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**

Motivation

Previous text-driven image editing methods:

(1) Rely on either user-provided masks (mask-based) or learning fine-grained pixel masks as editing regions (mask-free)

(2) Hinge on learning of precise pixel masks, inaccuracies in these pixel masks could unintentionally affect the global visual presentation

(3) Some text-to-image models like Muse [9], cannot support the pixel masks as regions for editing

Our method:

(1) Explores to learn intuitive box regions for image local editing

(2) Can be integrated with other text-to-image models

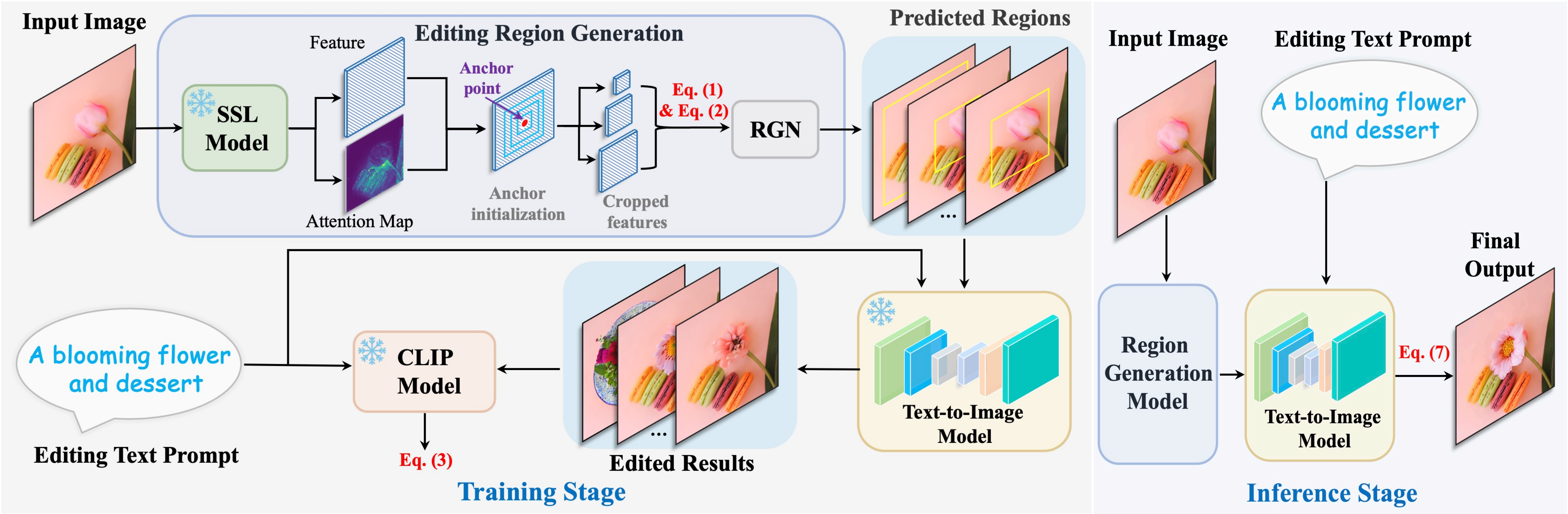
(3) Solves complex prompts with multiple objects and extended length

Variations in editing regions can significantly influence the edited results!

Overview


Input Image
Edited Image
Input Image
Edited Image
Input Image
Edited Image
donuts with ice cream
a cute sloth holds a box
a flowering cherry tree
a blooming flower and dessert
several lotus flowers growing in the water
a cup of coffee next to the bread

Method


Input Image
Editing Region Generation
Feature
SSL Model
Anchor point
Eq. (1) & Eq. (2)
RGN
Attention Map
Anchor initialization
Cropped features
Predicted Regions
...
A blooming flower and dessert
Editing Text Prompt
CLIP Model
Eq. (3)
Edited Results
...
Text-to-Image Model
Training Stage
Input Image
Editing Text Prompt
A blooming flower and dessert
Final Output
Region Generation Model
Text-to-Image Model
Eq. (7)
Inference Stage

