Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding

Abstract

- Large language transformers improve image generation
- Increasing size of LM >> Increasing size of generator
- Zero-Shot SotA on CoCo - 7.27 FID
- DrawBench - new comprehensive benchmark for text2img models
- dynamic thresholding — sampling technique for generating more photorealistic and detailed images
- Efficient U-Net — simpler, converges faster,more memory efficient

Введение в Classifier Free Guidance

1.Classifier Guidance - техника сэмплирования для улучшения качества получаемых изображений, путем снижения многообразия генерируемых изображений диффузионной моделью в задаче условной генерации, используя градиент предобученной модели $p(\mathbf{c}|\mathbf{z}_t)$

$$\mathbb{E}_{\boldsymbol{\epsilon},\lambda}\left[\|\boldsymbol{\epsilon}_\theta(\mathbf{z}_\lambda) - \boldsymbol{\epsilon}\|_2^2\right] \quad \boldsymbol{\epsilon}_\theta(\mathbf{z}_\lambda, \mathbf{c}) \approx -\sigma_\lambda \nabla_{\mathbf{z}_\lambda} \log p(\mathbf{z}_\lambda|\mathbf{c})$$

$$\boldsymbol{\epsilon}_\theta(\mathbf{z}_\lambda) \longrightarrow \boldsymbol{\epsilon}_\theta(\mathbf{z}_\lambda, \mathbf{c}).$$

Classifier-Free Guidance

$$\mathbb{E}_{\boldsymbol{\epsilon},\lambda}\left[\|\boldsymbol{\epsilon}_\theta(\mathbf{z}_\lambda) - \boldsymbol{\epsilon}\|_2^2\right] \qquad \boldsymbol{\epsilon}_\theta(\mathbf{z}_\lambda) \longrightarrow \boldsymbol{\epsilon}_\theta(\mathbf{z}_\lambda, \mathbf{c}).$$

$$\boldsymbol{\epsilon}_\theta(\mathbf{z}_\lambda, \mathbf{c}) \approx -\sigma_\lambda \nabla_{\mathbf{z}_\lambda} \log p(\mathbf{z}_\lambda|\mathbf{c})$$

Введение в Classifier Free Guidance

1.Classifier Free Guidance - вместо предобученной модели $p(\mathbf{c}|\mathbf{z}_t)$ обучаем нашу диффузионную модель с использованием двух функционалов: условного и безусловного. Во время обучения с некой вероятностью (10%) отбрасывается класс изображения.

$$\mathbb{E}_{\boldsymbol{\epsilon},\lambda}\left[\|\boldsymbol{\epsilon}_\theta(\mathbf{z}_\lambda) - \boldsymbol{\epsilon}\|_2^2\right] \quad \boldsymbol{\epsilon}_\theta(\mathbf{z}_\lambda, \mathbf{c}) \approx -\sigma_\lambda \nabla_{\mathbf{z}_\lambda} \log p(\mathbf{z}_\lambda|\mathbf{c})$$

$$\tilde{\boldsymbol{\epsilon}}_\theta(\mathbf{z}_t, \mathbf{c}) = w\boldsymbol{\epsilon}_\theta(\mathbf{z}_t, \mathbf{c}) + (1-w)\boldsymbol{\epsilon}_\theta(\mathbf{z}_t).$$

Проблемы связанные с CFG:

Повышение CFG scale способствует большему согласованию между запросом и изображением, однако при высоком значении параметра появляются артефакты в виде засветов, глитчей, тёмных пятен

