Blind Spatially-Varying Denoising in Image Diffusion Models

Anonymous authors
Paper under double-blind review

Abstract

Image diffusion models are usually trained with a single noise level for the whole image, so denoising can only proceed uniformly over all pixels at once. The few prior works that support spatially-varying denoising require an input noise map that specifies the noise level at every pixel location. In this paper, we propose blind spatially-varying denoising, which removes this requirement. We remove the timestep input of a pixel diffusion transformer and train it on spatially-varying noise. The model predicts the clean image, so subtracting its prediction from the input leaves the noise, and the size of this remainder gives the noise level of each patch. Any pixel-space model can measure these levels. Only one trained on spatially-varying noise can also denoise patch by patch. We show that we can add this ability to pretrained models by finetuning, including a text-to-image model. The result is still a generator with almost no loss in quality. Painting noise onto an image then becomes a single way to restore and edit it. The model removes strong spatially-varying noise, fills noise-covered regions, extends borders, and edits a photograph wherever noise is applied, without requiring a mask or noise map.

1 Introduction

A diffusion or flow model (Sohl-Dickstein et al., 2015; Ho et al., 2020; Song et al., 2021; Karras et al., 2022; Lipman et al., 2023) is told the noise level of its input. During training this level is fed to the network. During sampling a preset schedule sets it at every step. One level for the entire image means one can only denoise the whole image at once. The model cannot restore a photograph whose noise is heavy in one region and light in another, or regenerate one region while leaving the rest alone. Methods with spatially-varying levels (Schusterbauer et al., 2026; Papa et al., 2026) lift this restriction by taking a noise map, one level per location, as an extra input, which works when the map is known. But there are many situations when the noise map is unknown.

The network may not need to be told the noise level. Kadkhodaie et al. (2026) show that a diffusion model given no noise-level input can be competitive. Helbling et al. (2026) reach the same conclusion for flow matching, Sun et al. (2025) measure what removing noise conditioning costs, and Equilibrium Matching (Wang & Du, 2025) samples a noise-unconditional model by gradient descent, even from a partially noised image. All of them treat the noise level as one global number for the whole image. Instead, we ask whether a model can read and act on spatially-varying noise levels in images.

We build on pixel-space diffusion transformers that predict the clean image directly, such as JiT (Li & He, 2025). In pixel space, the noise the model sees is the noise in the image itself, and the model can compare its input with its prediction at every step. Where the noise exceeds the model's own error, the size of that residual in each patch gives the patch's noise level. The measured noise map is used once. It sets where each patch starts on the sampling path, and the sampler runs from there to the end, so clean patches barely move and noise-filled patches are generated from scratch. A clean-image predictor's update never looks at the noise map again, so an error in the map cannot compound.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

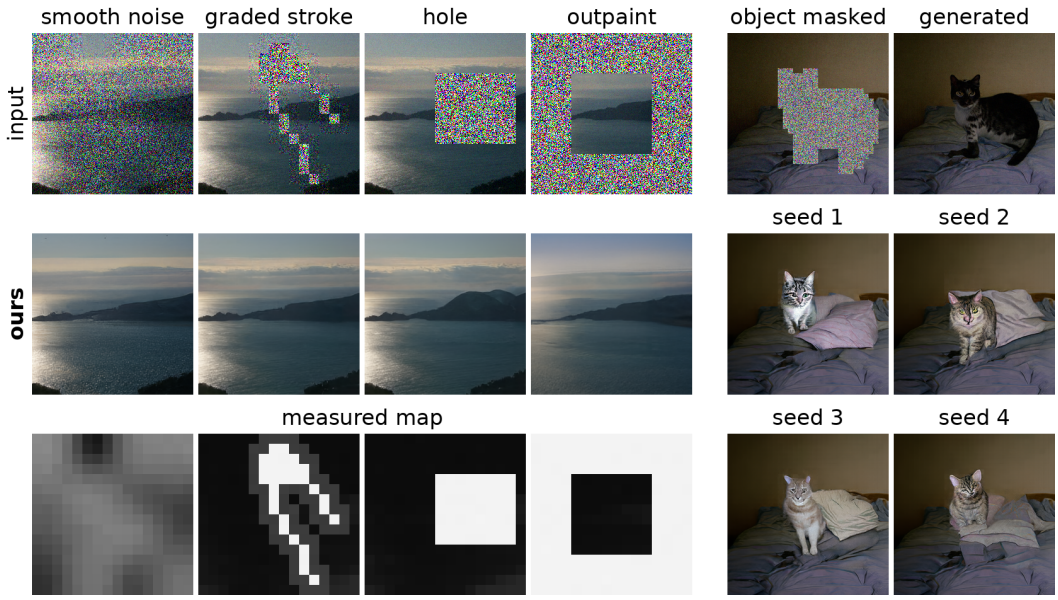


Figure 1: **Blind spatially-varying denoising:** the model reads the noise map from its input and acts on it. **Left:** restoration. Given only the corrupted image, with no noise map or mask, the model removes the noise in the inputs (JiT-B/16 trained on ImageNet (Deng et al., 2009), 256^2). The bottom row is the per-patch noise map $\mathbf{t}$ the model read from its input (brighter is noisier). **Right:** editing with painted noise, using the same recipe on a text-to-image model (finetuned MiniT2I-B/16, 512^2). A generated cat ("a cat sitting on an unmade bed with pillows") is replaced with a different cat for each seed, while the scene is unchanged.

Acting on the noise map must be trained. A model trained on uniformly noisy images reads a two-level input correctly, yet denoises the whole image at one level and leaves the noisy half corrupted (Figure 3). We show that the same model architecture trained without timestep input, and with spatially-varying noise, can restore each patch at its own noise level. The training does not have to start from scratch. Replacing a conditioned model's timestep embedding by its average and finetuning on spatially-varying noise closely matches the restoration of the model trained from scratch. The same recipe converts published JiT checkpoints and the timestep-free MiniT2I (Wang et al., 2026) text-to-image model.

The result gives restoration without having to tell what the corruption is. One model handles noise that varies spatially, holes filled with noise, brush strokes whose strength fades along the stroke, and regions beyond the border of the photograph (Figure 1). Filling a hole usually needed its location, from exemplar-based completion (Efros & Leung, 1999; Criminisi et al., 2004) to diffusion-based inpainting (Lugmayr et al., 2022). Here a hole is a region of pure noise, and the model finds it as it finds any other noise level. Because the map only sets the starting state, the noise inside the hole decides which fill the model draws. Different seeds give different objects of the same class in an unchanged scene (Figure 1, Appendix J). The painted noise still tells the model where to act. What goes away is the separate mask or map channel, and the second network built to consume it.

The same property makes the model an editor. Noise painted onto a region is the edit's starting state, so more noise means more change. Differential Diffusion (Levin & Fried, 2023) holds each pixel at the noised original until a release time. Ours starts every patch at its own level and sees the clean surroundings from the first step. Told nothing, it follows the painted strength nearly as closely as Differential Diffusion when handed the map (Table A11), while changing less of the photograph. Its clearest advantage is on images that are already noisy (Figure 7 and Table A12).

The contribution is the single network, not a better map. One model, with no extra input channel, both measures the map and acts on it. It comes within 1 dB of a model handed

a map on every noise family (Appendix K), and beats a conditioned model given the true mean noise level by 3.9–6.6 dB on spatially-varying noise (Table 2). The method is designed for strong white noise in pixel space, whether painted for an edit or present as corruption. Below $\sigma_{255} \approx 40$ the residual reflects the model's own error rather than the noise. Blacked-out holes are outside its reach. The recipe needs a pixel-space model that predicts the clean image, a setting that has only recently become competitive (Li & He, 2025; Wang et al., 2026). Every such released model is a candidate editor.

Overall, our contribution can be summarized as follows:

- We introduce blind spatially-varying denoising, a timestep-free pixel diffusion model that reads the noise level of every patch from its own residual and, once trained on spatially-varying noise, restores each patch at that level with no noise map as input.
- We show that this ability can be added to pre-trained models, including a text-to-image model, by finetuning at a small fraction of their training compute, with almost no loss in quality.
- We show that painted noise becomes a single interface for restoration and editing, such as noise removal, hole-filling, outpainting, and local image edits.

2 Related Work

Blind noise-level estimation. Kadkhodaie et al. (2026) remove the noise level σ input from a variance-exploding U-Net diffusion model and recover the noise level from the denoiser residual. Helbling et al. (2026) similarly show that the flow-matching time t is recoverable for flow matching. Sun et al. (2025) ask whether denoising models need noise conditioning at all. Wang & Du (2025) replace the time-conditional velocity with the time-invariant gradient of an implicit energy (LeCun et al., 2006; Du & Mordatch, 2019), samples by gradient descent with free step sizes and adaptive stopping, and denoises partially noised images. Energy Matching (Balcerak et al., 2025) reaches a similar landscape in two training stages. All of these assume a *single* noise level for the whole image. We port the residual estimator of Kadkhodaie et al. (2026) to x-prediction flow matching and, more importantly, apply it per patch, so that the model reads a noise map rather than a single level. Classical blind denoising also estimated a global σ, from weak-textured patches or principal components (Liu et al., 2013; Chen et al., 2015), but did so with a separate estimator run before denoising. We let the denoiser itself estimate the level, per patch, as part of denoising. Ambient Diffusion (Daras et al., 2023) and Ambient Omni (Daras et al., 2025) address a different problem: learning a clean distribution from corrupted training data, with each sample's noise level known or estimated and given to the model as a condition. Our model takes no such condition, at training or test time.

spatially-varying timesteps. Diffusion Forcing (Chen et al., 2024) trains a sequence model, for example over the frames of a video, with an independent noise level per token and conditions on those levels. Our IID training field draws per-patch levels the same way, but the model is never told them. Patch Forcing (Schusterbauer et al., 2026) and AsyncPatch (Papa et al., 2026) give an image network a per-token or 2-D timestep, enabling asynchronous and coarse-to-fine generation. The map is a conditioning *input*. We ask the complementary question: whether the map must be supplied, or can be measured by the network itself. The concurrent Equilibrium Forcing (Lillemark et al., 2026) also drops σ conditioning, but reads one level per video frame. We read and act on a level that varies across space instead.

Blind restoration. Classical blind denoisers estimate a noise level and then denoise according to that (Zhang et al., 2018; 2022; 2017; Zamir et al., 2022), and plug-and-play methods such as DiffPIR (Zhu et al., 2023) and DAPS (Zhang et al., 2025) pair a diffusion prior with a step that keeps the result close to the input, which requires knowing the noise level or mask. We benchmark them against our internally measured noise map in Section 5. They restore well when handed the true noise map. FFDNet (Zhang et al., 2018), DRUNet (Zhang et al., 2022), and DiffPIR take it as an input, and the blind models assume it is spatially homogeneous. For noise whose level varies in an unknown way they need a separate estimator. For editing, SDEdit (Meng et al., 2022) sets one noise level for the whole image, and Differential Diffusion (Levin & Fried, 2023) makes it spatial through

the sampling mechanism, holding each pixel at the noised original until a given release time. In contrast, our approach operates through the input, and the model reads the release times from the noise.

Inpainting and completion. Filling a hole by inpainting started from knowing where it is, from boundary propagation and patch copying (Bertalmio et al., 2000; Efros & Leung, 1999; Efros & Freeman, 2001; Criminisi et al., 2004; Barnes et al., 2009; Hays & Efros, 2007) to learned and diffusion inpainting (Pathak et al., 2016; Lugmayr et al., 2022), which take a binary mask. Blind inpainting (Wang et al., 2020) trains a separate stage to predict the mask. Instead, we measure a noise level. The price is that we find only holes filled with white noise.

3 Method

In this section, we describe how to train an image diffusion model for blind spatially-varying denoising. The model removes noise whose level changes across the image, without being told that level.

Setup. We train a pixel diffusion transformer (Li & He, 2025) with flow matching (Lipman et al., 2023; Liu et al., 2023; Albergo & Vanden-Eijnden, 2023) on noisy images

$$z = t\,x + (1-t)\,\varepsilon, \tag{1}$$

with x the clean image, ε white Gaussian noise, and $t \in [0,1]$ the noise level ($t = 1$ is no noise). The network outputs a prediction $\hat{x}$ of the clean image and is trained with a velocity loss and logit-normal timestep (Li & He, 2025). The only change from the standard recipe is that we remove the timestep embedding.

Residual noise estimator. The model is not told t, but it can measure it. We let d denote the number of values in the region being measured. From $z - x = (1-t)(\varepsilon - x)$ and $\mathbb{E}\|\varepsilon - x\|^2 = d + \|x\|^2$, the size of the residual $z - \hat{x}$ fixes t:

$$\hat{t} \;=\; 1 - \sqrt{\frac{\|z-\hat{x}\|^2}{d + \|\hat{x}\|^2}}. \tag{2}$$

This is exact when $\hat{x} = x$, and it uses only quantities the model computes anyway. We call this estimator the *residual internal clock*: it reads the time t off the residual. Writing $y = x + \sigma\varepsilon$, the same estimate is $\hat{\sigma} = \|y - \hat{x}\|/\sqrt{d}$ with $t = 1/(1+\sigma)$, the blind estimator of Kadkhodaie et al. (2026). The reading through the residual is reliable when the noise is larger than the model's own reconstruction error, and when it is white noise. A latent encoder would turn pixel noise into content-dependent perturbations, so we expect this approach to fail when the corruption is added in pixel space and then encoded. Noise added directly to latents, as proposed by Wang & Du (2025), is a different setting.