Training details

Training objective:

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

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

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

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

L_{Str}: Structural loss

L_{Dir}: Directional loss

L_{Clip}: Clip guidance loss

Training setting:

Anchor initialization: 8 anchor points & 7 region proposals

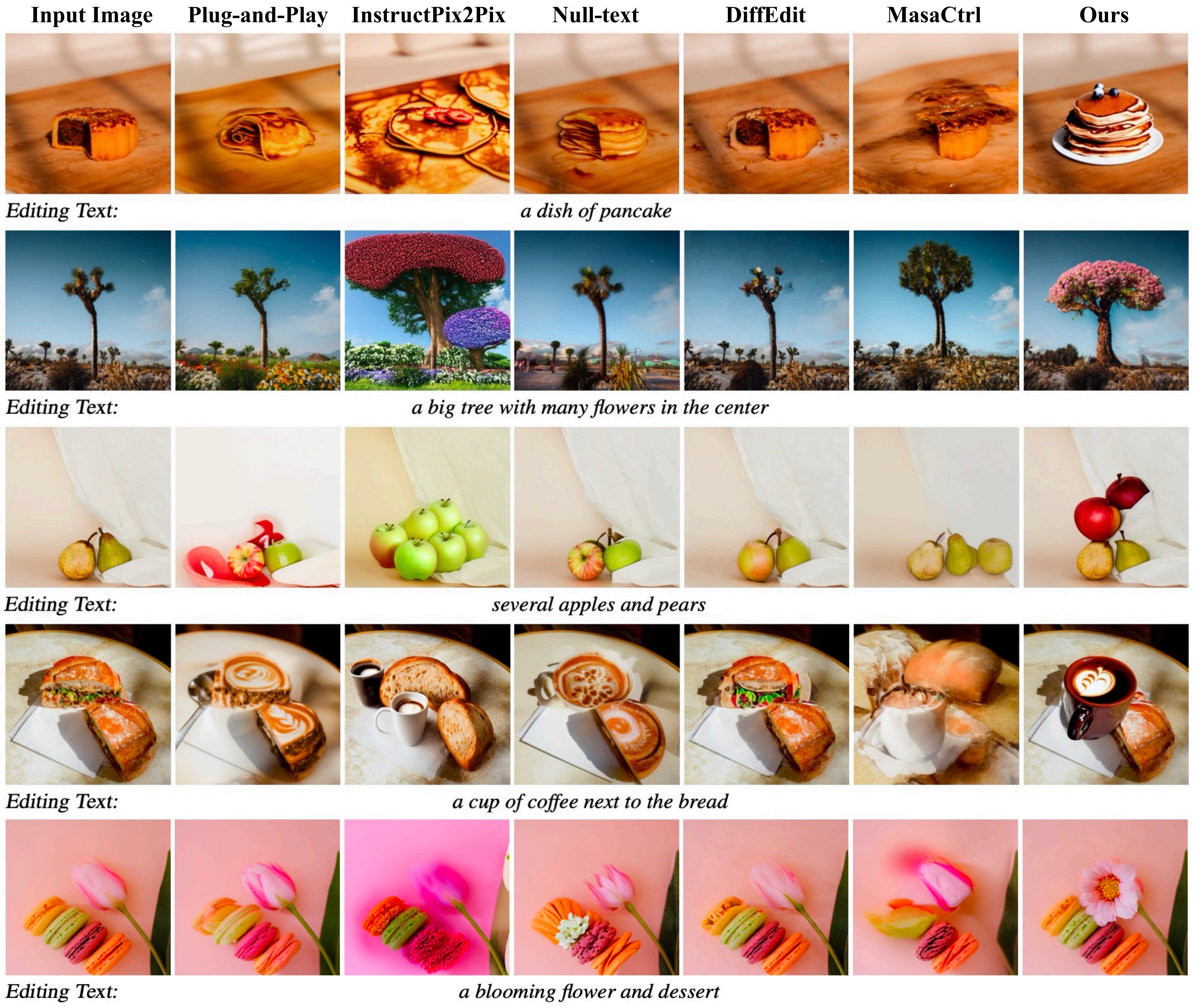
CLIP guidance model: ViT-B/16 weights

Editing model: Stable Diffusion-v-1-2

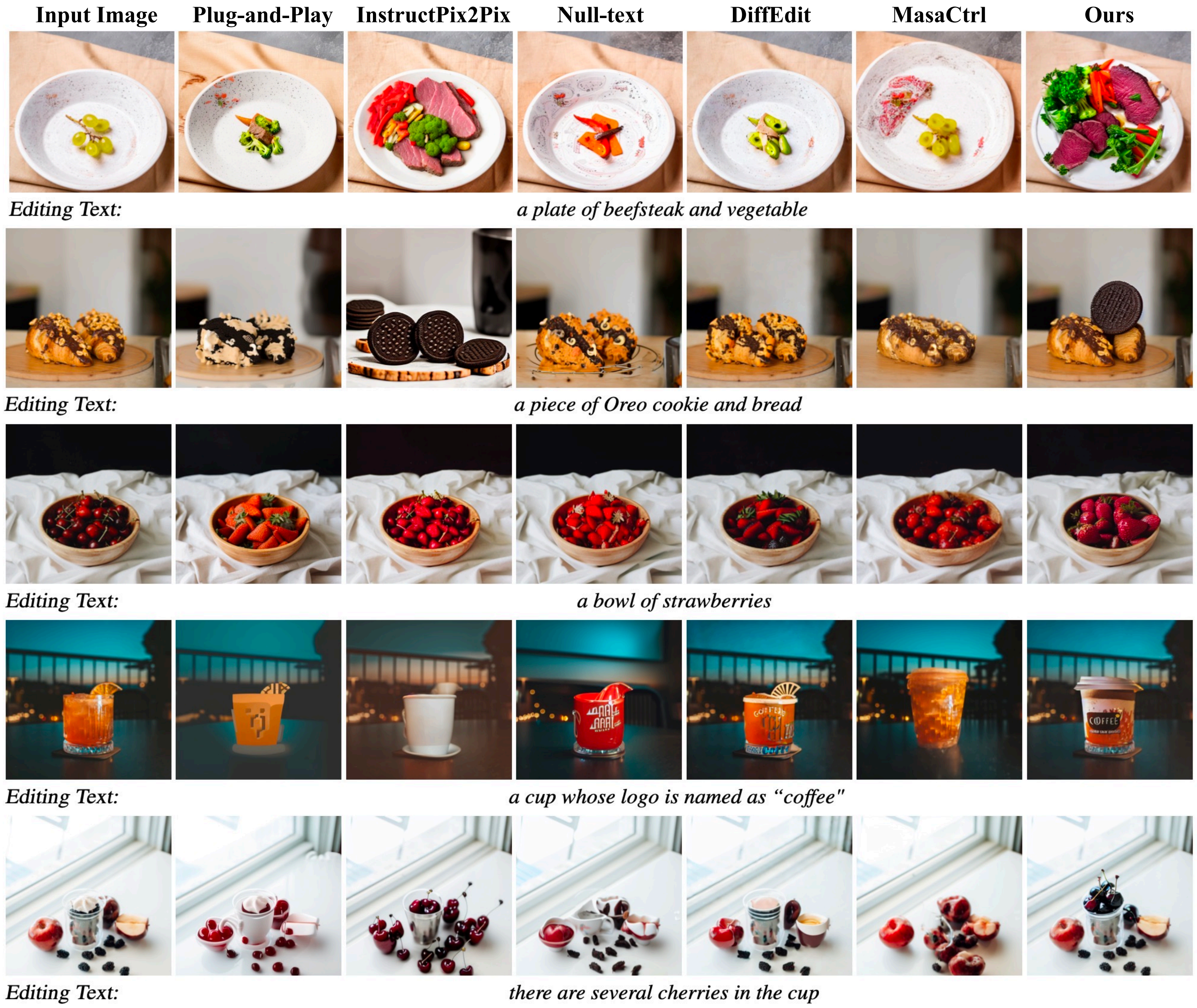
