Evaluating Generative Models

Jun-Yan Zhu

16-726 Learning-based Image Synthesis

Pairwise Comparison

(A/B test, user preference)

Which image best reflects this prompt?

A charming, old-fashioned bakery storefront with a hand-painted sign reading "Grandma's Bakery", colorful awnings, and a display of fresh pastries, photorealistic exterior

Prefer (← Key)

Prefer (→ Key)

https://artificialanalysis.ai/text-to-image/arena

Which video best reflects this prompt?

Waves rise higher and crash forward, sending spray and foam cascading through the air.

Prefer (← Key)

Prefer (→ Key)

https://artificialanalysis.ai/text-to-video/arena

Compute ELO ranking

E: expected outcome
R: current score
c: constant (=400)

$$E_A = \frac{1}{1 + 10^{(R_B - R_A)/c}}$$

$$E_B = \frac{1}{1 + 10^{(R_A - R_B)/c}} = 1 - E_A$$

Compute ELO ranking

E: expected outcome
R: current score
c: constant (=400)

$$E_A = \frac{1}{1 + 10^{(R_B - R_A)/c}}$$

$$E_B = \frac{1}{1 + 10^{(R_A - R_B)/c}} = 1 - E_A$$

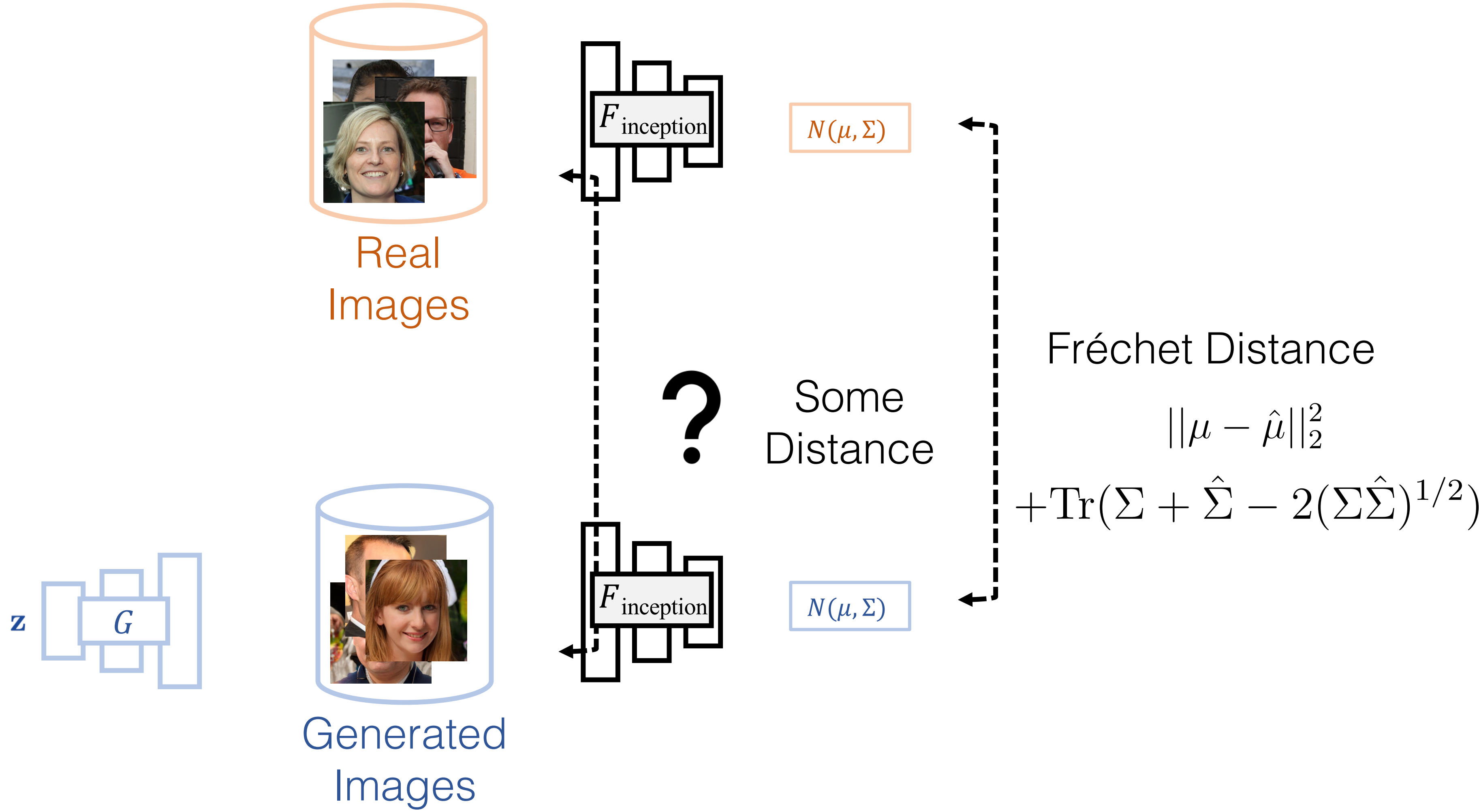
R': new score
R: current score
K: constant (=32)
S: actual outcome

$$R'_A = R_A + K \cdot (S_A - E_A)$$

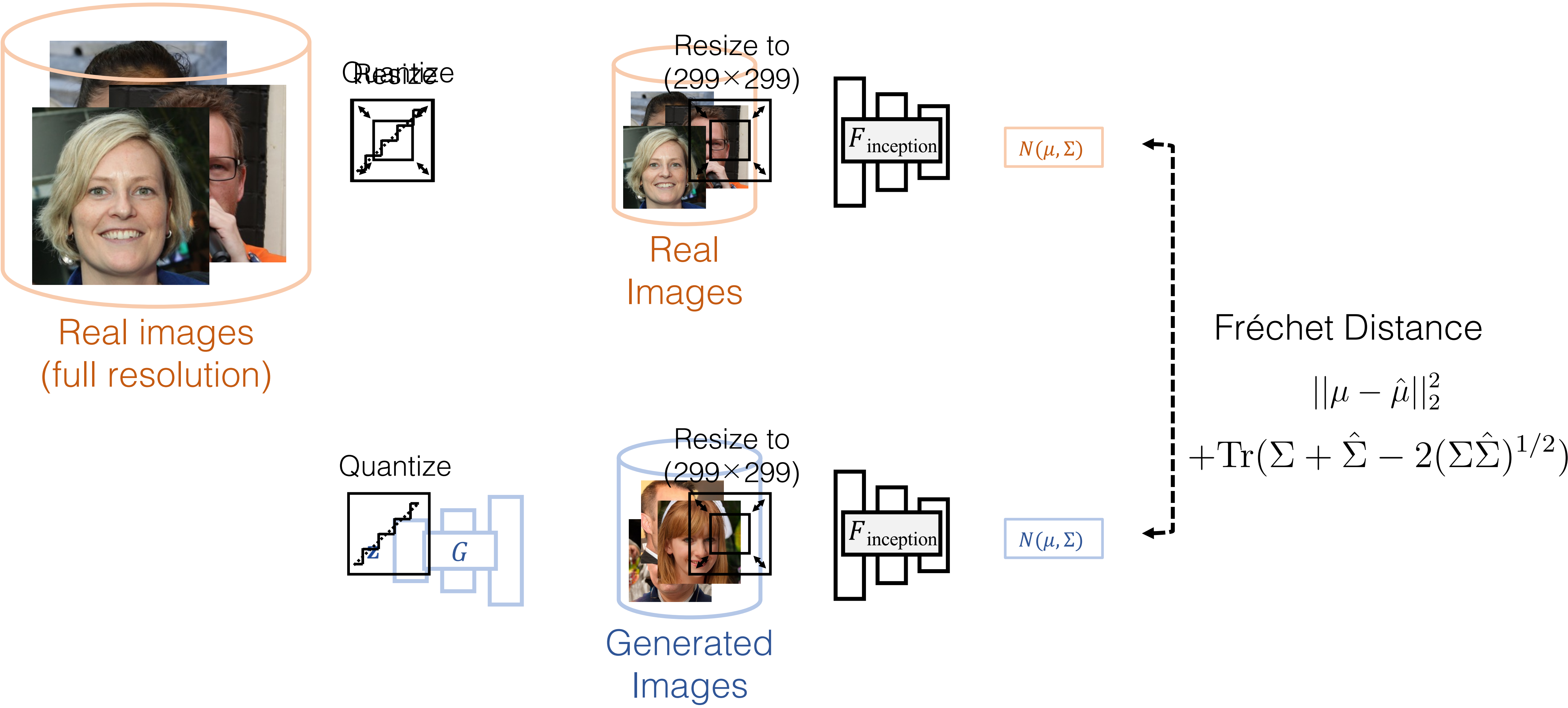
$$R'_B = R_B + K \cdot (S_B - E_B)$$

Automated Metrics

Fréchet Inception Distance (FID)



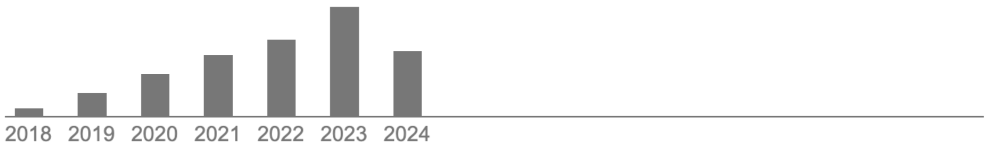
Fréchet Inception Distance (FID)



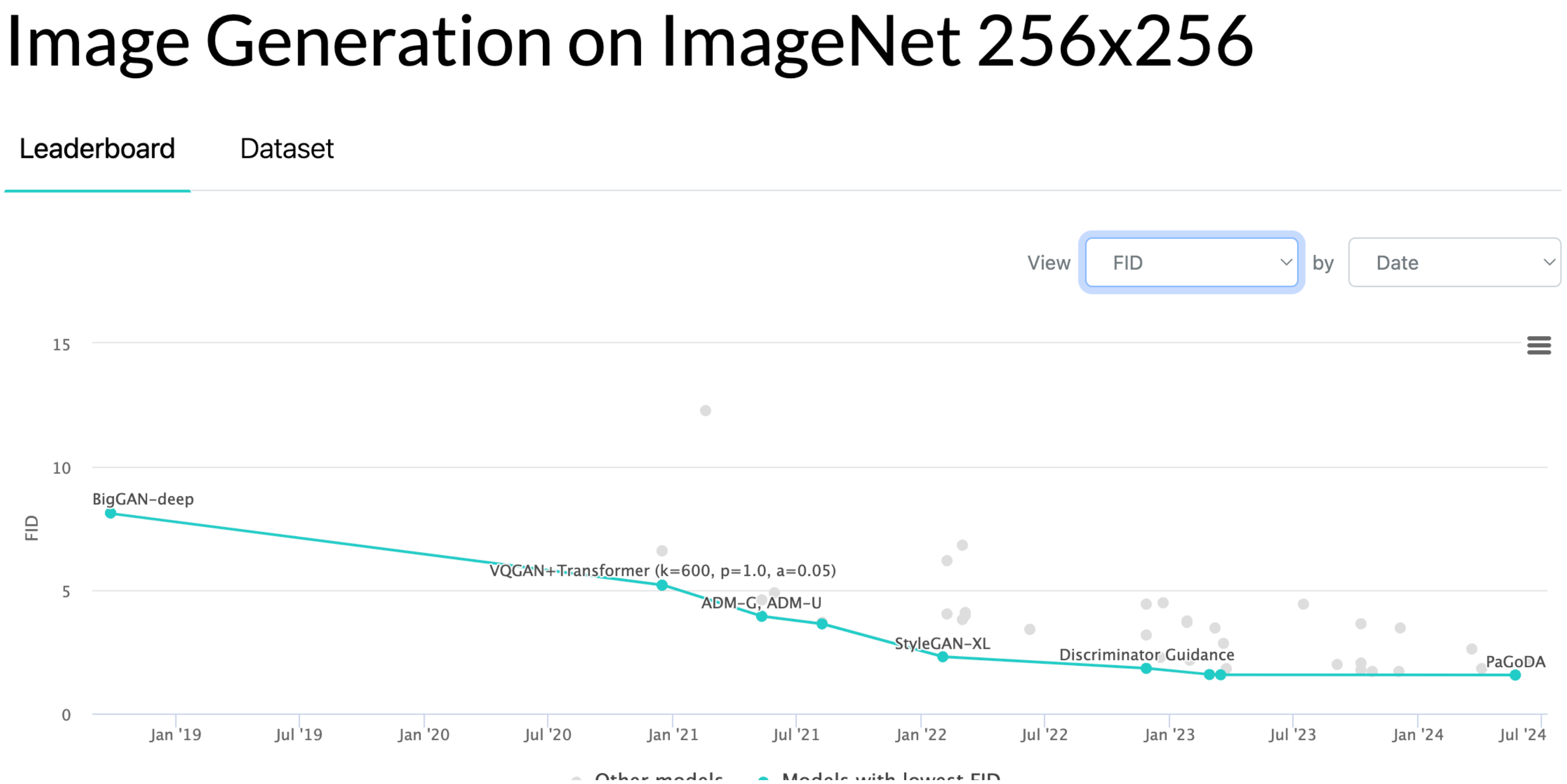
FID is being widely used

GANs trained by a two time-scale update rule converge to a local Nash equilibrium

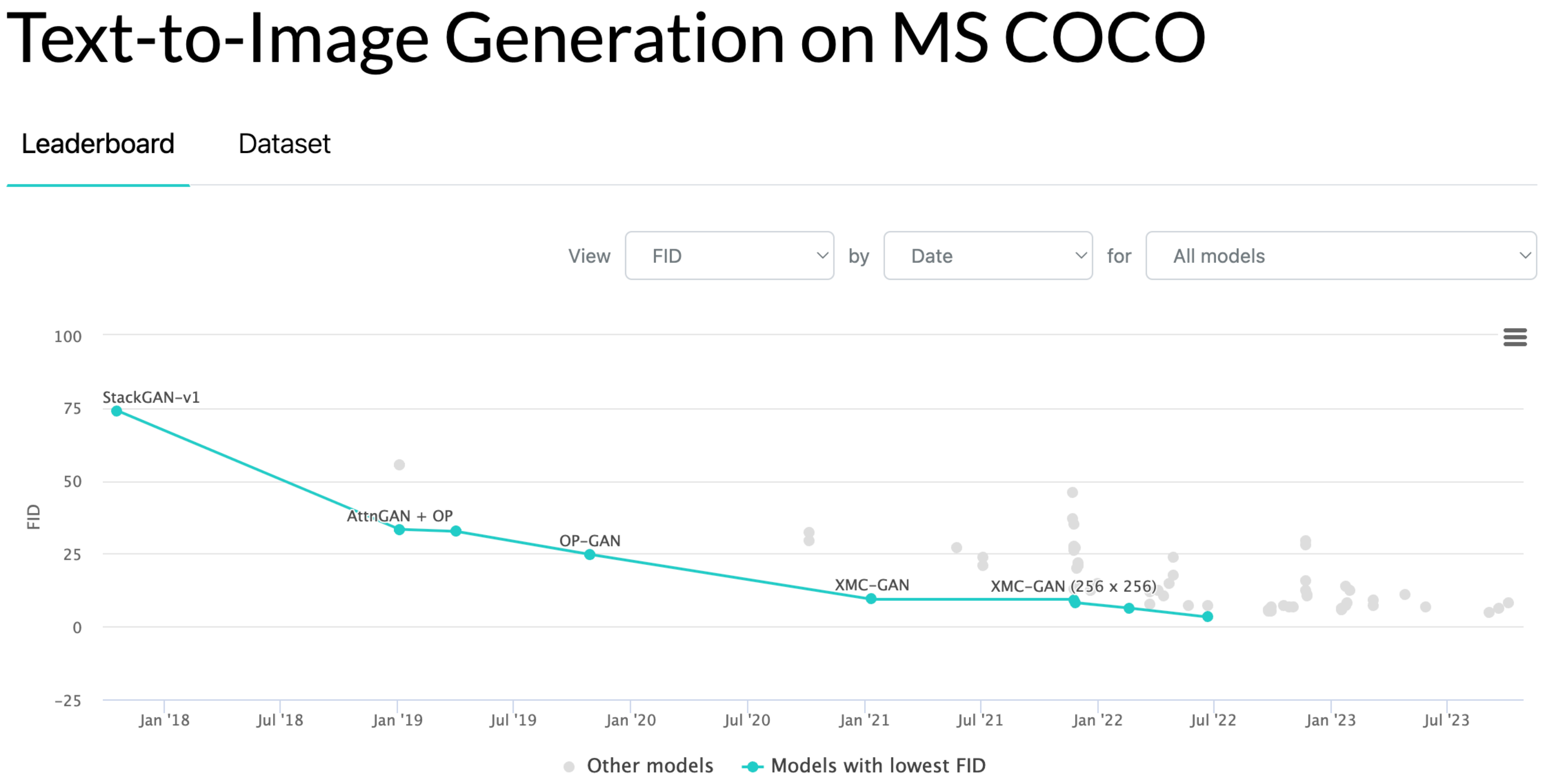
Authors	Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, Sepp Hochreiter
Publication date	2017
Conference	Advances in Neural Information Processing Systems
Pages	6626-6637
Description	Generative Adversarial Networks (GANs) excel at creating realistic images with complex models for which maximum likelihood is infeasible. However, the convergence of GAN training has still not been proved. We propose a two time-scale update rule (TTUR) for training GANs with stochastic gradient descent on arbitrary GAN loss functions. TTUR has an individual learning rate for both the discriminator and the generator. Using the theory of stochastic approximation, we prove that the TTUR converges under mild assumptions to a stationary local Nash equilibrium. The convergence carries over to the popular Adam optimization, for which we prove that it follows the dynamics of a heavy ball with friction and thus prefers flat minima in the objective landscape. For the evaluation of the performance of GANs at image generation, we introduce theFréchet Inception Distance"(FID) which captures the similarity of generated images to real ones better than the Inception Score. In experiments, TTUR improves learning for DCGANs and Improved Wasserstein GANs (WGAN-GP) outperforming conventional GAN training on CelebA, CIFAR-10, SVHN, LSUN Bedrooms, and the One Billion Word Benchmark.
Total citations	Cited by 12274



FID is being widely used



FID is being widely used



Why is FID so popular?