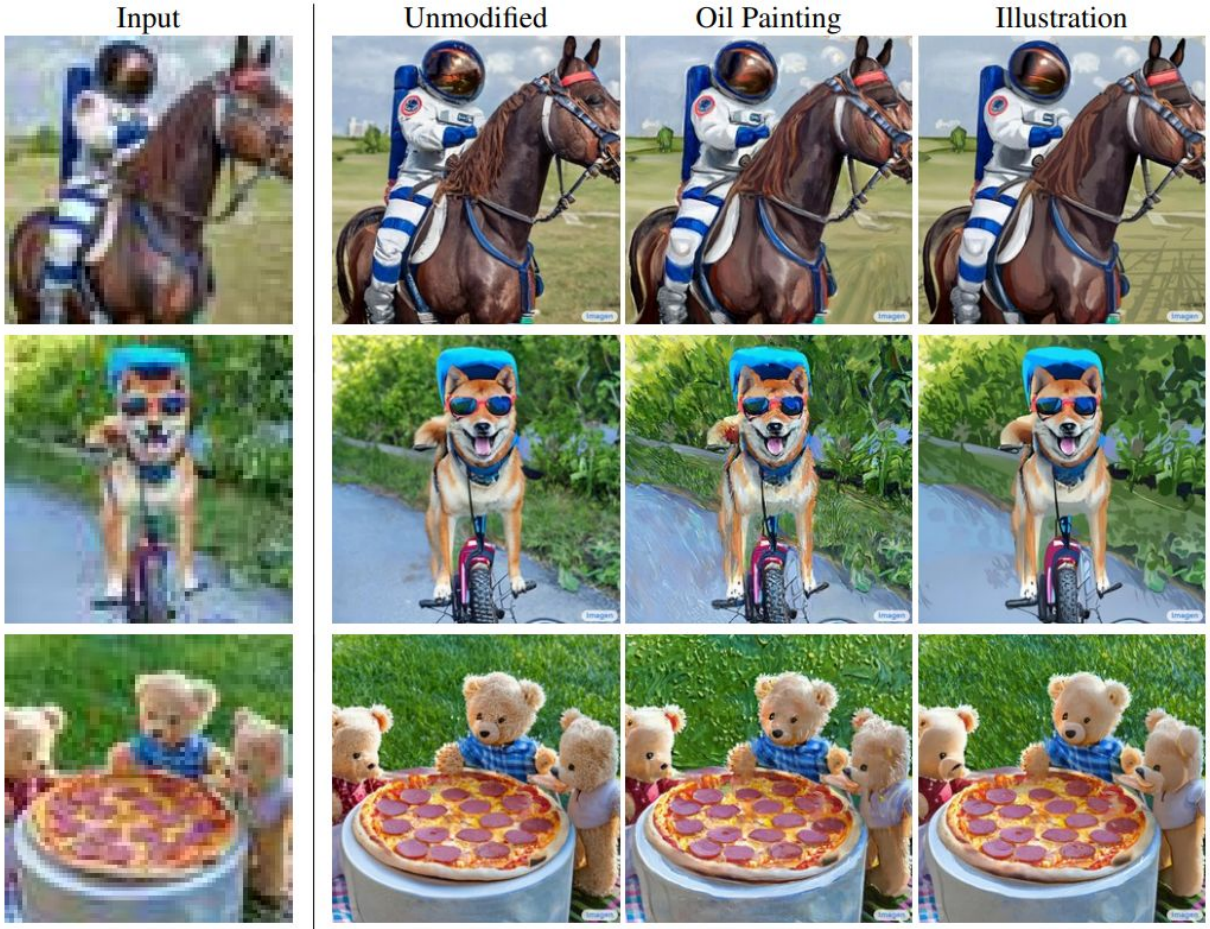
Static thresholding and dynamic thresholding

(a) No thresholding.

(b) Static thresholding.

(c) Dynamic thresholding.

