Google DeepMind

Rethinking Generative Image Pretraining: How Far Are We From Scaling Up Next-Pixel Prediction?

Xinchen Yan[1*] **Chen Liang**[2*] **Lijun Yu**[2] **Adams Wei Yu**[2] **Yifeng Lu**[2] **Quoc V. Le**[2]

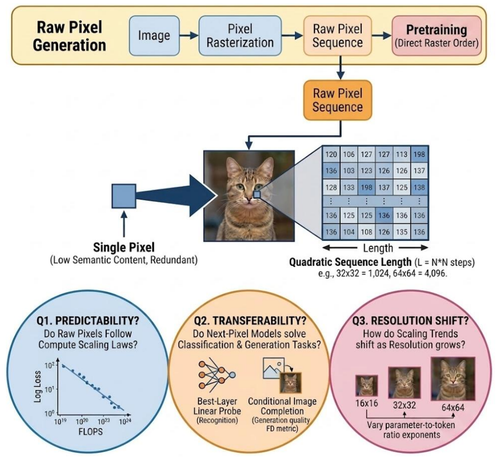
*Core contribution · [1]Work done at Google DeepMind · [2]Google DeepMind

ICML 2026
Seoul, South Korea

arXiv:2511.08704

1 Motivation

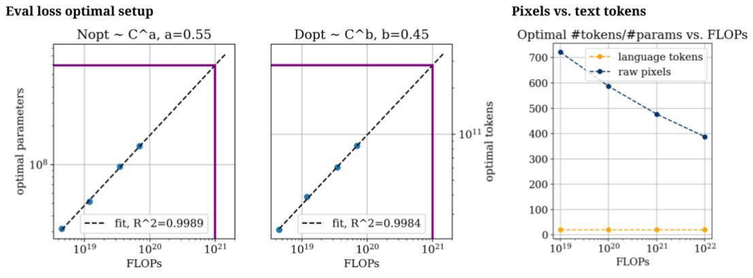
- Treat an image as a raw pixel sequence — end-to-end next-pixel prediction, **no tokenizer**.
- Hard: pixels carry little meaning, and sequences grow quadratically with resolution.
- Do scaling laws hold for pixels? Transfer to real tasks? Shift with resolution?



2 Experimental Setup

- Transformers from **10M to 449M** params (RoPE + GEGLU), trained on large-scale Internet image dataset JFT-300M.
- **IsoFLOP profiles:** 105 runs at 16^2, 32^2, 64^2, up to 2.8×10^{20} FLOPs.
- Metrics: next-pixel loss · linear-probe top-1 · completion Fréchet Distance.

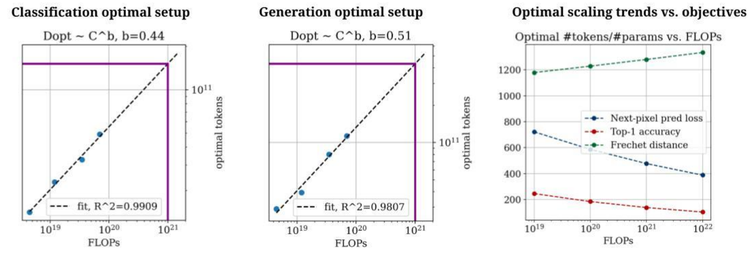
3 Scaling Laws Hold — but Pixels Are Data-Hungry



- Tight power-law fits for optimal model and data size ($R^2 \approx 0.999$).

On 32x32, **10–20×** more tokens per parameter than language models (>400 vs ~20)

