

# **Image editing in text2img models**

Igor Ryabkov

**Problem:** a small edit to the prompt causes a global impact on the result image



A robotic fox in a spacesuit flies through a snow swirl, with lights and fir branches all around him



A robotic bear in a spacesuit flies through a snow swirl, with lights and fir branches all around him

## **Solutions:**

- Prompt-to-prompt
- Dreambooth

# Dreambooth

Goal: Using 3-5 images of the object, be able to generate it in different context



Input images



in the Acropolis



swimming



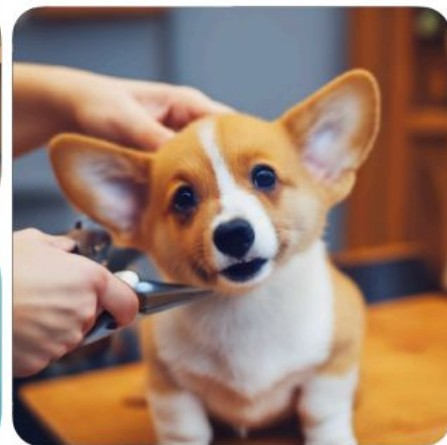
sleeping



in a doghouse



in a bucket



getting a haircut

# How can we specify our object to the model?

Using template: “[V] [class name]”

Example: “The [V] dog on the street”

## What exactly is [V]?

- Words like “special”, “my”, “unique”?
- Random characters: “xxy5syt00”?

