

Text-Driven Image Editing via Learnable Regions

Yuanze Lin¹ Yi-Wen Chen² Yi-Hsuan Tsai³ Lu Jiang³ Ming-Hsuan Yang^{2,3}

¹University of Oxford

²UC Merced

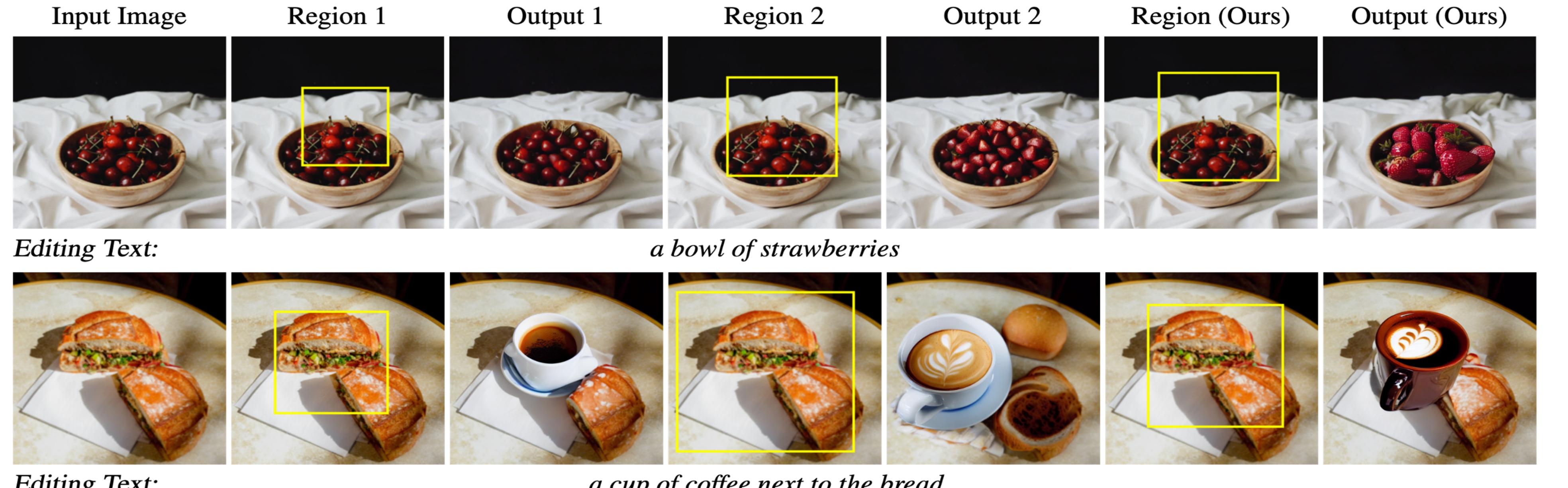
³Google



Source code

Motivation

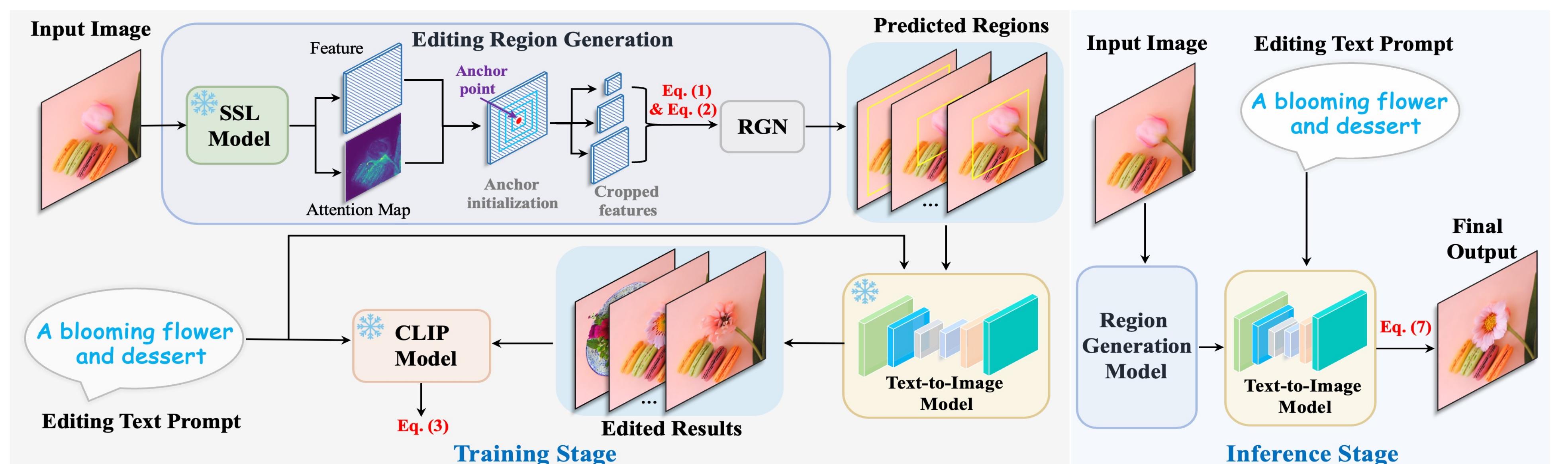
Variations in editing regions can significantly influence the edited results!



