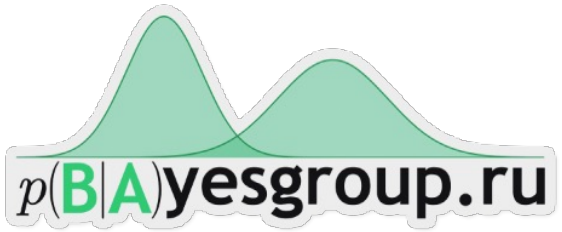
Image Manipulation by Diffusion Models

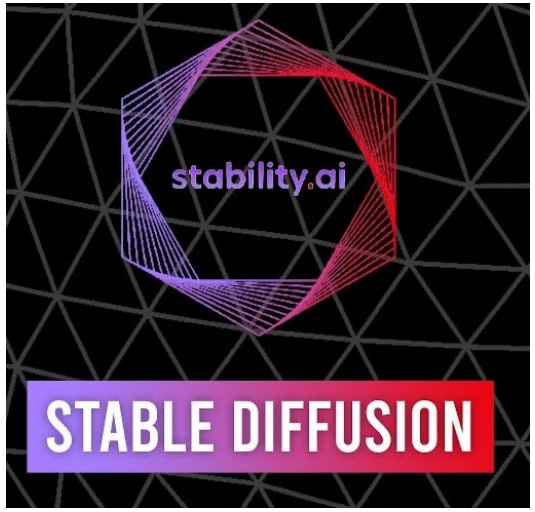
Aibek Alanov

Research scientist at Artificial Intelligence Research Institute (AIRI) and
Centre of Deep Learning and Bayesian Methods HSE University



Text-to-Image Diffusion Models

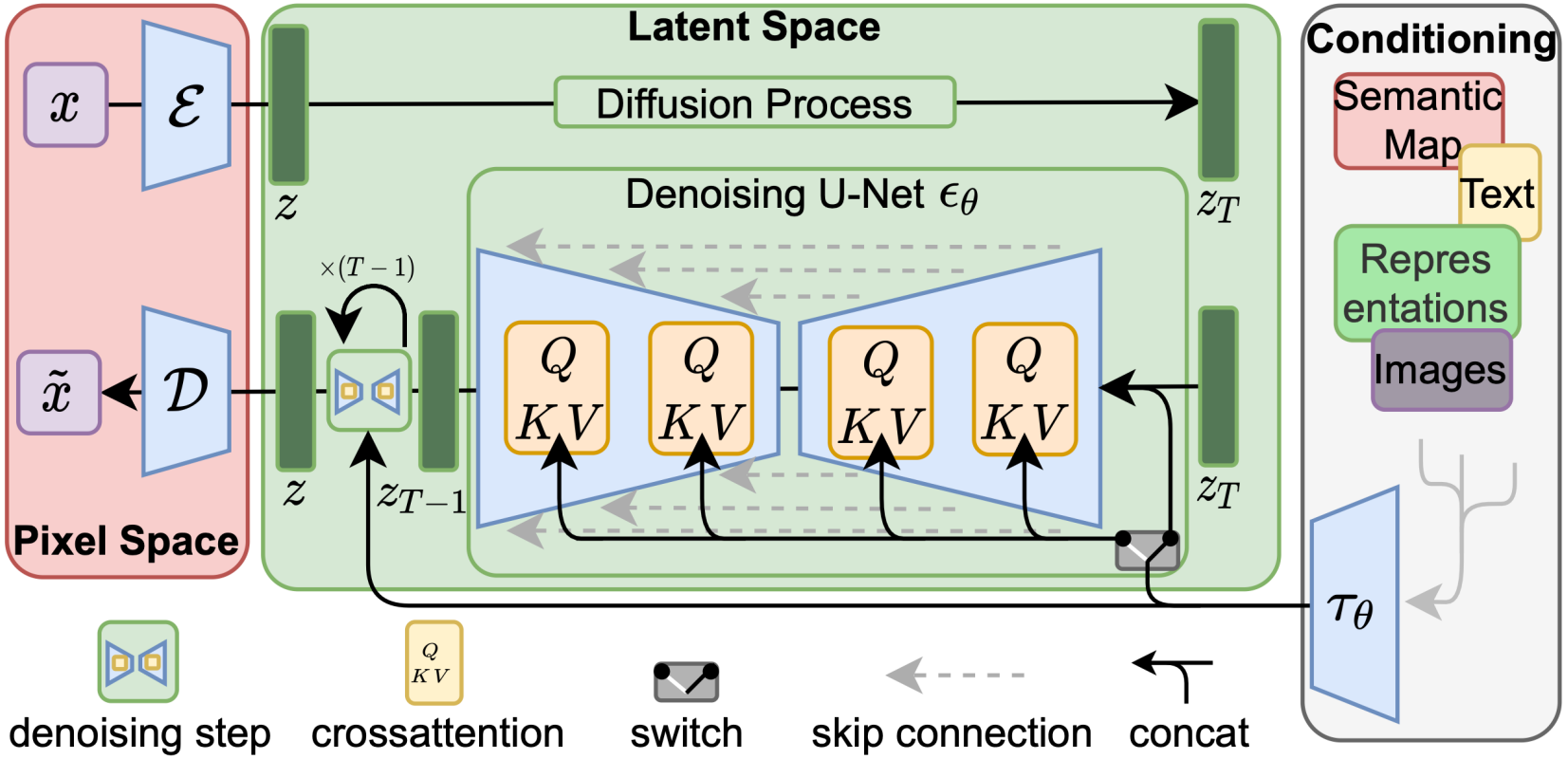
OpenAI
DALLE-2



ai-forever/
Kandinsky-2.0

Kandinsky 2.0 — multilingual text2image latent diffusion model

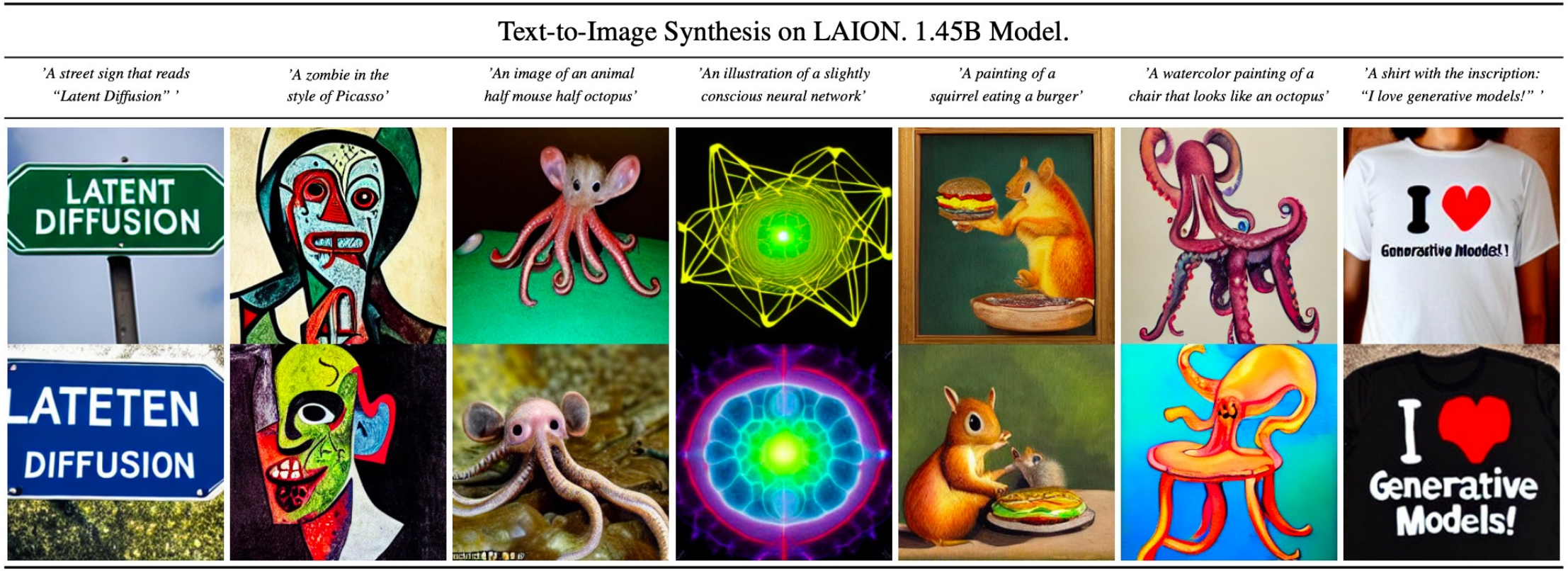
Stable Diffusion



$$L_{LDM} := \mathbb{E}_{\mathcal{E}(x),y,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(z_t, t, \tau_\theta(y))\|_2^2\right]$$

Rombach, Robin, et al. "High-resolution image synthesis with latent diffusion models." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

Stable Diffusion – Sample Examples



Rombach, Robin, et al. "High-resolution image synthesis with latent diffusion models." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