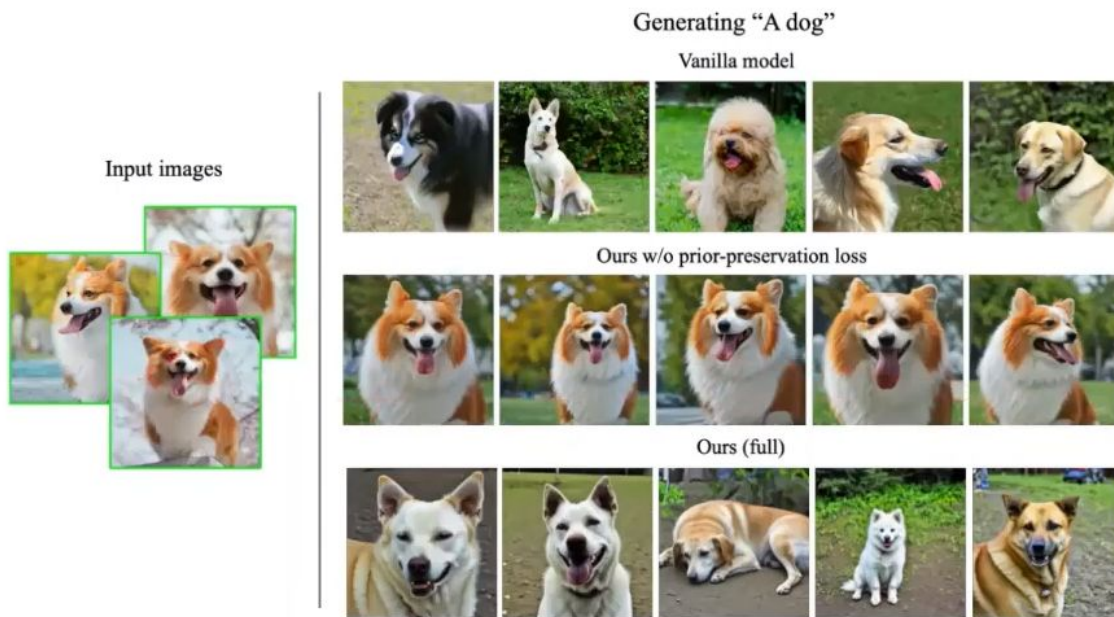
## **Best way to create [V]:**

- Choose rare tokens
- Find which words they refer to
- Concatenate them

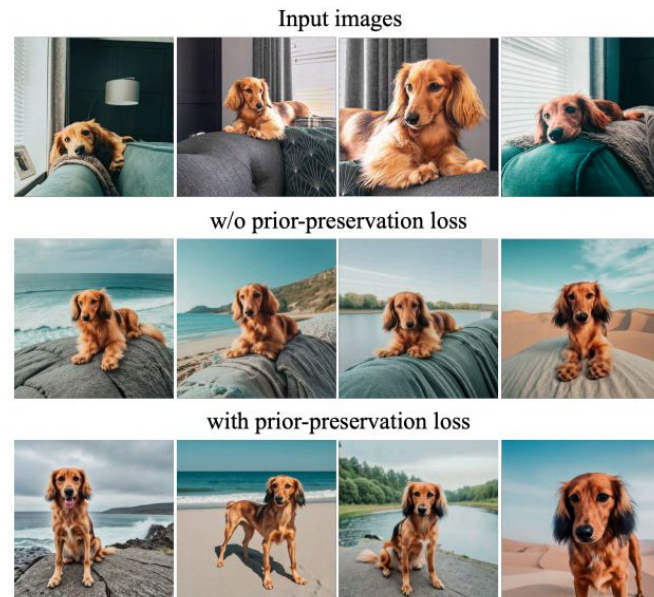
**Often used: “sks”**

Example: “The sks dog on the street”

