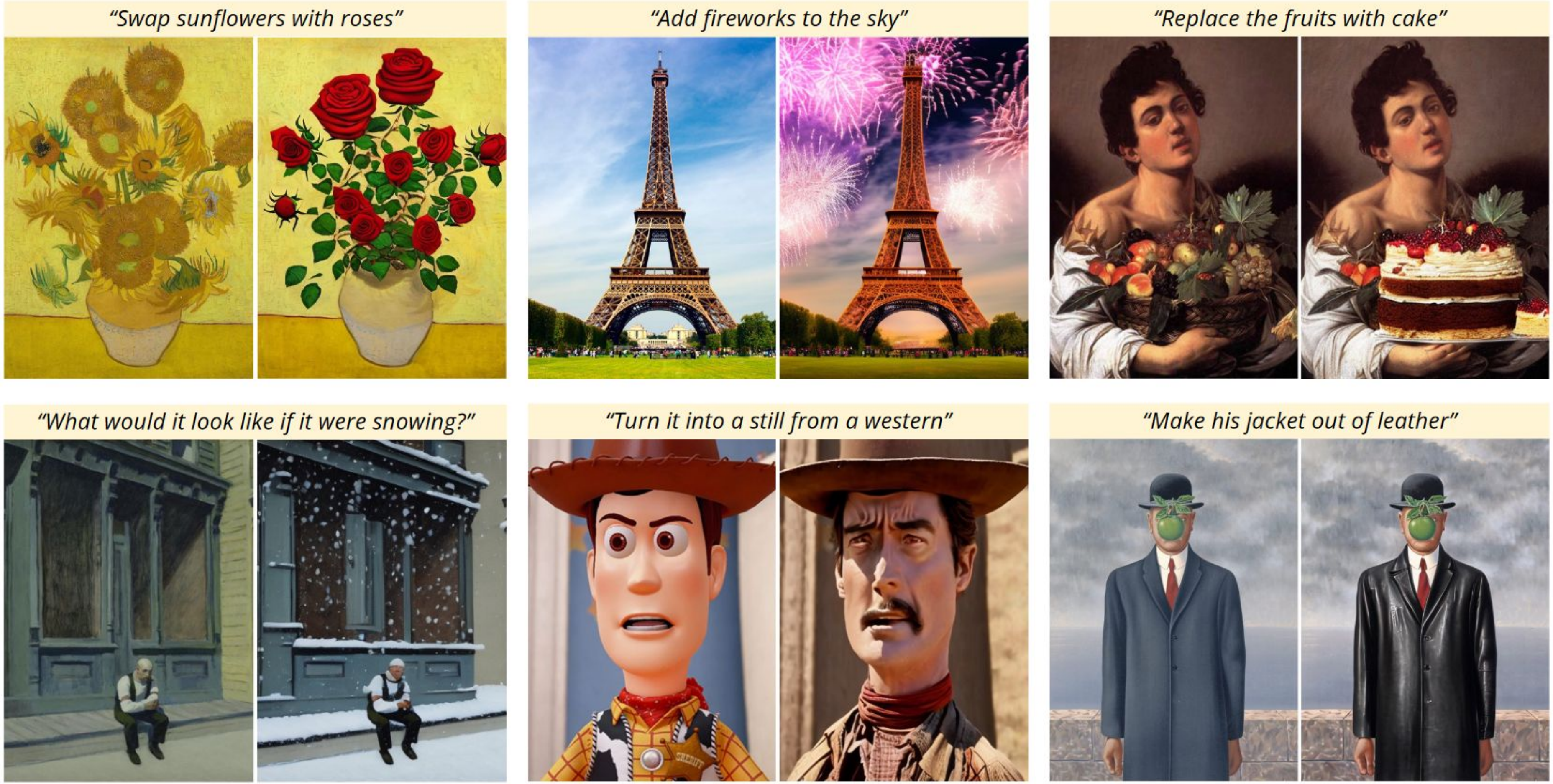
InstructPix2Pix: Learning to Follow Image Editing Instructions

Кондратьев Захар



"Swap sunflowers with roses"

"Add fireworks to the sky"

"Replace the fruits with cake"

"What would it look like if it were snowing?"

