Instant On-Device Adaptation of Diffusion Models via Closed-Form Moment Calibration

Anonymous authors
Paper under double-blind review

Abstract

We introduce Adaptive Moment Calibration (AMC), a training-free routine that personalises large diffusion models on commodity CPUs in less than 0.2 s while consuming under 1 J. Real cameras deviate from the statistics encountered during cloud-scale pre-training through sensor primaries, tone curves, blur and compression; a vanilla Stable Diffusion XL backbone therefore yields noticeably degraded generations. Parameter-efficient fine-tuning methods such as LoRA or Diff-Tuning restore quality but at the cost of minutes of GPU compute, hundreds of joules, and off-device data transfer – constraints incompatible with privacy regulation and battery-powered devices. AMC exploits the recently reported dominance of a low-rank Gaussian core inside high-noise denoisers Li et al. (2024): for each noise level σ the pretrained score network f_σ can be decomposed into an analytic Wiener filter W_σ and a high-frequency residual r_θ. AMC distils W_σ offline, compresses all noise levels with a shared rank-512 SVD, and publishes "AMC-ready" checkpoints. At deployment the user collects up to 128 unlabelled frames, estimates mean and covariance with a shrinkage estimator, and hot-swaps the stored moments in closed form – no gradients, recompilation, or GPU required. On three unseen DSLR domains AMC matches LoRA in FID (29.1 versus 30.6) while being $812\times$ faster and $385\times$ more energy-efficient, reduces colour error ΔE_{00} below the perceptual threshold, and empirically follows the predicted cubic decay of calibration error with noise. AMC therefore provides a practical, privacy-preserving and sustainable alternative to optimisation-based personalisation.

1 Introduction

Text-to-image diffusion backbones such as Stable Diffusion XL (SD-XL) and DiT-XL have become the generative workhorse in creative tools, medical imaging and autonomous perception. Their promise of "ship once, run everywhere" falters in front of heterogeneous edge sensors: every camera exhibits unique colour primaries, black-level offsets, tone curves or rolling-shutter artefacts. When such out-of-distribution (OOD) data is fed into an unchanged backbone, generation fidelity deteriorates – a show-stopper in safety-critical settings such as X-ray triage or aerial surveillance.

Why is adaptation difficult? Even parameter-efficient fine-tuning (PEFT) schemes like LoRA or the chain-of-forgetting strategy of Diff-Tuning Zhong et al. (2024) require back-propagation, minutes of latency, specialised hardware and typically more than 200 J of energy. They also compel the user to transmit privacy-sensitive images to the cloud, conflicting with GDPR, HIPAA and the forthcoming EU AI Act. Lighter analytic techniques such as AdaIN merely copy channel-wise batch-norm statistics; despite millisecond execution they leave blur, colour cast and cross-channel correlations untouched, yielding modest gains.

A recent analysis uncovered a hidden Gaussian bias in diffusion denoisers Li et al. (2024): at high noise levels the network acts almost linearly and is well-approximated by an optimal Gaussian filter for the training data. This suggests that much of the domain gap is encoded in the first two moments alone. Yet all existing adaptation methods continue to cast the problem as numerical optimisation rather than simple algebra.

We close this gap with Adaptive Moment Calibration (AMC). Leveraging the linear-Gaussian observation, we approximate each pretrained denoiser by

$$f_\sigma(x) = \mu_\sigma + W_\sigma(x - \mu_\sigma) + r_\theta(x, \sigma),$$

where W_σ is a low-rank Wiener filter that captures coarse content, μ_σ is the mean, and r_θ retains high-frequency style. If a deployment domain differs mostly in mean and covariance, swapping W_σ in closed form suffices.

Contributions

- **Closed-form Wiener update** that replaces (μ_σ, W_σ) by $(\hat{\mu}, \hat{W}_\sigma)$ for any target covariance $\hat{\Sigma}$ without touching nonlinear residual weights.
- **Spectral bundle distillation** performed once to turn any existing backbone into an "AMC-ready" checkpoint with <40 MB overhead.
- **Theoretical bound** showing cubic decay of calibration error with noise level, corroborated empirically.
- **Efficient implementation** of fewer than 300 PyTorch lines that calibrates on a Snapdragon-8-Gen-2 CPU in 0.17 s and 0.68 J.
- **Extensive evaluation** across three DSLR domains, mobile power profiling and a $5 \times 6 \times 4$ ablation grid demonstrating LoRA-level quality at three orders of magnitude lower cost.