# Problems:

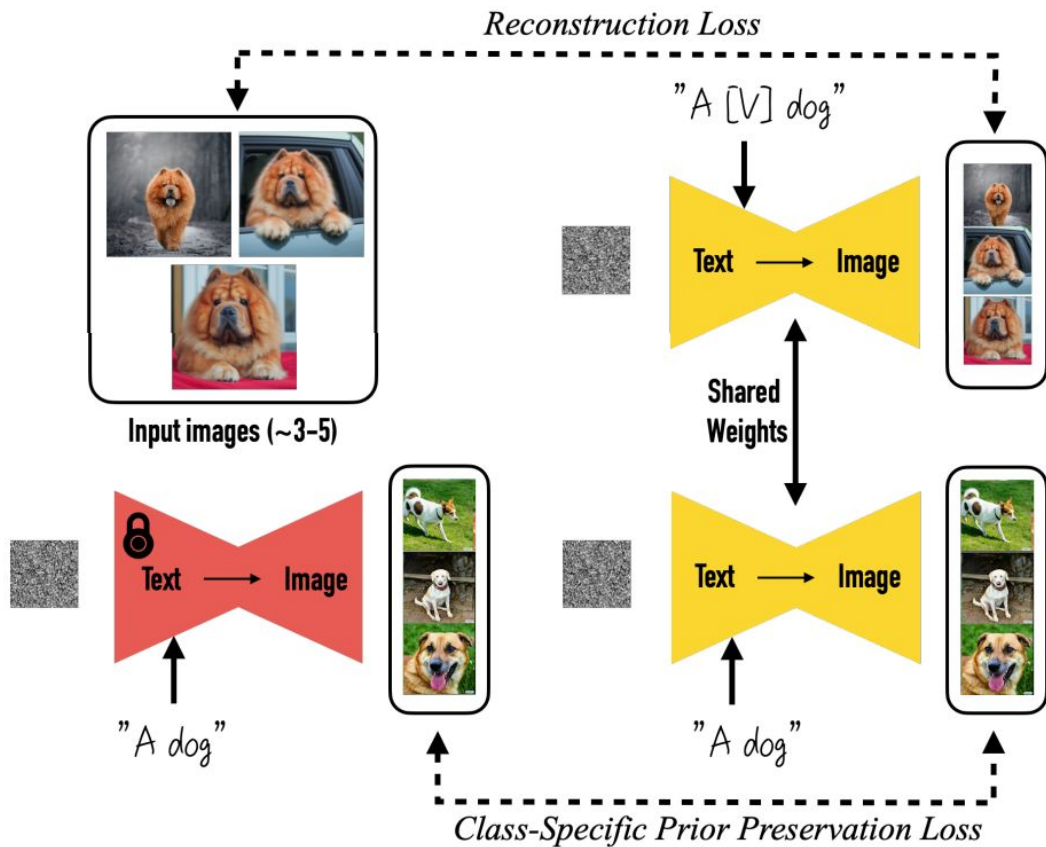


**Language drift**



**Overfitting**

# Fine-tuning:



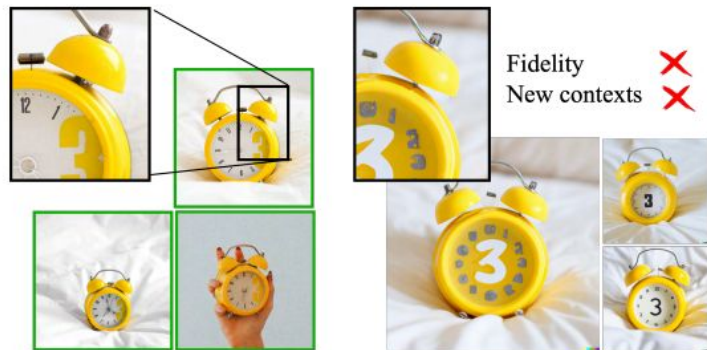
# Class-specific Prior Preservation Loss:

$$\mathbb{E}_{\mathbf{x}, \mathbf{c}, \epsilon, \epsilon', t} [w_t \|\hat{\mathbf{x}}_{\theta}(\alpha_t \mathbf{x} + \sigma_t \epsilon, \mathbf{c}) - \mathbf{x}\|_2^2 +$$

$$+ \lambda w_{t'} \|\hat{\mathbf{x}}_{\theta}(\alpha_{t'} \mathbf{x}_{\text{pr}} + \sigma_{t'} \epsilon', \mathbf{c}_{\text{pr}}) - \mathbf{x}_{\text{pr}}\|_2^2]$$



