# ADAPTIVE MOMENTS ARE SURPRISINGLY EFFECTIVE FOR PLUG-AND-PLAY DIFFUSION SAMPLING

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Guided diffusion sampling relies on approximating intractable likelihood scores, which introduces significant noise into the sampling dynamics. We propose using adaptive moment estimation to stabilize these noisy likelihood scores during sampling. Despite its simplicity, our approach achieves state-of-the-art results on image restoration and class-conditional generation tasks, outperforming more complicated methods, which are often computationally more expensive. We provide empirical analysis of our method on both synthetic and real data, demonstrating that mitigating gradient noise through adaptive moments offers an effective way to improve alignment.

## 1 INTRODUCTION

Diffusion models (Sohl-Dickstein et al., 2015; Song & Ermon, 2019; Ho et al., 2020) have become one of the most successful generative modeling approaches, achieving state-of-the-art results in text-to-image synthesis (Rombach et al., 2022; Ramesh et al., 2022; Saharia et al., 2022a), image-to-image translation (Saharia et al., 2022b), audio generation (Kong et al., 2020; Liu et al., 2023), video synthesis (Ho et al., 2022; Brooks et al., 2024), and molecular design (Hoogeboom et al., 2022; Watson et al., 2023).

Plug-and-play conditional generation enables sampling from a conditional distribution $p(x|y)$ using a diffusion model trained only on the marginal distribution $p(x)$. These methods guide the sampling process toward desired conditions $y$ without task-specific training. While some approaches like Classifier Guidance (Dhariwal & Nichol, 2021) train time-aware models directly on the diffusion latents, many plug-and-play methods leverage existing models that operate on clean data – whether analytical forward operators for inverse problems or pre-trained classifiers – making them highly flexible for diverse applications.

The plug-and-play guidance literature has evolved from early methods like Diffusion Posterior Sampling (DPS) (Chung et al., 2022) to increasingly sophisticated techniques. Recent work such as Universal Guidance for Diffusion Models (UGD) (Bansal et al., 2023) and Training-Free Guidance (TFG) (Ye et al., 2024) compose multiple algorithmic components, combining gradient computations in both latent and data spaces, Monte Carlo approximations, and iterative refinement procedures.

At the heart of plug-and-play guidance lies the challenge of incorporating the desired condition into the diffusion sampling process. Sampling in diffusion models can be understood as annealed Langevin dynamics using the score function $\nabla_{x_t} \log p(x_t)$ (Song & Ermon, 2019; Karras et al., 2022). For conditional sampling, Bayes' rule gives us:

$$\underbrace{\nabla_{x_t} \log p(x_t|y)}_{\text{Posterior Score}} = \underbrace{\nabla_{x_t} \log p(x_t)}_{\text{Prior Score}} + \underbrace{\nabla_{x_t} \log p(y|x_t)}_{\text{Likelihood Score}} \tag{1}$$

While the prior score is provided by the unconditional diffusion model, the likelihood score $\nabla_{x_t} \log p(y|x_t)$ is intractable to compute directly, necessitating approximation strategies (Chung et al., 2022; He et al., 2023; Song et al., 2023).

Prior work as predominately studied improving the likelihood score approximation. Orthogonal to this, we investigate whether information from earlier sampling steps can help mitigate approximation errors in later sampling steps. Instead of developing more sophisticated approximation methods,

we use adaptive moment estimation – a technique from stochastic optimization (Kingma, 2014) – to stabilize the noisy guidance gradients that arise in plug-and-play methods.

Our approach maintains exponential moving averages of the first and second moments of the likelihood gradients across sampling steps, effectively dampening noise while preserving the guidance signal. Despite its simplicity, this modification yields substantial improvements: DPS augmented with adaptive moments (AdamDPS) outperforms state-of-the-art methods across diverse benchmarks.

We also examine how task difficulty affects the relative performance of different methods. Existing evaluations typically employ relatively mild degradations that provide a strong guidance signal– for instance, 4x super-resolution or moderate blur kernels. We demonstrate that as task difficulty increases (e.g., 16x super-resolution or severe degradations), the performance landscape shifts significantly. Many recently proposed approaches that have been shown to outperform DPS in simpler settings degrade rapidly under challenging conditions and ultimately underperform the simpler DPS method. We find that adaptive moment estimation consistently improves upon DPS across difficulty levels and outperforms a comprehensive selection of recent methods.

Our contributions are: (i) we demonstrate that adaptive moment estimation can substantially improve plug-and-play guidance methods, achieving state-of-the-art results with minimal added complexity; (ii) we provide empirical analysis through synthetic Gaussian mixture experiments that illustrate how our method stabilizes noisy gradients; and (iii) we reveal that task difficulty significantly impacts relative method performance, suggesting the need for more comprehensive evaluation protocols. Our extensive empirical results across diverse tasks demonstrate its effectiveness and robustness.

## 2 BACKGROUND

**Diffusion Models and Score-Based Sampling.** Diffusion models (Sohl-Dickstein et al., 2015; Song & Ermon, 2019; Ho et al., 2020) learn to generate samples from a data distribution $p(x)$ by reversing a gradual noising process. The forward process progressively corrupts data into Gaussian noise over time $t \in [0, T]$, defining marginal distributions $p(x_t|x)$ where $x_t = \alpha_t x + \sigma_t \epsilon$ with $\epsilon \sim \mathcal{N}(0, \mathbb{I})$. The noise schedule parameters $\alpha_t$ and $\sigma_t$ satisfy $\alpha_t^2 + \sigma_t^2 = 1$, ensuring variance preservation. A neural network $\epsilon_\theta$ is trained to predict the noise at each timestep by minimizing:

$$\mathcal{L}_\epsilon(\theta) = \mathbb{E}_{t,x,\epsilon}\left[\|\epsilon_\theta(x_t, t) - \epsilon\|_2^2\right] \tag{2}$$

This objective implicitly trains the network to approximate the score function $\nabla_{x_t} \log p(x_t) \approx -\epsilon_\theta(x_t, t)/\sigma_t$. Importantly, the predicted noise also provides a minimum mean squared error estimate of the clean data:

$$x_{0|t} = \mathbb{E}[x_0|x_t] = \frac{x_t - \sigma_t \epsilon_\theta(x_t, t)}{\alpha_t} \tag{3}$$

under optimal training (Efron, 2011). This clean data estimate, denoted $x_{0|t}$, becomes crucial for plug-and-play guidance methods.

Sampling proceeds via *annealed Langevin dynamics* (Song & Ermon, 2019; Karras et al., 2022), iteratively denoising from $x_T \sim \mathcal{N}(0, \mathbb{I})$:

$$x_s = \frac{1}{\alpha_{t|s}} x_t + \frac{\sigma_{t|s}^2}{\alpha_{t|s}} \nabla_{x_t} \log p(x_t) + \frac{\sigma_{t|s}\sigma_s}{\sigma_t} \epsilon \tag{4}$$

where $\alpha_{t|s} = \alpha_t/\alpha_s$, $\sigma_{t|s}^2 = \sigma_t^2 - \alpha_{t|s}^2 \sigma_s^2$, and $\epsilon \sim \mathcal{N}(0, \mathbb{I})$ (Kingma et al., 2021). This process gradually transitions from high noise to low noise sampling as we move from $t = T$ to $t = 0$.

**The Plug-and-Play Guidance Challenge.** For conditional generation $p(x|y)$, Bayes' rule decomposes the posterior score as:

$$\nabla_{x_t} \log p(x_t|y) = \nabla_{x_t} \log p(x_t) + \nabla_{x_t} \log p(y|x_t) \tag{5}$$

The prior score is well approximated by the diffusion model $\epsilon_\theta$. The likelihood score $\nabla_{x_t} \log p(y|x_t)$ requires intractable marginalization to compute from a model for $p(y|x)$:

$$p(y|x_t) = \int p(y|x) p(x|x_t) dx \tag{6}$$

Plug-and-play methods often approximate this likelihood score using the denoised estimate: $\nabla_{x_t} \log p(y|x_t) \approx \nabla_{x_t} \log p(y|x_{0|t})$ which introduces error.

**Algorithm 1** Adam Diffusion Posterior Sampling

**Require:** Diffusion Model $\theta$, Guidance Model $f_\phi$, Condition $y$, Guidance Strength $\rho$, Sampling Timesteps $t_n, \ldots, t_0 \subseteq \mathcal{T}$, First Moment $m = \mathbf{0}$, Second Moment $v = \mathbf{0}$, Adam Step $k = 0$, First Moment Exponential Decay Rate $\beta_1$, Second Moment Exponential Decay Rate $\beta_2$

$x_{t_n} \sim \mathcal{N}(0, \mathbb{I})$
**for** $t = t_n, \ldots, t_1$ **do**
  $\Delta_t = \nabla_{x_t} \mathcal{L}(f_\phi(x_{0|t}, y)$
  $\tilde{\Delta}_t, m, v, k = \text{AdaptiveMomentEstimate}(\Delta_t, m, v, k, \beta_1, \beta_2)$
  $x_s = \text{Sample}(x_{0|t}, x_t, t, s) + \rho \tilde{\Delta}_t / \alpha_{t|s}$
**end for**
**return** $x_{t_0}$

**Algorithm 2** Adam Classifier Guidance

**Require:** Diffusion Model $\theta$, Guidance Model $f_\phi$, Condition $y$, Guidance Strength $\rho$, Sampling Timesteps $t_n, \ldots, t_0 \subseteq \mathcal{T}$, First Moment $m = \mathbf{0}$, Second Moment $v = \mathbf{0}$, Adam Step $k = 0$, First Moment Exponential Decay Rate $\beta_1$, Second Moment Exponential Decay Rate $\beta_2$

$x_{t_n} \sim \mathcal{N}(0, \mathbb{I})$
**for** $t = t_n, \ldots, t_1$ **do**
  $\Delta_t = \nabla_{x_t} \mathcal{L}(f_\phi(x_t, t), y)$
  $\tilde{\Delta}_t, m, v, k = \text{AdaptiveMomentEstimate}(\Delta_t, m, v, k, \beta_1, \beta_2)$
  $x_s = \text{Sample}(x_{0|t} + \rho \tilde{\Delta}_t \sigma_t^2 / \alpha_t, x_t, t, s)$
**end for**
**return** $x_{t_0}$

## 3 Guidance with Adaptive Moment Estimation

Plug-and-play conditional generation requires approximating the intractable likelihood score $\nabla_{x_t} \log p(y|x_t)$ to guide the diffusion sampling process toward desired conditions. In practice, this score is typically approximated by the gradient of a loss function $\mathcal{L}$ that measures alignment between the generated sample and the condition $y$.