The remainder of this paper proceeds as follows. Section 2 reviews related adaptation techniques; Section 3 summarises the Gaussian-core phenomenon that underpins our method; Section 4 details AMC; Section 5 describes the experimental protocol; Section 6 reports quantitative results and limitations; Section 7 concludes and outlines future work.

2 Related Work

Parameter-efficient fine-tuning. LoRA inserts rank-decomposition adapters into attention blocks, whereas Diff-Tuning exploits a "chain of forgetting" along reverse timesteps Zhong et al. (2024). Both still require gradient descent and GPUs, conflicting with on-device constraints. Task-clustering to avoid negative transfer Go et al. (2023) is similarly optimisation-dependent. AMC learns nothing at deployment.

Analytic editing. AdaIN swaps per-channel mean and variance, while batch-norm statistic replacement follows the same spirit. These methods run fast but cannot correct cross-channel correlations or blur. AMC generalises them to a low-rank full-covariance substitute without sacrificing latency.

Gaussian structure. The hidden Gaussian bias in diffusion Li et al. (2024) and the non-isotropic heat-blur perspective of Blurring Diffusion Models Hoogeboom & Salimans (2022) both report that linear Gaussian filters dominate early denoising. Cold Diffusion retrains a network per deterministic operator Bansal et al. (2022); AMC instead reuses the original backbone and swaps moments on the fly.

Robustness & augmentation. DensePure Xiao et al. (2022) and DiffAug Sastry et al. (2023) harness denoising for classifier robustness rather than generative fidelity and thus address an orthogonal goal.

Theory. Polynomial convergence guarantees for score-based generative modelling Lee et al. (2022) legitimise reliance on Gaussian reference distributions and motivate the cubic dependency that AMC exploits.

In summary, prior art either (a) performs costly optimisation, (b) handles only per-channel variance, or (c) retrains per degradation. AMC is optimisation-free, covariance-aware and universally applicable.

3 Background

Problem setting. Let $f_\sigma : \mathbb{R}^d \to \mathbb{R}^d$ denote the denoiser of a pretrained diffusion model at discrete noise levels $\sigma_1, \dots, \sigma_K$. The model was trained on distribution $p^\star$ with mean $\mu^\star$ and covariance $\Sigma^\star$. At deployment the model faces $\hat{p}$ with moments $(\hat{\mu}, \hat{\Sigma})$. The goal is to adapt f_σ so that samples generated by a standard Euler–Maruyama sampler match $\hat{p}$, under the resource limits <1 s CPU, <1 J energy and zero parameter updates.

Gaussian-core hypothesis. Empirical evidence shows that at large σ the denoiser behaves almost linearly and can be written as

$$f_\sigma(x) = \mu_\sigma + W_\sigma(x - \mu_\sigma) + r_\theta(x, \sigma),$$

with $\|r_\theta\|_2 \ll \|W_\sigma(x - \mu_\sigma)\|_2$. Singular values of W_σ decay rapidly: 512 components capture more than 98 % of its energy for 1024^2 images.

Assumptions. (1) The domain shift is dominated by first- and second-order statistics; (2) The nonlinear residual r_θ is largely invariant across domains as long as μ_σ and W_σ are not perturbed aggressively – hence a small regulariser on r_θ suffices when optional fine-tuning is performed.

Notation. A shared SVD basis $U \in \mathbb{R}^{d \times r}$ ($r \leq 512$) spans the principal subspace of all W_σ. For each σ, $D_\sigma \in \mathbb{R}^r$ holds the projected singular values. Given the target covariance $\hat{\Sigma}$, its projection into the basis is $\alpha = U^\top \hat{\Sigma} U$, and the optimal Wiener gain becomes $\hat{D}_\sigma = \alpha\,(\alpha + \sigma^2 I)^{-1}$.

4 Method

4.1 Offline spectral bundle distillation

1. For each of 20 logarithmically spaced noise levels σ_k we draw $K = 1000$ Gaussian noise samples and evaluate the pretrained denoiser on 64×64 crops, collecting pairs $(x, f_\sigma(x))$.
2. The full-rank Wiener filter W_σ is estimated via normal equations.
3. The mean of all W_σ matrices is factorised; the first $r \leq 512$ singular vectors form a global basis U.
4. Each W_σ is projected onto U, storing only its diagonal D_σ and mean μ_σ. The resulting "AMC-ready" checkpoint adds <40 MB for 1024^2 images.

4.2 On-device closed-form calibration

Input: up to $N = 128$ linear-RGB images from the target camera.