Upscalers noise conditioning augmentation

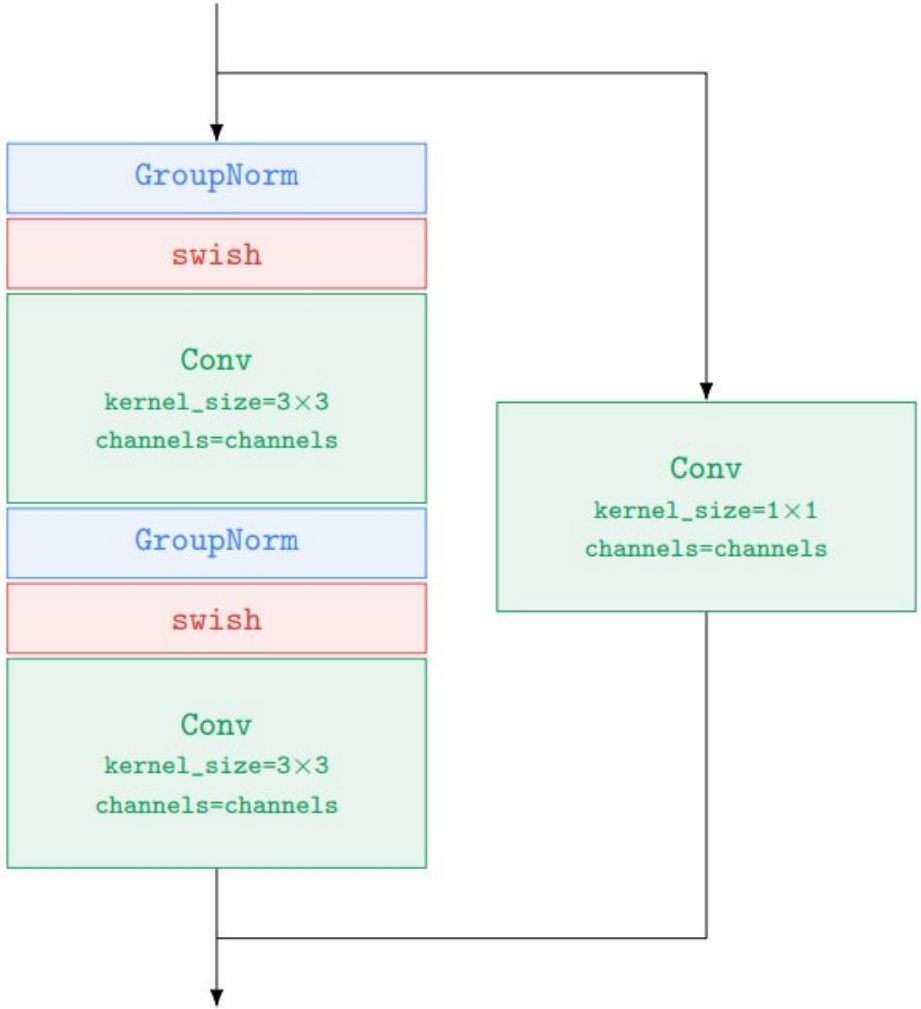


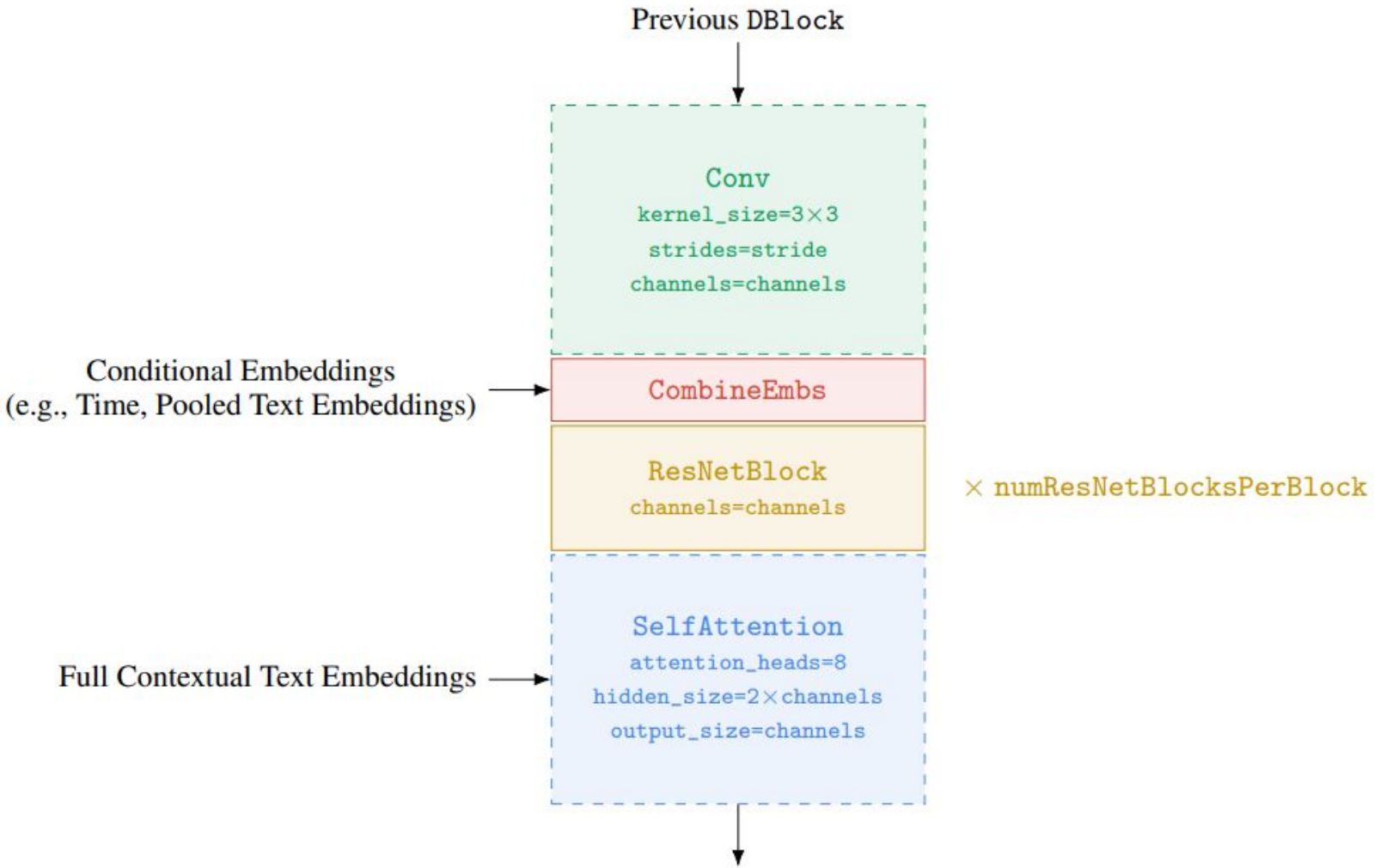
