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

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”?

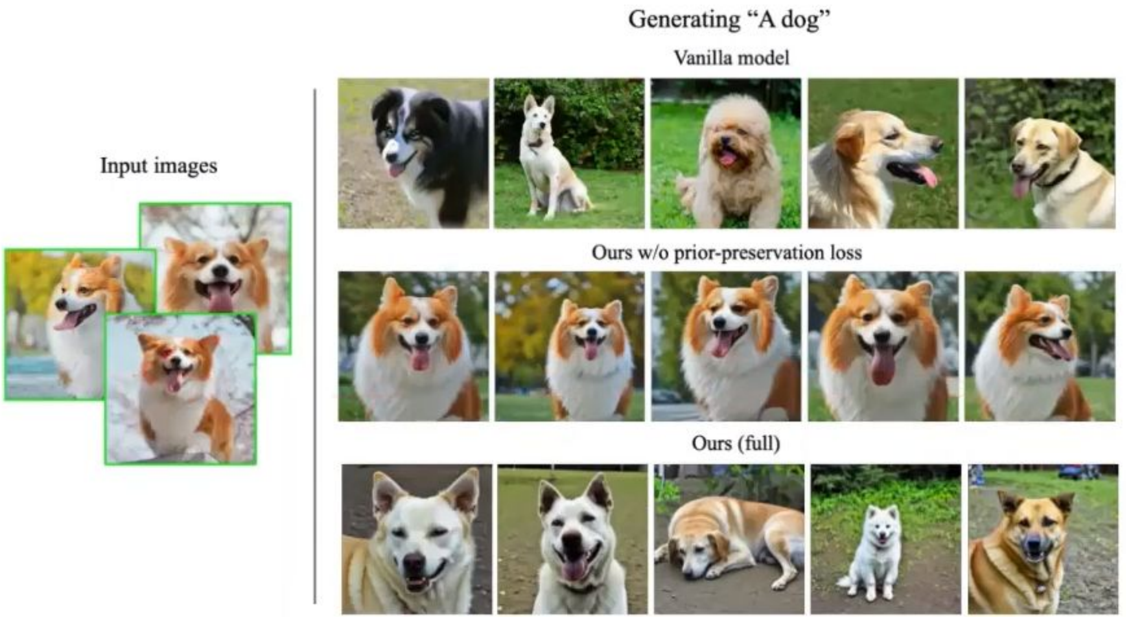
Best way to create [V]:

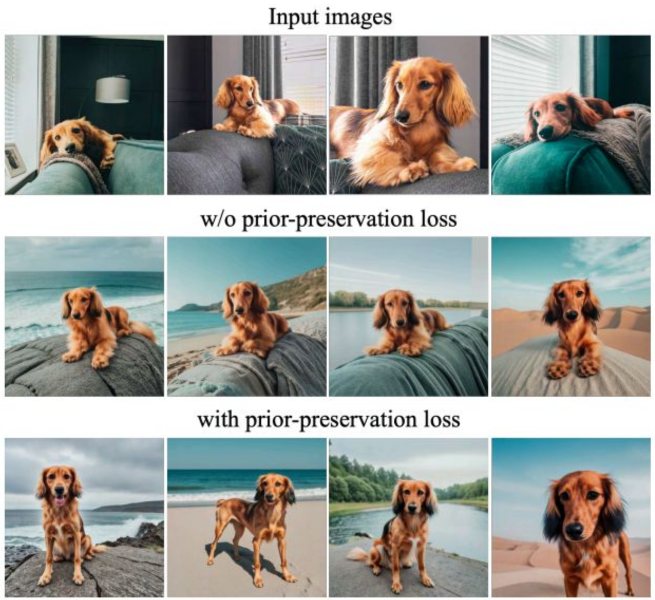
- Choose rare tokens
- Find which words they refer to
- Concatenate them

Often used: “sks”

Example: “The sks dog on the street”

Problems:



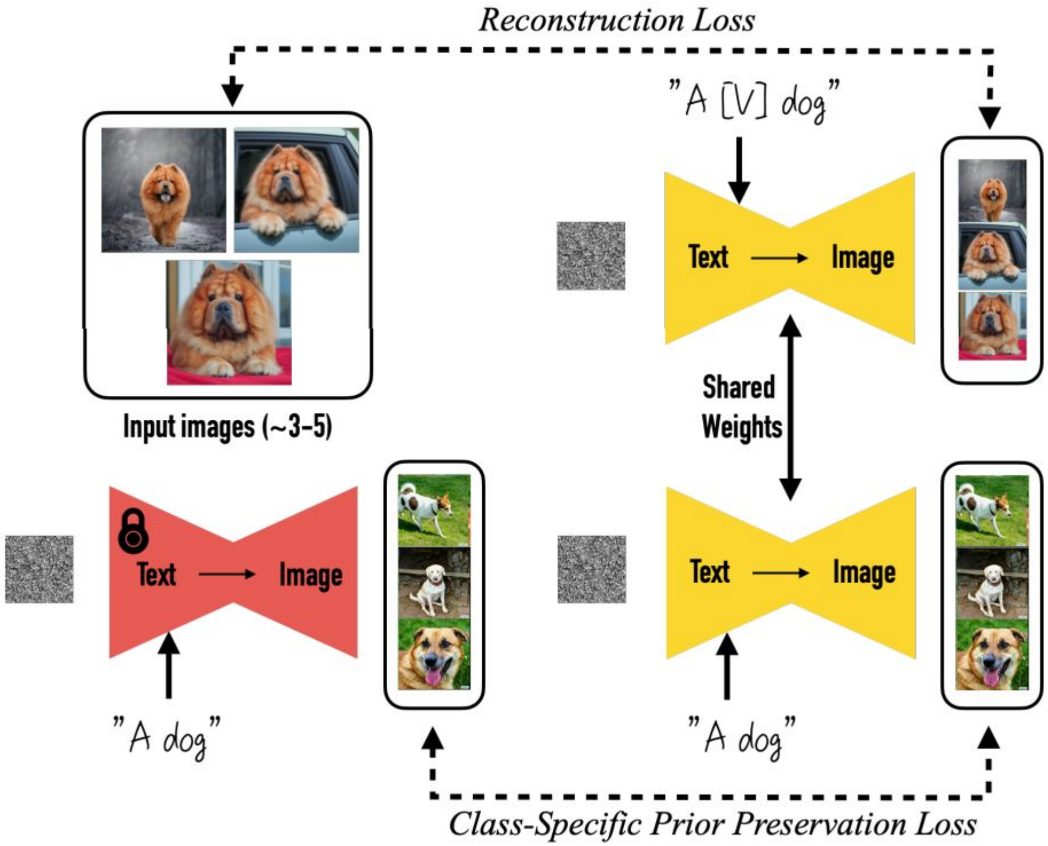


Language drift

Overfitting

Fine-tuning:

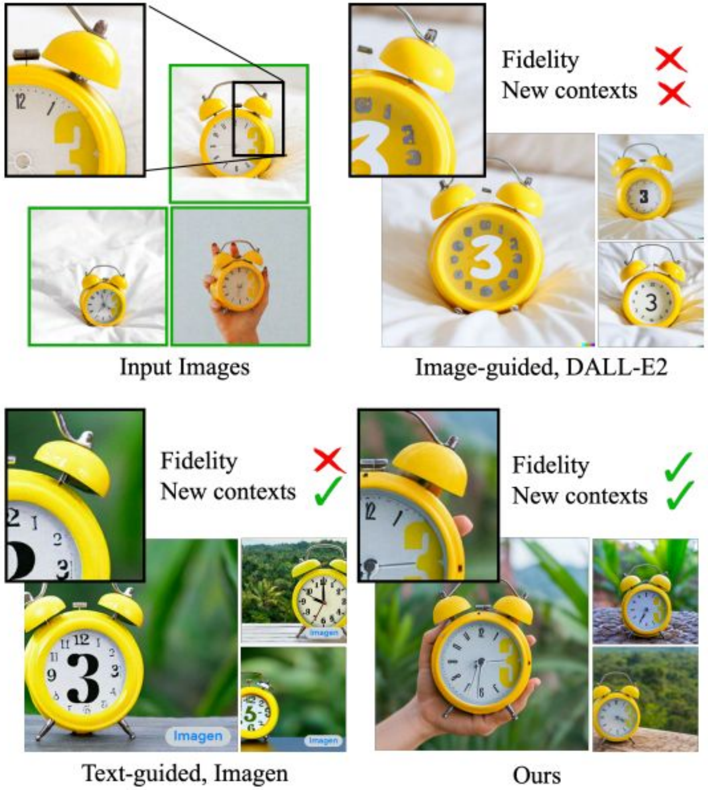
Class-specific Prior Preservation Loss:

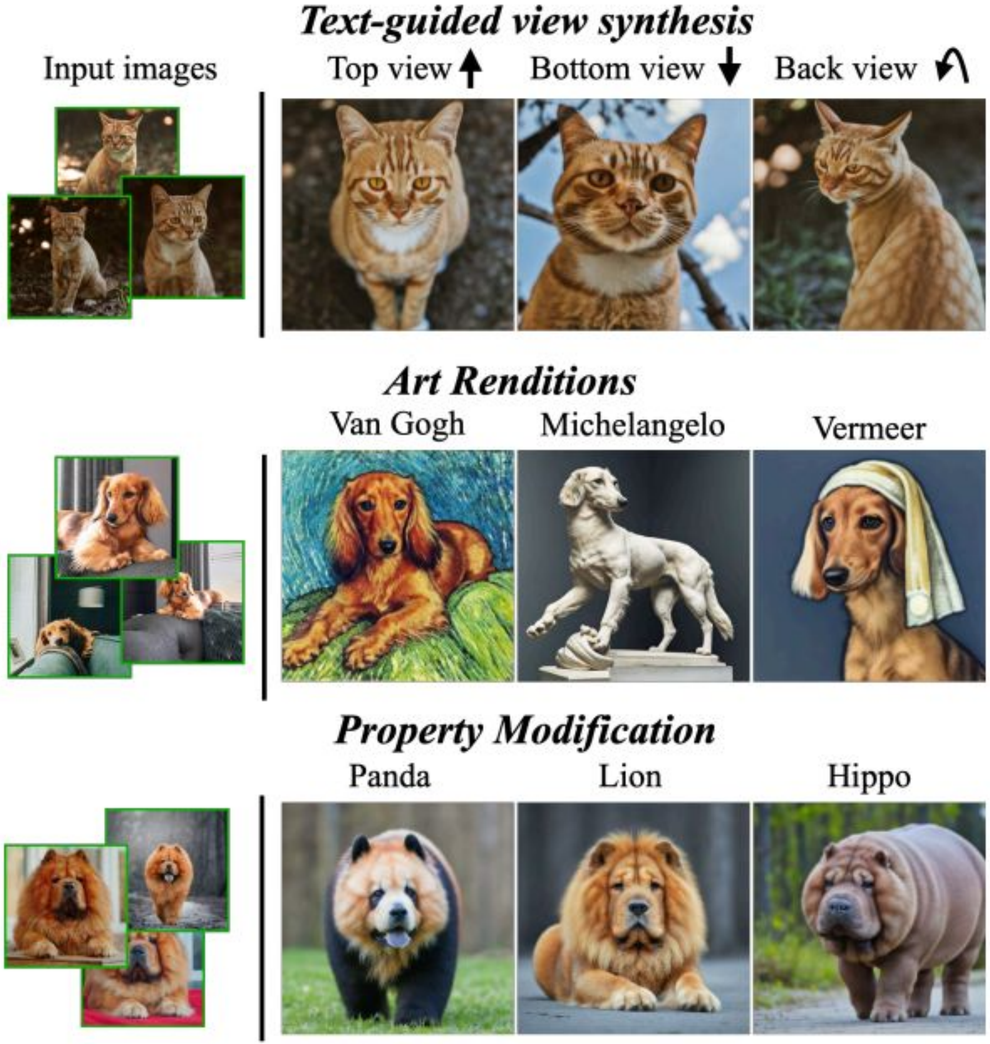


$$\mathbb{E}_{\mathbf{x},\mathbf{c},\boldsymbol{\epsilon},\boldsymbol{\epsilon}',t}[w_t\|\hat{\mathbf{x}}_\theta(\alpha_t\mathbf{x}+\sigma_t\boldsymbol{\epsilon},\mathbf{c})-\mathbf{x}\|_2^2+$$

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

Method capabilities:





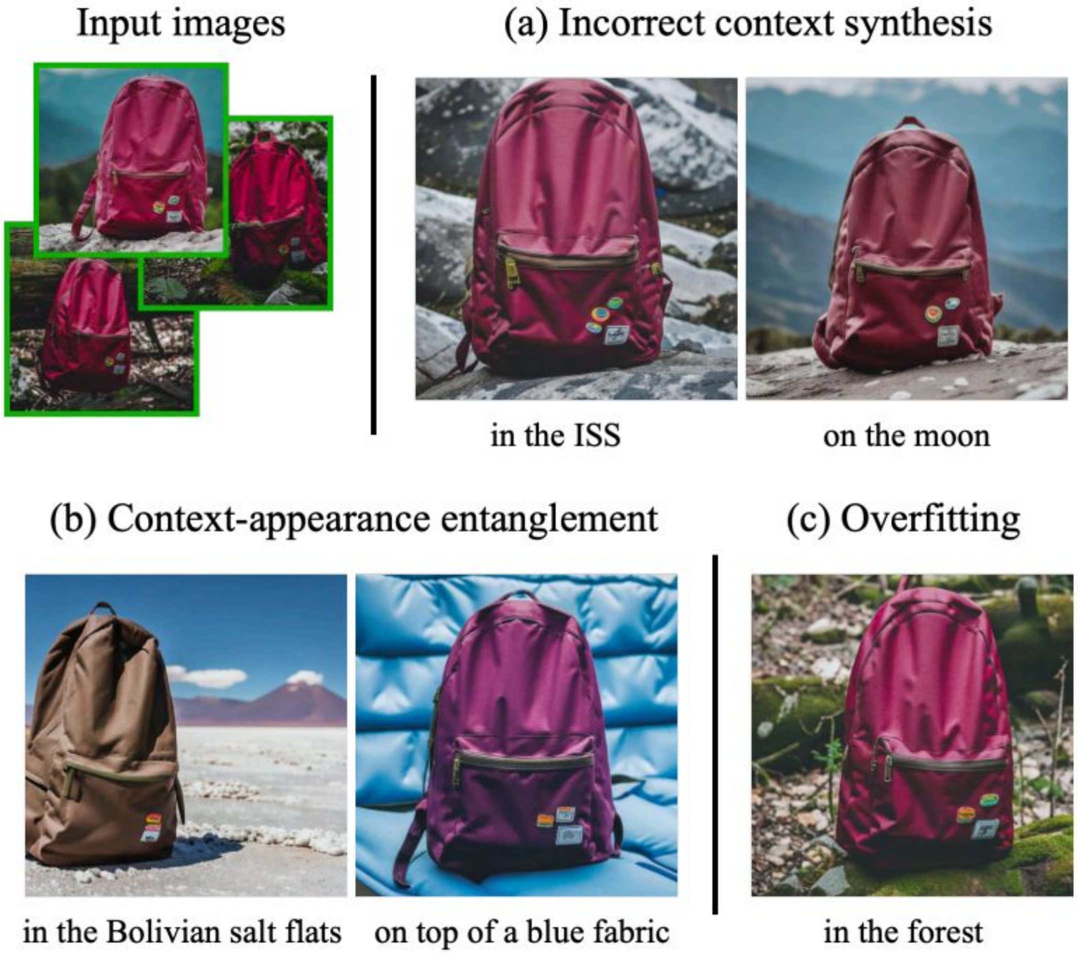
Method capabilities:

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

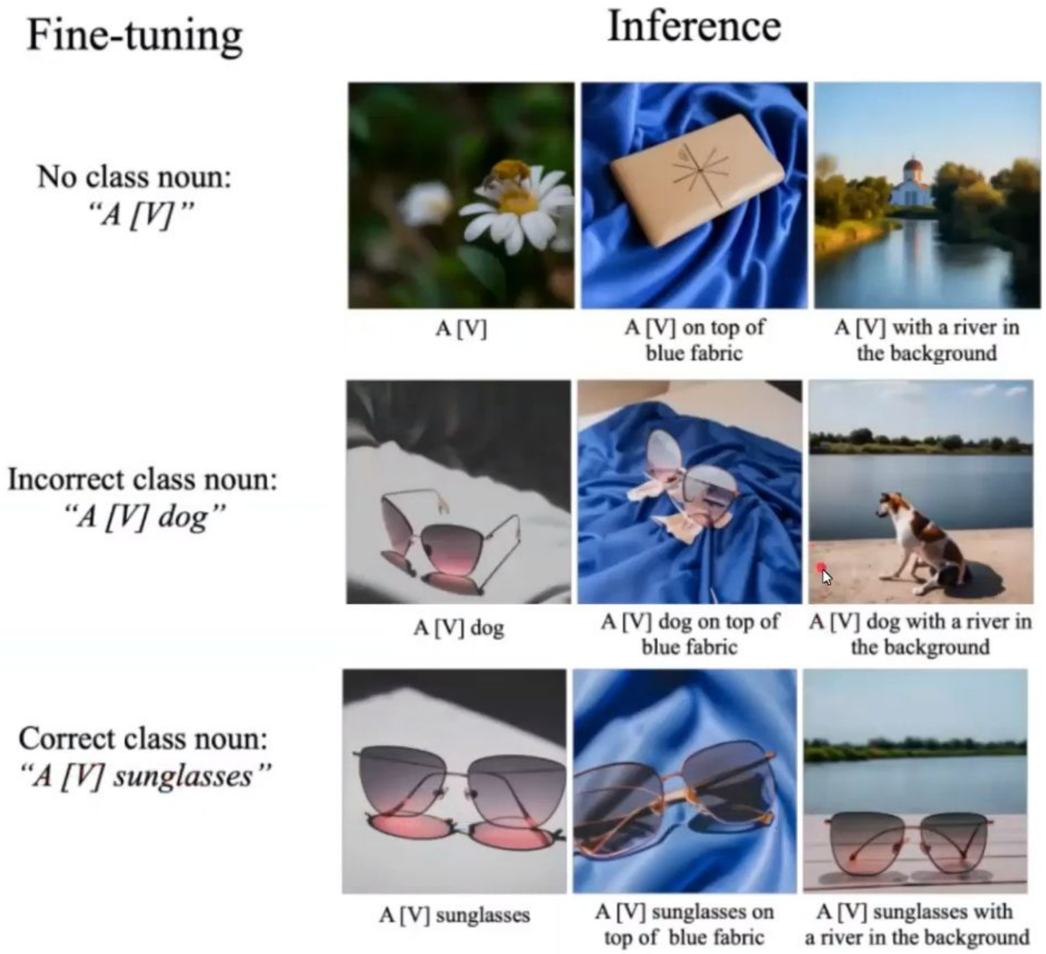


Comparison

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

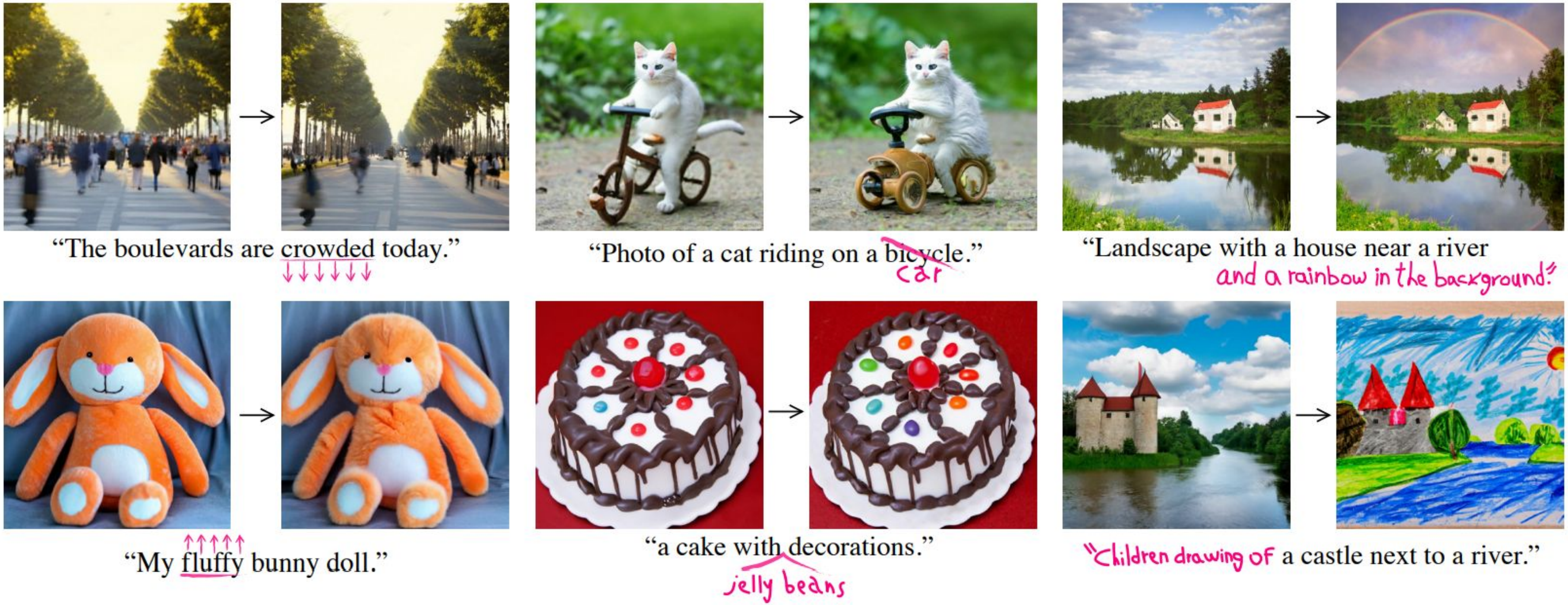
Comparison

Method	DINO ↑	CLIP-I ↑
