Can GAN's still compete?

"Scaling up GANs for Text-to-Image Synthesis"

Kang, Zhu, Zhang, Park, Shechtman, Paris, Park · POSTECH, CMU, Adobe Research

0.13 s to generate a 512 px image

GigaGAN samples, Fig. 1

GAN's: The Past and Present

Diffusion models trade inference speed for parameter scaling; GAN's preserve fast speed but stalled out at small scale.

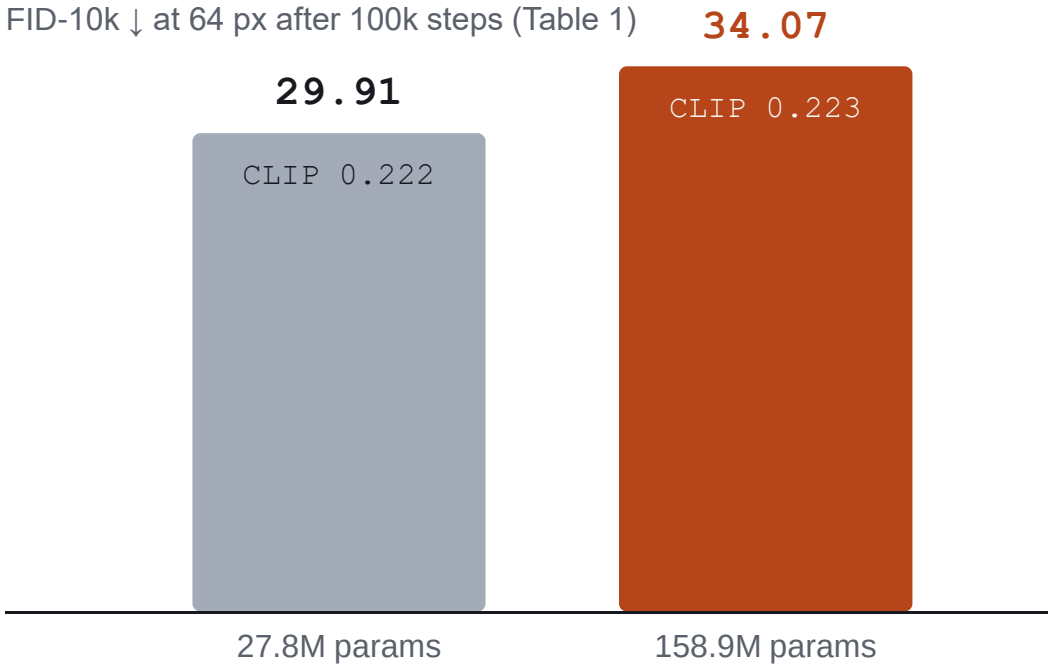
Diffusion · Stable Diffusion v1.5 50 steps · 2.9 s

GAN · GigaGAN 1 step · 0.13 s

The same conv filters must serve every prompt at every location, and low-resolution layers go inactive as models grow.