Per-patch noise estimator. Evaluating Equation (2) over the whole image gives one number. Evaluating it over each $P \times P$ patch gives a map $\mathbf{t}$ at patch resolution. To restore a corrupted image $y = x + \sigma\varepsilon$, we must first bring it to the form the model was trained on, $z = \mathbf{t} \odot y$. This needs the map, but the map is read from the prediction, which needs z. We therefore start from $\mathbf{t} = 0.5$ everywhere and alternate: scale, predict, re-read. Two rounds suffice, three forward passes in all, and the last $\hat{x}$ is the restored image.

What the measured noise map does in the sampler. For restoration, we place the corrupted image on the interpolant as $z = \mathbf{t} \odot y$ and integrate every patch from its own measured level to $t = 1$. We evenly space the steps over the measured remaining distance. For a model that predicts the clean image, the Euler step is then

$$dt \cdot v \;=\; \frac{1-\hat{t}}{n} \cdot \frac{\hat{x} - z}{1-\hat{t}} \;=\; \frac{\hat{x} - z}{n}, \tag{3}$$

so $\hat{t}$ cancels. Each step moves $1/n$ of the way toward $\hat{x}$, whatever the reading says. The Heun corrector cancels too when the reading at the predicted point agrees with the step just taken. A reading error, therefore, does not build up over the steps. On a uniform input, this sampler re-derives the designed schedule to within 0.006 in t and 0.01 FID at matched NFE (10k images, unguided).

The map therefore matters only for the starting state. A clean patch reads $\hat{t} \approx 1$ and does not move. A noise-filled hole reads $\hat{t} \approx 0$ and travels the full way. With classifier-free guidance (CFG) (Ho & Salimans, 2022), we read the noise level from the unguided prediction. Guidance pushes $\hat{x}$ past the model's best estimate of the clean image on purpose. The residual $z - \hat{x}$ then contains more than the noise. A reading taken from the guided prediction would make the image look noisier than it is. Once guidance switches on partway through sampling, reading from the unguided branch takes FID from 12.12 to 8.88 at cfg 4.0 (blind model, 10k images).

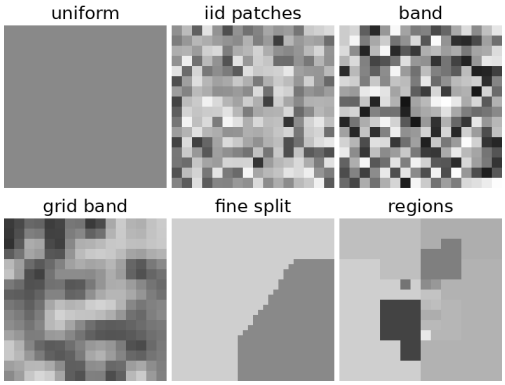


Figure 2: **Spatially-varying noise maps**: uniform, iid patches, band, grid band, fine split, and content-shaped regions. Ours (REGION) mixes all but iid patches, whose levels span the full range. In the band, patches are also drawn independently, but around one level for the whole image.

Training for locality. A model trained only on spatially uniform t has no reason to represent a local noise level (Section 4). We therefore train on spatially-varying noise maps instead of one scalar.

Our recipe (REGION) mixes the noise level maps of Figure 2. Half of all draws use content-shaped regions, which split the image along edges of the clean image and give each region its own noise level, so noise boundaries broadly follow objects. The other half use a uniform level, a per-patch band of levels around one global level, or a two-region split, drawn either per patch or at half-patch resolution (grid band, fine split), so that noise boundaries also cut through patches. Appendix D builds this mixture step by step. Unless stated otherwise, "ours" is REGION.

4 Does blindness hurt, and is the reading local?

We ask what removing the timestep input costs a generator, and whether the per-patch reading is local. As our base model, we build on JiT (Li & He, 2025) which makes the residual internal clock available in closed form at every step. All numbers are 256px JiT-B/16 on ImageNet, effective batch 1024. FID uses 50k samples against the standard ImageNet statistics, unless stated otherwise. "Tuned" means each model at its own best guidance scale.

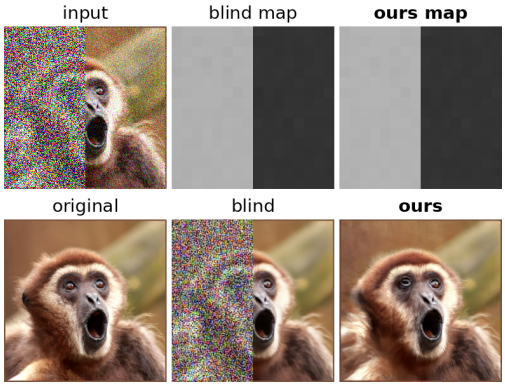


Figure 3: **Two-region noise level:** left half at t=0.3, right half at t=0.8. Both models read the two levels (top: brighter is noisier), but only the locality-trained model (ours) acts on them. The blind model leaves the noisy half full of speckle, ours restores it.

The cost is small and shrinks with training. Removing the timestep embedding costs little once each model is compared at its own best guidance scale (Table 1). At 600 epochs the blind model trails the conditioned baseline by only 0.29 tuned FID. Training from scratch on spatially-varying noise adds 0.88 tuned FID. Converting the conditioned model instead recovers it (3.65 vs 3.60).

Locality can be trained. Every model reads a uniform noise level almost exactly. On a two-region input (e.g. left half at t=0.3 and right half at t=0.8) the locality-trained model reads 0.300 and 0.799 and the blind model 0.308 and 0.799 (Figure 3). Notably, both read the two levels. In the one-pass probe of Table A5, the blind model (no timestep input, but trained with one noise level per image) separates the two halves less well (0.404 vs 0.495). This is not a misreading. The blind model denoises the whole image to a single

Table 1: **Blind, blind spatially-varying denoising training, and conversion for JiT.** *Converted* replaces the timestep embedding by its average, followed by 50k finetuning steps with a uniform (blind) or REGION field. *Restoration* follows the protocol of Table 2, n=5000. FID: 50k images, Heun 50, EMA 0.9999, tuned at each model's best guidance scale (cfg).

model					restoration, PSNR					FID	
size	res.	blind	spatial	conv.	unif.	smooth	piecew.	Pois.	hole	tuned (cfg)	no cfg
our JiT, 600 epochs											
B/16	256				26.57	23.13	21.77	27.36	14.98	3.60 (3.0)	26.0
B/16	256	✓			26.46	19.54	19.54	26.04	18.57	3.89 (3.5)	31.0
B/16	256	✓		✓	26.48	20.96	19.62	26.64	18.67	3.51 (3.5)	30.9
B/16	256	✓	✓		26.49	26.34	26.59	27.22	25.10	4.77 (3.5)	37.5
B/16	256	✓	✓	✓	26.47	26.36	26.67	27.65	25.20	3.65 (3.5)	36.2
released JiT checkpoints (Li & He, 2025)											
B/16	256				26.57	23.13	21.75	27.40	14.98	3.63 (3.0)	25.7
B/16	256	✓	✓	✓	26.47	26.37	26.68	27.66	25.27	3.73 (3.5)	35.6
L/16	256				26.71	23.03	21.50	27.40	15.61	2.44 (2.4)	13.8
L/16	256	✓	✓	✓	26.62	26.52	26.80	27.74	25.32	2.54 (3.0)	18.7
H/16	256				27.04	21.21	19.16	27.05	15.77	1.90 (2.2)	7.2
H/16	256	✓	✓	✓	26.95	26.82	27.14	28.19	25.95	2.92 (3.0)	16.7
B/32	512				27.14	25.69	25.14	27.87	15.21	4.10 (3.0)	28.6
B/32	512	✓	✓	✓	27.09	26.95	27.22	27.93	25.99	4.14 (3.5)	38.0

level, so its prediction $\hat{x}$ still contains some noise, which shrinks the residual the clock reads (Appendix G). The reading itself depends only on the estimator, so it works for every variant as long as the noise is above the crossover of Appendix C ($\sigma_{255} \approx 75$), above which the model's own error no longer enters the reading.

Both models read the map, but only one uses it. Given the two-level input, the blind model still removes noise as if the whole image had one level. Its error on the noisy half stays at 0.257 however often we iterate, while the locality-trained model's drops to 0.024 (Figure 3). The benchmark shows the same (Table 2). Scaling each patch of the input by its measured level helps the blind model by 1.3 dB on spatially-varying noise and 6.8 dB on holes, but it still trails the locality-trained model by 6.5–7 dB. It accepts a correctly scaled input yet still removes noise as if the image had a single level. The bottleneck is acting on the map. The leftover noise also explains why the blind model separates the two halves less well in the probe of Table A5.

Local denoising has to be trained. Unlike a convolutional network, whose view grows slowly with depth, a transformer lets every patch attend to all others, so nothing in the architecture favors local processing. Under uniform noise, the model never learns to treat patches differently. It learns local denoising only when training shows it spatially-varying noise.

Locality can be added to a timestep-conditioned model. Locality can also be added to an existing timestep-conditioned model. We first make the model blind: we replace its timestep embedding by a fixed vector, the embedding averaged over 16k values of t drawn from the training schedule. Every block thus receives the vector it saw on average during training. We then fine-tune with the REGION map. We train for 50k steps at batch size 1024, and learning rate 10^{-4}.

Table 1 shows that going blind by conversion costs nothing and the REGION map unlocks local restoration. The recipe also extends to checkpoints that we did not train. Converting the released JiT-B/16, L/16 and H/16 at 256px and JiT-B/32 at 512px (Li & He, 2025) with the same settings gains restoration quality, and restoration improves with model size (Table 1). The conversion costs at most 0.1 tuned FID against the source up to L/16, and 1.0 at H/16, whose FID is still falling at 50k steps (Appendix F).

5 Inpainting and editing without a map

A conditioned model cannot inpaint or edit on its own. It must be handed a noise level or mask. We therefore include two such pairings. In "cond., self-clocked", our clock supplies

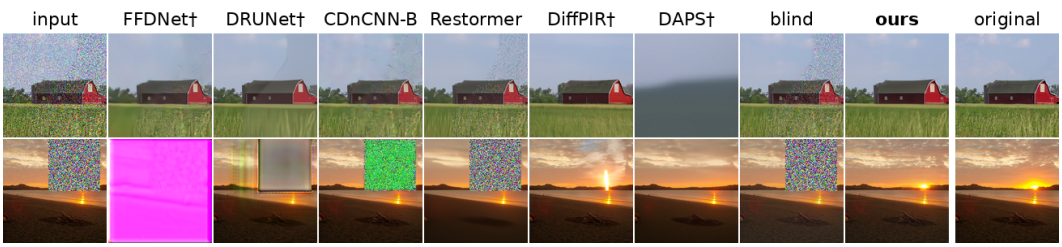


Figure 4: **Inpainting and restoration under unknown corruption.** *Top:* a piecewise noise map with two levels. *Bottom:* a mask-free hole. † marks the methods handed the true σ-map. The feedforward denoisers leave the hole untouched, DiffPIR and DAPS fill it given the mask, and ours fills it given no additional input. "Blind" is the uniform-noise model run through the same per-patch fixed point as ours (Figure 3). Ours is a 10-step Heun sample from the measured map.

one global level. In the second, a per-patch detector supplies the map (Appendix K). ImageNet carries the comparison with published methods. Models trained on FFHQ (Karras et al., 2019) and LSUN bedrooms (Yu et al., 2015) show that the results are not specific to ImageNet.

Setup. The input is $y = x + \sigma(u) \cdot \varepsilon$ with σ-noise level maps from four families no model saw in training: uniform, smooth, piecewise, and brightness-dependent noise with $\sigma^2 = a \cdot \text{intensity} + b$ (*poisson*). Unlike our models, baselines marked † are handed the true map. All methods run without class labels or guidance. "Blind, per-patch fixed point" is the uniform-noise model run through our per-patch fixed point. "blind, global clock" reads one level for the whole image. "cond., oracle mean" is the conditioned model told the true mean noise level. Our rows measure their own map and return $\hat{x}$ once without sampler. We report results on 5000 ImageNet validation images, five per class. The default band runs to $\sigma_{255} = 153$, above the crossover where our residual internal clock is clean, but outside most published denoisers' training bands (marked *).

