

Stochastic localization for generative modeling: beyond DDPMs

Stanford CS229 Project, Spring 2026

Elana Chen

Department of Computer Science
Stanford University
elanac25@stanford.edu

Ruochuan Xu

Department of Mathematics
Stanford University
ruochuan@stanford.edu

1 Introduction

Generative modeling widely uses denoising diffusion probabilistic models (DDPMs), with a range of applications in image synthesis, audio generation, etc. Despite the many theoretical guarantees for the convergence/stability of such models, the full force of the theoretical ideas behind DDPMs has not been sufficiently explored in applications. Several forward diffusion processes, such as the isotropic Gaussian process, are chosen by default due to their analytical simplicity, with no guarantee that they are good candidates for the data geometry in intended applications.

The theoretical work by Montanari (2023) shows that the seemingly canonical method of sampling by reversing a diffusion sits inside a much wider class of sampling methods inspired by the technique of stochastic localization in Markov chain mixing time analysis. This wide class of sampling schemes is based on estimating the posterior mean of a desired sample conditioned on an *observation process*. When the data distribution contains low-dimensional global structure, such as class identity, coarse layout, or global color statistics, an isotropic noising process may require the denoiser to infer that structure indirectly from noisy local observations. Thus, fixing the neural network architecture used for score-matching, models that use isotropic diffusion often fail to learn certain latent structures of the data. On the other hand, models that use carefully engineered observation processes can simplify the denoising task and improve sampling in controlled examples, as demonstrated in sections 5,6,7 of Montanari (2023).

Since Montanari demonstrated the superiority of tailored observation processes on synthetic data and Gaussian mixtures, we find it interesting and useful to test whether we can use similar ideas to improve sampling results in real-world datasets. On CIFAR-10 (Krizhevsky, 2009), we study whether discrete or continuous observations can expose class-like or low-frequency image structures to a small U-Net denoiser. The discrete observations include random labels, k-means pseudo-labels, PCA-tree pseudo-labels, and a true-label oracle. The continuous observations are low-resolution versions of the image, injected either through the time embedding or as spatial channels. On a landscape subset of MIT Places (Zhou et al., 2016), we study if explicit global color observations improve coherence in scene generation, so that the generated image is distinctly a beach or a mountain.

The input to our algorithms is a training set of RGB images x_i . Each model trains a U-Net denoiser under a DDPM noise-prediction objective. The baseline denoiser receives only a noisy image x_t and timestep t , and predicts the Gaussian noise ϵ . Observation-conditioned models additionally receive an observation $o(x_i)$, such as a pseudo-label, a true label, a low-resolution image, or a global color statistic. At generation time, we sample the observation from its empirical distribution and then run the learned reverse diffusion sampler conditioned on that observation. The output of each algorithm is a generator that produces synthetic images. Our goal is a controlled comparison in which the dataset, diffusion objective, sampler, and training budget are fixed as much as possible while the observation process is varied. We test the practical implications of the stochastic localization framework: when does revealing additional global structure help denoising, and when does it merely add unhelpful conditioning?

2 Related Work

Diffusion models generate data by learning to reverse a noising process. Early diffusion models were introduced by Sohl-Dickstein et al. (2015), and DDPMs popularized the practical noise-prediction objective for image synthesis (Ho et al., 2020). Score-based SDE models give a continuous-time view of the same idea (Song et al., 2021). Many improvements keep the standard isotropic observation but change other parts of the pipeline: DDIM and DPM-Solver accelerate sampling (Song et al., 2020; Lu et al., 2022), EDM improves schedules and preconditioning (Karras et al., 2022), and latent diffusion runs the model in a learned latent space (Rombach et al., 2022). In contrast, we focus on changing what information is revealed to the denoiser.

Our work is most closely connected to stochastic localization and conditional diffusion. Montanari (2023) interpreted reverse diffusion as one instance of a broader stochastic-localization framework, where sampling is driven by posterior means under an observation process. This viewpoint suggests that revealing latent variables or global statistics can simplify denoising. Conditional diffusion methods also show that useful side information can improve generation: classifier guidance and classifier-free guidance use labels or conditions to modify the learned score (Dhariwal and Nichol, 2021; Ho and Salimans, 2022). Our CIFAR-10 true-label oracle is related to class-conditional diffusion, but we are treating our datasets primarily as unlabeled, and our main question is whether pseudo-labels, low-resolution downsamples, or color averages can serve as useful observation processes.