**Existing Likelihood Score Approximations.** Two prominent approaches have emerged for this approximation, distinguished by the domain in which the guidance function operates. Diffusion Posterior Sampling (DPS) (Chung et al., 2022), a widely adopted plug-and-play method, leverages existing models $f_\phi : \mathcal{X} \to \mathcal{Y}$ that operate on clean data. DPS approximates the likelihood score as:

$$\nabla_{x_t} \log p(y|x_t) \approx \nabla_{x_t} \mathcal{L}(f_\phi(x_{0|t}), y) \tag{7}$$

where $x_{0|t}$ is the predicted clean sample from Equation 3. This approach enables the use of pre-trained models without modification but requires backpropagation through the denoising network.

Alternatively, Classifier Guidance (CG) (Dhariwal & Nichol, 2021) trains time-aware models $f_\phi : \mathcal{X}_t \times \mathcal{T} \to \mathcal{Y}$ directly on the noisy latents of the diffusion process, approximating:

$$\nabla_{x_t} \log p(y|x_t) \approx \nabla_{x_t} \mathcal{L}(f_\phi(x_t, t), y) \tag{8}$$

While this requires training a specialized model, it provides a more direct approximation by conditioning on the actual noisy latent rather than a point estimate.

**Stabilization via Adaptive Moments.** Examining the guided sampling update which follows from Equation 4 and Equation 5, we observe that it performs gradient ascent on the likelihood component at each timestep:

$$x_s = \frac{1}{\alpha_{t|s}} x_t + \frac{\sigma_{t|s}^2}{\alpha_{t|s}} \Big( s_\theta(x_t, t) + \nabla \mathcal{L}(\cdot) \Big) + \frac{\sigma_{t|s} \sigma_s}{\sigma_t} \epsilon \tag{9}$$

Drawing from stochastic optimization, where adaptive moment estimation has proven effective at stabilizing noisy gradients (e.g. Adam (Kingma, 2014), AdaGrad (Duchi et al., 2011)), we propose

maintaining exponential moving averages of the likelihood gradients across sampling steps:

$$g_t = \nabla_{x_t} \mathcal{L}(f_\phi(\cdot), y) \quad \text{(DPS or CG)} \tag{10}$$

$$m_k = \beta_1 m_{k-1} + (1 - \beta_1) g_t \tag{11}$$

$$v_k = \beta_2 v_{k-1} + (1 - \beta_2) g_t^2 \tag{12}$$

where $g_t^2$ denotes element-wise squaring and $k$ is a step counter. The stabilized likelihood score is then computed as:

$$\hat{g}_t = \frac{\hat{m}_k}{\sqrt{\hat{v}_k} + \epsilon} \tag{13}$$

with bias-corrected moments $\hat{m}_k = m_k/(1 - \beta_1^k)$ and $\hat{v}_k = v_k/(1 - \beta_2^k)$, and $\epsilon$ a small constant for numerical stability (Kingma, 2014).

This adaptive moment estimation serves multiple purposes: the first moment (momentum) smooths the optimization trajectory by accumulating gradient information across steps, while the second moment adaptively scales the updates based on the historical variance of each gradient component. For conditional sampling, where the likelihood approximation can vary dramatically across noise levels, this stabilization is particularly beneficial. We denote the resulting methods as **AdamDPS** (Algorithm 1) when applied to DPS and **AdamCG** (Algorithm 2) when applied to classifier guidance, demonstrating that this simple modification yields substantial improvements in both sampling stability and final sample quality.

## 4 Synthetic Study

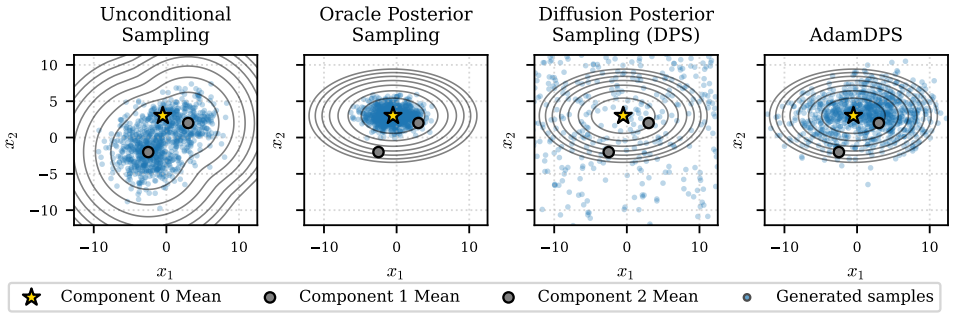


Figure 1: Final sample distributions on the 2D GMM for different sampling methods when conditioning on component 0 (target indicated by yellow star).

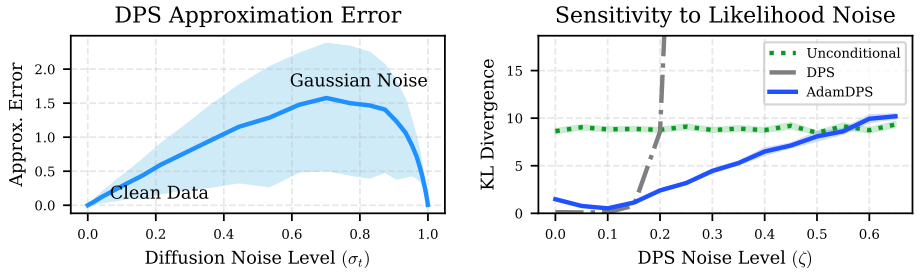


Figure 2: Analysis of DPS and AdamDPS on a 2D Gaussian Mixture Model.

We first analyze our method in a tractable 2D Gaussian Mixture Model (GMM), see Figure 1, where both the unconditional score $\nabla_{x_t} \log p_t(x_t)$, the true conditional score $\nabla_{x_t} \log p_t(x_t|y)$, and the DPS likelihood approximation $\nabla_{x_t} \log p(y|x_{0|t})$ are available in closed form. This allows direct evaluation of approximation error. We construct a GMM with three components and condition on one target component for conditional sampling.

DPS replaces $\nabla_{x_t} \log p(y|x_t)$ with $\nabla_{x_t} \log p(y|x_{0|t})$. Even with perfect $x_{0|t}$, this substitution is imperfect. Figure 2 (left) shows the L2 error across noise levels $\sigma_t$: negligible at very low or high

noise, but peaking at intermediate levels. This inherent error motivates more robust guidance. To mimic imperfect guidance models, we inject Gaussian noise of magnitude $\zeta\|\nabla_{x_t} \log p(y|x_{0|t})\|$ into the DPS gradient. We observe that DPS trajectories oscillate and fail to converge under noise, yielding diffuse samples that deviate from the oracle posterior, see Figure 1 ($\zeta = .18$).

**AdamDPS Stabilization.** AdamDPS smooths and rescales these noisy gradients. Trajectories are more stable and converge directly to the target, producing concentrated samples close to the oracle posterior. Quantitatively, Figure 2 shows AdamDPS consistently achieves lower KL in settings with non-negligible gradient noise. The GMM study highlights two points: (i) the DPS approximation itself introduces error, and (ii) DPS is highly sensitive to gradient noise. By contrast, AdamDPS dampens noise, stabilizes trajectories, and yields more faithful conditional samples.

## 5 Benchmarks

We conducted extensive experiments across a range of tasks to validate the effectiveness of adaptive moments for plug-and-play guidance. All methods benchmarked were tuned extensively using 150 trials of Bayesian Optimization (Jones et al., 1998) on a held-out validation set of 32 images. Reconstruction tasks were tuned to minimize LPIPs (Zhang et al., 2018), while class conditional sampling was tuned for CMMD (Jayasumana et al., 2024) since tuning for accuracy encourages generating adversarial examples. To measure alignment with the desired condition, we report LPIPs for reconstuction tasks and accuracy for class conditional sampling. We also report FID (Heusel et al., 2017) as a measure of fidelity, computed on 2048 samples following the evaluation procedure from Ye et al. (2024). We benchmark against Loss Guided Diffusion (LGD) (Song et al., 2023), Manifold Preserving Guided Diffusion (MPGD) (He et al., 2023), DPS, UGD, and the state-of-the-art TFG on ImageNet (Deng et al., 2009), CIFAR10 (Krizhevsky et al., 2009), and the Cats subset of Cats vs. Dogs (Elson et al., 2007). We set $N_{recur} = 1$ for TFG and sweep $N_{iter} = 1, 2, 4$ for UGD and TFG, while tuning the remaining hyperparameters as recommended by Ye et al. (2024). We provide additional quantitative and qualitative results in the appendix.