- Better than other metrics
  - vs. Inception Score (IS), density estimate with Parzen window
- Model agnostic
  - vs. Perceptual Path Length (PPL) and log likelihood
- Cheap and fast to compute
  - vs. Classification Accuracy Score
- Cover both diversity and realism
  - vs. precision and recall
- Easy to reproduce
  - vs. user studies

Known issues with FID

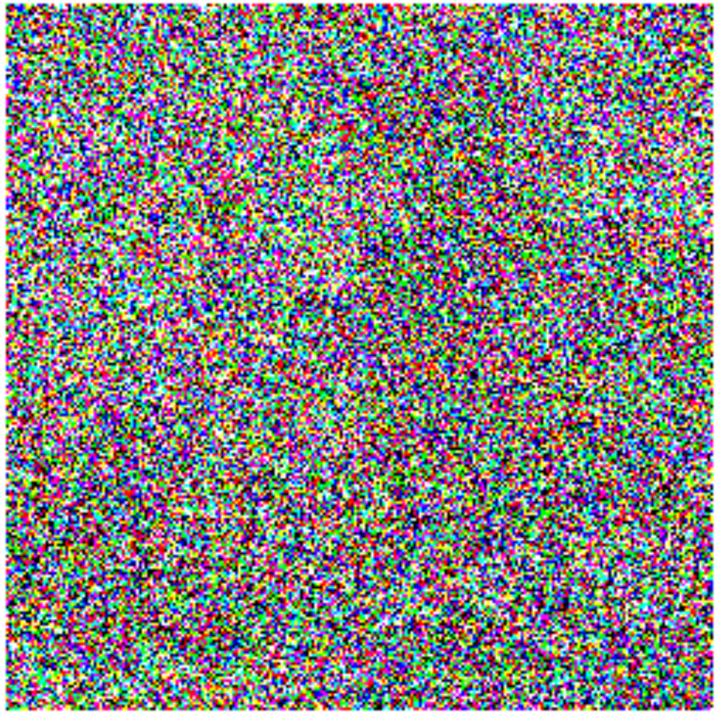
- The Gaussian Assumption.
- The large number of images required.
- The low-level image processing details.
- The choice of feature extractor.

Known issues with FID

- The Gaussian Assumption.

Our goal is to model complex distribution

- Two Gaussian Toy Example



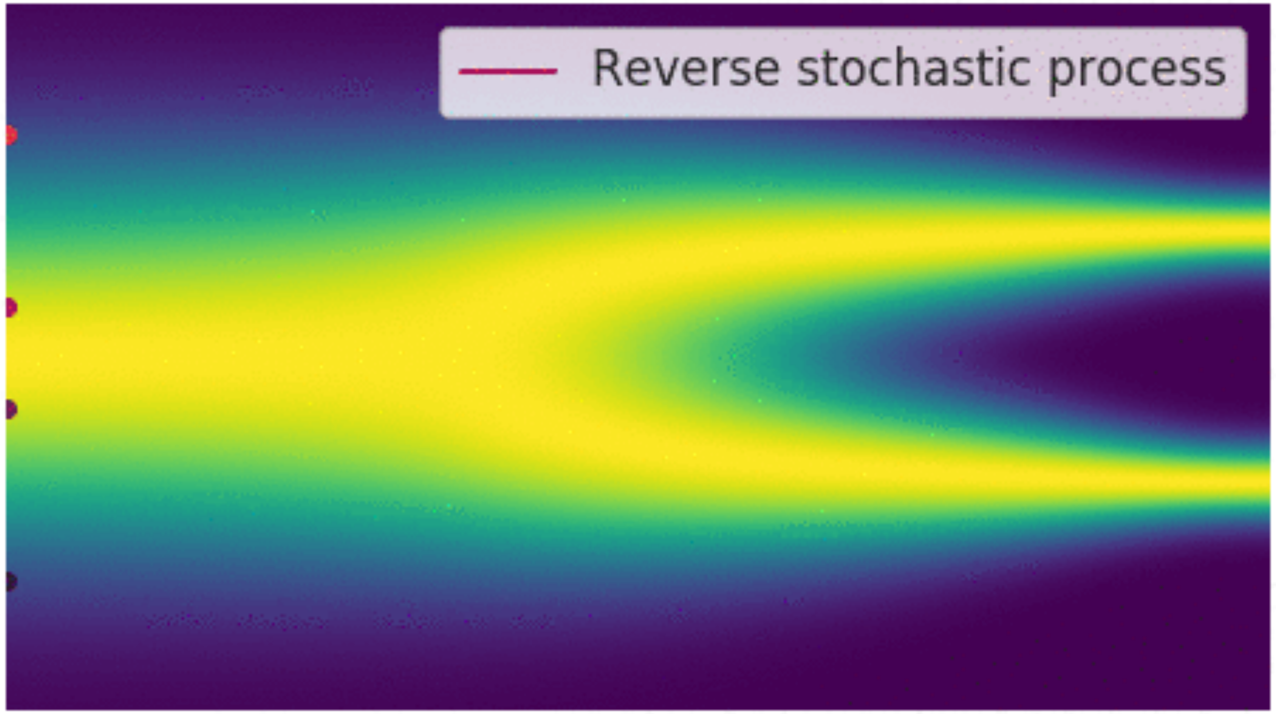


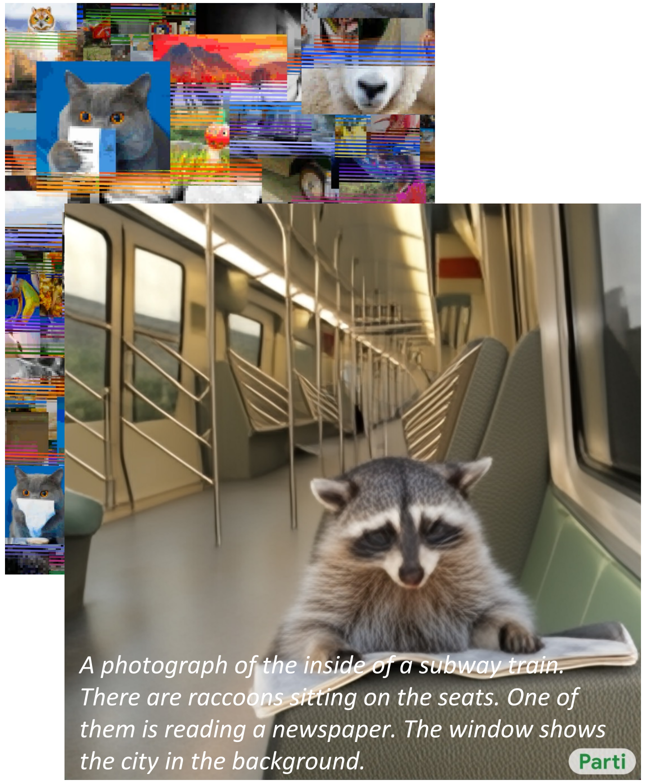
Image credit: Yang Song

Single-category dataset

Flickr-Faces-HQ Dataset (FFHQ) [Karras et al., 2018]

In the wild text-to-image synthesis

Diffusion models
(DALL-E 2, Imagen, SD)



Autoregressive models
(Image GPT, Parti)

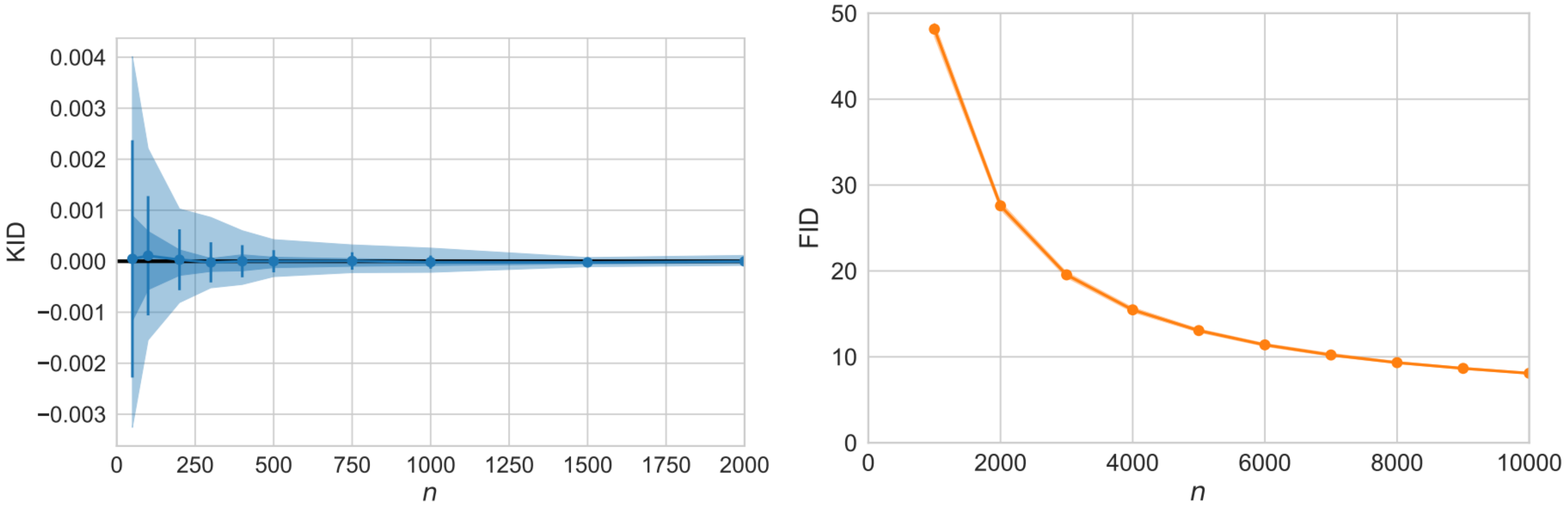
GANs, Masked GIT
(GigaGAN, MUSE)

Known issues with FID

- The Gaussian Assumption.
- The large number of images required.

FID vs. Kernel Inception Distance (KID)

- Computing covariance matrix requires lots of samples.
  - At least 2048 (for 2048d features), preferably 10K-50K.
  - Use KID if you have a small training/test set.



(a) KID estimates are unbiased, and standard deviations shrink quickly even for small n.

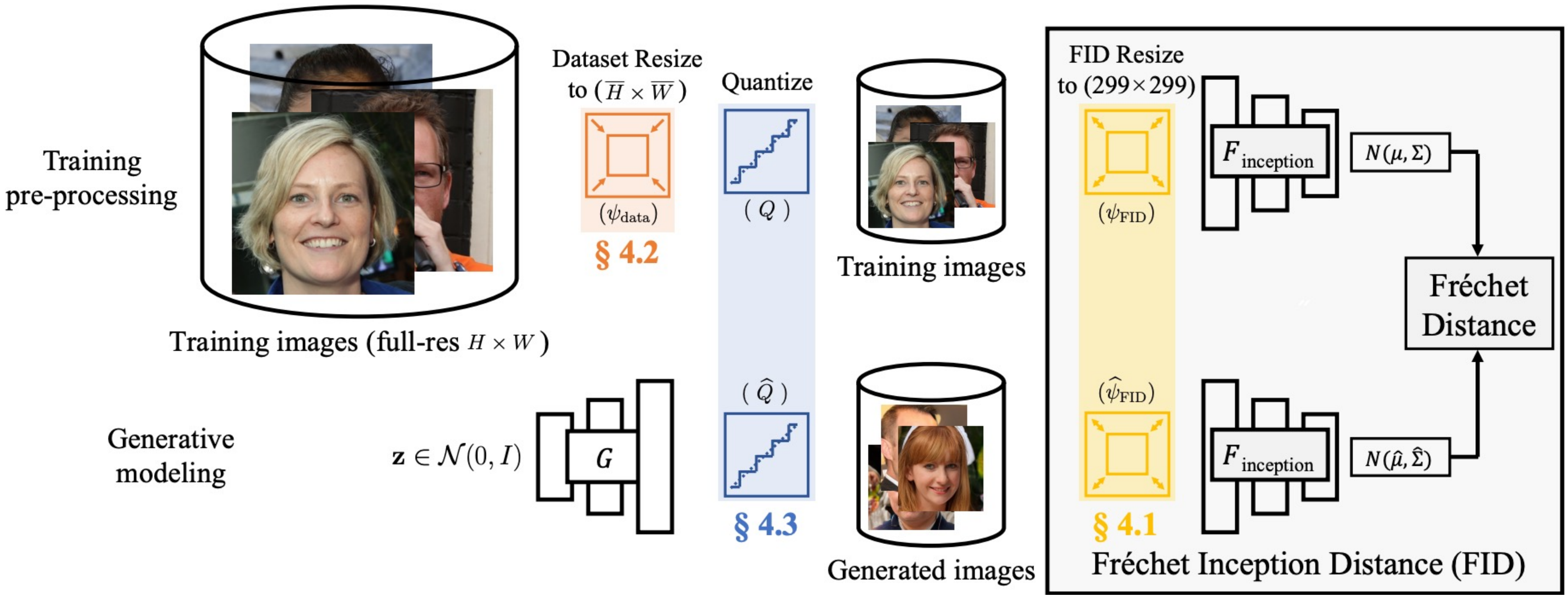
(b) FID estimates exhibit strong bias for n even up to 10 000. All standard deviations are less than 0.5.

[Binkowski et al., ICLR 2018], [Chong and Forsyth., CVPR 2020]

Known issues with FID

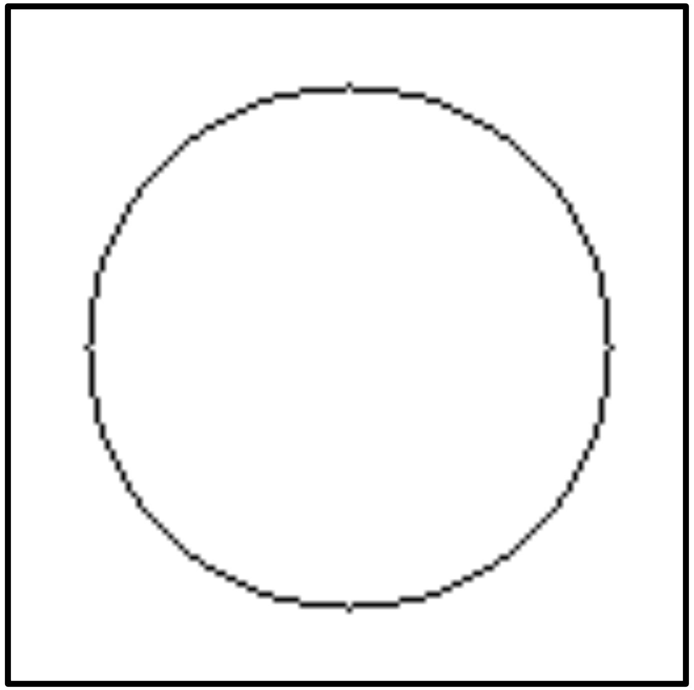
- The Gaussian Assumption.
- The large number of images required.
- The low-level image processing details.