The closest empirical work to our discrete-observation experiments is cluster-conditioned diffusion. Adaloglou et al. (2025) studied cluster assignments as label-free conditions and show that carefully designed clusters can improve image synthesis. Our goal is more modest: we use simple k-means, PCA-tree, random-label, low-resolution observations, and color averages as controlled probes of the observation-process hypothesis under a fixed small DDPM architecture.

3 Dataset and Features

CIFAR-10. We use CIFAR-10 (Krizhevsky, 2009): 50,000 training and 10,000 test images at 32×32 RGB resolution. Pixels are scaled to $[-1, 1]$, with no data augmentation. For k-means and PCA-tree observations, we extract unsupervised features by area-downsampling each image to 8×8 , flattening, standardizing, and reducing to 64 PCA components. K-means labels are obtained with MiniBatch k-means using $K = 10$. PCA-tree labels are obtained by recursively splitting the largest leaf by the median projection onto its first principal component until ten leaves are formed. For continuous low-resolution observations, we use $A_4(x) = \text{AvgPool}_{4 \times 4}(x)$ for the time-embedding injection and $A_8(x) = \text{AvgPool}_{8 \times 8}(x)$ for the spatial-channel model. CIFAR labels of the ten classes such as ships and cats are used only for the true-label oracle and for post-hoc semantic evaluation.

MIT Places (Landscape Subset) We use a four-category landscape subset of Places365 (Zhou et al., 2016): beach, mountain, rainforest, and badlands, yielding 20,000 training and 400 validation images resized to 64×64 RGB and normalized to $[-1, 1]$. They were selected for their strong and visually distinct dominant color signatures: beach scenes are characterized by blue water and tan sand, mountain scenes by grey rock and white snow, rainforest scenes by saturated green canopy, and badlands scenes by red and ochre arid terrain. If per-channel spatial means are globally informative, the denoiser should learn to commit to the dominant tone of each scene type rather than producing patchwork outputs. For the linear observation arm, per-channel spatial means $\mu \in \mathbb{R}^3$ are broadcast as three additional input channels, giving a six-channel input; these mean channels are noised more slowly than pixel channels, implementing the linear observation process of Montanari Section 4.6. The validation set is used for all evaluation metrics.

4 Methods

DDPM baseline. All models use the standard DDPM ϵ -prediction objective with forward process $x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon$, where $\bar{\alpha}_t = \prod_{s=1}^t (1 - \beta_s)$, minimizing

$$\mathcal{L}_{\text{DDPM}} = \mathbb{E}_{x_0, t, \epsilon} [\|\epsilon - \epsilon_\theta(x_t, t)\|_2^2].$$

At generation time we initialize $x_T \sim \mathcal{N}(0, I)$ and run the standard DDPM reverse update. The baseline treats the image as an unstructured vector with no explicit latent variable.

CIFAR-10 observations. For discrete observations $O = c \in \{0, \dots, 9\}$, we use a shared conditional U-Net $\epsilon_\theta(x_t, t, c) \approx \epsilon$, where c is embedded and added to the timestep embedding. We compare random labels, k-means pseudo-labels, PCA-tree pseudo-labels, and a true-label oracle. For continuous observations we define $A_r(x) = \text{AvgPool}_{r \times r}(x)$ and test two injection mechanisms: a time-embedding version where $A_4(x)$ is flattened through an MLP and added to the timestep embedding, and a spatial-channel version where $A_8(x)$ is upsampled and concatenated with the noisy image as $\tilde{x}_t = [x_t, \text{Upsample}(A_8(x_0))]$ $\in \mathbb{R}^{6 \times 32 \times 32}$. Random-label and randomly permuted low-resolution observations serve as negative controls.

