Generative AI with Diffusion Models: Toward Efficient, Fair and Scalable Content Synthesis

Samridhhi Negi[1], Shubhra Komal[2], Vishal Kumar Sinha[3], Arup Gope[4], Dr. Savitha Choudhary[5]
[1,2,3,4] *Department of Computer Science and Engineering,*
Sir M. Visvesvaraya Institute of Technology, Bengaluru, Karnataka, India [5]*Associate Professor, Department of Computer Science and Engineering, Sir M. Visvesvaraya Institute of Technology, Bengaluru, Karnataka, India*

***Abstract*—Diffusion-based generative models have eclipsed GAN and VAE architectures in fidelity, robustness and mode coverage, but their iterative denoising chains impose steep com- putational and energy budgets. We present a complete stack that (i) fuses Transformer self-attention and convolution into a lightweight denoiser, (ii) compresses a 1 000-step teacher into a four-step student via progressive distillation, (iii) applies loss- aware pruning, mixed-precision kernels and adaptive timestep scheduling, and (iv) embeds a real-time bias-detection guardrail. Trained on a 550 k image–text corpus filtered for legal and ethical compliance, the system delivers an FID of 6.9 on MS- COCO while running 4.6× faster and 4.5× greener than a 50- step baseline, and it surpasses Stable Diffusion 1.5 by 1.2 FID at 38 % lower energy. Experiments on desktop GPUs, laptop GPUs and edge NPUs confirm viability for interactive design, AR filters and mobile creativity apps, moving diffusion models closer to trustworthy, resource-aware deployment.**

***Index Terms*—Generative AI, diffusion models, model com-pression, image synthesis, AI fairness, energy efficiency**

I. INTRODUCTION

Context. The last three years have witnessed an explosion of denoising diffusion probabilistic models (DDPMs) that beat GANs and VAEs in every major perceptual metric [1], [2]. Their advantages stem from stable log-likelihood training and explicit noise scheduling, yet practical adoption stalls where real-time latency or constrained power envelopes are manda- tory (e.g., mobile UIs, AR headsets, point-of-care medical imaging).

Challenge. A vanilla DDPM typically runs 50–250 reverse denoising steps per sample, each resembling a UNet forward pass: at 512px this is hundreds of milliseconds on a high- end GPU and seconds on edge hardware. The dual mandate is therefore clear: slash compute without harming fidelity—and do so while preventing demographic or toxic bias from large crawled datasets.

Contributions.

- Hybrid Transformer–Diffusion architecture that re-duces parameters by 18 % while widening receptive fields.
- Four-step progressive distillation augmented with loss- aware channel pruning, float16 kernels and timestep adap-

tation, yielding 54 ms latency at 512px on an RTX 3080- Laptop.

- Real-time fairness guardrail built on CLIP embeddings, adding < 3 ms overhead with > 90 % sensitive-content recall.
- Comprehensive evaluation across quality, efficiency, fairness and edge deployment, including an ablation study disentangling each optimization.

II. BACKGROUND AND RELATED WORK

A. Foundations of Diffusion Modeling

DDPMs treat generation as gradual noise removal; improved variants tighten ELBO bounds [3], frame the process as SDEs [4], or accelerate sampling with deterministic solvers such as DDIM and DPMSolver++ [5], [6].

B. Latent Diffusion and Large-Scale Text Conditioning

Rombach *et al.* compress pixel-space diffusion into a learned latent code, enabling 768^2 image synthesis on

gaming GPUs [7]. GLIDE, DALL·E 2, Imagen and Stable Diffusion connect these backbones with transformer encoders to follow arbitrary text prompts [10]– [12].

C. Efficiency: Distillation and Consistency

Progressive distillation collapses hundreds of steps into 4–8 with marginal fidelity loss [8]. Consistency models push to single-step generation via self-supervised objectives [9], albeit with quality trade-offs in high-resolution regimes.

D. Fairness and Safety in Diffusion

FairDiffusion, BiasAudit and related studies highlight gen- der and ethnic skew in synthetic faces and occupational stereo- types [13], [14]. Techniques range from post-hoc filtering to dataset balancing; our guardrail opts for lightweight in-loop scoring to meet real-time budgets.

III. PROPOSED FRAMEWORK

A. Problem Statement

Formally, given a condition c (text, low-res image, or multi-
modal embedding), generate $\hat{x} \in \mathrm{R}^{H\times W\ \times 3}$ that minimizes
perceptual divergence from the data manifold, under latency $L^{\max}$ and energy $E^{\max}$ constraints, while satisfying a fairness risk score $r(c, \hat{x}) \leq \tau$.