Low-level image processing details



Downsampling a circle

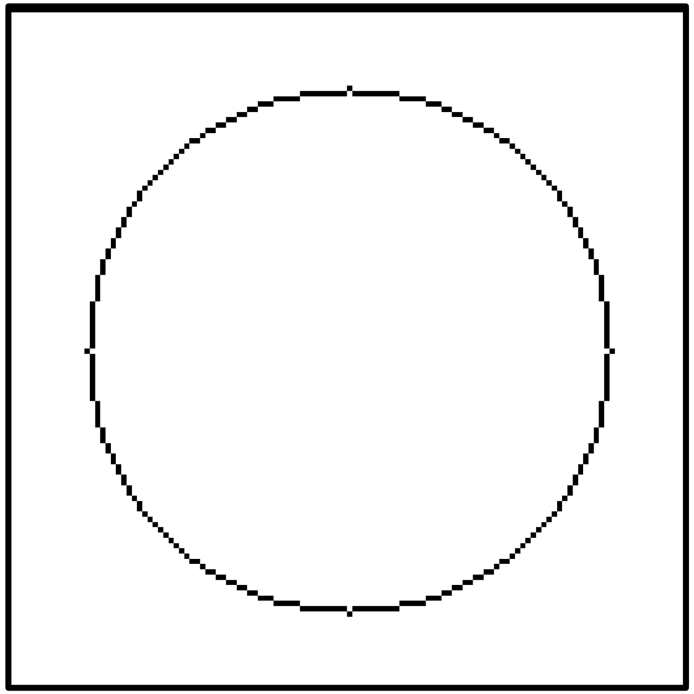
input image



Downsampling a circle

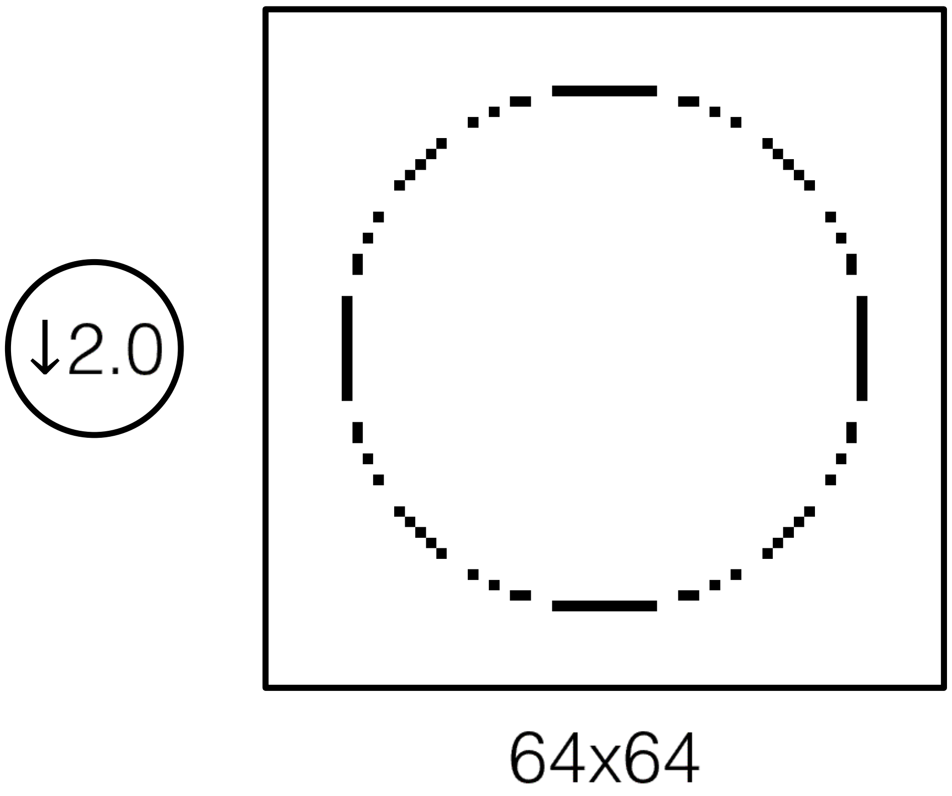
naïve nearest

↓1.0

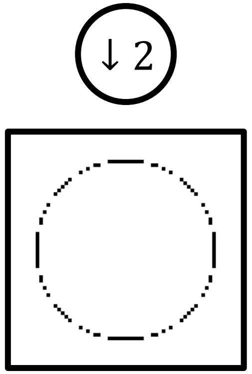


128x128

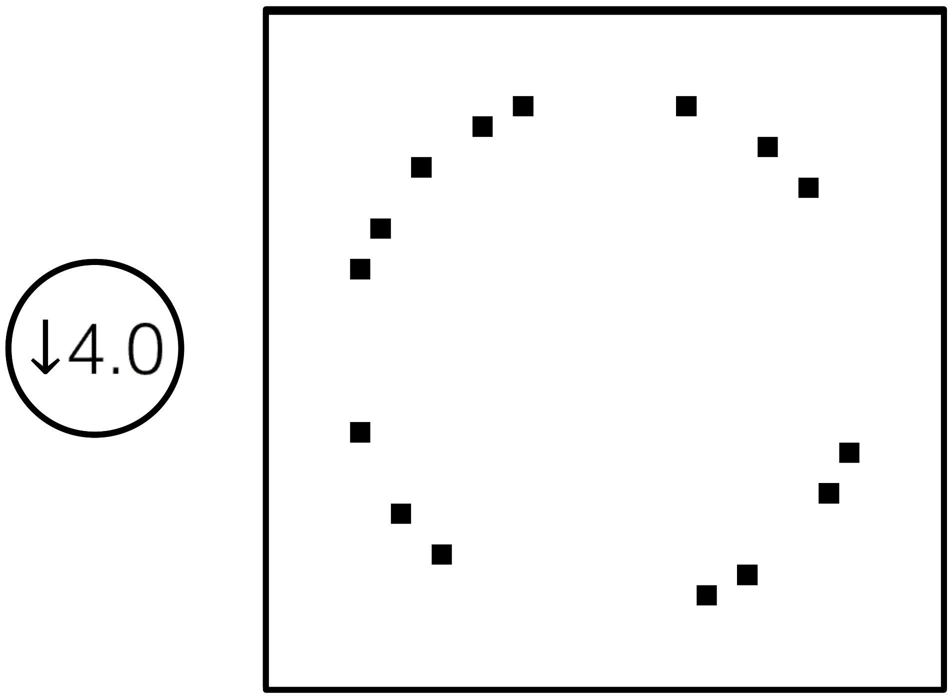
Downsampling a circle



naïve
nearest

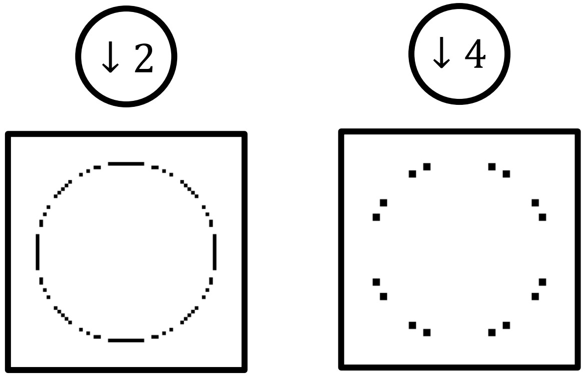


Downsampling a circle



32x32

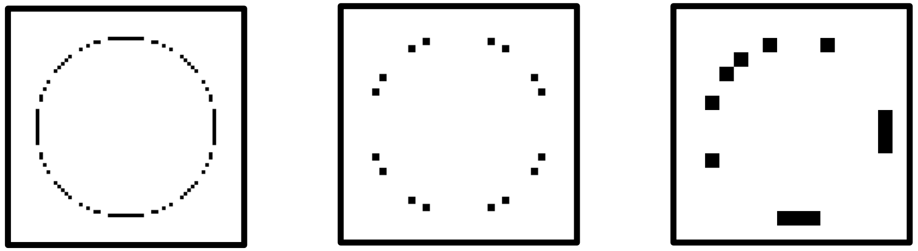
naïve
nearest



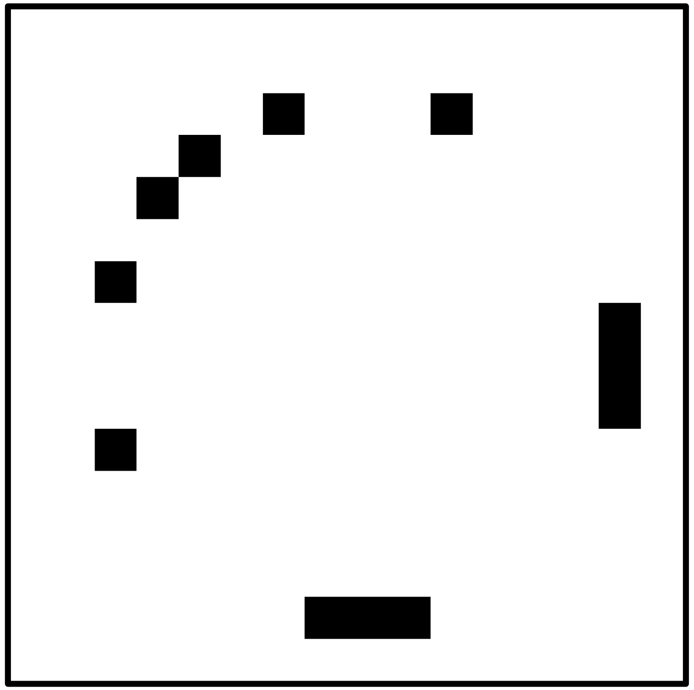
Downsampling a circle

↓ 2 ↓ 4 ↓ 8

naïve
nearest



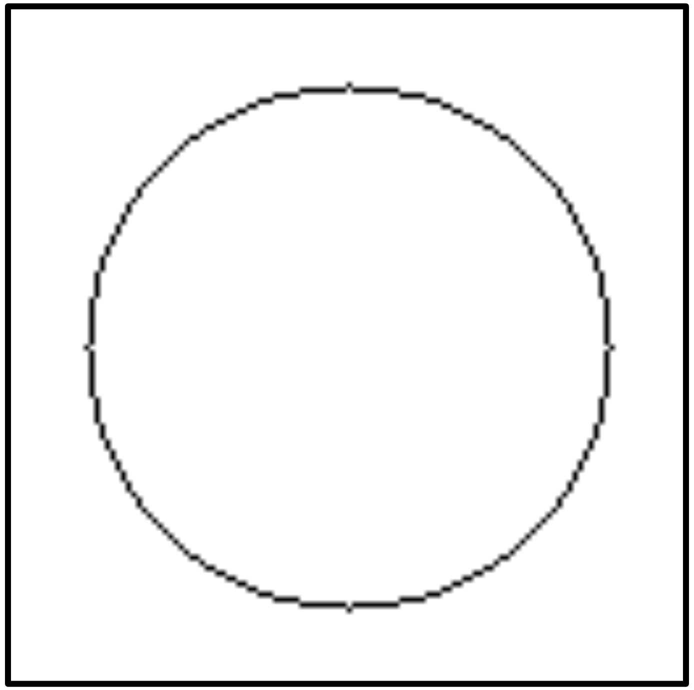
↓8.0



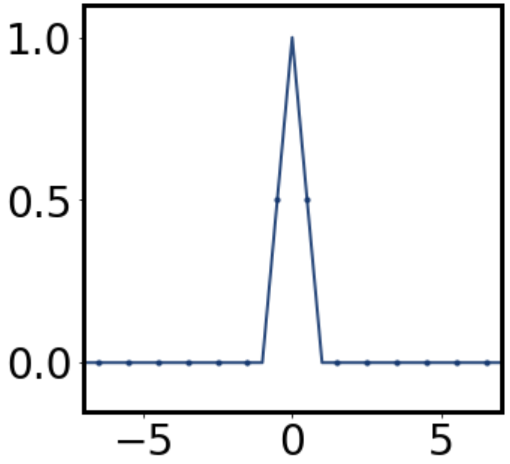
16x16

Downsampling a circle

prefiltering the image

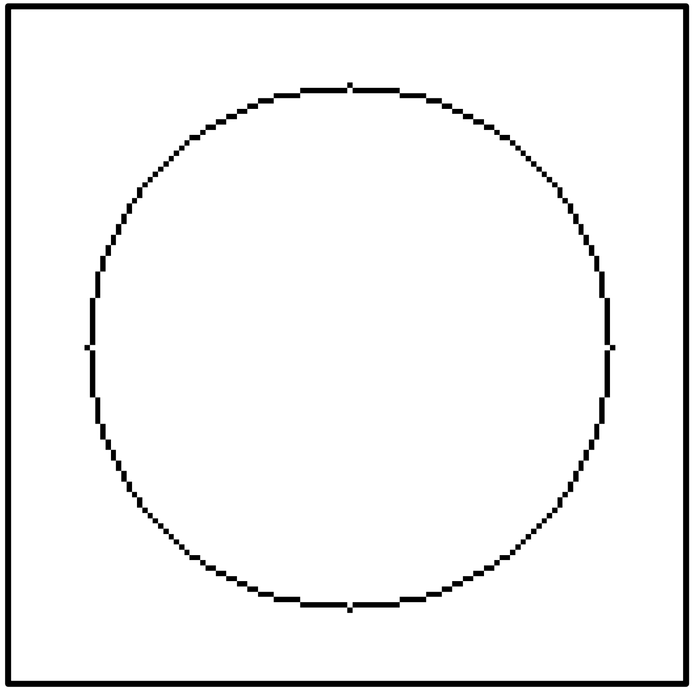


*



↓1.0

128x128

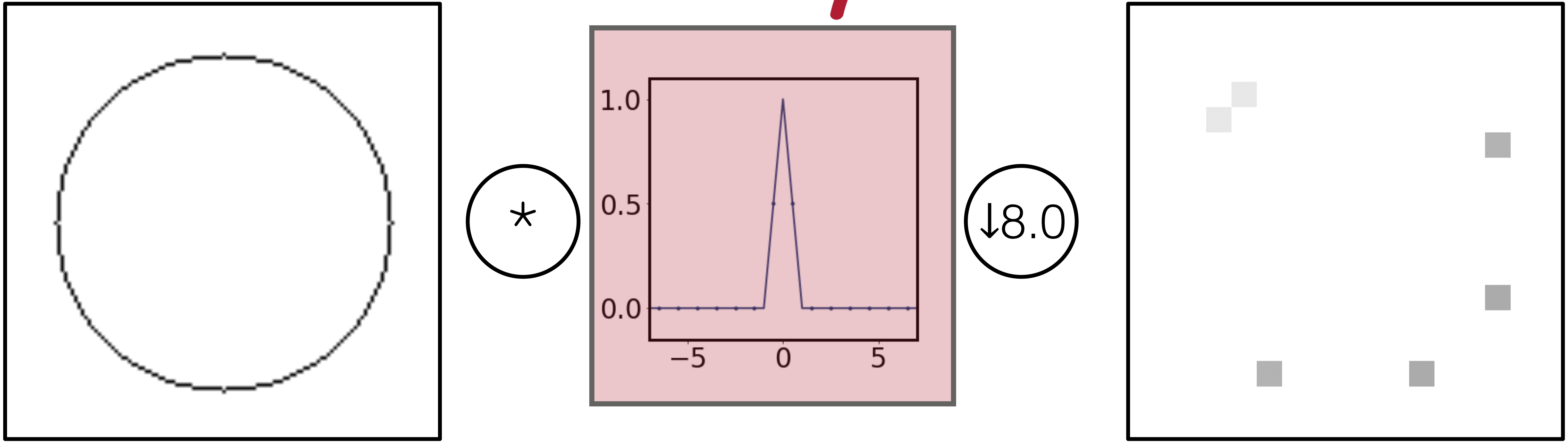


Downsampling a circle

prefiltering the image

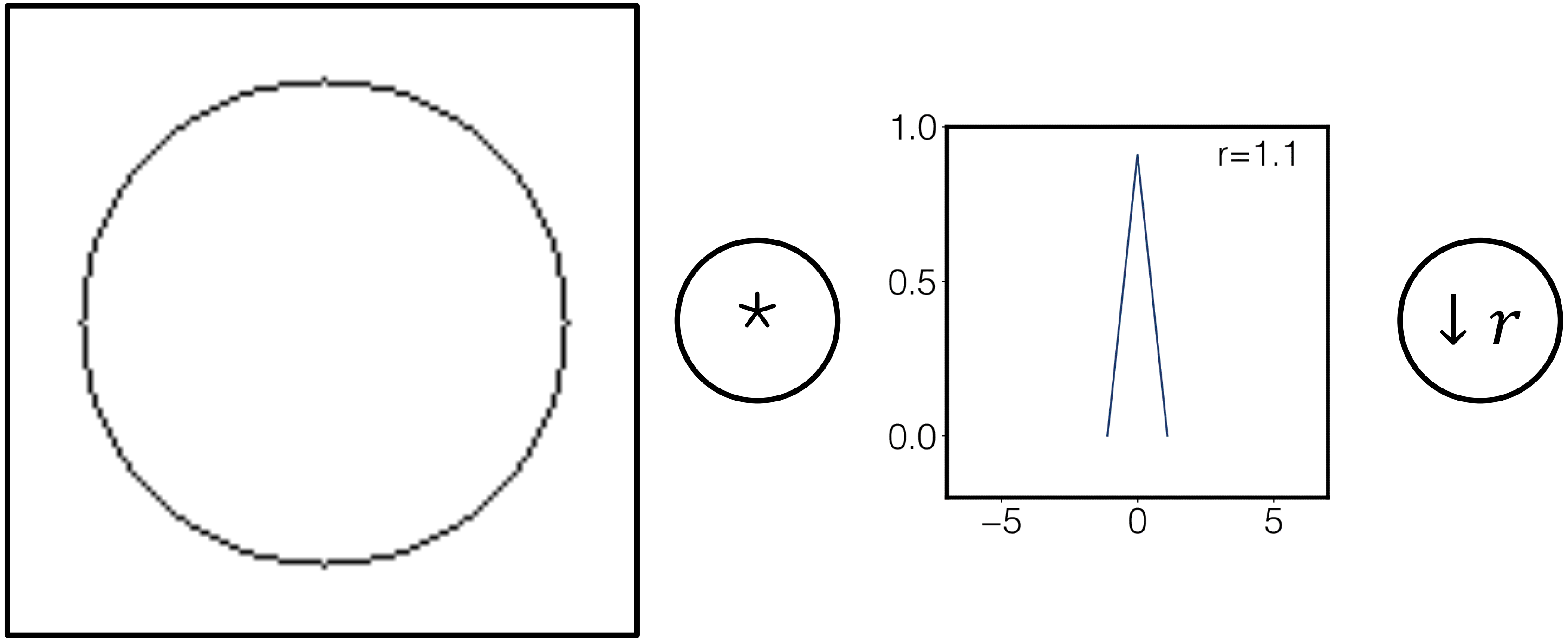
should not be fixed!