Архитектура модели

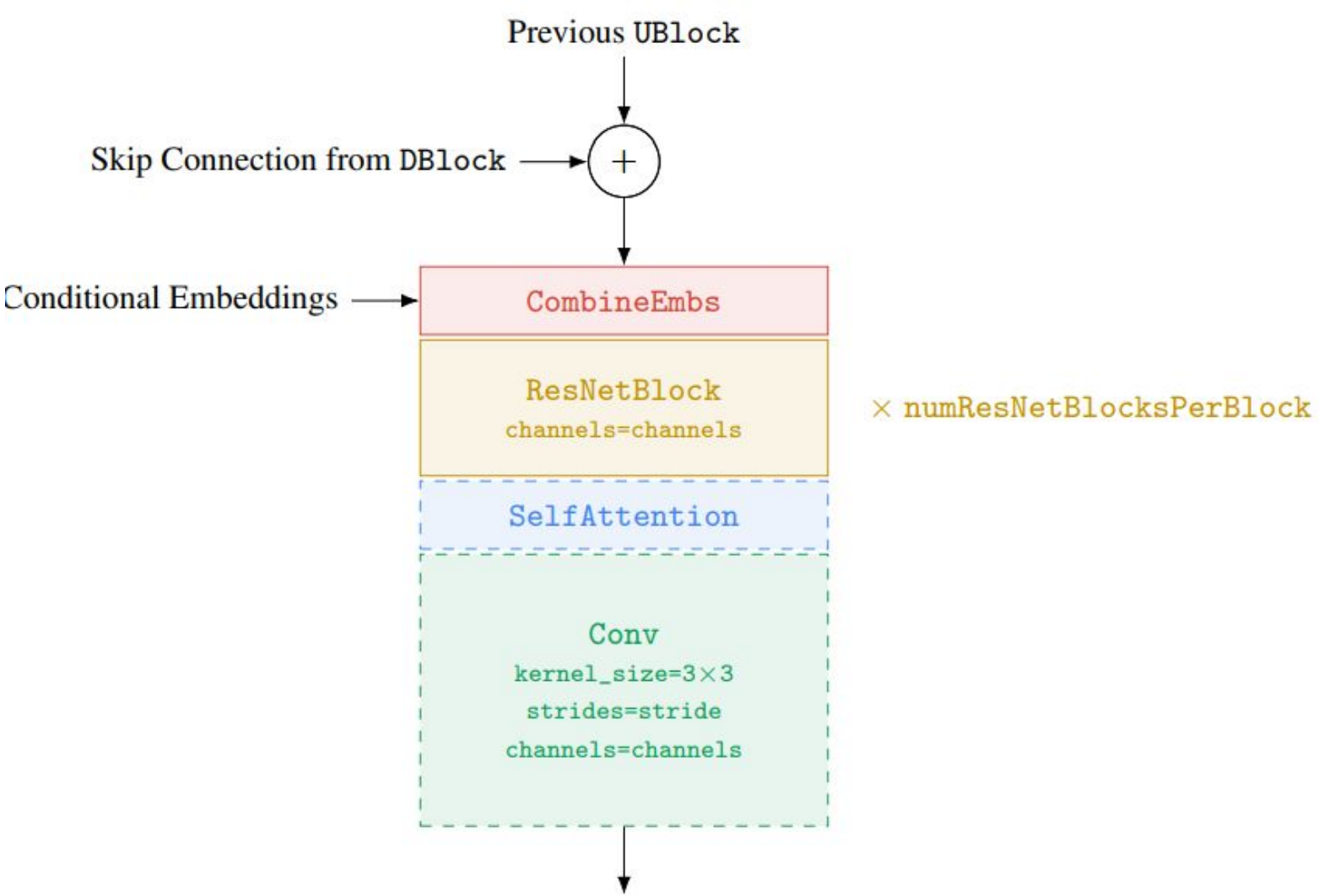
- Base model - U-net –> 64x64
- Super resolution models - модифицированная U-net. Увеличена скорость сходимости, инференса, лучше эффективность по памяти - Efficient U-net.