# Method capabilities:



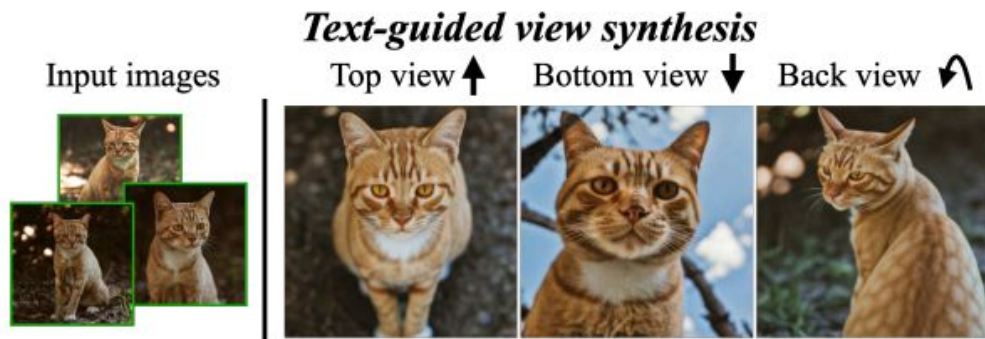
Input Images

Image-guided, DALL-E2



Text-guided, Imagen

Ours



Input images

*Text-guided view synthesis*

Top view ↑

Bottom view ↓

Back view ↗



*Art Renditions*

Van Gogh

Michelangelo

Vermeer



*Property Modification*

Panda

Lion

Hippo



# Method capabilities:



Input images



A [V] backpack in the Grand Canyon



A wet [V] backpack in water



A [V] backpack in Boston



A [V] backpack with the night sky



Input images



A [V] teapot floating in milk



A transparent [V] teapot with milk inside



A [V] teapot pouring tea



A [V] teapot floating in the sea



# Method limitations:

- incorrect context synthesis
- Context-appearance entanglement
- Overfitting

Input images



(a) Incorrect context synthesis



in the ISS



on the moon

(b) Context-appearance entanglement



in the Bolivian salt flats



on top of a blue fabric

(c) Overfitting



in the forest

# Comparison

Method	DINO ↑	CLIP-I ↑	CLIP-T ↑
Real Images	0.774	0.885	N/A
DreamBooth (Imagen)	<b>0.696</b>	<b>0.812</b>	<b>0.306</b>
DreamBooth (Stable Diffusion)	0.668	0.803	0.305
Textual Inversion (Stable Diffusion)	0.569	0.780	0.255

Input Images



DreamBooth (Imagen)



DreamBooth (Stable Diffusion)



Textual Inversion (Stable Diffusion)



# Comparison

Method	DINO $\uparrow$	CLIP-I $\uparrow$
Correct Class	<b>0.744</b>	<b>0.853</b>
No Class	0.303	0.607
Wrong Class	0.454	0.728

## Fine-tuning

No class noun:  
“A [V]”

Incorrect class noun:  
“A [V] dog”

Correct class noun:  
“A [V] sunglasses”

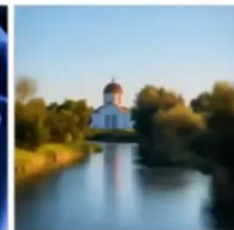
## Inference



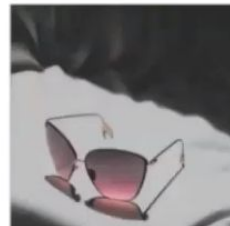
A [V]



A [V] on top of  
blue fabric



A [V] with a river in  
the background



A [V] dog



A [V] dog on top of  
blue fabric



A [V] dog with a river in  
the background



A [V] sunglasses



A [V] sunglasses on  
top of blue fabric

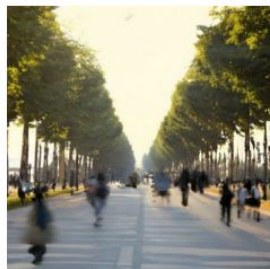


A [V] sunglasses with  
a river in the background



# Prompt-to-prompt

**Goal:** without fine-tuning, be able to change a small aspect in a picture: change a word, add some context, change the intensity of a word



“The boulevards are crowded today.”



“Photo of a cat riding on a ~~bicycle~~ car.”

car



“Landscape with a house near a river

and a rainbow in the background?”



“My fluffy bunny doll.”



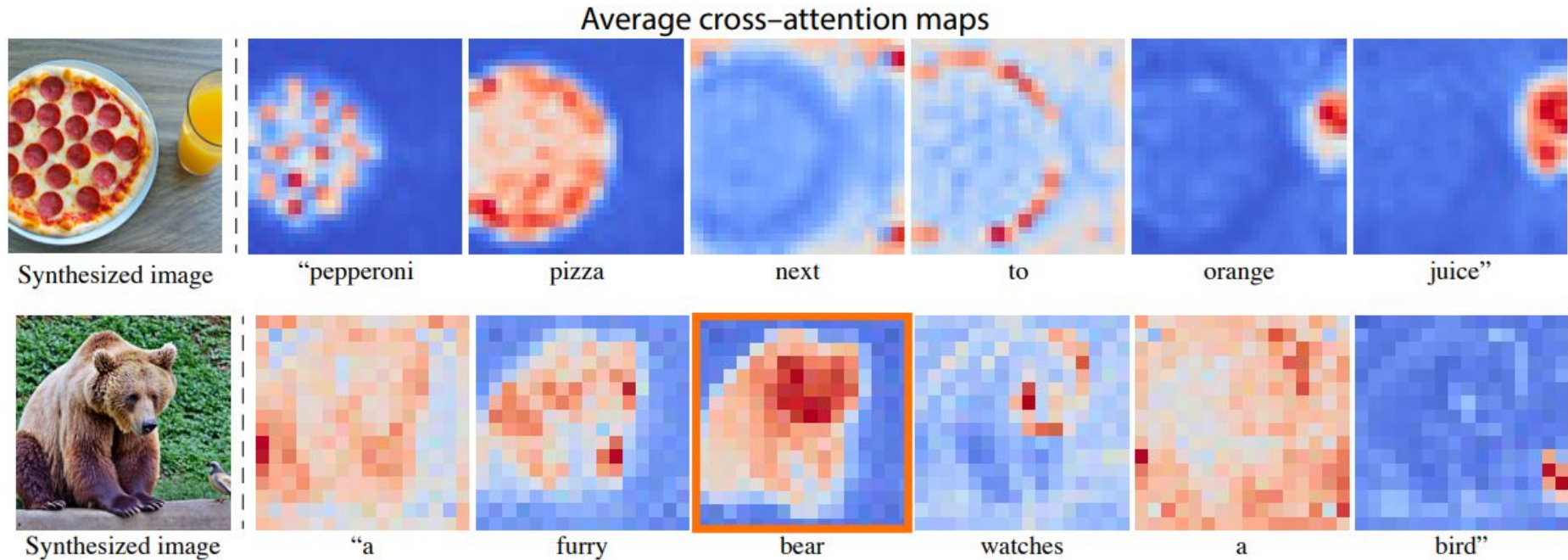
“a cake with decorations.”