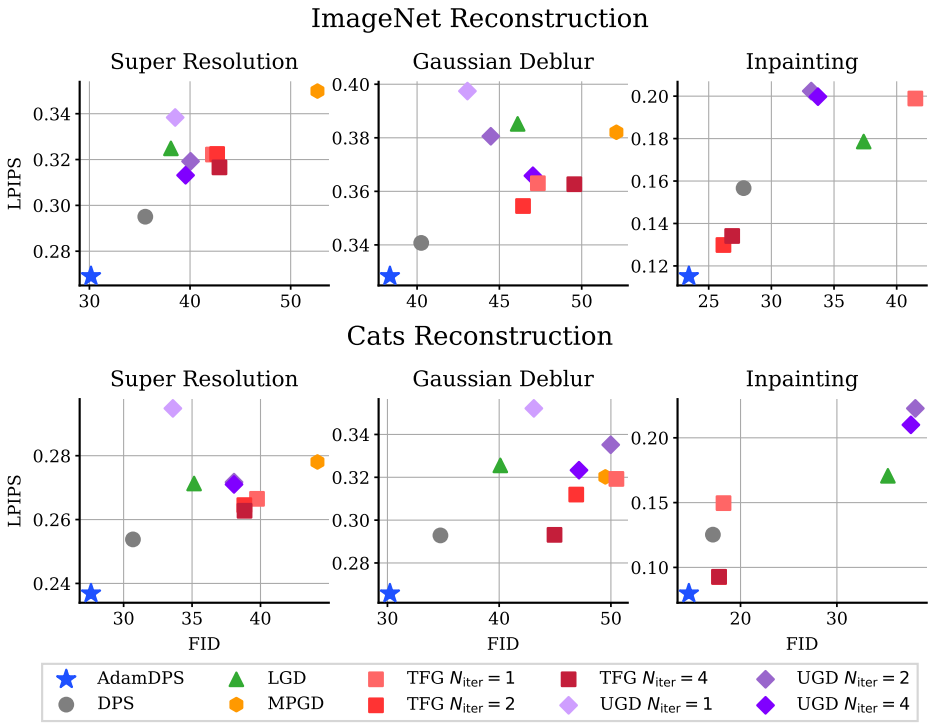


Figure 3: Comparison of all methods on ImageNet and Cats for Super Resolution at 16x Downsampling, Gaussian Deblur at Blur Intensity 12, and Inpainting at 90% Uniform Random Mask.

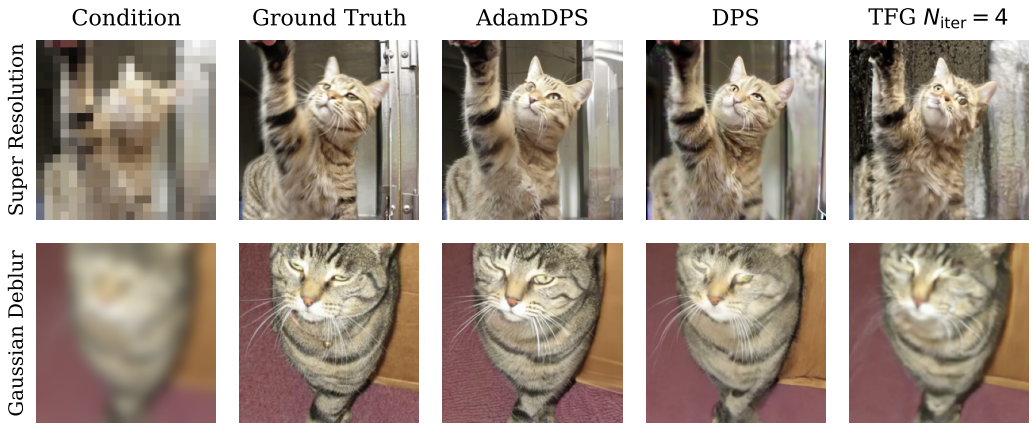


Figure 4: Qualitative examples from the Cats dataset: Super Resolution at 12x Downsampling and Gaussian Deblurring at Blur Intensity 9.

**Reconstruction** Qualitatively, AdamDPS generates images with superior perceptual fidelity, preserving finer details particularly in challenging regions where other methods struggle. TFG reconstructions frequently exhibit visual artifacts, while DPS lacks the refined detail recovery of our approach, see Figure 4. Quantitatively, for both ImageNet and the Cats dataset AdamDPS outperforms all other methods on all reconstruction tasks: super resolution at 16x downsample, Gaussian deblur at blur intensity 12, and inpainting at 90% uniform random masking, see Figure 3. Interestingly, the second best performing method across datasets in super resolution and Gaussian deblurring is DPS. We observe this happens in challenging settings where conditioning information provides limited direct correspondence to the target image, causing other methods to degrade significantly while DPS remains robust. Inpainting at 90% uniform random masking proves easier than super resolution at 16x downsample and Gaussian deblur at blur intensity 12, as evidenced by lower LPIPS scores across all methods. In this simpler setting, TFG $N_{\text{iter}} = 4$ outperforms DPS.

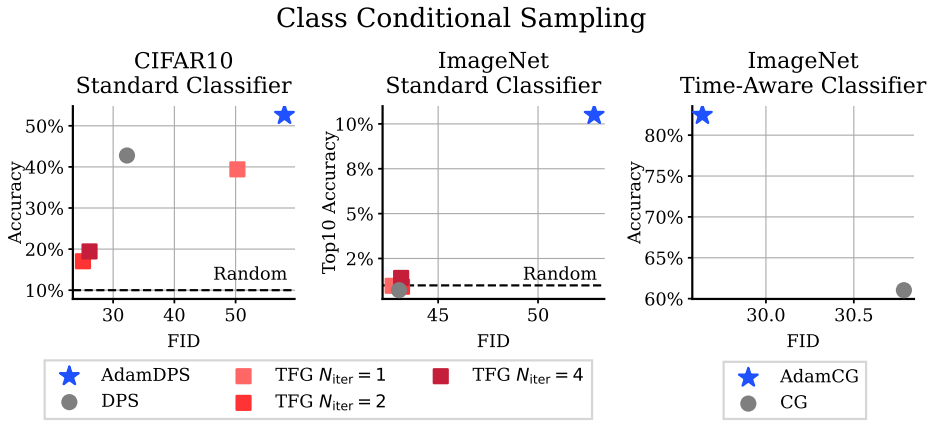


Figure 5: Comparison of DPS, TFG, and AdamDPS on CIFAR10 and ImageNet with a Standard Classifier, as well as CG and AdamCG on ImageNet with a Time-Aware Classifier.

**Class Conditional Sampling.** AdamDPS demonstrates strong performance on class conditional tasks, see Figure 5. On CIFAR-10, AdamDPS outperforms the next best method, DPS, by 9.86% in classification accuracy, establishing clear superiority over existing approaches. The more striking results is on ImageNet, where all other approaches fail to exceed random guessing, achieving top-10 classification accuracies at or below the 1% random baseline. In contrast, AdamDPS achieves 10.49% top-10 accuracy, a substantial improvement that demonstrates the method's ability to gen-

erate samples satisfying the desired condition in challenging settings. In the time-aware classifier setting, AdamCG improves upon CG by more than 20% in classification accuracy.

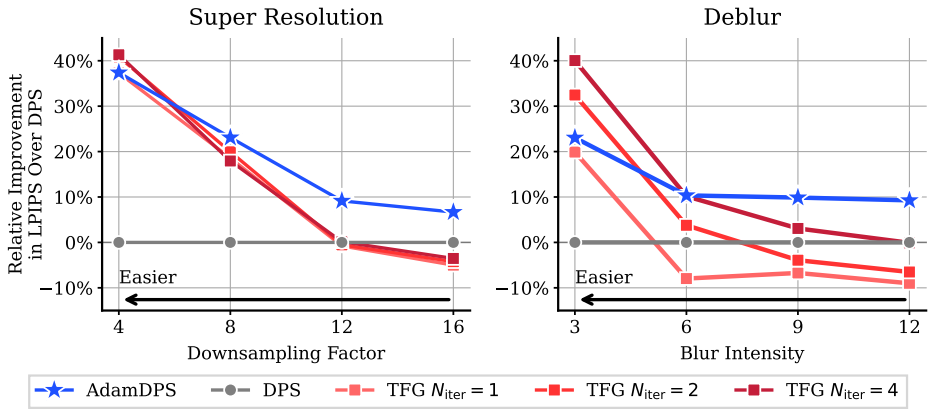


Figure 6: Task Difficulty Ablation on Cats comparing AdamDPS and TFG to DPS.

**Task Difficulty Ablation.** We examine how task difficulty affects the relative performance of AdamDPS and TFG compared to DPS. Figure 6 shows the relative LPIPS improvement of AdamDPS and TFG over DPS on the Cats dataset as super resolution and deblurring tasks increase in difficulty. For super resolution, AdamDPS consistently maintains substantial positive improvement over DPS across all difficulty levels, achieving nearly 10% relative improvement at 12x downsampling. While certain TFG variants ($N_{\text{iter}} \geq 2$) marginally outperform AdamDPS in the easiest settings, their advantage rapidly diminishes as task difficulty increases, with performance eventually falling below the DPS baseline, as we saw in Figure 3. Deblurring tasks exhibit a similar pattern: while TFG variants ($N_{\text{iter}} \geq 2$) outperform AdamDPS in the easiest setting, AdamDPS demonstrates robust positive improvement that becomes more pronounced at higher blur intensities, as TFG variants show declining performance and ultimately underperform DPS on the most challenging tasks. These results underscore AdamDPS's robustness, maintaining effectiveness even when conditioning information is limited.

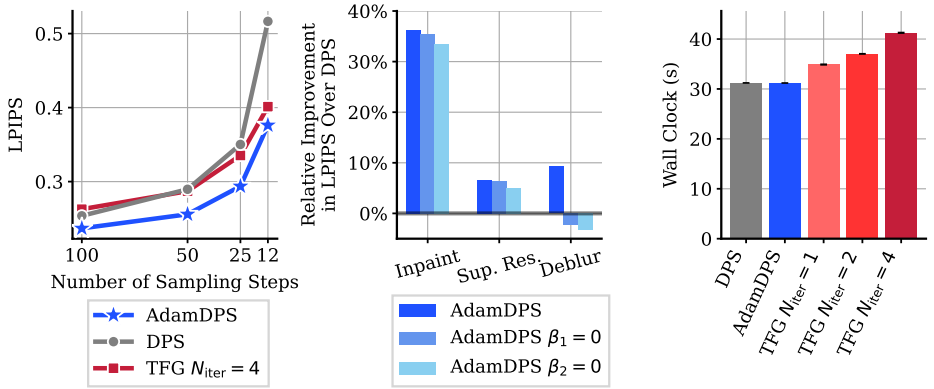


Figure 7: Left: Ablation of sampling steps for Super Resolution 16x Downsampling on Cats dataset. Middle: Ablation of Adam $\beta_1$ $\beta_2$ for Super Resolution 16x Downsampling, Gaussian Deblur as Blur Intensity 12 and Inpainting at 90% Uniform Random Mask on the Cats dataset. Right: Wall clock comparison on 1 H100 GPU class conditional sampling a batch of 8 256x256 Images.