Restoration results (varying noise). On *uniform* noise every JiT model that returns a single $\hat{x}$ ties at 26.5–26.6 dB (Table 2), so none of the gaps below comes from one model simply being a better denoiser. Where the noise varies, the measured map beats the conditioned model handed the true *mean* noise level by 3.9 dB (smooth) and 6.6 dB (piecewise). The blind model runs through the same per-patch fixed point trails by 6.5–7 dB. A map-conditioned model (Schusterbauer et al., 2026; Papa et al., 2026) does as well when fed our or the true map (Table A10). As can be seen in Figure 5, our advantage is largest where the noise is strong and structured. At low noise our own reconstruction error limits us, and denoisers trained on that band lead by up to 12 dB at σ_{255}=5 (Table A16), where our output is also about 6 dB below the noisy input. This is why clean photographs are edited with the patches read as clean copied back. From about σ_{255}=40 ours leads, by 2–8 dB above σ_{255}=75, where our map reads only noise (the

Table 2: **Inpainting and restoration results** (ImageNet val). †handed the true σ-map; "our map": handed our measured map. *outside the published training band. ‡DAPS assumes a near-noiseless measurement and is reported on holes only. "Mask": handed the hole as a binary mask, 20 steps. "Sampled": 10 Heun steps from the measured map.

	varying noise, PSNR				hole	
method	unif.	smooth	piec.	pois.	PSNR	LPIPS
corrupted input	12.25	11.63	12.34	14.98	14.67	0.318
published restorers						
FFDNet†*	21.71	21.05	21.21	25.70	12.68	0.784
DRUNet†*	22.30	21.56	21.96	26.24	19.73	0.291
DRUNet*, our map	22.30	21.58	21.96	26.23	19.26	0.323
CDnCNN-B*	21.35	20.50	20.63	25.83	17.18	0.244
Restormer*	22.84	20.45	18.54	25.69	15.37	0.350
DiffPIR†	22.67	22.69	23.49	24.30	24.45	0.084
DAPS†‡	–	–	–	–	24.48	0.176
JiT-B/16, conditioned						
oracle mean $\bar{t}$	26.58	22.40	20.01	27.25	16.25	0.329
self-clocked	26.57	23.13	21.77	27.36	14.98	0.749
Blended Diffusion, mask	–	–	–	–	17.04	0.245
RePaint, mask	–	–	–	–	23.99	0.135
JiT-B/16, blind (told nothing)						
global clock	26.51	18.22	18.28	25.45	11.79	0.726
per-patch fixed point	26.46	19.54	19.54	26.04	18.57	0.283
ours, one $\hat{x}$	26.49	26.34	26.59	27.22	25.10	0.111
ours, sampled	25.77	25.32	25.72	25.89	24.97	0.089

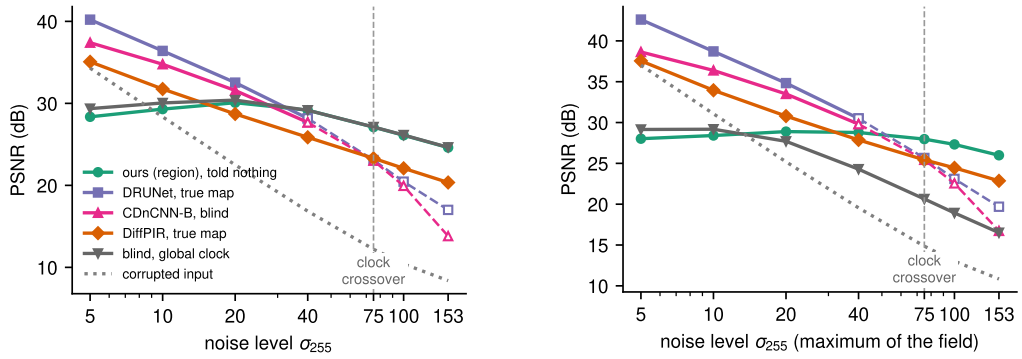


Figure 5: **PSNR against noise level** (ImageNet, n=200 per point). **Left:** uniform noise. **Right:** piecewise noise. Ours and the blind model are told nothing, DRUNet and DiffPIR get the true map. Dashed: outside a model's training band. At low noise our own reconstruction error limits us and trained denoisers lead. From about σ_{255}=40 ours leads. The vertical line marks where our measured internal map starts reading only noise (Appendix C). The blind model matches ours on uniform noise but not on piecewise noise.

vertical line, Appendix C). This is mostly against denoisers outside their training band, and the fair comparison is a model handed a per-patch map (Table A10). The blind model matches ours on uniform noise but falls 7–9 dB behind on piecewise noise. Models trained on FFHQ and LSUN show the same pattern (Table A7).

Inpainting (hole). We replace the noise map with patch-aligned holes (pure noise inside, $\sigma_{255} \approx 2500$; $\sigma_{255} \approx 6$ outside) and give no model a mask. The point is the interface: one image carries both content and corruption. Inside a hole, the one-step prediction $\hat{x}$ averages all plausible fills into a blur, so we sample to $t = 1$ instead. This costs 0.13 dB of PSNR and improves LPIPS from 0.111 to 0.089 (Figure 4). DiffPIR, handed the true σ-map (for a hole, the mask), is 0.5 dB below our sample and slightly better in LPIPS (0.084 vs 0.089). A single global noise level falls below the corrupted input (11.79 vs 14.67 dB). It reads the hole's noise as the image's and over-denoises everything. Holes without noise are not found, since the clock measures noise, not missing content.

Strokes and borders. A graded stroke has no mask, only a continuous field. On weak strokes DRUNet (with map) leads. As the stroke gets stronger, ours barely drops while the denoisers collapse (Table A15). Padding the canvas with noise makes outpainting the same problem, and ours beats DiffPIR baseline handed the mask.

Editing by painted noise. Because the map only sets the starting state, noise painted onto a region is the edit's initial condition. The model measures where the noise is and how strong it is, and changes the photograph there and by that much. A stroke of noise over the border between two spliced photographs is filled from both sides (Figure 6).

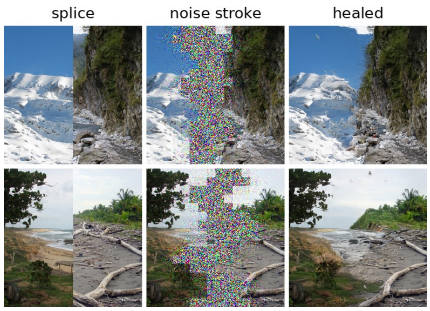


Figure 6: **Editing with noise.** Two photographs are spliced (left), a stroke of noise is painted over the middle. The model fills it from the measured map (right, 10-step sample, REGION).

In Figure 7, we compare our model with SDEdit (Meng et al., 2022), which uses one strength for the whole image and pastes the original back outside the painted region, and with Differential Diffusion (Levin & Fried, 2023), which is handed the painted strength map and holds each pixel at the noised original until its release time. In a second variant we also hand Differential Diffusion the photograph's true noise map, so held pixels carry only the noise they lack (our modification). Blended Diffusion (Avrahami et al., 2022) and RePaint (Lugmayr et al., 2022) get the painted region as a binary mask, with 5 resampling passes per step for RePaint. All four are our implementations on the conditioned JiT-B/16 with the class label and 20 steps. DiffPIR (Zhu

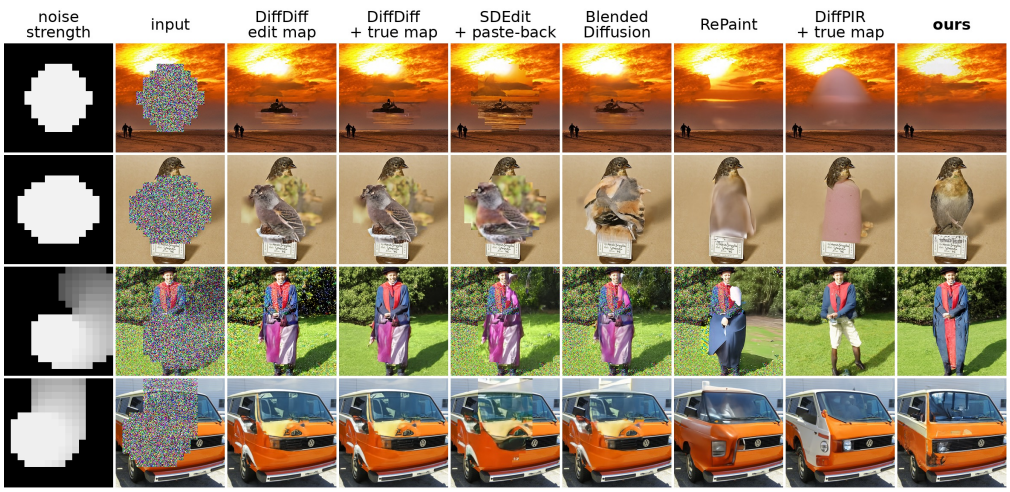


Figure 7: **Editing examples compared to baselines.** Columns: the painted noise strength, the input, Differential Diffusion handed the edit map, and also the true noise map (our modification), SDEdit with the original pasted back, Blended Diffusion and RePaint handed the painted region as a mask, DiffPIR handed the true noise map, and ours, handed nothing. In row 3 the photograph is already noisy. The baselines invent new objects (a boat, a different bird), keep the photograph's noise or change the gown. DiffPIR restores instead of editing. Images from ImageNet val.

et al., 2023) gets the true noise map. Handed only the edit map, Differential Diffusion treats the photograph's noise as content and keeps it. Handed the true noise map, it cleans the photograph, but it starts every pixel of the stroke near pure noise with the surroundings noised too, so the edit is invented without its context and objects change identity (Figure 7). Ours starts each patch at its own level and sees the clean surroundings from the first step.

On 1000 validation images per setting with a random thick graded stroke, ours is closer to the clean photograph than Differential Diffusion handed the edit map on 90% of clean photographs and on 99–100% from σ_{255}=25 up, and SDEdit trails both on every edit (Table A12). On 250 images with large graded strokes, ours reaches LPIPS 0.136 against 0.176–0.229 for the mask-based methods and DiffPIR on clean photographs, and 0.227 against 0.408–0.782 under spatially-varying noise. DiffPIR does well on clean photographs because it barely edits. On Differential Diffusion's own adherence benchmark, ours follows the painted strength almost as closely without being handed the map (Table A11).

LPIPS to the original measures preservation, not the edit. Human raters therefore saw the noisy input and two edits of it, ours and one baseline, and asked which is the better edit (60 images per setting, 25 votes per pair, 13,506 votes in total). Details are provided in Appendix B. On photographs with noise of σ_{255}=25 and 50, 60–62% of votes prefer ours over Differential Diffusion handed the true noise map, 64–65% over DiffPIR and 71–76% over RePaint. On clean photographs, 68% prefer ours over RePaint, but only 55% over Differential Diffusion and DiffPIR.

Conclusion

We proposed blind spatially-varying denoising: a pixel diffusion transformer without a timestep input, trained on spatially-varying noise. Since the model predicts the clean image, subtracting its prediction from the input gives the noise level of every patch, so no noise map is needed. Any pixel-space model that predicts the clean image can measure these levels, but only one trained on spatially-varying noise also denoises patch by patch. We showed that this ability can be added to pretrained models, including a text-to-image model, by fine-tuning, with almost no loss in guided sample quality. The result is one model that removes strong spatially-varying noise, fills noise-covered regions, extends borders, and edits a photograph wherever noise is painted onto it, including photographs that are already noisy, with neither a mask nor a noise map.

AI use statement

We used generative AI tools to polish the writing of the manuscript, and as a coding assistant. All code was reviewed by the authors, and will be publicly released. We take responsibility for the final content of this work, including text produced with the help of AI.

Reproducibility statement

Code and models will be publicly released upon acceptance of the paper.

References

Abdelrahman Abdelhamed, Stephen Lin, and Michael S Brown. A high-quality denoising dataset for smartphone cameras. In *CVPR*, 2018.

Michael Samuel Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. In *ICLR*, 2023.

Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *CVPR*, 2022.

Michal Balcerak, Tamaz Amiranashvili, Antonio Terpin, Suprosanna Shit, Lea Bogensperger, Sebastian Kaltenbach, Petros Koumoutsakos, and Bjoern Menze. Energy matching: Unifying flow matching and energy-based models for generative modeling. *arXiv preprint arXiv:2504.10612*, 2025.

Connelly Barnes, Eli Shechtman, Adam Finkelstein, and Dan B. Goldman. PatchMatch: A randomized correspondence algorithm for structural image editing. *ACM Transactions on Graphics (SIGGRAPH)*, 28(3), 2009.

Marcelo Bertalmio, Guillermo Sapiro, Vicent Caselles, and Coloma Ballester. Image inpainting. In *SIGGRAPH*, 2000.

Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. In *NeurIPS*, 2024.

Guangyong Chen, Fengyuan Zhu, and Pheng Ann Heng. An efficient statistical method for image noise level estimation. In *ICCV*, 2015.

Antonio Criminisi, Patrick Pérez, and Kentaro Toyama. Region filling and object removal by exemplar-based image inpainting. *IEEE Transactions on Image Processing*, 13(9): 1200–1212, 2004.

Giannis Daras, Kulin Shah, Yuval Dagan, Aravind Gollakota, Alexandros G. Dimakis, and Adam Klivans. Ambient diffusion: Learning clean distributions from corrupted data. In *NeurIPS*, 2023.

Giannis Daras, Adrian Rodriguez-Munoz, Adam Klivans, Antonio Torralba, and Constantinos Daskalakis. Ambient diffusion omni: Training good models with bad data. *arXiv preprint arXiv:2506.10038*, 2025.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*, 2009.

Yilun Du and Igor Mordatch. Implicit generation and modeling with energy based models. In *NeurIPS*, 2019.

Alexei A. Efros and William T. Freeman. Image quilting for texture synthesis and transfer. In *SIGGRAPH*, 2001.

Alexei A. Efros and Thomas K. Leung. Texture synthesis by non-parametric sampling. In *ICCV*, 1999.

Anne Harrington, A Koepke, Shyamgopal Karthik, Trevor Darrell, and Alexei A Efros. It's never too late: Noise optimization for collapse recovery in trained diffusion models. In *CVPR*, 2026.

James Hays and Alexei A. Efros. Scene completion using millions of photographs. *ACM Transactions on Graphics (SIGGRAPH)*, 26(3), 2007.

Alec Helbling, Sebastian Gutierrez Hernandez, Benjamin Hoover, Duen Horng Chau, and Parikshit Ram. What time is it? how data geometry makes time conditioning optional for flow matching. *arXiv preprint arXiv:2605.08344*, 2026.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020.