jelly beans



“Children drawing of a castle next to a river.”

# The importance of Cross-Attention Layer



It specifies the pixels that need to be changed

# Algorithm

---

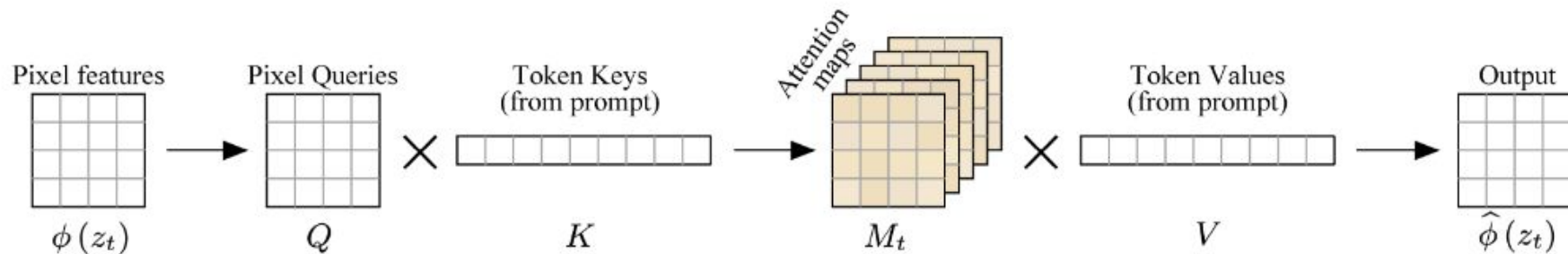
**Algorithm 1:** Prompt-to-Prompt image editing

---

- 1 **Input:** A source prompt  $\mathcal{P}$ , a target prompt  $\mathcal{P}^*$ , and a random seed  $s$ .
  - 2 **Optional for local editing:**  $w$  and  $w^*$ , words in  $\mathcal{P}$  and  $\mathcal{P}^*$ , specifying the editing region.
  - 3 **Output:** A source image  $x_{src}$  and an edited image  $x_{dst}$ .
  - 4  $z_T \sim N(0, I)$  a unit Gaussian random variable with random seed  $s$ ;
  - 5  $z_T^* \leftarrow z_T$ ;
  - 6 **for**  $t = T, T - 1, \dots, 1$  **do**
    - 7  $z_{t-1}, M_t \leftarrow DM(z_t, \mathcal{P}, t, s)$ ;
    - 8  $M_t^* \leftarrow DM(z_t^*, \mathcal{P}^*, t, s)$ ;
    - 9  $\widehat{M}_t \leftarrow Edit(M_t, M_t^*, t)$ ;
    - 10  $z_{t-1}^* \leftarrow DM(z_t^*, \mathcal{P}^*, t, s)\{M \leftarrow \widehat{M}_t\}$ ;
    - 11 **if** *local* **then**
      - 12  $\alpha \leftarrow B(\overline{M}_{t,w}) \cup B(\overline{M}_{t,w}^*)$ ;
      - 13  $z_{t-1}^* \leftarrow (1 - \alpha) \odot z_{t-1} + \alpha \odot z_{t-1}^*$ ;
    - 14 **end**
  - 15 **end**
  - 16 **Return**  $(z_0, z_0^*)$
-