**Sampling Steps Ablation.** We analyze the effect of reducing sampling steps on LPIPS performance for Super Resolution at 16x Downsampling on the Cats dataset, as shown in Figure 7. AdamDPS consistently achieves superior LPIPS compared to both DPS and TFG $N_{\text{iter}} = 4$ across all step

counts. AdamDPS's advantage persists at lower step counts, maintaining a width gap even at 25 steps, underscoring its efficiency and suitability for few-step sampling.

**Wall Clock.** Figure 7 shows average wall clock time from 5 trials of ImageNet class conditional sampling using 1 H100 GPU with a batch size 8. AdamDPS and DPS substantially outpace TFG, whose cost scales with $N_{\text{recur}}(1 + N_{\text{iter}})$ gradient computations and $N_{\text{recur}}$ backpropagation steps.

## 6 Analysis

**Sampling Loss Trajectories.** We track the measurement loss $\mathcal{L}(f_\phi(\hat{x}_\theta(x_t, t)), y)$ along the reverse trajectory to explore the behavior of the different methods. Figure 8 aggregates losses over images for four settings: 4x SR, 16x SR, class conditional sampling with a standard or time-aware classifier. We observe that for challenging settings (16x SR and class conditional sampling) AdamDPS reduces the loss early and steadily compared to TFG, and consistently acheives a lower terminal loss than DPS. On the other hand, TFG typically stalls early in sampling for challenging tasks. This demonstrates that adaptive moments are particularly beneficial in stabilizing guidance early when the signal is noisiest. For easy inverse setting (4x SR) TFG descends quickly and reaches competitive terminal losses, while AdamDPS remains smoothly monotone and competitive. In these settings the conditioning information provides very rich guidance signals. For the class conditional setting, TFG and DPS often fail to make meaningful progress while AdamDPS exhibits reliable, monotonic descent and consistently converges to a lower terminal loss (right panel). Representative per-sample curves in Figure 9 (16x SR) visualize these behaviors. Consistent with the loss trajectory visualization, AdamDPS makes consistent progress towards the conditioning information early in the sampling process.

**Adam $\beta_1$, $\beta_2$ Ablation.** Figure 7 shows an ablation of Adam hyperparameters $\beta_1$ and $\beta_2$ for AdamDPS across inpainting, super-resolution, and deblurring tasks. We compare the relative LPIPS improvement over DPS for three configurations: default AdamDPS, AdamDPS with $\beta_1 = 0$, and AdamDPS with $\beta_2 = 0$. AdamDPS consistently outperforms both ablated variants across all tasks, demonstrating that both momentum and adaptive scaling are essential for optimal performance, with their relative importance varying by task.

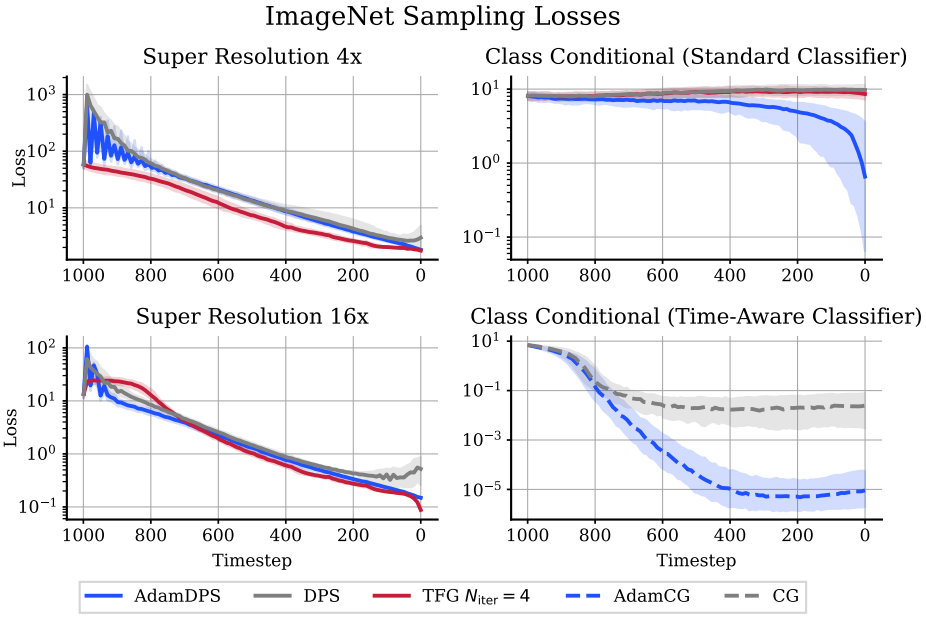


Figure 8: Sample figure caption.

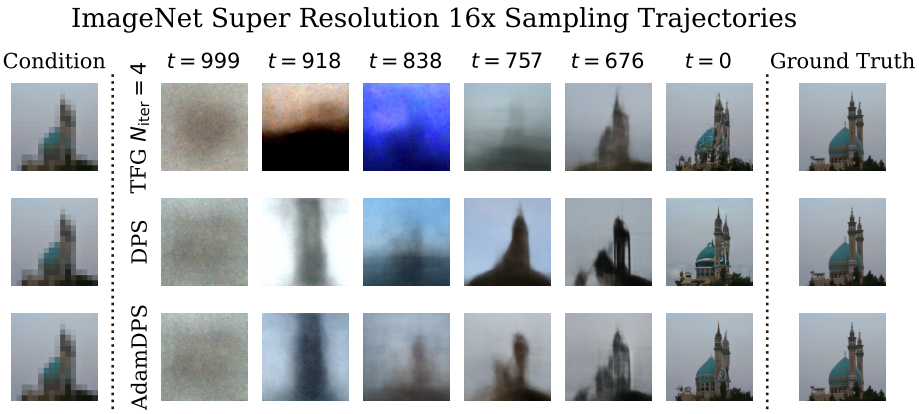


Figure 9: Sample figure caption.

## 7 Related Work

Conditional generation with diffusion models, often termed *classifier guidance*, was initially proposed for class-conditional sampling using a learned time-aware classifier with an unconditional diffusion model (Dhariwal & Nichol, 2021). This concept has since been broadly applied, particularly to inverse problems involving known analytical models $f_\phi(x)$ that map data $x$ to a condition $y$. For these inverse problems, Diffusion Posterior Sampling (DPS) (Chung et al., 2022) uses the denoised estimate $\hat{x}_\theta(x_t, t) \approx \mathbb{E}[x_0 \mid x_t]$ (Efron, 2011) and applies $\nabla_{x_t} \log p(y|x_t) \approx \nabla_{x_t} \mathcal{L}(f_\phi(\hat{x}_\theta(x_t, t), y))$. Related approaches include PiGDM (Song et al.), which handles non-differentiable measurements via pseudoinverse guidance, and DDRM (Kawar et al., 2022), which leverages variational inference for efficient posterior sampling. MPGD sidesteps backprop through the diffusion model by optimizing in data space on $\hat{x}_\theta$ (He et al., 2023). Orthogonal improvements include timestep resampling ("time travel") and efficient recurrence (Mokady et al., 2023; Wang et al., 2022; Lugmayr et al., 2022; Du et al., 2023; Yu et al., 2023), Monte-Carlo smoothing of the DPS surrogate via Loss-Guided Diffusion (LGD) (Song et al., 2023), and covariance estimation as in TMPD (Boys et al., 2023) and FreeHunch (Rissanen et al., 2024). Compositional frameworks such as UGD (Bansal et al., 2023) and TFG (Ye et al., 2024) unify DPS/MPGD with reoccurrence and LGD.

## 8 Conclusion

Guidance in diffusion models often involves navigating noisy likelihood estimates, particularly when dealing with complex, real-world conditional generation tasks where analytical guidance models can be computationally expensive. While recent approaches have combined multiple sophisticated techniques, increasing algorithmic complexity, we have demonstrated that a simple adaptation can yield significant improvements. AdamDPS, by incorporating adaptive moment estimation from stochastic optimization, effectively stabilizes the noisy gradients inherent in the Diffusion Posterior Sampling approximation. Our experiments show that this straightforward modification leads to comparable or superior performance against more complex methods, especially in challenging settings and with limited computational budgets where multiple guidance steps per iteration are infeasible. The key advantage of AdamDPS lies in its simplicity: it requires minimal changes to existing DPS frameworks while leveraging a well-understood and robust optimization technique. This work suggests that focusing on effectively managing the inherent noise in fundamental guidance signals can be as, or more, beneficial than composing increasingly elaborate guidance schemes.

### Ethics

Our work focuses on improving plug-and-play diffusion sampling; an important problem with many practical applications. While its possible for this technology to be misused, we advocate for responsible use in line with established guidelines.

### Reproducibility

All experiments are conducted on publicly available data with publicly available pretrained models. Upon acceptance we will release the necessary code to reproduce all of our results.

## References


Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 843–852, 2023.

Benjamin Boys, Mark Girolami, Jakiw Pidstrigach, Sebastian Reich, Alan Mosca, and O Deniz Akyildiz. Tweedie moment projected diffusions for inverse problems. *arXiv preprint arXiv:2310.06721*, 2023.

Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, et al. Video generation models as world simulators. *OpenAI Blog*, 1(8):1, 2024.

