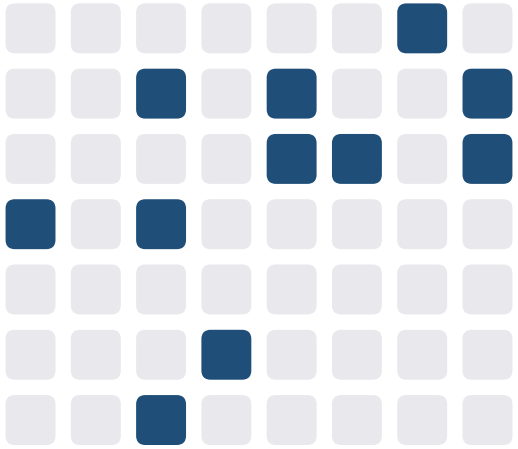
NeurIPS 2026

Sparsely Supervised Diffusion

Supervise less. Generate consistently.



Wenshuai Zhao[1,4] · Zhiyuan Li[1] · Yi Zhao[2] · Mohammad Hassan Vali[1]
Martin Trapp[3] · Joni Pajarinen[1] · Juho Kannala[1,5] · Arno Solin[1,4]

[1] Aalto University [2] Imperial College London [3] KTH Royal Institute of Technology
[4] ELLIS Institute Finland [5] University of Oulu

Project page

The problem

Locally plausible. Globally inconsistent.

Diffusion models can pick up accidental correlations from finite data. The parts look right; the whole does not fit together.

Baseline



SSD · η = 0.8



Same initial noise in each column · CelebA-50K

Outline

01

Method

Sparse supervision and partial views

02

Analysis

How masking attenuates correlations

03

Experiments

Generation quality and spatial consistency

Method

Keep the input.
Sparsify the supervision.

Input

full noisy image x_t

unchanged

Predict

any network $f\theta(x_t, t)$

unchanged

Supervise

retain $p = 1 - \eta$ of the loss



```
error = (prediction - target).square()
mask = (torch.rand_like(error) > eta).float()
loss = (error * mask).sum() / mask.sum().clamp(min=1)
```

Method

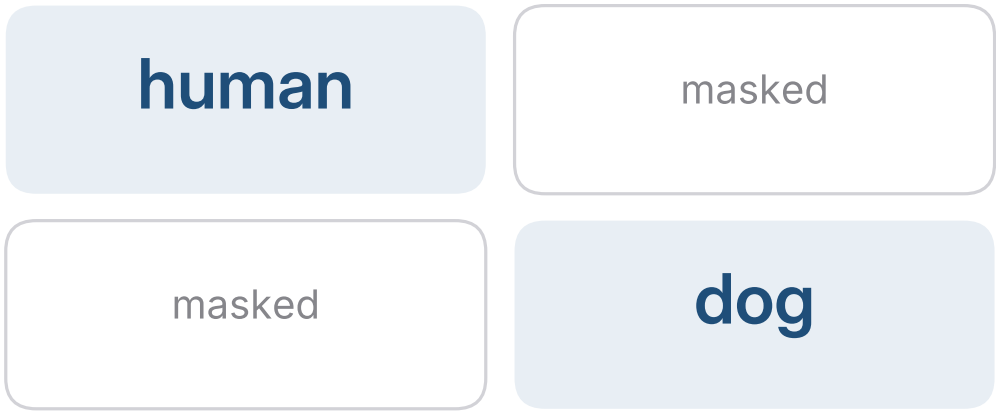
Partial views weaken an accidental association

Standard supervision



Learns: human ⇒ dog

SSD: random partial views



Each is also supervised on its own

The model still receives the full input. A masked entry is unobserved for that loss, not evidence of absence.

Analysis

Masking attenuates cross-coordinate covariance

$$\Sigma_{mask} = p^2\Sigma + p(1-p)\,\mathrm{diag}(\Sigma)$$

Diagonal · coordinate variance

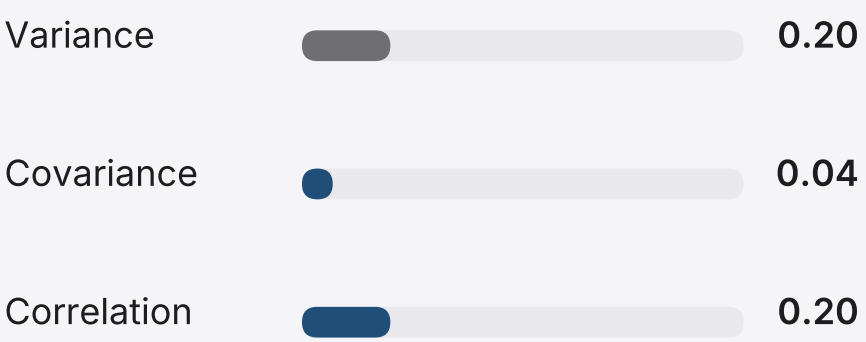
$$p\,\Sigma_{ii}$$

Off-diagonal · cross-coordinate covariance

$$p^2\,\Sigma_{ij}$$

At η = 0.8, so p = 0.2

relative to the unmasked value



Analytical model: zero-mean Gaussian data, independent Bernoulli masks, p = 1 − η. Every correlation shrinks by a factor p.