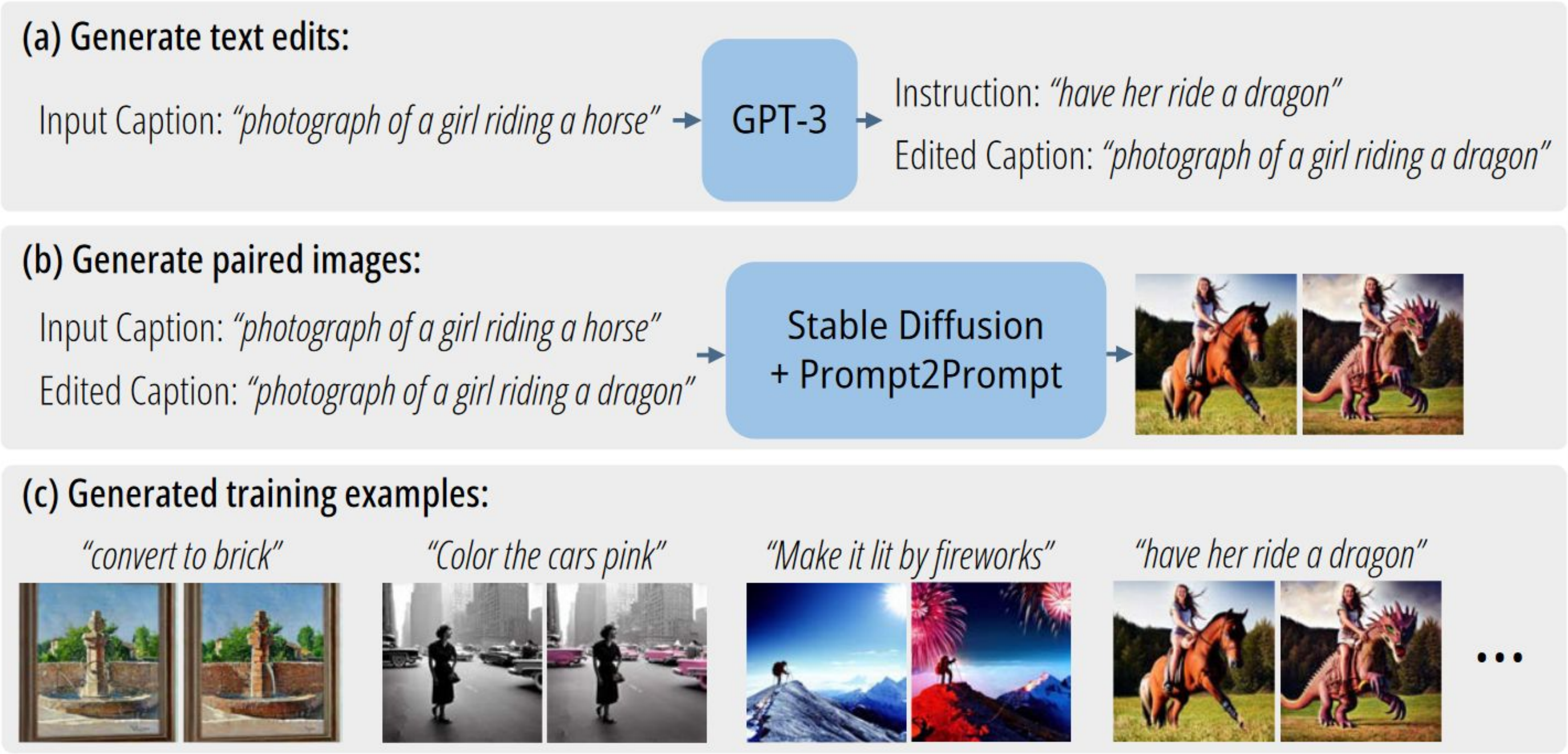
"Turn it into a still from a western"

"Make his jacket out of leather"

Генерация данных

Набор данных LAION-Aesthetics из картинок и подписей к ним.
(625K пар)

Генерация данных



(a) Generate text edits:

Input Caption: *"photograph of a girl riding a horse"* → GPT-3 →

Instruction: *"have her ride a dragon"*

Edited Caption: *"photograph of a girl riding a dragon"*

(b) Generate paired images:

Input Caption: *"photograph of a girl riding a horse"*

Edited Caption: *"photograph of a girl riding a dragon"*

→ Stable Diffusion + Prompt2Prompt →

(c) Generated training examples:

"convert to brick"

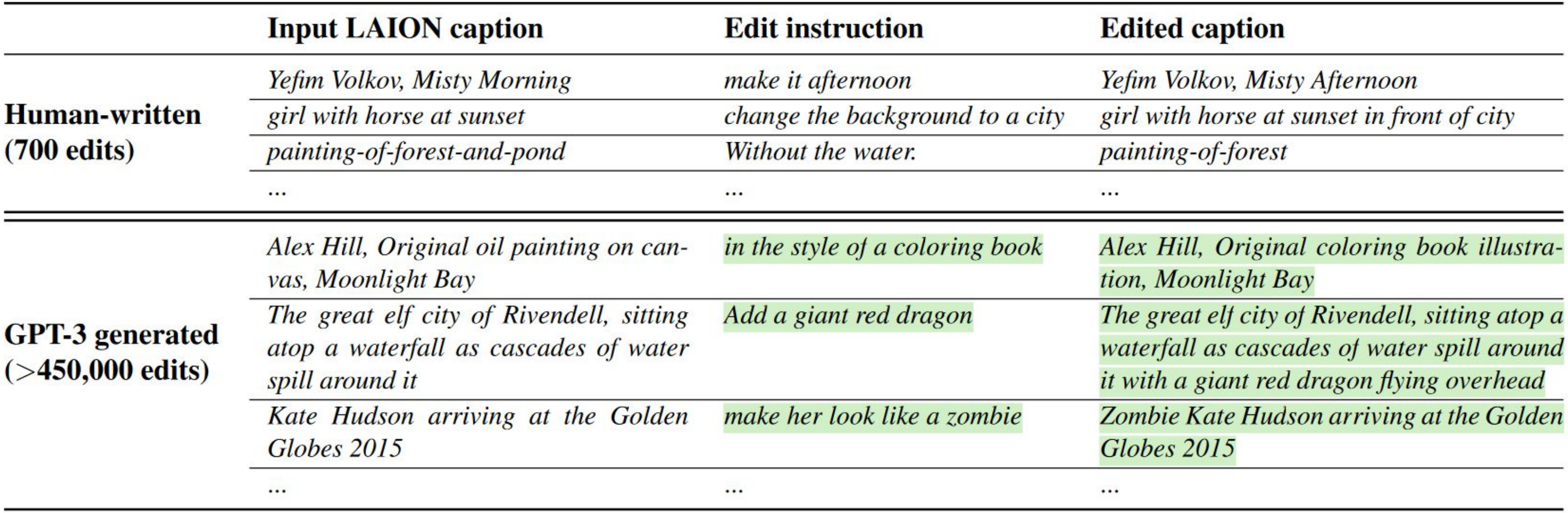
"Color the cars pink"

