Text-Driven Image Editing via Learnable Regions

Yuanze Lin[1] Yi-Wen Chen[2] Yi-Hsuan Tsai[3] Lu Jiang[3] Ming-Hsuan Yang[2,3]

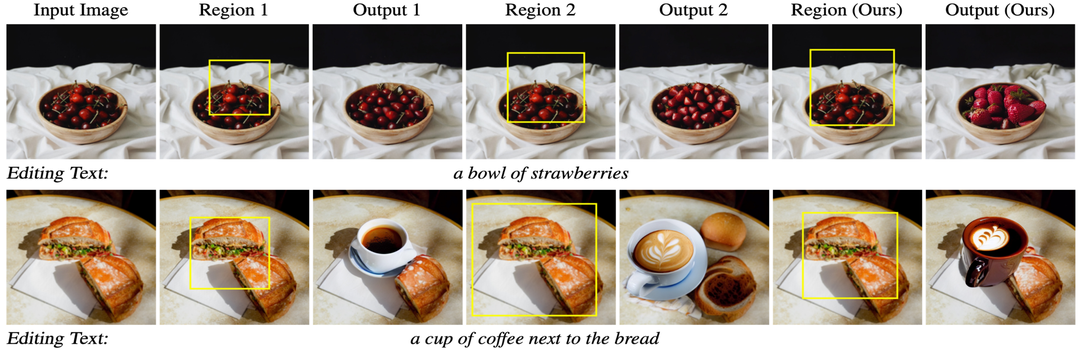
[1]University of Oxford [2]UC Merced [3]Google

Source code

CVPR SEATTLE, WA JUNE 17-21, 2024

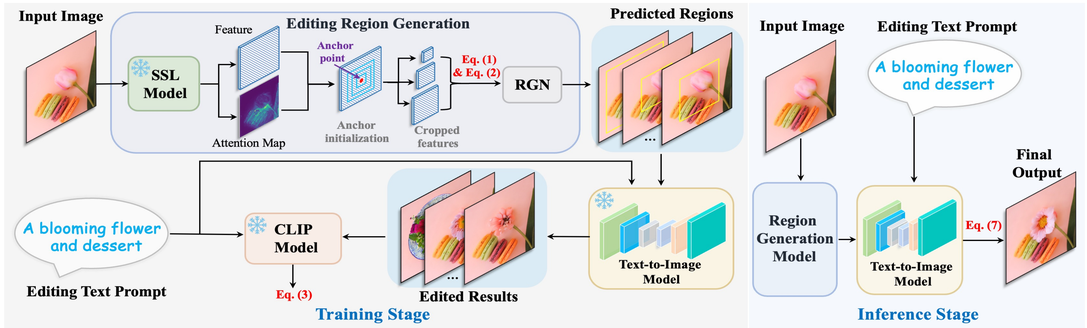
Motivation

Variations in editing regions can significantly influence the edited results!



- Explores to learn intuitive box regions for image local editing
- It can be integrated with other text-to-image models
- Solves complex prompts with multiple objects and extended length

Method



- Feature and anchor initialization from the SSL model
- Train region generation network to obtain editing regions
- Inference by quality score: $S = \alpha \cdot S_{\text{t2i}} + \beta \cdot S_{\text{i2i}}$

$$\mathcal{L} = \lambda_C \mathcal{L}_{Clip} + \lambda_S \mathcal{L}_{Str} + \lambda_D \mathcal{L}_{Dir},$$

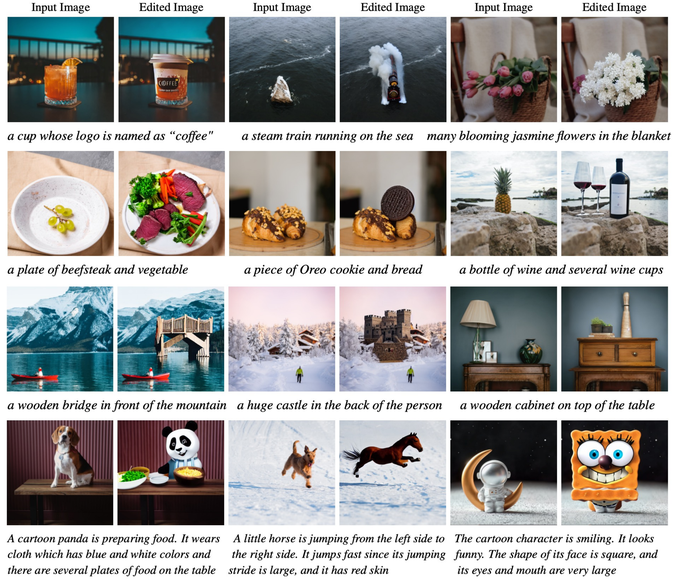
$$\mathcal{L}_{Clip} = \mathcal{D}_{cos}(E_v(X_o), E_t(T)),$$

$$\mathcal{L}_{Str} = ||Q(f_{X_o}) - Q(f_X)||_2,$$

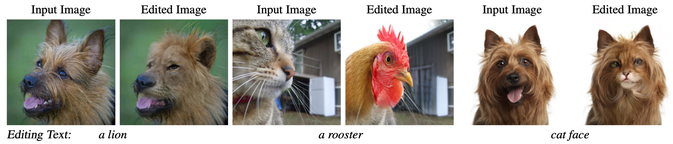
$$\mathcal{L}_{Dir} = \mathcal{D}_{cos}(E_v(X_o) - E_v(X), E_t(T) - E_t(T_{\text{ROI}}))$$

Experiments

Image editing results with simple and complex prompts

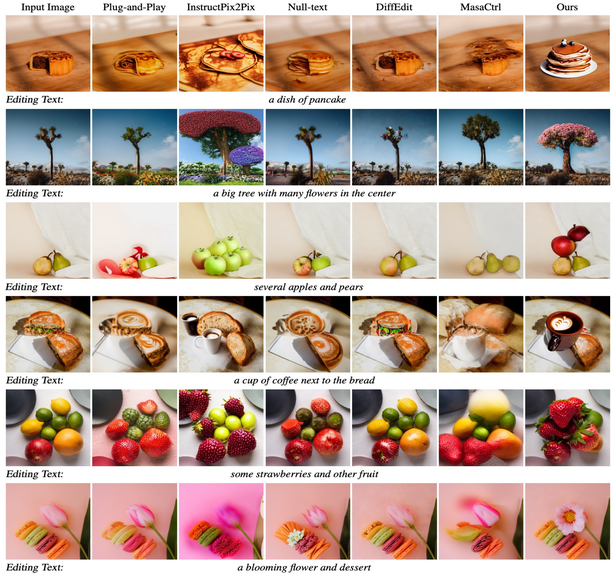


Compatibility with image synthesis models (MaskGiT)



Effect of different loss components

Comparison with existing methods



User study

Compared Methods	Preference for Ours
vs. Plug-and-Play	80.5% ±1.9%
vs. InstructPix2Pix	73.2% ±2.2%
vs. Null-text	88.2% ±1.6%
vs. DiffEdit	91.9% ±1.3%
vs. MasaCtrl	90.8% ±1.4%
Average	**84.9%**

Effect of region generation methods

Compared Methods	Preference for Ours
vs. Random-anchor-random-size	83.9% ±2.6%
vs. DINO-anchor-random-size	71.0% ±3.2%