Zahra Kadkhodaie, Aram-Alexandre Pooladian, Sinho Chewi, and Eero P. Simoncelli. Blind denoising diffusion models and the blessings of dimensionality. *arXiv preprint arXiv:2602.09639*, 2026.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. In *NeurIPS*, 2022.

Yann LeCun, Sumit Chopra, Raia Hadsell, Marc'Aurelio Ranzato, and Fujie Huang. A tutorial on energy-based learning. *Predicting Structured Data*, 2006.

Eran Levin and Ohad Fried. Differential diffusion: Giving each pixel its strength. *arXiv preprint arXiv:2306.00950*, 2023.

Tianhong Li and Kaiming He. Back to basics: Let denoising generative models denoise. *arXiv preprint arXiv:2511.13720*, 2025.

Hansen Jin Lillemark, Alex Rojas, Zachary Novack, Runqian Wang, Yilun Du, Yian Ma, Taylor Berg-Kirkpatrick, and Rose Yu. Equilibrium forcing: Adaptive video generation without noise conditioning. *arXiv preprint arXiv:2608.14706*, 2026.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *ICLR*, 2023.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *ICLR*, 2023.

Xinhao Liu, Masayuki Tanaka, and Masatoshi Okutomi. Single-image noise level estimation for blind denoising. *IEEE Transactions on Image Processing*, 22(12):5226–5237, 2013.

Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. RePaint: Inpainting using denoising diffusion probabilistic models. In *CVPR*, 2022.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *ICLR*, 2022.

Samuele Papa, Valentin De Bortoli, Guillaume Couairon, Daniel Sýkora, Romuald Elie, and Klaus Greff. AsyncPatch diffusion: Spatially-flexible image generation. *arXiv preprint arXiv:2606.07079*, 2026.

Deepak Pathak, Philipp Krähenbühl, Jeff Donahue, Trevor Darrell, and Alexei A. Efros. Context encoders: Feature learning by inpainting. In *CVPR*, 2016.

Johannes Schusterbauer, Ming Gui, Yusong Li, Pingchuan Ma, Felix Krause, and Björn Ommer. Denoising, fast and slow: Difficulty-aware adaptive sampling for image generation. In *CVPR*, 2026.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *ICML*, 2015.

Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2021.

Qiao Sun, Zhicheng Jiang, Hanhong Zhao, and Kaiming He. Is noise conditioning necessary for denoising generative models? *arXiv preprint arXiv:2502.13129*, 2025.

Runqian Wang and Yilun Du. Equilibrium matching: Generative modeling with implicit energy-based models. *arXiv preprint arXiv:2510.02300*, 2025.

Xianbang Wang, Hanhong Zhao, Yiyang Lu, Kangyang Zhou, Linrui Ma, and Kaiming He. Minit2i: A minimalist baseline for text-to-image generation, 2026. URL `https://peppaking8.github.io/#/post/minit2i`.

Yi Wang, Ying-Cong Chen, Xin Tao, and Jiaya Jia. VCNet: A robust approach to blind image inpainting. In *ECCV*, 2020.

Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *CVPR*, 2022.

Bingliang Zhang, Wenda Chu, Julius Berner, Chenlin Meng, Anima Anandkumar, and Yang Song. Improving diffusion inverse problem solving with decoupled noise annealing. In *CVPR*, 2025.

Kai Zhang, Wangmeng Zuo, Yunjin Chen, Deyu Meng, and Lei Zhang. Beyond a Gaussian denoiser: Residual learning of deep CNN for image denoising. *IEEE Transactions on Image Processing*, 2017.

Kai Zhang, Wangmeng Zuo, and Lei Zhang. FFDNet: Toward a fast and flexible solution for CNN-based image denoising. *IEEE Transactions on Image Processing*, 2018.

Kai Zhang, Yawei Li, Wangmeng Zuo, Lei Zhang, Luc Van Gool, and Radu Timofte. Plug-and-play image restoration with deep denoiser prior. *IEEE TPAMI*, 2022.

Yuanzhi Zhu, Kai Zhang, Jingyun Liang, Jiezhang Cao, Bihan Wen, Radu Timofte, and Luc Van Gool. Denoising diffusion models for plug-and-play image restoration. In *CVPR Workshops*, 2023.

Appendix: Blind Spatially-Varying Denoising in Image Diffusion Models

A Limitations

The residual internal clock needs white noise stronger than the model's own reconstruction error. Below $\sigma_{255} \approx 40$ the reading is dominated by that error and denoisers trained on that band are stronger (Appendix P); on near-uniform real noise, a single global reading is the better use of the clock (Appendix N). Holes that carry no noise are not found. Restoration and editing work at the resolution of the patch grid, and the method requires a pixel-space model that predicts the clean image. Training on spatially-varying noise costs some unguided sample quality (Appendix E). However, this might amortize with longer training.

B User study

Stimuli. We drew 180 ImageNet validation images, 60 per setting, from the 1000-image sets of Table A12: an edit of a clean photograph and edits of photographs with white noise of σ_{255}=25 and 50. Each image keeps the stroke, noise, class label and outputs of that benchmark exactly; rerunning the benchmark to save them reproduced every entry of Table A12. Ours on clean photographs copies back the patches it measures as clean, as in Table A12. Each baseline gets what it needs: Differential Diffusion the true noise map (our modification, Section 5), RePaint the stroke as a binary mask (5 resampling passes), both on the conditioned JiT-B/16 with the class label and 20 steps, and DiffPIR the true noise map on its own unconditional ADM prior (100 steps). Images were drawn in two rounds (seeds 2027 and 2029) with no overlap between settings or rounds.

Task. Each trial shows the photograph with the painted noise as the reference and two edits of it, ours and one baseline, with the instruction "Which is the better edit of the reference?". The platform shuffles left and right per trial, and we additionally placed ours first in about half the uploaded pairs. Images were uploaded under opaque hashed filenames, and no method name was visible to raters. Every pair received 25 votes.

Raters. The study ran on a commercial crowdsourcing platform with its general pool of raters, without screening or catch trials: 13,506 votes from raters in 107 countries. The platform scores each rater's reliability; 92% of votes came from raters scoring at least 0.5, and keeping only those votes changes no entry of Table A1 by more than 0.01.

Analysis. For each image we take the share of its votes that prefer ours. A cell of Table A1 reports the mean over its images with a 95% bootstrap interval over images (10,000 resamples), and a cell is marked significant when a two-sided binomial test of the per-image majority against 0.5 rejects at 0.0056 (0.05 with Bonferroni over the nine cells). The platform returns no rater identity, so agreement is measured per pair: Fleiss' κ is 0.07–0.14. Single votes are noisy, and the preferences below come from averaging many of them.

Results. On noisy photographs raters prefer ours over every baseline: 60–62% of votes over Differential Diffusion handed the true noise map, 64–65% over DiffPIR and 71–76% over RePaint (Table A1 and Figure A1). On clean photographs they prefer ours over RePaint (68%) but are split between ours and Differential Diffusion (55%) or DiffPIR (55%), neither significant. The two rounds agree to within 0.08 in every cell.

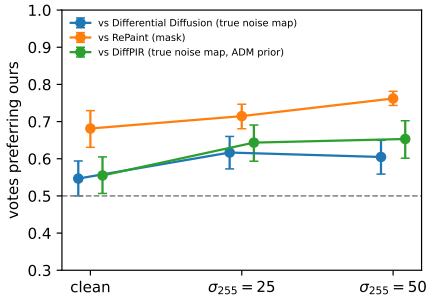


Figure A1: Share of votes preferring ours per setting and baseline, 60 images per point. Dashed line is no preference.

The figure rows. We also rated the four rows of Figure 7 against each baseline (Table A2). Against Differential Diffusion, 0.88–0.94 of votes prefer ours in every row. Against RePaint the majority prefers ours in every row (0.62–0.86), though

Table A1: User study. Share of votes preferring ours when raters see the photo with the painted noise and are asked "Which is the better edit of the reference?". 60 images per setting from the sets of Table A12, same strokes, noise and outputs; 25 votes per pair; 95% bootstrap intervals over images. Bold: per-image majority differs from chance (binomial test, Bonferroni over the nine cells). ‡handed the true noise map. §handed the stroke as a mask. †handed the true noise map, own ADM prior.

setting	vs DD‡	vs RePaint§	vs DiffPIR†
edit, clean photo	0.55 [0.50, 0.59]	**0.68** [0.63, 0.73]	0.55 [0.51, 0.60]
edit, noise σ_{255}=25	**0.62** [0.57, 0.66]	**0.71** [0.68, 0.75]	**0.64** [0.59, 0.69]
edit, noise σ_{255}=50	**0.60** [0.56, 0.65]	**0.76** [0.74, 0.78]	**0.65** [0.60, 0.70]
all settings	0.59 [0.56, 0.62]	0.72 [0.70, 0.74]	0.62 [0.59, 0.65]

for the large graded stroke the interval includes
0.5. Against DiffPIR, raters prefer ours in the two strong edits of clean photographs (0.94 and 0.90) but not in the other two (0.50 and 0.42), where DiffPIR restores the photograph instead of editing it and raters do not count that against it.

Table A2: The four rows of Figure 7, rated separately and not pooled with Table A1, since they were chosen as failures of the baselines. Share of 50 votes preferring ours, 95% Clopper–Pearson interval over votes.

row	setting	vs DD‡	vs RePaint§	vs DiffPIR†
1	strong edit, clean	0.88 [0.76, 0.95]	0.70 [0.55, 0.82]	0.94 [0.83, 0.99]
2	strong edit, clean	0.88 [0.76, 0.95]	0.86 [0.73, 0.94]	0.90 [0.78, 0.97]
3	smooth varying noise	0.88 [0.76, 0.95]	0.74 [0.60, 0.85]	0.50 [0.36, 0.64]
4	large graded stroke, clean	0.94 [0.83, 0.99]	0.62 [0.47, 0.75]	0.42 [0.28, 0.57]

C What the clock measures: derivation and controls

The clock measures the size of the residual, which holds two parts: the noise and the model's own error. On a *flat* σ the per-patch reading follows $\hat{\sigma}_p^2 = \sigma^2 + c\,\mathrm{mse}_p$ (Equation (5)). We measure $c = +0.998$ at $\sigma_{255} = 5$: hard patches read noisier by exactly their error. The leak is small at $\sigma_{255} \approx 40$ and gone above $\sigma_{255} \approx 75$, which we call the *crossover*. Above it the map is a pure noise map. Every training field leaks the same way, so the leak belongs to the estimator, including Kadkhodaie et al.'s.

A residual-magnitude clock is not a noise meter everywhere. This section characterizes what it reads, exactly, and where the characterization matters.

Setup. We corrupt held-out images with a *flat* σ map. The true t field is then constant by construction, so *every* spatial structure in the measured map is leakage. We measure difficulty directly as the per-patch reconstruction MSE rather than proxying it by image content, and compare against a null obtained by spatially permuting each image's own values, which is the finite-grid floor ($\xi_{\mathrm{null}} = 0.63$ patches on a 16×16 grid).

C.1 A closed form for the leak

Let $e = \hat{x} - x$ and $M_p = \|e_p\|^2$. Expanding the residual,

$$\|y_p - \hat{x}_p\|^2 = d\sigma^2 - 2\sigma\langle \varepsilon_p, e_p\rangle + \|e_p\|^2. \tag{4}$$

The cross term decides the sign. At the MMSE solution the orthogonality principle gives $\mathbb{E}\langle y - \hat{x}, e\rangle = 0$, hence $\mathbb{E}\langle \sigma\varepsilon, e\rangle = \mathbb{E}\|e\|^2$, and the cross term *overcancels*, leaving $\hat{\sigma}_p^2 = \sigma^2 - \mathrm{mse}_p$. If instead the error is a *content bias*—a function of the image, independent of the realized noise—then $\langle \varepsilon, e\rangle \approx 0$, the cross term vanishes, and the residual energy *adds*.

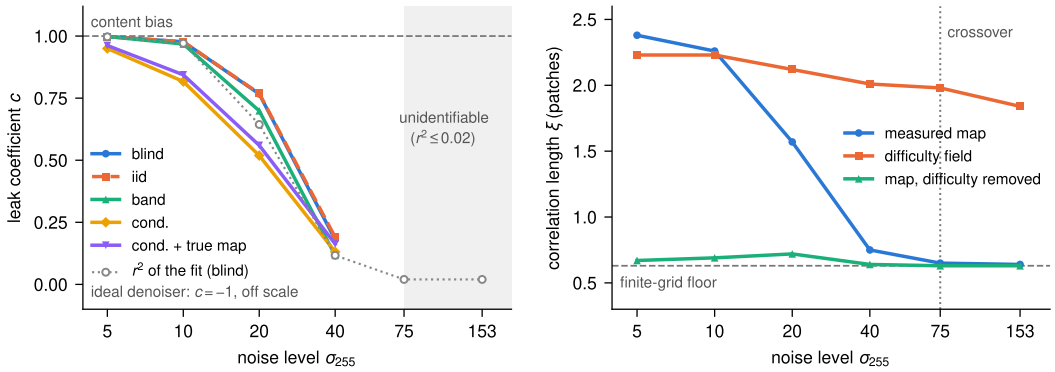


Figure A2: What the clock measures. Left: the leak coefficient c of Equation (5) per training variant: every blind variant sits at $+1$, pure content bias, at low noise and decays with σ. Conditioning moves the model measurably toward the ideal denoiser, whose $c = -1$ is off the scale. The dotted line is the r^2 of the fit: above $\sigma_{255} \approx 75$ the map is white and c cannot be fitted. Right: correlation length of the measured map on a flat σ field, of the model's difficulty field, and of the map after difficulty is removed. Below the crossover the map inherits difficulty's structure. Above it, it sits on the finite-grid floor and is a noise map.