Widening StyleGAN2 5.7× makes FID worse

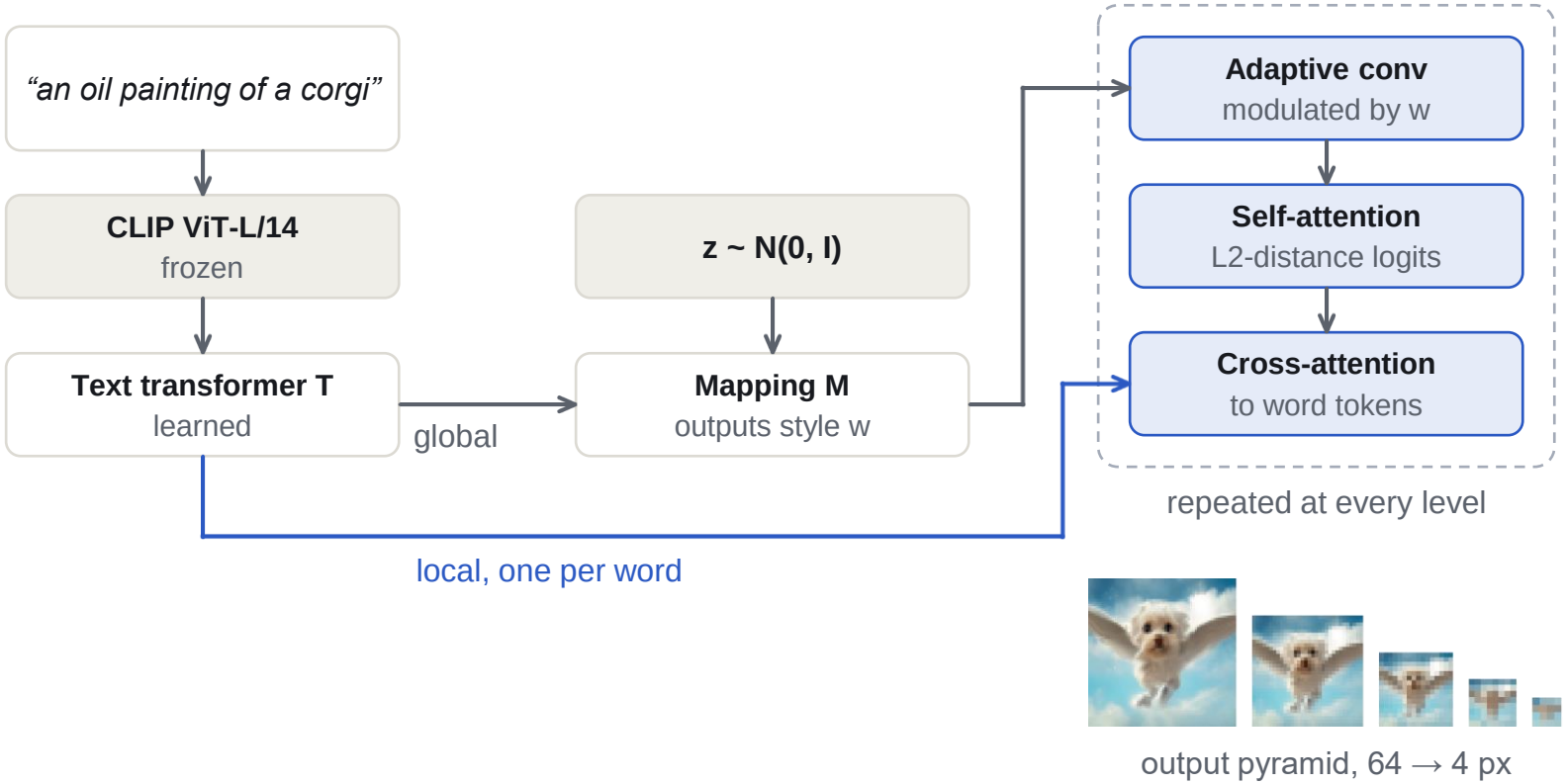
FID-10k ↓ at 64 px after 100k steps (Table 1)



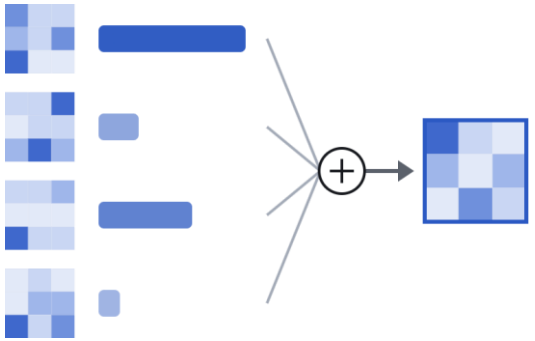
GigaGAN · Kang et al., CVPR 2023

Adaptive Convolution Kernel Selection

Filters are mixed per sample from a bank; attention adds long-range context and word-level text.



Sample-adaptive kernel selection



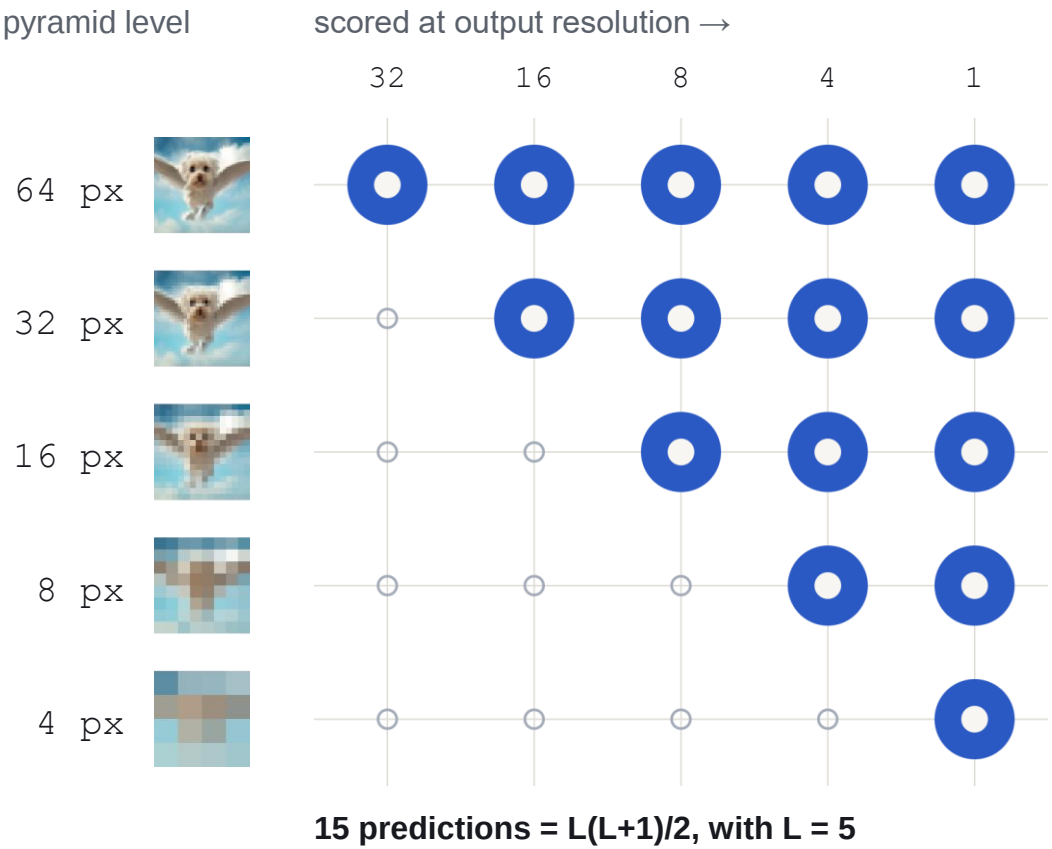
N kernels → softmax weights from w → one mixed kernel K