Linear global observation for MIT Places. For $x \in \mathbb{R}^{3 \times H \times W}$, we compute per-channel spatial means $\mu_c(x) = \frac{1}{HW} \sum_{i,j} x_{cij}$ and form the augmented state

$$z = Lx = [x; \text{broadcast}(\mu(x))] \in \mathbb{R}^{6 \times H \times W}.$$

Pixel channels follow the standard noise schedule; global channels use reduced noise $\mu_t = \sqrt{\bar{\alpha}_t} \text{broadcast}(\mu) + \lambda \sqrt{1 - \bar{\alpha}_t} \eta$ with $\lambda = 0.35$, so global color statistics remain informative at higher noise levels. We train a six-channel U-Net predicting pixel noise and refresh global channels from the current pixel estimate during sampling. See Appendix A for the full stochastic-localization viewpoint and Appendix D for the complete MIT Places construction.

5 Experiments / Results / Discussion

Experimental setup We trained all models with AdamW at learning rate 2×10^{-4} , diffusion horizon $T=1000$, and a linear β -schedule from 10^{-4} to 0.02. CIFAR-10 models use batch size 128 at 32×32 ; MIT Places models use batch size 64 at 64×64 and `num_workers=0` on GPU. Hyperparameters follow standard DDPM defaults and Montanari-style small U-Net settings so differences across arms reflect the observation process rather than tuning. CIFAR-10 evaluation uses the 50,000-image training set; MIT Places uses a fixed 400-image validation holdout (100 per scene) while training on 20,000 images. Each generative arm produces 5,000 samples from the final checkpoint.

Model	Observation process	Overall FID	Semantic TV	Training epochs	Backbone	Observation Injection
Baseline	none	50.099	0.242	20	Small U-Net	None
Random label	Random discrete label	51.114	0.253	20	Small U-Net	Label embedding
K-means	Learned discrete pseudo-label	51.910	0.240	20	Small U-Net	Label embedding
PCA-tree	Learned half-space pseudo-label	52.224	0.239	20	Small U-Net	Label embedding
True-label oracle	CIFAR label	45.446	0.152	20	Small U-Net	Label embedding
Low-res spatial	8 x 8 low-resolution image	39.477	0.093	20	Small U-Net	Upsampled, concatenated channels
Random low-res spatial	Randomly permuted 8 x 8 low-resolution image	51.300	0.169	20	Small U-Net	Upsampled, concatenated channels

(a) CIFAR-10 results with smaller U-Net, 20 epochs

Model	Observation process	Overall FID	Semantic TV	Training epochs	Backbone	Observation Injection
Baseline	none	37.824	0.174	40	Wide U-Net	None
Random label	Random discrete label	38.138	0.174	40	Wide U-Net	Label embedding
K-means	Learned discrete pseudo-label	37.275	0.165	40	Wide U-Net	Label embedding
PCA-tree	Learned half-space pseudo-label	38.852	0.177	40	Wide U-Net	Label embedding
True-label oracle	CIFAR label	31.312	0.075	40	Wide U-Net	Label embedding
Low-res spatial	8 x 8 low-resolution image	28.930	0.093	40	Wide U-Net	Upsampled, concatenated channels
Random low-res spatial	Randomly permuted 8 x 8 low-resolution image	38.677	0.169	40	Wide U-Net	Upsampled, concatenated channels

(b) CIFAR-10 results with wider U-Net, 40 epochs

Figure 1: Evaluation of generative models trained on CIFAR-10 with different observation processes. Both FID and semantic TV are computed from 5,000 generated samples against the 50,000 training set.

Evaluation metrics *Fréchet Inception Distance (FID)*. We measure distributional similarity between generated and real images using FID on Inception-v3 features:

$$\text{FID} = \|\mu_r - \mu_g\|_2^2 + \text{Tr}(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{1/2}), \quad (1)$$

where (μ_r, Σ_r) and (μ_g, Σ_g) are the mean and covariance of feature vectors for real and generated sets. Lower is better. FID captures overall visual quality but is not sensitive to global color coherence alone.

Semantic TV (evaluated for CIFAR-10). This is the total variation distance between the predicted class distribution of generated images under a pretrained CIFAR-10 classifier and the CIFAR-10 training distribution, which is uniform on the 10 classes.

Global color spatial variance (evaluated for MIT Places). For image x with channels $c \in R, G, B$, we compute

$$V(x) = \frac{1}{3} \sum_{c \in \{R, G, B\}} \text{Var}_{i,j}[x_{c,i,j}], \quad (2)$$

and average $V(x)$ over a sample set. Lower values indicate more uniform scene-wide tint; real validation images provide a reference level rather than a target minimum, since legitimate sky/ground structure also induces spatial variation.