Hyungjin Chung, Jeongsol Kim, Michael T Mccann, Marc L Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. *arXiv preprint arXiv:2209.14687*, 2022.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Yilun Du, Conor Durkan, Robin Strudel, Joshua B Tenenbaum, Sander Dieleman, Rob Fergus, Jascha Sohl-Dickstein, Arnaud Doucet, and Will Sussman Grathwohl. Reduce, reuse, recycle: Compositional generation with energy-based diffusion models and mcmc. In *International conference on machine learning*, pp. 8489–8510. PMLR, 2023.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(61):2121–2159, 2011. URL `http://jmlr.org/papers/v12/duchi11a.html`.

Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011.

Jeremy Elson, John (JD) Douceur, Jon Howell, and Jared Saul. Asirra: A captcha that exploits interest-aligned manual image categorization. In *Proceedings of 14th ACM Conference on Computer and Communications Security (CCS)*. Association for Computing Machinery, Inc., October 2007.

Yutong He, Naoki Murata, Chieh-Hsin Lai, Yuhta Takida, Toshimitsu Uesaka, Dongjun Kim, Wei-Hsiang Liao, Yuki Mitsufuji, J Zico Kolter, Ruslan Salakhutdinov, and Stefano Ermon. Manifold preserving guided diffusion. *arXiv preprint arXiv:2311.16424*, 2023.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in neural information processing systems*, 35:8633–8646, 2022.

Emiel Hoogeboom, Vıctor Garcia Satorras, Clément Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3d. In *International conference on machine learning*, pp. 8867–8887. PMLR, 2022.

Sadeep Jayasumana, Srikumar Ramalingam, Andreas Veit, Daniel Glasner, Ayan Chakrabarti, and Sanjiv Kumar. Rethinking fid: Towards a better evaluation metric for image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9307–9315, 2024.

Donald R Jones, Matthias Schonlau, and William J Welch. Efficient global optimization of expensive black-box functions. *Journal of Global optimization*, 13(4):455–492, 1998.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022.

Bahjat Kawar, Michael Elad, Stefano Ermon, and Jiaming Song. Denoising diffusion restoration models. *Advances in neural information processing systems*, 35:23593–23606, 2022.

Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. *Advances in neural information processing systems*, 34:21696–21707, 2021.

Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Zhifeng Kong, Wei Ping, Jiaji Huang, Kexin Zhao, and Bryan Catanzaro. Diffwave: A versatile diffusion model for audio synthesis. *arXiv preprint arXiv:2009.09761*, 2020.

Alex Krizhevsky et al. Learning multiple layers of features from tiny images. 2009.

Haohe Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Xubo Liu, Danilo Mandic, Wenwu Wang, and Mark D Plumbley. Audioldm: Text-to-audio generation with latent diffusion models. *arXiv preprint arXiv:2301.12503*, 2023.

Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11461–11471, 2022.

Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6038–6047, 2023.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

Severi Rissanen, Markus Heinonen, and Arno Solin. Free hunch: Denoiser covariance estimation for diffusion models without extra costs. *arXiv preprint arXiv:2410.11149*, 2024.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022a.

Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE transactions on pattern analysis and machine intelligence*, 45(4):4713–4726, 2022b.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pp. 2256–2265. pmlr, 2015.

Jiaming Song, Arash Vahdat, Morteza Mardani, and Jan Kautz. Pseudoinverse-guided diffusion models for inverse problems. In *International Conference on Learning Representations*.

Jiaming Song, Qinsheng Zhang, Hongxu Yin, Morteza Mardani, Ming-Yu Liu, Jan Kautz, Yongxin Chen, and Arash Vahdat. Loss-guided diffusion models for plug-and-play controllable generation. In *International Conference on Machine Learning*, pp. 32483–32498. PMLR, 2023.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. In *Advances in Neural Information Processing Systems*, volume 32, pp. 11895–11907, 2019.

Yinhuai Wang, Jiwen Yu, and Jian Zhang. Zero-shot image restoration using denoising diffusion null-space model. *arXiv preprint arXiv:2212.00490*, 2022.

Joseph L Watson, David Juergens, Nathaniel R Bennett, Brian L Trippe, Jason Yim, Helen E Eisenach, Woody Ahern, Andrew J Borst, Robert J Ragotte, Lukas F Milles, et al. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976):1089–1100, 2023.

Haotian Ye, Haowei Lin, Jiaqi Han, Minkai Xu, Sheng Liu, Yitao Liang, Jianzhu Ma, James Y Zou, and Stefano Ermon. Tfg: Unified training-free guidance for diffusion models. *Advances in Neural Information Processing Systems*, 37:22370–22417, 2024.

Jiwen Yu, Yinhuai Wang, Chen Zhao, Bernard Ghanem, and Jian Zhang. Freedom: Training-free energy-guided conditional diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23174–23184, 2023.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 586–595, 2018.

# A Additional Quantitative Results

Table 1: Super Resolution at 16x Downsampling, Gaussian Deblur at Blur Intensity 12, and Inpainting at 90% Uniform Random Mask on ImageNet.

| Method | Super Resolution | | Gaussian Deblur | | Inpainting | |
|---|---|---|---|---|---|---|
| | LPIPs ↓ | FID ↓ | LPIPs ↓ | FID ↓ | LPIPs ↓ | FID ↓ |
| DPS | 0.30 | 35.55 | 0.34 | 40.25 | 0.16 | 27.79 |
| LGD | 0.32 | 38.10 | 0.39 | 46.11 | 0.18 | 37.37 |
| MPGD | 0.35 | 52.66 | 0.38 | 52.11 | 0.45 | 142.47 |
| $\text{UGD}_{N_{\text{iter}}=1}$ | 0.34 | 38.52 | 0.40 | 43.06 | 0.29 | 47.30 |
| $\text{UGD}_{N_{\text{iter}}=2}$ | 0.32 | 40.05 | 0.38 | 44.48 | 0.20 | 33.18 |
| $\text{UGD}_{N_{\text{iter}}=4}$ | 0.31 | 39.56 | 0.37 | 47.04 | 0.20 | 33.72 |
| $\text{TFG}_{N_{\text{iter}}=1}$ | 0.32 | 42.27 | 0.36 | 47.34 | 0.20 | 41.49 |
| $\text{TFG}_{N_{\text{iter}}=2}$ | 0.32 | 42.69 | 0.35 | 46.44 | 0.13 | 26.17 |
| $\text{TFG}_{N_{\text{iter}}=4}$ | 0.32 | 42.92 | 0.36 | 49.55 | 0.13 | 26.87 |
| AdamDPS | 0.27 | 30.16 | 0.33 | 38.35 | 0.12 | 23.42 |

Table 2: Super Resolution at 4x Downsampling, Gaussian Deblur at Blur Intensity 3 on ImageNet.

| Method | Super Resolution | | Gaussian Deblur | |
|---|---|---|---|---|
| | LPIPs ↓ | FID ↓ | LPIPs ↓ | FID ↓ |
| DPS | 0.19 | 26.43 | 0.23 | 27.66 |
| $\text{TFG}_{N_{\text{iter}}=1}$ | 0.14 | 25.37 | 0.20 | 27.21 |
| $\text{TFG}_{N_{\text{iter}}=2}$ | 0.13 | 25.03 | 0.17 | 26.01 |
| $\text{TFG}_{N_{\text{iter}}=4}$ | 0.14 | 26.62 | 0.16 | 25.74 |
| AdamDPS | 0.12 | 20.92 | 0.17 | 23.45 |

Table 3: Super Resolution at 16x Downsampling, Gaussian Deblur at Blur Intensity 12, and Inpainting at 90% Uniform Random Mask on Cats.

| Method | Super Resolution | | Gaussian Deblur | | Inpainting | |
|---|---|---|---|---|---|---|
| | LPIPs ↓ | FID ↓ | LPIPs ↓ | FID ↓ | LPIPs ↓ | FID ↓ |
| DPS | 0.25 | 30.68 | 0.29 | 34.74 | 0.13 | 17.13 |
| LGD | 0.27 | 35.15 | 0.33 | 40.10 | 0.17 | 35.29 |
| MPGD | 0.28 | 44.16 | 0.32 | 49.50 | 0.42 | 76.60 |
| $\text{UGD}_{N_{\text{iter}}=1}$ | 0.29 | 33.60 | 0.35 | 43.10 | 0.32 | 64.22 |
| $\text{UGD}_{N_{\text{iter}}=2}$ | 0.27 | 38.07 | 0.34 | 49.98 | 0.22 | 38.14 |
| $\text{UGD}_{N_{\text{iter}}=4}$ | 0.27 | 38.07 | 0.32 | 47.14 | 0.21 | 37.68 |
| $\text{TFG}_{N_{\text{iter}}=1}$ | 0.27 | 39.75 | 0.32 | 50.49 | 0.15 | 18.24 |
| $\text{TFG}_{N_{\text{iter}}=2}$ | 0.26 | 38.82 | 0.31 | 46.90 | 0.09 | 17.78 |
| $\text{TFG}_{N_{\text{iter}}=4}$ | 0.26 | 38.83 | 0.29 | 44.95 | 0.09 | 17.77 |
| AdamDPS | 0.24 | 27.62 | 0.27 | 30.22 | 0.08 | 14.64 |

Table 4: Super Resolution at 4x Downsampling, Gaussian Deblur at Blur Intensity 3 on Cats.

| Method | Super Resolution LPIPs ↓ | Super Resolution FID ↓ | Gaussian Deblur LPIPs ↓ | Gaussian Deblur FID ↓ |
|---|---|---|---|---|
| DPS | 0.14 | 17.74 | 0.18 | 23.58 |
| $\text{TFG}_{N_{\text{iter}}=1}$ | 0.09 | 14.81 | 0.15 | 20.78 |
| $\text{TFG}_{N_{\text{iter}}=2}$ | 0.08 | 13.54 | 0.12 | 18.20 |
| $\text{TFG}_{N_{\text{iter}}=4}$ | 0.08 | 14.09 | 0.11 | 16.42 |
| AdamDPS | 0.09 | 13.19 | 0.14 | 20.43 |

Table 5: Class Conditional Sampling with Standard Classifier on CIFAR10.