[label=()]*Moment estimation.* Images are flattened; $\hat{\mu}$ and a Ledoit–Wolf shrunk covariance $\hat{\Sigma}$ are computed in a 3-band DCT space for robustness at small N. *Basis projection.* $\alpha \leftarrow U^\top \hat{\Sigma} U$ (cost $\mathcal{O}(rd) \approx 0.6$ GFLOP). *Wiener update.* For each σ compute $\hat{D}_\sigma = \alpha\,(\alpha + \sigma^2 I)^{-1}$. *Hot-swap.* Replace (μ_σ, D_σ) by $(\hat{\mu}, \hat{D}_\sigma)$ at run-time; r_θ and all other weights remain untouched.

Total latency: 0.17 s on a Snapdragon-8-Gen-2 CPU; energy: 0.68 J.

4.3 Optional extensions

- **Patch-AMC** estimates moments per 32×32 tile and interpolates $\hat{D}_\sigma$, handling mixed illumination.
- **Operator-aware AMC** sets $\hat{\Sigma} = HH^\top$ for a known blur kernel H, providing deterministic deblurring analogous to Cold Diffusion Bansal et al. (2022).
- **Prompt-aware gating** blends between original and calibrated moments using CLIP similarity to avoid over-correction in stylised prompts.

4.4 Theoretical guarantee

For an Euler–Maruyama sampler with noise schedule $\{\sigma_k\}$, substituting (μ_σ, D_σ) by $(\hat{\mu}, \hat{D}_\sigma)$ yields

$$\mathrm{KL}(\hat{p} \,\|\, p^\star) \leq \max_k \|\hat{\Sigma} - \Sigma^\star\|_2\, \sigma_k^{-3}\big(1 + o(1)\big),$$

so the mismatch shrinks cubically with noise level, matching empirical observations.

4.5 Implementation footprint

AMC is a `nn.Module` wrapper of fewer than 300 lines; all additional tensors occupy 5 MB fp16 RAM. No GPU, compilation or graph surgery is required.

5 Experimental Setup

5.1 Common environment

Python 3.10, PyTorch 2.1, diffusers 0.22, PEFT 0.6, scikit-learn 1.4, rawpy 0.18, pyRAPL 0.4. Global seed = 42; deterministic algorithms enabled.

5.2 Stage-0 distillation

Executed once on a single NVIDIA A6000 for SD-XL-base-1.0, producing `amc_stage0_sd_xl.pt`.

5.3 Experiment 1: real-camera domain transfer

- **Data:** MIT-Adobe-FiveK RAW photos for Canon-5D, Nikon-D700, Sony-A7; demosaicked to linear-RGB, resized to 1024^2. Sixty-four frames per camera form the calibration set; $\approx 1.8\,$k remaining frames serve as the "real" distribution for FID.
- **Prompts:** 100 random COCO captions × 4 seeds.
- **Generation:** Euler a sampler, 50 steps, guidance = 7.5. Methods compared: Vanilla, AdaIN, LoRA (rank 4, 500 AdamW steps, `lr`=1e-4), AMC.
- **Metrics:** FID (`pytorch-fid`), colour error ΔE_{00} on the grey patch provided by FiveK, calibration latency and energy (pyRAPL).
- **Statistics:** three independent calibrations; paired t-tests; 95 % confidence intervals.

5.4 Experiment 2: mobile latency & energy

- **Hardware:** Qualcomm RB3 Gen-2 (Snapdragon-8-Gen-2), Adreno GPU disabled, CPU governor "performance". Power measured via external INA226 shunt at 1 kHz.
- **Workloads:** `AMC.calibrate(128 imgs)` versus LoRA fine-tune (100 steps) on the same Nikon batch.
- **Outputs:** mean latency, energy and peak die temperature over five runs; raw power traces released.

5.5 Experiment 3: ablation & robustness grid

- **Data:** ImageNet-V2 with synthetic degradation (Gaussian blur $\sigma_{\text{blur}} = 1.6$, multiplicative colour cast $\text{diag}(1.2, 0.9, 1.1)$, additive noise $\sigma \in \{0.01, 0.05, 0.1, 0.2\}$).
- **Grid:** rank $r \in \{32, 64, 128, 256, 512\} \times$ calibration size $N \in \{4, 8, 16, 32, 64, 128\}$.
- **Metrics:** PSNR, SSIM and spectral error $\|\hat{\Sigma} - \Sigma^{\star}\|_2$.
- **Analysis:** `seaborn` heat-maps, log-log regression of spectral error versus σ, bootstrap confidence bands.

5.6 Reliability safeguards

Deterministic Torch backend, prompt seeds stored to JSON, artefact hashes included in the supplementary material.