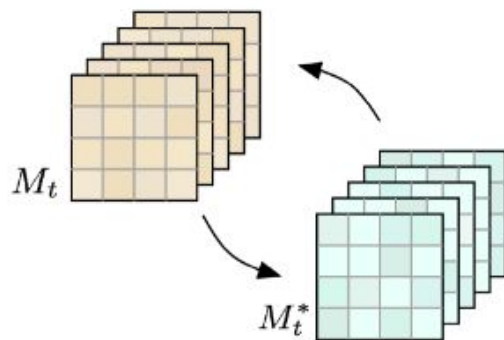


# Cross-Attention Layer

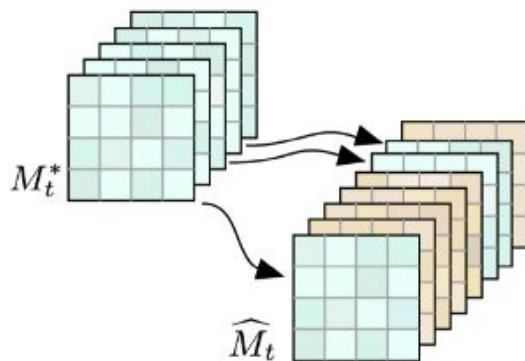


Text to Image Cross Attention

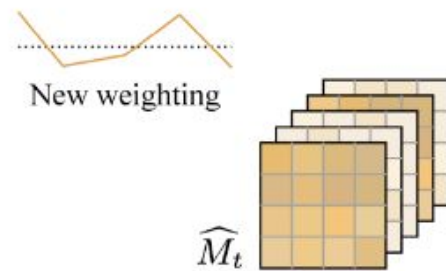
Cross Attention Control



Word Swap



Prompt Refinement



Attention Re-weighting

**Word swap:**

$$\text{Edit}(M_t, M_t^*, t) := \begin{cases} M_t^* & \text{if } t < \tau \\ M_t & \text{otherwise,} \end{cases}$$

fixed attention maps and random seed



fixed random seed



# Prompt refinement:

$$(Edit(M_t, M_t^*, t))_{i,j} := \begin{cases} (M_t^*)_{i,j} & \text{if } A(j) = None \\ (M_t)_{i,A(j)} & \text{otherwise.} \end{cases}$$

“A photo of a bear wearing sunglasses and having a drink.”



Source image



“...wearing a **squared** sunglasses...”



“...**colorful** sunglasses...”



“...**ski** sunglasses...”



“...**geeky** sunglasses...”



“...**beer** drink.”



“...**coffee** drink.”



“...**wheatgrass** drink.”

Local description



Source image



“...mat black car...”



“...sport car...”



“...old car...”



“...the blossom street.”

Global description



“...at sunset.”



“...in Manhattan.”