128x128



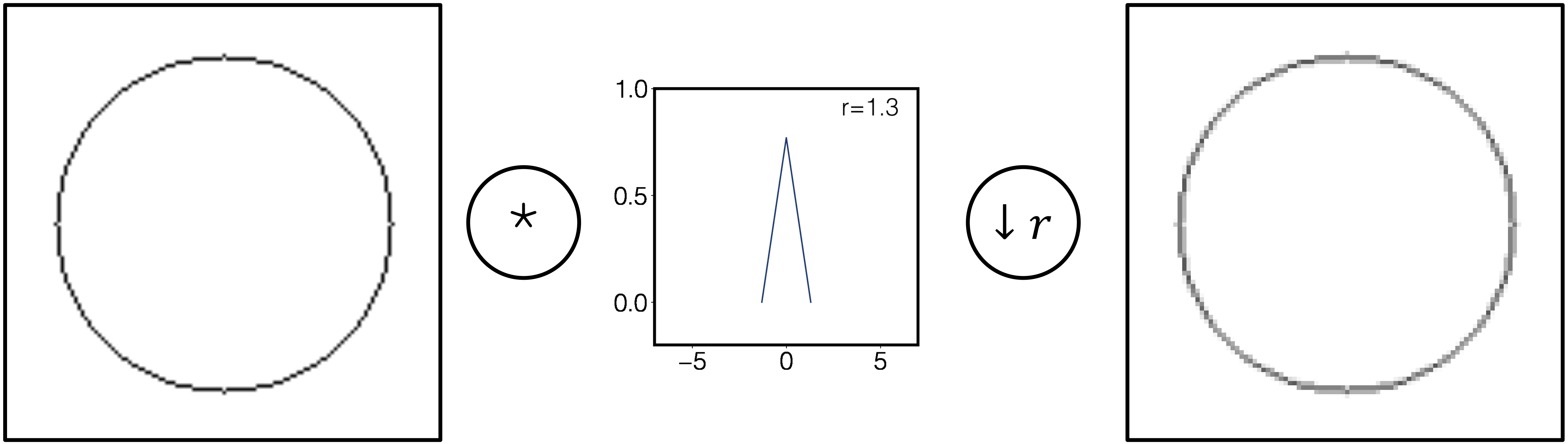
Downsampling a circle

prefiltering the image, adapting the width

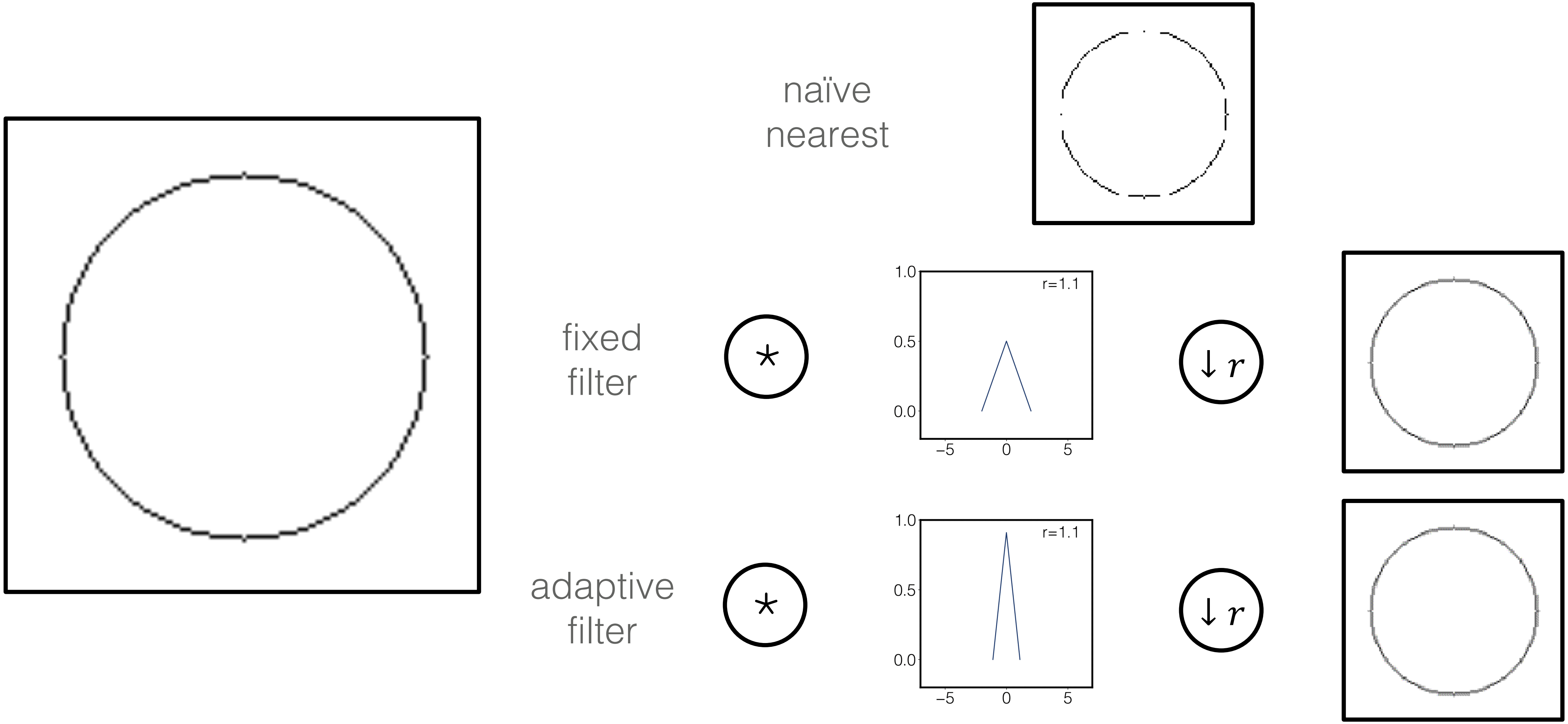


Downsampling a circle

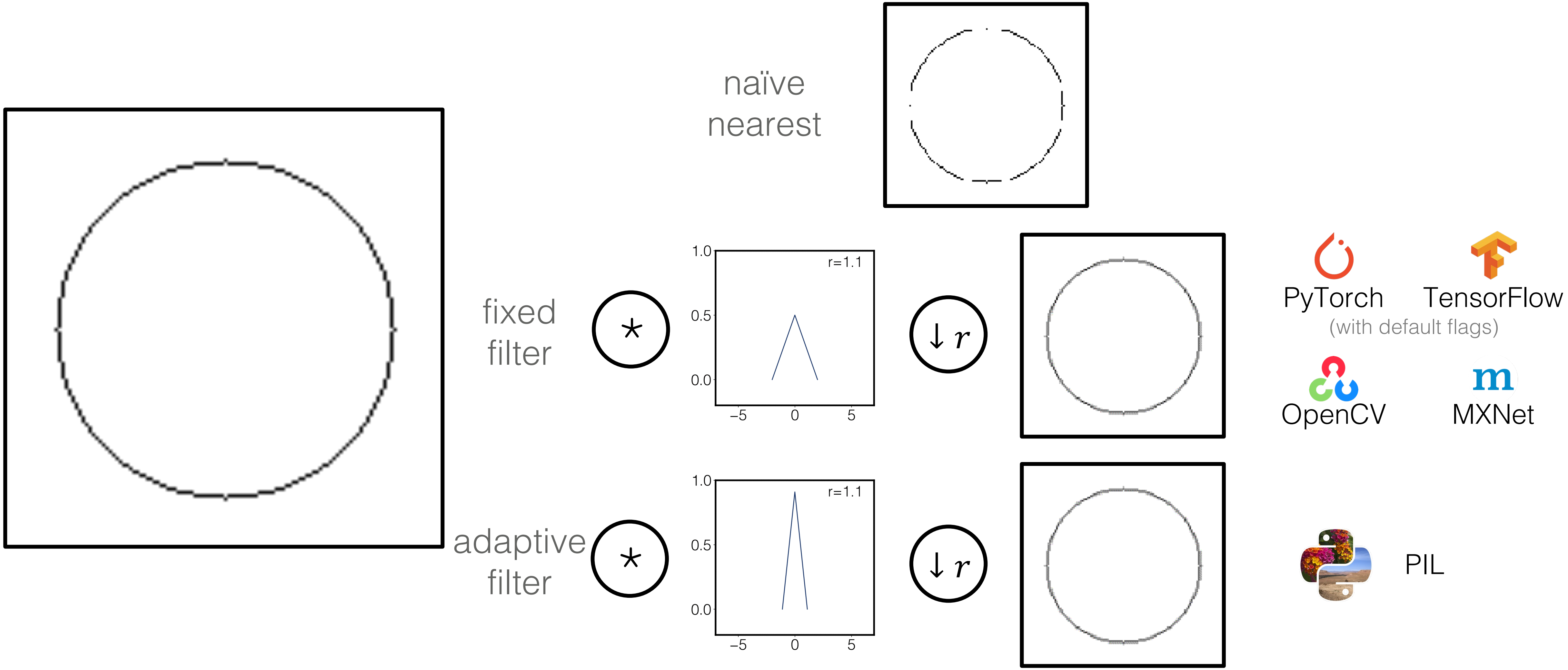
prefiltering the image, adapting the width



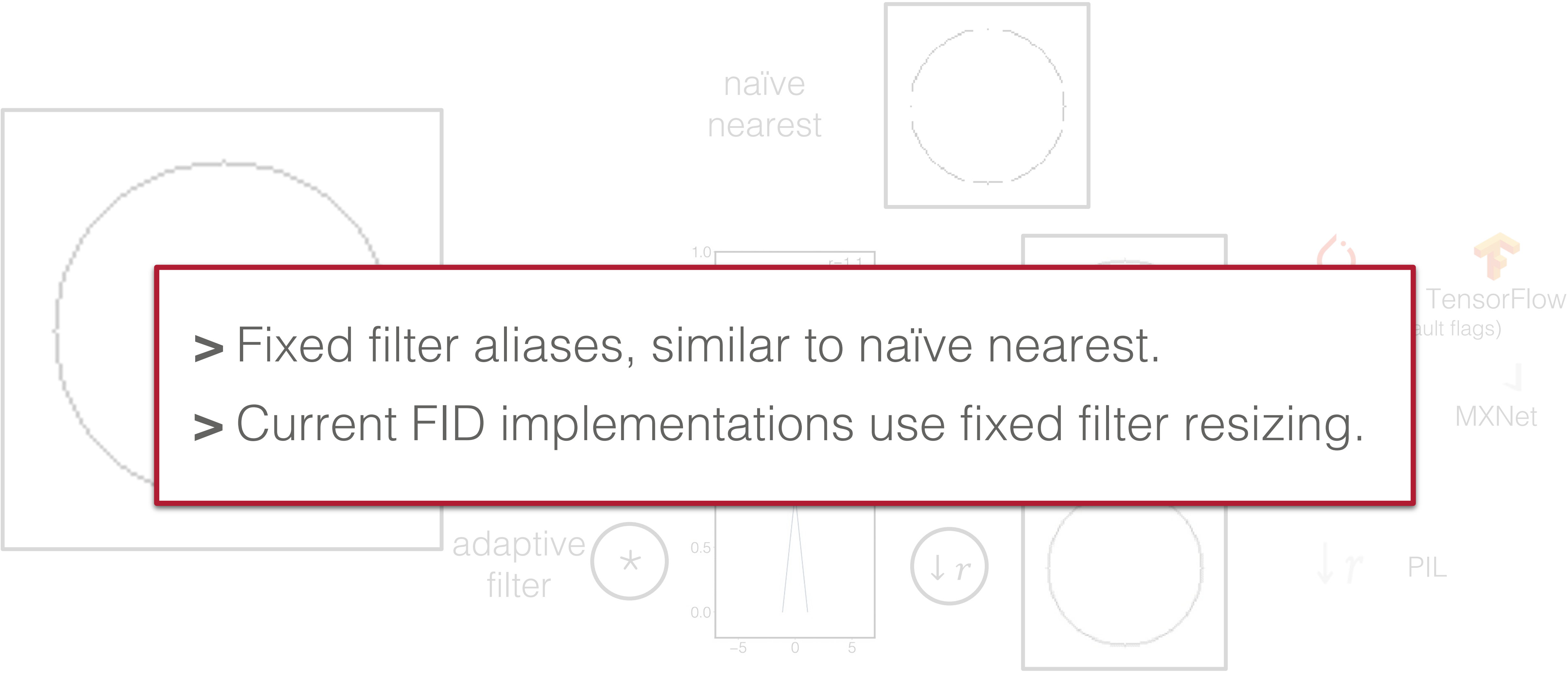
Downsampling a circle



Downsampling a circle



Downsampling a circle



> Fixed filter aliases, similar to naïve nearest.

> Current FID implementations use fixed filter resizing.