"Make it lit by fireworks"

"have her ride a dragon"

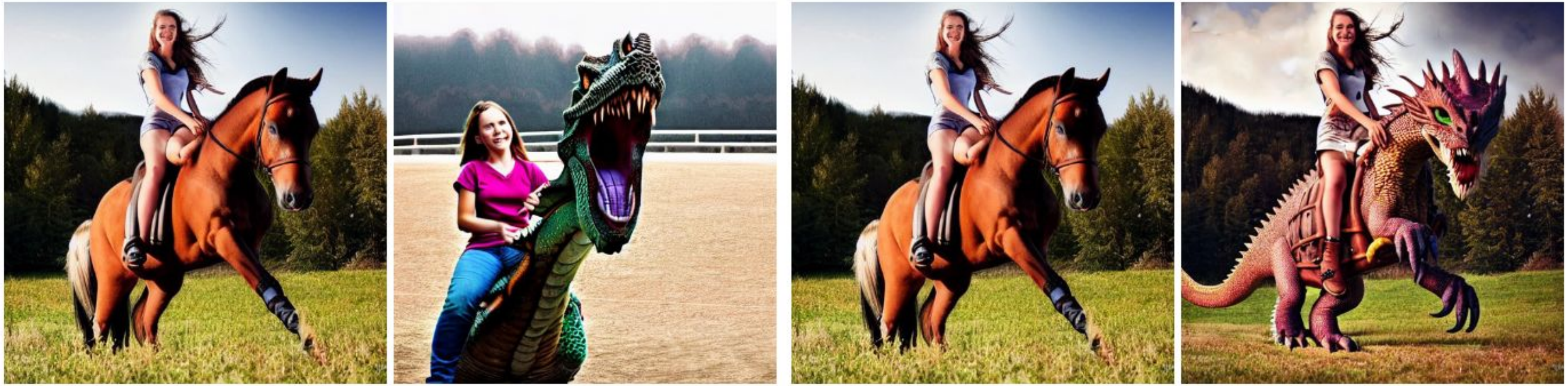
...

Генерация данных (a)



	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*
	...	...	...

Генерация данных (b)



(a) Without Prompt-to-Prompt.

(b) With Prompt-to-Prompt.

Генерация данных (b)

Для каждой пары подписей получают 100 пар картинок с разными значениями параметра p (отвечает за схожесть двух изображений).

Отбирают с помощью CLIP directional similarity (сравниваются разности текстов и картинок в пространстве CLIP)

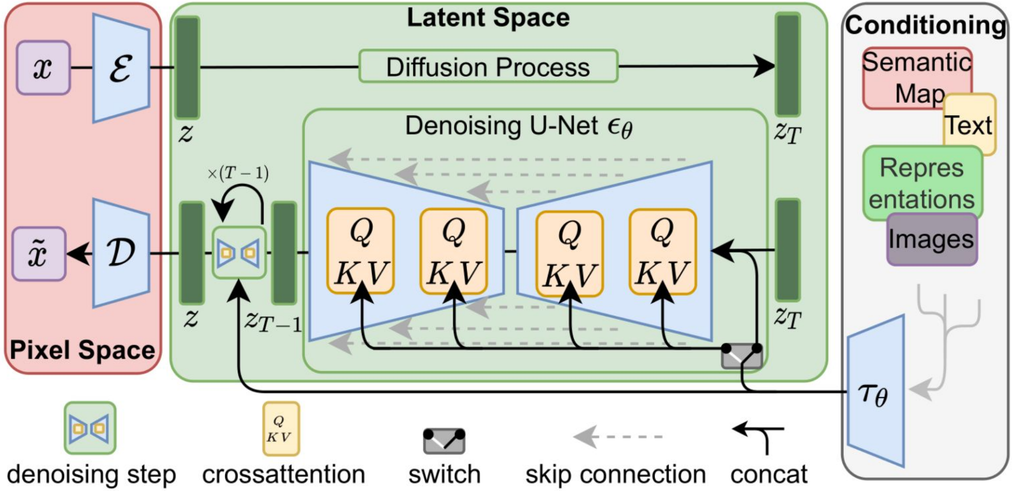
Получили более 450K примеров для обучения.