Editing Text:
a bowl of strawberries

- 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

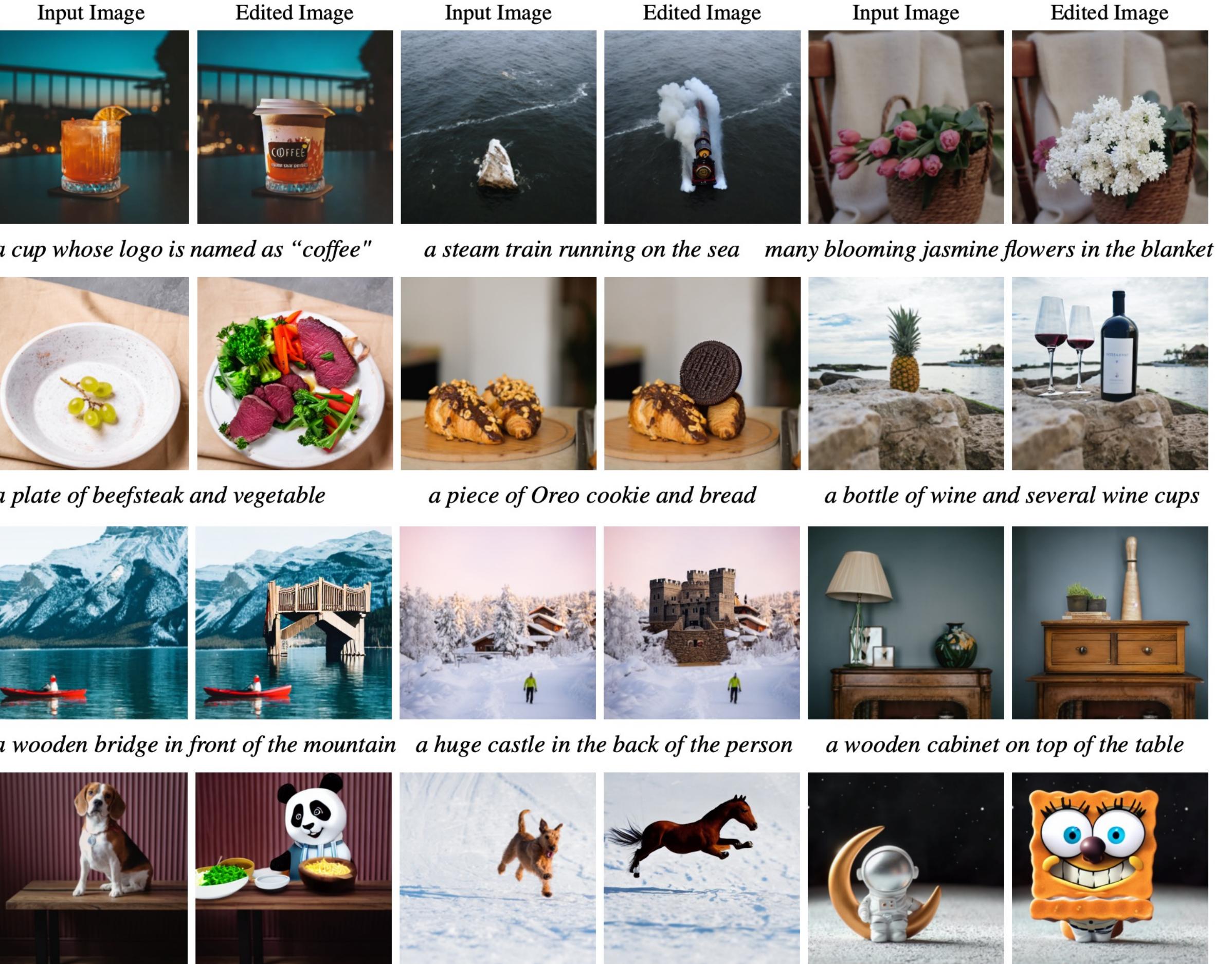


- Feature and anchor initialization from the SSL model
- Train region generation network to obtain editing regions
- Inference by quality score: $S = \alpha \cdot S_{t2i} + \beta \cdot S_{i2i}$

$$\begin{aligned}\mathcal{L} &= \lambda_C \mathcal{L}_{Clip} + \lambda_S \mathcal{L}_{Str} + \lambda_D \mathcal{L}_{Dir}, \\ \mathcal{L}_{Clip} &= \mathcal{D}_{cos}(E_v(X_o), E_t(T)), \\ \mathcal{L}_{Str} &= \|Q(f_{X_o}) - Q(f_X)\|_2, \\ \mathcal{L}_{Dir} &= \mathcal{D}_{cos}(E_v(X_o) - E_v(X), E_t(T) - E_t(T_{ROI}))\end{aligned}$$

