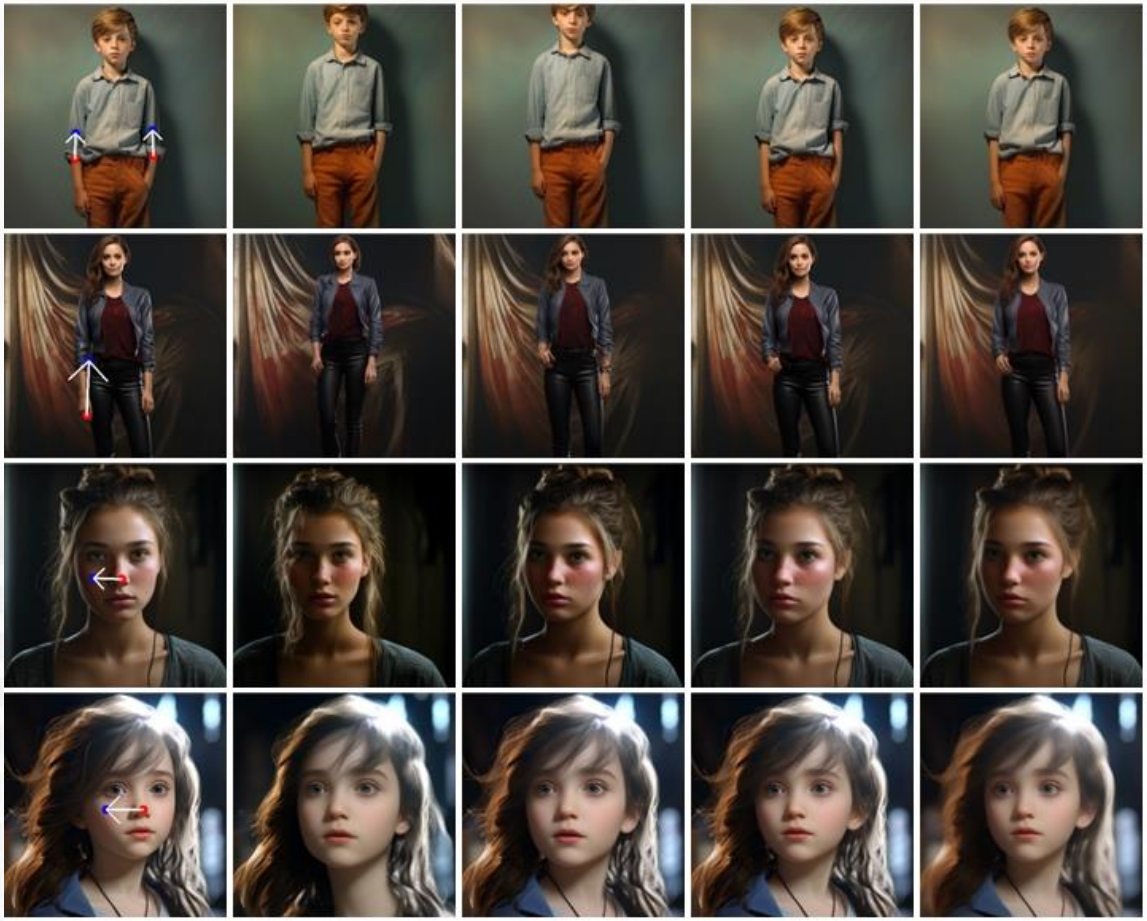

Drag Instruction
Output
Drag Instruction
InstantDrag
DirectDrag (ours)
InstantDrag with
single drag
InstantDrag with
multiple drag
DirectDrag with
single drag