InstructPix2Pix

За основу взяли Stable Diffusion (веса тоже)

Особенности:

- Latent diffusion
- Classifier-free Guidance

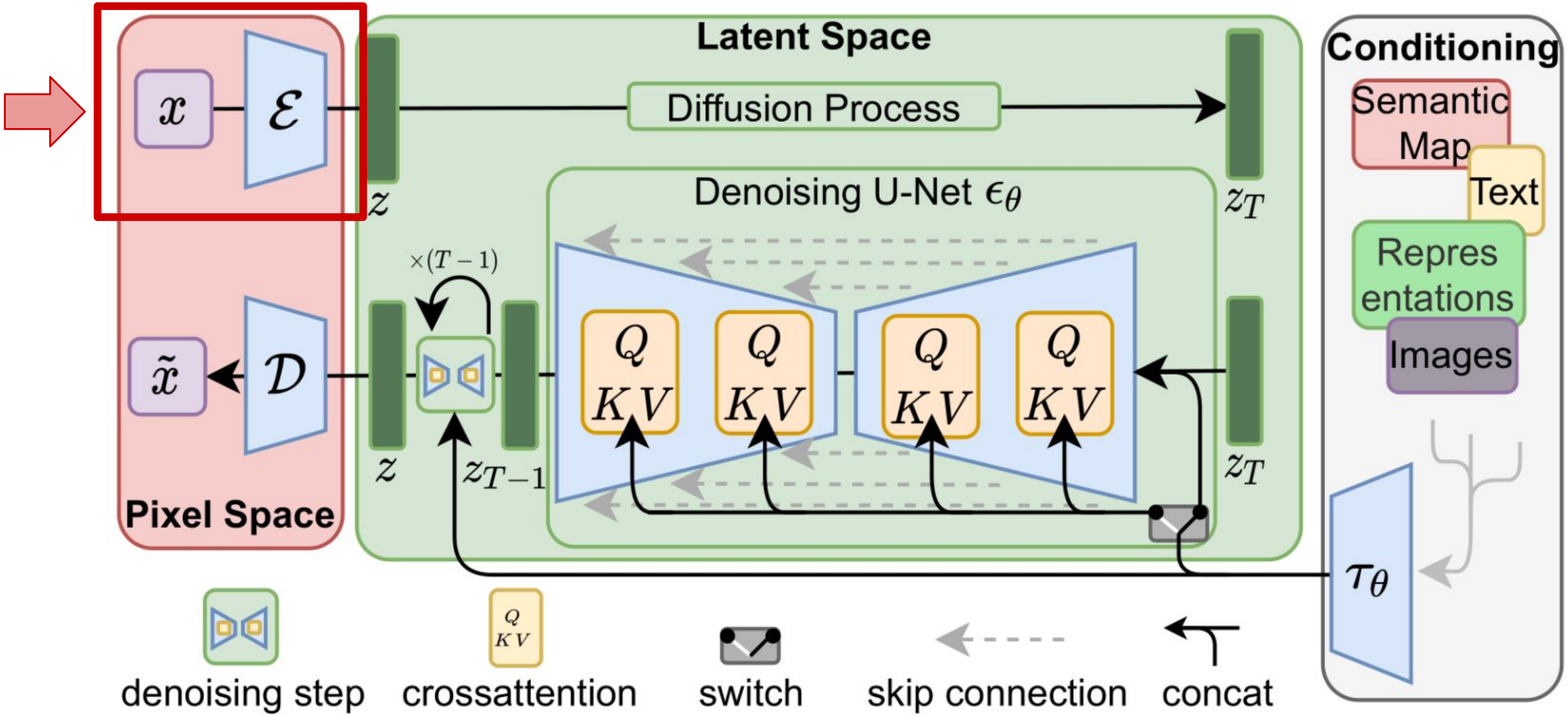


Latent diffusion

Кратко:

- Вместо картинок работаем с их представлениями.
- Для представлений предобученный автоэнкодер.