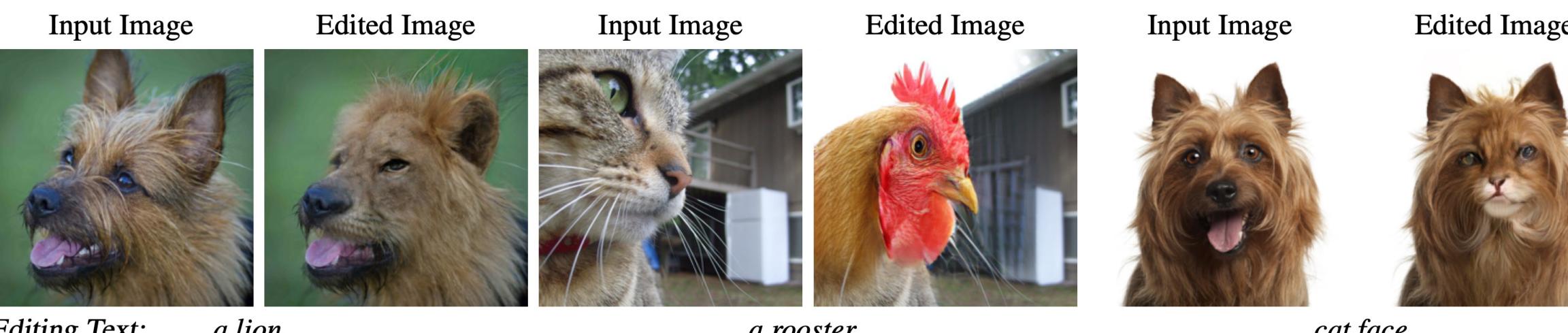
Experiments

Image editing results with simple and complex prompts



A cartoon panda is preparing food. It wears cloth which has blue and white colors and there are several plates of food on the table
A little horse is jumping from the left side to the right side. It jumps fast since its jumping stride is large, and it has red skin
The cartoon character is smiling. It looks funny. The shape of its face is square, and its eyes and mouth are very large