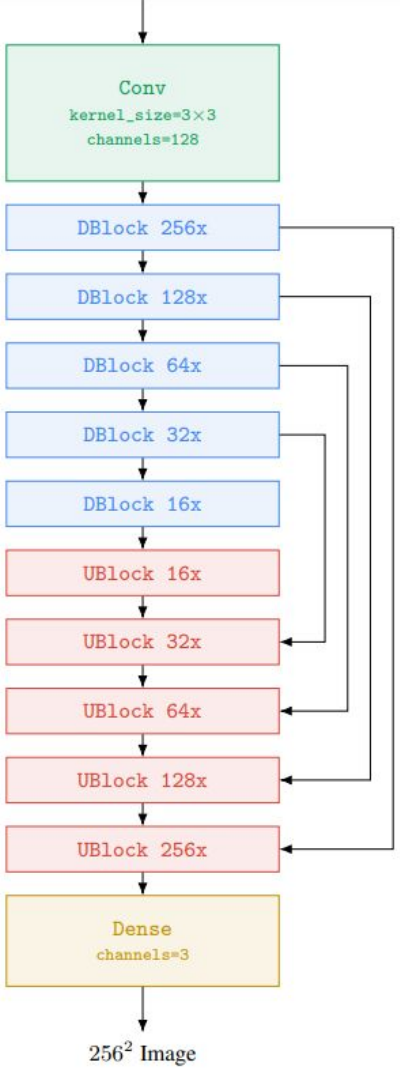
Efficient U-net

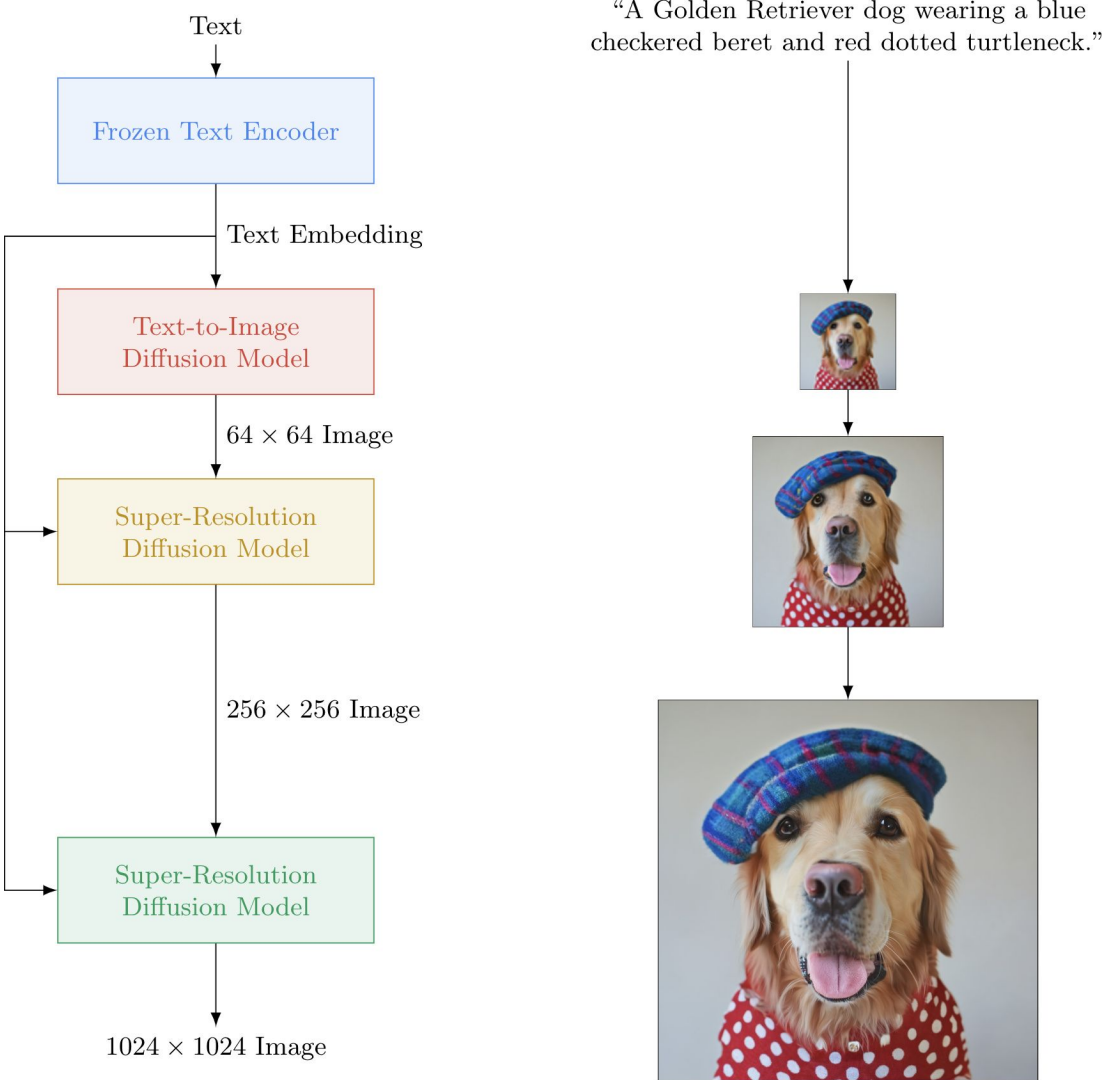
- Перенос параметров модели с блоков с высоким разрешением к low res блокам.
- Масштабируем skip-connections на 1/sqrt(2)
- Операции апсемплинга\даунсемплинга выполняются в обратном порядке для повышения скорости


GroupNorm
swish
Conv
kernel_size=3×3
channels=channels
GroupNorm
swish
Conv
kernel_size=3×3
channels=channels
Conv
kernel_size=1×1
channels=channels


Previous DBlock
Conv
kernel_size=3×3
strides=stride
channels=channels
Conditional Embeddings
(e.g., Time, Pooled Text Embeddings)
CombineEmbs
ResNetBlock
channels=channels
× numResNetBlocksPerBlock
SelfAttention
attention_heads=8
hidden_size=2×channels
output_size=channels
Full Contextual Text Embeddings