Spectral PSD distance (evaluated for MIT Places). We compute the radially averaged power spectral density (PSD) of grayscale images and report the ℓ_2 distance between the mean generated PSD curve and the mean real-validation curve.

Part 1: CIFAR-10 results Figure 1 (a) (b) report CIFAR-10 results using 5,000 generated samples per model. FID measures overall image distribution quality, while semantic TV measures class-balance error using a pretrained CIFAR-10 classifier. Lower is better for both. The 20-epoch runs use a smaller U-Net, while the 40-epoch runs use a wider U-Net; within each table, all models share the same training budget and differ only in the observation process and its injection mechanism.

The discrete-observation results show that useful conditioning depends on the quality of the revealed information. Random labels do not help, confirming that the gains are not caused merely by adding an embedding. K-means pseudo-labels become mildly useful in the 40-epoch wide-U-Net setting, improving FID from 37.824 to 37.275 and semantic TV from 0.174 to 0.165. However, this improvement is small compared with the true-label oracle, which reaches 31.312 FID and 0.075 semantic TV. PCA-tree labels perform worse than the baseline, showing that not every unsupervised partition gives a useful observation.

The strongest CIFAR-10 result comes from the continuous low-resolution spatial observation. Injecting an 8×8 low-resolution image as aligned spatial channels gives the best FID, 28.930, compared with 37.824 for the baseline and 38.677 for the randomly permuted low-resolution control. Since the random-lowres model has the same six-channel architecture and the same marginal observation distribution, this gap shows that the improvement comes from the meaningful correspondence between the observation and the target image, not merely from extra input channels. The lowres spatial model also improves semantic TV to 0.093, though the true-label oracle remains best on semantic mode balance.

Part 2: MIT Places landscape results Table reports 5,000 generated images per arm vs. 400 real validation landscapes. Figure shows radial PSD curves.

Table 1: Places results (5k samples vs. 400 val images).			
Model	FID ↓	Spatial var. ↓	PSD ℓ_2 ↓
Baseline	152.6	0.170	2466
Linear global obs.	96.8	0.131	2543
Real val (ref.)	—	0.182	—

Linear global observation substantially improves FID (96.8 vs. 152.6), indicating better overall match to the validation distribution. Spatial color variance drops from 0.170 (baseline) to 0.131 (linear), consistent with reduced patchy global tint;

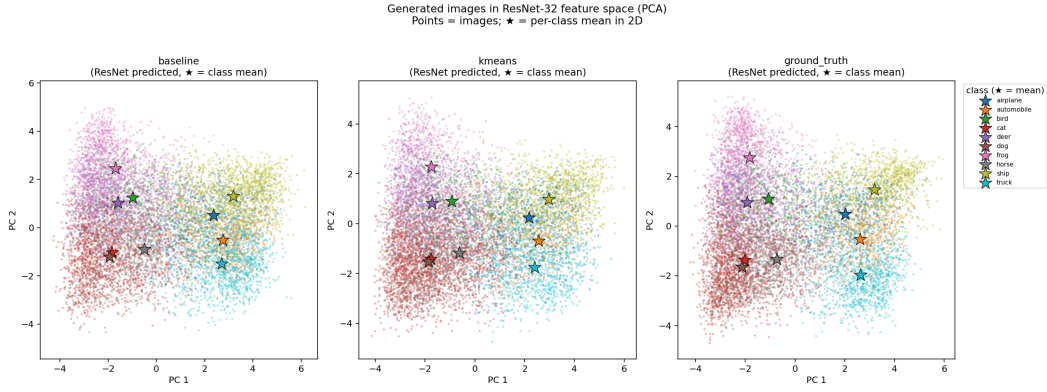


Figure 2: Post-hoc visualization of CIFAR-10 generated samples in a pretrained ResNet feature space, projected to two dimensions by PCA and colored by CIFAR semantic class. The pseudo-label-conditioned samples do not form clean semantic clusters: points of different true classes remain heavily mixed, and the pseudo-label structure does not align well with CIFAR categories: for instance, it fails to cleanly separate cats and dogs. This supports our quantitative label diagnostics and explains why k-means pseudo-labels help only modestly compared with the true-label oracle.