Table A3: Fitted c of Equation (5) (r^2 of the fit for blind: 0.997/0.971/0.644/0.117 top to bottom). At low σ the error is pure content bias in every blind variant. Conditioning ("cond.") moves the estimator measurably toward MMSE (+0.950 vs +0.998), and being handed the true map ("+map") changes nothing further. Above $\sigma_{255} = 75$ the field is white and c is unidentifiable ($r^2 \leq 0.02$), so entries are omitted rather than reported as a sign flip.

σ_{255}	blind	IID	BAND	cond.	+map
5	**+0.998**	+0.998	+0.998	+0.950	+0.962
10	+0.977	+0.977	+0.968	+0.817	+0.844
20	+0.767	+0.769	+0.699	+0.519	+0.560
40	+0.180	+0.190	+0.117	+0.131	+0.163
$\geq$75	—	—	—	—	—

The two hypotheses differ only in a sign, which an unsigned r^2 cannot distinguish. Fitting the coefficient in

$$\hat{\sigma}_p^2 \;=\; \sigma^2 + c\,\mathrm{mse}_p \tag{5}$$

per image settles it: $c = -1$ is exact MMSE, $c = +1$ is pure bias.

Table A3 gives $c = +0.998$ with $r^2 = 0.997$ at $\sigma_{255} = 5$, from a hypothesis with *no free parameters*. The clock reads *noisier* on hard patches, by exactly their reconstruction error. The model is nowhere near MMSE in this band—consistent with the finding that 74% of patches at $\sigma_{255} = 5$ have $\mathrm{mse}_p > \sigma^2$, i.e. lose outright to the trivial estimator $\hat{x} = y$—and c decays toward the MMSE limit as σ grows and the model returns to its training distribution.

C.2 Where the crossover is

Above $\sigma_{255} \approx 75$ the clock map is white ($\xi = 0.64$ against a null of 0.63). The default corruption families of Section 5 run to $\sigma_{255} = 153$ and sit mostly in this regime, so those results are largely unaffected. Below the crossover the clock's correlation length *is* difficulty's—2.38 vs 2.23 patches, a ratio of 1.07—which puts a number on the resolution limit: at $\sigma_{255} = 5$ the clock manufactures $\sim$38 px of correlation on perfectly uniform input, so nothing finer is resolvable there. Removing difficulty per image nonparametrically collapses ξ to the null; a *linear* control leaves 1.21.

Table A4: Correlation length ξ in patches ($\times 16$ for px), $1/e$ crossing of the unbiased radial ACF, against a null of 0.63.

σ_{255}	5	10	20	40	75	153
ξ (blind)	**2.38**	2.26	1.57	0.75	0.65	0.64
ξ of difficulty	2.23	2.23	2.12	2.01	1.98	1.84
ξ after removal	**0.67**	0.69	0.72	0.64	0.63	0.63
$r(\hat{t}-t, \text{mse})$	-0.98	-0.98	-0.77	-0.20	$+0.03$	$+0.04$

Table A5: The training arms. FID: tuned, 200 epochs, 50k images. "sep." and t-MAE come from the two-region probe (left half t=0.3, right half t=0.8, clock read once, 200 images): separation between the halves (truth 0.500) and mean map error. The blind arm's lower separation reflects noise left in $\hat{x}$, not a misreading (Appendix G). "Fine field": the BAND mixture at half-patch cells (Figure 2).

arm	training field (share of draws)	FID	sep.	t-MAE
BLIND	uniform	5.22	0.404	0.053
IID	uniform $\frac{1}{2}$, iid patches $\frac{1}{2}$	5.39	0.448	0.028
BAND	uniform $\frac{1}{4}$, band $\frac{1}{2}$, split $\frac{1}{4}$	6.71	0.495	0.010
BAND+FINE	BAND $\frac{1}{2}$, fine field $\frac{1}{2}$	7.11	0.495	0.010
REGION (ours)	regions $\frac{1}{2}$, fine field $\frac{1}{4}$, BAND $\frac{1}{4}$	6.64	0.495	0.010

D Ablations: which training field buys locality

Every arm below is the same architecture, data and budget; the only difference is the t-field the model was trained under. Figure 2 shows the six field types the arms are built from, and Table A5 says how each arm mixes them. All fields are drawn per training image on the 16×16 patch grid, in image coordinates unless the name says otherwise, and every arm keeps a share of spatially uniform draws so generation is never abandoned.

Any spatial field buys locality; the split makes it sharp. Against BLIND, IID costs +0.17 tuned FID and BAND +1.49, while the probe orders them the other way: IID separates the two halves to 0.448 at a map error of 0.028, whereas BAND reaches 0.495 at 0.010. IID is the cheapest field and the least local one. What BAND adds over it is a global level that the per-patch values scatter around, and the two-region split; our reading is that the split is what teaches a sharp boundary, which is the case a corrupted photograph presents. Because BAND draws per patch, its field fragments further as resolution grows at a fixed patch size, so we specify the fine field in image coordinates.

Fields that cut through patches. BAND never places a noise boundary inside a patch, and real boundaries do (Section 3). Drawing fields at half-patch cells fixes that. On its own, the fine field improves smooth restoration by 0.7 dB but breaks hole *sampling* (LPIPS 0.093 $\rightarrow$ 0.165, speckle left in the fill), and the regression persists at 600 epochs, so it is the recipe, not the budget. Mixing the fine field with BAND's per-patch draw, which is BAND+FINE, keeps the gain and repairs sampling (0.097) at no cost in FID. Hole-sampling robustness lives in the per-patch band.

Fields shaped like content. REGION replaces half of BAND+FINE's draws with content-shaped regions. The patch grid is walked in Hilbert order, cut at the $R-1$ largest jumps in the clean image's gradient energy ($R \in \{4, \ldots, 64\}$), and each region receives one level, so training visits fields whose boundaries follow image structure rather than the grid. At matched 600 epochs it is best or tied on every synthetic restoration family (+0.4 dB over BAND+FINE on smooth fields), and matches BAND+FINE in generation within seed noise (tuned FID 6.64 vs 7.11 at 200 epochs, 4.77 vs 5.10 at 600), with the guidance optimum moving down as training proceeds (5.0 $\rightarrow$ 3.5). It is the locality-trained arm closest to BLIND (3.89 at 600 epochs).

Table A6: Generation cost of blindness and of field training, trained from scratch. 50k images, Heun 50 steps, CFG interval $[0.1, 1.0]$, EMA weights. "Tuned" is each model at its own best guidance scale, given in brackets. FFHQ and LSUN carry a single label, so guidance is inert and only the unguided number is reported; they are scored against Inception statistics of their own training sets (70k FFHQ, 200k LSUN).

data	FID	baseline (cond.)	blind	ours (REGION)
ImageNet, 600 ep.	unguided	26.00	31.03	37.50
	tuned (cfg)	3.60 (3.0)	3.89 (3.5)	4.77 (3.5)
FFHQ, 1200 ep.	unguided	23.26	28.55	24.11
LSUN, 400 ep.	unguided	23.54	19.63	22.34

Seeds. A second BAND seed puts error bars under the tables: restoration agrees to ≤ 0.21 dB across all four families (inpainting to 0.01), while tuned FID spreads 0.46 with the guidance optimum moving half a step. Restoration claims in this paper are seed-stable.

E Generation cost in detail

Compared at one shared guidance scale, dropping the timestep looks expensive. Compared at each model's own best scale, it is not (Table A6): removing the embedding mostly moves the useful guidance range. At 600 epochs the blind model trails the conditioned baseline by 0.29 tuned FID, and the gap has shrunk with training (0.70 at 200 epochs, 0.39 at 400). The baseline beats the published JiT-B/16 number (3.66), so this is not an internal-only comparison. Field training adds its own cost on top: the REGION model trails blind by 0.88 and the baseline by 1.16 tuned, and unguided by 6.5 and 11.5.

Other datasets. The same three models trained from scratch on FFHQ and LSUN bedrooms give the same picture. Both datasets carry a single label, so guidance has nothing to guide toward: FID is flat across cfg 1–3 to within 0.35 at 10k images, and we report the unguided number. On FFHQ the blind model trails the conditioned one by 5.3 and the locality-trained model recovers all but 0.8 of that. On LSUN the order flips: both blind models beat the conditioned baseline, the plain blind one by 3.9, from one seed each, so the sign is probably real and the size is not to be trusted.

F Converting a conditioned model: details

How the timestep is removed. We tried three ways on a 5k-step pilot with the REGION field: *delete* the timestep embedding outright, *anneal* it to its average over the first 20% of the run, or replace it by the *mean* embedding from the start. Mean and anneal give the same restoration from 1k steps on; delete trails by about 1.5 dB at 5k steps, because every adaLN layer loses the vector it was trained with. We use the mean. The embedder is frozen in all three, which keeps the mean matched to it.

Which weights to score. Intermediate checkpoints are scored on the 0.999 EMA, because after 10k steps the 0.9999 average still holds 37% of the source weights and is not a working model; Table 1 reports the 0.9999 weights at 50k steps, as for every other model in the paper.

How fast. On the 500-image subset the pilot's probe separation reaches 0.477 at 3k steps and 0.490 at 5k (reference 0.495), and hole PSNR is flat from 5k steps on. FID at cfg 3.5 is 4.50 at 10k steps (1% of the from-scratch compute); tuned, it is 3.77 at 25k.

Released checkpoints. Table 1 applies the recipe unchanged to four released JiT checkpoints: B/16, L/16 and H/16 at 256px and B/32 at 512px; for B/32 the only changes are the 512px data and its noise scale of 2. Each source is run as a conditioned model through the self-clocked fixed point, the "cond., self-clocked" row of Table 2. Every conversion gains 9.7–10.8 dB on holes over its source, and 3.2–5.6 dB on smooth and 4.9–8.0 dB on piecewise noise at 256px (1.3 and 2.1 dB for B/32, whose source is already better on these fields).

Restoration improves with size: the converted H/16 is 0.45–0.7 dB above the converted B/16 on every family, and above our B/16 trained from scratch (Table 2) on every family as well. Generation does not scale as cleanly. Against its own source the conversion costs at most 0.1 tuned FID at B/16, B/32 and L/16, but 1.0 at H/16. The H/16 conversion is still improving at 50k steps (FID at cfg 3.0 on the 0.999 EMA: 4.52, 3.76, 3.25 at 10k, 25k, 50k), so part of that gap is budget.

G Why both models read the two-region input, and only one acts on it

Section 4 states that the blind model reads a two-level input correctly through the restoration fixed point but appears not to in the one-pass probe of Table A5. Figure A3 traces both readings on the same 64 images (left half t=0.3, σ=2.33; right half t=0.8, σ=0.25; null class; 600-epoch checkpoints). We log, per half and per iteration $k = 0, \dots, 4$ of the fixed point, the read $\hat{t}$, the reconstruction error of $\hat{x}$, and the correlation ρ between the error $\hat{x} - x$ and the realized noise ε; and we run the one-pass protocol, $z = t \odot x + (1 - t) \odot \varepsilon$ fed once, reading its $\hat{x}$ both with the interpolant clock of Equation (2) from $z - \hat{x}$ and with the VE clock $\hat{\sigma} = \|y - \hat{x}\|/\sqrt{d}$ from $y - \hat{x}$.

Three measurements explain the difference. First, the one-pass $\hat{x}$ read from $y - \hat{x}$ gives 0.308 and 0.799 for the blind model, the same as the fixed point; read from $z - \hat{x}$ it gives 0.400 and 0.801. The reading changes with the residual, not with the iteration. Second, the fixed point does not repair anything: at k=0, before any scaling, the blind model already reads 0.312 and 0.743, and at k=1 it reads 0.308 and 0.792. The hypothesis that scaling each half by its own t presents a uniformly noisy input to a global denoiser is false as stated: after scaling, the noise amplitude in z is 0.73 on one half and 0.19 on the other, and the signal-to-noise ratio of each half is what it was in the one-pass input. Third, the cause is a noise-correlated error. On the noisy half the blind model's error is correlated with ε at ρ=0.30 at every iteration; the locality-trained model's at 0.05. Write $\hat{x} = x + \delta$ with δ partly aligned with ε. In $z - \hat{x} = (1 - t)(\varepsilon - \hat{x}) + t(x - \hat{x})$ the aligned part of δ cancels a fraction of $(1 - t)\varepsilon$ outright, and the normalizer $d + \|\hat{x}\|^2$ of Equation (2) has nothing large to hide it behind: with ρ=0.30 and an error of rms 0.5 the reading moves from 0.30 to 0.39, which is the 0.400 of the table. In $y - \hat{x} = \sigma\varepsilon - \delta$ the same cancellation is measured against σ^2, and at σ=2.33 it moves $\hat{\sigma}^2$ by 13% and $\hat{t}$ by 0.01. This is the leak of Appendix C seen from the other side: the interpolant clock is the VE clock with the noise term shrunk by $(1 - t)^2/\sigma^2$, and below that shrinkage the model's own error is visible.