| Method | Accuracy ↑ | Top3 Accuracy ↑ | FID ↓ |
|---|---|---|---|
| DPS | 42.77 | 64.65 | 32.26 |
| LGD | 21.83 | 46.88 | 31.02 |
| MPGD | 26.66 | 56.40 | 34.11 |
| $\text{UGD}_{N_{\text{iter}}=1}$ | 9.28 | 28.86 | 25.40 |
| $\text{UGD}_{N_{\text{iter}}=2}$ | 24.61 | 56.10 | 32.44 |
| $\text{UGD}_{N_{\text{iter}}=4}$ | 31.88 | 63.67 | 37.18 |
| $\text{TFG}_{N_{\text{iter}}=1}$ | 39.40 | 67.09 | 50.29 |
| $\text{TFG}_{N_{\text{iter}}=2}$ | 17.04 | 42.53 | 25.07 |
| $\text{TFG}_{N_{\text{iter}}=4}$ | 19.43 | 44.97 | 26.15 |
| AdamDPS | 52.64 | 79.00 | 57.98 |

Table 6: Class Conditional Sampling with Standard Classifier on ImageNet.

| Method | Top10 Accuracy ↑ | FID ↓ |
|---|---|---|
| DPS | 0.73 | 43.04 |
| LGD | 0.68 | 42.51 |
| MPGD | 0.88 | 43.63 |
| $\text{UGD}_{N_{\text{iter}}=1}$ | 0.83 | 42.35 |
| $\text{UGD}_{N_{\text{iter}}=2}$ | 1.42 | 45.29 |
| $\text{UGD}_{N_{\text{iter}}=4}$ | 1.71 | 43.58 |
| $\text{TFG}_{N_{\text{iter}}=1}$ | 0.98 | 42.72 |
| $\text{TFG}_{N_{\text{iter}}=2}$ | 0.93 | 43.19 |
| $\text{TFG}_{N_{\text{iter}}=4}$ | 1.42 | 43.15 |
| AdamDPS | 10.50 | 52.81 |

Table 7: Class Conditional Sampling with Time-Aware Classifier on ImageNet.

| Method | Accuracy ↑ | FID ↓ |
|---|---|---|
| CG | 61.04 | 30.78 |
| AdamCG | 82.47 | 29.64 |

## B Additional Qualitative Results

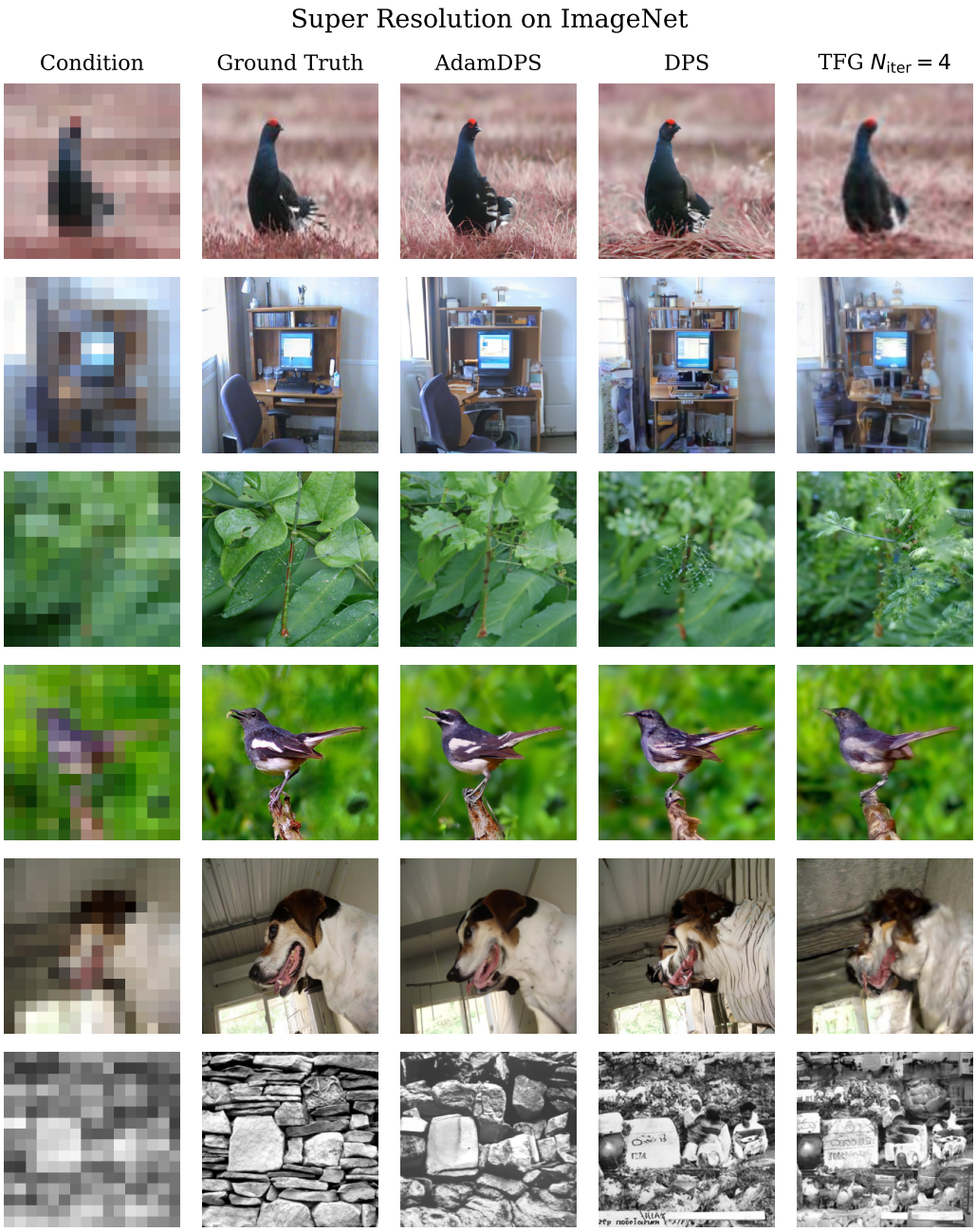


Figure 10: Additional Qualitative Results for Super Resolution at 16x Downsampling on Imagenet.

## C LLM Usage

Large language models were used for proofreading and revising the wording of the paper. All claims and arguments were drafted and verified by the authors.

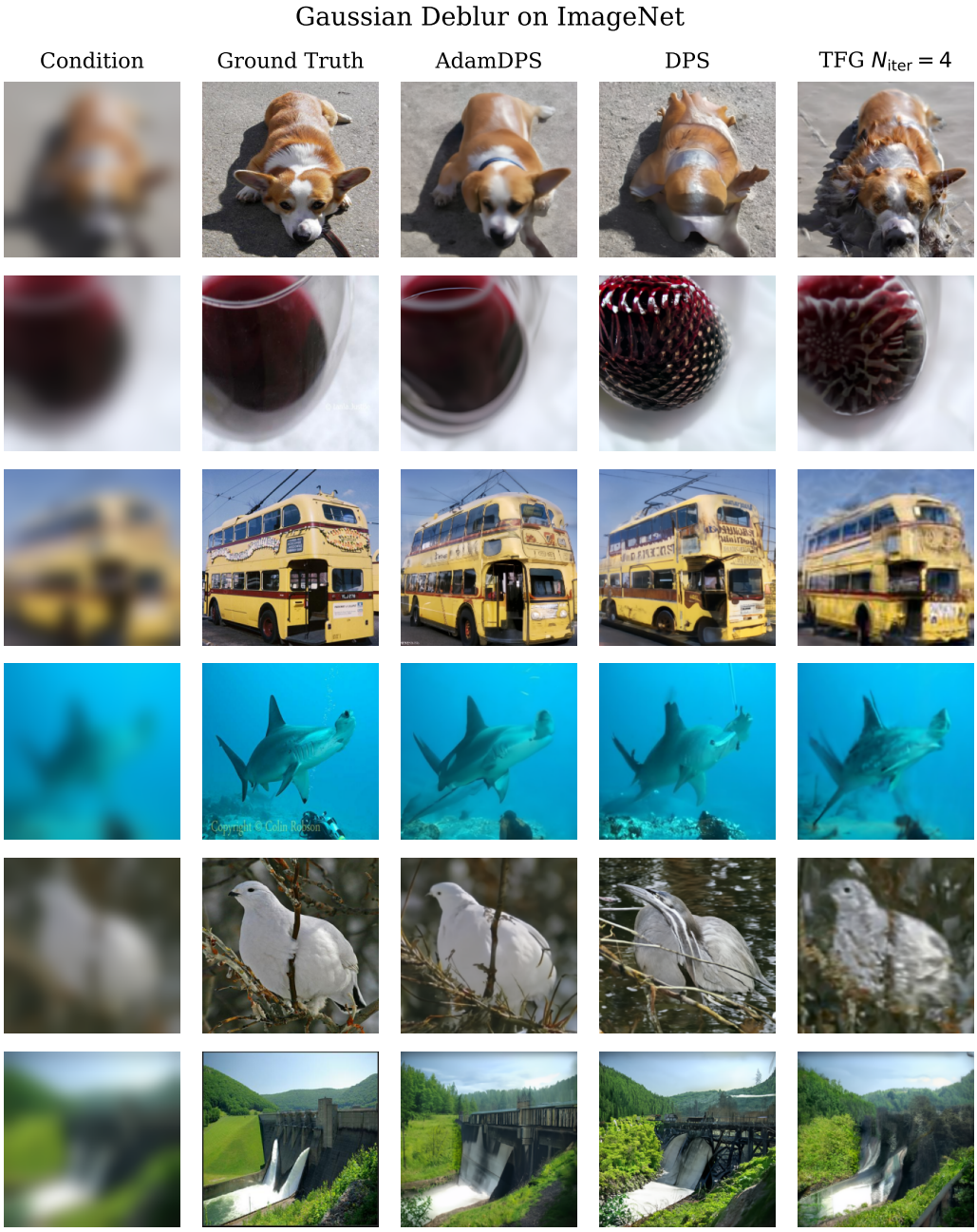


Figure 11: Additional Qualitative Results for Gaussian Deblur at Blur Intensity 12 on Imagenet.

