Revisiting the Evaluation of Image Synthesis with GANs

Mengping Yang[1,*] Ceyuan Yang[1,*] Yichi Zhang[1] Qingyan Bai[2] Yujun Shen[2] Bo Dai[1]

[1]Shanghai AI Laboratory, [2]Ant Group

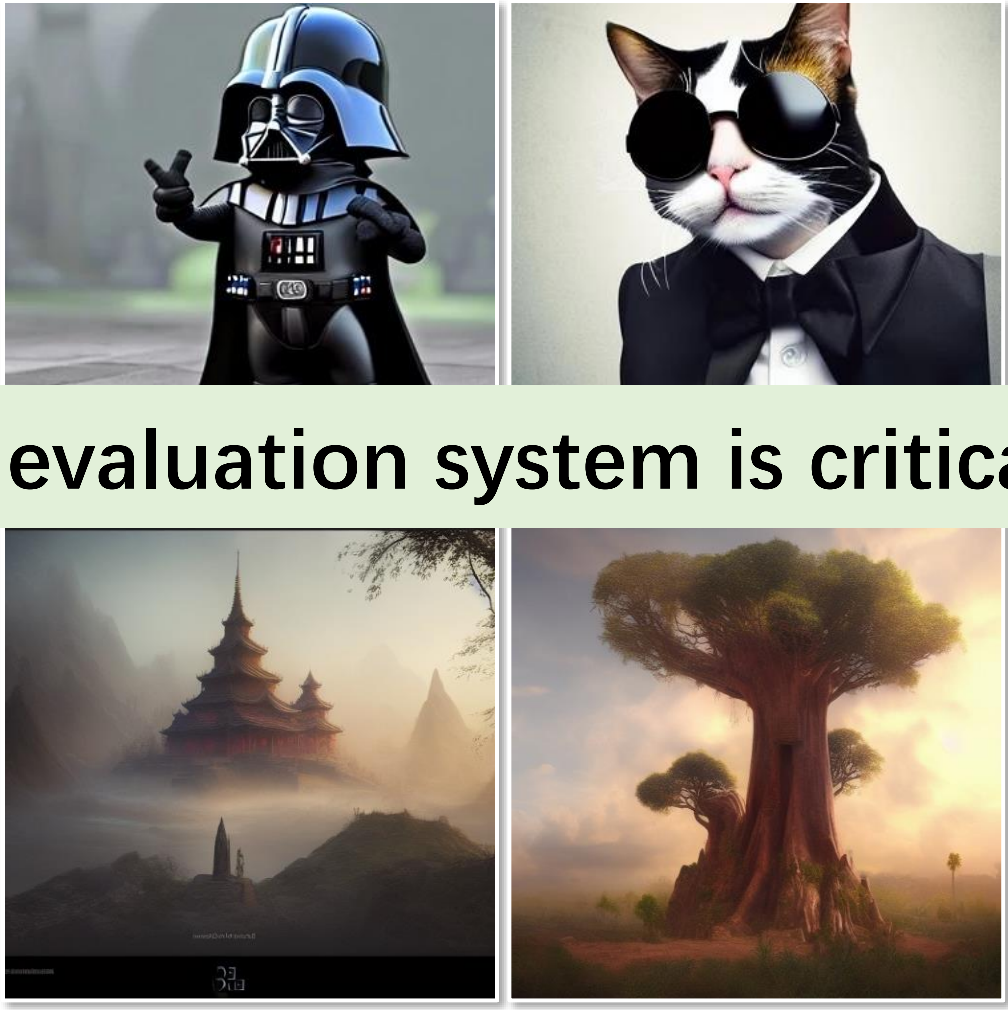
* Denotes Equal Contribution

https://github.com/kobeshegu/Synthesis-Measurement-CKA

Explosive developments of generative models

A consistent and comprehensive evaluation system is critical!