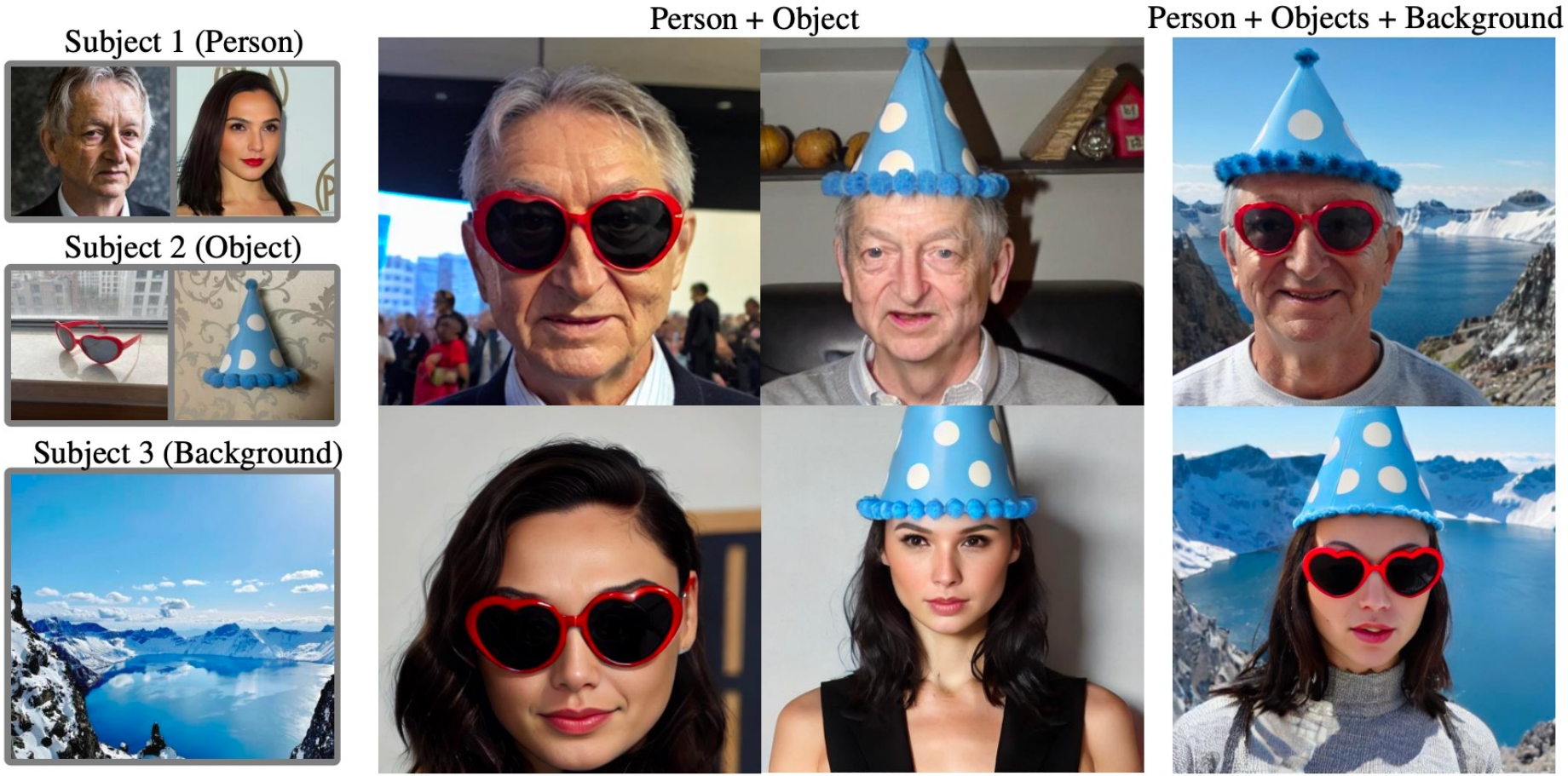
Image Manipulation Problems

- Generating unique concepts in novel scenes
- Editing the given real image

Concept Generation

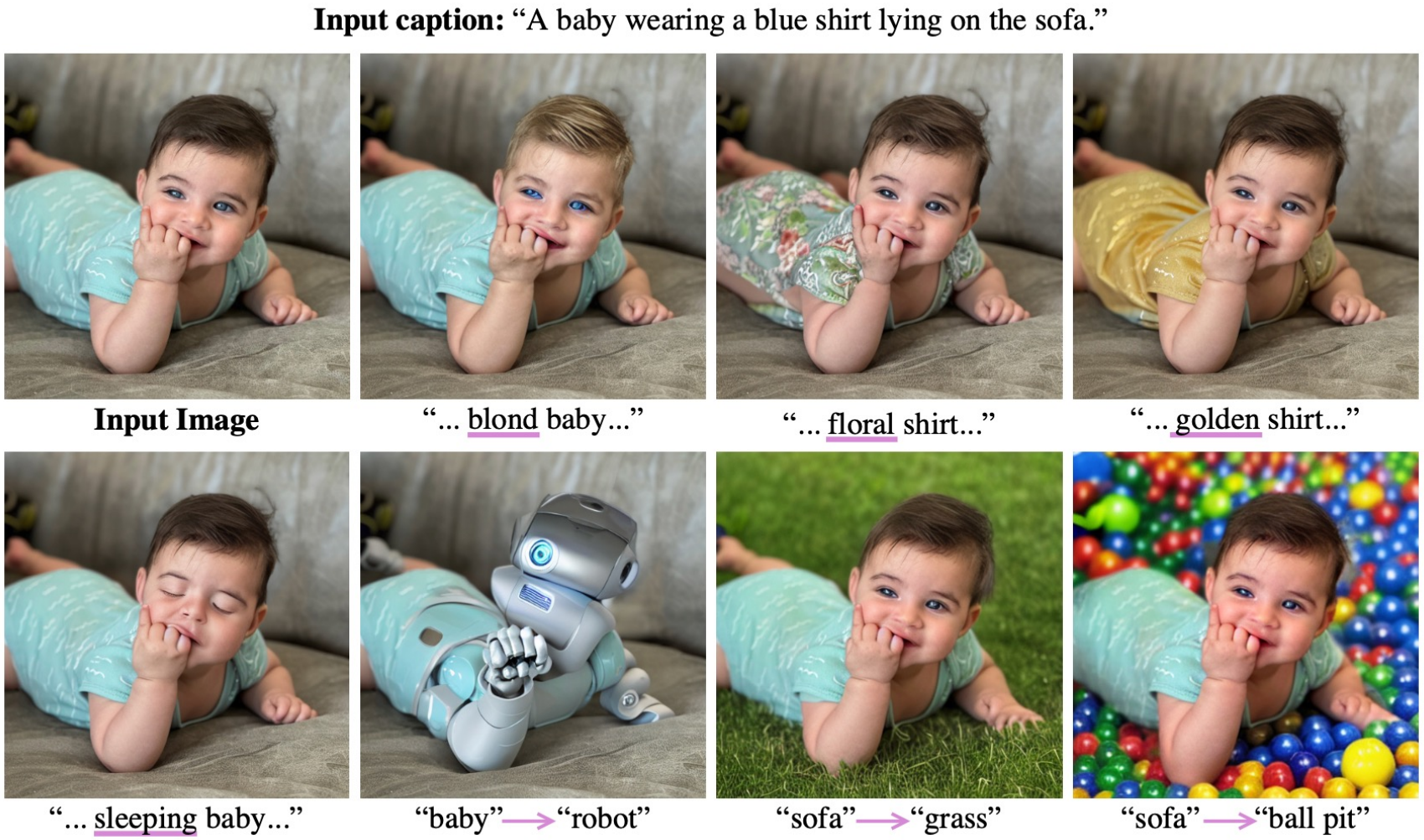
Ruiz, Nataniel, et al. "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation." arXiv preprint arXiv:2208.12242 (2022).

Concept Generation



Liu, Zhiheng, et al. "Cones: Concept Neurons in Diffusion Models for Customized Generation." arXiv preprint arXiv:2303.05125 (2023).

Editing Real Image

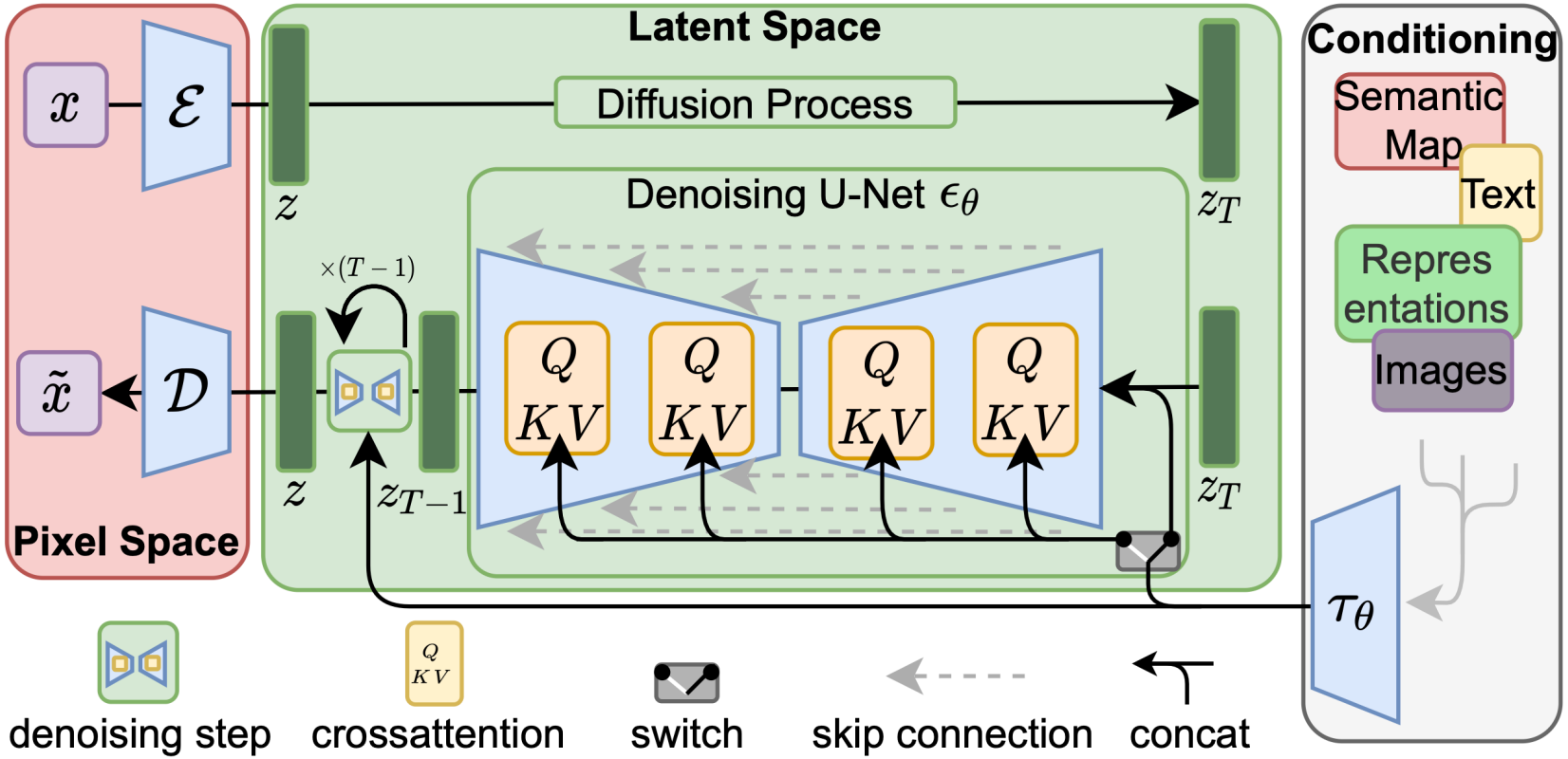


Mokady, Ron, et al. "Null-text Inversion for Editing Real Images using Guided Diffusion Models." arXiv preprint arXiv:2211.09794 (2022).

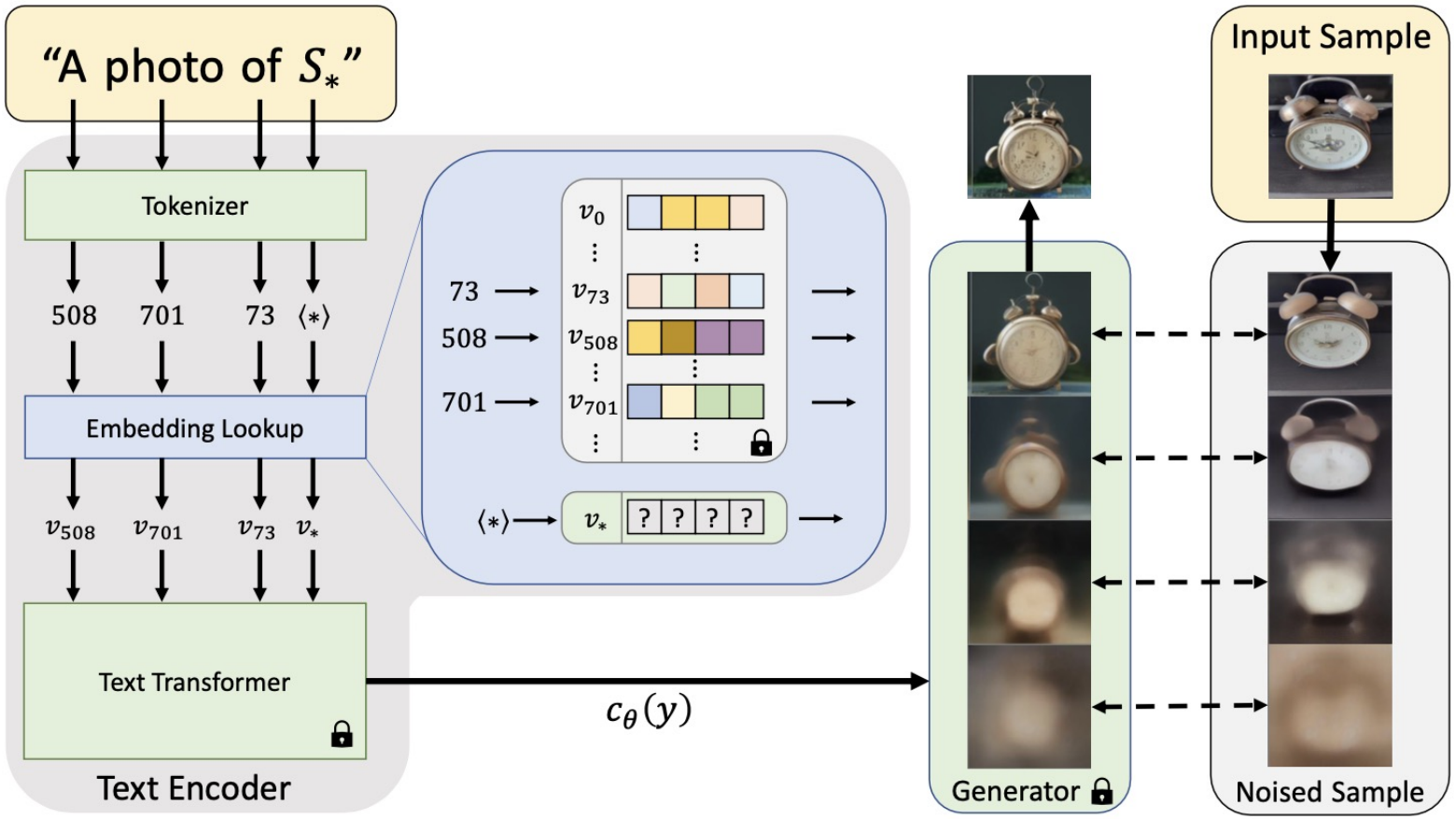
Textual Inversion

Rombach, Robin, et al. "High-resolution image synthesis with latent diffusion models." Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022.

Textual Inversion



Textual Inversion



Gal, Rinon, et al. "An image is worth one word: Personalizing text-to-image generation using textual inversion." arXiv preprint arXiv:2208.01618 (2022).