Downsampling an FFHQ image

1024

Downsampling an FFHQ image

1024

1019

adaptive-width
prefilter

Downsampling an FFHQ image

1024

1019

fixed-width
prefilter

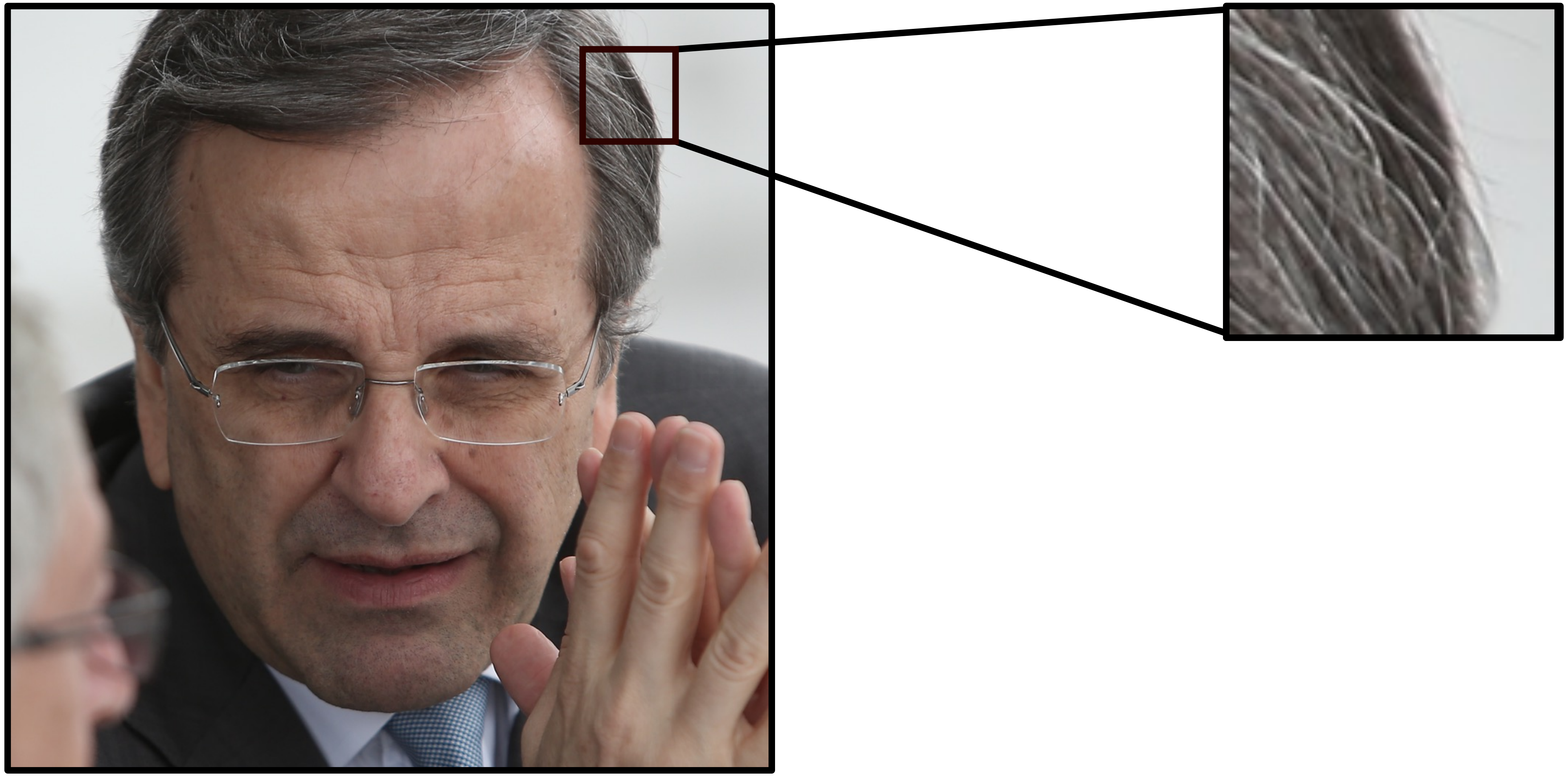
Downsampling an FFHQ image

1024

Downsampling an FFHQ image

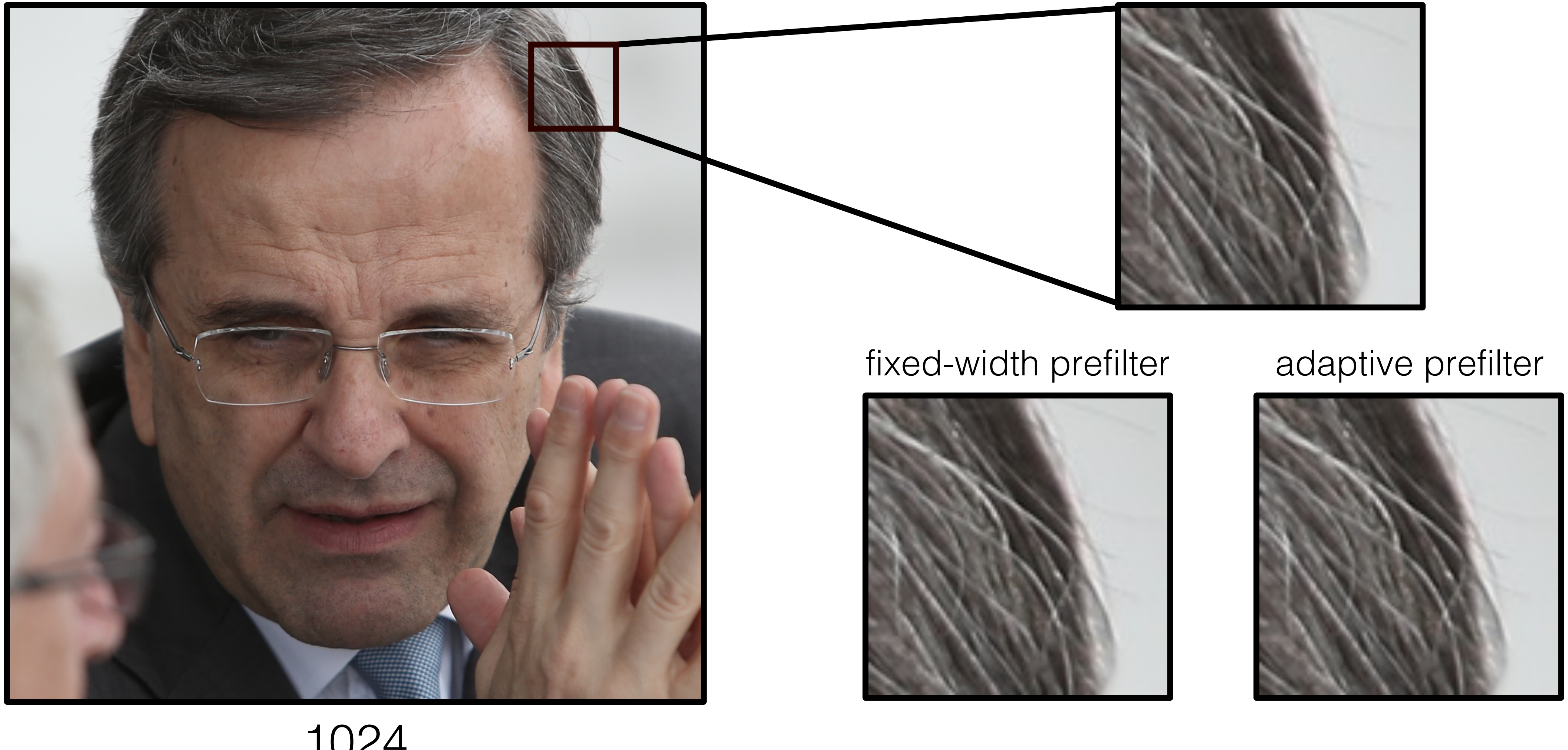
1024

Downsampling an FFHQ image

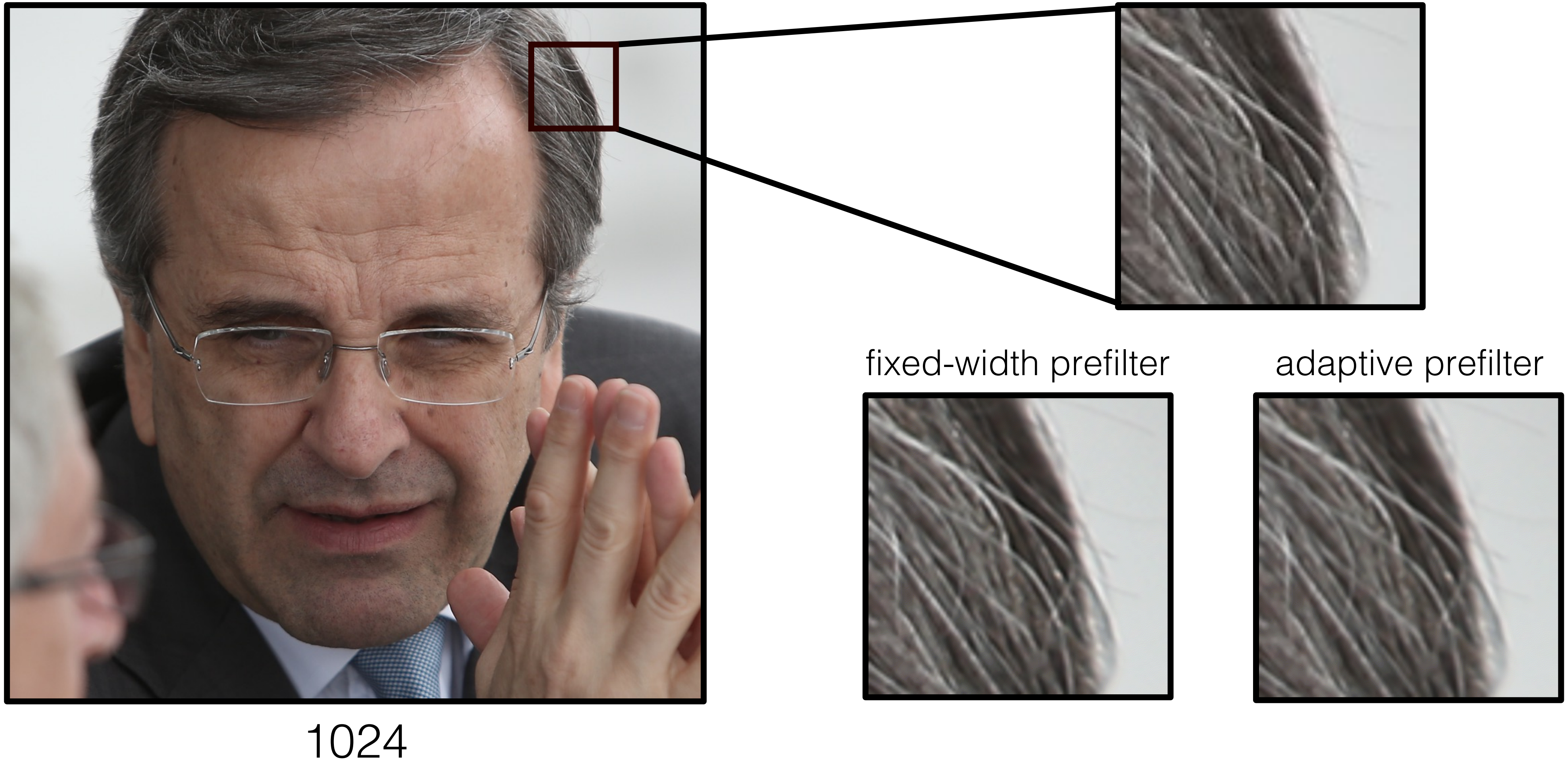


1024

Downsampling an FFHQ image



Downsampling an FFHQ image



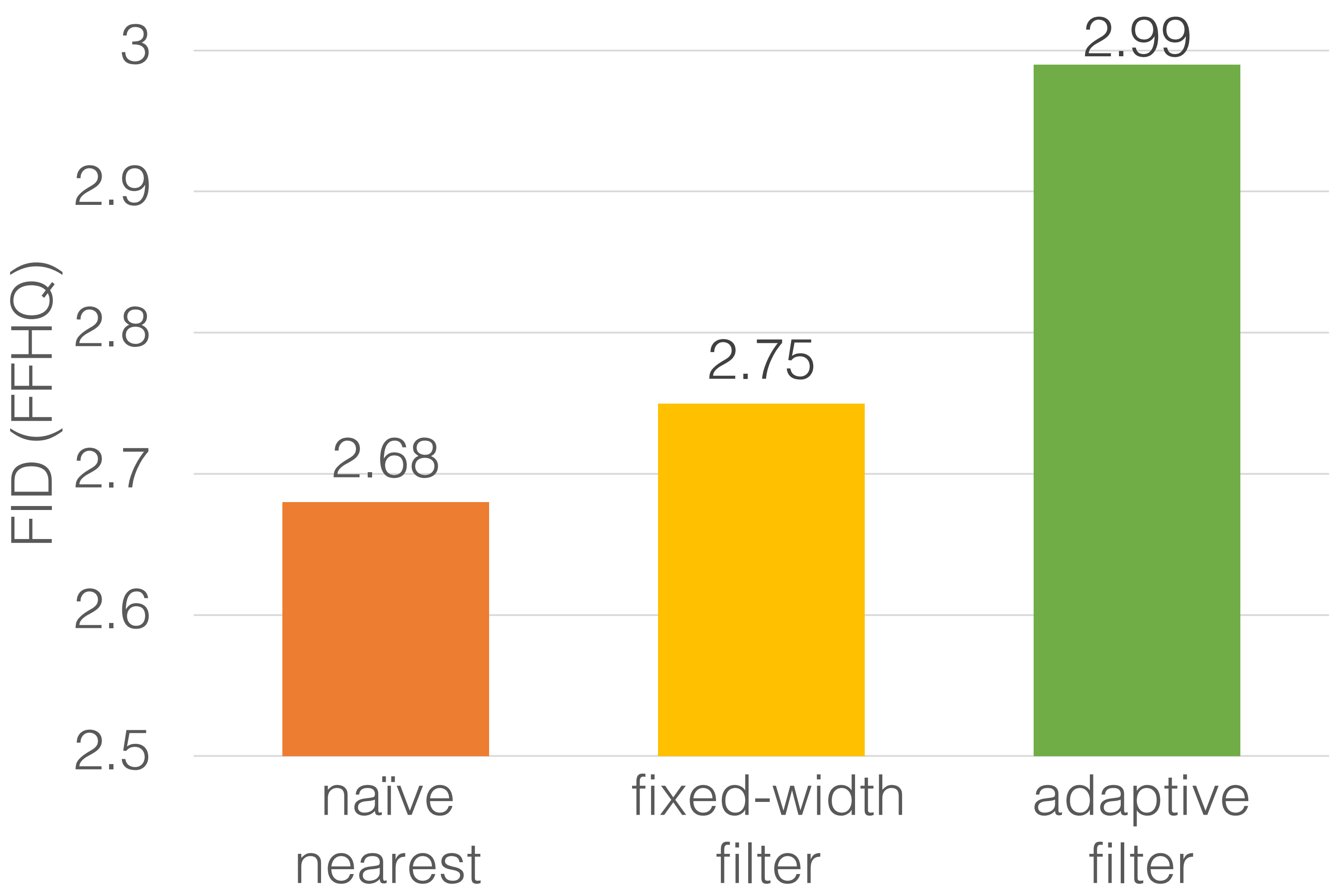
Changes in Inception Features

adaptive prefilter

fixed-width prefilter

Changes in Inception Features

Changes in FID



> Different resizing functions result in vastly different evaluation scores.

> aliased resizing deceptively causes improvements in the metric.

JPEG Compression

PNG
(uncompressed)

JPEG Compression

PNG
(uncompressed)

↓

JPEG
`quality = 99`

JPEG Compression

PNG
(uncompressed)

↓

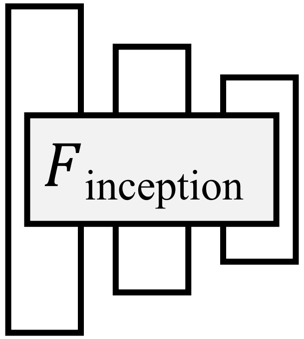
JPEG
`quality = 75`

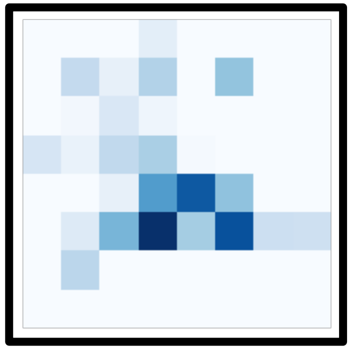
JPEG Compression

PNG
(uncompressed)

↓

JPEG
`quality = 75`





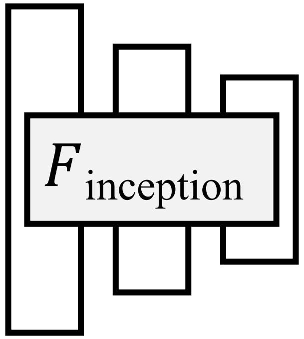
JPEG-75

JPEG Compression

PNG
(uncompressed)

↓

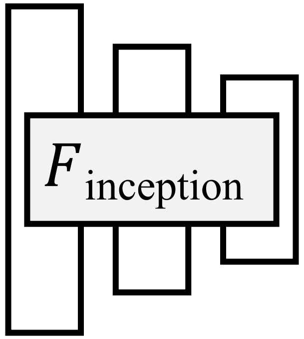
JPEG
`quality = 75`



JPEG 90 JPEG 75

JPEG Compression

PNG
(uncompressed)



