

817 A Theoretical Results

818 A.1 Kernel Estimation

819 **Assumption A.1.** The density p is λ lipschitz.

820 Let $\{X^{(i)}\}_{i=1}^n$ a set of n independent samples from a density p that satisfies Assumption A.1. Let \hat{p}
821 be the empirical density on those samples.

822 We are interested in bounding the total variation distance between $p_\sigma := p \otimes \mathcal{N}(0, \sigma^2)$ and $\hat{p}_\sigma =$
823 $\hat{p} \otimes \mathcal{N}(0, \sigma^2)$. In particular,

$$\hat{p}_\sigma(x) = \frac{1}{n\sigma} \sum_{i=1}^n \phi\left(\frac{X^{(i)} - x}{\sigma}\right), \quad (\text{A.1})$$

824 where $\phi(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2}$ is the Gaussian kernel. We want to argue that the TV distance between
825 p_σ and \hat{p}_σ is small given sufficiently many samples n . For simplicity, let's fix the support of p to be
826 $[0, 1]$. We have:

$$d_{\text{TV}}(p_\sigma, \hat{p}_\sigma) = \frac{1}{2} \int_0^1 |p_\sigma(x) - \hat{p}_\sigma(x)| dx = \sum_{l=0}^{L-1} \int_{l/L}^{(l+1)/L} |p_\sigma(x) - \hat{p}_\sigma(x)| dx \quad (\text{A.2})$$

827 Now let us look at one of the terms of the summation.

$$\int_{l/L}^{(l+1)/L} |p_\sigma(x) - \hat{p}_\sigma(x)| dx = \int_{l/L}^{(l+1)/L} |p_\sigma(x) - p_\sigma(l/L) + p_\sigma(l/L) - \hat{p}_\sigma(x)| dx \quad (\text{A.3})$$

$$\leq \int_{l/L}^{(l+1)/L} |p_\sigma(x) - p_\sigma(l/L)| dx + \int_{l/L}^{(l+1)/L} |p_\sigma(l/L) - \hat{p}_\sigma(x)| dx. \quad (\text{A.4})$$

828 We first work on the first term. Using Lemma A.6:

$$\int_{l/L}^{(l+1)/L} |p_\sigma(x) - p_\sigma(l/L)| dx \leq \lambda \int_{l/L}^{(l+1)/L} |x - l/L| dx \quad (\text{A.5})$$

$$= \frac{\lambda}{2L^2}. \quad (\text{A.6})$$

829 Next, we work on the second term.

$$\int_{l/L}^{(l+1)/L} |p_\sigma(l/L) - \hat{p}_\sigma(x)| dx = \int_{l/L}^{(l+1)/L} |p_\sigma(l/L) - \hat{p}_\sigma(l/L) + \hat{p}_\sigma(l/L) - \hat{p}_\sigma(x)| dx \quad (\text{A.7})$$

$$\leq \int_{l/L}^{(l+1)/L} |p_\sigma(l/L) - \hat{p}_\sigma(l/L)| dx + \int_{l/L}^{(l+1)/L} |\hat{p}_\sigma(l/L) - \hat{p}_\sigma(x)| dx. \quad (\text{A.8})$$

830 According to Lemma A.5, we have that \hat{p}_σ is $\hat{\lambda} = \frac{1}{\sigma^2 \sqrt{2\pi e}}$ Lipschitz. Then, the second term becomes:

$$\int_{l/L}^{(l+1)/L} |\hat{p}_\sigma(l/L) - \hat{p}_\sigma(x)| dx \leq \hat{\lambda} \int_{l/L}^{(l+1)/L} |l/L - x| dx = \frac{\hat{\lambda}}{2L^2}. \quad (\text{A.9})$$

831 It remains to bound the following term

$$\int_{l/L}^{(l+1)/L} |p_\sigma(l/L) - \hat{p}_\sigma(l/L)| dx = \frac{|p_\sigma(l/L) - \hat{p}_\sigma(l/L)|}{L} \quad (\text{A.10})$$

832 We will be applying Hoeffding's Inequality, stated below:

833 **Theorem A.2** (Hoeffding's Inequality). *Let Y_1, \dots, Y_n be independent random variables in $[a, b]$ with*
 834 *mean μ . Then,*

$$\Pr \left(\left| \frac{1}{n} \sum_{i=1}^n Y_i - \mu \right| \geq t \right) \leq 2 \exp(-2nt^2/(b-a)^2). \quad (\text{A.11})$$

835 Recall that \hat{p}_σ can be written as

$$\hat{p}_\sigma(x) = \frac{1}{n} \sum_{i=1}^n \frac{\phi((X^{(i)} - x)/\sigma)}{\sigma} = \frac{1}{n} \sum_{i=1}^n Y_i, \quad (\text{A.12})$$

836 in terms of the random variables $Y_i := \frac{\phi((X^{(i)} - x)/\sigma)}{\sigma}$. These random variables are supported in
 837 $\left[0, \frac{1}{\sqrt{2\pi\sigma^2}}\right]$. So, for any x , we have that:

$$\Pr(|\hat{p}_\sigma(x) - \mathbb{E}[\hat{p}_\sigma(x)]| \geq t) \leq 2 \exp(-4\pi\sigma^2 nt^2). \quad (\text{A.13})$$

838 Taking $t = \sqrt{\frac{\log(2L/\delta)}{4\pi\sigma^2 n}}$ and using the above inequality and the union bound, we have that, with
 839 probability at least $1 - \delta$, for all $l \in \{0, 1, \dots, L-1\}$:

$$|\hat{p}_\sigma(l/L) - \mathbb{E}[\hat{p}_\sigma(l/L)]| \leq \sqrt{\frac{\log(2L/\delta)}{4\pi\sigma^2 n}}. \quad (\text{A.14})$$

840 Let us now compute the expected value of $\hat{p}_\sigma(x)$.

$$\mathbb{E}[\hat{p}_\sigma(x)] = \mathbb{E} \left[\frac{1}{n\sigma} \sum_{i=1}^n \phi \left(\frac{X^{(i)} - x}{\sigma} \right) \right] \quad (\text{A.15})$$

$$= \frac{1}{n\sigma} \sum_{i=1}^n \mathbb{E} \left[\phi \left(\frac{X^{(i)} - x}{\sigma} \right) \right] \quad (\text{A.16})$$

$$= \frac{1}{\sigma} \int p(u) \phi \left(\frac{x-u}{\sigma} \right) du \equiv (p \otimes \mathcal{N}(0, \sigma^2))(x) = p_\sigma(x). \quad (\text{A.17})$$

841 Combining equation A.14 and equation A.17, we get:

$$|\hat{p}_\sigma(l/L) - p_\sigma(x)| \leq \sqrt{\frac{\log(2L/\delta)}{4\pi\sigma^2 n}}. \quad (\text{A.18})$$

Putting everything together we have:

$$d_{\text{TV}}(p_\sigma, \hat{p}_\sigma) \leq \frac{\lambda}{2L} + \frac{1}{2L\sigma^2\sqrt{2\pi e}} + \sqrt{\frac{\log(2L/\delta)}{4\pi\sigma^2 n}}.$$

Choosing $L = n \cdot \max\{\lambda, 1\}$ we get that:

$$d_{\text{TV}}(p_\sigma, \hat{p}_\sigma) \lesssim \frac{1}{n} + \frac{1}{\sigma^2 n} + \sqrt{\frac{\log n + \log(1 \vee \lambda) + \log 2/\delta}{\sigma^2 n}}.$$

842 A.2 Evolution of parameters under noise

843 *Proof of theorem 4.2:* We will use the following facts:

844 **Fact 1** (Direct corollary of the optimal coupling theorem). There exists a coupling γ of P and Q ,
 845 which samples a pair of random variables $(X, Y) \sim \gamma$ such that $\Pr_\gamma[X \neq Y] = d_{\text{TV}}(P, Q)$.