Acting is where the models part. The blind model's error on the noisy half stays at 0.257 from k=0 to k=4 while the locality-trained model's falls from 0.301 to 0.024 (Figure A3, right). On the clean half both reach 0.006. A model trained on uniform noise reads a two-level map through its residual and denoises as if the map were one number. The probe column of Table A5 therefore ranks the arms by how much noise their one-level denoising leaves in $\hat{x}$, which is the quantity that decides restoration, and its ordering matches the restoration ordering of Table A13.

H Restoration on FFHQ and LSUN

Table A7 repeats the three-way comparison of Section 5 with models trained from scratch on FFHQ and LSUN bedrooms, one architecture and one training budget per dataset, and the same protocol in every cell. The shape does not change with the data. On uniform noise the three rows tie to within 0.16 dB on every dataset, so no row is a better denoiser than the others. Where the field varies (smooth and piecewise), the blind global clock loses 5.7–9.3 dB against its own uniform number and the oracle-mean model loses 3.6–7.1 dB, while the locality-trained model holds its uniform level to within 0.4 dB on FFHQ and LSUN and 0.2 dB on ImageNet. The absolute levels differ, FFHQ is about 2 dB easier than ImageNet at every row because faces are smoother than object photographs, but the gaps do not: the local map leads the oracle mean by 3.4–6.9 dB on the noise families and by 7.9–10.1 dB on holes and strokes on all three datasets.

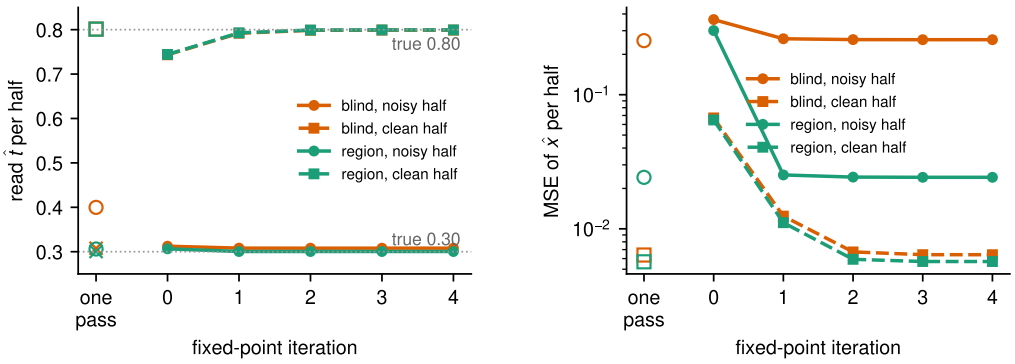


Figure A3: The two-region probe across fixed-point iterations, blind and locality-trained (REGION) models, 64 images. *Left:* read $\hat{t}$ per half. Filled markers: the fixed point, read from $y - \hat{x}$. Open markers at "one pass": the same one-pass $\hat{x}$ read from $z - \hat{x}$ with Equation (2); crosses: read from $y - \hat{x}$. Only the blind model's noisy half moves between the two, from 0.40 to 0.31. *Right:* reconstruction error per half. The blind model's noisy half never improves; the locality-trained model's drops 12× at the first scaled iteration.

Table A7: PSNR (dB) across datasets, default band, $n = 200$, one protocol for every cell: fixed-point measurement and one $\hat{x}$, no sampler. Holes are patch-aligned and filled with noise; the stroke is a random walk of width 2 tokens covering 25% of the image; no mask is supplied. FFHQ and LSUN models are trained from scratch on each dataset (FFHQ 1200 epochs, LSUN 400; ImageNet 600). The three rows share one architecture and one training budget per dataset. The ImageNet rows are a separate 200-image draw from Table 2, which is why its hole numbers differ slightly.

data	method	uniform	smooth	piecewise	poisson	holes	stroke
ImageNet	blind, global clock	26.95	18.31	18.36	26.06	11.82	10.62
	REGION (ours)	26.93	**26.71**	**26.88**	27.70	**25.22**	**23.71**
	cond., oracle mean $\bar{t}$	**27.00**	22.57	20.13	**27.77**	16.11	12.63
FFHQ	blind, global clock	28.60	22.88	19.65	29.39	12.36	11.64
	REGION (ours)	**28.76**	**28.53**	**28.85**	**30.00**	**26.79**	**24.97**
	cond., oracle mean $\bar{t}$	28.68	25.09	22.32	29.87	18.52	14.89
LSUN	blind, global clock	28.00	19.18	18.73	27.47	12.20	11.16
	REGION (ours)	27.99	**27.65**	**27.98**	**28.95**	**25.25**	**24.05**
	cond., oracle mean $\bar{t}$	**28.10**	23.68	21.05	28.91	17.36	13.94

The hole and stroke columns are the sharpest test, because a mean level says nothing about where the damage is. There the oracle-mean model lands 2–3 dB above the corrupted input and the locality-trained model 11–14 dB above it, on faces, bedrooms and objects alike. The single-label datasets also settle a side question: the restoration in Section 5 runs without class information, and here there is none to run with, so the reconstruction rests on the measured map alone.

I A SECOND PIXEL-SPACE ARCHITECTURE: MINIT2I

Section 3 argues that any pixel-space x-prediction transformer should read its own noise level. MiniT2I (Wang et al., 2026) is a public text-to-image model of that kind: an MM-DiT denoiser on RGB pixels at 512px, trained with flow matching, x-prediction and a logit-normal timestep, conditioned on frozen T5 text tokens. Two details make it a clean second test. Its released weights are timestep-blind by construction: the timestep vector is computed and consumed by no block, so no embedding has to be removed. And it trains with noise of standard deviation 2, $x_t = t\,x + (1 - t)\,2\varepsilon$, so the on-interpolant scaling for

Table A8: PSNR (dB) on MiniT2I-B/16 at 512px, n=200 held-out ImageNet training crops, default band, one $\hat{x}$ per cell, no sampler, text nulled at evaluation; same families as Table A7. The released weights are timestep-blind and uniform-trained; the finetune (COCO captions, 10k steps) adds the REGION field and nothing else.

MiniT2I-B/16, told nothing	uniform	smooth	piecewise	poisson	holes	stroke
released, global clock	28.20	21.20	19.77	28.25	12.44	10.93
released, per-patch fixed point	28.19	21.76	20.40	28.40	17.47	13.86
REGION finetune, global clock	28.28	25.93	24.91	29.05	13.01	12.79
REGION finetune, per-patch	28.28	27.97	28.49	29.45	26.62	25.71
map t-MAE, released	0.005	0.012	0.014	0.012	0.016	0.021
map t-MAE, finetuned	0.005	0.007	0.009	0.012	0.013	0.011

a VE input $y = x + \sigma\varepsilon$ is $z = t\,y$ with $t = 2/(2+\sigma)$; we confirmed the factor from the model itself (reconstruction error at t=0.5 is 0.020 with the factor and 0.043 without) and the residual internal clock is otherwise unchanged. Text is nulled throughout (every text token replaced by the model's mask token), as class is nulled in our restoration. Evaluation runs on 512px centre crops of full-resolution ImageNet training photographs held out from the finetune, since the validation pack is 256px.

Reading. On the released B/16 weights, uniform noise across $\sigma_{255} \in \{13, 25, 50, 75, 100, 153\}$ is read to a global t-MAE of 0.0004–0.003 and a per-patch t-MAE of 0.002–0.005 (σ_{255} error 0.7–3.1), the precision of our blind JiT. On the two-region probe the released model reads 0.308 and 0.799 through the fixed point and 0.379 and 0.798 in one pass with the interpolant clock, the same pooled one-pass reading and correct VE reading as blind JiT (Appendix G). Reading transfers unchanged.

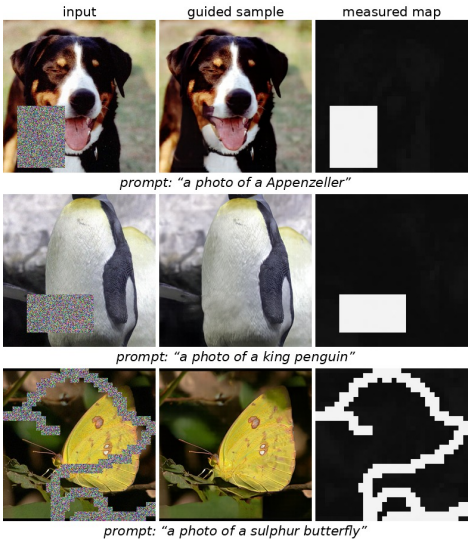
Acting. As for JiT, the released model reads a two-level input and cannot act on it: on the two-region probe it restores the noisy half at 9.4 dB. We therefore finetune it with the REGION training field, borrowed unchanged from our code (half content-region fields, a quarter fine band/split, a quarter band/split), with MiniT2I's own loss and per-pixel t entering the loss only, and with its text conditioning kept on: 10k steps at an effective batch of 128 on the 118k COCO 2017 training photographs at 512px, one of the five captions per image drawn at random and dropped to the null prompt 10% of the time, learning rate 10^{-4}, EMA 0.999. Table A8 repeats the protocol of Table A7 on 200 held-out ImageNet crops. A second finetune on 60k ImageNet training crops with the text nulled gives the same numbers to within 0.07 dB on every family, so neither the data nor the text is what the field adds.

The picture is the one of Table A7. All four rows tie on uniform noise. The released model reads every field to within 0.02 and acts on none of them: the per-patch fixed point gains it 0.6 dB on the varying noise families and 3–5 dB on holes and strokes over its global clock, and it still trails by 6–12 dB. The locality-trained model holds its uniform level on smooth and piecewise noise (28.0, 28.5 against 28.3) and fills holes and strokes at 26.6 and 25.7 dB, from a map it measured itself, with the same global-clock control collapsing to 13 dB. The gains are the JiT gains: +6–8 dB on the varying fields and +9–12 dB on holes and strokes over the same model read through the same fixed point without the field. Brief finetuning on a public checkpoint suffices, because reading was already there and only acting had to be trained.

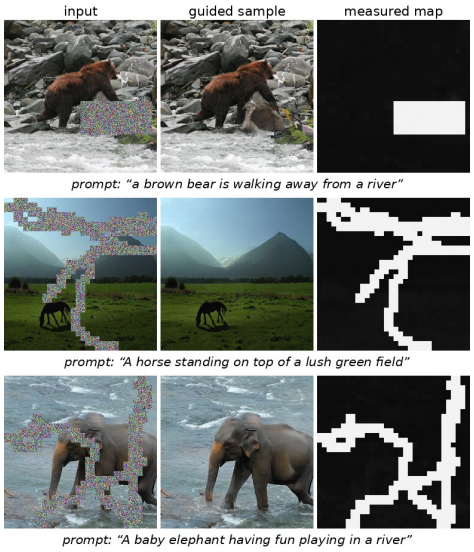
What the finetune costs in generation. Table A9 reports FID of the released and the finetuned model as text-to-image generators, 5000 COCO val2017 captions against the real val images at 512px, and 10k ImageNet class prompts ("a photo of a *class*") against the ImageNet-512 reference statistics, 50 Euler steps, without guidance. The finetune is 4.5 FID worse on COCO and 2.3 worse on ImageNet prompts, which is the price of 10k steps on a spatially-varying field with no change to the generative objective. A longer finetune on the same COCO data, 50k steps at batch 512, removes that cost: FID 30.2 against the released model's 32.7.

Table A9: Generation quality of MiniT2I before and after the REGION finetune, 512px, 50 Euler steps, unguided. COCO: 5000 val2017 captions, scored against the 5000 real val images. ImageNet: 10 prompts per class, scored against the ImageNet-512 statistics used for JiT. Finetunes run 10k steps at batch 128 unless marked.

model	finetune data	COCO, FID-5k	ImageNet, FID-10k
released	–	32.7	29.2
REGION finetune	COCO, captions on	37.2	–
REGION finetune, 50k steps	COCO, captions on	30.2	–
REGION finetune	ImageNet, text nulled	–	31.5



(a) Held-out ImageNet training crops, the class name as the prompt.



(b) COCO val2017 photographs, the image's own caption as the prompt.

Figure A4: MiniT2I-B/16 with the REGION finetune, 512px, no mask supplied. Columns: the input, a guided sample from the measured map (guidance 2, 50 Heun steps), and the measured map (white is noisy). The prompt is written under each row.

Figure A4a shows guided fills. The one-shot $\hat{x}$ is the posterior mean and blurs inside a hole, as it does for JiT; sampling from the measured map with the class name as the text prompt ("a photo of a *class*", guidance 2, 50 Heun steps, the teaser's setting) fills holes and strokes with plausible content. On 64 held-out images, the sample costs 1.9 dB of PSNR against the mean on holes (24.8 vs 26.7) and 2.0 dB on strokes (23.4 vs 25.5), the price of committing to one fill instead of averaging them, as in Table 2.

The same works with the captions the model was finetuned on: Figure A4b fills holes and strokes in COCO val2017 photographs with each image's own caption as the prompt. On 64 val images the mean reaches 26.7 dB (LPIPS 0.120) on holes and 26.0 (0.163) on strokes, the caption-guided sample 24.9 (0.152) and 24.0 (0.171).

Figure A5 applies the object-inpainting protocol of Appendix J to this model: the main object of a COCO val photograph (RMBG mask on the patch grid) is replaced by noise, the model measures the map, and the caption-guided sampler (guidance 6, frozen at the measured map) fills the hole from nine seeds.