PNG

JPEG Compression

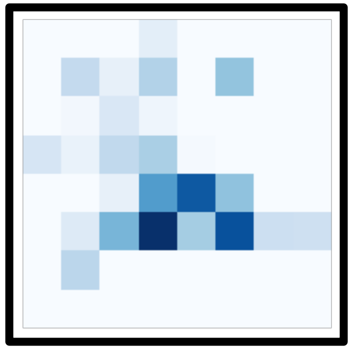
PNG

FID: 0

JPEG-90

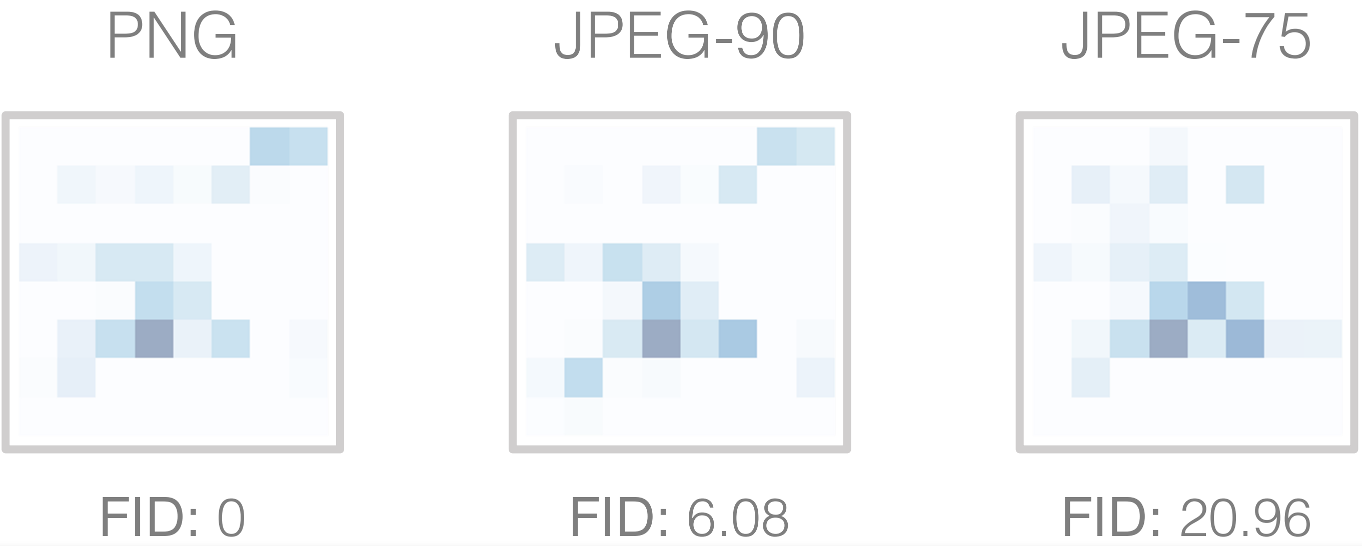
FID: 6.08

JPEG-75



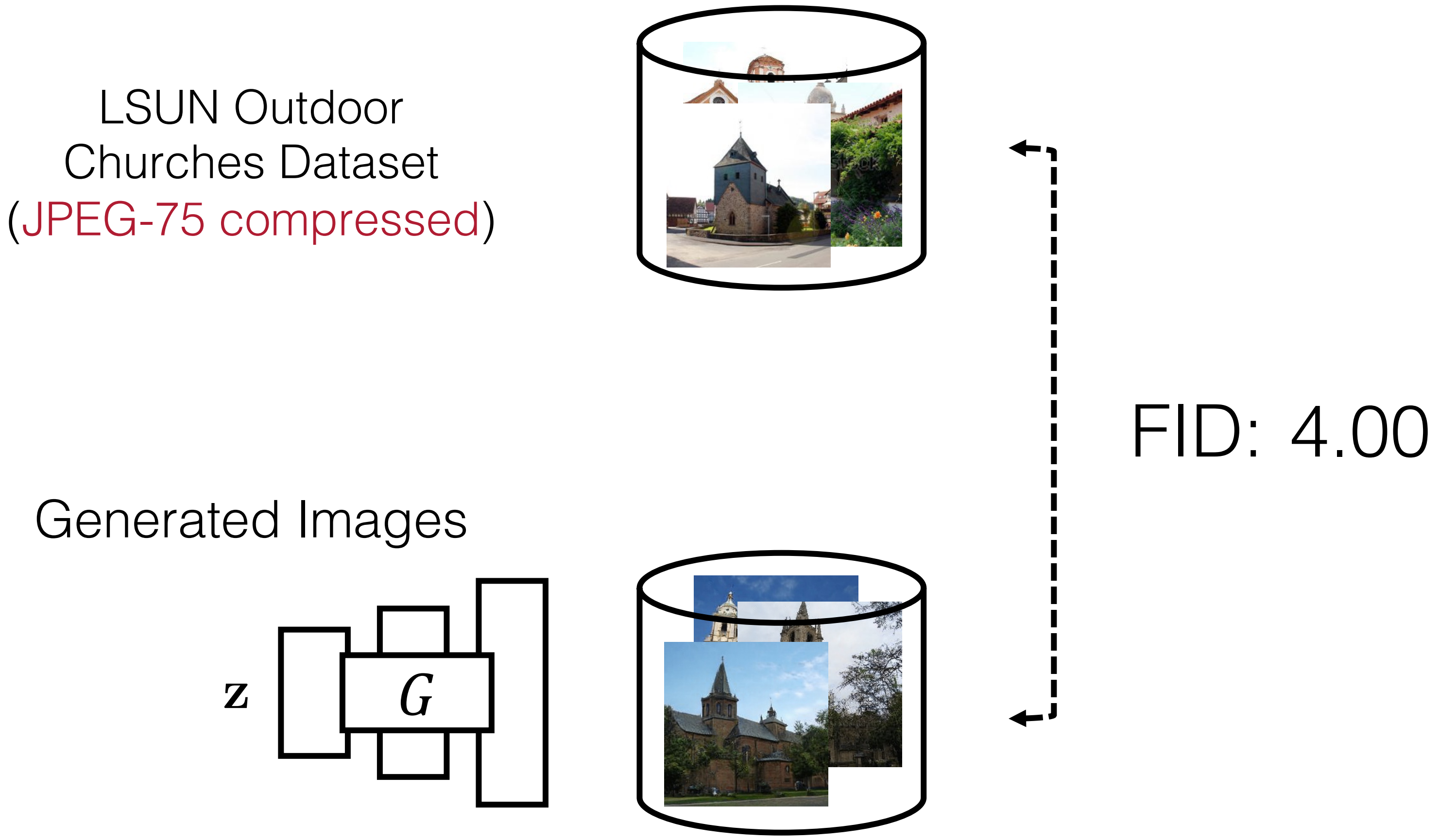
FID: 20.96

JPEG Compression



> Compressed images look near identical to original.
> Compression changes features and downstream metrics.

JPEG Compression - Dataset



JPEG Compression - Dataset

Discussion

> Evaluating generative models involves many steps.

> Image resizing and compression are crucial.

Recommendations

> Pre-filter the image adaptively when resizing.

> Avoid Lossy compression schemes.

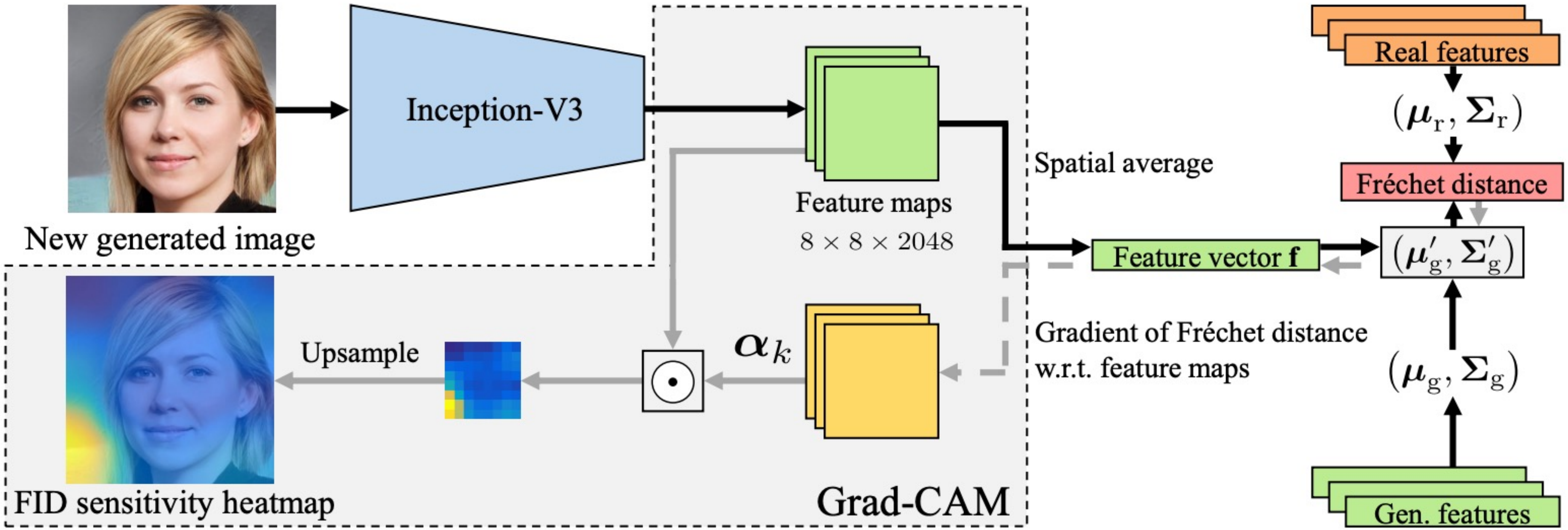
> Try out our library. (downloaded 20M+ times)

```
pip install clean-fid
```

Known issues with FID

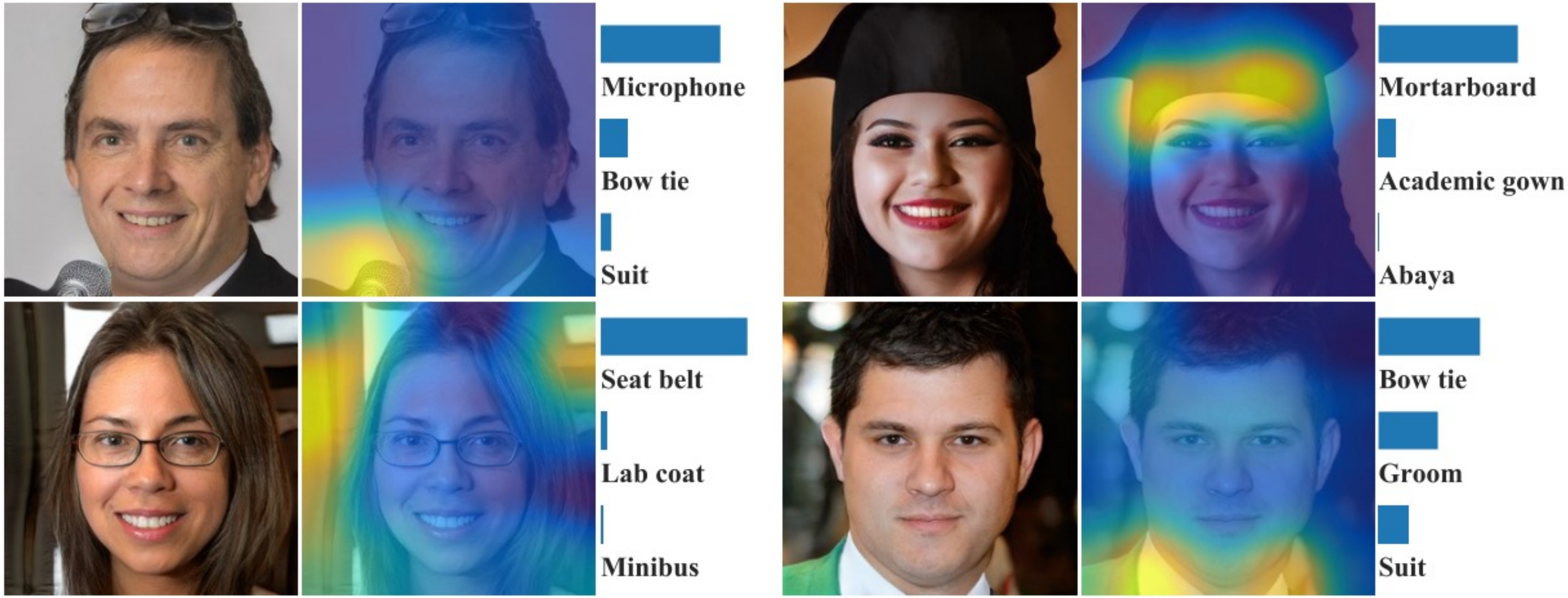
- The Gaussian Assumption.
- The large number of images required.
- The low-level image processing details.
- The choice of feature extractor.

The choice of feature extractor



The Role of ImageNet Classes in Fréchet Inception Distance.
Tuomas Kynkäänniemi, Tero Karras, Miika Aittala, Timo Aila, Jaakko Lehtinen. ICLR 2023

The choice of feature extractor



The Role of ImageNet Classes in Fréchet Inception Distance.
Tuomas Kynkäänniemi, Tero Karras, Miika Aittala, Timo Aila, Jaakko Lehtinen. ICLR 2023

Replace Inception Network with CLIP

FID = 5.28, Recall = 0.45, $\mathrm{FID}_{\mathrm{CLIP}}$ = 4.67 | FID = 5.30, Recall = 0.46, $\mathrm{FID}_{\mathrm{CLIP}}$ = 2.76

(a) Projected FastGAN | (b) StyleGAN2

FID-CLIP

The Role of ImageNet Classes in Fréchet Inception Distance.
Tuomas Kynkäänniemi, Tero Karras, Miika Aittala, Timo Aila, Jaakko Lehtinen. ICLR 2023

What about Video Generation?

Fréchet Video Distance (FVD)

- f_i: N features extracted from a pretrained I3D network
- (μ_r, Σ_r): mean and covariance of the features from real videos
- (μ_g, Σ_g): mean and covariance of the features from generated videos

The performance of the video generator is then measured as:

$$\mathbf{FVD} = \|\mu_r - \mu_g\|_2^2 + \mathbf{Tr}\left(\mathbf{\Sigma}_r + \mathbf{\Sigma}_g - 2\left(\mathbf{\Sigma}_r\mathbf{\Sigma}_g\right)^{\frac{1}{2}}\right)$$

FVD [Unterthiner, et al. arXiv, 2022]

Does FVD Align with Human Perception?

There are several issues with FVD metric on its own. First, it does not capture motion collapse, which can be ob-

datasets. We conjecture the unusual decrease of the FVD w.r.t. the duration of DIGAN and TATS-hierarchical on Sky Time-lapse can be explained by that the I3D model [7] used to calculate FVD is trained on Kinetics-400 dataset, and the sky videos can be outliers of the training data and lead to weak activation in the logit layers and therefore such unusual behaviors. We further perform qualitative

for our model. Another issue with FVD calculation is that it is biased towards image quality. If one trains a good image generator, i.e. a model which is not able to generate any videos at all, then FVD will still be good for it even despite the fact that it would have degenerate motion.

The commonly used Fréchet video distance (FVD) [57] attempts to measure similarity between real and generated video distributions. We find that FVD is sensitive to the realism of individual frames and motion over short segments, but that it does not capture long-term realism. For example, FVD is

StyleGAN-v [Skorokhodov, et al. CVPR, 2022]
TATS [Ge, et al. ECCV, 2022]
LongVideoGAN [Brooks, et al. NeurIPS, 2022]
VideoPhy [Bansal, et al. arXiv, 2024]

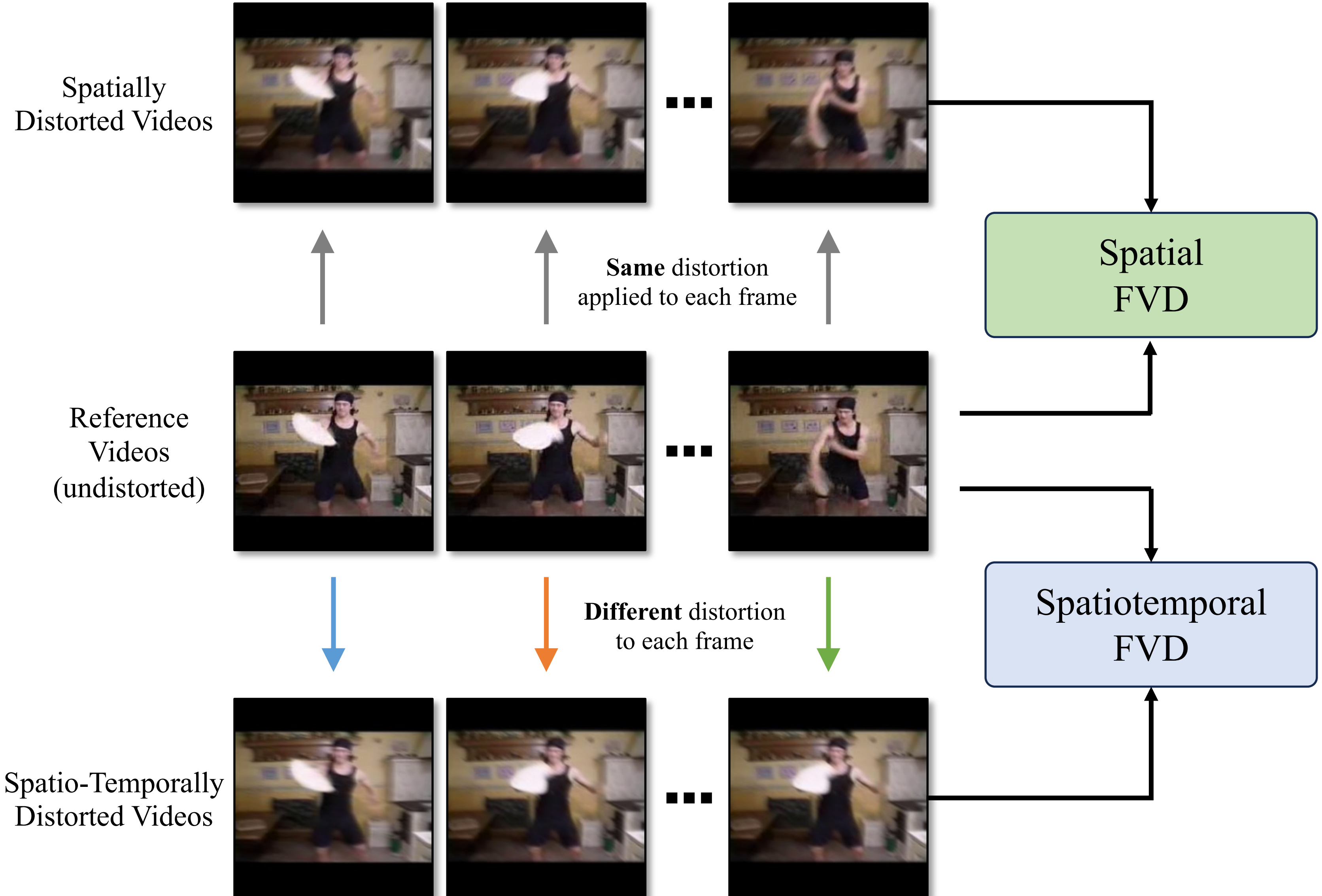
FVD is biased towards per-frame quality than temporal consistency

Reference Videos

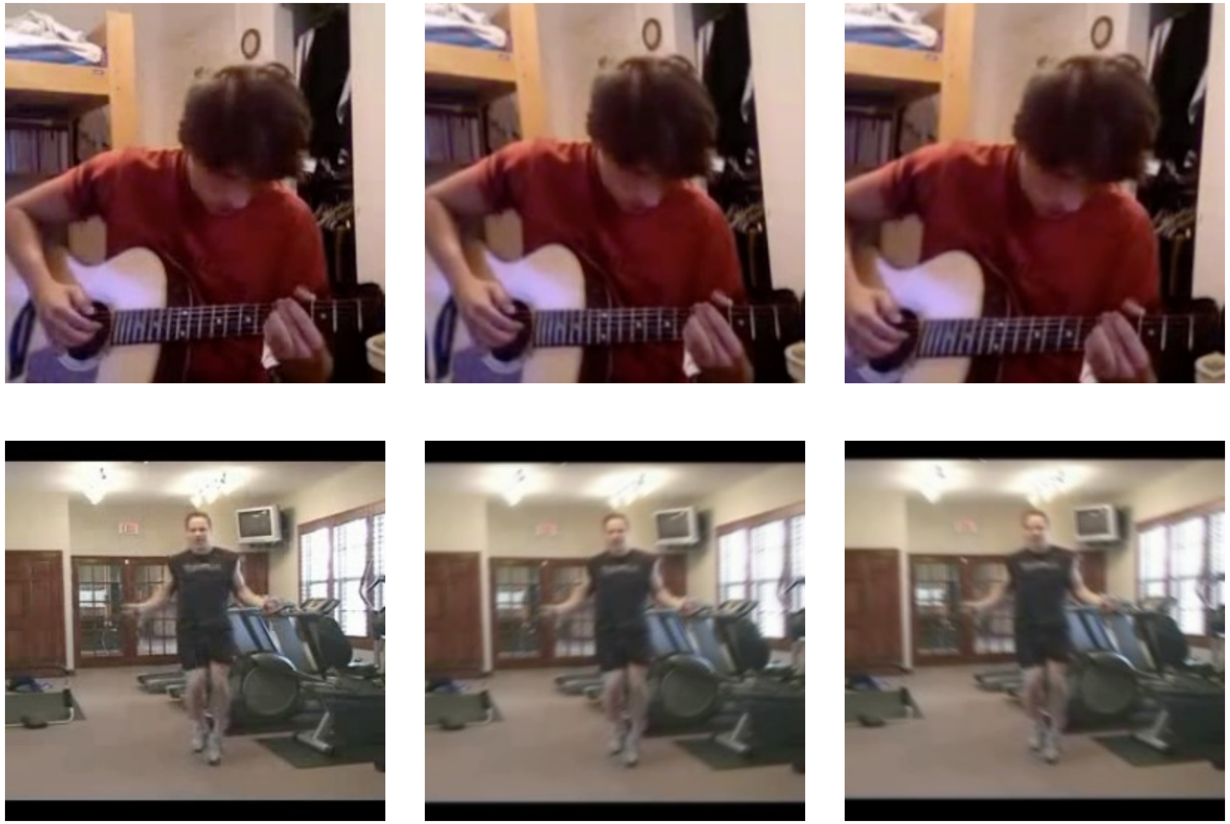
Medium Spatial Corruption
No Temporal Corruption
FVD=317.10