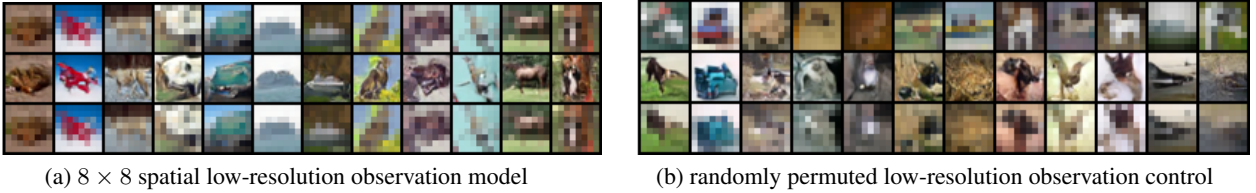


Figure 3: Qualitative effect of the spatial low-resolution observation. Each column is one generated sample. The rows show, from top to bottom, the sampled condition A_{cond} , the generated image $\hat{X}$, and the generated image’s low-resolution projection $A(\hat{X})$. In the clean low-resolution observation model, $A(\hat{X})$ tracks A_{cond} , showing that the generator uses the observation to control global color and layout. In the randomly permuted control, the condition and projection are unrelated, as expected.

however, linear variance is below real validation (0.182), suggesting some over-smoothing relative to natural landscapes. PSD ℓ_2 slightly favors baseline over linear (2466 vs. 2543), so low-frequency structure is mixed; we weight FID and qualitative grids more heavily for the main claim. Visually, baseline samples exhibit inconsistent color patches across sky and ground; linear samples show more coherent scene-wide color.

Discussion

CIFAR interpretation. The CIFAR experiments suggest two lessons. First, discrete observations help only when they are informative: the true-label oracle helps strongly, k-means helps modestly with more capacity, and random or PCA-tree labels do not help. Second, continuous spatial observations can be more effective than hard pseudo-labels. Unlike a discrete label, an 8×8 low-resolution image preserves color, layout, and coarse shape information. The qualitative panels in Figure 3 show that the clean lowres model uses this observation to guide global structure, while the randomly permuted control does not. Thus, observation design includes both what information is revealed and how it is injected into the denoiser.

Observation injection matters. The CIFAR low-resolution experiments show that the neural architecture used to present an observation is itself an important design choice. Flattening a 4×4 low-resolution image, passing it through an MLP, and adding it to the timestep embedding did not improve FID. In contrast, upsampling an 8×8 low-resolution image and concatenating it as spatial channels gave the best CIFAR FID. This comparison is not perfectly isolated, since it changes both the observation resolution and the injection mechanism, but it suggests a clear principle: spatial observations should be presented spatially. A low-resolution image contains aligned color and layout information, and concatenating it to x_t makes that information directly accessible to local convolutional filters.

Why linear global observation helps. Augmenting the state with spatial means noised on a weaker schedule gives the U-Net direct access to global color early in denoising, addressing a failure mode of isotropic diffusion on structured scenes without requiring scene labels.

Overfitting. We did not perform early stopping or extensive hyperparameter search; instead, we compare arms under matched training budgets and use random-observation controls to separate informative observations from extra conditioning capacity.

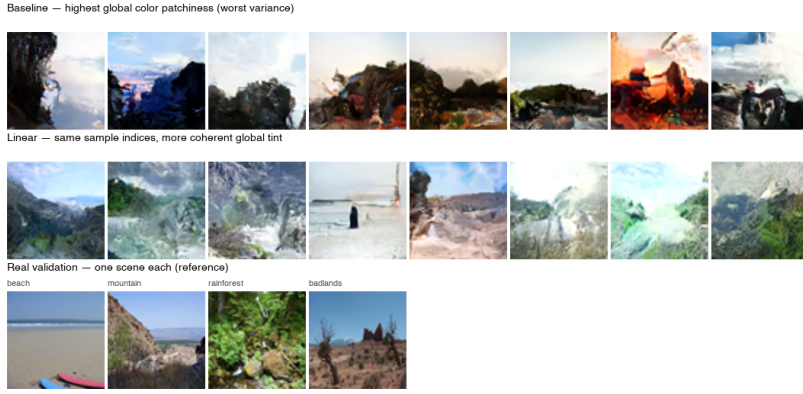


Figure 4: Places samples with 200 epochs: baseline (patchiest global color, top), linear global observation at the same indices (middle), and real validation references by scene (bottom)

For CIFAR-10, the randomly permuted low-resolution model has the same six-channel architecture and marginal observation distribution as the clean lowres model but performs much worse, suggesting that the lowres improvement is not simply due to additional input channels. For MIT Places, the validation FID gap between baseline and linear global observation is large, and the qualitative samples show visibly more coherent global color. However, our CIFAR FID uses the training set as the real reference, so those FID values should be interpreted as within-CIFAR comparisons rather than as held-out generalization estimates.