Mixing happens once per layer, so its cost does not grow with resolution (Eq. 1).

$$K = \Sigma\ \alpha_i(w)\ \cdot K_i$$

Pyramid Outputs

Each pyramid level enters at its own depth and is scored at every coarser output.



15 predictions = L(L+1)/2, with L = 5

Matching-aware loss · extended

+ “a teddy bear on tabletop” → **labeled fake**

Applied to generated images as well as real ones (Table 1).

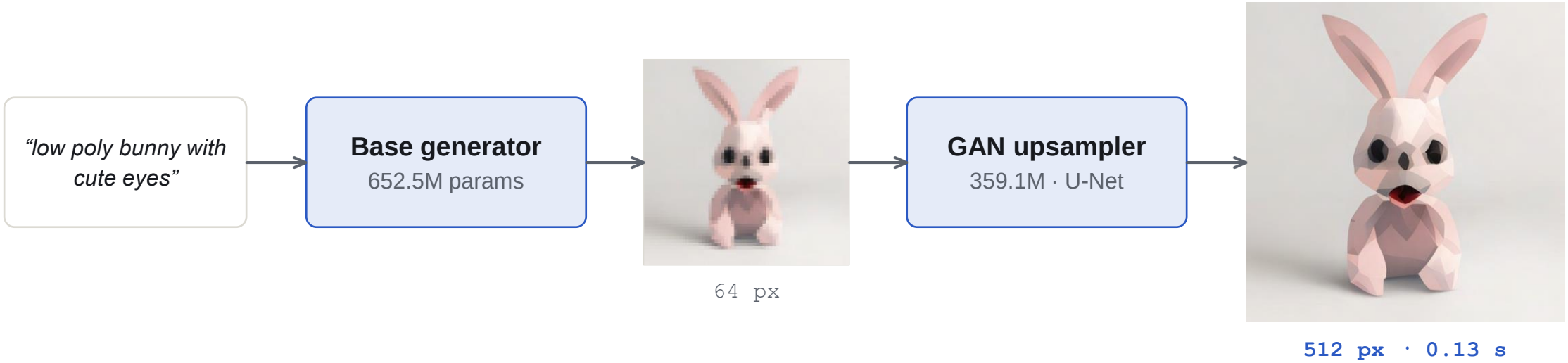
CLIP contrastive loss · borrowed

Each output must be matched to its own caption within the batch, scored by frozen CLIP.

Vision-aided discriminator · borrowed

A second discriminator on frozen CLIP ViT-B/32 features (Kumari et al. 2022).

Dual Stage Design



Upsampler losses

GAN, matching and CLIP losses, plus LPIPS against the real high-res image; noise is added to inputs.

128 → 1024 variant

Trained on Adobe Stock images. Applied twice, it reaches 4K (16 MP) in 3.66 s.