846 **Fact 2.** For any $x, y \in \mathbb{R}^d$: $d_{\text{TV}}(\mathcal{N}(x, \sigma^2 \mathbf{I}), \mathcal{N}(y, \sigma^2 \mathbf{I})) \leq \|x - y\|/2\sigma$

Proof. The KL divergence between $\mathcal{N}(\mu_1, \Sigma_1)$ and $\mathcal{N}(\mu_2, \Sigma_2)$ is

$$\text{KL}(\mathcal{N}(\mu_1, \Sigma_1), \mathcal{N}(\mu_2, \Sigma_2)) = \frac{1}{2} \left(\text{tr}(\Sigma_2^{-1} \Sigma_1) + (\mu_2 - \mu_1) \Sigma_2^{-1} (\mu_2 - \mu_1) - d + \log \frac{|\Sigma_2|}{|\Sigma_1|} \right).$$

Applying this general result to our case:

$$\text{KL}(\mathcal{N}(x, \sigma^2 \mathbf{I}), \mathcal{N}(y, \sigma^2 \mathbf{I})) = \frac{1}{2} \left(\frac{\|x - y\|^2}{\sigma^2} \right).$$

847 We conclude by applying Pinsker's inequality. □

848 A corollary of Fact 2 and the optimal coupling theorem is the following:

849 *Fact 3.* Fix arbitrary $x, y \in \mathbb{R}^d$. There exists a coupling $\gamma_{x,y}$ of $\mathcal{N}(0, \sigma^2 \mathbf{I})$ and $\mathcal{N}(0, \sigma^2 \mathbf{I})$, which
850 samples a pair of random variables $(Z, Z') \sim \gamma_{x,y}$ such that $\Pr_{\gamma_{x,y}}[x + Z \neq y + Z'] = \|x - y\|/2\sigma$.

851 Now let us denote by $\tilde{P} = P \otimes \mathcal{N}(0, \sigma^2 \mathbf{I})$ and $\tilde{Q} = Q \otimes \mathcal{N}(0, \sigma^2 \mathbf{I})$. To establish our claim in the
852 theorem statement, it suffices to exhibit a coupling $\tilde{\gamma}$ of \tilde{P} and \tilde{Q} which samples a pair of random
853 variables $(\tilde{X}, \tilde{Y}) \sim \tilde{\gamma}$ such that: $\Pr_{\tilde{\gamma}}[\tilde{X} \neq \tilde{Y}] \leq d_{\text{TV}}(P, Q) \cdot \frac{D}{2\sigma}$. We define coupling $\tilde{\gamma}$ as follows:

854 Let us argue the following:

855 **Lemma A.3.** *The afore-described sampling procedure $\tilde{\gamma}$ is a valid coupling of \tilde{P} and \tilde{Q} .*

856 *Proof.* We need to establish that the marginals of $\tilde{\gamma}$ are \tilde{P} and \tilde{Q} . □

857 **Lemma A.4.** *Under the afore-described coupling $\tilde{\gamma}$: $\Pr_{\tilde{\gamma}}[\tilde{X} \neq \tilde{Y}] \leq d_{\text{TV}}(P, Q) \cdot \frac{D}{2\sigma}$.*

858 *Proof.* Event $\tilde{X} \neq \tilde{Y}$ happens, when $X \neq Y$ and, conditioning on this event, when $X + Z \neq$
859 $Y + Z'$ happens. By Fact 1, $\Pr_{\gamma}[X \neq Y] = d_{\text{TV}}(P, Q)$. By Fact 3, for any realization of (X, Y) ,
860 $\Pr_{\gamma_{X,Y}}[X + Z \neq Y + Z'] = \frac{\|X - Y\|}{2\sigma} \leq \frac{D}{2\sigma}$, where we used that P and Q are supported on a set
861 with diameter D . Putting the above together, the claim follows. □

862 □

863 A.3 Auxiliary Lemmas

864 **Lemma A.5** (Lipschitzness of the empirical density). *For a collection of points $X^{(1)}, \dots, X^{(n)}$
865 consider the function $\hat{p}_\sigma(x) = \frac{1}{n\sigma} \sum_{i=1}^n \phi\left(\frac{X^{(i)} - x}{\sigma}\right)$, where $\phi(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2}$ is the Gaussian
866 kernel. Then p_σ is $\left(\frac{1}{\sigma^2 \sqrt{2\pi e}}\right)$ -Lipschitz.*

867 *Proof.* Let us compute the derivative of \hat{p}_σ :

$$\hat{p}'_\sigma(x) = \frac{1}{n\sigma} \sum_{i=1}^n \frac{d}{dx} \phi\left(\frac{x - X^{(i)}}{\sigma}\right) \tag{A.19}$$

$$= \frac{1}{\sqrt{2\pi}n\sigma} \sum_{i=1}^n \exp\left(-\frac{(X^{(i)} - x)^2}{2\sigma^2}\right) \frac{X^{(i)} - x}{\sigma^2} \tag{A.20}$$

$$\leq \frac{1}{\sqrt{2\pi}\sigma^2} \max_u \exp(-u^2/2) u \tag{A.21}$$

$$\leq \frac{1}{\sigma^2 \sqrt{2\pi e}}. \tag{A.22}$$

868 □

869 **Lemma A.6** (Lipschitzness of a density convolved with a Gaussian). *Let p be a density that is
870 λ -Lipschitz. Let $p_\sigma = p \otimes \mathcal{N}(0, \sigma^2 \mathbf{I})$. Then, p_σ is also λ -Lipschitz.*

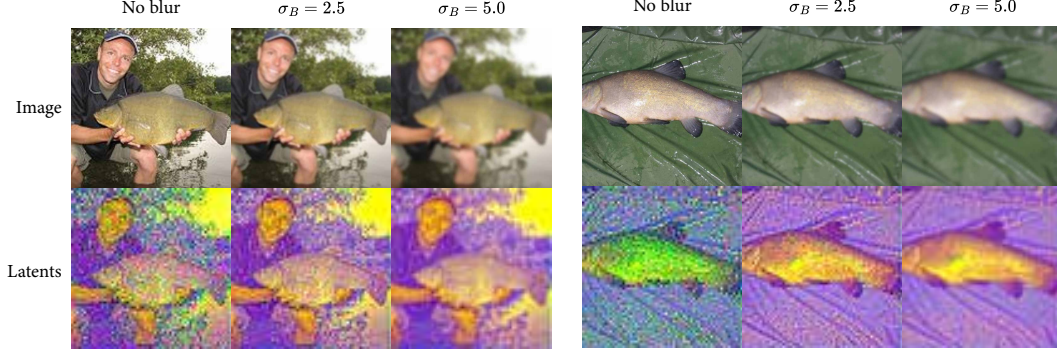


Figure 8: Visualization of images, latents, and decoded latents as image is progressively blurred with increasing Kernel Standard Deviation σ_B

871 *Proof.* Let us denote with $\phi_\sigma(\cdot)$ the Gaussian density with variance σ^2 . We have that:

$$p_\sigma(x) - p_\sigma(y) = \int (p(x - \tau) - p(y - \tau)) \phi_\sigma(\tau) d\tau \Rightarrow \quad (\text{A.23})$$

$$|p_\sigma(x) - p_\sigma(y)| \leq \int |p(x - \tau) - p(y - \tau)| \phi_\sigma(\tau) d\tau \quad (\text{A.24})$$

$$\leq \lambda |x - y| \cdot \int \phi_\sigma(\tau) d\tau \quad (\text{A.25})$$

$$= \lambda |x - y|. \quad (\text{A.26})$$

872

□

873 B Additional corruption visualizations

874 B.1 ImageNet corruptions

875 Figure 8 shows how blurring corruptions manifest in the compressed latent space of the Stable
876 Diffusion VAE Encoder [34]. As the blurring strength is increased, the high frequency details,
877 observed as noise in the latent space, disappear. Moreover, the visualized latent also becomes more
878 purple, which indicates that the second component of the latents, visualized as green, approaches
879 zero. This likely suggests that this component encodes high-frequency information.

880 B.2 CIFAR-10 corruptions

881 Figures 9a, 9b and 10 show gaussian blur, motion blur, and JPEG corrupted CIFAR-10 images
882 respectively at different levels of severity. Table 3 shows results for JPEG compressed data at
883 different levels of compression. We also tested our method for motion blurred data using severity 3,
884 obtaining a best FID of 5.85 (compared to 8.79 of training on only the clean data).