B. Hybrid Transformer–Diffusion Denoiser

Each denoising stage alternates depth-wise conv blocks (local detail) with 8-head self-attention (global context). FiLM layers modulate channels using T5-XL text embeddings. Weight sharing across blocks further trims memory.

C. Optimization Pipeline

1) Progressive Distillation: teacher trajectories (1 000 steps) supervise a 4-step student via MSE + noise-pred loss.
2) Loss-Aware Pruning: group-Lasso removes 30 % of filters whose gradients correlate least with FID improve- ment.
3) Mixed Precision & FP8 Kernels: float16 ops every- where except LayerNorm; TensorRT FP8 accelerates attention by 1.3×.
4) Adaptive Step Scheduling: a tiny policy net allocates extra denoise budget to high-entropy spatial regions.

D. Bias-Detection Guardrail

Three CLIP-derived classifiers flag nudity, profanity and demographic imbalance. Trajectories exceeding threshold τ =
0.4 are either resampled (if early) or pixel-masked (if late), keeping FID drop ¡0.02.

IV. EXPERIMENTAL SETUP

A. Hardware & Runtime Environment

Training: 4× RTX 4090, PyTorch 2.2, CUDA 12.4. Deploy- ment: RTX 3080-Laptop, Apple M2 NPU (16-core), Snap- dragon 8-Gen-3 (Hexagon NPU).

B. Dataset and Pre-processing

We merge LAION-Aesthetics v2, OpenImages, VisualGenome to 550 k pairs. A filter pipeline removes water- marks, synthetic text, and explicit content. Images are center- cropped/ resized to 512^2 then VQ-VAE-2 encoded to $64^2\times4$ latent tensors; captions are tokenized with SentencePiece (32 k vocab) and embedded via frozen T5-XL.

C. Evaluation Metrics

Quality: FID, IS, CLIP-FID, CLIP-R. *Efficiency:* single- sample latency, FPS, energy (J/img) via Nvidia-SMI + NPU counters. *Fairness:* KL divergence between gender/ethnicity proportions in prompts vs. generations; toxic-content recall.

V. RESULTS AND DISCUSSION

A. Quantitative Results

Table I benchmarks the four-step student against the 50-step teacher and Stable Diffusion 1.5 on the MS-COCO 30 k split.

TABLE I
QUALITY–EFFICIENCY TRADE-OFF (512PX, RTX 3080-LAPTOP)

Model	FID ↓	Latency (ms) ↓	Energy (J) ↓
50-step Teacher	6.71	248	81.3
Stable Diffusion 1.5	8.12	87	27.4
Ours (4-step)	6.89	54	18.0

B. Ablation Study

Latency rises to 165ms if distillation is removed (25- step student); energy climbs 31 % when mixed

precision is disabled. Excluding the guardrail leaves FID unchanged but raises fairness KL divergence from 0.06 to 0.21—well above policy threshold.

C. Qualitative Assessment

Figure omitted for brevity. Human raters (n=25) preferred our samples to Stable Diffusion by 62 % and found no significant realism gap relative to the teacher (p¿0.1).

D. Edge Deployment

Apple M2 delivers 0.9 FPS (8 W system draw); Snapdragon 8-Gen-3 yields 0.4 FPS (5 W). Both sustain interactive AR or thumbnail use-cases.

ACKNOWLEDGMENT

We thank the Department of Computer Science and En- gineering, Sir M. Visvesvaraya Institute of Technology, for compute resources.

REFERENCES

[1] J. Ho et al., “Denoising diffusion probabilistic models,” NeurIPS, 2020.
[2] P. Dhariwal and A. Nichol, “Diffusion models beat GANs on image synthesis,” NeurIPS, 2021.
[3] A. Nichol and P. Dhariwal, “Improved DDPMs,” ICML, 2021.
[4] Y. Song et al., “Score-based generative modeling via SDEs,” ICLR, 2021.
[5] J. Song et al., “DDIM,” ICLR, 2021.
[6] C. Lu et al., “DPMSolver++,” arXiv:2206.00927, 2022.
[7] R. Rombach et al., “Latent diffusion models,” CVPR, 2022.
[8] T. Salimans and J. Ho, “Progressive distillation,” arXiv:2202.00512, 2022.
[9] Y. Song et al., “Consistency models,” NeurIPS, 2023.
[10] A. Nichol et al., “GLIDE,” ICML, 2022.
[11] A. Ramesh et al., “DALL·E 2,” arXiv:2204.06125, 2022.
[12] C. Saharia et al., “Imagen,” ICML, 2022.
[13] P. Schramowski et al., “Fairness in diffusion,” AAAI, 2023.
[14] C. Xu et al., “FairDiffusion,” arXiv:2402.00333, 2024.