Small Spatial Corruption
Severe Temporal Corruption
FVD=310.52

Quantifying Temporal Sensitivity



Quantifying Temporal Sensitivity



Clean Videos | Spatial Corruption | Spatiotemporal Corruption

Temporal corruption doesn't affect frame quality

Metric	Distortion	UCF-101	Sky Time-lapse	FaceForencis	Taichi-HD	SSv2	Kinectics-400
FID	Spatial	133.15	79.11	80.42	169.76	100.65	112.22
	Spatiotemporal	133.69 (+0.4%)	79.35 (+0.3%)	79.57 (−1.1%)	170.10 (+0.2%)	100.62 (−0.0%)	112.85 (+0.6%)
FVD	Spatial	1460.18	211.08	354.49	1016.78	594.68	996.71
	Spatiotemporal	1705.27 (+16.8%)	286.39 (+35.7%)	367.35 (+3.6%)	1201.35 (+18.2%)	678.08 (+14.0%)	1155.53 (+15.9%)

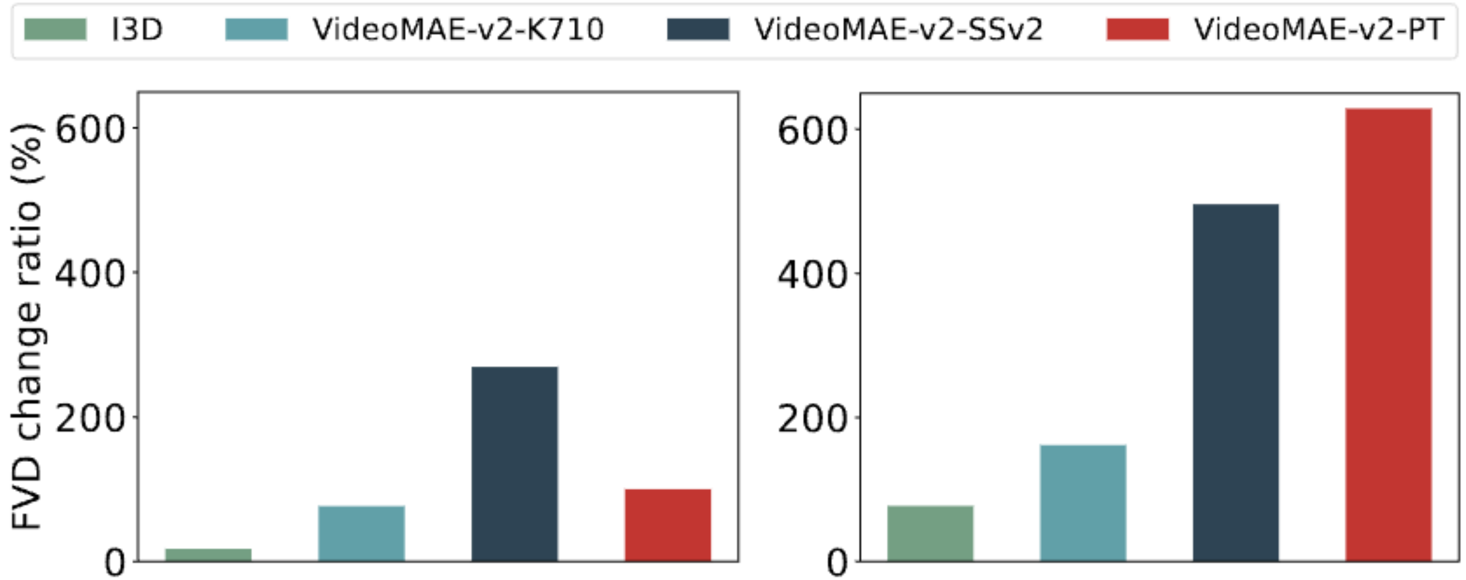
[6] FID [Heusel, et al. arXiv, 2022]

Temporal corruption doesn't affect frame quality

Metric	Distortion	UCF-101	Sky Time-lapse	FaceForencis	Taichi-HD	SSv2	Kinectics-400
FID	Spatial	133.15	79.11	80.42	169.76	100.65	112.22
	Spatiotemporal	133.69 (+0.4%)	79.35 (+0.3%)	79.57 (−1.1%)	170.10 (+0.2%)	100.62 (−0.0%)	112.85 (+0.6%)
FVD	Spatial	1460.18	211.08	$\boxed{354.49}$	1016.78	594.68	996.71
	Spatiotemporal	1705.27 (+16.8%)	286.39 (+35.7%)	$\boxed{367.35 \text{ (+3.6\%)}}$	1201.35 (+18.2%)	678.08 (+14.0%)	1155.53 (+15.9%)

FVD's temporal sensitivity = $\frac{\text{FVD}_{\text{spatiotemporal}} - \text{FVD}_{\text{spatial}}}{\text{FVD}_{\text{spatial}}} \times 100\%$

Understand temporal sensitivity by comparing with self-supervised video features



(a) Motion Blur (b) Elastic Transformation

CLIP-FID [Kynkäänniemi, et al. arXiv, 2022]
I3D [Carreira et al. CVPR, 2017]
VideoMAE-v2 [Wang et al. CVPR, 2023]

Case Study

Case study: StyleGAN-v

Default
StyleGAN-V

With LSTM
codes

StyleGAN-v [Skorokhodov, et al. CVPR, 2022]

Case study: StyleGAN-v

FVD Feature	StylegGAN-v	w/ LSTM codes
I3D	190.82	172.71(−18.11%)
VideoMAE-SSv2	332.80	616.74(+283.94%)
VideoMAE-K710	155.51	191.48(+35.97%)

StyleGAN-v [Skorokhodov, et al. CVPR, 2022]

Discussion

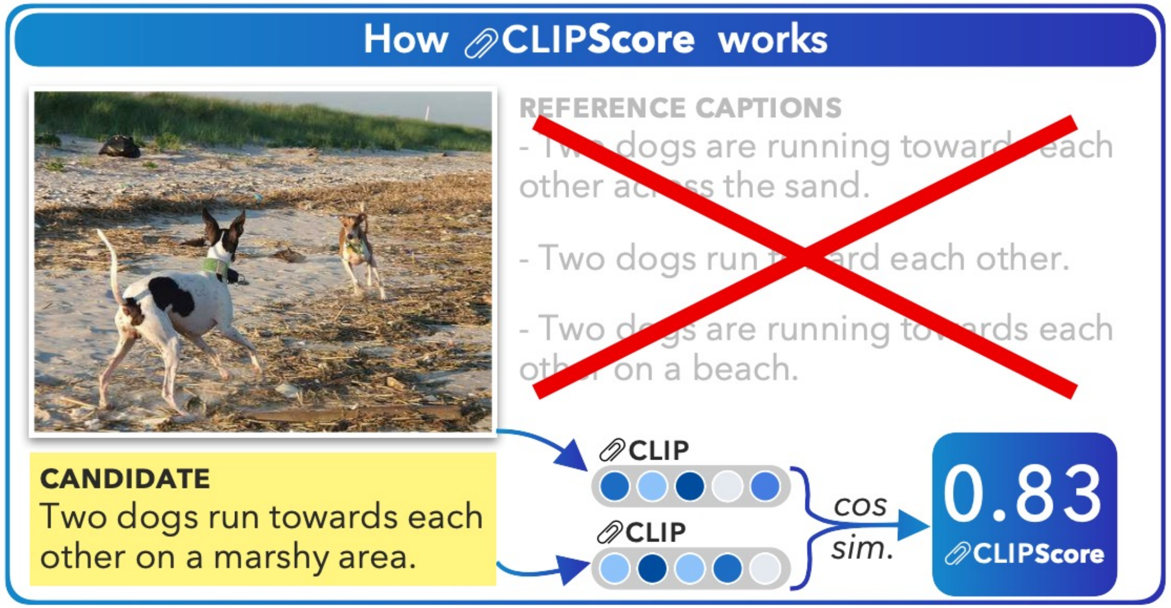
- FVD is highly biased towards per-frame quality over temporal consistency.
- Using self-supervised features improve its sensitivity to the temporal quality.
- Our new FVD toolbox (https://github.com/songweige/content-debiased-fvd) is available with `pip install cd-fvd`.

Summary

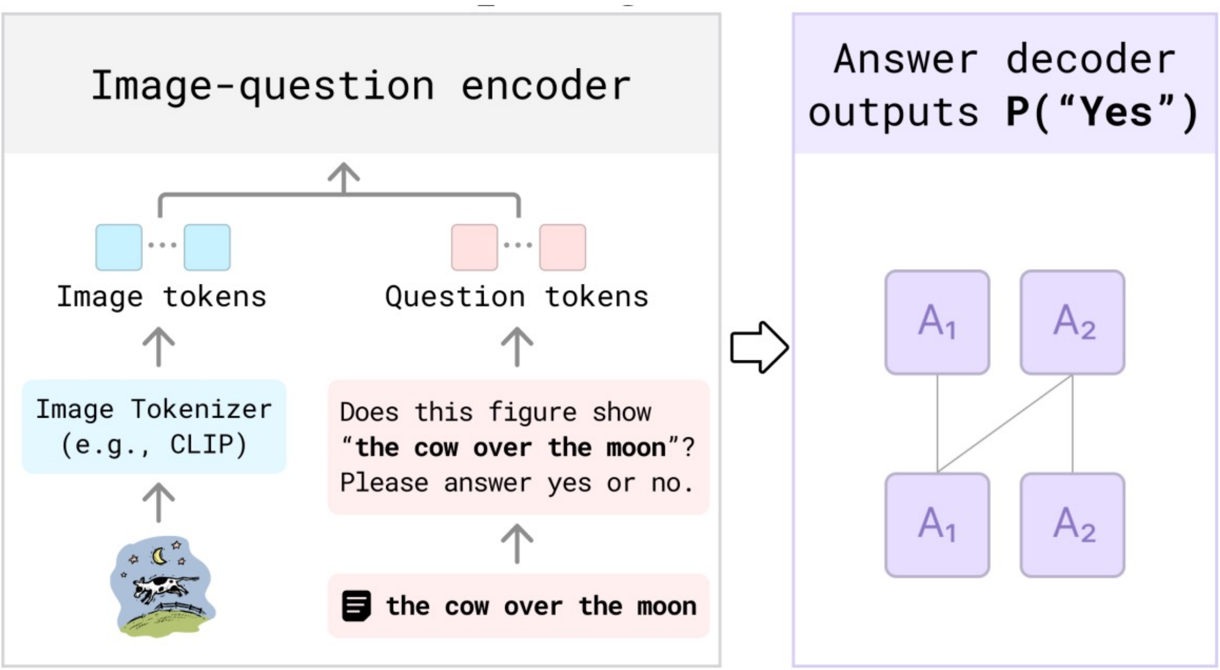
- FID lovers: Our state-of-the-art model improves MSCOCO FID from 6.8 to 6.75.
- FID haters: we should stop using metrics and just look at the pixels.
- Current takes: (1) use metrics with careful implementations; (2) use multiple evaluation protocols. (3) evaluate it on downstream applications.

Evaluation with (Multimodal) LLM

CLIPScore and VQAScore



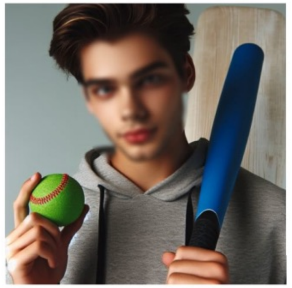
CLIPScore: A Reference-free Evaluation Metric for Image Captioning. Jack Hessel e al., 2021



Evaluating Text-to-Visual Generation with Image-to-Text Generation. Zhiqiu Lin e al., 2024.

CLIPScore and VQAScore

Text Prompt	DALL·E 3	Midjourney v6	SD-XL	DeepFloyd-IF
The brown dog chases the black dog around the tree.				
VQAScore (Ours)	**0.90**	0.69	0.60	0.32
Human	**4.67**	4.00	3.00	2.67
CLIPScore	0.27	**0.31**	0.28	0.25

Text Prompt	DALL·E 3	Midjourney v6	SD-XL	DeepFloyd-IF
A young man is holding a blue bat and a green ball.				
VQAScore (Ours)	**0.97**	0.96	0.87	0.52
Human	**4.33**	3.67	2.33	2.33
CLIPScore	0.28	0.30	**0.34**	0.31

CLIPScore and VQAScore

Text Prompt	DALL·E 3	Midjourney v6	SD-XL	DeepFloyd-IF
A snowy landscape with a cabin, but no smoke from the chimney.				
VQAScore (Ours)	0.15	0.10	**0.74**	**0.74**
Human	2.67	2.33	**4.67**	**4.67**
CLIPScore	0.28	**0.32**	0.30	0.26

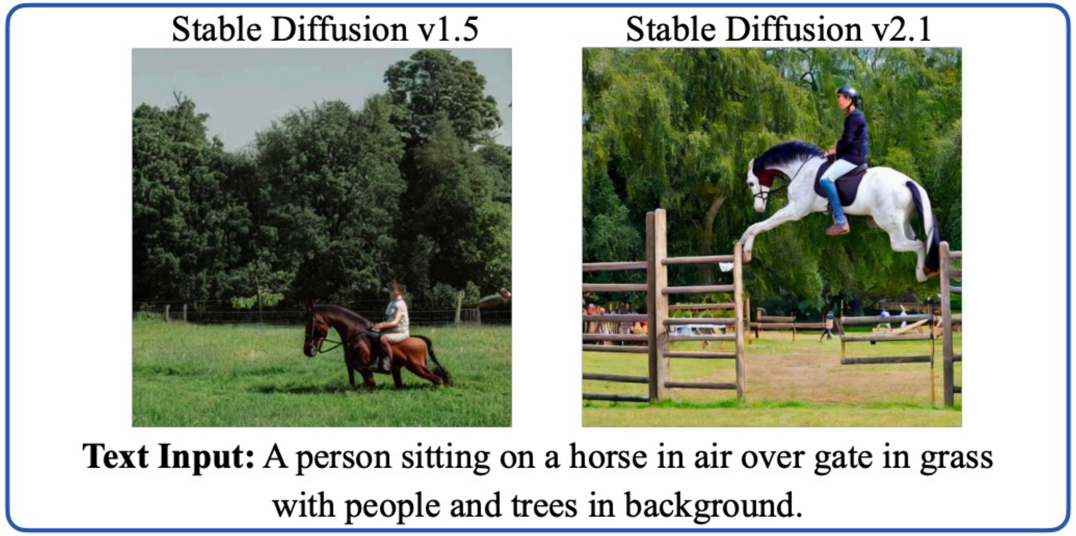
Text Prompt	DALL·E 3	Midjourney v6	SD-XL	DeepFloyd-IF
Two bicycles leaning against a wall with three windows.				
VQAScore (Ours)	0.94	0.94	0.95	**0.96**
Human	2.67	2.67	4.00	**4.67**
CLIPScore	0.30	**0.35**	0.30	0.30