Latent diffusion



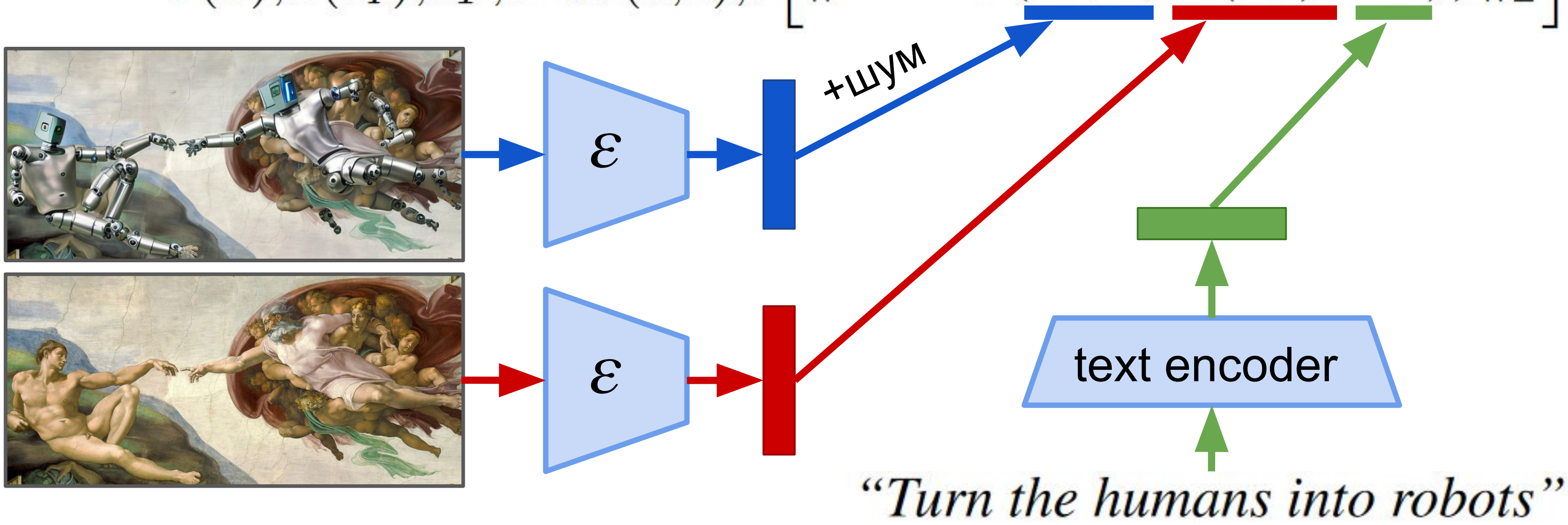
Latent diffusion

Зачем?

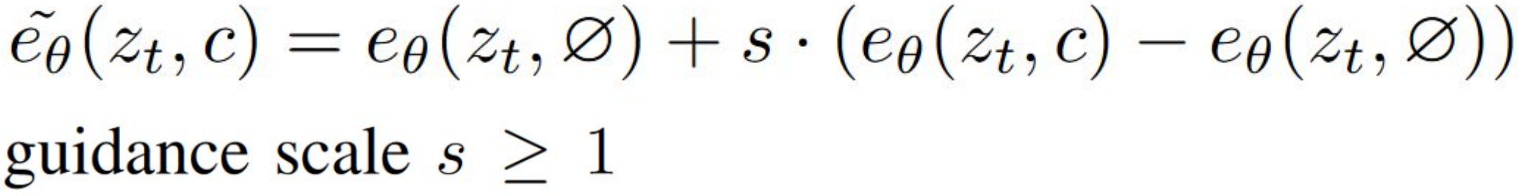
- Быстрее
- Лучше качество

Обучение

$$L = \mathbb{E}_{\mathcal{E}(x),\mathcal{E}(c_I),c_T,\epsilon\sim\mathcal{N}(0,1),t}\Big[\|\epsilon - \overbrace{\epsilon_\theta(z_t, t, \mathcal{E}(c_I), c_T))}^{\text{модель}}\|_2^2\Big]$$



Classifier-free Guidance



$$\tilde{e_\theta}(z_t, c) = e_\theta(z_t, \varnothing) + s \cdot (e_\theta(z_t, c) - e_\theta(z_t, \varnothing))$$

guidance scale $s \geq 1$

Чем больше коэффициент, тем “ближе” мы к конкретному классу и “дальше” от обобщённого предсказания.