Training compute, A100-days

Improvements

1.0B
parameters: 652.5M base + 359.1M upsampler

0.98B
training images seen, LAION2B-en + COYO-700M

36×
larger than StyleGAN2, 6× StyleGAN-XL

New

Sample-adaptive kernels
Filter bank mixed per sample from style w

Multi-scale I/O discriminator
Each pyramid level scored at every coarser scale

Extended

Matching-aware loss on fakes
Mismatched captions applied to generated images too

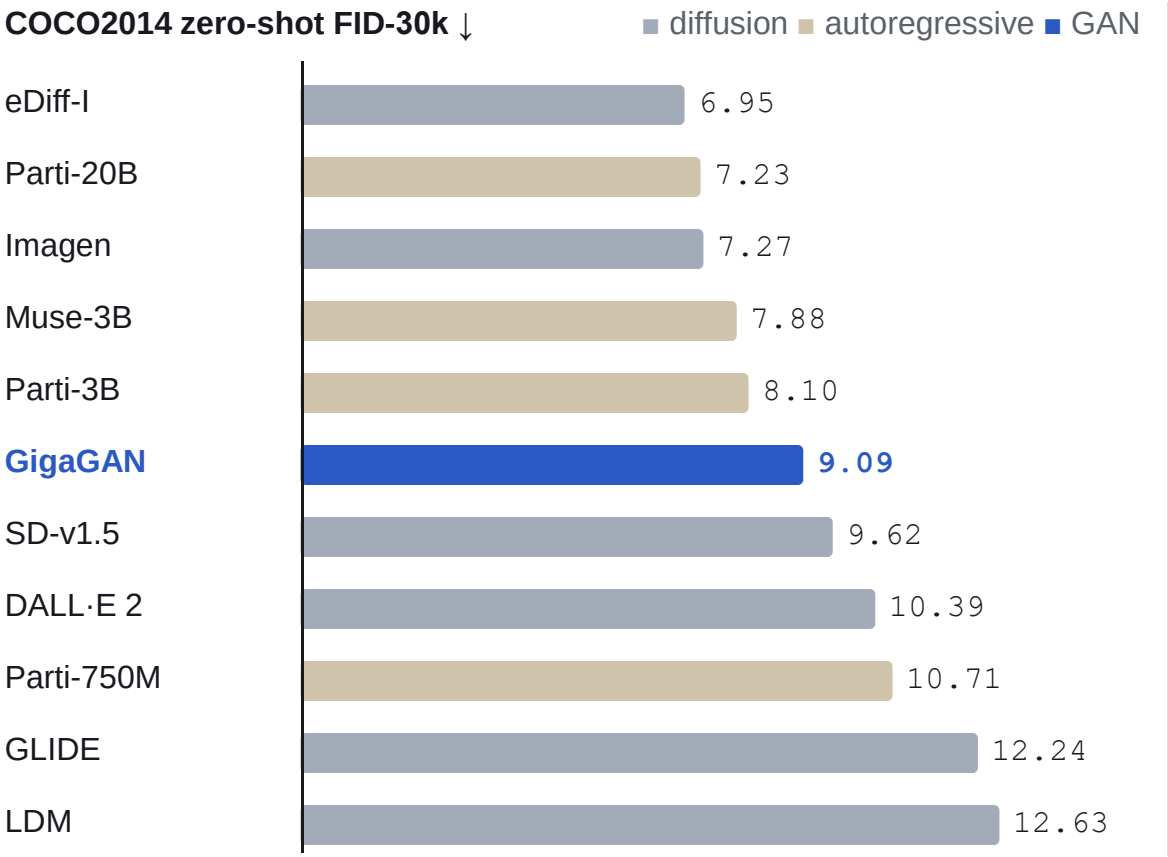
Text-conditioned truncation
Truncate toward the prompt's own mean style