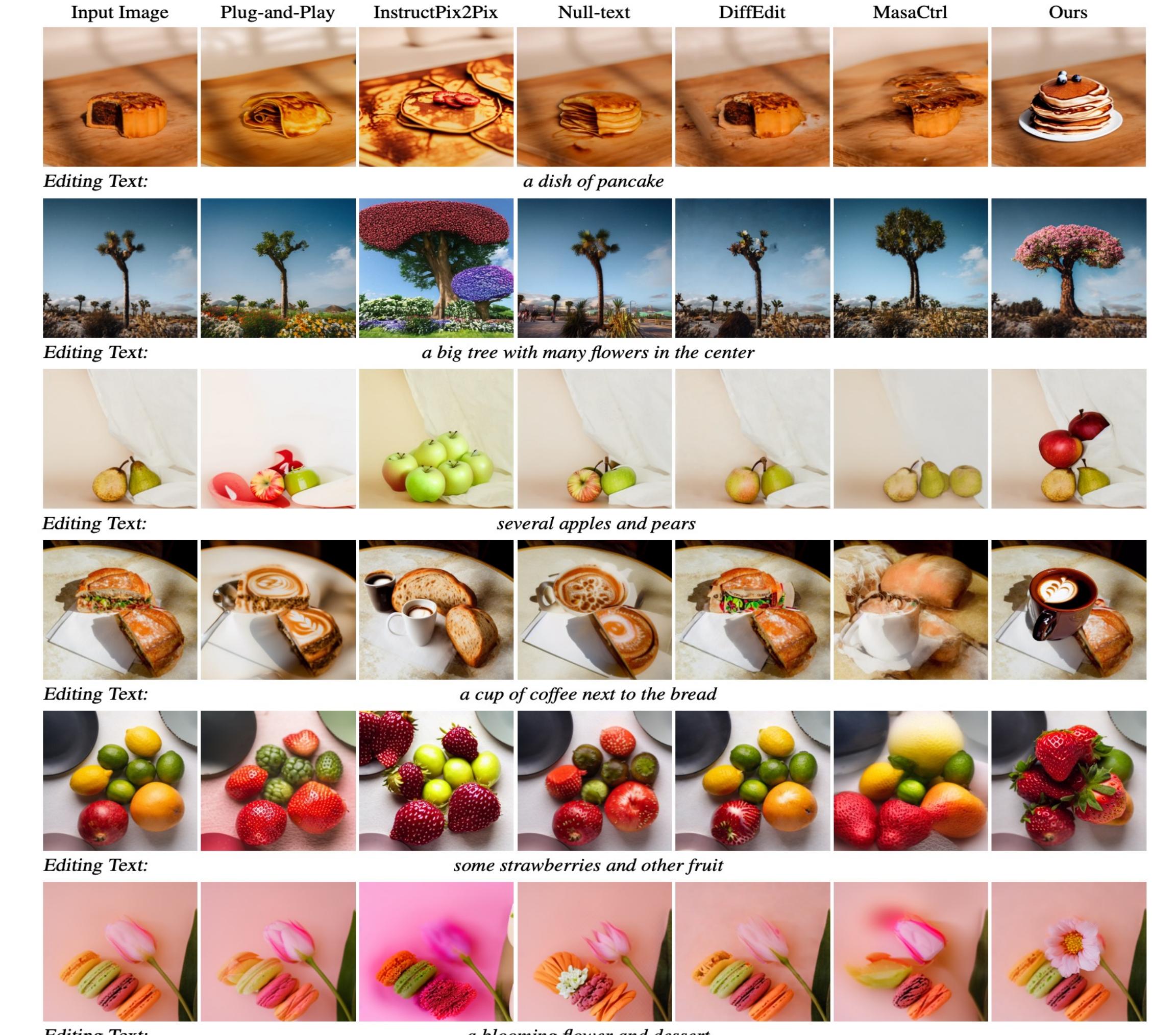
Compatibility with image synthesis models (MaskGiT)



Effect of different loss components



Comparison with existing methods



User study

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

Effect of region generation methods

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