Classifier-free Guidance

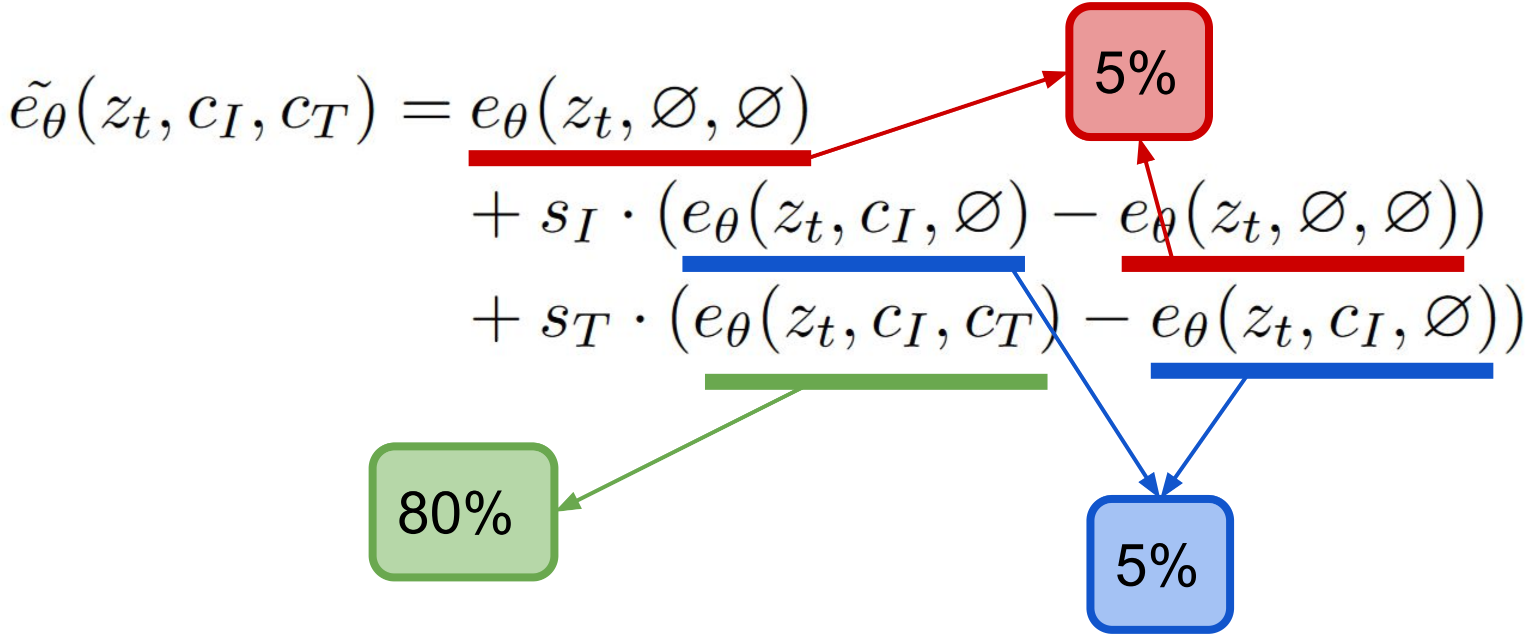
$$\begin{aligned}\tilde{e_\theta}(z_t, c_I, c_T) = e_\theta(z_t, \varnothing, \varnothing) \\ + s_I \cdot (e_\theta(z_t, c_I, \varnothing) - e_\theta(z_t, \varnothing, \varnothing)) \\ + s_T \cdot (e_\theta(z_t, c_I, c_T) - e_\theta(z_t, c_I, \varnothing))\end{aligned}$$

Согласно формулам получается:

$$s_I \longrightarrow p_\theta(c_I | z_t)$$

$$s_T \longrightarrow p_\theta(c_T | c_I, z_t)$$

Обучение

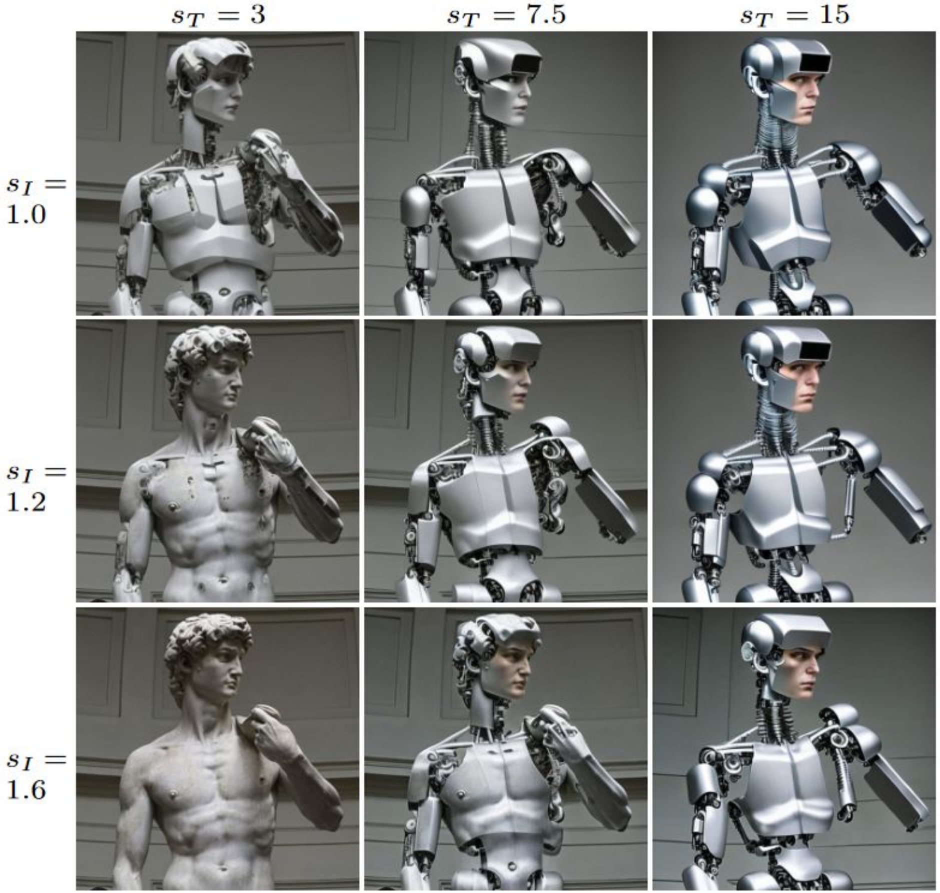


$$\tilde{e}_\theta(z_t, c_I, c_T) = e_\theta(z_t, \varnothing, \varnothing) + s_I \cdot (e_\theta(z_t, c_I, \varnothing) - e_\theta(z_t, \varnothing, \varnothing)) + s_T \cdot (e_\theta(z_t, c_I, c_T) - e_\theta(z_t, c_I, \varnothing))$$