Data source: Unsplash

Training strategy: 2 A5000 GPUs, 5 epochs, Adam optimizer, 0.003 learning rate, batch size is 1

Experiments



Experiments



Experiments

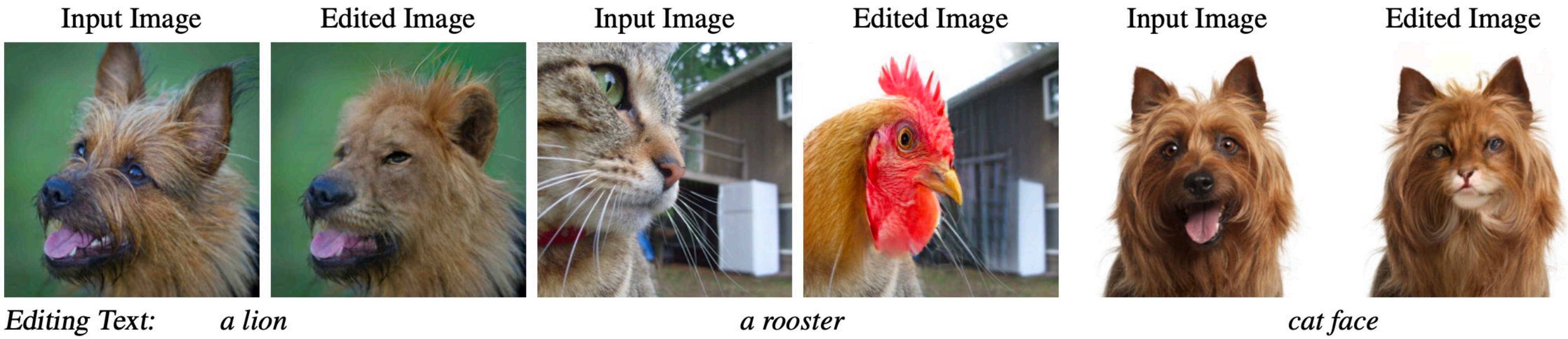
Collect 60 samples from **Unsplash** for this experiment

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%**

203 participants in this user study!

Experiments (ablation studies)

Generalizability of proposed method (integrating with MaskGiT):



Effect of different loss components:

L_{Dir}: Directional loss

L_{Str}: Structural loss

Experiments (ablation studies)

Effect of region generation methods (user study)

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

Effect of loss components

Loss Component	$S_{t2i} \uparrow$	$S_{i2i} \uparrow$
$\mathcal{L}_{Clip}$	0.301	0.801
$\mathcal{L}_{Clip} + L_{Str}$	0.294	0.806
$\mathcal{L}_{Clip} + L_{Str} + L_{Dir}$	0.301	0.805

L_{Str}: Structural loss L_{Dir}: Directional loss L_{Clip}: Clip guidance loss

Impact of the number of region proposals

# of region proposals	$S_{t2i} \uparrow$	$S_{i2i} \uparrow$
1	0.231	0.915
3	0.273	0.837
5	0.295	0.809
7	0.300	0.805
9	0.301	0.802