Random generated Churches

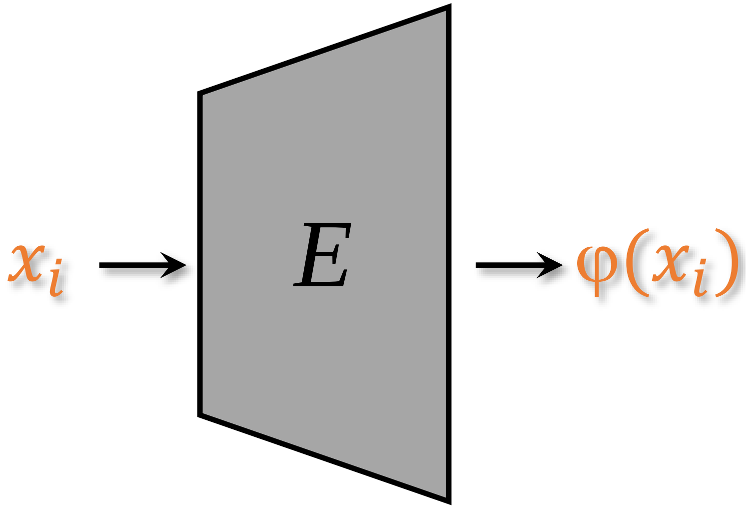


Artwork generated by Stable Diffusion

Credit: https://stablediffusion.fr/

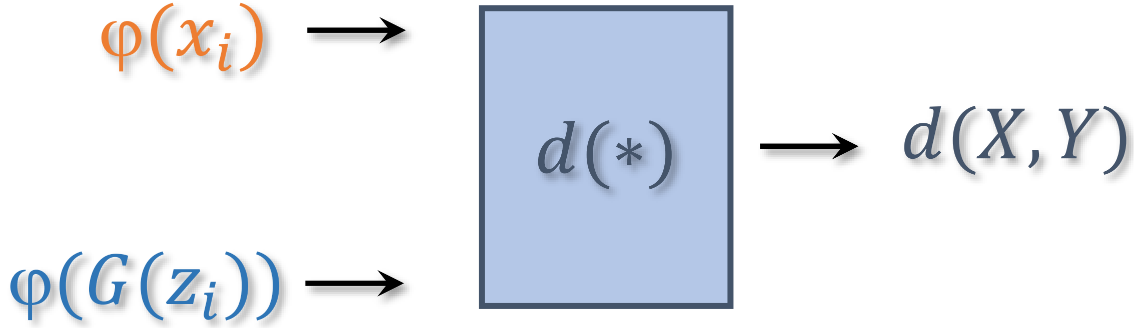
Two essential components for synthesis evaluation

Feature Extractor $\varphi(*)$



Extracting samples' features

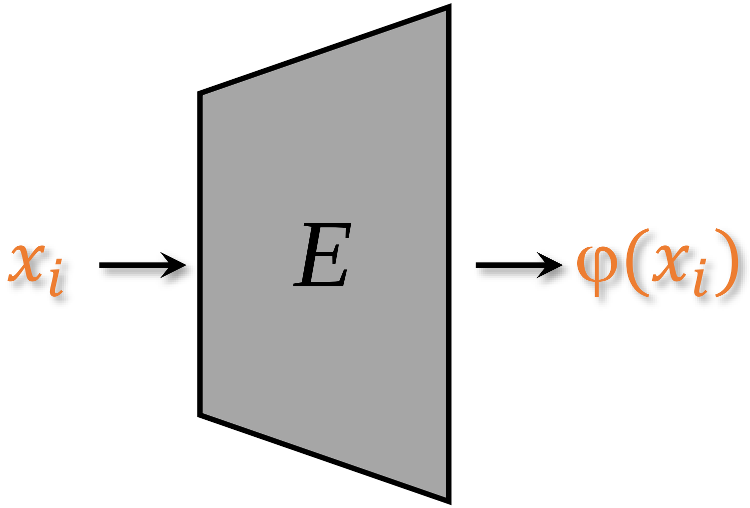
Distributional Distance $d(*)$



Delivering the distribution divergence

Several key factors *w.r.t* feature extractors

Feature Extractor $\varphi(*)$



Extracting samples' features

Features extractors define measurement spaces for evaluation, they differ in:

- **Supervision** (Fully/Self-supervised)
- **Network architectures** (CNN vs. ViT)
- **Representation spaces** (Similarity)

Extractors yield *different* focus on *various semantics*

- CNN-based extractors highlight **objects related to the pre-trained domain** (*e.g.*, microphone, hat, and sunglasses)
- ViT-based extractors capture **larger** regions
- Multiple extractors **complement** each other

Original Inception ConvNeXt SWAV MoCo-R CLIP-R RepVGG Swin-Trans ViT DeiT CLIP-V MoCo-V ResMLP

Extractors may define similar (homogeneous) spaces

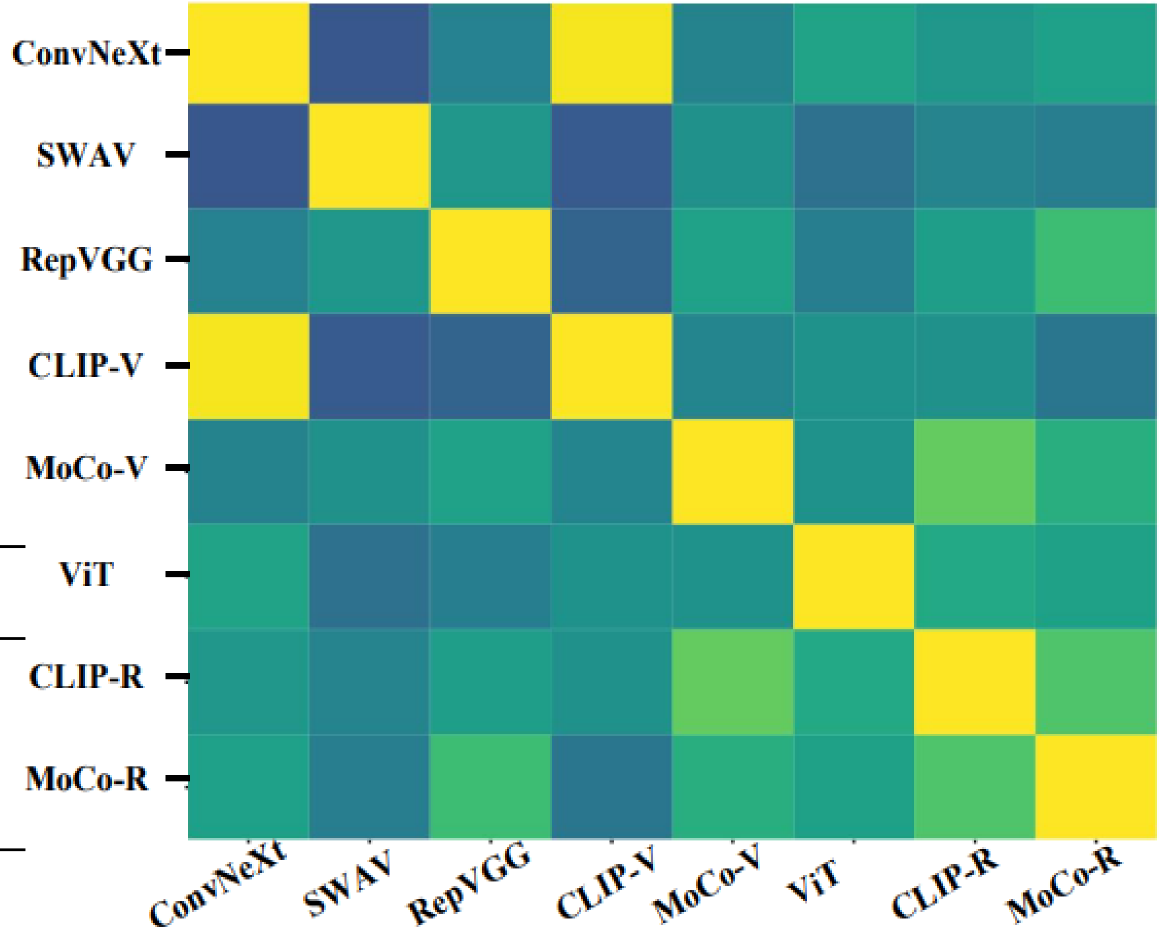
- Similar representation spaces are redundant in practice
- Remaining extractors:

CNN-based extractors	**ConvNeXt, SWAV, RepVGG**
ViT-based extractors	**CLIP-ViT, MoCo-ViT, ViT**

- These extractors provide reliable rank:

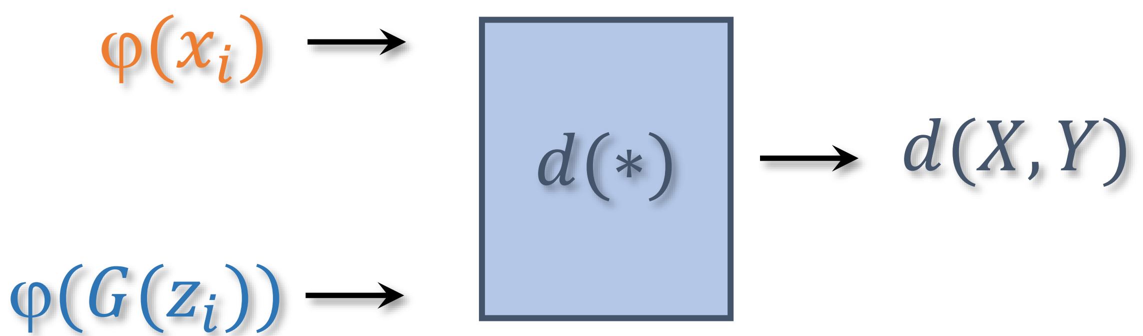
Model	ConvNeXt	RepVGG	SWAV	ViT	MoCo-V	CLIP-V
BigGAN	140.04	67.53	1.12	29.95	238.78	3.35
-deep	102.26	58.85	0.87	23.98	85.83	3.22
StyleGAN-XL	19.22	15.93	0.18	8.51	29.38	1.85

StyleGAN-XL > BigGAN-deep > BigGAN



Investigation on different distributional distances

Distributional Distance



Delivering the distribution divergence

Various distances reflect different divergence, they are influenced by:

- **Source of features** (features from different layers and spaces)
- **The amount of synthesized samples**