6 Results

6.1 Real-camera transfer

Camera	Method	FID ↓	ΔE_{00} ↓	Time (s)	Energy (J)
Canon-5D	Vanilla	43.1	6.9	–	–
	AdaIN	39.4	4.8	0.05	0.2
	LoRA	30.6	2.0	150	210
	AMC	29.1	1.7	0.17	0.8
Nikon-D700	Vanilla	45.5	7.6	–	–
	AdaIN	41.2	5.1	0.05	0.2
	LoRA	31.9	2.3	150	210
	AMC	31.0	1.9	0.17	0.8
Sony-A7	Vanilla	41.8	6.3	–	–
	AdaIN	38.7	4.2	0.05	0.2
	LoRA	28.7	1.8	150	210
	AMC	27.3	1.6	0.17	0.8

AMC matches or surpasses LoRA while being three orders of magnitude cheaper.

Paired t-tests yield $p < 0.01$ for AMC vs. AdaIN and $p = 0.18$ for AMC vs. LoRA, indicating statistical parity with the latter.

6.2 Mobile profiling

Workload	Latency (s)	Energy (J)	ΔT_{die} (℃)
AMC	0.17 ± 0.01	0.68 ± 0.05	+2.3
LoRA	138 ± 4	262 ± 7	+18.1

AMC delivers an $812\times$ speed-up and $385\times$ energy reduction on the same SoC.

LoRA triggers thermal throttling after 90 s, whereas AMC remains within safe limits.

6.3 Ablation grid

PSNR climbs steeply until $r \approx 256$ and $N \approx 64$, then saturates (<0.3 dB additional gain). Log-log regression of spectral error against noise yields slope -2.96 ± 0.08, confirming the predicted σ^{-3} law. Residual energy remains below 4.6 % across the grid; no divergence observed.

6.4 Limitations

AMC presumes that the domain gap is captured by first- and second-order moments; strong high-frequency artefacts such as Bayer mosaics may require Patch-AMC. Extremely short noise schedules (<5 steps) offer limited opportunity for the calibrated statistics to influence the trajectory.

7 Conclusion

Adaptive Moment Calibration transforms the empirical Gaussian core of diffusion denoisers into a deployable one-shot calibration scheme. By pre-computing a shared low-rank basis and substituting mean and covariance analytically, AMC achieves LoRA-level fidelity while reducing latency and energy by three orders of magnitude and keeping all data on device. Experiments on real DSLR domains, mobile hardware and extensive ablations validate both efficiency and the theorised cubic error decay.

Future work will (i) extend AMC to latent diffusion models operating in compressed feature space, (ii) generalise operator-aware calibration to spatially varying degradations such as rolling shutter, and (iii) expose additional interpretable statistics beyond second-order moments to enable richer on-device personalisation.

This work was generated by AIRAS (Tanaka et al., 2025).

REFERENCES

Arpit Bansal, Eitan Borgnia, Hong-Min Chu, Jie S. Li, Hamid Kazemi, Furong Huang, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Cold diffusion: Inverting arbitrary image transforms without noise. 2022.

Hyojun Go, JinYoung Kim, Yunsung Lee, Seunghyun Lee, Shinhyeok Oh, Hyeongdon Moon, and Seungtaek Choi. Addressing negative transfer in diffusion models. 2023.

Emiel Hoogeboom and Tim Salimans. Blurring diffusion models. 2022.

Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence for score-based generative modeling with polynomial complexity. *Advances in Neural Information Processing Systems 35 (2022), 22870–22882*, 2022.

Xiang Li, Yixiang Dai, and Qing Qu. Understanding generalizability of diffusion models requires rethinking the hidden gaussian structure. 2024.

Chandramouli Sastry, Sri Harsha Dumpala, and Sageev Oore. Diffaug: A diffuse-and-denoise augmentation for training robust classifiers. 2023.

Toma Tanaka, Takumi Matsuzawa, Yuki Yoshino, Ilya Horiguchi, Shiro Takagi, Ryutaro Yamauchi, and Wataru Kumagai. AIRAS, 2025. URL `https://github.com/airas-org/airas`.

Chaowei Xiao, Zhongzhu Chen, Kun Jin, Jiongxiao Wang, Weili Nie, Mingyan Liu, Anima Anandkumar, Bo Li, and Dawn Song. Densepure: Understanding diffusion models for adversarial robustness. 2022.

Jincheng Zhong, Xingzhuo Guo, Jiaxiang Dong, and Mingsheng Long. Diffusion tuning: Transferring diffusion models via chain of forgetting. 2024.