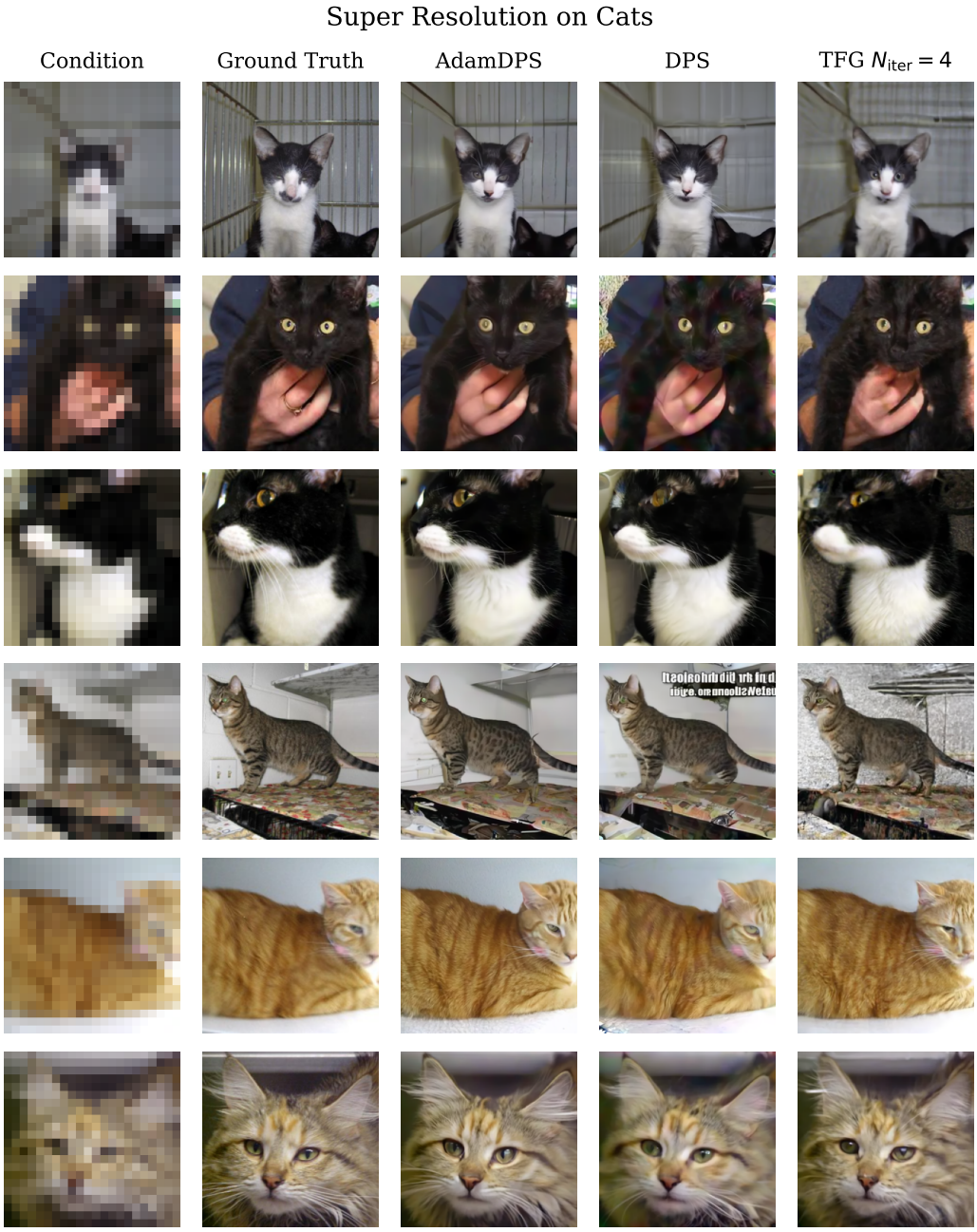


Figure 12: Additional Qualitative Results for Super Resolution at 12x Downsampling on Cats.

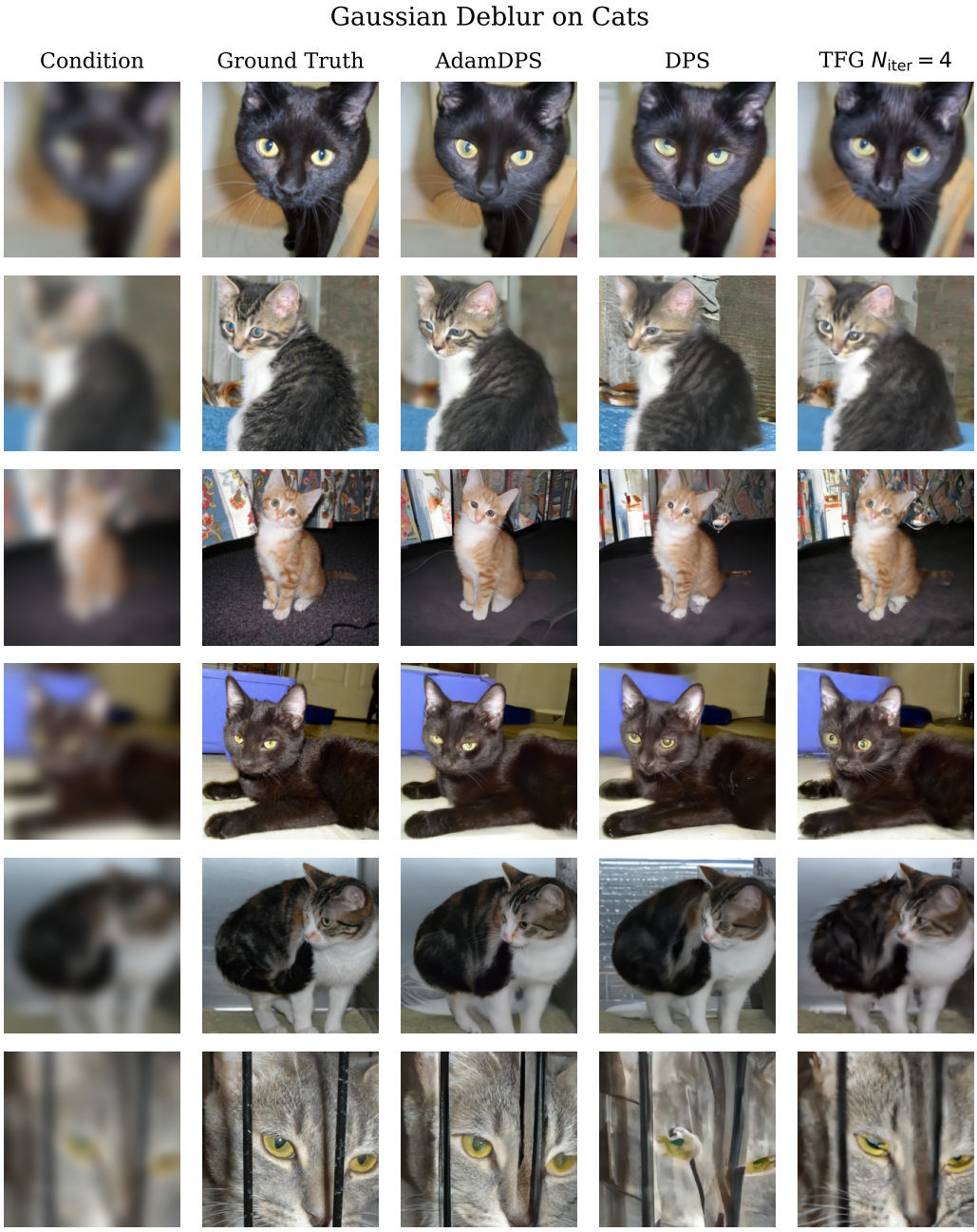


Figure 13: Additional Qualitative Results for Gaussian Deblur at Blur Intensity 9 on Cats.