Table 3: Results for learning from JPEG compressed data on CIFAR-10.

Method	Dataset	Clean (%)	Corrupted (%)	JPEG Compression (Q)	$\bar{\sigma}_{t_n}^{\min}$	FID
Only Clean	Cifar-10	10	0	—	—	8.79
				15%	1.60	6.67
				18%	1.40	6.43
Ambient Omni	Cifar-10	10	90	25%	1.27	6.34
				50%	1.03	5.94
				75%	0.81	5.57
				90%	0.63	4.72

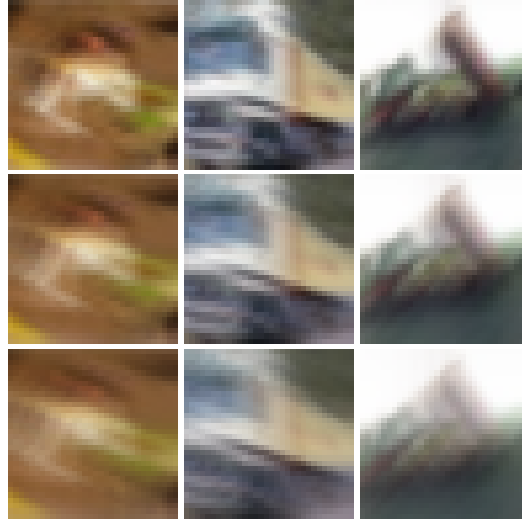
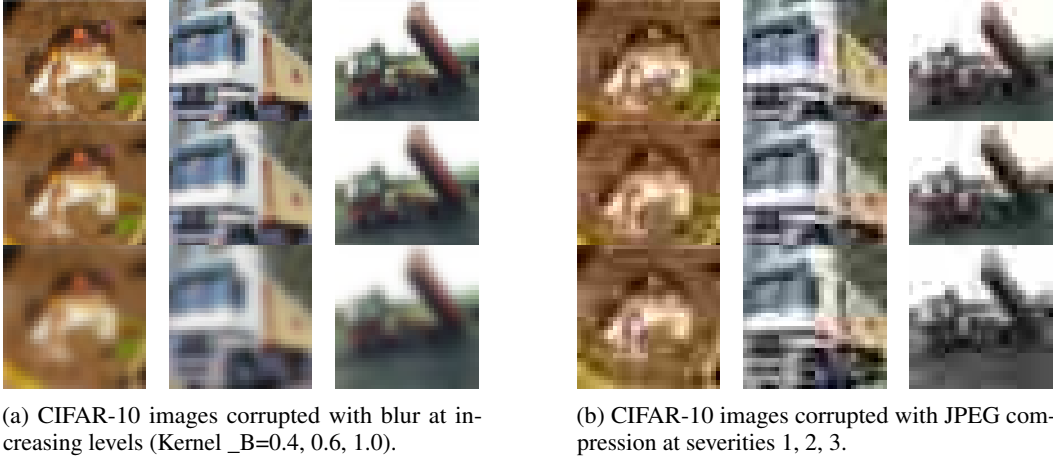


Figure 10: CIFAR-10 images corrupted with motion blur at increasing levels of corruption. The top row shows severity 1, the middle row shows severity 2, and the bottom row shows severity 3.

885 B.3 FFHQ-64x64 corruptions

886 Figures 11 and 12 show motion blur and JPEG corrupted FFHQ images respectively at different
887 levels of severity. Table 4 shows results for the gaussian blur case.

Table 4: Results for learning from blurred data, FFHQ.

Method	Dataset	Clean (%)	Corrupted (%)	Parameters Values (σ_B)	$\bar{\sigma}_{t_n}^{\min}$	FID
Only Clean	FFHQ	10	0	-	-	5.12
Ambient Omni	FFHQ	10	90	0.8	2.89	4.95
		10	90	0.6	2.12	4.65
		10	90	0.4	0.63	3.32



Figure 11: FFHQ-64x64 images corrupted with motion blur at increasing levels of corruption. The top row shows severity 1, the middle row shows severity 2, and the bottom row shows severity 3.



Figure 12: FFHQ-64x64 images corrupted with JPEG compression at increasing levels of corruption. The top row shows severity 1, the middle row shows severity 2, and the bottom row shows severity 3.

Table 5: Ablation study of ambient weight and stability buffer on Cifar-10 with 10% clean data and 90% corrupted data with blur of 0.6.

Method	FID ↓
<i>No ambient preconditioning weight and no buffer:</i>	
$\lambda_{\text{amb}}(\sigma, \sigma_{\min}) = 1 \ \& \ \sigma > \sigma_{\min}$	5.49
<i>Adding ambient preconditioning weight:</i>	
+ Weight $\lambda_{\text{amb}}(\sigma, \sigma_{\min}) = \sigma^4 / (\sigma^2 - \sigma_{\min}^2)^2$	5.36
<i>Adding stability buffer/clipping:</i>	
+ Clip $\lambda_{\text{amb}}(\sigma, \sigma_{\min})$ at 2.0	5.35
+ Clip $\lambda_{\text{amb}}(\sigma, \sigma_{\min})$ at 4.0	5.69
+ Buffer $\lambda_{\text{amb}}(\sigma, \sigma_{\min})$ at 2.0 i.e. $\sigma > \sqrt{2}\sigma_{\min}$	5.40
+ Buffer $\lambda_{\text{amb}}(\sigma, \sigma_{\min})$ at 4.0 i.e. $\sigma > (2/\sqrt{3})\sigma_{\min}$	5.34

C Ambient diffusion ablations

In this section, we ablate the ambient pre-conditioning loss weight

$$\lambda_{\text{amb}}(\sigma, \sigma_{\min}) = \sigma^4 / (\sigma^2 - \sigma_{\min}^2)^2 \quad (\text{C.1})$$

and its stability buffer, summarized in table 5. Using no ambient pre-conditioning and no buffer, we obtain an FID of 5.56 training with 10% clean data and 90% corrupted data with gaussian blur 0.6 using classifier annotations. In the same setting, adding the ambient pre-conditioning weight $\lambda_{\text{amb}}(\sigma, \sigma_{\min})$ improves FID by 0.13 points. Next, we ablated two strategies to mitigate the impact of the singularity of $\lambda_{\text{amb}}(\sigma, \sigma_{\min})$ at $\sigma = \sigma_{\min}$. The first strategy clips the ambient pre-conditioning weight at a specified maximum value $\lambda_{\text{amb}}^{\text{MAX}}$, but still trains for σ arbitrarily close to σ_{\min} . The second strategy also specifies a maximum value, but imposes a buffer

$$\sigma > \sqrt{1 + \frac{1}{\lambda_{\text{amb}}^{\text{MAX}} - 1}} \sigma_{\min} \quad (\text{C.2})$$

that restricts training to noise levels σ such that $\lambda_{\text{amb}}(\sigma, \sigma_{\min}) \leq \lambda_{\text{amb}}^{\text{MAX}}$. Clipping the ambient weight to $\lambda_{\text{amb}}^{\text{MAX}} = 2.0$ minimally improves FID to 5.35, but clipping to 4.0 significantly worsens it to 5.69. Adding a buffer at $\lambda_{\text{amb}}^{\text{MAX}} = 2.0$ slightly worsens FID to 5.40, but slackening the buffer to 4.0 minimally improves FID to 5.34. We opt for the buffering strategy in favor of the clipping strategy since performance appears convex in the buffer parameter, and because it obtains the best FID.

D Classifier annotation ablations

Balanced vs unbalanced data: We ablate the impact of classifier training data on the setting of CIFAR-10 with 10% clean data and 90% corrupted data with gaussian blur with $\sigma_B = 0.6$. When annotating with a classifier trained on the same unbalanced dataset we train the diffusion model on we obtained a best FID of 6.04, compared to the 5.34 obtained if we train on a balanced dataset.