5%

80%

5%



Edit instruction: *“Turn him into a cyborg!”*

Сравнение

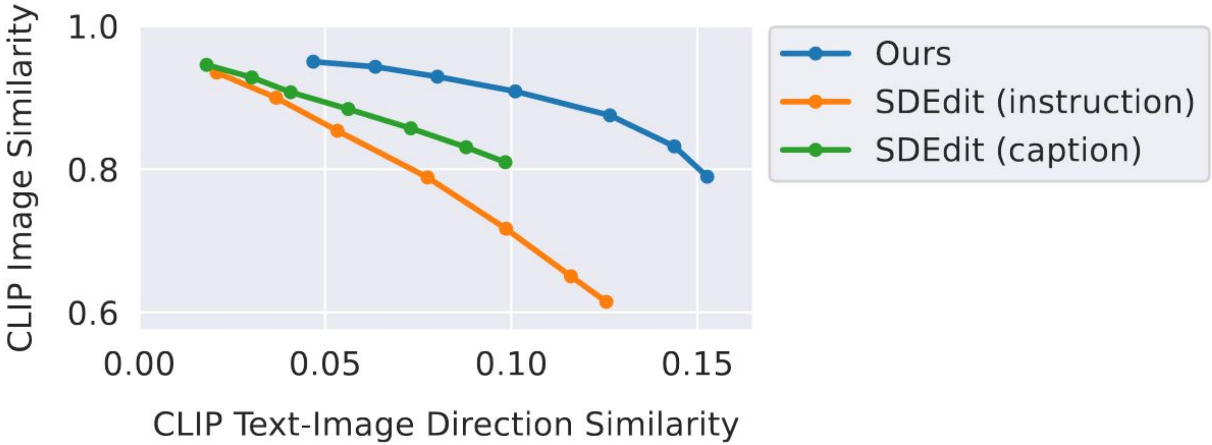
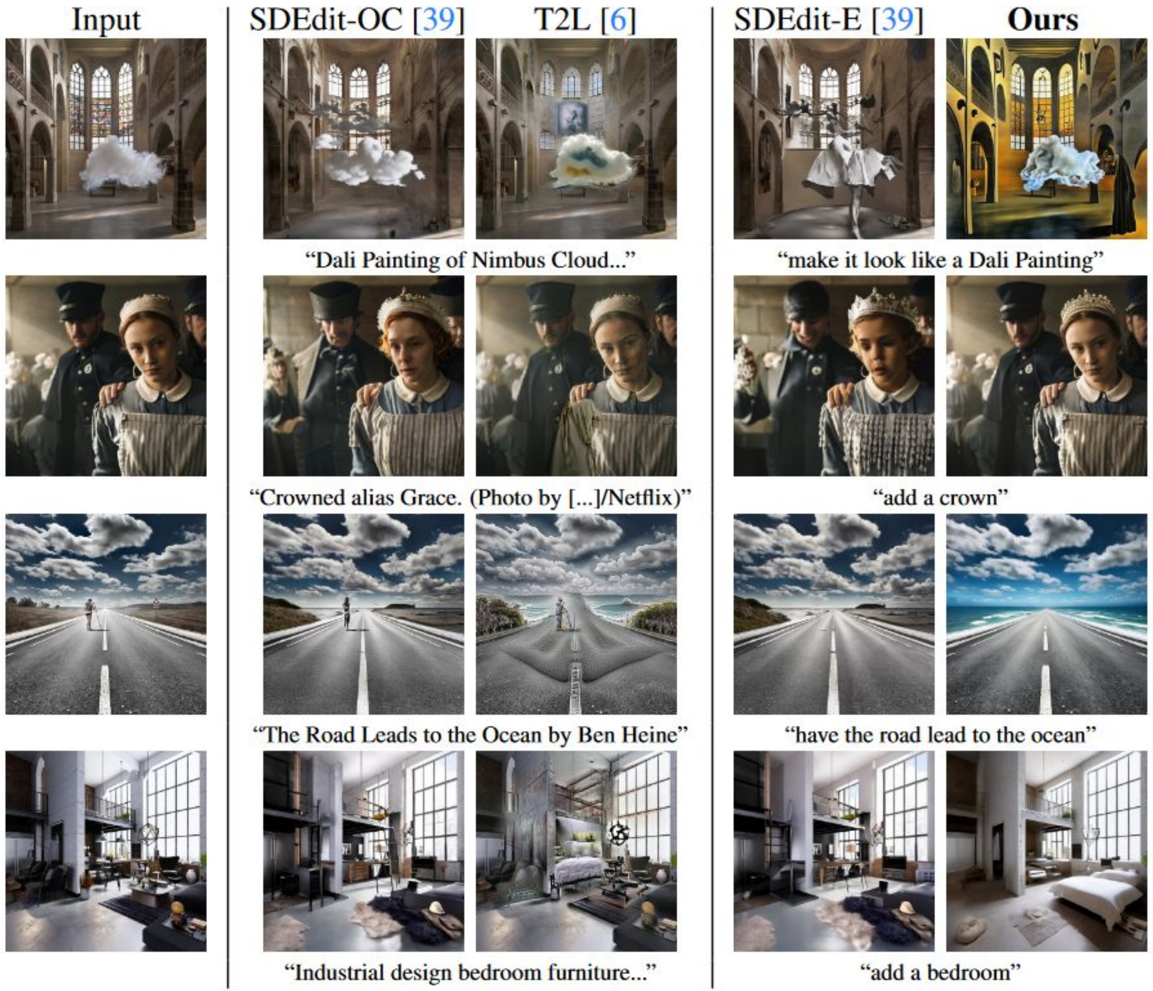
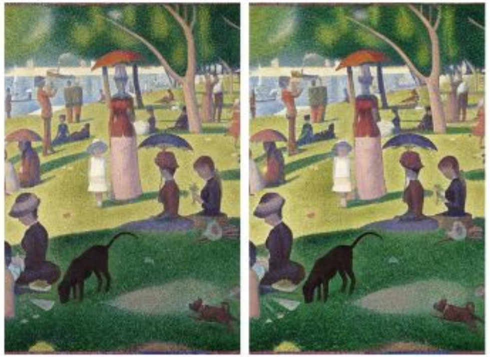