Textual Inversion

Our optimization goal can then be defined as:

$$v_* = \arg\min_v \mathbb{E}_{z\sim\mathcal{E}(x),y,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(z_t, t, c_\theta(y))\|_2^2\right], \quad (2)$$

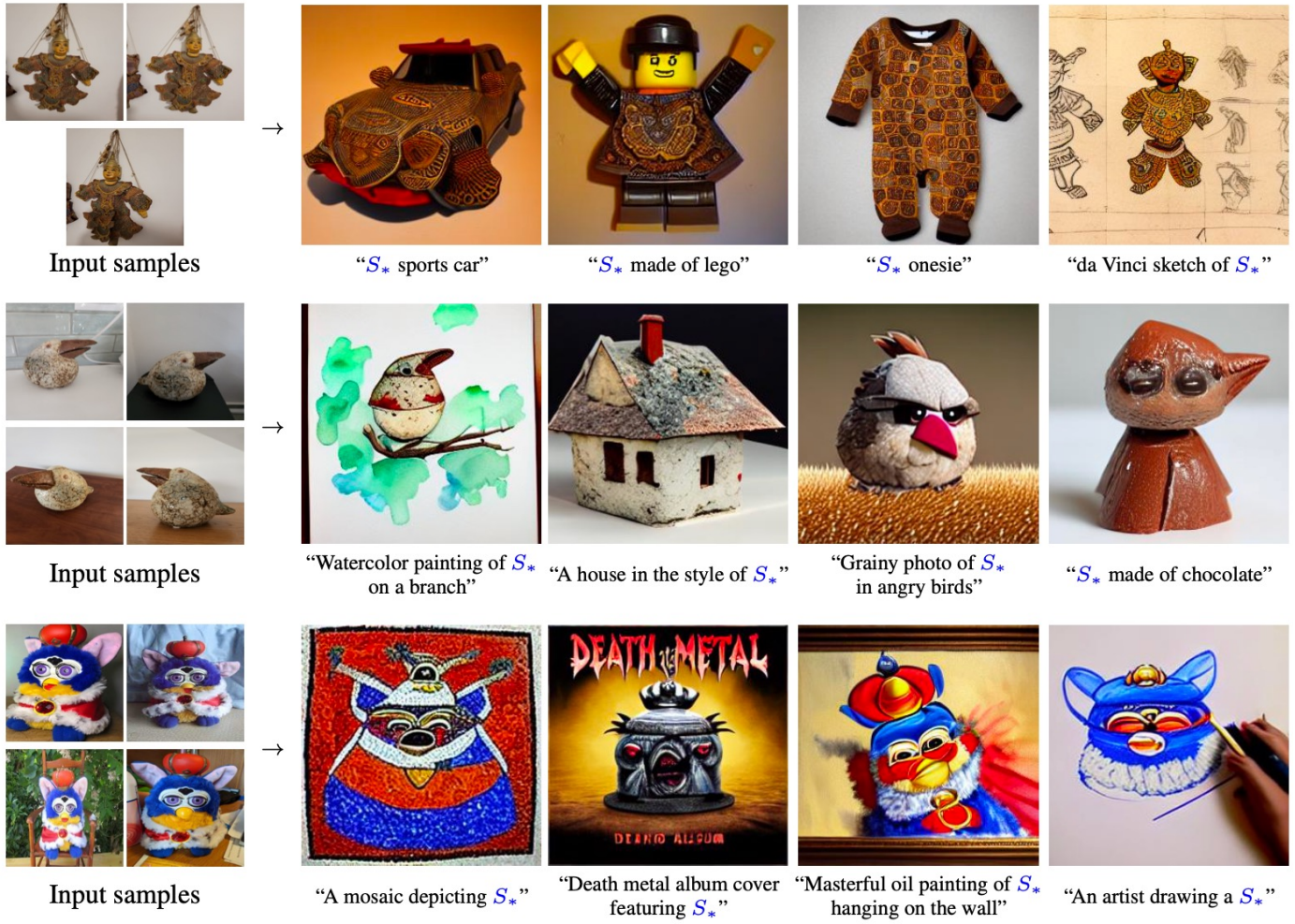
and is realized by re-using the same training scheme as the original LDM model, while keeping both c_θ and ϵ_θ fixed. Notably, this is a reconstruction task. As such, we expect it to motivate the learned embedding to capture fine visual details unique to the concept.

Gal, Rinon, et al. "An image is worth one word: Personalizing text-to-image generation using textual inversion." arXiv preprint arXiv:2208.01618 (2022).

Textual Inversion

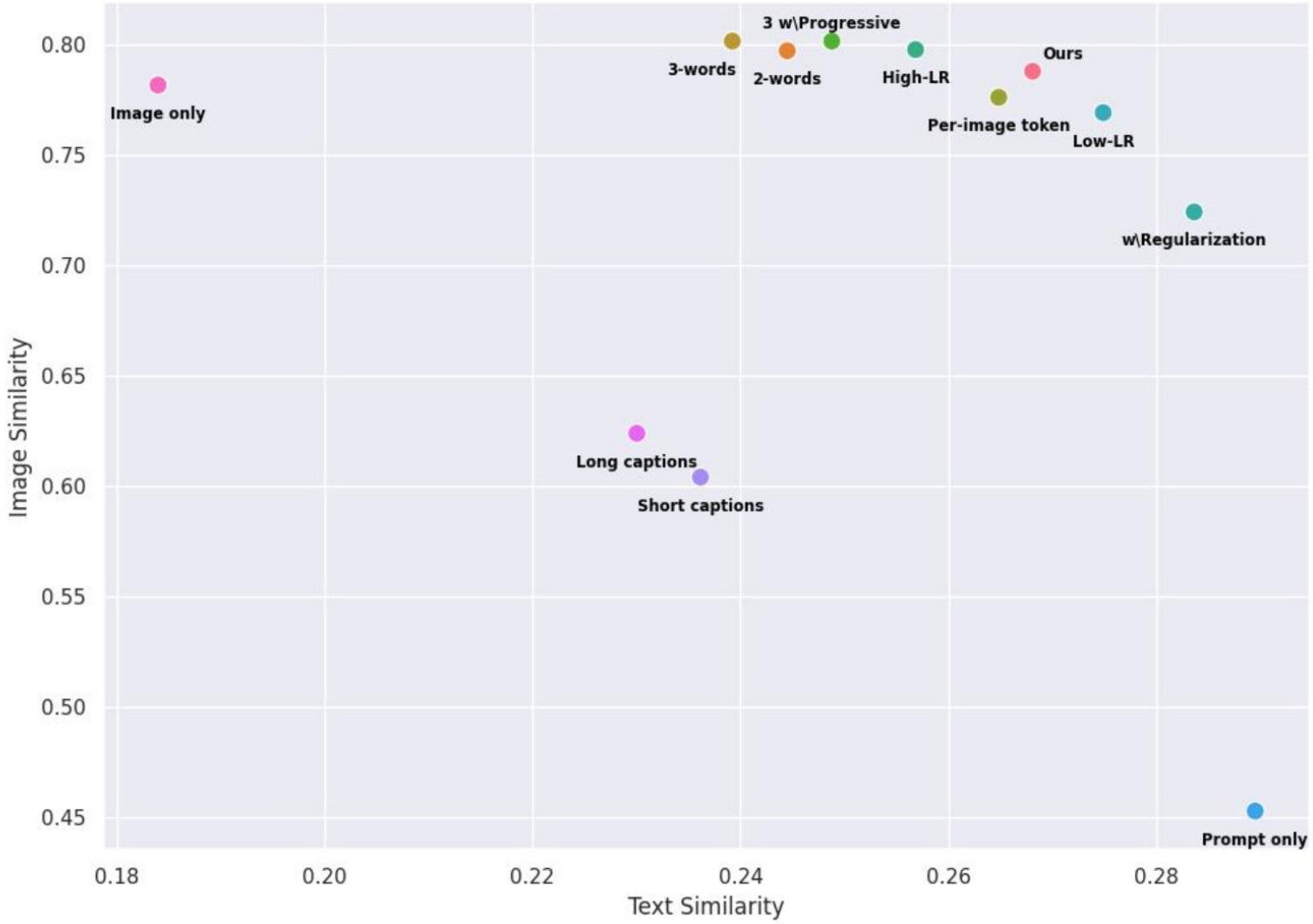
Gal, Rinon, et al. "An image is worth one word: Personalizing text-to-image generation using textual inversion." arXiv preprint arXiv:2208.01618 (2022).

Textual Inversion

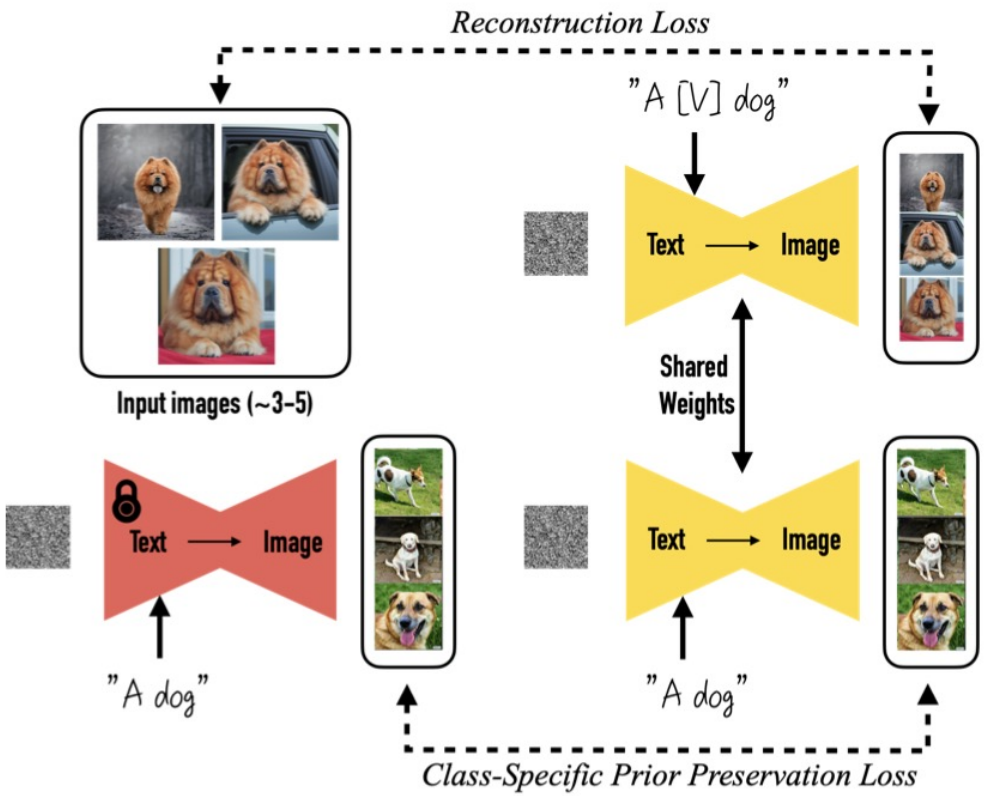


Gal, Rinon, et al. "An image is worth one word: Personalizing text-to-image generation using textual inversion." arXiv preprint arXiv:2208.01618 (2022).

Textual Inversion



DreamBooth



English words. Our approach is to find rare tokens in the vocabulary, and then invert these tokens into text space, in order to minimize the probability of the identifier having a strong prior. We perform a rare-token lookup in the vocab-

$\mathbf{c}_{\text{pr}} := \Gamma(f(\text{"a [class noun]"}))$. The loss becomes:

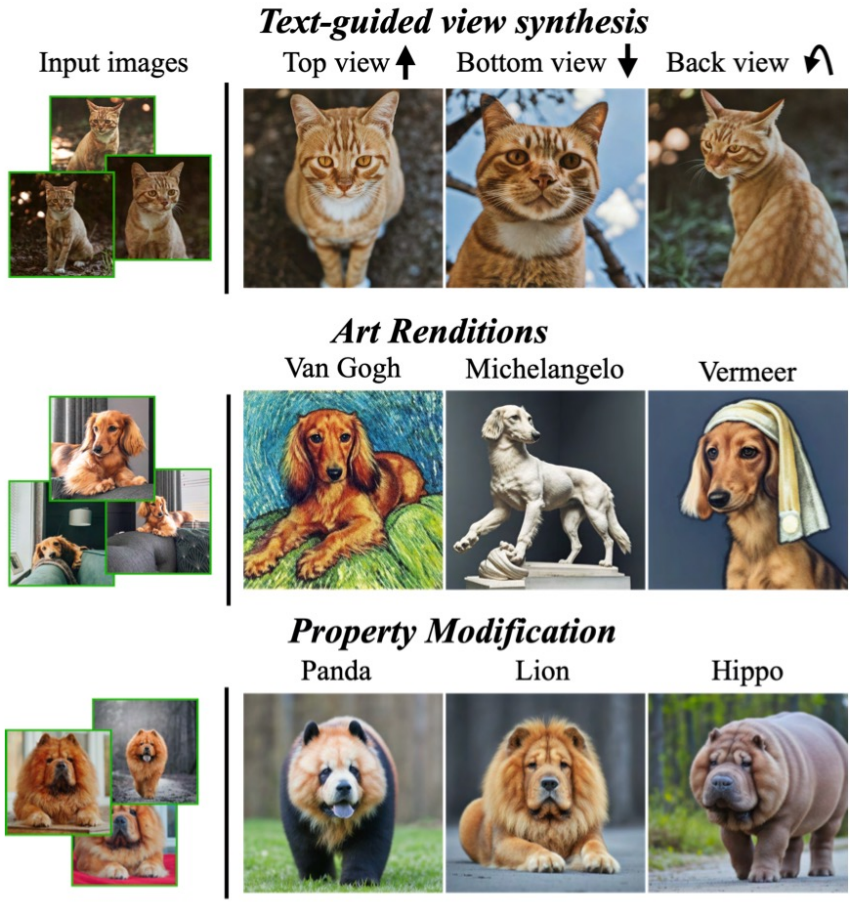
$$\mathbb{E}_{\mathbf{x},\mathbf{c},\boldsymbol{\epsilon},\boldsymbol{\epsilon}',t}[w_t \|\hat{\mathbf{x}}_\theta(\alpha_t \mathbf{x} + \sigma_t \boldsymbol{\epsilon}, \mathbf{c}) - \mathbf{x}\|_2^2 + \lambda w_{t'} \|\hat{\mathbf{x}}_\theta(\alpha_{t'} \mathbf{x}_{\text{pr}} + \sigma_{t'} \boldsymbol{\epsilon}',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

Ruiz, Nataniel, et al. "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation." arXiv preprint arXiv:2208.12242 (2022).

DreamBooth

Ruiz, Nataniel, et al. "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation." arXiv preprint arXiv:2208.12242 (2022).

DreamBooth



Ruiz, Nataniel, et al. "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation." arXiv preprint arXiv:2208.12242 (2022).

DreamBooth vs Textual Inversion

Figure 4. **Comparisons with Textual Inversion [20]** Given 4 input images (top row), we compare: DreamBooth Imagen (2nd row), DreamBooth Stable Diffusion (3rd row), Textual Inversion (bottom row). Output images were created with the following prompts (left to right): "a [V] vase in the snow", "a [V] vase on the beach", "a [V] vase in the jungle", "a [V] vase with the Eiffel Tower in the background". DreamBooth is stronger in both subject and prompt fidelity.