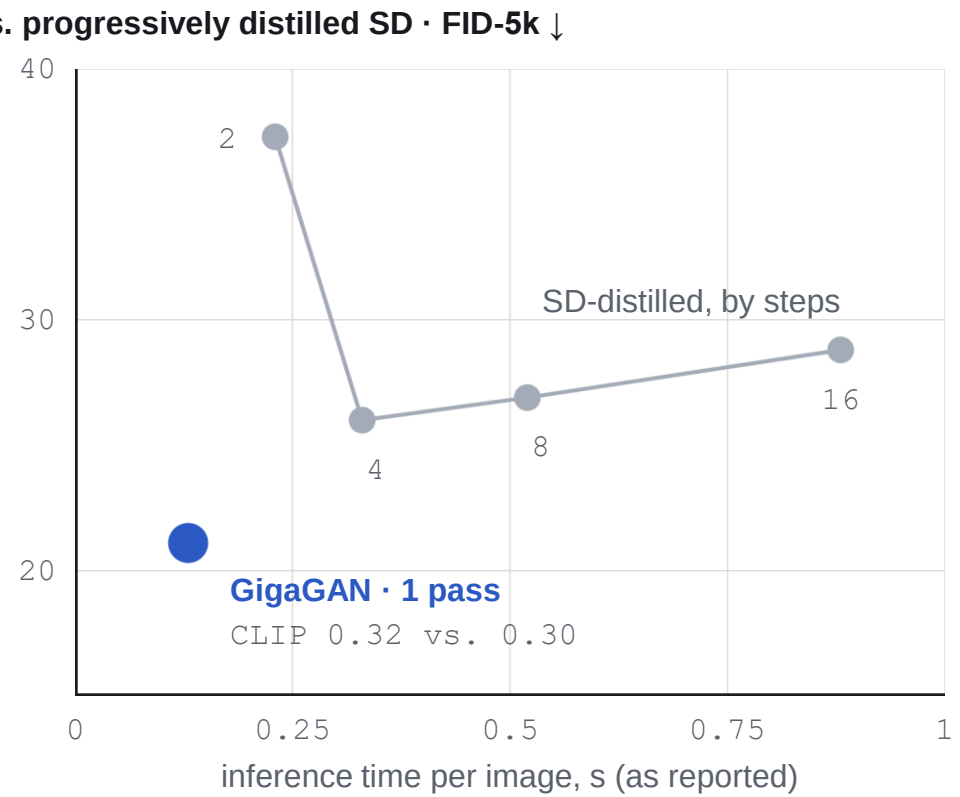
Borrowed and integrated

- L2 self-attention · ViTGAN
- Word-token cross-attention · diffusion
- CLIP contrastive loss · prior GANs
- Vision-aided D · Kumari et al.
- 64 px cascade · Imagen, DALL·E 2
- StyleGAN2 backbone · Karras et al.

Comparative Performance

COCO FID-30k across published models (Table 2) and against distilled Stable Diffusion (Table 3).



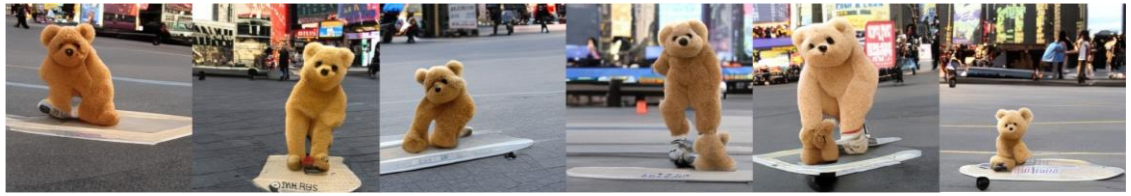


Comparative Performance

Random samples for one prompt (Fig. A12) beside the paper's FID–CLIP trade-off (Fig. A3).

Prompt: "A teddy bear on a skateboard in Times Square"

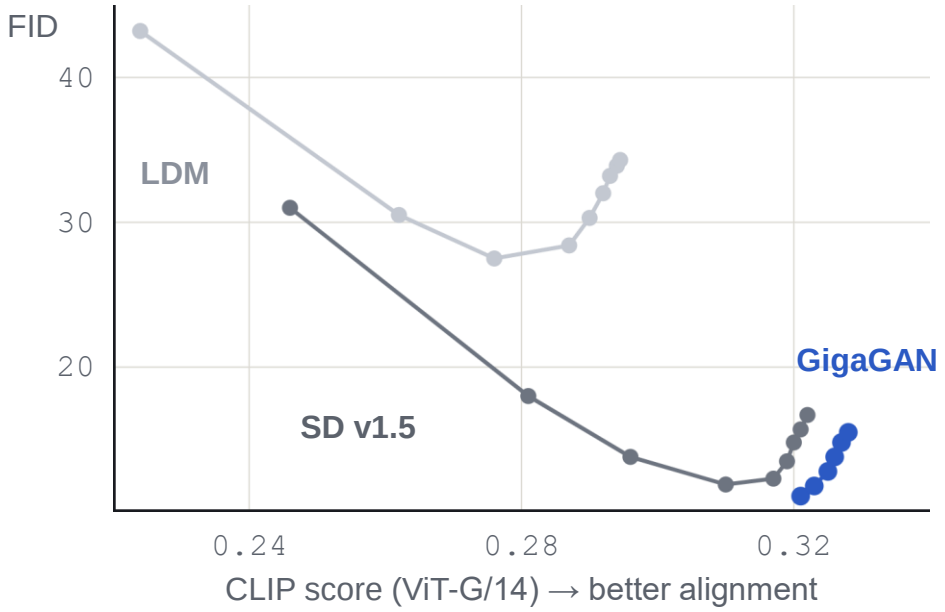
GigaGAN · ψ = 0.8 · 0.14 s



Stable Diffusion v1.5 · 50 steps · 2.9 s

DALL·E 2 · official service

FID–CLIP trade-off · Fig. A3 redrawn (approx.)