Limitations. (i) Small U-Net capacity and 64×64 Places resolution. (ii) Only using training instead of testing samples for evaluation on CIFAR-10; only 400 validation images for Places FID. (iii) Global variance alone does not distinguish good structured variation from blur. (iv) We omit the full matrix denoiser from Montanari and use a finite-receptive-field U-Net instead. (v) A limitation of the low-resolution spatial CIFAR model is that $A(x_i)$ is sampled cleanly from the empirical training distribution. Since an 8×8 RGB observation is already fairly informative, the model may behave partly like stochastic super-resolution from training-set sketches, improving FID while potentially reducing diversity or increasing nearest-neighbor similarity.

Overall, results support the project thesis: *matching the observation process to latent structure* matters—ground-truth latents help on CIFAR, and linear global observations help on Places—while naive isotropic or poorly estimated latents leave performance near baseline.

6 Conclusion / Future Work

We empirically studied Montanari’s observation-process view of diffusion sampling on CIFAR-10 and MIT Places. Our results support a qualified version of the thesis: observation processes help when they reveal structure that is both informative about the data and usable by the denoising architecture. On CIFAR-10, arbitrary random labels do not help, while the true-label oracle substantially improves both FID and semantic mode balance. K-means pseudo-labels provide a smaller gain in the wider 40-epoch model, suggesting that simple unsupervised clusters capture some useful structure but fall far short of semantic labels. The strongest CIFAR result is the 8×8 spatial low-resolution observation, which improves FID from 37.824 for the baseline to 28.930, while its randomly permuted control performs much worse at 38.677. This shows that coarse spatial observations can be highly useful when they are meaningfully paired with the target image and injected into the U-Net as aligned channels.

On MIT Places, the linear global color observation improves FID from 152.6 to 96.8 and visibly reduces patchy color artifacts. This is the experiment closest to Montanari’s linear observation process: global color channels are included in the noised state and remain informative at higher noise levels. Together, the CIFAR and Places experiments show that observation-process design is not merely a theoretical reformulation of diffusion sampling. It leads to practical choices about what global information to reveal and how to make it accessible to a finite-capacity convolutional denoiser.

Several directions remain open. First, the CIFAR low-resolution spatial model conditions on clean 8×8 observations sampled from the empirical training set, so it may partly behave like stochastic super-resolution from training-set sketches. A more faithful implementation would diffuse the augmented state $(x, A(x))$, add noise to both pixel and low-resolution coordinates, and enforce consistency during reverse sampling. Second, better unsupervised observations, such as self-supervised or pretrained visual features, could test whether learned pseudo-labels can approach the true-label oracle without using CIFAR labels. Third, scaling to higher-resolution images and stronger U-Nets would clarify whether the gains from observation design persist when the baseline architecture has more capacity to infer global structure on its own. Finally, a systematic comparison of linear, discrete, and nonlinear observation processes would help turn the stochastic-localization perspective into a practical design toolkit for diffusion models.

7 Appendices

A Observation-Process Viewpoint

Let $X \sim \mu$ denote a data sample. In the stochastic-localization view, a sampler is specified by an observation process Y_t that gradually reveals information about X . The ideal denoising drift is a posterior mean

$$m(y, t) = \mathbb{E}[X \mid Y_t = y].$$

Standard isotropic diffusion is the special case where Y_t is a uniformly noised version of the image. More general observation processes can expose selected latent or global structure earlier than local pixel details.