J The noise decides the fill

Because the map only sets the starting state, the noise in a hole is the fill's initial condition: the map says where, the class label or caption says what, and the noise says which.



Figure A5: Object inpainting from nine seeds, no mask supplied. Per block: the original, the input with the object replaced by noise, the map the model measured, and below nine fills. Left: the MiniT2I COCO finetune on a COCO image, prompted with the caption written under the block, guidance 6 frozen at the measured map, 50 Heun steps. Right: our JiT model (REGION) on an ImageNet validation image at 256^2, guidance 2, 50 steps.

Harrington et al. (2026) make the complementary point that the initial noise is a control input for generation. With nine seeds per hole, JiT fills each hole with a different object of the class in an unchanged scene (Figure A5, right). Over 158 images of seven classes (barn, castle, kite, bookcase, coffee mug, organ, flowerpot; objects covering 4–60% of the frame), the mean pairwise DINOv2 dissimilarity between the nine fills of a hole is 0.48 (0.33–0.58 per class), against 0.80 between a fill and real objects of its class, while the context outside the hole stays at 31.8 dB to the original. The MiniT2I finetune behaves the same way under caption guidance (0.334 at guidance 6 on 127 COCO val photographs; Figure A5, left).

Table A10: The measured map as input to map-fed restorers. PSNR (dB) and hole/stroke LPIPS of three consumers handed a classical per-patch detector's map (MAD, LoG; Appendix L), our measured map, or the true map. Ours takes no map. Default band, $n = 5000$ (strokes $n = 200$). The map-conditioned model was not run on strokes.

	spatially-varying noise, PSNR			hole		stroke	
method	smooth	piecew.	poisson	PSNR	LPIPS	PSNR	LPIPS
DRUNet, MAD map	21.55	21.79	26.17	18.27	0.311	14.69	0.522
DRUNet, LoG map	21.35	21.62	25.59	18.33	0.379	14.93	0.567
DRUNet, our map	21.59	21.97	26.25	19.36	0.319	15.70	0.518
DRUNet, true map	21.57	21.96	26.25	19.83	0.287	16.05	0.488
DiffPIR, MAD map	22.50	22.93	24.24	24.15	0.093	21.37	0.143
DiffPIR, true map	22.69	23.46	24.30	24.54	0.083	21.59	0.132
map-cond., MAD map	26.43	26.16	27.93	27.06	0.114	–	–
map-cond., LoG map	26.27	26.11	27.65	26.45	0.186	–	–
map-cond., our map	26.46	26.88	27.95	27.04	0.124	–	–
map-cond., true map	26.47	26.39	27.98	27.07	0.106	–	–
ours, one $\hat{x}$ (no map)	26.33	26.57	27.25	25.17	0.110	23.83	0.167
ours, sampled (no map)	25.32	25.70	25.93	25.06	0.088	23.21	0.118

K Measured and detected maps as input to map-fed restorers

A natural alternative is to estimate each patch's noise level directly from its pixels, without a network. We test that with two classical detectors on the 16 px patch grid. *MAD*: the Immerkaer high-pass residual $K = [[1, -2, 1], [-2, 4, -2], [1, -2, 1]]$ with the robust scale $\hat{\sigma} = 1.4826\,\mathrm{median}|K * y|/\|K\|_2$ over the 768 residual values of a patch. *LoG*: a Laplacian of Gaussian (σ_g=1 px) with the same robust scale. Both are exact for white noise on a flat patch and read any high-pass energy of the content as noise. In the default band, where white noise dominates texture inside a 16 px patch, MAD reads the map about as accurately as our clock; LoG is worse everywhere, most on holes, because the Gaussian pre-blur spreads the hole boundary. Table A10 feeds each map, and the true map, to the three consumers that take one, on the 5000-image protocol of Table 2 for the noise families and holes and on 200 images for strokes.

Fed our measured map, DRUNet and the map-conditioned model match or exceed their true-map results on every noise family (Table A10), so the clock also works as a plug-in noise estimator for existing restorers. Either detector supplies a number per patch that a second network must be built to consume. Our model measures the same field from its own residual and restores each patch at its own level in the same forward pass.

L Differential Diffusion's adherence benchmark

Levin & Fried (2023) measure whether the amount of change follows the painted map. For a change map M they average the per-pixel LPIPS map between input and output over 1000 ImageNet images, subtract the same average for a full-change map, and report the correlation with M (CAM) and the residual of the best affine fit (DAM). We rebuild the benchmark on JiT-B/16 with no class label (their empty prompt) and 20 steps, on 1000 validation images and three maps drawn from their description on our patch grid: a left-to-right gradient, four shapes at strengths 0.33–1, and four triangles at 0–1. Our port of Differential Diffusion scores close to their published Stable Diffusion numbers (0.97/0.81/0.93), so the port behaves like the original. Told nothing, ours follows the map almost as closely as Differential Diffusion handed it, within 0.01–0.016 in CAM, and stays closer to the original on every map (Table A11). Maps are on the patch grid because our model acts per patch; Differential Diffusion could follow a finer map.

Table A11: The adherence benchmark of Levin & Fried (2023) on JiT-B/16, 1000 ImageNet validation images, no class label, 20 steps. CAM: correlation between the change map and the measured change; DAM: residual of the best affine fit, per pixel; LPIPS: distance to the original. CAM 95% bootstrap intervals are within ±0.012. Ours copies back the patches it measures as clean.

map	method	CAM ↑	DAM/px ↓	LPIPS ↓
gradient	Differential Diffusion (handed the map)	**0.960**	**0.086**	0.243
	ours, told nothing	0.947	0.098	**0.218**
shapes	Differential Diffusion (handed the map)	**0.736**	**0.244**	0.174
	ours, told nothing	0.727	0.247	**0.133**
triangles	Differential Diffusion (handed the map)	**0.921**	**0.141**	0.238
	ours, told nothing	0.905	0.154	**0.199**

Table A12: Editing corrupted photographs, and repairing them. LPIPS to the clean photograph, 1000 ImageNet validation images per row, JiT-B/16, class label, 20 steps. For edits this measures how well the photograph is preserved; adherence to the painted map is in Table A11. Edits are random thick graded strokes; repairs have no edit, and Differential Diffusion (DD) is handed the corruption as its map. ‡Also handed the photograph's true noise map, with held pixels carrying only the noise they lack. *Ours copies back the patches it measures as clean ($\hat{t} > 0.88$); on the other rows it has no paste-back. Last column: share of images where ours is closer than DD with the edit map, and than DD‡.

	ours	Differential Diffusion		SDEdit	ours closer
setting	(no map)	edit map	+ true map‡	+ paste	than DD / DD‡
edit, clean photo*	**0.177**	0.220	0.220	0.297	90% / 90%
edit, noise σ_{255}=10	**0.192**	0.231	0.228	0.367	84% / 87%
edit, noise σ_{255}=25	**0.199**	0.365	0.243	0.524	99% / 91%
edit, noise σ_{255}=50	**0.221**	0.585	0.264	0.735	100% / 90%
edit, smooth noise	**0.261**	0.763	0.287	0.853	100% / 75%
edit, brightness noise, per pixel	0.279	0.594	**0.265**	0.734	100% / 54%
edit, brightness noise, per patch	**0.245**	0.594	0.265	0.735	100% / 71%
repair, graded noise stroke	0.170	0.527	**0.160**	0.431	100% / 50%
repair, noise-filled holes	**0.090**	0.219	0.106	0.244	100% / 81%

M Training-field variants on the restoration benchmark

The default-band restoration numbers of the training arms of Appendix D, on the 200-image draw of Table A7. All arms tie on uniform noise to within 0.23 dB, as they must: the field type only matters where the field varies. The order on the varying families follows how much sub-patch and irregular structure the training field contains. BAND draws one level per patch and trails REGION by 0.84 dB on the smooth field and 0.47 dB on piecewise; BAND+FINE, which splits patches in half during training, recovers half of the smooth gap; REGION, whose regions follow image content and cut through patches, is best or tied on all four families. Brightness-dependent noise separates nothing, since its field is set by the image and every arm has seen fields of that scale.

The low-noise BAND arm, trained on $\sigma_{255} \in [5, 50]$, is the control for the band choice: shifting the training band down costs 1.0 dB on smooth and 0.6 dB on brightness-dependent noise in the default band and, as Table A16 shows, buys at most 0.15 dB in its own band. The field type, not the band, is what the training decides.

N Real sensor noise, zero-shot

Real camera noise, evaluated without any training on it.

Table A13: PSNR (dB) of the training-field variants, default band $\sigma_{255} \in [13, 153]$, $n = 200$, same protocol as Table 2. Discussed in Appendix D.

method	uniform	smooth	piecewise	poisson
REGION (ours)	26.93	26.71	26.88	27.70
BAND+FINE	26.85	26.35	26.80	27.61
BAND	26.84	25.87	26.41	27.69
BAND, low-noise arm	26.70	24.87	26.36	27.10

Table A14: SIDD validation, sRGB, 1280 blocks at 256px, PSNR on $[0, 255]$ and SSIM over all blocks, 600-epoch checkpoints. No model was trained on SIDD. The published denoisers run blind; DRUNet is handed the map REGION measured. "Blind, global clock" is the uniform-trained blind model reading one level for the whole crop.

method	PSNR	SSIM
noisy input	23.66	0.328
CDnCNN-B (Zhang et al., 2017)	26.20	0.430
Restormer (Zamir et al., 2022)	23.99	0.344
DRUNet (Zhang et al., 2022), fed our measured map	27.59	0.525
blind, global clock	**27.69**	**0.633**
REGION (ours), per-patch map	26.10	0.535

SIDD (Abdelhamed et al., 2018) contains corruptions we did not design: real, signal-dependent, and nearly uniform within a crop. The strongest SIDD systems are trained on SIDD itself, a different claim from the zero-exposure transfer we test. Zero-shot, the clock's single global reading reaches 27.69 dB, ahead of every published denoiser we ran and of DRUNet fed our map (Table A14). Because the field is nearly uniform, a per-patch map only adds estimator noise here, and reading per patch costs 1.6 dB (26.10), about the level of the published blind denoisers (CDnCNN-B 26.20, Restormer 23.99). The measured map still drives DRUNet, a map-fed denoiser, to 27.59: the clock as a plug-in instrument for an existing restorer. The deployment rule is therefore: on noise that is close to uniform, read one global level; measure per patch only where the field is expected to vary.

O Graded strokes and outpainting

Table A15 gives the numbers behind the "Strokes and borders" paragraph of Section 5, with the 600-epoch models and 200 ImageNet validation images per cell. The graded stroke is 4 tokens wide with a 2-token halo, covers 15% of the image, and ramps from $\sigma_{255} \approx 13$ at its edge to the core strength given in the table. Outpainting pads the canvas with pure noise, either around the centre quarter (4× uncrop, 75% invented) or on the right half. Ours returns one $\hat{x}$ for the strokes and a 20-step sample for outpainting; no model is given a mask.

P Low-noise band

In the low band, $\sigma_{255} \in [5, 50]$, the published denoisers are in their training band and the residual internal clock reads the model's own error rather than the noise (Appendix C). Denoisers trained on this band lead, while we remain ahead of the conditioned model given the oracle mean on the smooth and piecewise fields, and perceptual quality is close (LPIPS 0.08–0.10 for us against DRUNet's 0.09–0.10).

Q Qualitative results on three datasets

Two photographs per dataset under six corruptions, one model per dataset told nothing. Three things are worth checking in each page. First, the measured map (last row of each block) follows the corruption and not the content. It is flat under uniform noise, a smooth

Table A15: Graded strokes (left) and outpainting (right). †handed the true σ-map, which for outpainting is the mask. "Drop": PSNR lost from the weakest to the strongest stroke. Ours is one $\hat{x}$ on strokes and a 20-step sample on outpainting. 600-epoch models, n=200.

	graded stroke, PSNR, core σ (in $[-1, 1]$ units)					outpainting, LPIPS	
method	0.39	1.6	6.4	20	drop	4× uncrop	right half
corrupted input	24.95	19.16	15.97	14.64	–	1.170	0.841
DRUNet†	**34.78**	**28.52**	23.71	21.18	13.6	0.779	0.655
CDnCNN-B	33.48	25.89	19.89	17.97	15.5	1.007	0.736
DiffPIR†	–	–	–	–	–	0.454	0.306
ours	28.60	27.92	**26.94**	**26.12**	**2.5**	**0.349**	**0.264**

Table A16: PSNR (dB) in the low band, $\sigma_{255} \in [5, 50]$, $n = 200$; same protocol and rows as Table 2. Here the published denoisers are in their training band and we are below the crossover of Appendix C. * marks a cell drawn outside that model's published training band.

method	uniform	smooth	piecewise	poisson
corrupted input	20.61	19.48	20.01	19.49
FFDNet (Zhang et al., 2018)	30.44	28.00	26.67	29.42
DRUNet (Zhang et al., 2022)	**30.94**	**30.40**	**30.91**	**30.50***
CDnCNN-B (Zhang et al., 2017)	30.08	29.70	30.16	29.67
Restormer (Zamir et al., 2022)	**30.98**	27.36	25.64	29.43*
DiffPIR (Zhu et al., 2023)	27.52	27.23	28.01	26.98
blind, global clock	29.44	26.06	24.73	27.51
REGION (ours)	29.17	29.05	28.63	28.34
cond., oracle mean $\bar{t}$	**30.23**	28.79	27.86	**29.53**
variants (Appendix D)				
BAND+FINE	29.36	29.19	28.72	28.44
BAND	29.40	28.84	28.68	28.51
BAND, low-noise arm	29.44	28.23	28.83	27.91

ramp under the smooth field, two flat plateaus under the piecewise field, brightest where the image is bright under brightness-dependent noise, and a hard patch-aligned region under the hole and the stroke, with the surrounding photograph read as clean. Second, the blind model with its global clock (second row) shows the failure that motivates the paper. One level for the whole image over-smooths the clean part and leaves speckles in the noisy part, and a hole is denoised as if it were mild noise and stays a block of color. Third, our restoration (third row) has no visible boundary where the noise level changes, which is what a correct per-patch entry into the sampler should give.