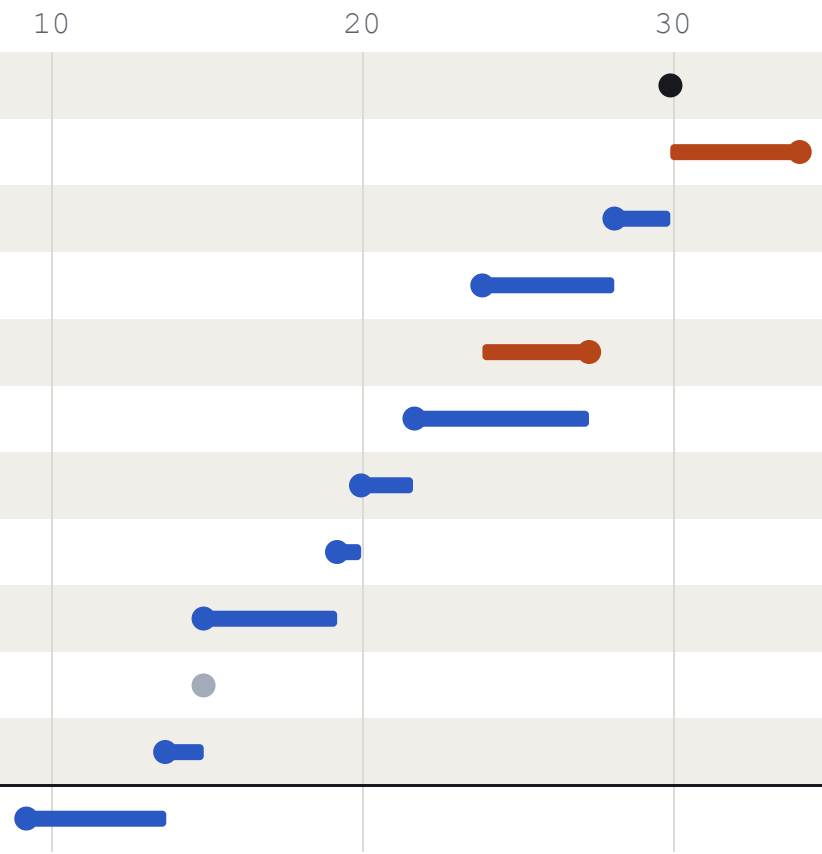


Better on the curve, not in the samples. Suspects: CLIP-based losses and data filtering, 512 → 256 px evaluation.

Comparative Performance

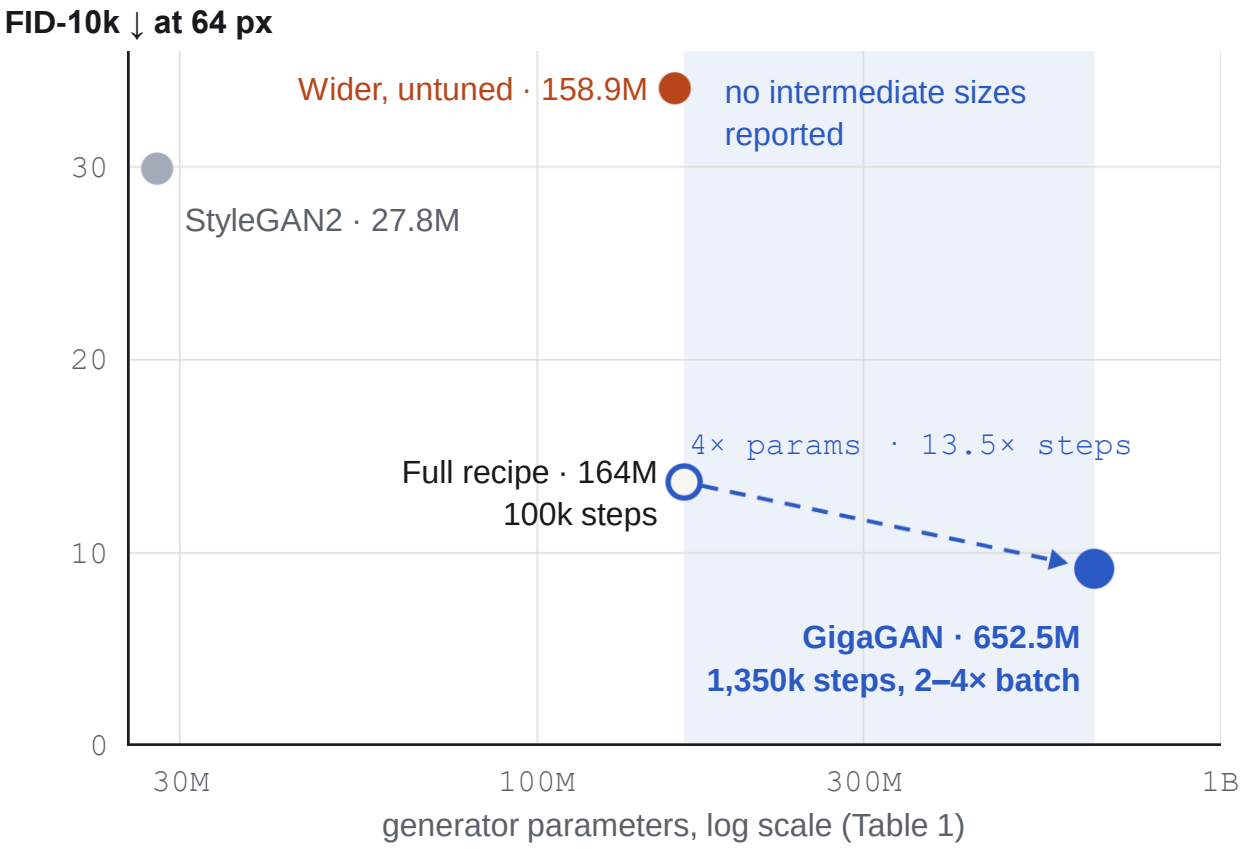
Table 1: cumulative, one run each, 100k steps at 64 px; bars show change from the row above.