Correct Class	**0.744**	**0.853**
No Class	0.303	0.607
Wrong Class	0.454	0.728

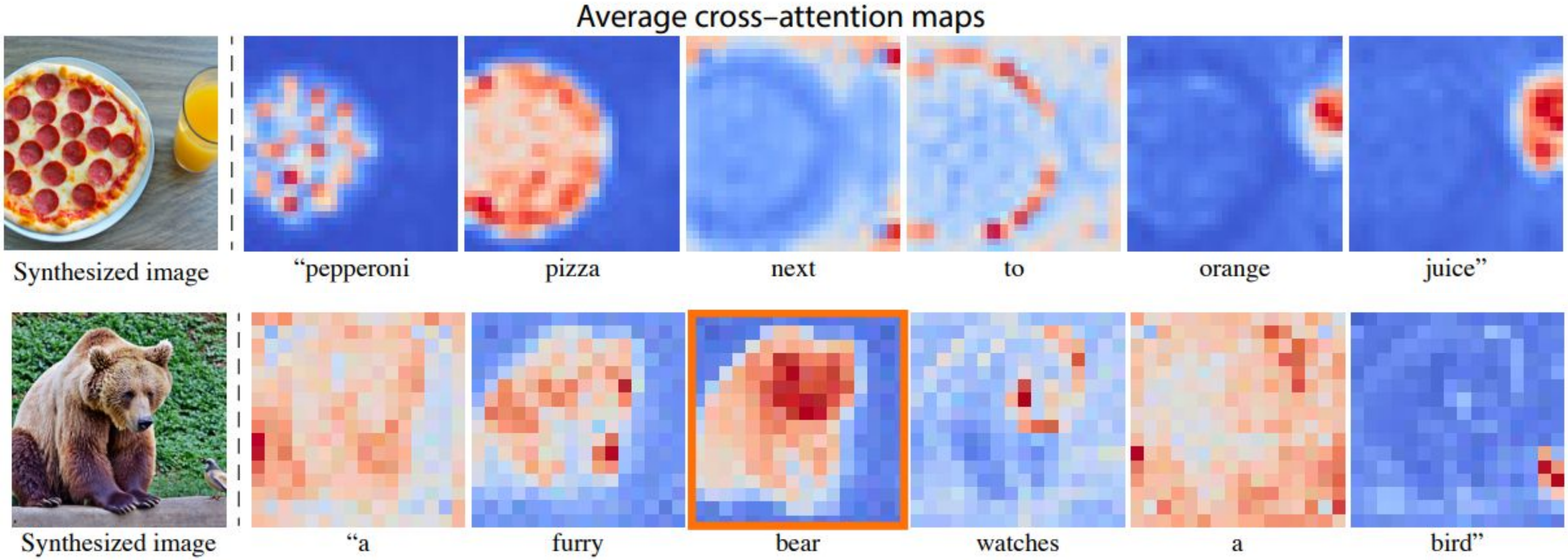


Prompt-to-prompt

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 importance of Cross-Attention Layer