Training iterations: We ablate the impact of classifier training iterations on the setting of cifar-10 with 10% clean data and 90% corrupted data with JPEG compression of severity 2, training the classifier with a balanced dataset. We report minute variations in the best FID, obtaining 6.50, 6.58, and 6.49 when training the classifier for 5e6, 10e6, and 15e6 images worth of training respectively.

E Training Details

E.1 Formation of the high-quality and low-quality sets.

In the theoretical problem setting we assumed the existence of a good set S_G from the clean distribution and a bad set S_B from the corrupted distribution. In practice, we do not actually possess these sets initially, but we can construct them so long as we have access to a measure of "quality". Given a function on images which tells us whether its good enough to generate or not e.g. CLIP quality

917 [46] greater than some threshold, we can define our good set S_G as the good enough images and S_B
918 as the complement. From this point on we can apply the methodology of ambient-o as developed,
919 either employing classifier annotations as in our pixel diffusion experiments, or fixed annotations as
920 in our large scale ImageNet and text-to-image experiments.

921 E.2 Datasets

922 **CIFAR-10.** CIFAR-10 [28] consists of 60,000 32x32 images of ten classes (airplane, automobile,
923 bird, cat, deer, dog, frog, horse, ship, and truck).

924 **FFHQ.** FFHQ [24] consists of 70,000 512x512 images of faces from Flickr. We used the dataset at
925 64x64 resolution for our experiments.

926 **AFHQ.** AFHQ [6] consists of 5,653 images of cats, 5,239 images of dogs and 5,000 images of
927 wildlife, for a total of 15,892 images.

928 **ImageNet.** ImageNet [12] consists of 1,281,167 images of variable resolution from 1000 classes.

929 **Conceptual Captions.** Conceptual Captions [39] consists of 12M (image url, caption) pairs.

930 **Segment Anything-1B.** Segment Anything [27] consists of 11.1M high-resolution images anno-
931 tated with segmentation masks. Since the original dataset did not have real captions, we use the same
932 LLaVA generated captions created by the MicroDiffusion [37] paper.

933 **JourneyDB.** JourneyDB consists of 4.4M synthetic image-caption pairs from Midjourney [44].

934 **DiffusionDB.** DiffusionDB consists of 14M synthetic image-caption pairs, mostly generated from
935 Stable Diffusion models [47]. We use the same 10.7M quality-filtered subset created by the MicroD-
936 iffusion paper [37].

937 E.3 Diffusion model training

938 **CIFAR-10.** We use the EDM [22] codebase as a reference to train class-conditional diffusion
939 models on CIFAR-10. The architecture is a Diffusion U-Net [41] with ~55M parameters. We use
940 the Adam optimizer [26] with learning rate 0.001, batch size 512, and no weight decay. While the
941 original EDM paper trained for 200×10^6 images worth of training, when training with corrupted
942 data we saw best results around 20×10^6 images. On a single 8xV100 node we achieved a throughput
943 of 0.8s per 1k images, for an average of 4.4h per training run.

944 **FFHQ.** Same as for CIFAR-10, except learning was set to $2e - 4$, we trained for a maximum of
945 100×10^6 images worth of training, and saw best results around 30×10^6 images worth.

946 **AFHQ.** Same as FFHQ.

947 **ImageNet.** We use the EDM2 [23] codebase as a reference to train class-conditional diffusion
948 models on ImageNet. The architecture is a Diffusion U-Net [41] with ~125M parameters. We use
949 the Adam optimizer [26] with reference learning rate 0.012, batch size 2048, and no weight decay.
950 Same as the original codebase, we trained for ~2B worth of images. On 32 H200 GPUs, XS models
951 took ~3 days to train, while XXL models took ~7 days.

952 **MicroDiffusion.** We use the MicroDiffusion codebase as a reference to train text-to-image models
953 on an academic budget [37]. We follow their recipe exactly, changing only the standard denoising
954 diffusion loss to the ambient diffusion loss. The architecture is a Diffusion Transformer [32] utilizing
955 Mixture-of-Experts (MoE) feedforward layers [40, 21], with ~1.1B parameters. We use the
956 AdamW optimizer [26] with reference learning rates $2.4e - 4/8e - 5/8e - 5/8e - 5$ for each of the
957 four phases and batch size 2048 for all phases. On 8 H200 GPUs, training takes ~2 days to train.

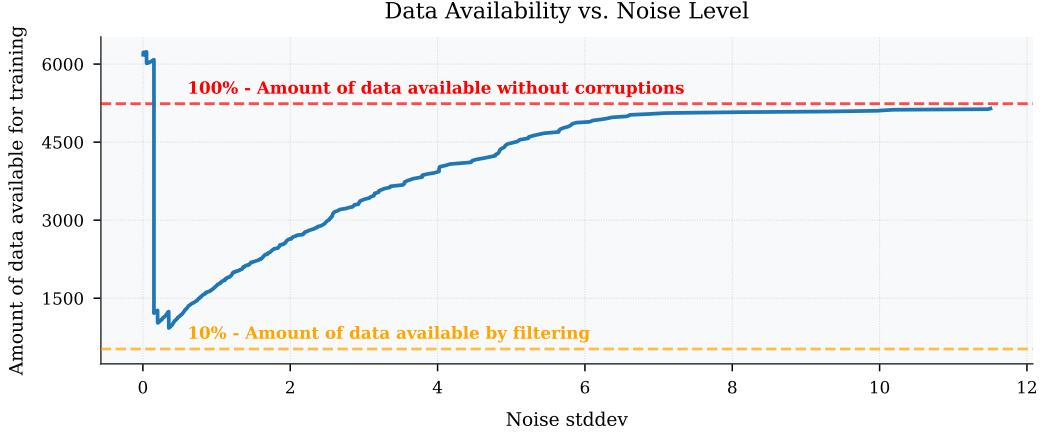


Figure 13: Amount of samples available at each noise level when training a generative model for dogs in the following setting: (1) we have 10% of the dogs dataset uncorrupted, (2) we have the other 90% of the dogs dataset corrupted with gaussian blur with $\sigma_B = 0.6$, and (3) we have 100% of the clean dataset of cats. At low noise levels, we can train on both the high quality dogs and a lot of the cats, resulting in $> 100\%$ of samples available relative to the original dogs dataset size. As the noise level starts to increase, we stop being able to use the out-of-distribution cat samples, but start gaining some blurry dog samples. As the noise level approaches the maximum all the blurry dogs become available for training, such that the amount of data available approaches 100%.

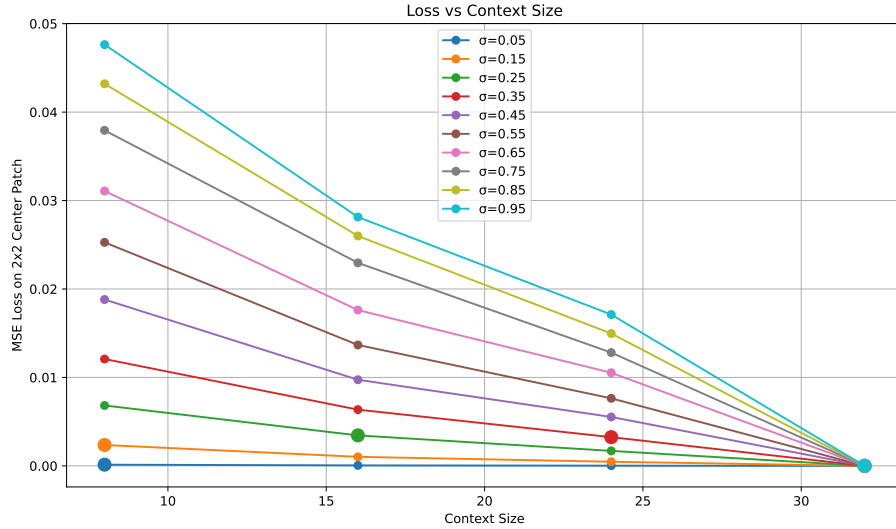


Figure 14: ImageNet-512x512: denoising loss of an optimally trained model, measured at 2×2 center patch, as we increase the context size given to the model (horizontal axis) and the noise level (different curves). As expected, for higher noise, more context is needed for optimal denoising. The large dot on each curve marks the point where the loss nearly plateaus.