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

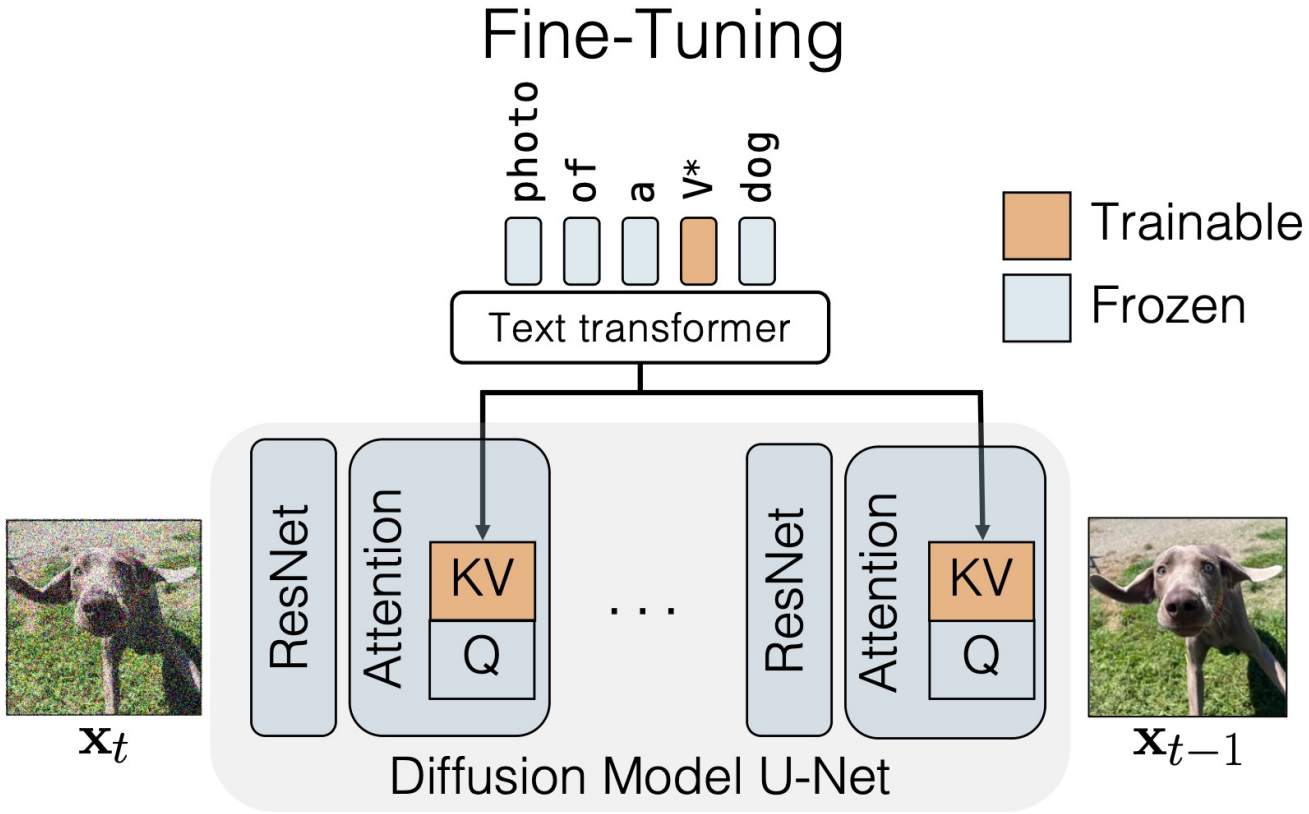
Table 1. Subject fidelity (DINO, CLIP-I) and prompt fidelity (CLIP-T, CLIP-T-L) quantitative metric comparison.

Method	Subject Fidelity ↑	Prompt Fidelity ↑
DreamBooth (Stable Diffusion)	**68%**	**81%**
Textual Inversion (Stable Diffusion)	22%	12%
Undecided	10%	7%

Table 2. Subject fidelity and prompt fidelity user preference.

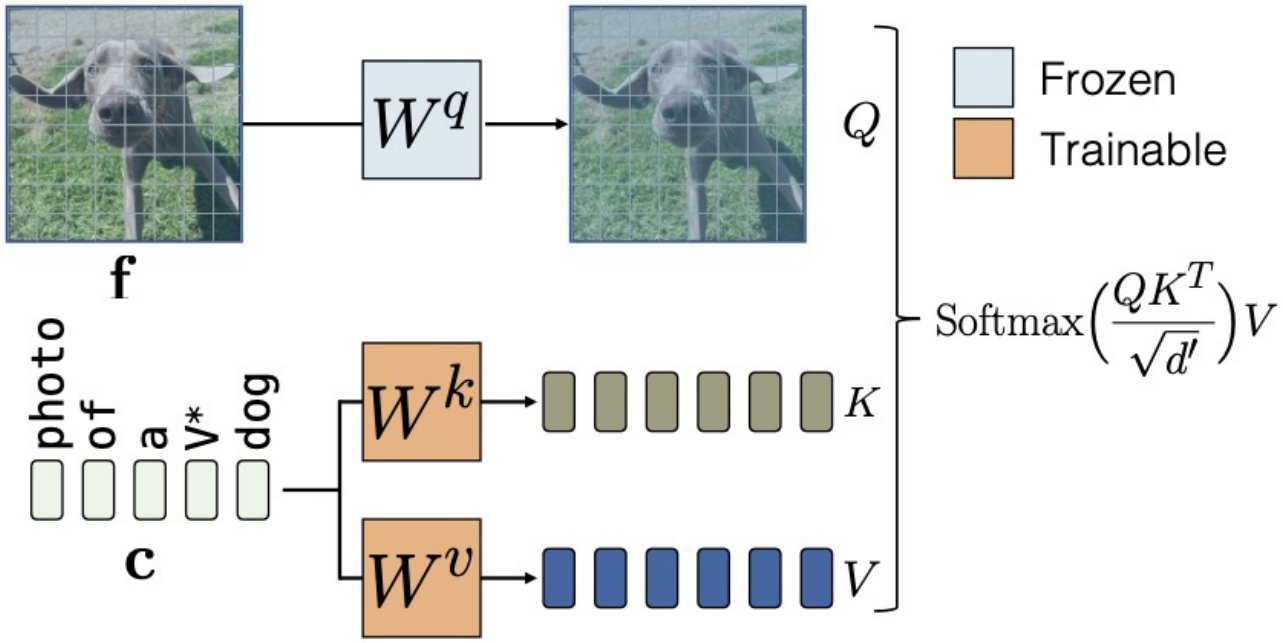
Ruiz, Nataniel, et al. "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation." arXiv preprint arXiv:2208.12242 (2022).

Custom Diffusion



Kumari, Nupur, et al. "Multi-Concept Customization of Text-to-Image Diffusion." arXiv preprint arXiv:2212.04488 (2022).

Custom Diffusion



Kumari, Nupur, et al. "Multi-Concept Customization of Text-to-Image Diffusion." arXiv preprint arXiv:2212.04488 (2022).

Custom Diffusion – Multiple Concept Training

- Joint training on multiple concepts
- Constrained optimization to merge concepts

$$\hat{W} = \arg\min_{W} ||W C_{\text{reg}}^{\top} - W_0 C_{\text{reg}}^{\top}||_F$$
$$\text{s.t. } W C^{\top} = V, \text{ where } C = [\mathbf{c}_1 \cdots \mathbf{c}_N]^{\top} \quad (4)$$
$$\text{and } V = [W_1 \mathbf{c}_1^{\top} \cdots W_N \mathbf{c}_N^{\top}]^{\top}.$$

Kumari, Nupur, et al. "Multi-Concept Customization of Text-to-Image Diffusion." arXiv preprint arXiv:2212.04488 (2022).

Custom Diffusion – Constrained Optimization

$$\hat{W} = W_0 + \mathbf{v}^\top \mathbf{d}, \text{ where } \mathbf{d} = C(C_{\text{reg}}^\top C_{\text{reg}})^{-1}$$
$$\text{and } \mathbf{v}^\top = (V - W_0 C^\top)(\mathbf{d}C^\top)^{-1}.$$

By using the method of Lagrange multipliers [8], we need to minimize the following objective:

$$L = \frac{1}{2}||WC_{\text{reg}}^\top - W_0 C_{\text{reg}}^\top|| - \text{trace}(\mathbf{v}(WC^\top - V)), \quad (7)$$

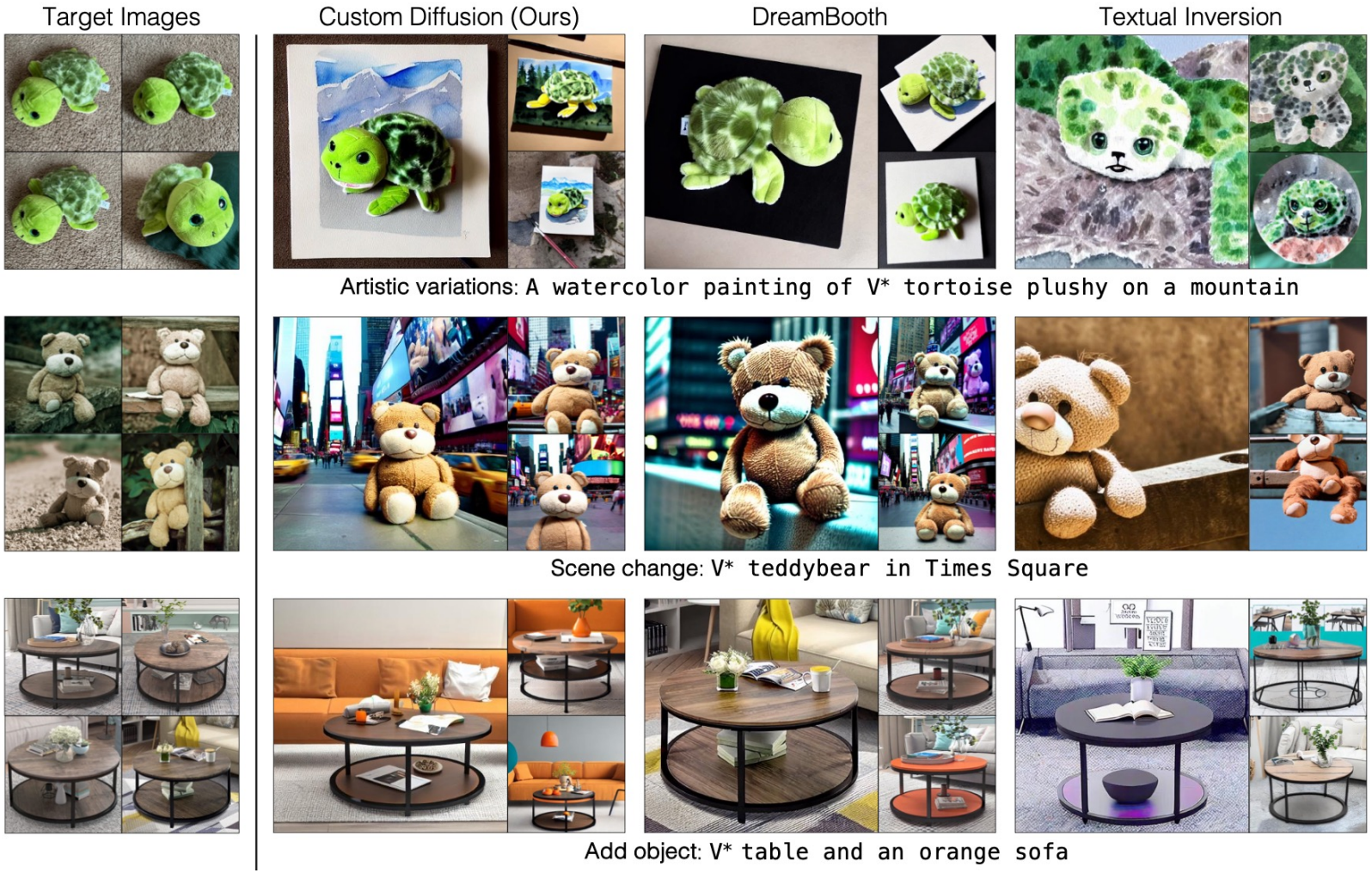
here $\mathbf{v} \in \mathbb{R}^{s \times o}$ is the Lagrangian multiplier corresponding to the constraints. Differentiating the above objective and equating it to 0, we obtain:

$$\begin{aligned} &WC_{\text{reg}}^\top C_{\text{reg}} - W_0 C_{\text{reg}}^\top C_{\text{reg}} - \mathbf{v}^\top C = 0 \\ &\implies W = W_0 + \mathbf{v}^\top C(C_{\text{reg}}^\top C_{\text{reg}})^{-1}. \end{aligned} \quad (8)$$

We assume C_{reg} is non-degenerate. Using the above solution in Eqn. 6, $WC^\top = V$, we obtain:

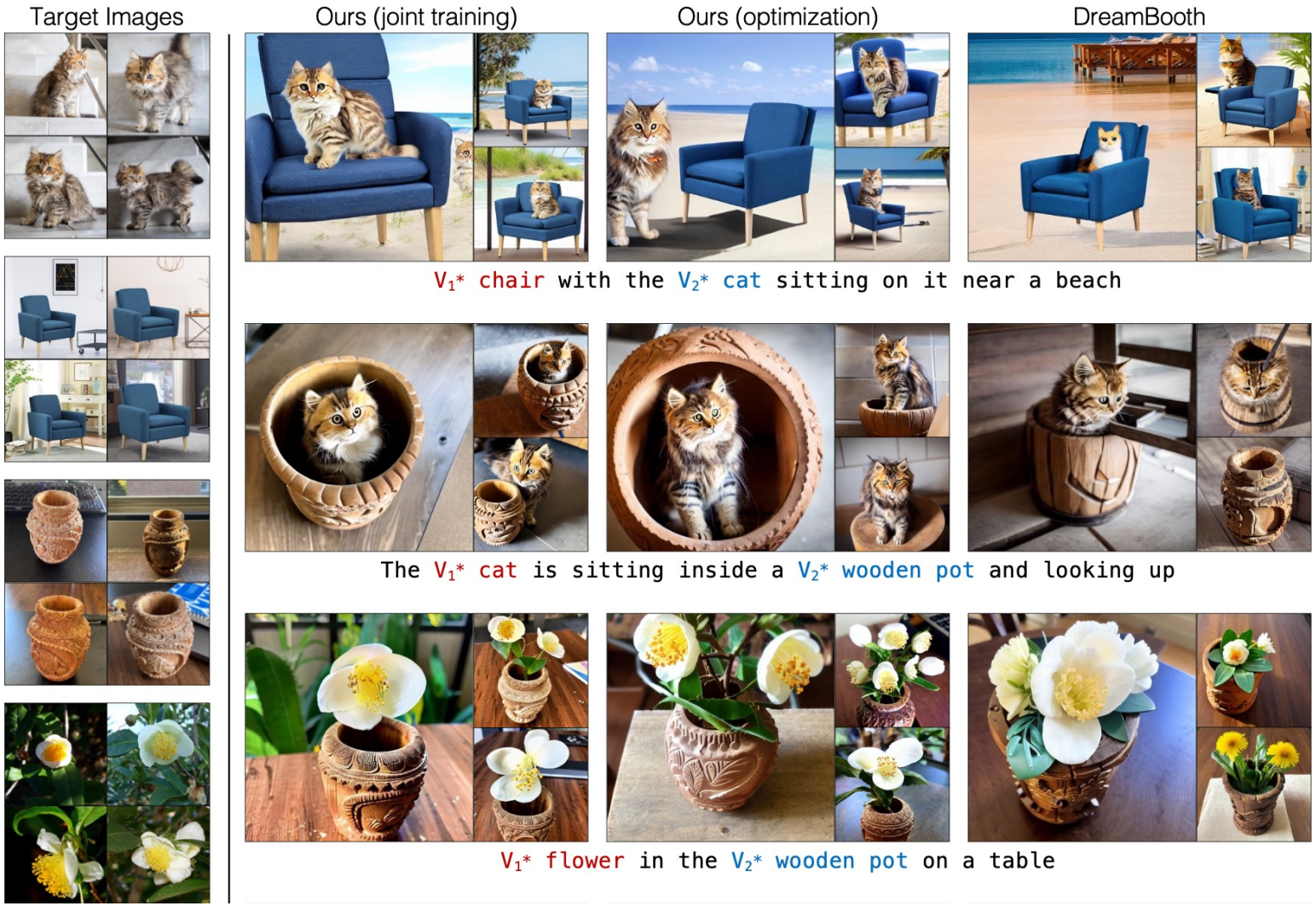
$$\begin{aligned} &(W_0 + \mathbf{v}^\top C(C_{\text{reg}}^\top C_{\text{reg}})^{-1})C^\top = V \\ &\text{Let } \mathbf{d} = C(C_{\text{reg}}^\top C_{\text{reg}})^{-1} \\ &\mathbf{v}^\top = (V - W_0 C^\top)(\mathbf{d}C^\top)^{-1}. \end{aligned} \quad (9)$$

Custom Diffusion – Results



Kumari, Nupur, et al. "Multi-Concept Customization of Text-to-Image Diffusion." arXiv preprint arXiv:2212.04488 (2022).

Custom Diffusion – Results



Custom Diffusion – Results

	Method	Text-alignment	Image-alignment	KID (validation)
Single-Concept	Textual Inversion	0.670	**0.827**	22.27
	DreamBooth	0.781	0.776	32.53
	Ours (w/ fine-tune all)	**0.795**	0.748	**19.27**
	Ours	**0.795**	0.775	20.96
Multi-Concept	Textual Inversion	0.544	0.630	
	DreamBooth	0.783	0.695	
	Ours (w/ fine-tune all)	0.787	0.691	—
	Ours (Sequential)	0.797	0.700	
	Ours (Optimization)	0.800	0.695	
	Ours (Joint)	**0.801**	**0.706**	

Kumari, Nupur, et al. "Multi-Concept Customization of Text-to-Image Diffusion." arXiv preprint arXiv:2212.04488 (2022).

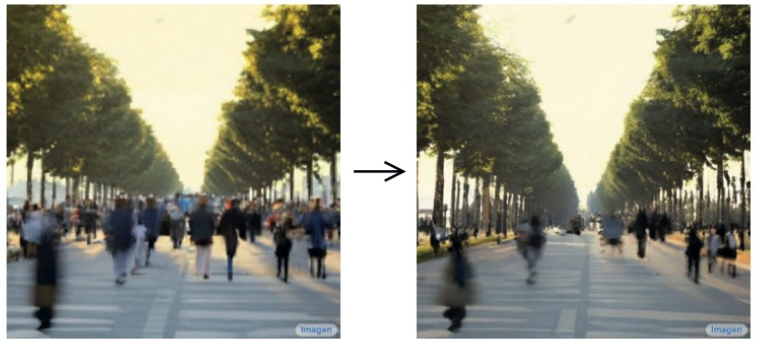
Other Works on Concept Generation

“ELITE: Encoding Visual Concepts into Textual Embeddings for Customized Text-to-Image Generation” (2023)

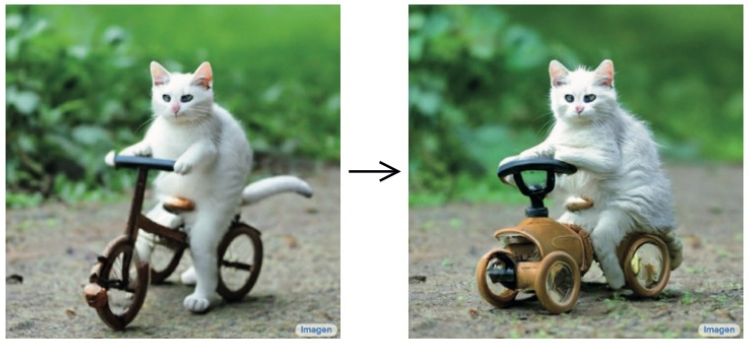
“Cones: Concept Neurons in Diffusion Models for Customized Generation” (2023)

“SVDiff: Compact Parameter Space for Diffusion Fine-Tuning” (2023)

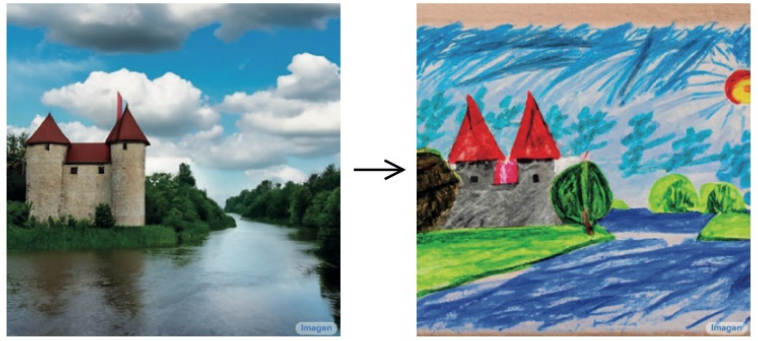
Prompt-to-Prompt Image Editing



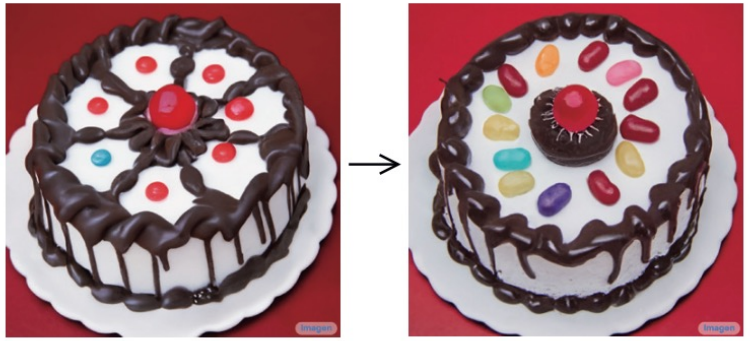
"The boulevards are crowded today."



"Photo of a cat riding on a bicycle."



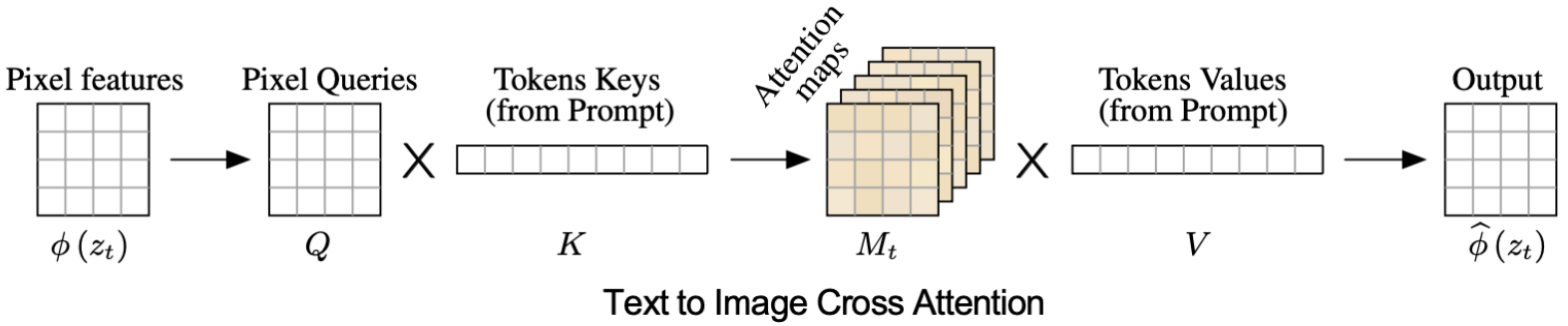
"Children drawing of a castle next to a river."



"a cake with decorations."

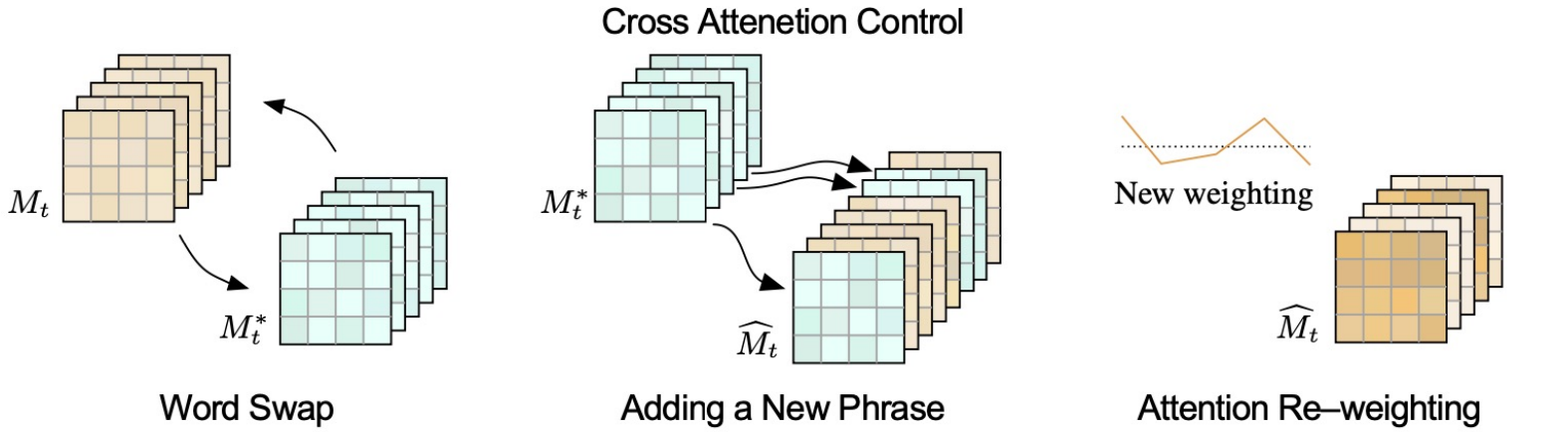
Hertz, Amir, et al. "Prompt-to-prompt image editing with cross attention control." arXiv preprint arXiv:2208.01626 (2022).

Prompt-to-Prompt Image Editing



Hertz, Amir, et al. "Prompt-to-prompt image editing with cross attention control." arXiv preprint arXiv:2208.01626 (2022).