Table 1 row	FID ↓	CLIP ↑	note
StyleGAN2	29.91	0.222	
+ Larger (5.7×)	34.07	0.223	bigger, worse
+ Tuned	28.11	0.228	shrinks to 26.2M
+ Attention	23.87	0.235	
+ Matching-aware D	27.29	0.250	FID worsens
+ Matching-aware G and D	21.66	0.254	
+ Adaptive convolution	19.97	0.261	
+ Deeper	19.18	0.263	
+ CLIP loss	14.88	0.280	
+ Multi-scale training	14.92	0.300	FID flat
+ Vision-aided GAN	13.67	0.287	CLIP drops
+ Scale-up (GigaGAN)	9.18	0.307	also 13.5× steps



Potential Limitations in Scale

§5 calls this “a conclusive answer about the scalability of GANs.” Here is the evidence.



ImageNet 256 · FID ↓ · Table A1

StyleGAN-XL · 166M	2.32
GigaGAN · 569M	3.45
BigGAN-Deep · 112M	6.95

Table A2 lists ranges, not values

```
R1 weight 0.2048 – 2.048
batch 512 – 1024
CLIP loss 0.2 – 1.0
```

What would settle it

4–5 sizes at matched compute, slopes against diffusion, and the share of runs that diverged.

Upscaler

8× text-conditioned upsampling of a 128 px input (Fig. 2), same eye crop from each method.

Input · 128 px | Real-ESRGAN · 0.06 s | SD Upscaler · 7.75 s | **GigaGAN 1K · 0.13 s** | **GigaGAN 4K · 3.66 s**

FID-10k ↓ · LAION 128 → 1024 (Table 4)

Real-ESRGAN	8.6
SD Upscaler	9.39
GigaGAN	1.54

FID-50k ↓ · ImageNet 64 → 256 (Table 5)