958 E.4 Classifier training

959 Classifier training is done using the same optimization recipe (optimizer, learning rate, batch size,
 960 etc.) as diffusion model training, except we change the architecture to an encoder-only "Half-Unet",
 961 simply by removing the decoder half of the original UNet architecture.

962 F Additional Figures

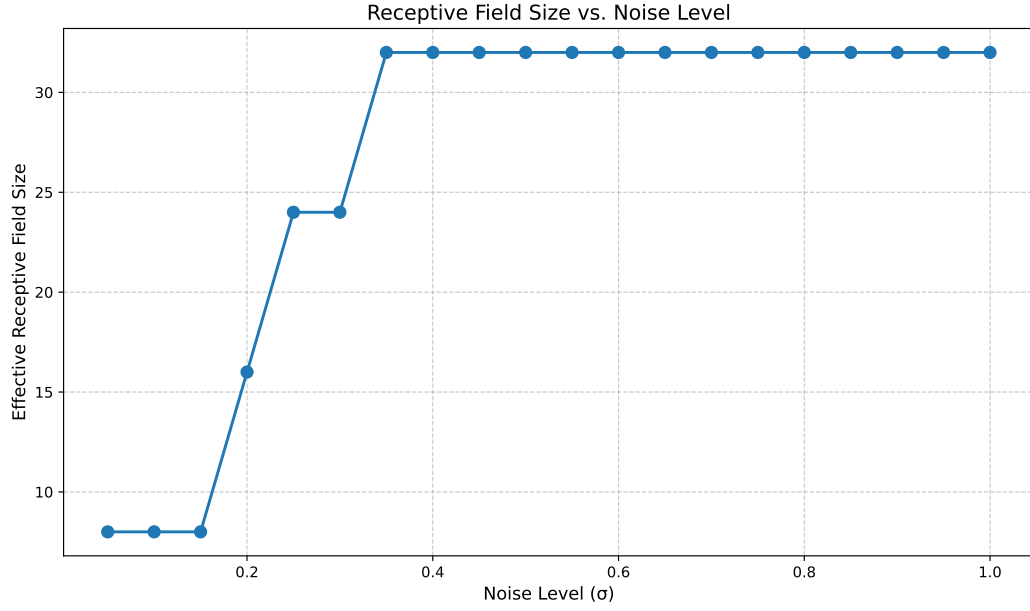


Figure 15: ImageNet-512x512: context size needed to be within $\epsilon = 1e - 3$ of the optimal loss for different noise levels. As expected, for higher noise, more context is needed for optimal denoising.

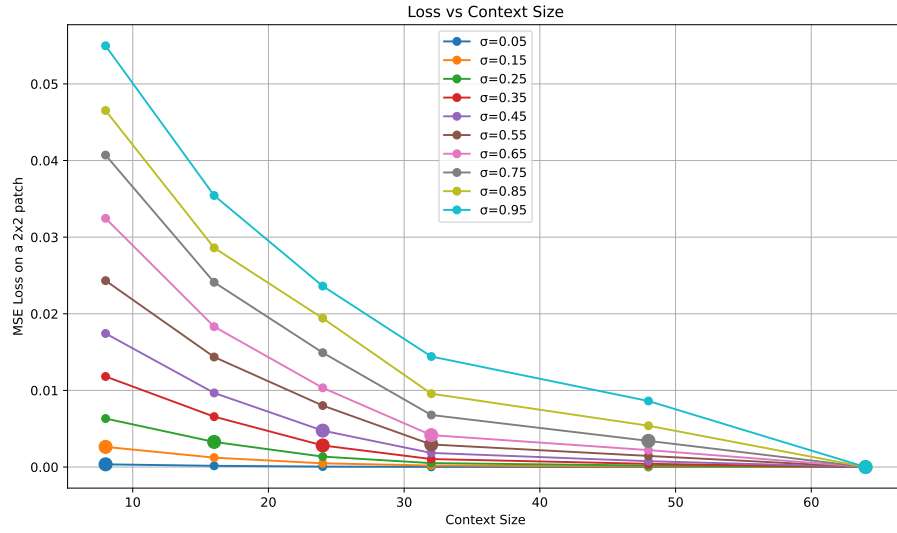


Figure 16: FFHQ: denoising loss of an optimally trained model, measured at 2×2 center patch, as we increase the context size given to the model (horizontal axis) and the noise level (different curves). As expected, for higher noise, more context is needed for optimal denoising. The large dot on each curve marks the point where the loss nearly plateaus.

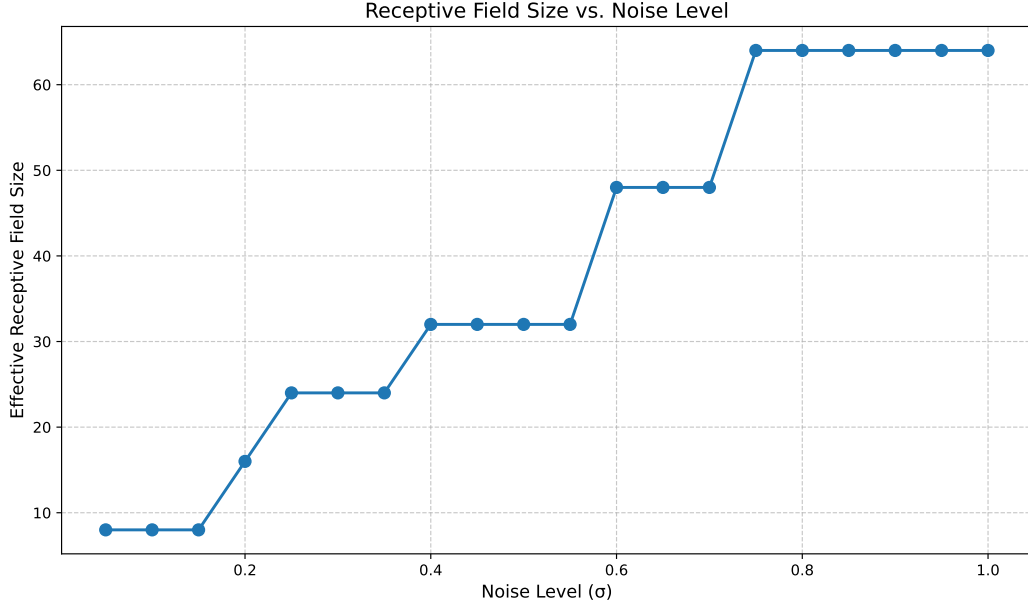


Figure 17: FFHQ: context size needed to be within $\epsilon = 1e - 3$ of the optimal loss for different noise levels. As expected, for higher noise, more context is needed for optimal denoising.

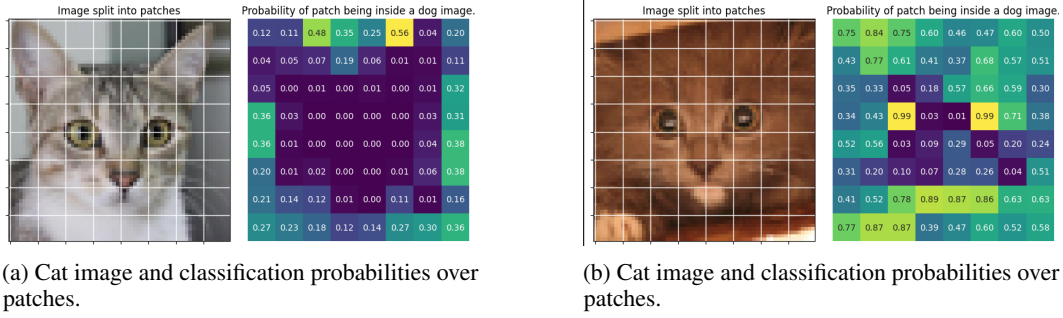


Figure 18: Two examples of cats from the AFHQ dataset. We partition each cat into non overlapping patches and we compute the probabilities of the patch belonging to an image of a dog using a cats vs dogs classifier trained on patches. The cat on the right has a lot more patches that could belong to a dog image according to the classifier, possibly due to the color or the texture of the fur.