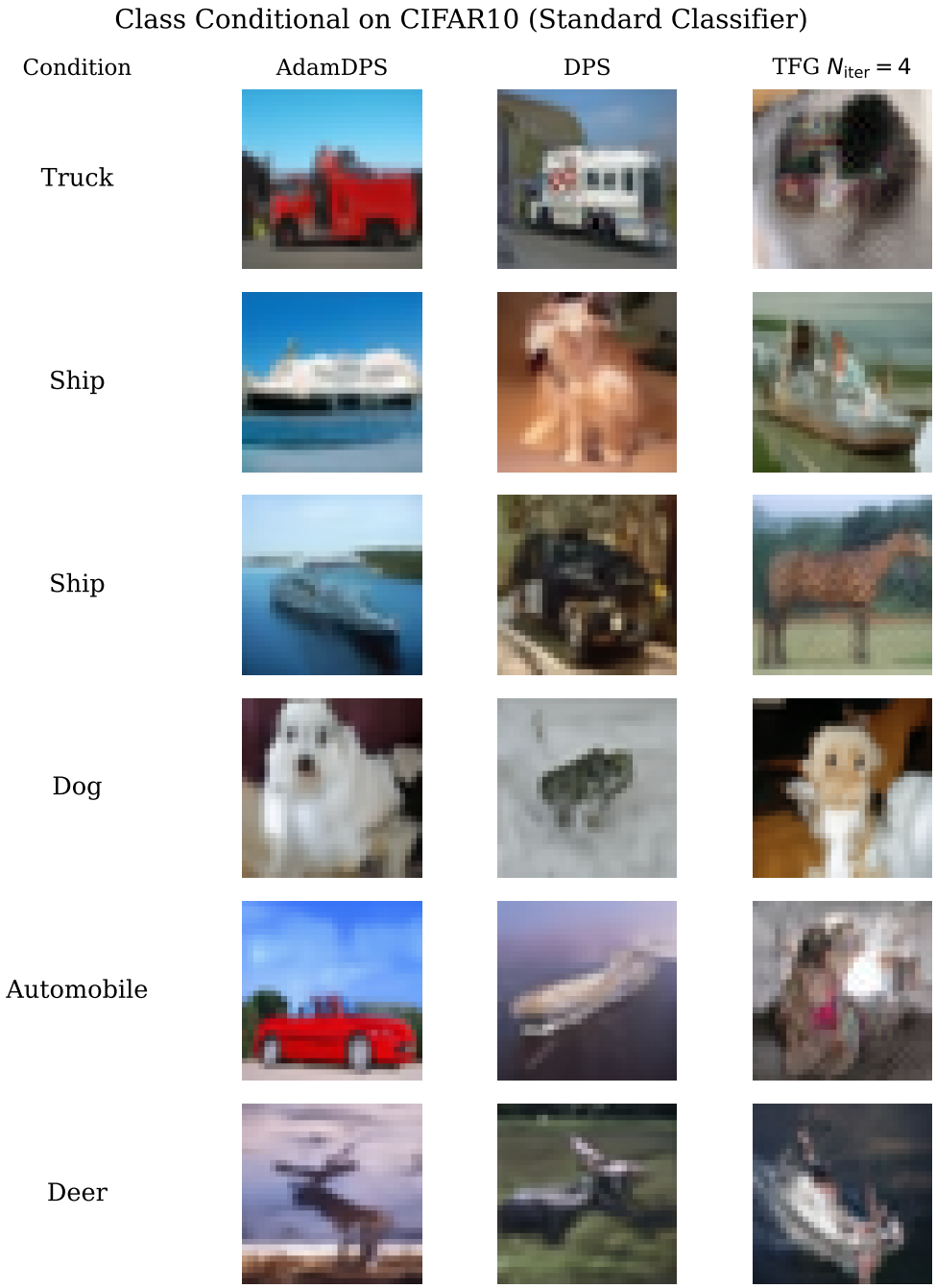


Figure 14: Additional Qualitative Results for Class Conditional Sampling with a Standard Classifier on CIFAR10.

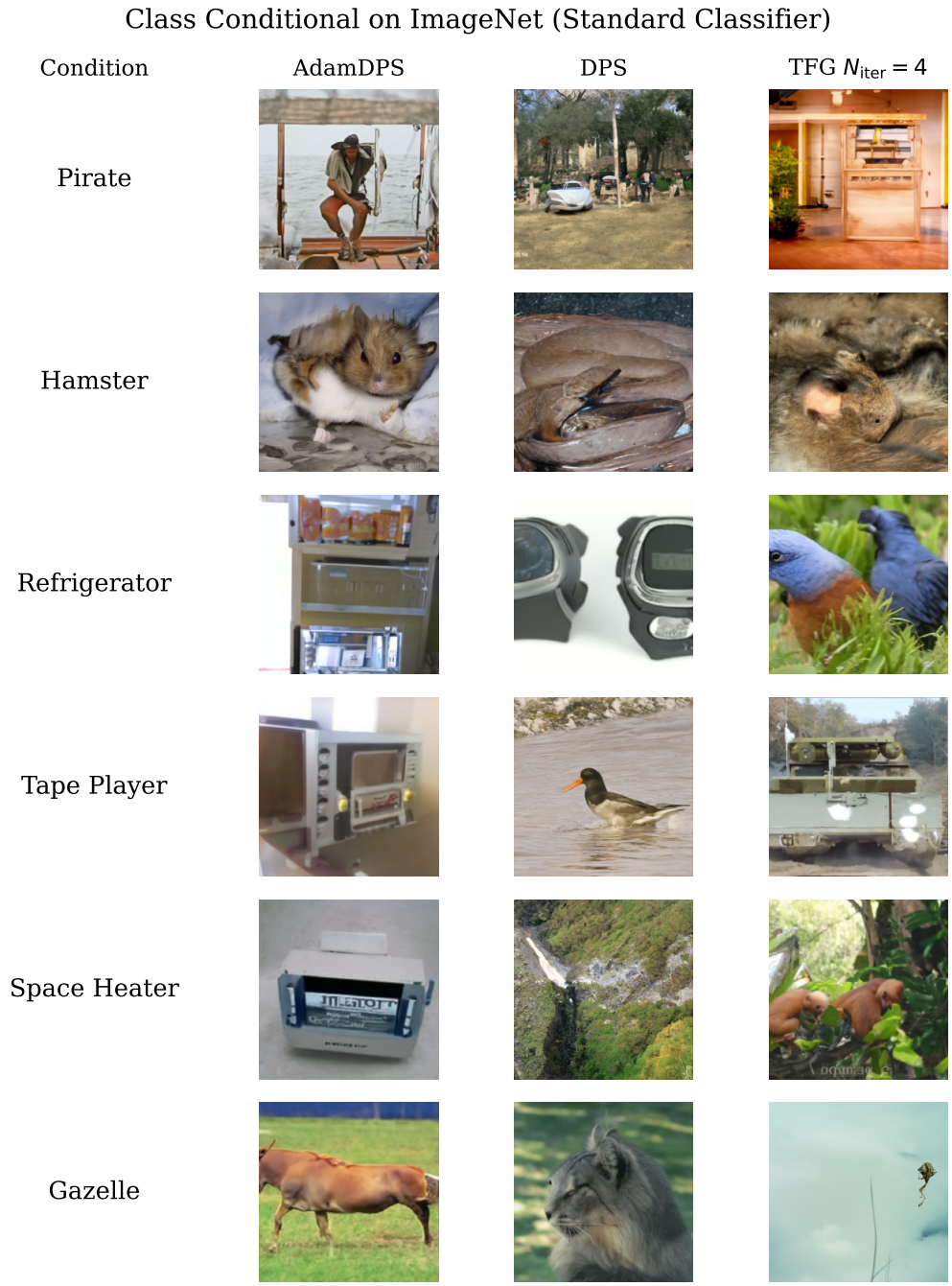


Figure 15: Additional Qualitative Results for Class Conditional Sampling with a Standard Classifier on ImageNet.

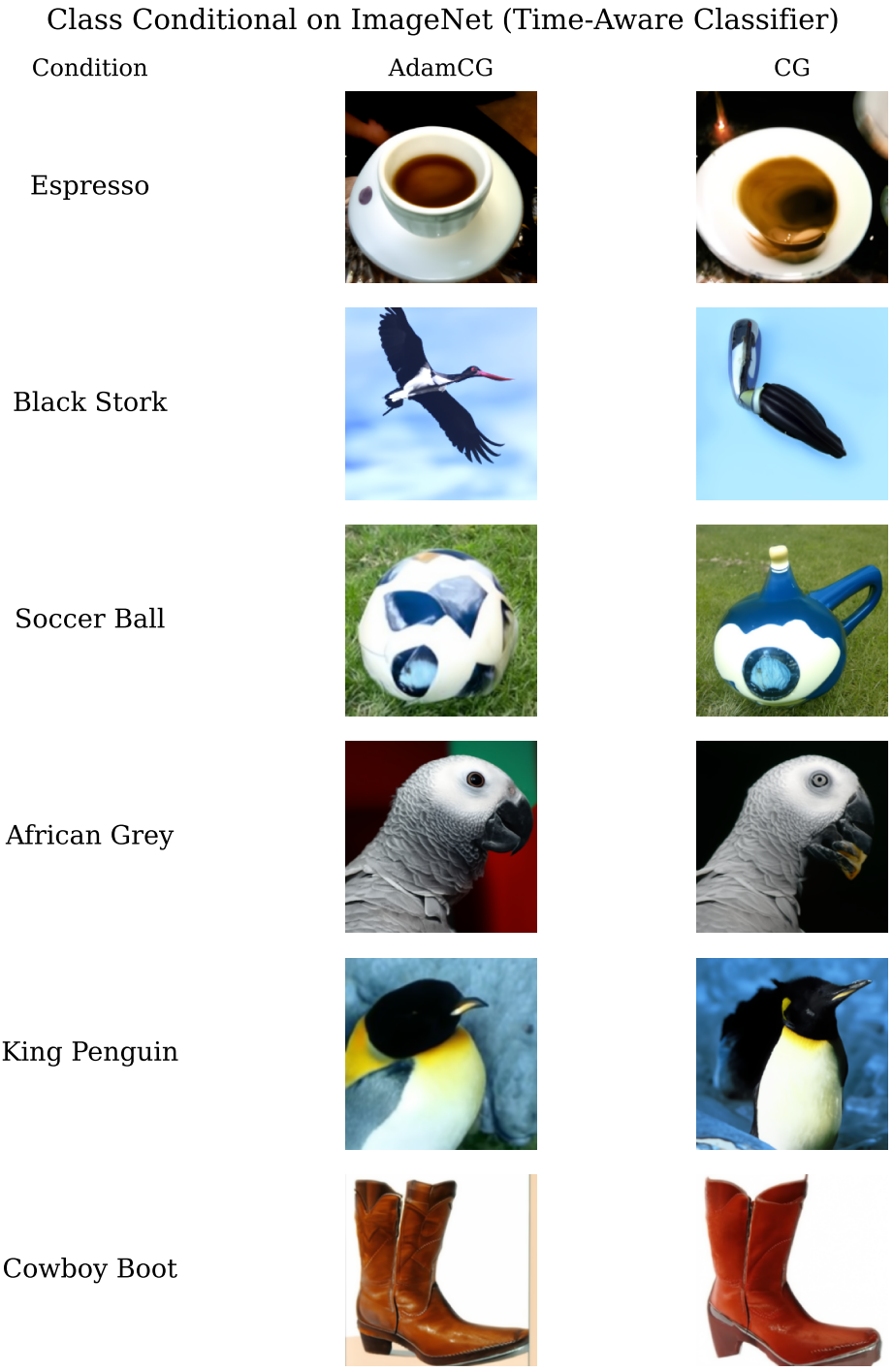


Figure 16: Additional Qualitative Results for Class Conditional Sampling with a Time-Aware Classifier on ImageNet.