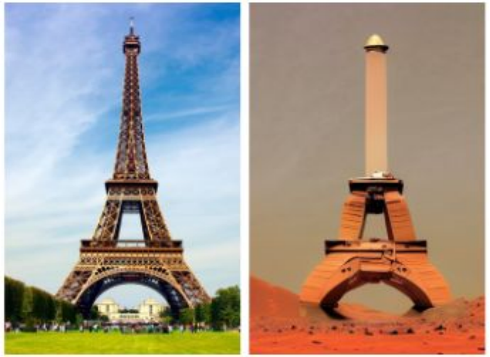
Figure 8. We plot the trade-off between consistency with the input image (Y-axis) and consistency with the edit (X-axis). For both metrics, higher is better. For both methods, we fix text guidance to 7.5, and vary our $s_I \in [1.0, 2.2]$ and SDEdit's strength (the amount of denoising) between $[0.3, 0.9]$.


Input
SDEdit-OC [39]
T2L [6]
SDEdit-E [39]
Ours
“Dali Painting of Nimbus Cloud...”
“make it look like a Dali Painting”
“Crowned alias Grace. (Photo by [...]/Netflix)”
“add a crown”
“The Road Leads to the Ocean by Ben Heine”
“have the road lead to the ocean”
“Industrial design bedroom furniture...”
“add a bedroom”

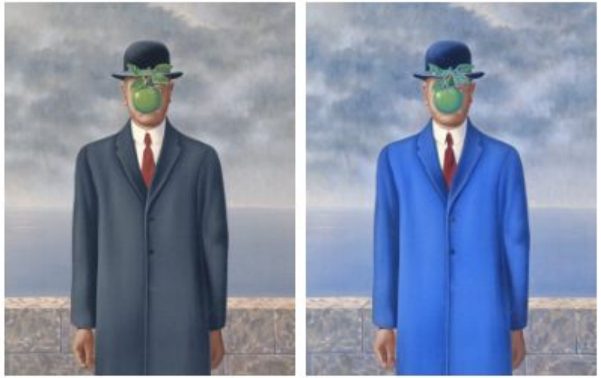
Что не получается



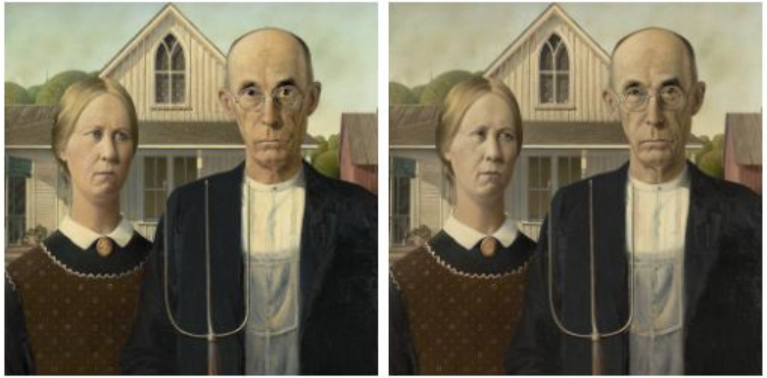
“Zoom into the image”



“Move it to Mars”



“Color the tie blue”



“Have the people swap places”

- https://arxiv.org/pdf/2211.09800.pdf
- https://arxiv.org/pdf/2112.10752.pdf