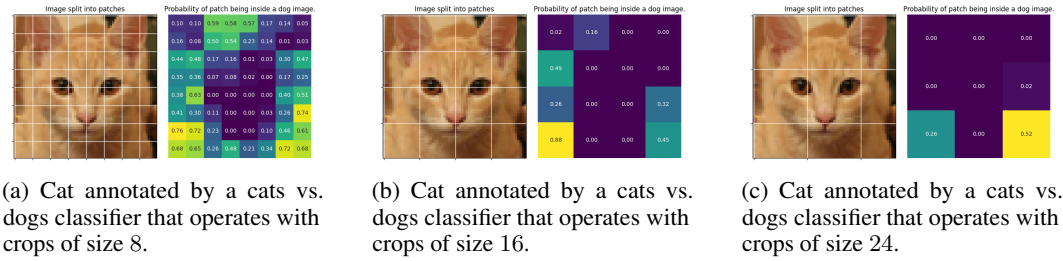


Figure 19: Patch-based annotations of a cat image from AFHQ using cats vs. dogs classifiers trained on different patch sizes.

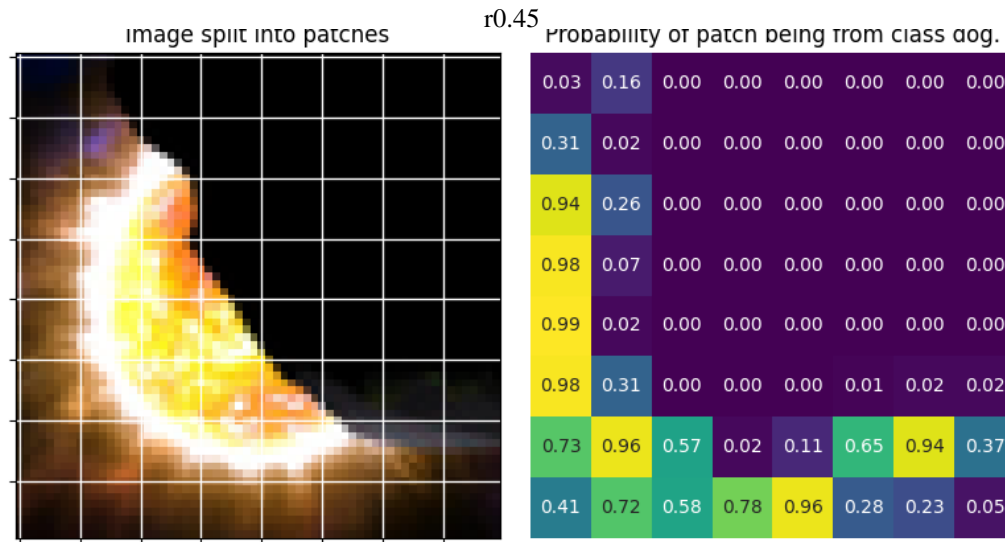
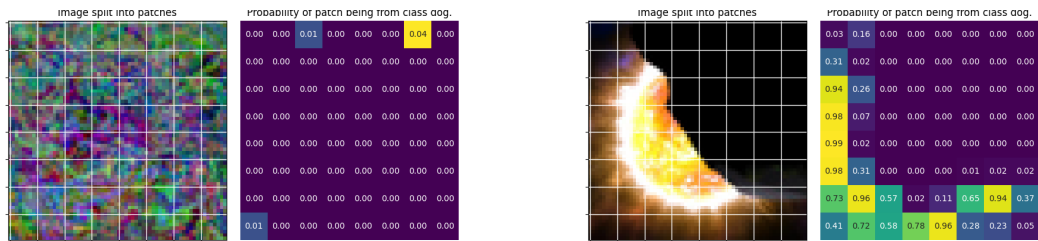


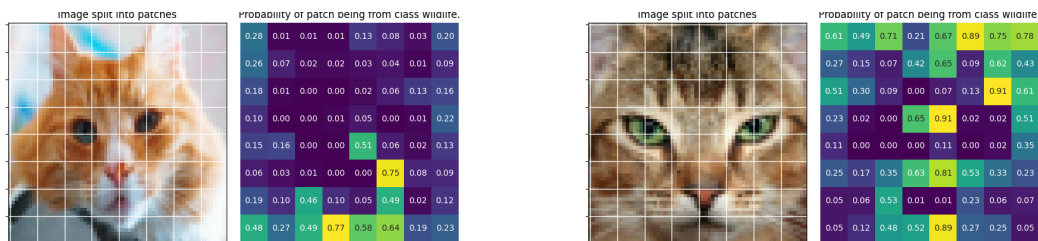
Figure 20: Patch level probabilities for dogness in a synthetic dog image (procedural program). The cat has more useful patches than this non-realistic procedural program.



(a) Synthetic image and classification probabilities over patches.

(b) Synthetic image and classification probabilities over patches.

Figure 21: Two examples of procedurally generated images. We partition each image into non overlapping patches and we compute the probabilities of the patch belonging to an image of a dog using a synthetic image vs dogs classifier trained on patches. The image on the right has a lot more patches that could belong to a dog image according to the classifier, possibly due to the color or the texture.



(a) Cat image and classification probabilities over patches.

(b) Cat image and classification probabilities over patches.

Figure 22: Two examples of cat images. We partition each image into nonoverlapping patches and we compute the probabilities of the patch belonging to an image of wildlife using a cats vs wildlife classifier trained on patches. The image on the right has a lot more patches that could belong to a wildlife image according to the classifier, possibly due to the color or the texture.



(a) Example batch.

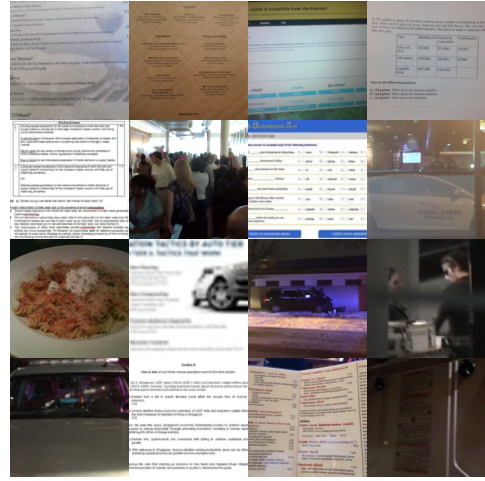


(b) Noisy batch.

Figure 23: Example batch.



(a) Highest quality images from CC12M according to CLIP.

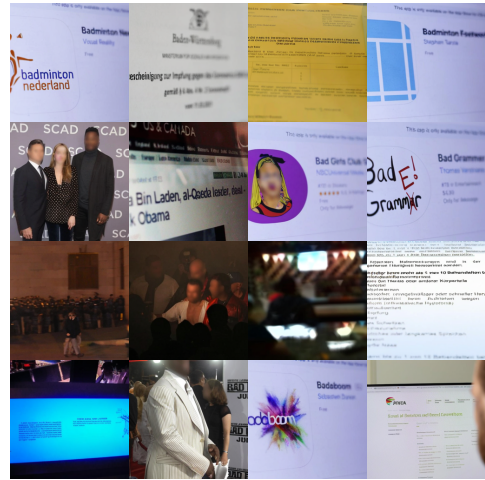


(b) Lowest quality images from CC12M according to CLIP.

Figure 24: CLIP annotations for quality of images from CC12M.



(a) Highest quality images from SA1B according to CLIP.

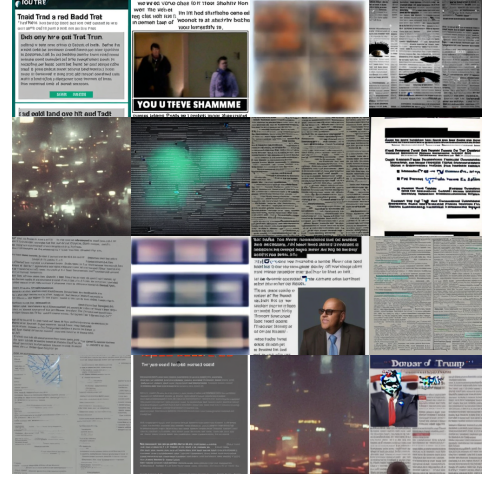


(b) Lowest quality images from SA1B according to CLIP.