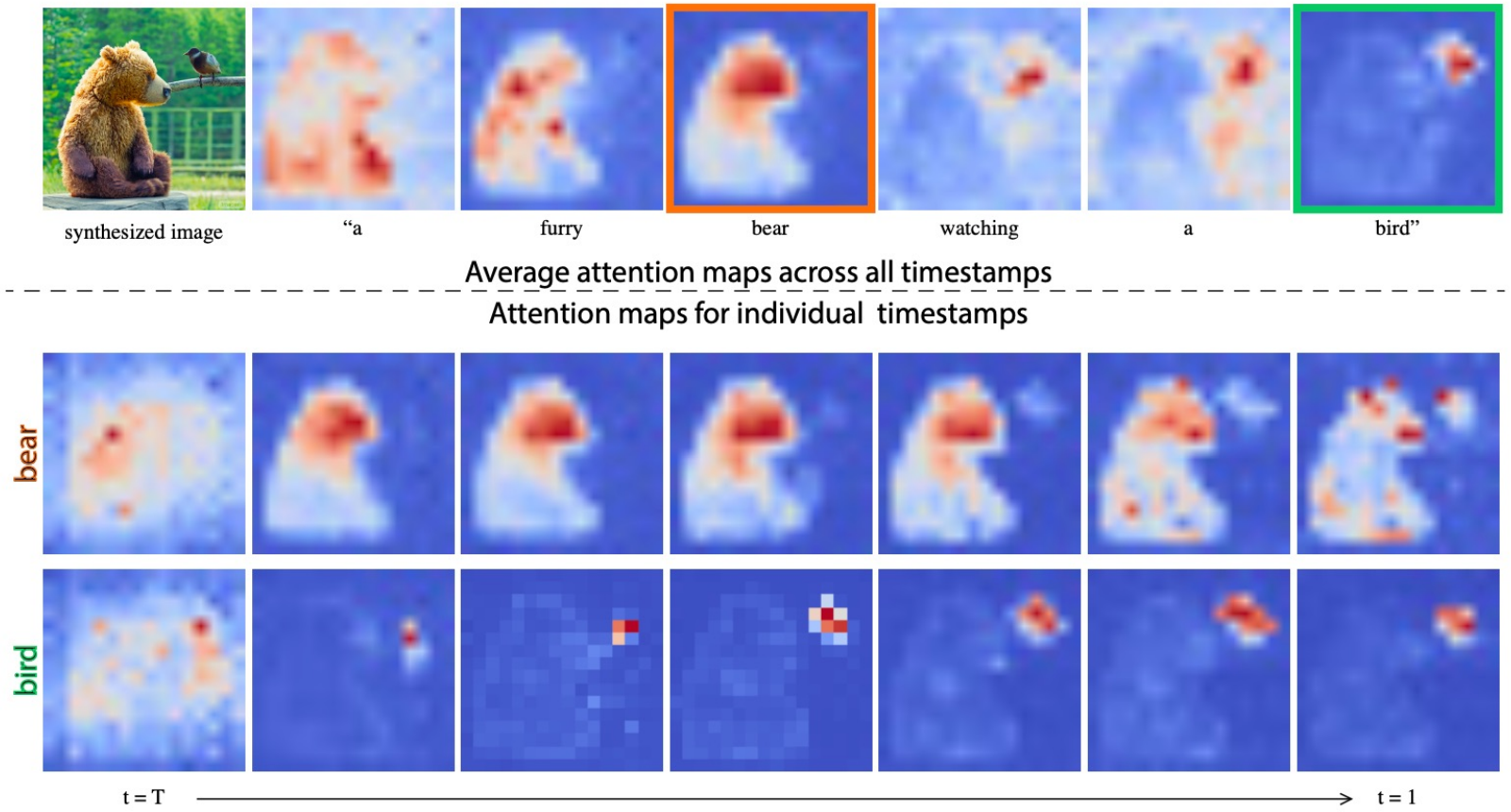
Prompt-to-Prompt Image Editing



Fixed attention maps and random seed

Hertz, Amir, et al. "Prompt-to-prompt image editing with cross attention control." arXiv preprint arXiv:2208.01626 (2022).

Prompt-to-Prompt Image Editing



Hertz, Amir, et al. "Prompt-to-prompt image editing with cross attention control." arXiv preprint arXiv:2208.01626 (2022).

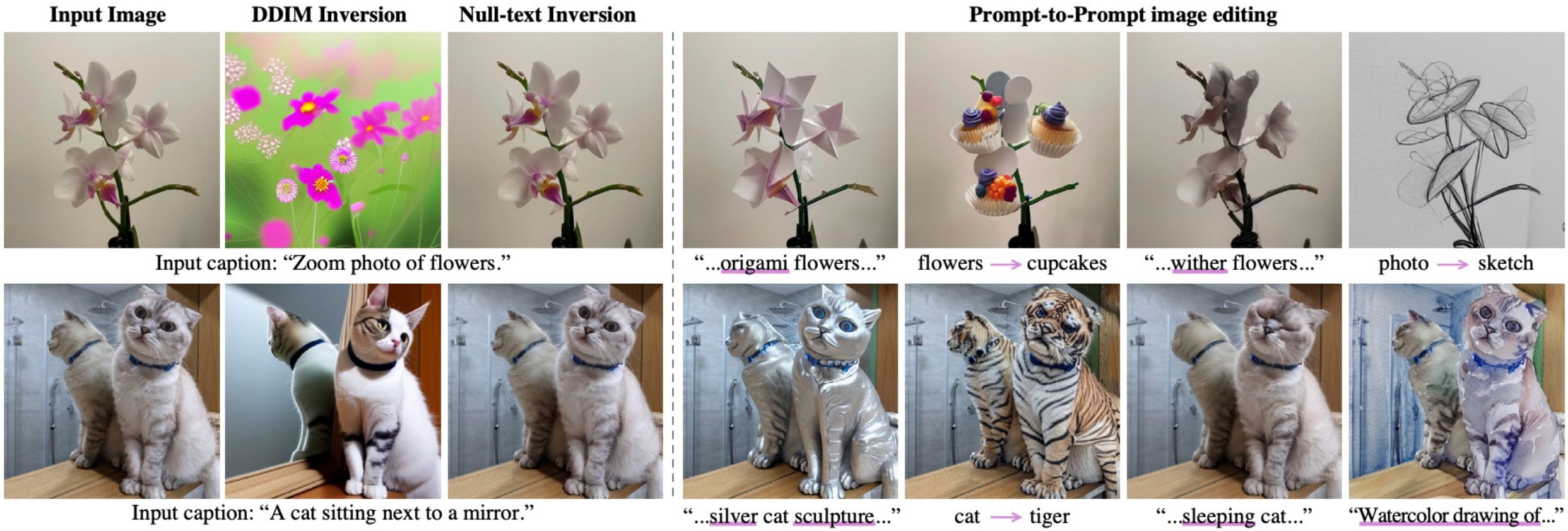
Prompt-to-Prompt Image Editing

Algorithm 1: Prompt-to-Prompt image editing

Input: A source prompt $\mathcal{P}$, a target prompt $\mathcal{P}^*$, and a random seed s.
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**
    $z_{t-1}, M_t \leftarrow DM(z_t, \mathcal{P}, t, s)$;
    $M_t^* \leftarrow DM(z_t^*, \mathcal{P}^*, t, s)$;
    $\widehat{M}_t \leftarrow Edit(M_t, M_t^*, t)$;
    $z_{t-1}^* \leftarrow DM(z_t^*, \mathcal{P}^*, t, s_t)\{M \leftarrow \widehat{M}_t\}$;
end
Return (z_0, z_0^*)

Hertz, Amir, et al. "Prompt-to-prompt image editing with cross attention control." arXiv preprint arXiv:2208.01626 (2022).

Null-Text Inversion for Editing Real Images



Mokady, Ron, et al. "Null-text Inversion for Editing Real Images using Guided Diffusion Models." arXiv preprint arXiv:2211.09794 (2022).

Null-Text Inversion for Editing Real Images

$$\min_{\theta} E_{z_0,\varepsilon\sim N(0,I),t\sim\text{Uniform}(1,T)} \|\varepsilon - \varepsilon_\theta(z_t, t, \mathcal{C})\|_2^2. \quad (1)$$

$$z_{t-1} = \sqrt{\frac{\alpha_{t-1}}{\alpha_t}} z_t + \left(\sqrt{\frac{1}{\alpha_{t-1}} - 1} - \sqrt{\frac{1}{\alpha_t} - 1}\right) \cdot \varepsilon_\theta(z_t, t, \mathcal{C}).$$

$$\tilde{\varepsilon}_\theta(z_t, t, \mathcal{C}, \varnothing) = w \cdot \varepsilon_\theta(z_t, t, \mathcal{C}) + (1 - w) \cdot \varepsilon_\theta(z_t, t, \varnothing).$$

E.g., $w = 7.5$ is the default parameter for Stable Diffusion.

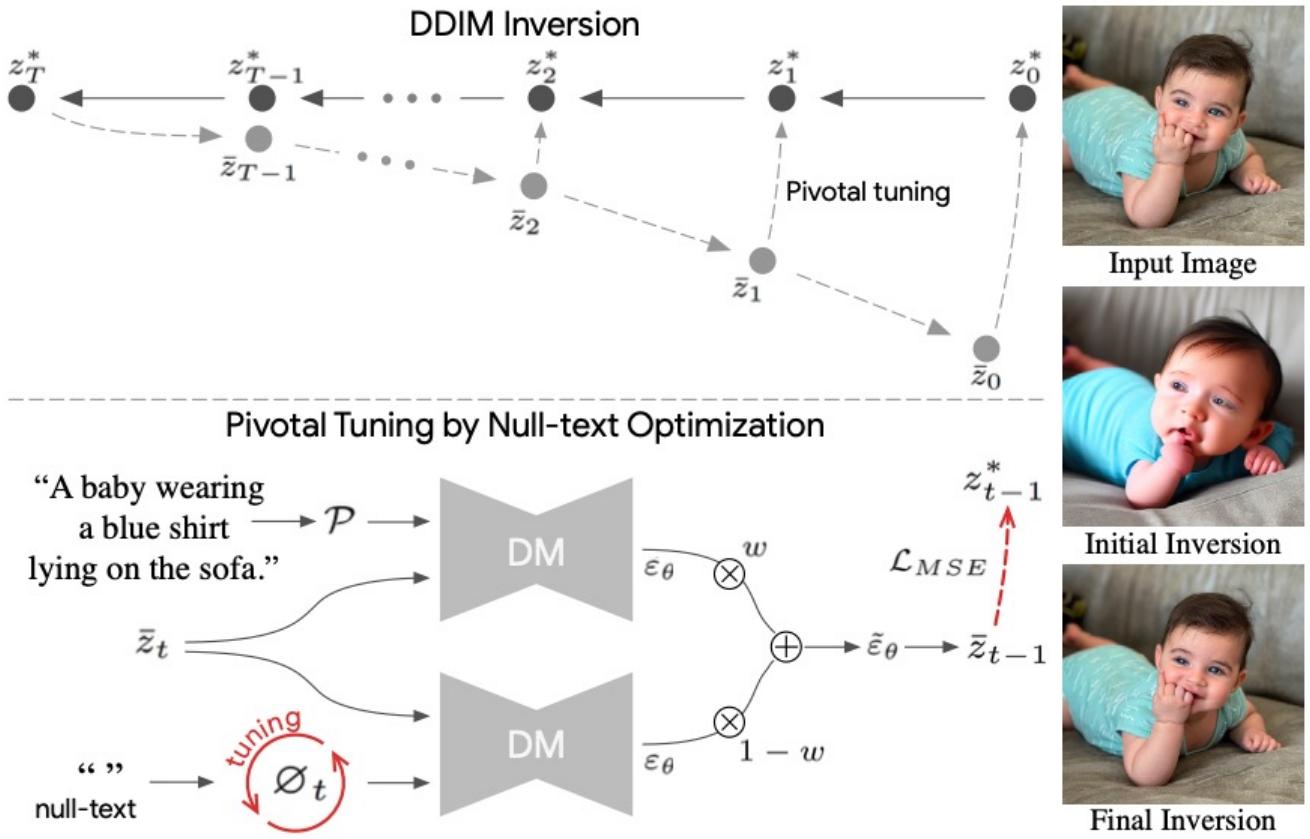
DDIM inversion. A simple inversion technique was suggested for the DDIM sampling [13, 35], based on the assumption that the ODE process can be reversed in the limit of small steps:

$$z_{t+1} = \sqrt{\frac{\alpha_{t+1}}{\alpha_t}} z_t + \left(\sqrt{\frac{1}{\alpha_{t+1}} - 1} - \sqrt{\frac{1}{\alpha_t} - 1}\right) \cdot \varepsilon_\theta(z_t, t, \mathcal{C}).$$

In other words, the diffusion process is performed in the reverse direction, that is $z_0 \rightarrow z_T$ instead of $z_T \rightarrow z_0$, where z_0 is set to be the encoding of the given real image.

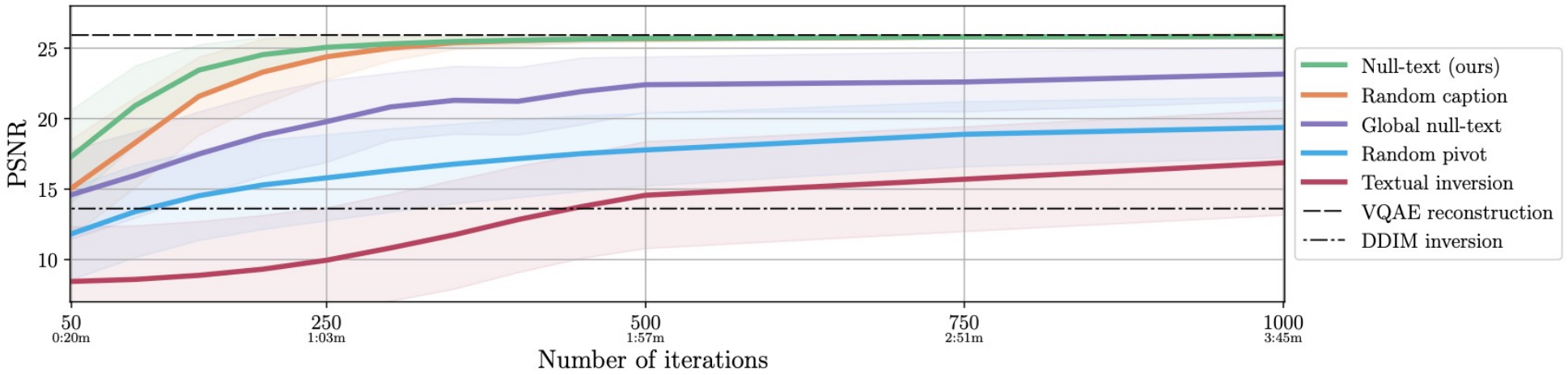
Mokady, Ron, et al. "Null-text Inversion for Editing Real Images using Guided Diffusion Models." arXiv preprint arXiv:2211.09794 (2022).