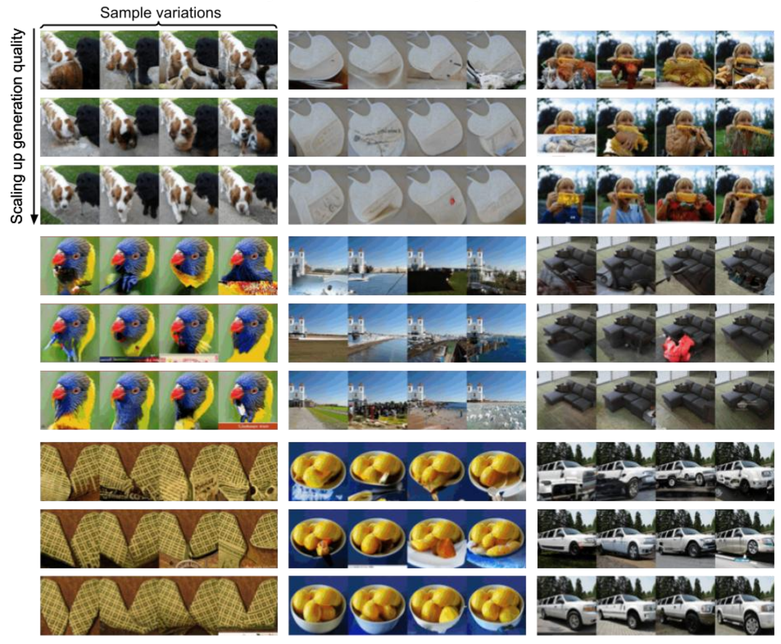
4 Optimal Scaling Is Task-Dependent



- Generation-optimal data grows **3–5× faster** than classification-optimal.
- The best scaling recipe depends on the task you care about.

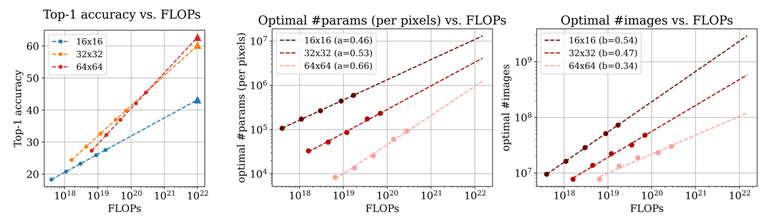
Verified: predicted 46.39% top-1 → measured 46.41% at 5× compute.

5 Generation Quality Scales Smoothly

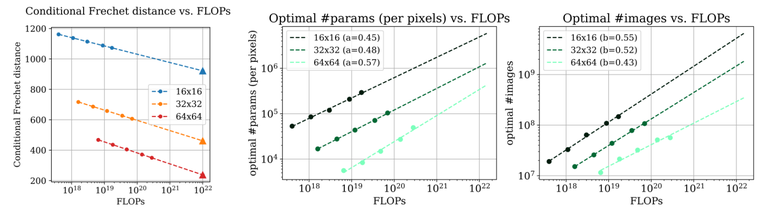


ImageNet 64×64 — top half is the prompt; bottom half is generated pixel by pixel. Bigger models (top → bottom), better completions.

6 Resolution Shifts the Optimal Strategy



(a) Image classification vs. FLOPs (b) Compute-optimal $N_{\text{opt}}^{\text{pp}}$ vs. FLOPs (left) and $D_{\text{opt}}^{\text{pp}}$ vs. FLOPs (right)



(c) Image completion vs. FLOPs (d) Compute-optimal $N_{\text{opt}}^{\text{pp}}$ vs. FLOPs (left) and $D_{\text{opt}}^{\text{pp}}$ vs. FLOPs (right)

- Classification saturates beyond 32×32; **generation keeps improving.**
- Higher resolution shifts compute toward model size, away from data.

7 Forecast: How Far Are We?

Predicted ImageNet top-1 (model size, #images)

Acc ↑ (N_{opt}, $D_{\text{opt}}^{\text{pp}}$) ‖ FLOPs	10^{22}	10^{23}	10^{24}
16 × 16 Projection	43.1% (2.78B, 2.33B)	48.8% (7.95B, 8.18B)	54.4% (22.73B, 28.6B)
32 × 32 Projection	60.1% (3.39B, 0.58B)	69.6% (11.6B, 1.4B)	79.0% (39.64B, 4.1B)
64 × 64 Projection	62.7% (3.88B, 0.1B)	73.8% (17.7B, 0.22B)	84.9% (80.72B, 0.5B)

Predicted completion Fréchet Distance (model size, #images)

FD ↓ (N_{opt}, $D_{\text{opt}}^{\text{pp}}$) ‖ FLOPs	10^{22}	10^{23}	10^{24}
16 × 16 Projection	921.7 (1.25B, 5.19B)	867.1 (3.51B, 18.51B)	812.6 (9.86B, 65.97B)
32 × 32 Projection	461.3 (1.1B, 1.47B)	394.0 (3.3B, 4.92B)	326.7 (9.90B, 16.4B)
64 × 64 Projection	237.4 (1.39B, 0.29B)	165.2 (5.24B, 0.77B)	92.9 (19.67B, 2.06B)

- **At 10^{24} FLOPs: 84.9% top-1** and FD 92.9 at 64×64.
- Needs only ~0.5B images — **datasets that size already exist.**

Compute — not data — is the bottleneck.

Pixel-by-pixel pretraining becomes feasible within five years.