Figure 25: CLIP annotations for quality of images from SA1B.



(a) Highest quality images from DiffDB according to CLIP.

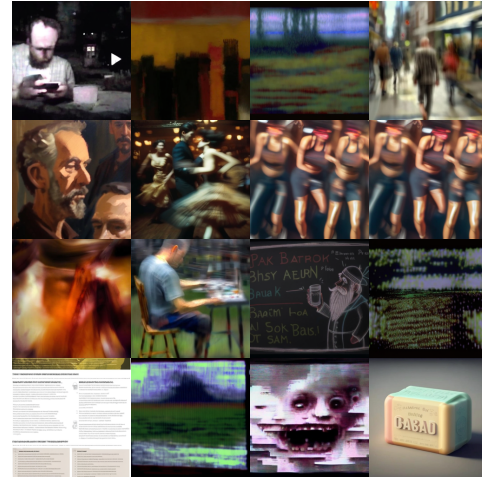


(b) Lowest quality images from DiffDB according to CLIP.

Figure 26: CLIP annotations for quality of images from DiffDB.



(a) Highest quality images from JDB according to CLIP.



(b) Lowest quality images from JDB according to CLIP.

Figure 27: CLIP annotations for quality of images from JDB.

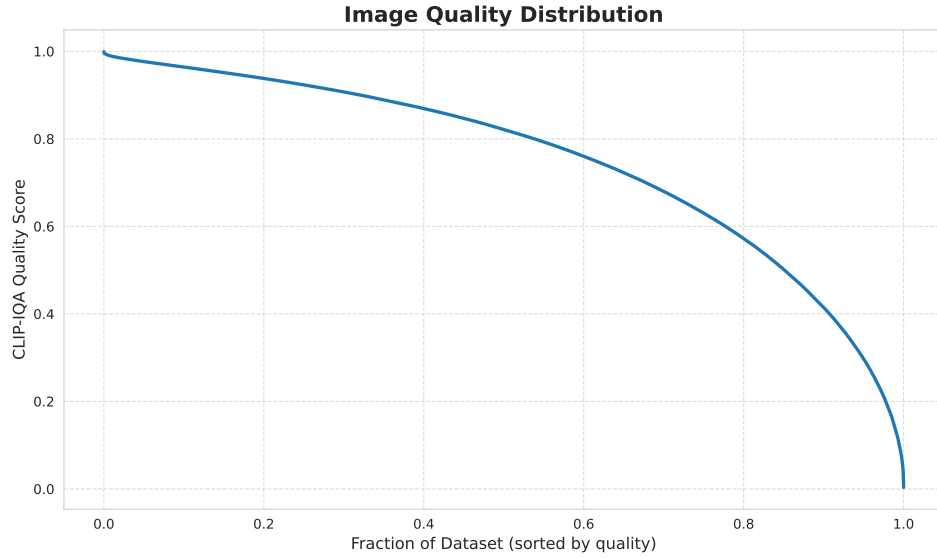


Figure 28: Distribution of image qualities according to CLIP for ImageNet-512.



Figure 29: Examples of mode collapse due to fine-tuning in contrast with our ambient-o model.

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: No research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [NA]

Justification: No important, original, or non-standard usage of LLMs in the paper.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.

References

- [1] Asad Aali, Marius Arvinte, Sidharth Kumar, and Jonathan I Tamir. “Solving Inverse Problems with Score-Based Generative Priors learned from Noisy Data”. In: *arXiv preprint arXiv:2305.01166* (2023) (cit. on pp. 1, 3).
- [2] Asad Aali, Giannis Daras, Brett Levac, Sidharth Kumar, Alex Dimakis, and Jon Tamir. “Ambient Diffusion Posterior Sampling: Solving Inverse Problems with Diffusion Models Trained on Corrupted Data”. In: *The Thirteenth International Conference on Learning Representations*. 2025. URL: <https://openreview.net/forum?id=qeXcMutEZY> (cit. on p. 1).
- [3] Weimin Bai, Yifei Wang, Wenzheng Chen, and He Sun. “An Expectation-Maximization Algorithm for Training Clean Diffusion Models from Corrupted Observations”. In: *arXiv preprint arXiv:2407.01014* (2024) (cit. on p. 1).
- [4] Manel Baradad, Chun-Fu Chen, Jonas Wulff, Tongzhou Wang, Rogerio Feris, Antonio Torralba, and Phillip Isola. “Procedural Image Programs for Representation Learning”. In: *Advances in Neural Information Processing Systems*. Ed. by Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho. 2022. URL: <https://openreview.net/forum?id=wJHTgIoEOP> (cit. on p. 8).
- [5] Ashish Bora, Eric Price, and Alexandros G Dimakis. “AmbientGAN: Generative models from lossy measurements”. In: *International conference on learning representations*. 2018 (cit. on p. 1).
- [6] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. “StarGAN v2: Diverse image synthesis for multiple domains”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2020, pp. 8188–8197 (cit. on p. 27).

- 693 [7] Xiaoliang Dai et al. *Emu: Enhancing Image Generation Models Using Photogenic Needles in*
694 *a Haystack*. 2023. arXiv: [2309.15807 \[cs.CV\]](#) (cit. on pp. 1, 3, 9).
- 695 [8] Giannis Daras, Yeshwanth Cherapanamjeri, and Constantinos Costis Daskalakis. “How Much
696 is a Noisy Image Worth? Data Scaling Laws for Ambient Diffusion.” In: *The Thirteenth*
697 *International Conference on Learning Representations*. 2025. URL: [https://openreview.net/forum?id=qZwtPEw2qN](#) (cit. on p. 1).
- 698 [9] Giannis Daras, Yuval Dagan, Alexandros G Dimakis, and Constantinos Daskalakis. “Consistent
699 diffusion models: Mitigating sampling drift by learning to be consistent”. In: *arXiv preprint*
700 *arXiv:2302.09057* (2023) (cit. on pp. 1, 3).
- 701 [10] Giannis Daras, Alexandros G Dimakis, and Constantinos Daskalakis. “Consistent Diffusion
702 Meets Tweedie: Training Exact Ambient Diffusion Models with Noisy Data”. In: *arXiv preprint*
703 *arXiv:2404.10177* (2024) (cit. on pp. 1, 3).
- 704 [11] Giannis Daras, Kulin Shah, Yuval Dagan, Aravind Gollakota, Alex Dimakis, and Adam Klivans.
705 “Ambient Diffusion: Learning Clean Distributions from Corrupted Data”. In: *Thirty-seventh*
706 *Conference on Neural Information Processing Systems*. 2023. URL: [https://openreview.net/forum?id=wBJBLy9kBY](#) (cit. on pp. 1, 3).
- 707 [12] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. “ImageNet: A large-
708 scale hierarchical image database”. In: *2009 IEEE Conference on Computer Vision and Pattern*
709 *Recognition*. 2009, pp. 248–255. DOI: [10.1109/CVPR.2009.5206848](#) (cit. on p. 27).
- 710 [13] Prafulla Dhariwal and Alexander Nichol. “Diffusion models beat gans on image synthesis”. In:
711 *Advances in neural information processing systems* 34 (2021), pp. 8780–8794 (cit. on p. 4).
- 712 [14] Sander Dieleman. *Diffusion is spectral autoregression*. 2024. URL: [https://sander.ai/2024/09/02/spectral-autoregression.html](#) (cit. on p. 2).
- 713 [15] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao
714 Nguyen, Ryan Marten, Mitchell Wortsman, Dhruva Ghosh, Jieyu Zhang, et al. “DataComp: In
715 search of the next generation of multimodal datasets”. In: *arXiv preprint arXiv:2304.14108*
716 (2023) (cit. on p. 1).
- 717 [16] Dhruva Ghosh, Hanna Hajishirzi, and Ludwig Schmidt. *GenEval: An Object-Focused Frame-*
718 *work for Evaluating Text-to-Image Alignment*. 2023. arXiv: [2310.11513 \[cs.CV\]](#). URL:
719 [https://arxiv.org/abs/2310.11513](#) (cit. on p. 9).
- 720 [17] Sachin Goyal, Pratyush Maini, Zachary C Lipton, Aditi Raghunathan, and J Zico Kolter.
721 “Scaling Laws for Data Filtering—Data Curation cannot be Compute Agnostic”. In: *Proceedings*
722 *of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2024, pp. 22702–
723 22711 (cit. on p. 1).
- 724 [18] Dan Hendrycks and Thomas Dietterich. “Benchmarking neural network robustness to common
725 corruptions and perturbations”. In: *arXiv preprint arXiv:1903.12261* (2019) (cit. on p. 1).
- 726 [19] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter.
727 “Gans trained by a two time-scale update rule converge to a local nash equilibrium”. In:
728 *Advances in neural information processing systems* 30 (2017) (cit. on pp. 7, 8).
- 729 [20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. “Denoising diffusion probabilistic models”. In:
730 *Advances in Neural Information Processing Systems* 33 (2020), pp. 6840–6851 (cit. on p. 1).
- 731 [21] Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey E. Hinton. “Adap-
732 tive Mixtures of Local Experts”. In: *Neural Computation* 3.1 (Mar. 1991). eprint:
733 [https://direct.mit.edu/neco/article-pdf/3/1/79/812104/neco.1991.3.1.79.pdf](#), pp. 79–87. ISSN:
734 0899-7667. DOI: [10.1162/neco.1991.3.1.79](#). URL: [https://doi.org/10.1162/neco.1991.3.1.79](#) (cit. on p. 27).
- 735 [22] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. “Elucidating the design space of
736 diffusion-based generative models”. In: *arXiv preprint arXiv:2206.00364* (2022) (cit. on pp. 7,
737 27).
- 738 [23] Tero Karras, Miika Aittala, Jaakko Lehtinen, Janne Hellsten, Timo Aila, and Samuli Laine.
739 “Analyzing and Improving the Training Dynamics of Diffusion Models”. In: *Proc. CVPR*. 2024
740 (cit. on pp. 2, 9, 27).
- 741 [24] Tero Karras, Samuli Laine, and Timo Aila. “A style-based generator architecture for generative
742 adversarial networks”. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and*
743 *Pattern Recognition*. 2019, pp. 4401–4410 (cit. on p. 27).