Null-Text Inversion for Editing Real Images



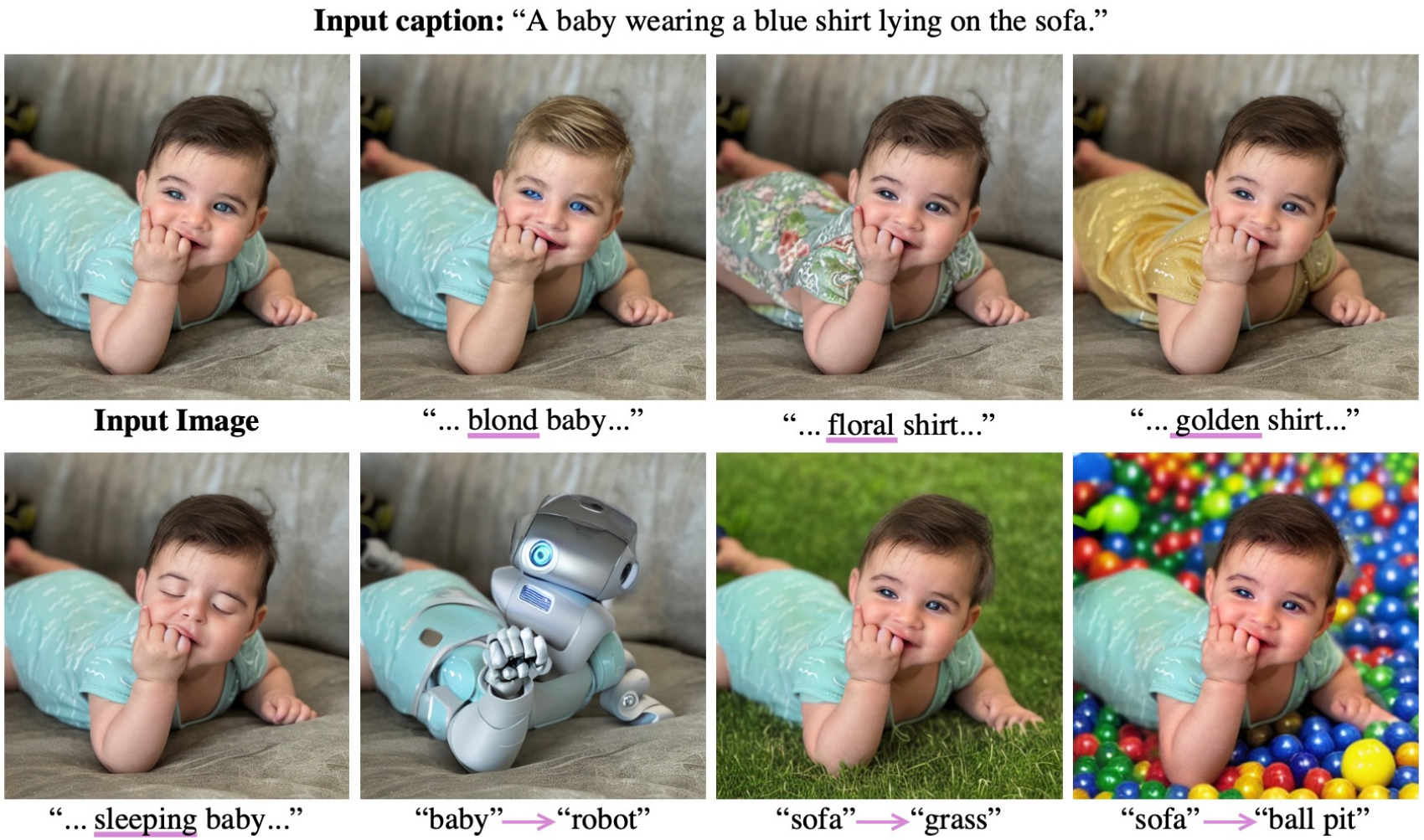
Mokady, Ron, et al. "Null-text Inversion for Editing Real Images using Guided Diffusion Models." arXiv preprint arXiv:2211.09794 (2022).

Null-Text Inversion for Editing Real Images



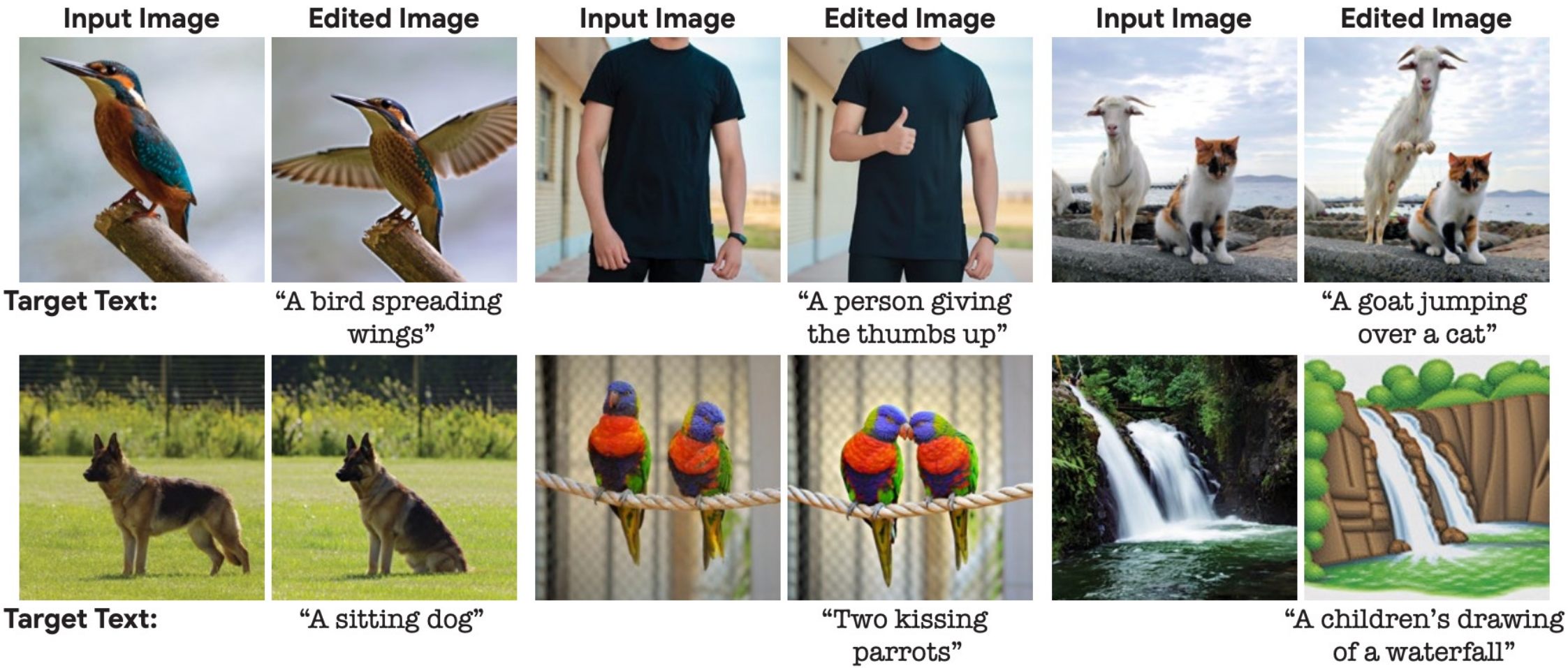
Mokady, Ron, et al. "Null-text Inversion for Editing Real Images using Guided Diffusion Models." arXiv preprint arXiv:2211.09794 (2022).

Null-Text Inversion for Editing Real Images



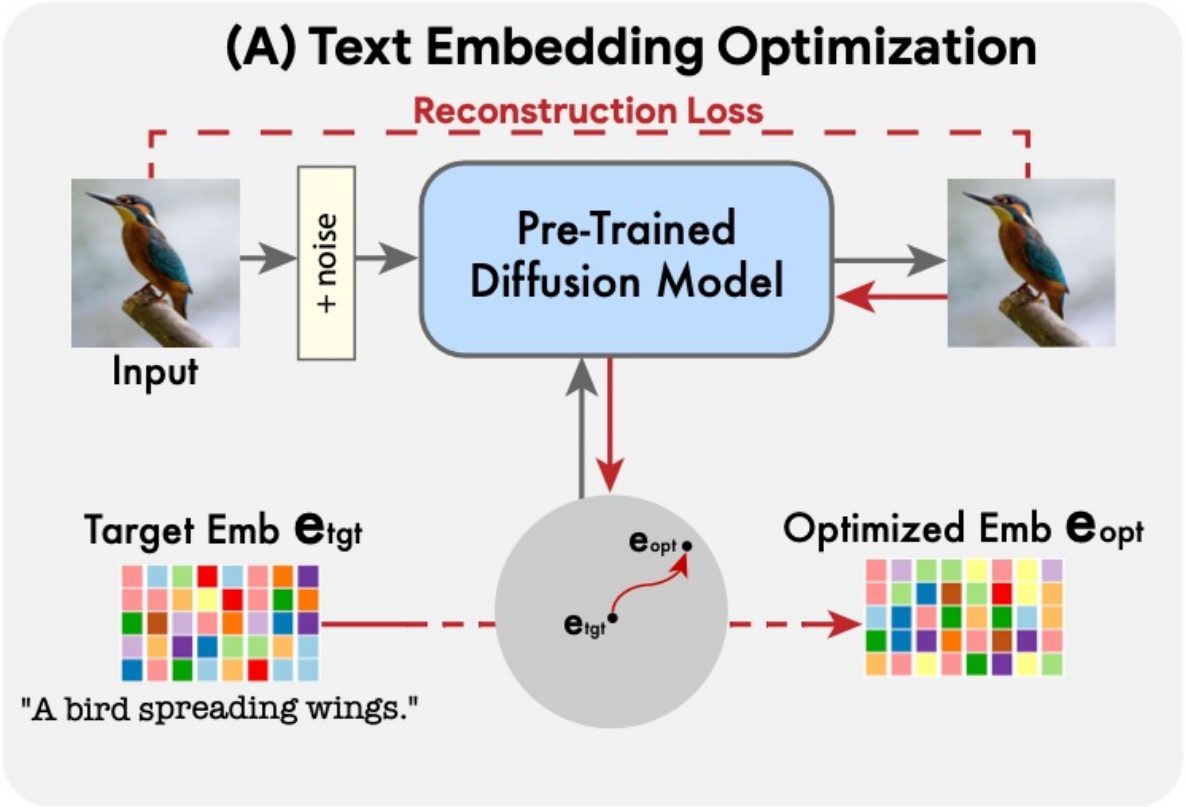
Mokady, Ron, et al. "Null-text Inversion for Editing Real Images using Guided Diffusion Models." arXiv preprint arXiv:2211.09794 (2022).

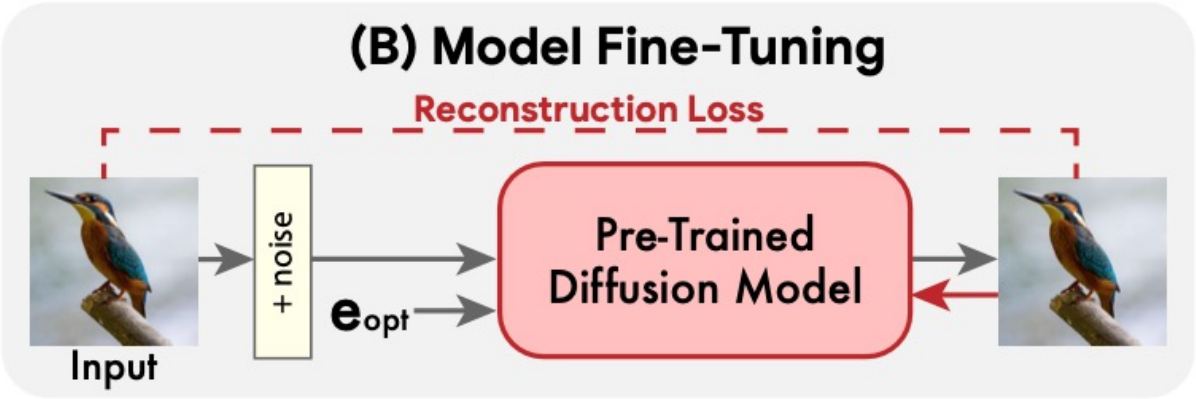
Imagic: Text-Based Real Image Editing

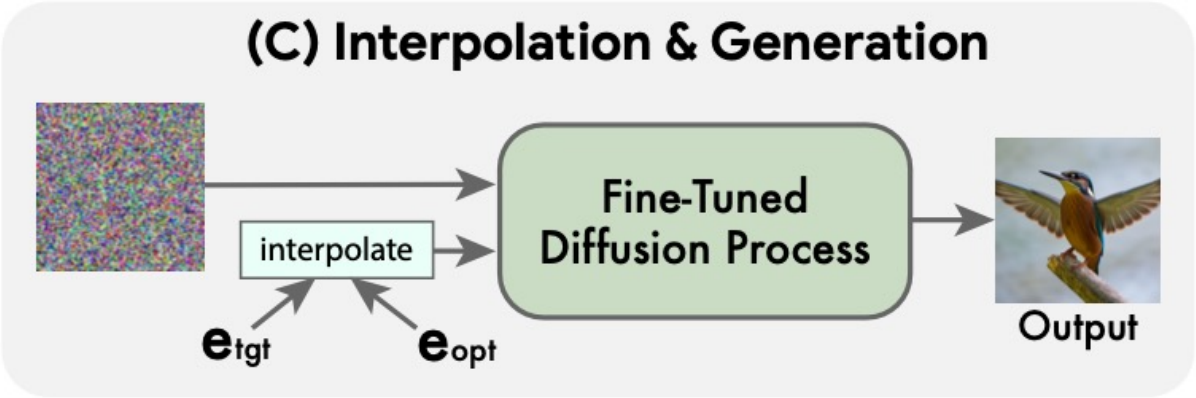


Mokady, Ron, et al. "Null-text Inversion for Editing Real Images using Guided Diffusion Models." arXiv preprint arXiv:2211.09794 (2022).