Impact of the number of anchor points

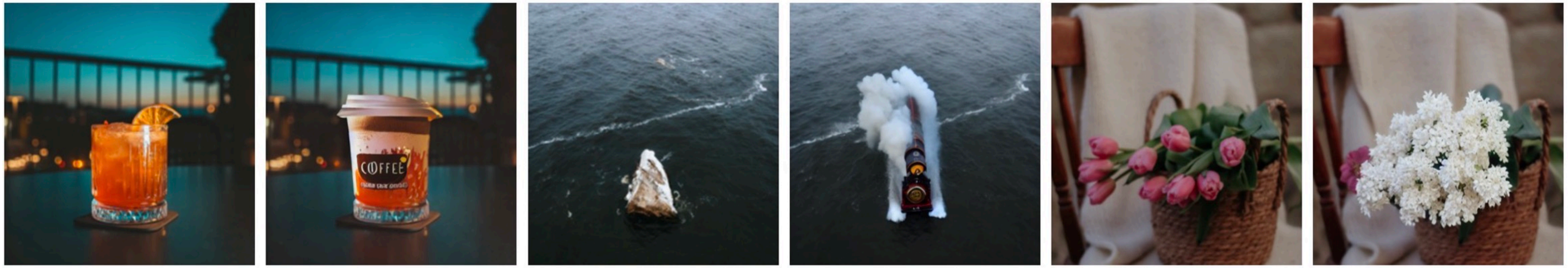
# of anchor points	$S_{t2i} \uparrow$	$S_{i2i} \uparrow$
1	0.275	0.824
4	0.296	0.805
6	0.301	0.802
8	0.300	0.805
10	0.298	0.803

Default settings are highlighted with blue!

S_{t2i}: CLIP's text-to-image similarity score S_{i2i}: CLIP's image-to-image similarity score

Results with diverse prompts

Various prompts for one kind of object:



a cup whose logo is named as "coffee"   *a steam train running on the sea*   *many blooming jasmine flowers in the blanket*

Prompts featuring multiple objects:

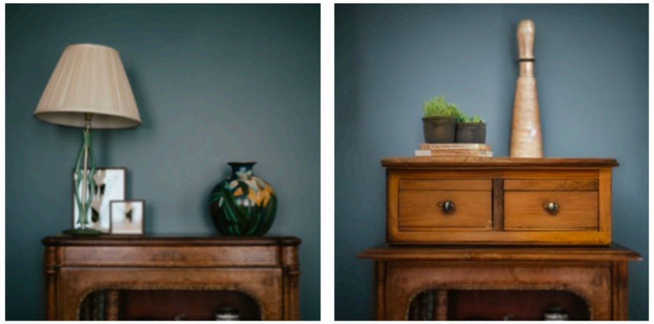
a plate of beefsteak and vegetable   *a piece of Oreo cookie and bread*   *a bottle of wine and several wine cups*

Results with diverse prompts

Prompts with geometric relations:

a wooden bridge in front of the mountain

a huge castle in the back of the person

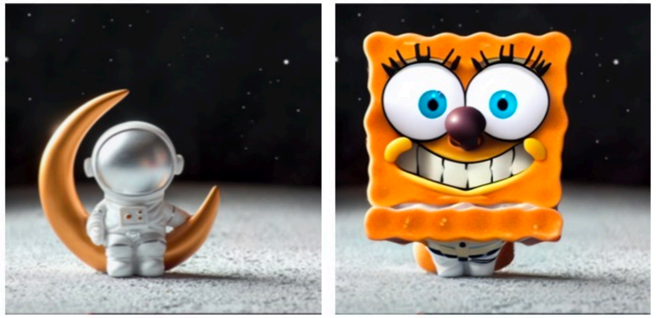


a wooden cabinet on top of the table

Prompts with long paragraphs:

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

Additional results

a cheerful snowman adorned with a carrot nose

a large steamer sails accompanied by sea gulls

the Chinese lantern adds cultural elegance, radiating a soft glow

a Christmas tree adorned with twinkling lights

a colossal whale shark floating on the top of the deep ocean

a parked caravan epitomizes travel tales

a securely moored white ship on the solid ground

beautiful cars with sleek lines and polished exterior

a delicate spider weaves an intricate web

Failure cases


Thank you!