- [25] Varun A Kelkar, Rucha Deshpande, Arindam Banerjee, and Mark A Anastasio. “Ambient-Flow: Invertible generative models from incomplete, noisy measurements”. In: *arXiv preprint arXiv:2309.04856* (2023) (cit. on p. 1).
- [26] Diederik P Kingma and Jimmy Ba. “Adam: A Method for Stochastic Optimization”. In: *arXiv preprint arXiv:1412.6980* (2014) (cit. on p. 27).
- [27] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. *Segment Anything*. 2023. arXiv: 2304.02643 [cs.CV]. URL: <https://arxiv.org/abs/2304.02643> (cit. on p. 27).
- [28] Alex Krizhevsky and Geoffrey Hinton. “Learning multiple layers of features from tiny images”. In: (2009) (cit. on p. 27).
- [29] Jeffrey Li et al. *DataComp-LM: In search of the next generation of training sets for language models*. 2024. arXiv: 2406.11794 [cs.LG] (cit. on pp. 1, 3).
- [30] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. *Microsoft COCO: Common Objects in Context*. 2015. arXiv: 1405.0312 [cs.CV]. URL: <https://arxiv.org/abs/1405.0312> (cit. on p. 9).
- [31] Haoye Lu, Qifan Wu, and Yaoliang Yu. “SFBD: A Method for Training Diffusion Models with Noisy Data”. In: *Frontiers in Probabilistic Inference: Learning meets Sampling*. 2025. URL: <https://openreview.net/forum?id=6HN14zuHRb> (cit. on p. 1).
- [32] William Peebles and Saining Xie. *Scalable Diffusion Models with Transformers*. 2023. arXiv: 2212.09748 [cs.CV]. URL: <https://arxiv.org/abs/2212.09748> (cit. on p. 27).
- [33] Guilherme Penedo, Hynek Kydlíček, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, Thomas Wolf, et al. “The FineWeb Datasets: Decanting the Web for the Finest Text Data at Scale”. In: *arXiv preprint arXiv:2406.17557* (2024) (cit. on p. 1).
- [34] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. *High-Resolution Image Synthesis with Latent Diffusion Models*. 2022. arXiv: 2112.10752 [cs.CV]. URL: <https://arxiv.org/abs/2112.10752> (cit. on p. 23).
- [35] François Rozet, G r me Andry, Fran ois Lanusse, and Gilles Louppe. “Learning Diffusion Priors from Observations by Expectation Maximization”. In: *arXiv preprint arXiv:2405.13712* (2024) (cit. on pp. 1, 3).
- [36] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. “Laion-5b: An open large-scale dataset for training next generation image-text models”. In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 25278–25294 (cit. on p. 1).
- [37] Vikash Sehwal, Xianghao Kong, Jingtao Li, Michael Spranger, and Lingjuan Lyu. “Stretching Each Dollar: Diffusion Training from Scratch on a Micro-Budget”. In: *arXiv preprint arXiv:2407.15811* (2024) (cit. on pp. 2, 3, 9, 27).
- [38] Kulin Shah, Alkis Kalavasis, Adam R. Klivans, and Giannis Daras. *Does Generation Require Memorization? Creative Diffusion Models using Ambient Diffusion*. 2025. arXiv: 2502.21278 [cs.LG] (cit. on p. 1).
- [39] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. “Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning”. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Ed. by Iryna Gurevych and Yusuke Miyao. Melbourne, Australia: Association for Computational Linguistics, July 2018, pp. 2556–2565. DOI: 10.18653/v1/P18-1238. URL: <https://aclanthology.org/P18-1238/> (cit. on p. 27).
- [40] Noam Shazeer, Azalia Mirhoseini, Krzysztof Mazi r, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. *Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer*. 2017. arXiv: 1701.06538 [cs.LG]. URL: <https://arxiv.org/abs/1701.06538> (cit. on p. 27).
- [41] Jiaming Song, Chenlin Meng, and Stefano Ermon. “Denoising diffusion implicit models”. In: *arXiv preprint arXiv:2010.02502* (2020) (cit. on p. 27).
- [42] Yang Song and Stefano Ermon. “Generative modeling by estimating gradients of the data distribution”. In: *Advances in Neural Information Processing Systems* 32 (2019) (cit. on p. 1).

- 803 [43] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and
804 Ben Poole. “Score-based generative modeling through stochastic differential equations”. In:
805 *arXiv preprint arXiv:2011.13456* (2020) (cit. on p. 1).
- 806 [44] Keqiang Sun, Junting Pan, Yuying Ge, Hao Li, Haodong Duan, Xiaoshi Wu, Renrui Zhang,
807 Aojun Zhou, Zipeng Qin, Yi Wang, Jifeng Dai, Yu Qiao, Limin Wang, and Hongsheng Li.
808 *JourneyDB: A Benchmark for Generative Image Understanding*. 2023. arXiv: 2307.00716
809 [cs.CV]. URL: <https://arxiv.org/abs/2307.00716> (cit. on p. 27).
- 810 [45] Antonio Torralba, Phillip Isola, and William T Freeman. *Foundations of computer vision*. MIT
811 Press, 2024 (cit. on p. 2).
- 812 [46] Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. “Exploring CLIP for Assessing the
813 Look and Feel of Images”. In: *AAAI*. 2023 (cit. on p. 27).
- 814 [47] Zijie J Wang, Evan Montoya, David Munechika, Haoyang Yang, Benjamin Hoover, and
815 Duen Horng Chau. “DiffusionDB: A Large-scale Prompt Gallery Dataset for Text-to-Image
816 Generative Models”. In: *arXiv preprint arXiv:2210.14896* (2022) (cit. on p. 27).