Imagic: Text-Based Real Image Editing







Mokady, Ron, et al. "Null-text Inversion for Editing Real Images using Guided Diffusion Models." arXiv preprint arXiv:2211.09794 (2022).

Imagic: Text-Based Real Image Editing

Text embedding optimization The target text is first passed through a text encoder [43], which outputs its corresponding text embedding $\mathbf{e}_{tgt} \in \mathbb{R}^{T \times d}$, where T is the number of tokens in the given target text, and d is the token embedding dimension. We then freeze the parameters of the generative diffusion model f_θ, and optimize the target text embedding $\mathbf{e}_{tgt}$ using the denoising diffusion objective [19]:

$$\mathcal{L}(\mathbf{x}, \mathbf{e}, \theta) = \mathbb{E}_{t,\boldsymbol{\epsilon}} \left[\|\boldsymbol{\epsilon} - f_\theta(\mathbf{x}_t, t, \mathbf{e})\|_2^2 \right], \qquad (2)$$

where $t \sim Uniform[1, T]$, $\mathbf{x}_t$ is a noisy version of $\mathbf{x}$ (the input image) obtained using $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$ and Equation 1, and θ are the pre-trained diffusion model weights. This results in a text embedding that matches our input image as closely as possible. We run this process for relatively few steps, in order to remain close to the initial target text embedding, obtaining $\mathbf{e}_{opt}$. This proximity enables meaningful linear

Text embedding interpolation Since the generative diffusion model was trained to fully recreate the input image $\mathbf{x}$ at the optimized embedding $\mathbf{e}_{opt}$, we use it to apply the desired edit by advancing in the direction of the target text embedding $\mathbf{e}_{tgt}$. More formally, our third stage is a simple linear interpolation between $\mathbf{e}_{tgt}$ and $\mathbf{e}_{opt}$. For a given hyperparameter $\eta \in [0, 1]$, we obtain

$$\bar{\mathbf{e}} = \eta \cdot \mathbf{e}_{tgt} + (1 - \eta) \cdot \mathbf{e}_{opt}, \qquad (3)$$

Mokady, Ron, et al. "Null-text Inversion for Editing Real Images using Guided Diffusion Models." arXiv preprint arXiv:2211.09794 (2022).

Imagic: Text-Based Real Image Editing

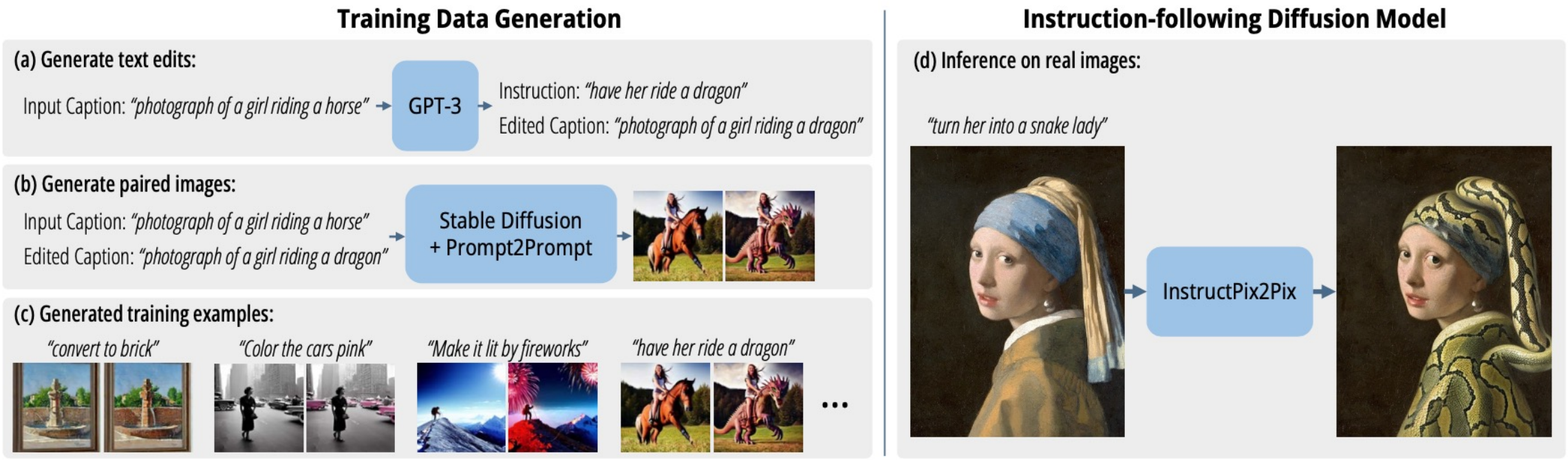
Figure 7. **Embedding interpolation.** *Varying η with the same seed, using the pre-trained (top) and fine-tuned (bottom) models.*

Mokady, Ron, et al. "Null-text Inversion for Editing Real Images using Guided Diffusion Models." arXiv preprint arXiv:2211.09794 (2022).

InstructPix2Pix

Brooks, Tim, Aleksander Holynski, and Alexei A. Efros. "Instructpix2pix: Learning to follow image editing instructions." arXiv preprint arXiv:2211.09800 (2022).

InstructPix2Pix



Brooks, Tim, Aleksander Holynski, and Alexei A. Efros. "Instructpix2pix: Learning to follow image editing instructions." arXiv preprint arXiv:2211.09800 (2022).

InstructPix2Pix

	Input LAION caption	Edit instruction	Edited caption
Human-written (700 edits)	*Yefim Volkov, Misty Morning*	*make it afternoon*	*Yefim Volkov, Misty Afternoon*
	girl with horse at sunset	*change the background to a city*	*girl with horse at sunset in front of city*
	painting-of-forest-and-pond	*Without the water.*	*painting-of-forest*
	...	...	...
GPT-3 generated (>450,000 edits)	*Alex Hill, Original oil painting on canvas, Moonlight Bay*	*in the style of a coloring book*	*Alex Hill, Original coloring book illustration, Moonlight Bay*
	The great elf city of Rivendell, sitting atop a waterfall as cascades of water spill around it	*Add a giant red dragon*	*The great elf city of Rivendell, sitting atop a waterfall as cascades of water spill around it with a giant red dragon flying overhead*
	Kate Hudson arriving at the Golden Globes 2015	*make her look like a zombie*	*Zombie Kate Hudson arriving at the Golden Globes 2015*
	...	...	...

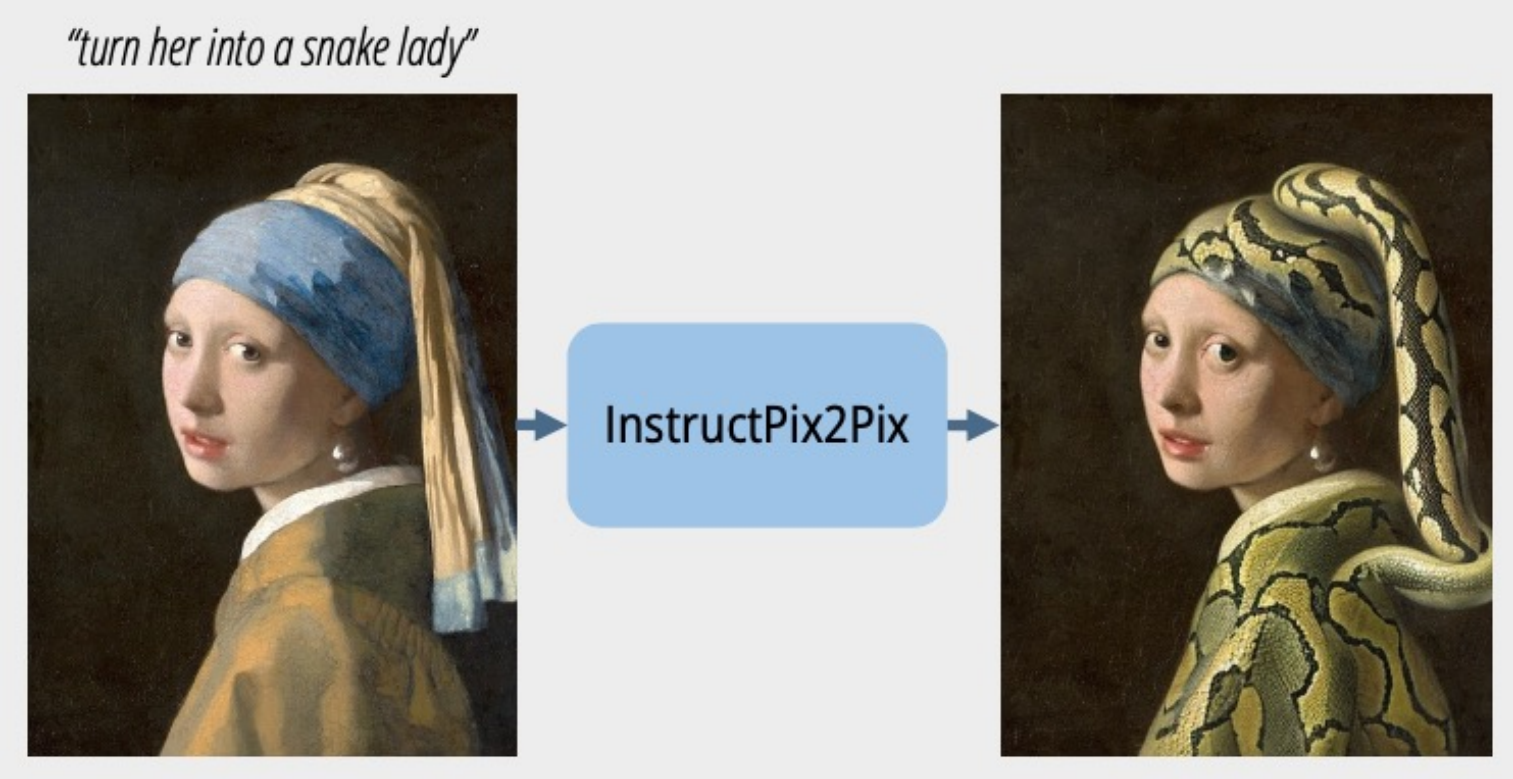
Brooks, Tim, Aleksander Holynski, and Alexei A. Efros. "Instructpix2pix: Learning to follow image editing instructions." arXiv preprint arXiv:2211.09800 (2022).

InstructPix2Pix

(a) Without Prompt-to-Prompt. (b) With Prompt-to-Prompt.

Brooks, Tim, Aleksander Holynski, and Alexei A. Efros. "Instructpix2pix: Learning to follow image editing instructions." arXiv preprint arXiv:2211.09800 (2022).

InstructPix2Pix



Brooks, Tim, Aleksander Holynski, and Alexei A. Efros. "Instructpix2pix: Learning to follow image editing instructions." arXiv preprint arXiv:2211.09800 (2022).

InstructPix2Pix

Figure 5. *Mona Lisa* transformed into various artistic mediums.

Figure 6. *The Creation of Adam* with new context and subjects (generated at 768 resolution).

Brooks, Tim, Aleksander Holynski, and Alexei A. Efros. "Instructpix2pix: Learning to follow image editing instructions." arXiv preprint arXiv:2211.09800 (2022).

InstructPix2Pix

Figure 11. Applying our model recurrently with different instructions results in compounded edits.

Input

Edit instruction: *"in a race car video game"*

Figure 12. By varying the latent noise, our model can produce many possible image edits for the same input image and instruction.

Brooks, Tim, Aleksander Holynski, and Alexei A. Efros. "Instructpix2pix: Learning to follow image editing instructions." arXiv preprint arXiv:2211.09800 (2022).

InstructPix2Pix

"Zoom into the image" *"Move it to Mars"* *"Color the tie blue"* *"Have the people swap places"*

Figure 13. Failure cases. Left to right: our model is not capable of performing viewpoint changes, can make undesired excessive changes to the image, can sometimes fail to isolate the specified object, and has difficulty reorganizing or swapping objects with each other.

Brooks, Tim, Aleksander Holynski, and Alexei A. Efros. "Instructpix2pix: Learning to follow image editing instructions." arXiv preprint arXiv:2211.09800 (2022).

Other Works on Real Image Editing

“UniTune: Text-Driven Image Editing by Fine Tuning an Image Generation Model on a Single Image” (2022)

“SINE: SINgle Image Editing with Text-to-Image Diffusion Models” (2022)

“Zero-shot Image-to-Image Translation” (2023)