We use two practical forms of this idea. For CIFAR-10, most observations are clean side observations $O = o(X)$, such as a pseudo-label or a low-resolution image. The denoiser is trained as

$$\epsilon_\theta(x_t, t, O) \approx \epsilon,$$

and at sampling time we draw O from its empirical training distribution before running conditional reverse diffusion. This corresponds to the decomposition of the target distribution μ as

$$\mu(dx) = \int \mathcal{L}(X \in dx \mid O = o) d(o_\# \mu)(o).$$

For MIT Places, we instead use an augmented linear observation $z = Lx$, where global color channels are included in the noised state. This is closer to Montanari’s linear-observation construction because the global channels evolve during the forward and reverse processes rather than being revealed cleanly at the start of denoising.

B Full DDPM Forward Process

Using the standard forward process

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon,$$

where $\bar{\alpha}_t = \prod_{s=1}^t (1 - \beta_s)$, the baseline minimizes

$$\mathcal{L}_{\text{DDPM}} = \mathbb{E}_{x_0, t, \epsilon} [\|\epsilon - \epsilon_\theta(x_t, t)\|_2^2].$$

At generation time, we initialize $x_T \sim \mathcal{N}(0, I)$ and use the standard DDPM reverse update.

C Full CIFAR-10 Continuous Observation Construction

For continuous CIFAR observations we define

$$A_r(x) = \text{AvgPool}_{r \times r}(x).$$

In the time-embedding version, $A_4(x)$ is flattened, passed through an MLP, and added to the timestep embedding. In the spatial-channel version, $A_8(x)$ is upsampled to 32×32 and concatenated with the noisy image:

$$\tilde{x}_t = [x_t, \text{Upsample}(A_8(x_0))] \in \mathbb{R}^{6 \times 32 \times 32}.$$

The model then predicts pixel noise from $(\tilde{x}_t, t)$.

D Full MIT Places Linear Observation Construction

For $x \in \mathbb{R}^{3 \times H \times W}$, the per-channel spatial means are defined as

$$\mu_c(x) = \frac{1}{HW} \sum_{i,j} x_{cij}, \quad \mu(x) = (\mu_R, \mu_G, \mu_B).$$

The augmented observation state is

$$z = Lx = [x; \text{broadcast}(\mu(x))] \in \mathbb{R}^{6 \times H \times W}.$$

Global channels are noised as

$$\mu_t = \sqrt{\bar{\alpha}_t} \text{broadcast}(\mu) + \lambda \sqrt{1 - \bar{\alpha}_t} \eta,$$

with $\lambda = 0.35$ and η drawn once per channel and broadcast spatially. ““

E Contributions

Elana Chen: implemented the CIFAR-10 pipeline for the baseline, K-means, and ground-truth label observation processes, trained all three models for 200 epochs, and generated and evaluated the corresponding samples ;implemented the full MIT Places pipeline including dataset filtering, the baseline isotropic DDPM, and the linear global observation process, trained both models for 200 epochs, and ran all Places evaluation metrics (FID, global color variance, spectral PSD).

Ruochuan Xu: implemented and trained the extended suite of CIFAR-10 observation processes: baseline, random label, K-means, PCA-tree, true-label oracle, low-resolution spatial, and random low-resolution spatial; led the theoretical and mathematical framing of the project, including connecting the experimental designs to Montanari’s stochastic localization framework.

References

- Nikolas Adaloglou, Tim Kaiser, Felix Michels, and Markus Kollmann. 2025. Rethinking cluster-conditioned diffusion models for label-free image synthesis. In *IEEE/CVF Winter Conference on Applications of Computer Vision*.
- Prafulla Dhariwal and Alex Nichol. 2021. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems*.
- Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *arXiv preprint arXiv:2006.11239*.
- Jonathan Ho and Tim Salimans. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.
- Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. 2022. Elucidating the design space of diffusion-based generative models. In *Advances in Neural Information Processing Systems*.
- Alex Krizhevsky. 2009. Learning multiple layers of features from tiny images. Technical report, University of Toronto.
- Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. 2022. DPM-Solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps. *arXiv preprint arXiv:2206.00927*.
- Andrea Montanari. 2023. Sampling, diffusions, and stochastic localization. *arXiv preprint arXiv:2305.10690*.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*.
- Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. 2015. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*.
- Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.
- Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. 2021. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*.
- Bolei Zhou, Aditya Khosla, Agata Lapedriza, Antonio Torralba, and Aude Oliva. 2016. Places: An image database for deep scene understanding.