Previous UBlock
Skip Connection from DBlock
+
Conditional Embeddings
CombineEmbs
ResNetBlock
channels=channels
× numResNetBlocksPerBlock
SelfAttention
Conv
kernel_size=3×3
strides=stride
channels=channels


Conv
kernel_size=3×3
channels=128
DBlock 256x
DBlock 128x
DBlock 64x
DBlock 32x
DBlock 16x
UBlock 16x
UBlock 32x
UBlock 64x
UBlock 128x
UBlock 256x
Dense
channels=3
256^2 Image


Text
Frozen Text Encoder
Text Embedding
Text-to-Image Diffusion Model
64 × 64 Image
Super-Resolution Diffusion Model
256 × 256 Image
Super-Resolution Diffusion Model
1024 × 1024 Image
"A Golden Retriever dog wearing a blue checkered beret and red dotted turtleneck."

Evaluating text2image models

- COCO - базовый бенчмарк для text2img моделей
- Тестируем на 30к сэмплах
- FID имеет свои недостатки как метрика. Хотим тестировать качество на реальных людях.


A brown bird and a blue bear.


One cat and two dogs sitting on the grass.

A sign that says 'NeurIPS'.


A small blue book sitting on a large red book.


A blue coloured pizza.


A wine glass on top of a dog.


A pear cut into seven pieces arranged in a ring.


A photo of a confused grizzly bear in calculus class.


A small vessel propelled on water by oars, sails, or an engine.

DrawBench

- COCO ограничен по промптам, поэтому не позволяет протестировать модель на все 100.
- 11 категорий промптов
- Тестируем модель на способность генерации конкретных цветов объектов, определенного количества, отношений между предметами, текста на изображении
- Длинные промпты
- Промпты со словами, которые редко используются
- Промпты, в которых отсутсвует часть слов

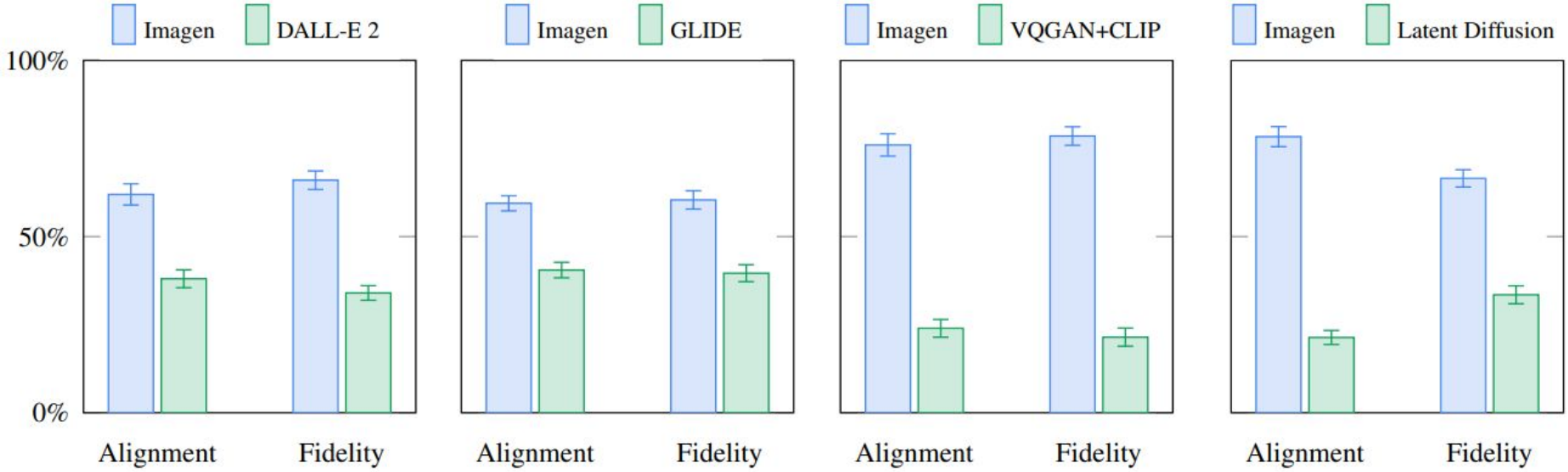
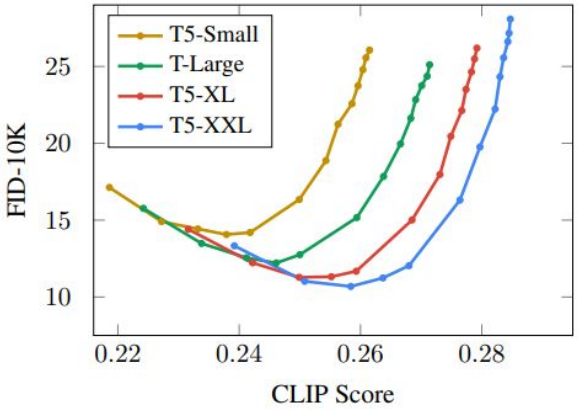