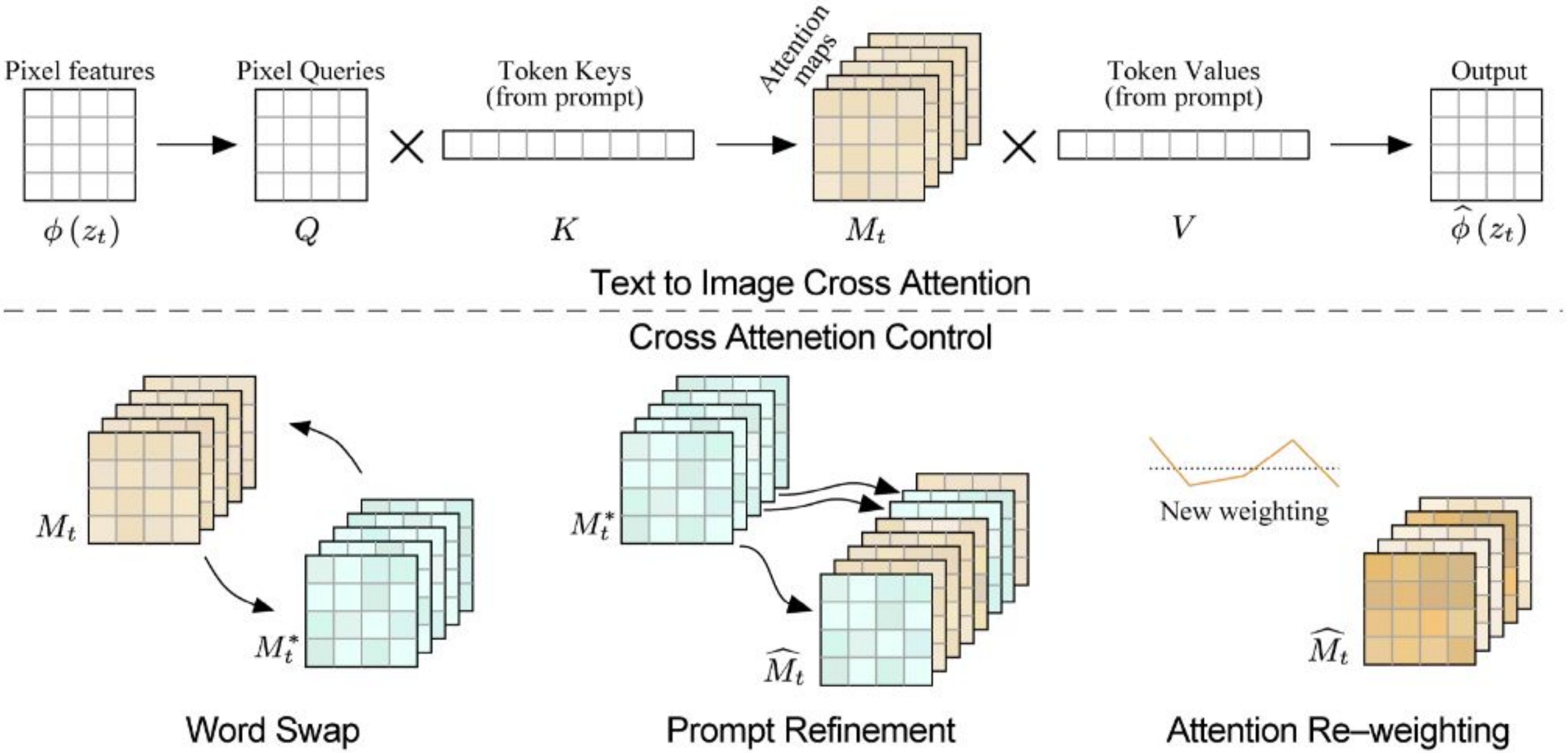
It specifies the pixels that need to be changed

Algorithm

Algorithm 1: Prompt-to-Prompt image editing

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

Cross-Attention Layer



Word swap:

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

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.”

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 ↑	1.84 ± 1.11	4.15 ± 1.09	3.38 ± 1.12	4.64 ± 0.64
(2) Text Alignment ↑	2.46 ± 1.16	2.89 ± 1.22	4.26 ± 1.03	4.55 ± 0.71
(3) Realism ↑	1.32 ± 0.70	2.36 ± 1.12	4.11 ± 0.93	4.42 ± 0.82