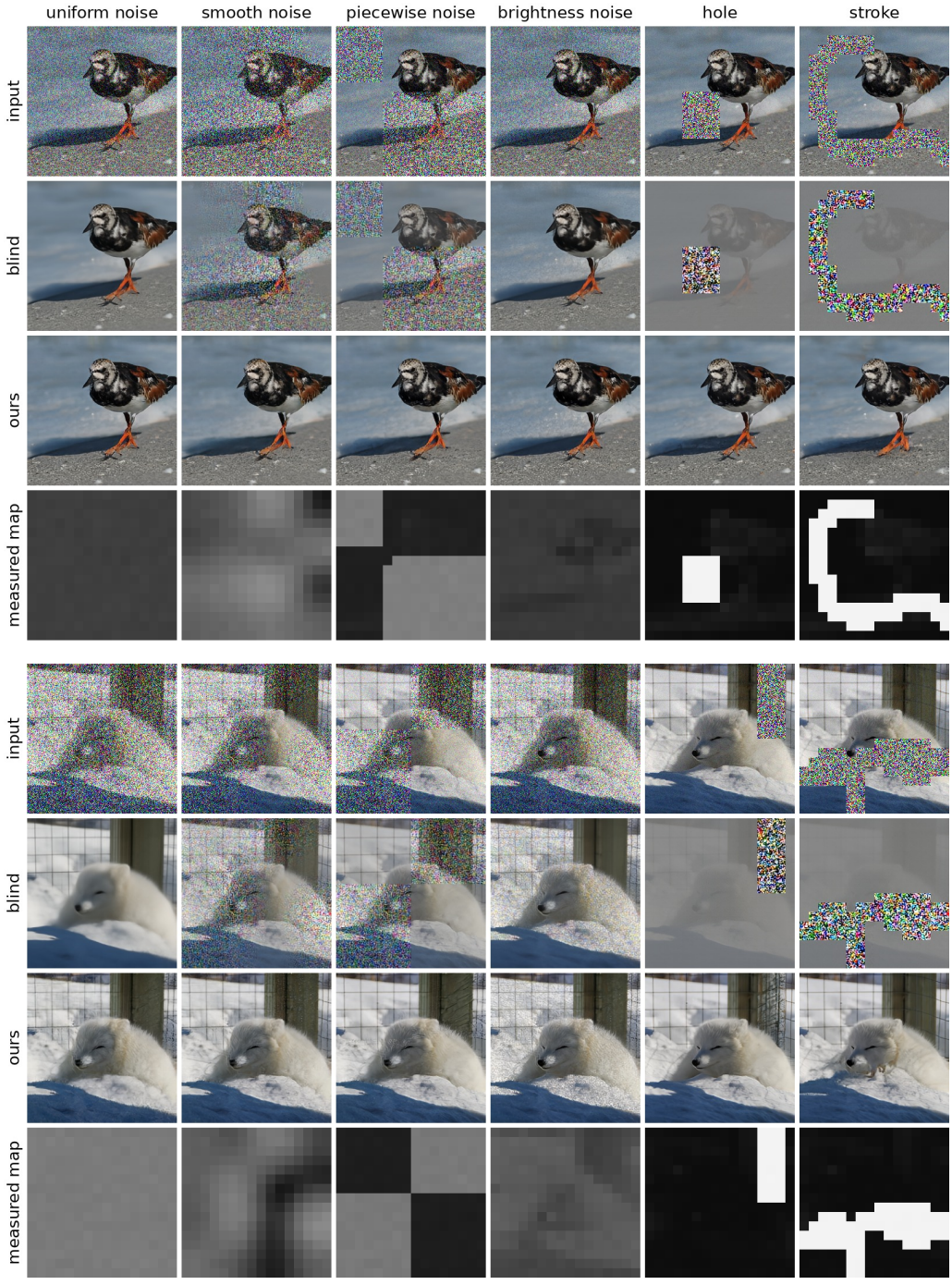


Figure A6: ImageNet, one model told nothing (REGION). Two photographs under six corruptions: uniform noise, smooth noise, piecewise noise, brightness-dependent noise, a noise-filled hole, and a noise stroke. Rows per photograph: the corrupted input, the blind model with its global clock, our restoration, and the per-patch map our model measured (brighter is noisier). Holes and strokes are sampled; the rest is one $\hat{x}$.

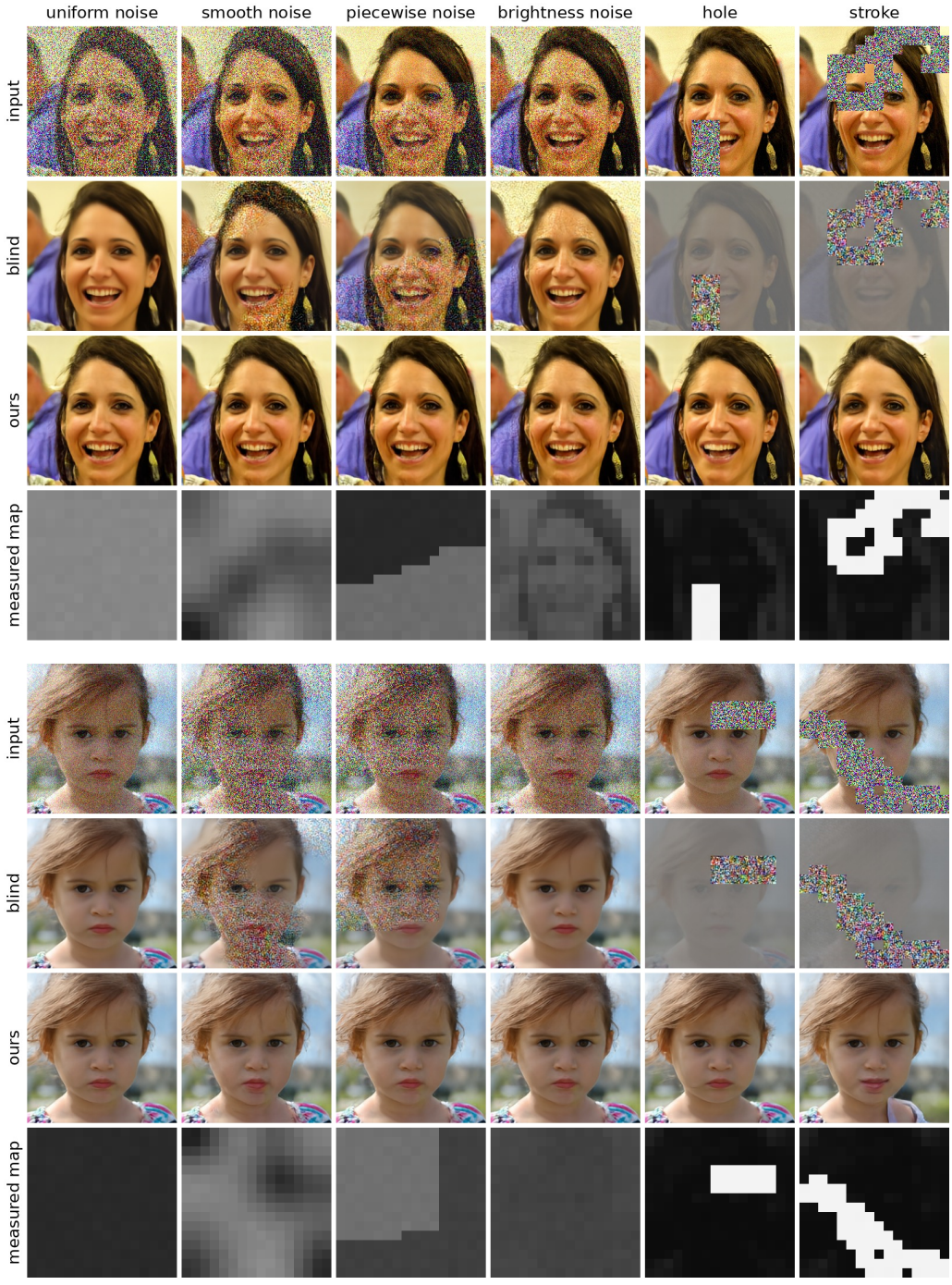


Figure A7: FFHQ, one model told nothing (REGION, 1200 epochs). Two photographs under five corruptions: smooth noise, piecewise noise, brightness-dependent noise, a noise-filled hole, and a noise stroke. Rows per photograph: the corrupted input, our restoration, the blind model with its global clock, and the per-patch map our model measured (brighter is noisier). Holes and strokes are sampled; the rest is one $\hat{x}$.

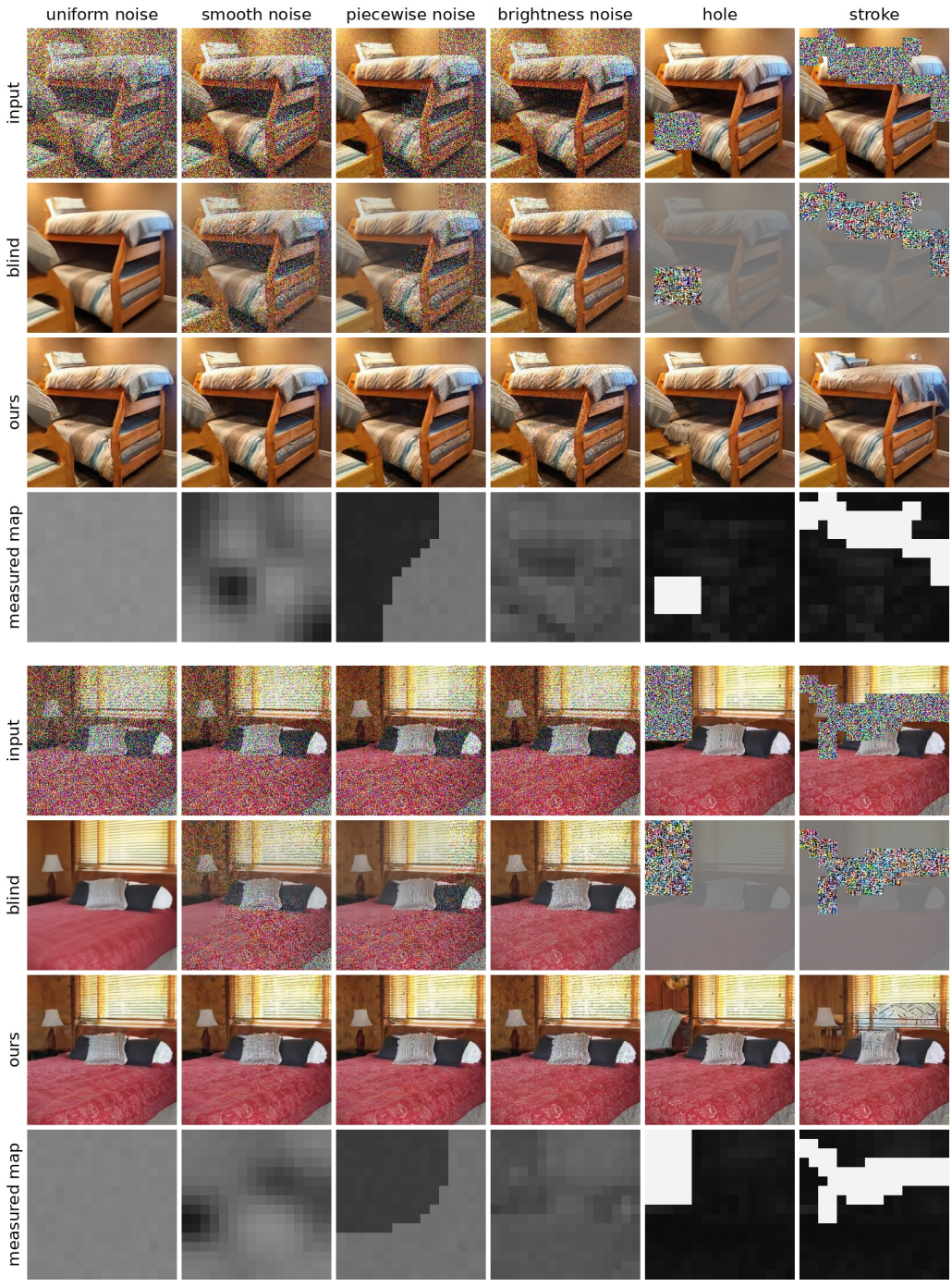


Figure A8: LSUN bedrooms, one model told nothing (REGION, 400 epochs). Two photographs under five corruptions: smooth noise, piecewise noise, brightness-dependent noise, a noise-filled hole, and a noise stroke. Rows per photograph: the corrupted input, our restoration, the blind model with its global clock, and the per-patch map our model measured (brighter is noisier). Holes and strokes are sampled; the rest is one $\hat{x}$.