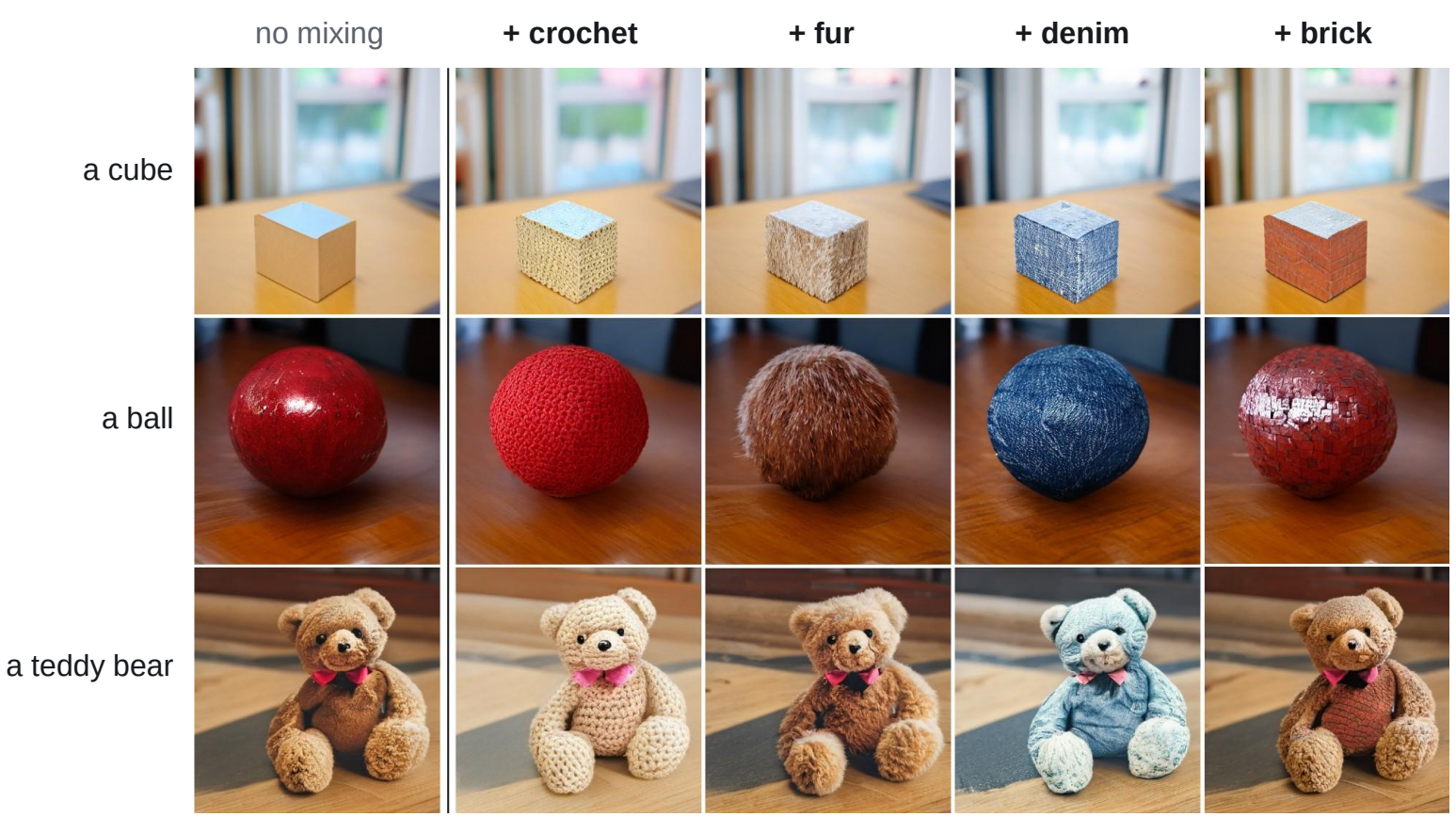
SR3 · 100 steps	5.2
LDM-4 · 100 steps	2.4
GigaGAN · 1 step	1.2

Caveats

- LPIPS is a loss and a metric
- Train Adobe Stock, test LAION
- No faithfulness check

Latent Space Steering

Prompt mixing (Fig. 8): layout from “a X on tabletop”, fine layers from “…texture of Y”.



Prompt interpolation, Fig. 7

Strengths and Weaknesses

GANs train stably at billion scale on web data 1.0B params and 0.98B images seen; one successful run (Table 2)	Strong
The upsampler is a fast, high-quality drop-in Tables 4–5 and Figs. 2–3; best FID in both settings	Strong
Much faster than diffusion and AR models 0.13 s holds; the margin depends on baseline and hardware	Supported
Text-to-image quality competitive with diffusion Lower COCO FID than SD-v1.5, but samples lag (Figs. 9, A9–A14)	Mixed
Each proposed component improves the model Cumulative, single-run ablation with regressions (Table 1)	Partial
Naively scaling StyleGAN is unstable One untuned wide run; tuning and size never separated	Partial
The latent space is disentangled and controllable Style and prompt mixing shown, never measured (Figs. 6–8)	Qualitative only
Quality has not saturated with model size Two points that also differ in steps and batch (Table 1)	Not shown

In Practice

Standalone GANs did not take over; adversarial objectives did, inside diffusion distillation.

Jan 2023 · StyleGAN-T
Concurrent large-scale text-to-image GAN

Mar 2023 · GigaGAN
This paper: a 1B-parameter GAN on web data

Nov 2023 · SDXL-Turbo
GAN loss distills diffusion to 1–4 steps

Apr 2024 · VideoGigaGAN
GigaGAN's upsampler extended to video

May 2024 · DMD2
Distribution matching plus a GAN loss

NeurIPS 2024 · R3GAN
A minimal, modern GAN baseline

A real scaling study

4–5 model sizes at matched data and compute; compare slopes with diffusion.

Generate in a latent space

Swap the 64 px pixel base for a VAE latent, as latent diffusion did.

Stronger text encoders

"…with a robotic half face" is ignored (Fig. 9)

Edit real images

Invert photos into w and t so mixing works beyond generated samples.