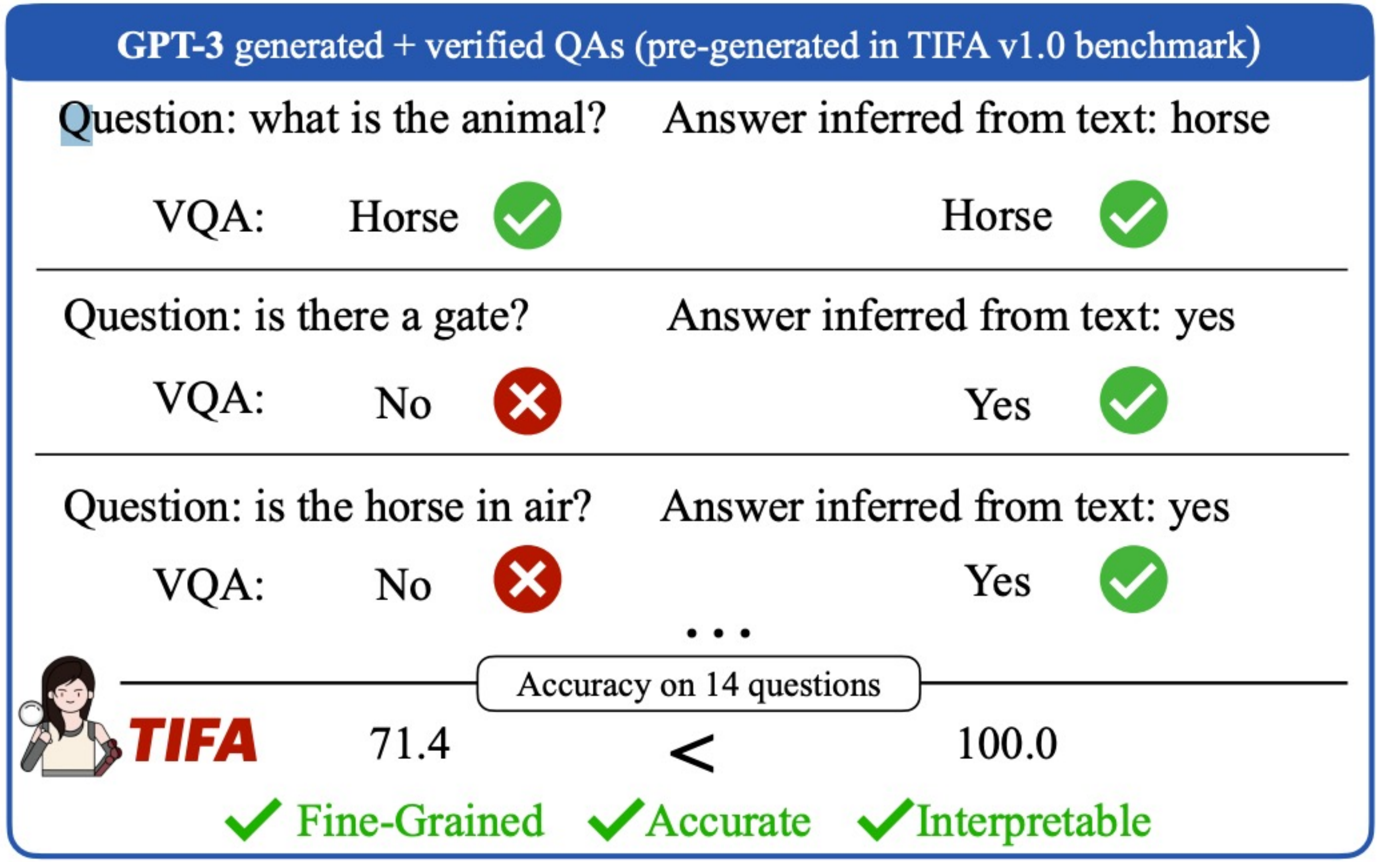
CLIPScore and VQAScore

Text Prompt	DALL·E 3	Midjourney v6	SD-XL	DeepFloyd-IF
A snowy landscape with a cabin, but no smoke from the chimney.				
VQAScore (Ours)	0.15	0.10	**0.74**	**0.74**
Human	2.67	2.33	**4.67**	**4.67**
CLIPScore	0.28	**0.32**	0.30	0.26

Text Prompt	DALL·E 3	Midjourney v6	SD-XL	DeepFloyd-IF
Two bicycles leaning against a wall with three windows.				
VQAScore (Ours)	0.94	0.94	0.95	**0.96**
Human	2.67	2.67	4.00	**4.67**
CLIPScore	0.30	**0.35**	0.30	0.30

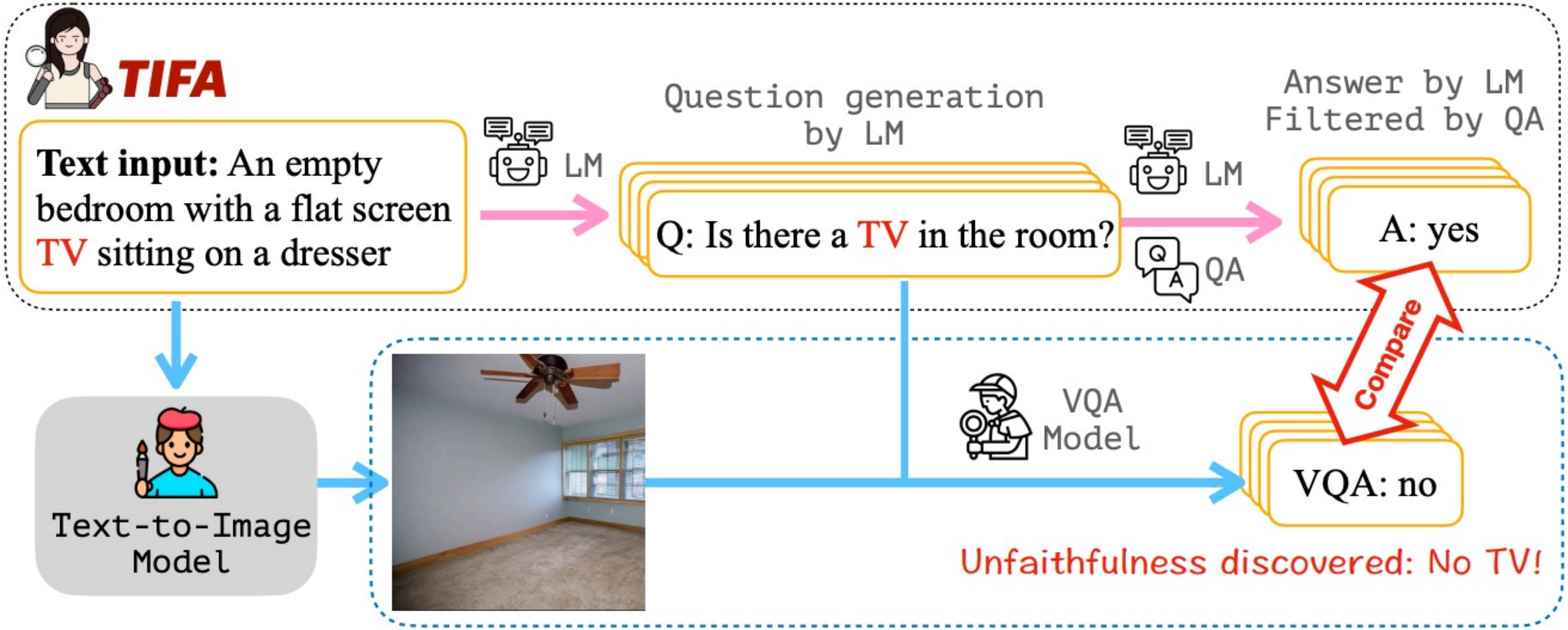
TIFA with Question Answering





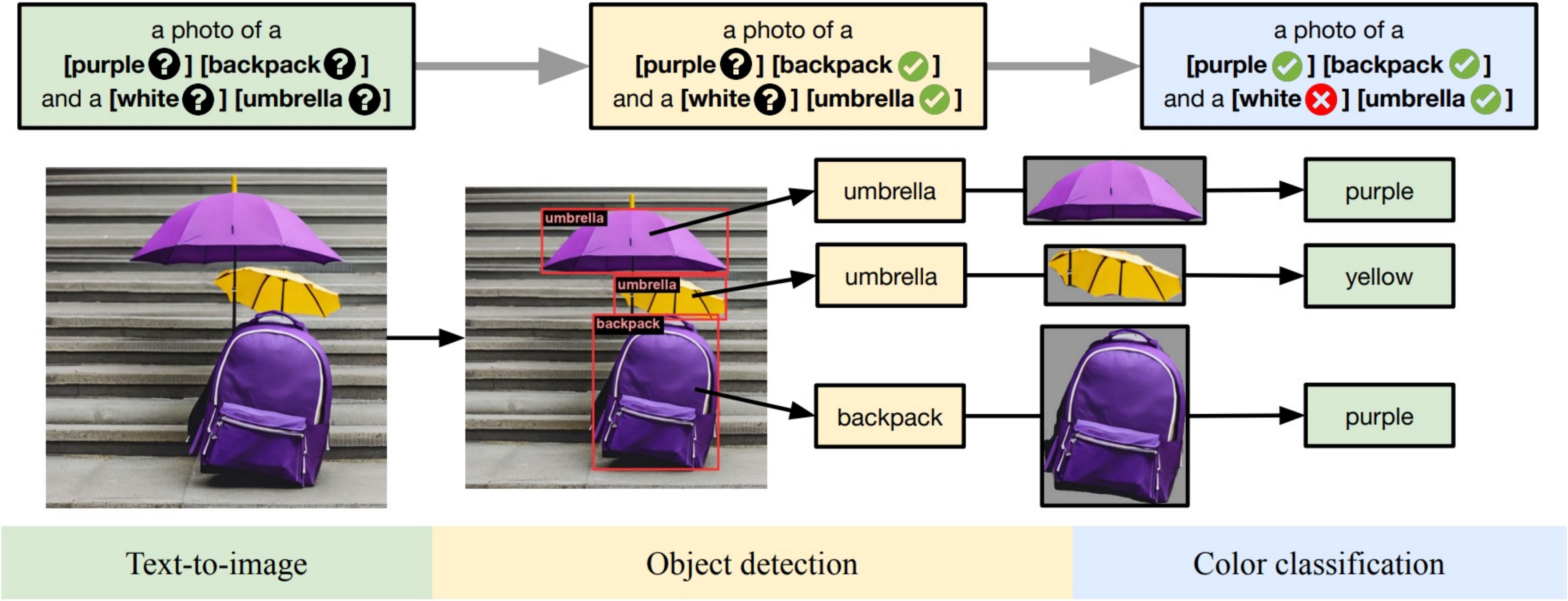
TIFA: Text-to-Image Faithfulness Evaluation with Question Answering
Yushi Hu e al., 2023.

TIFA with Question Answering



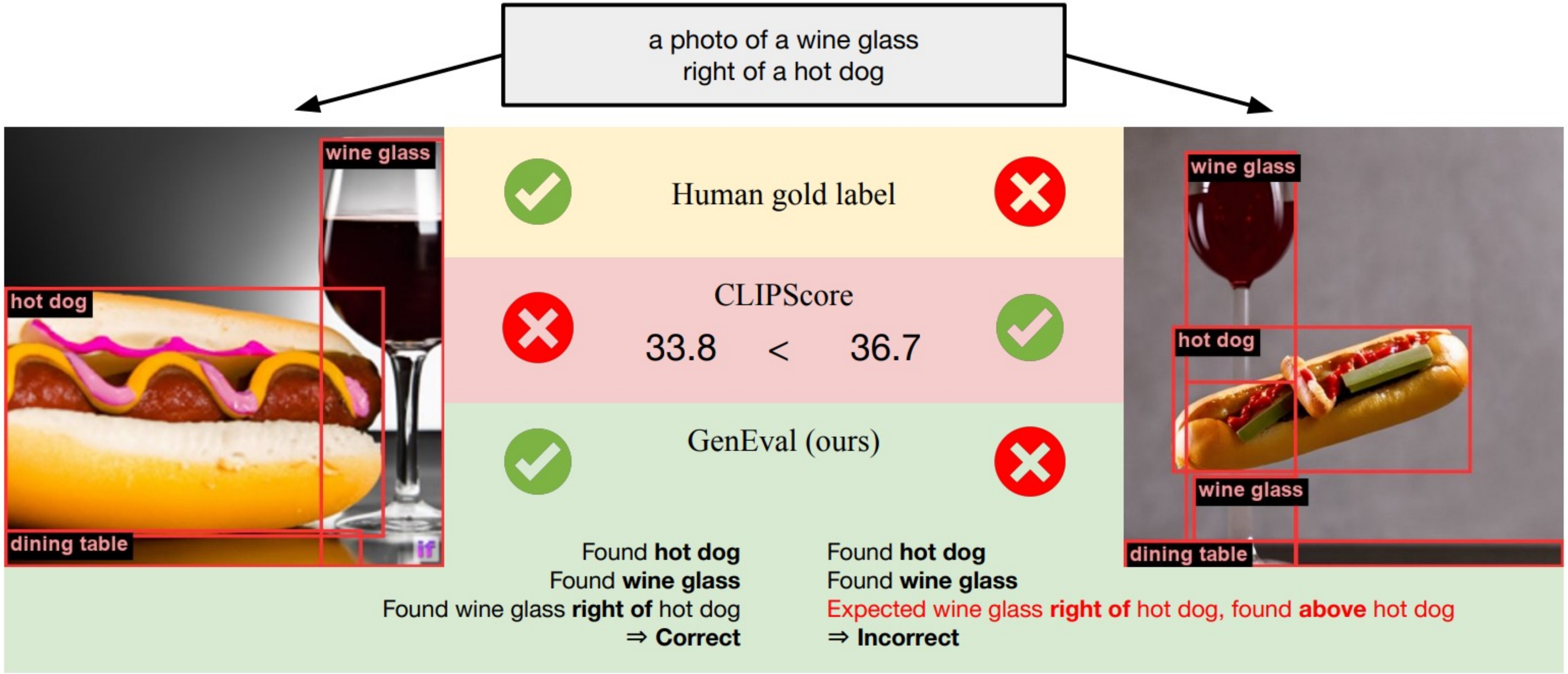
TIFA: Text-to-Image Faithfulness Evaluation with Question Answering
Yushi Hu e al., 2023.

GenEval: Object-Focused Evaluation



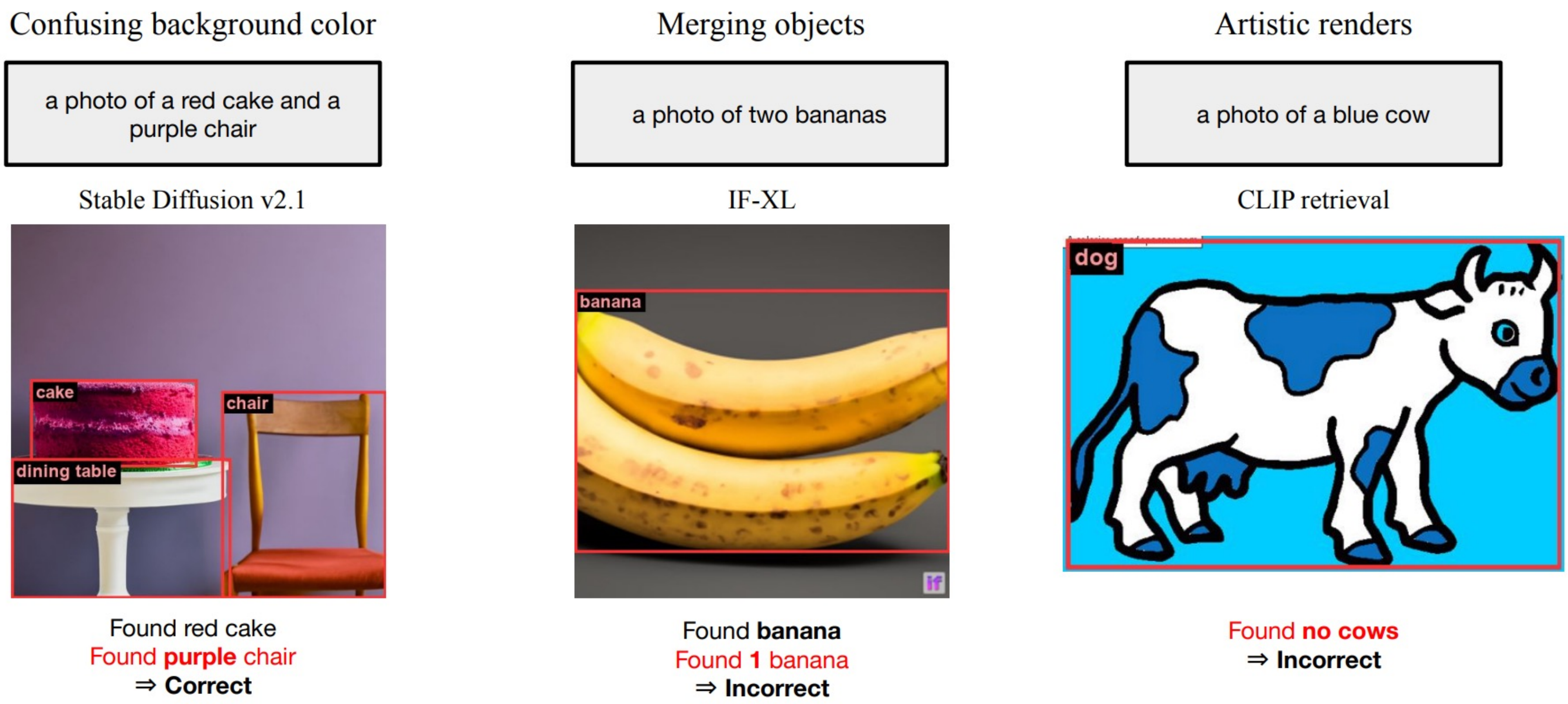
GenEval: An Object-Focused Framework for Evaluating Text-to-Image Alignment
Dhruba Ghosh et al., 2023.

GenEval: Object-Focused Evaluation



GenEval: An Object-Focused Framework for Evaluating Text-to-Image Alignment
Dhruba Ghosh et al., 2023.

GenEval: Object-Focused Evaluation



GenEval: An Object-Focused Framework for Evaluating Text-to-Image Alignment
Dhruba Ghosh et al., 2023.

Which one shall we use?

How do we evaluate
“evaluation metrics”?