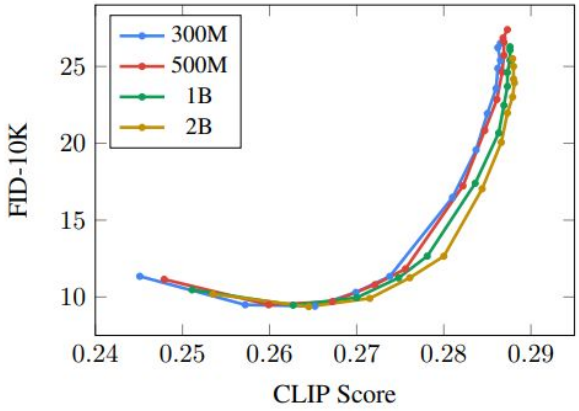
Figure 3: Comparison between Imagen and DALL-E 2 [54], GLIDE [41], VQ-GAN+CLIP [12] and Latent Diffusion [57] on DrawBench: User preference rates (with 95% confidence intervals) for image-text alignment and image fidelity.

Model	FID-30K	Zero-shot FID-30K
AttnGAN [76]	35.49	
DM-GAN [83]	32.64	
DF-GAN [69]	21.42	
DM-GAN + CL [78]	20.79	
XMC-GAN [81]	9.33	
LAFITE [82]	8.12	
Make-A-Scene [22]	7.55	
DALL-E [53]		17.89
LAFITE [82]		26.94
GLIDE [41]		12.24
DALL-E 2 [54]		10.39
Imagen (Our Work)		**7.27**

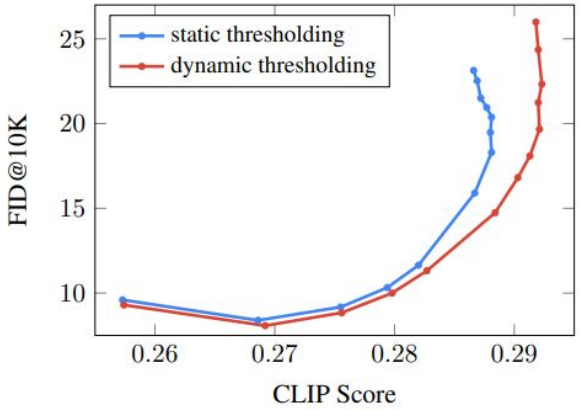
Model	Photorealism ↑	Alignment ↑
Original		
Original	50.0%	91.9 ± 0.42
Imagen	$39.5 \pm 0.75\%$	91.4 ± 0.44
No people		
Original	50.0%	92.2 ± 0.54
Imagen	$43.9 \pm 1.01\%$	92.1 ± 0.55



(a) Impact of encoder size.



(b) Impact of U-Net size.



(c) Impact of thresholding.

Figure 4: Summary of some of the critical findings of Imagen with pareto curves sweeping over different guidance values. See Appendix D for more details.

Выводы:

- Scaling text encoder size is extremely effective
- Scaling text encoder size is more important than U-Net size
- Dynamic thresholding is critical
- Human raters prefer T5-XXL over CLIP on DrawBench
- Noise conditioning augmentation is critical
- Text conditioning method is critical
- Efficient U-Net is critical