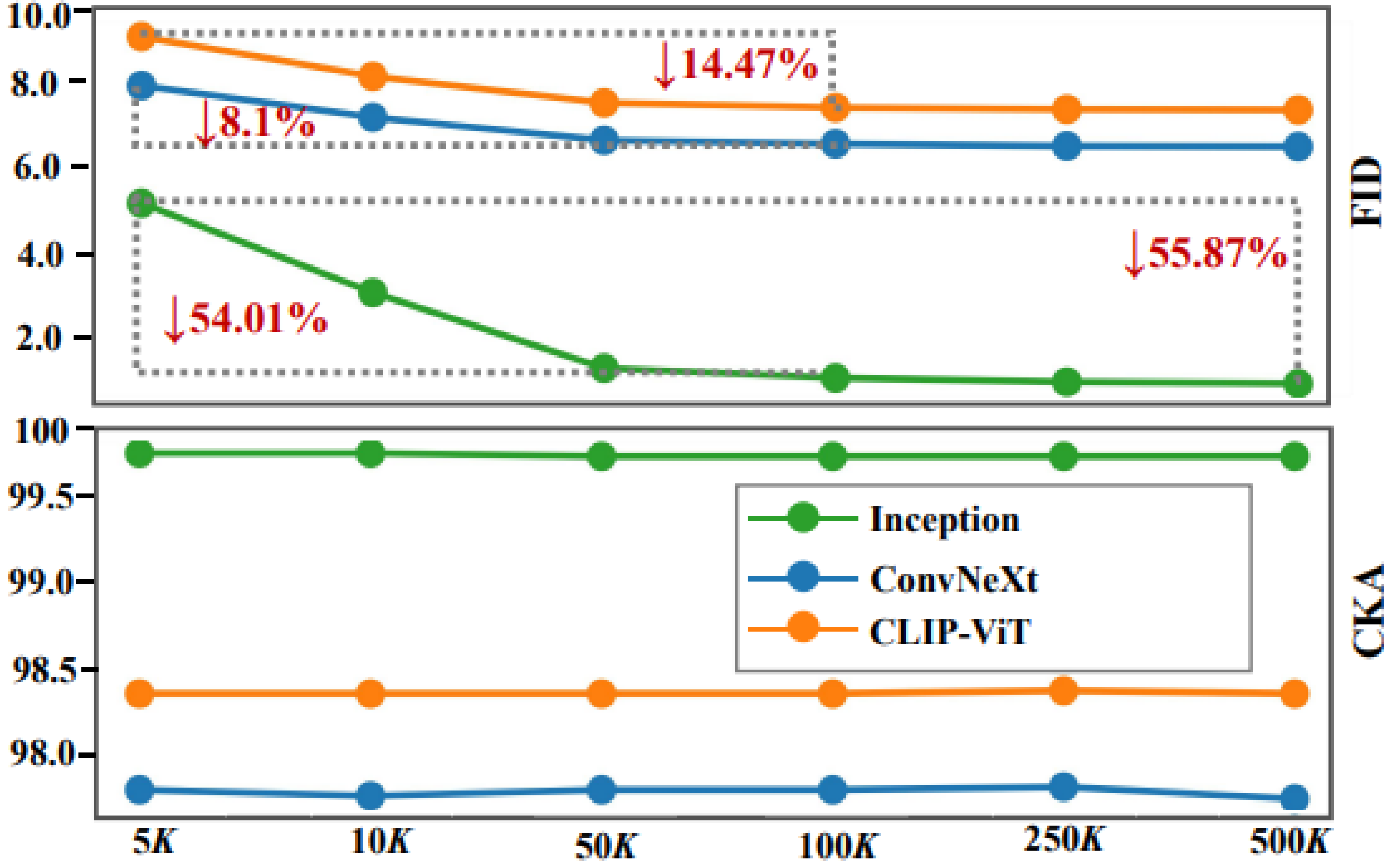
CKA provides *normalized scores* in various spaces

- Comparable between **hierarchical layers** and **representation spaces**
- Easier to **combine** scores from **different extractors**
- The FD scores of various layers **fluctuate dramatically**

Features from shallow to deep layers

Model	BigGAN		StyleGAN-XL	
Layer	FD ↓	CKA ↑	FD ↓	CKA ↑
Layer_1	0.60	99.06	0.05	99.84
Layer_2	7.45	86.89	0.77	91.06
Layer_3	30.24	82.80	6.11	85.75
Layer_4	104.10	80.13	35.77	83.55
Overall	N/A	**87.22**	N/A	**90.05**

CKA shows satisfactory *sample-efficiency* and *stability*

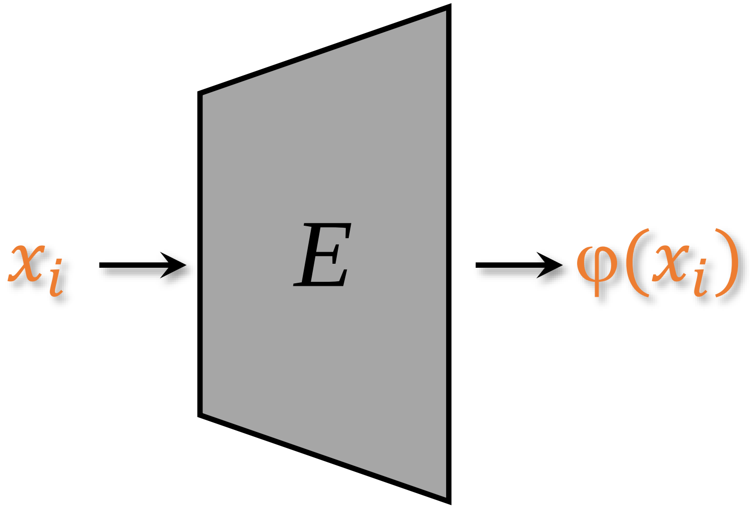


Centered Kernel Alignment:

- **Stable** under different data amount
- **Less samples** are required for reliable evaluation
- FID Scores could be altered by synthesizing more samples

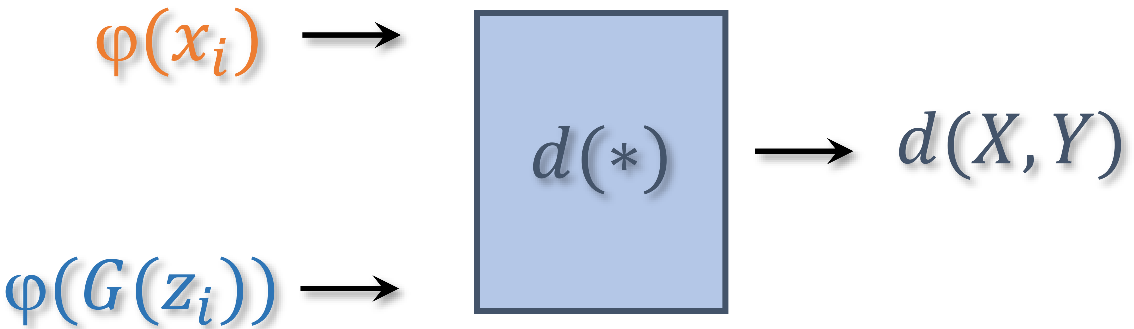
Our new measurement system

Multiple feature extractors



Extracting samples' features

Center Kernel Alignment

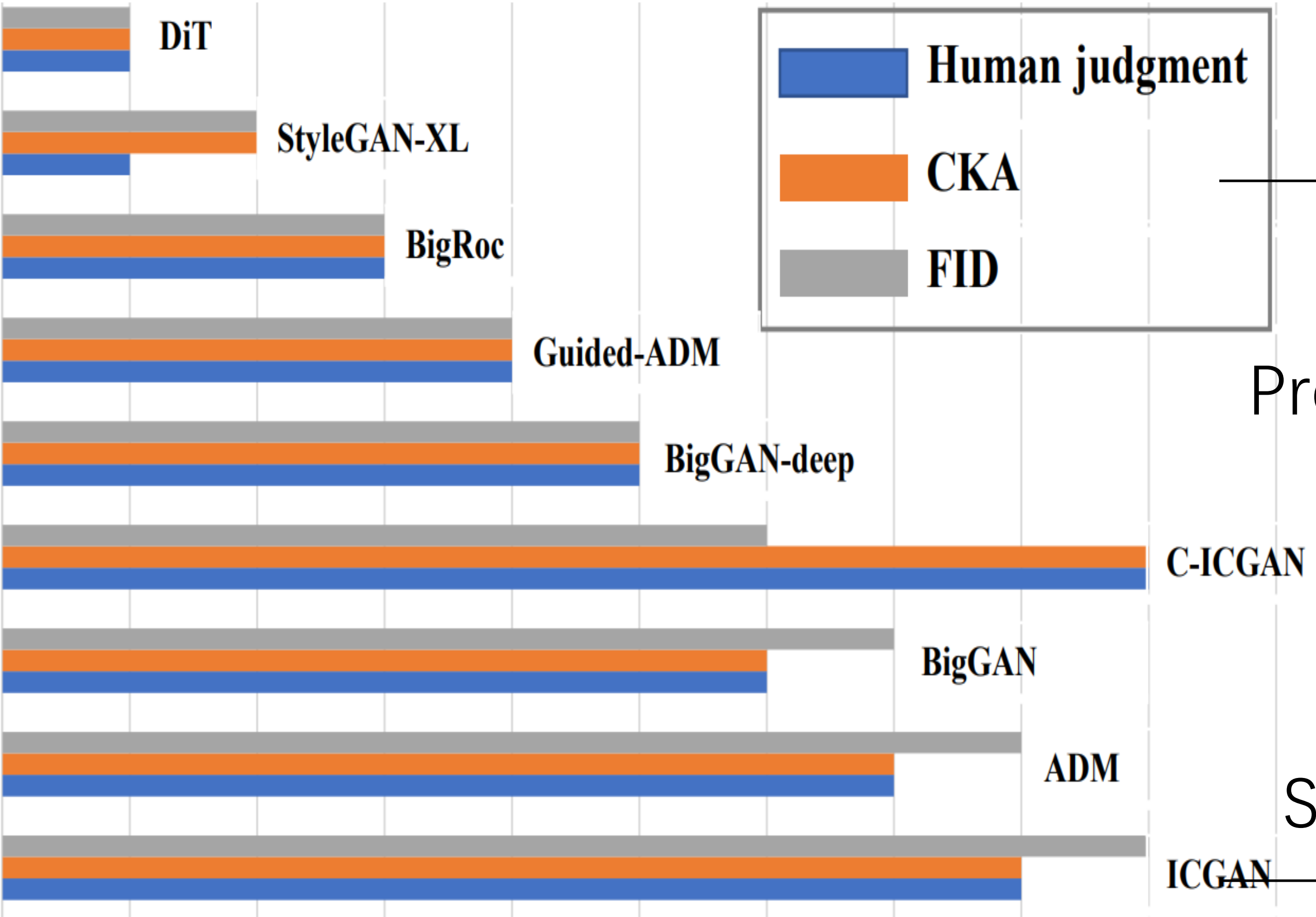


Delivering the distribution divergence

Our evaluation system facilitates more comprehensive evaluation!

Benchmark 1: Re-evaluate existing generative models

Our evaluation ***correlates well*** **with** ***human visual judgment***



Model	FID↓	CKA↑	Human eval
StyleGAN2	3.66	91.61	45%
Projected-GAN	3.39	91.41	39%
InsGen	3.31	92.58	58%
EqGAN	2.89	92.63	62%
StyleGAN-XL	2.19	92.85	66%

Benchmark 2: GANs *v.s.* Diffusion models

GANs achieve **better trade-offs** between efficiency and quality
Designing **computation-efficient** diffusion models is essential

Model	FID$_{\downarrow}$	CKA$_{\uparrow}$	Human eval	#Params	Sec/Kimg(s)
BigGAN	8.70	82.82	53%	158.3 M	33.6
BigGAN-deep	6.95	83.65	55%	85 M	27.6
StyleGAN-XL	2.30	86.52	67%	166.3 M	64.8