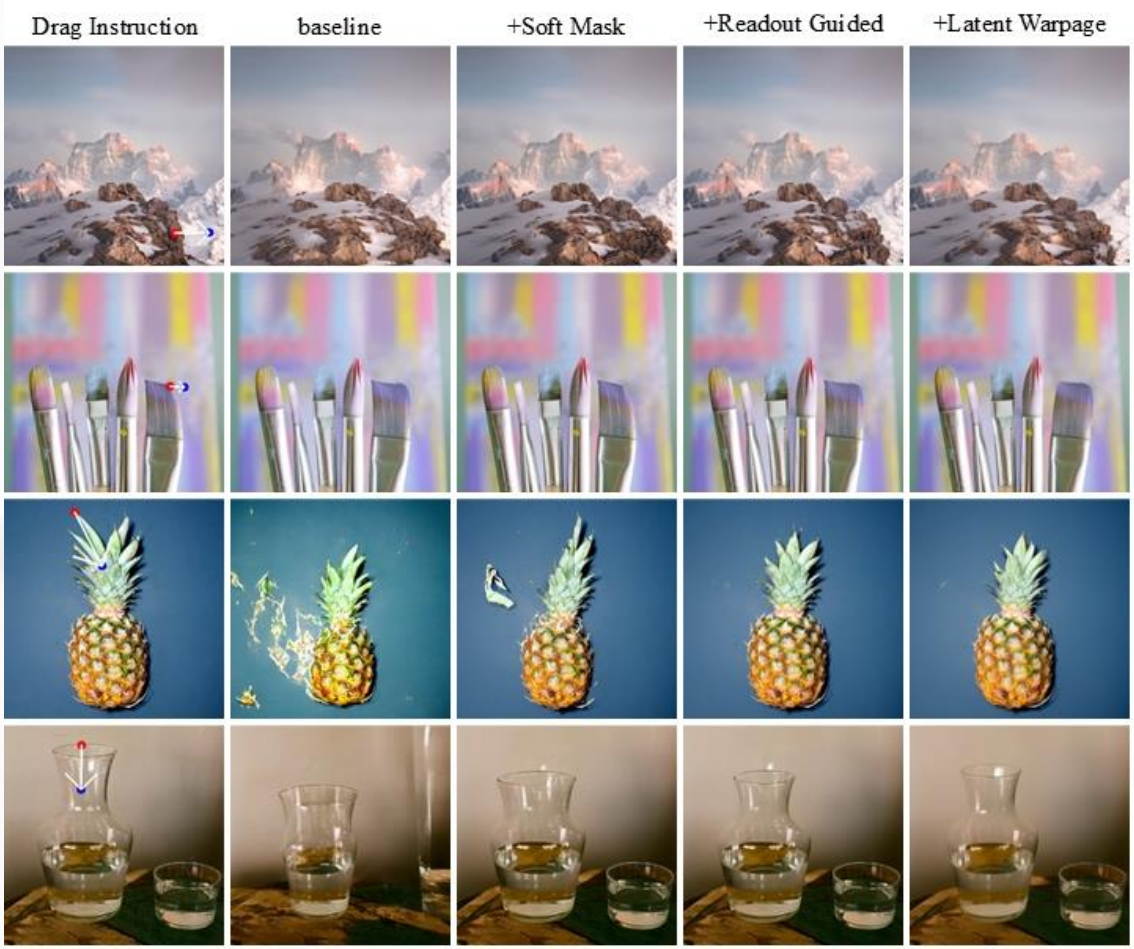
Extended Qualitative Examples

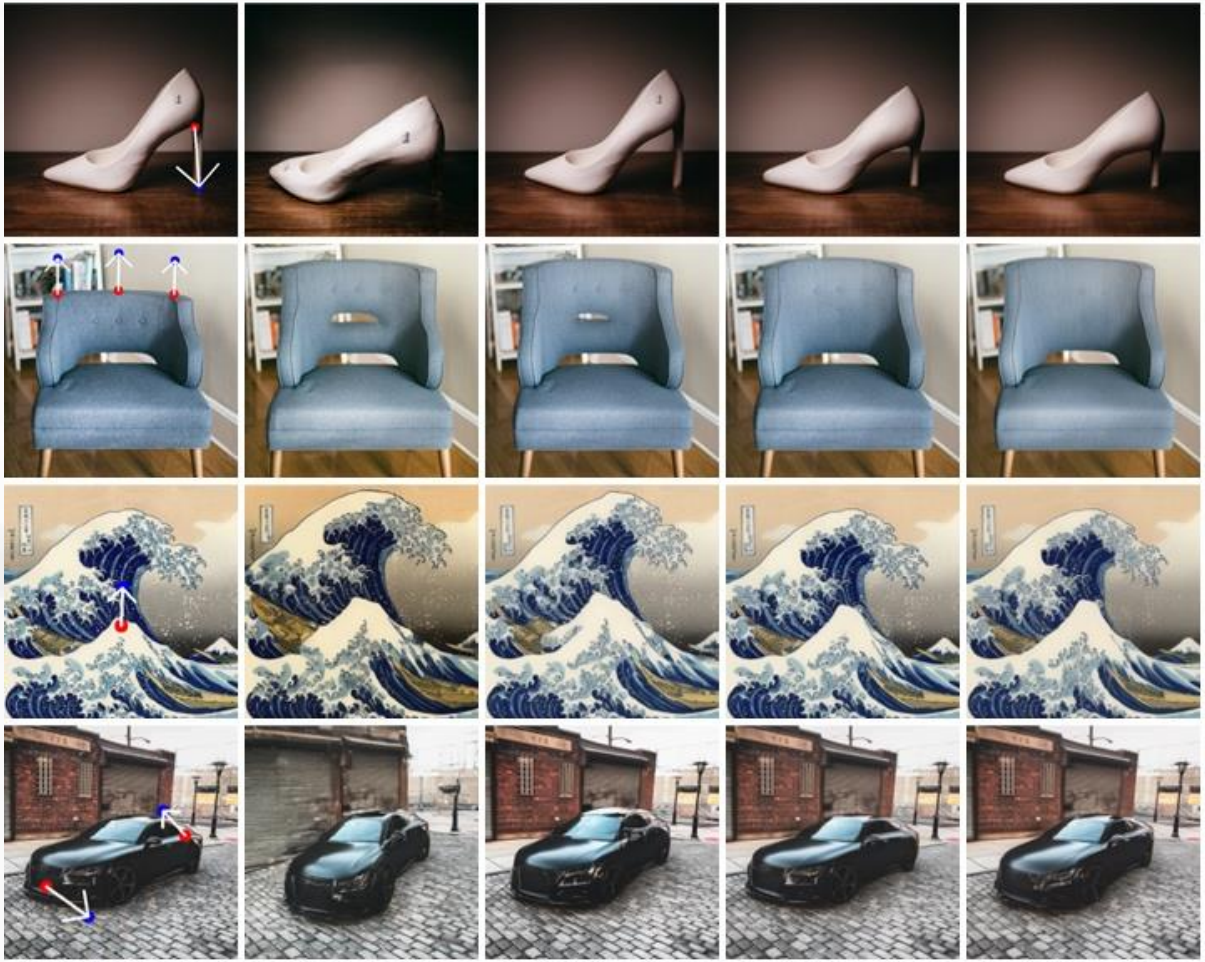


Qualitative Results of the Ablation Study






Drag Instruction
baseline
+Soft Mask
+Readout Guided
+Latent Warpage



Thank you for your attention!

國立臺灣科技大學

NATIONAL TAIWAN UNIVERSITY OF SCIENCE AND TECHNOLOGY