# Attention Re-weighting:

$$(Edit(M_t, M_t^*, t))_{i,j} := \begin{cases} c \cdot (M_t)_{i,j} & \text{if } j = j^* \\ (M_t)_{i,j} & \text{otherwise.} \end{cases}$$

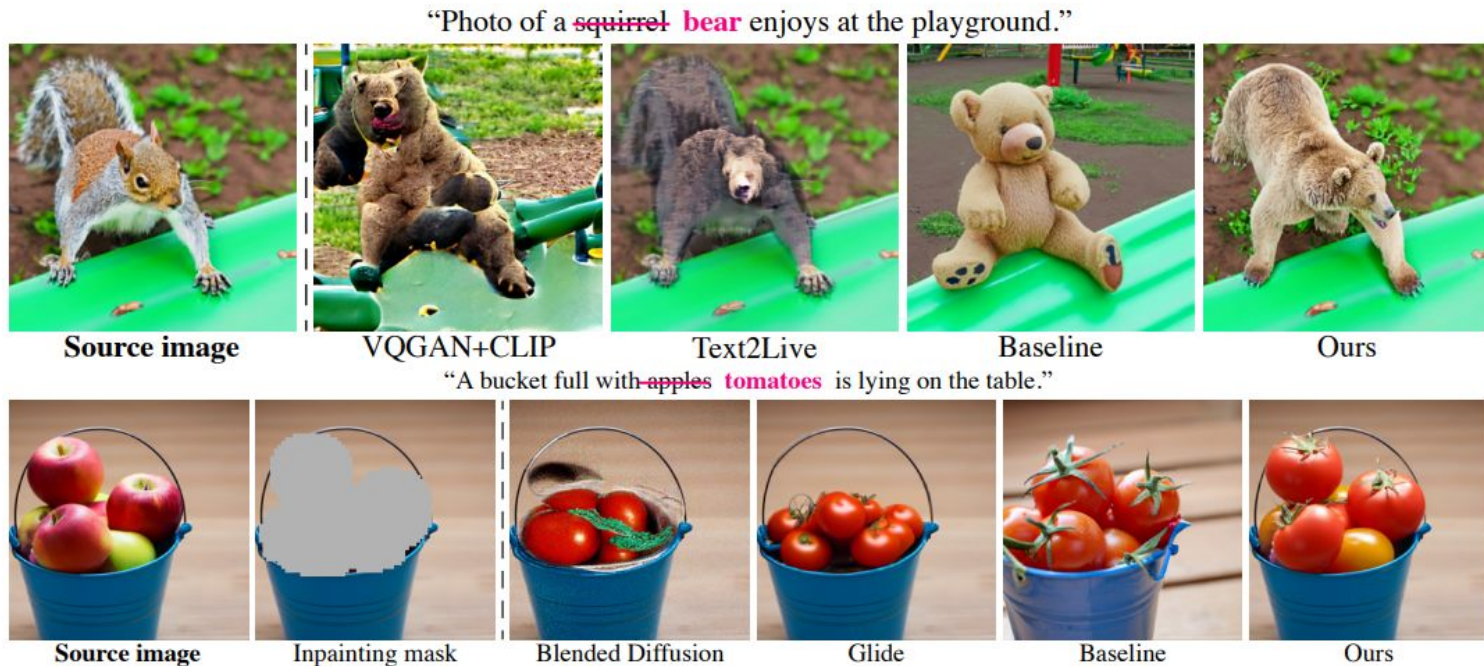


“The picnic is ready under a blossom(↓) tree.”



“My colorful(↓) bedroom.”

# Comparison



**Table 1: User Study results.** *The participants were asked to rate: (1) background / structure preservation with respect to the source image, (2) alignment to the text, and (3) realism.*

	VQGAN+CLIP	Text2Live	baseline	Ours
(1) Background / Structure $\uparrow$	$1.84 \pm 1.11$	$4.15 \pm 1.09$	$3.38 \pm 1.12$	$4.64 \pm 0.64$
(2) Text Alignment $\uparrow$	$2.46 \pm 1.16$	$2.89 \pm 1.22$	$4.26 \pm 1.03$	$4.55 \pm 0.71$
(3) Realism $\uparrow$	$1.32 \pm 0.70$	$2.36 \pm 1.12$	$4.11 \pm 0.93$	$4.42 \pm 0.82$