ADM	10.94	82.12	45%	500 M	17274
Guided-ADN	4.59	84.66	57%	554 M	17671
DiT	2.27	86.61	67%	675 M	3736.8

Benchmark 2: Image-to-Image translation

Our system is **generalizable** for different synthesis tasks

Horse-to-Zebra dataset

Model	FID	ConvNeXt	RepVGG	SWAV	ViT	MoCo-ViT	CLIP-ViT	Overall
CycleGAN [71]	83.32	73.55	88.67	85.82	83.96	74.72	73.74	80.08
AttentionGAN [57]	76.05	75.59	91.73	86.37	85.16	76.65	75.49	81.83
CUT [43]	51.29	78.48	93.22	88.83	87.84	78.75	77.36	84.08

Cat-to-Dog

Model	FID	ConvNeXt	RepVGG	SWAV	ViT	MoCo-ViT	CLIP-ViT	Overall
CUT [43]	74.95	84.93	78.75	88.83	84.31	93.56	70.91	83.55
GP-UNIT [68]	60.96	90.45	87.79	94.05	90.12	95.91	75.32	88.94

Cat-to-Dog

Model	FID	ConvNeXt	RepVGG	SWAV	ViT	MoCo-ViT	CLIP-ViT	Overall
GP-UNIT [68]	31.66	79.58	78.18	96.79	86.93	93.92	77.42	85.47
MUNIT [25]	18.88	84.87	84.11	98.51	88.11	95.95	86.10	89.61

Code

ArXiv

Revisiting the Evaluation of Image Synthesis with GANs

Mengping Yang[1,*] Ceyuan Yang[1,*] Yichi Zhang[1] Qingyan Bai[2] Yujun Shen[2] Bo Dai[1]

[1]Shanghai AI Laboratory, [2]Ant Group

[*] Denotes Equal Contribution