Experiments

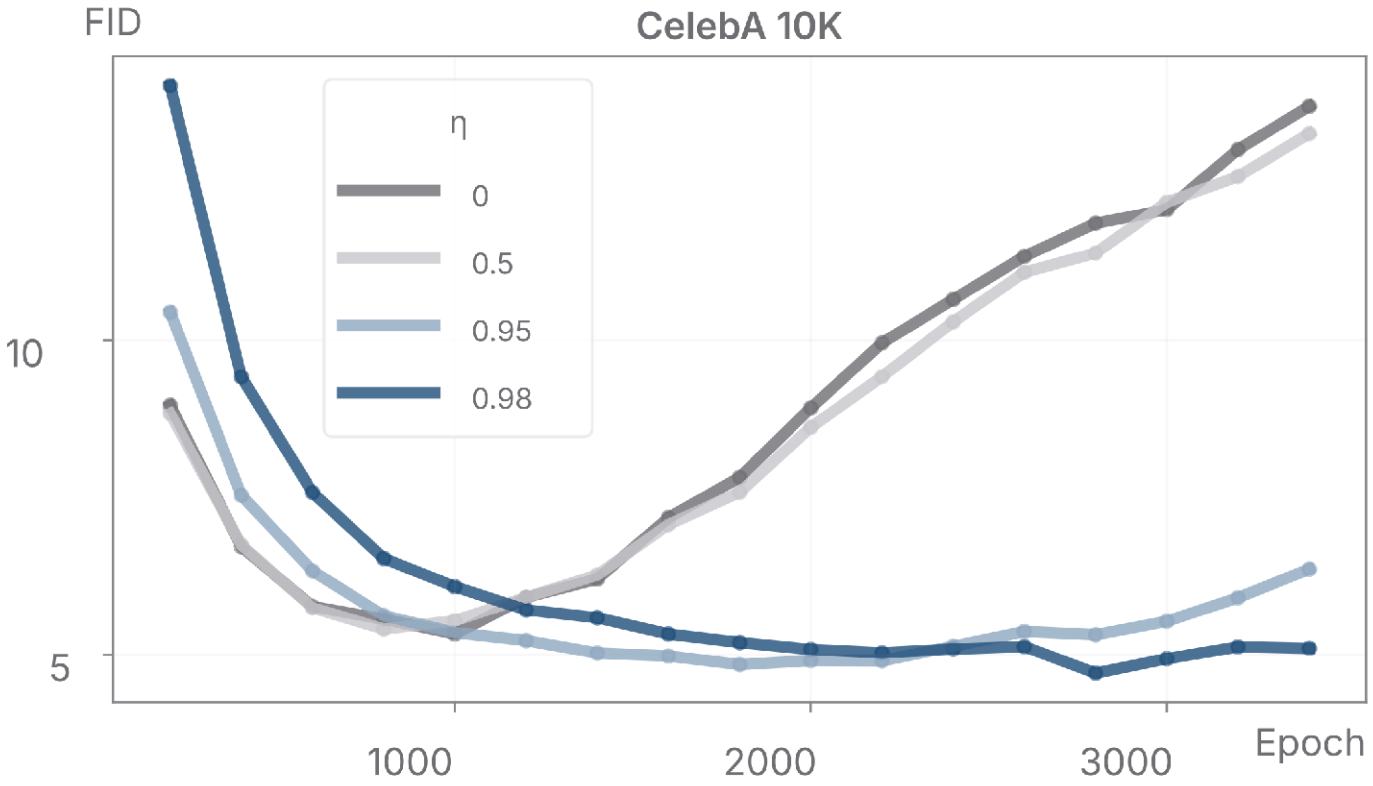
Competitive generation quality

Dataset	Baseline	SSD
CIFAR-10	2.28 ± 0.04	**2.24 ± 0.05**
LSUN Bedroom	1.40 ± 0.04	**1.39 ± 0.01**
CelebA-50K	2.23 ± 0.23	**1.85 ± 0.14**
CelebA-10K	13.72	**5.10**
ImageNet 256	4.98	**4.83**

FID ↓ · mean ± s.d. for the first three rows · ImageNet: DiT-XL/2, CFG = 1.5

Experiments

More stable training on limited data



FID on CelebA-10K, η = 0.98

5.10

baseline 13.72

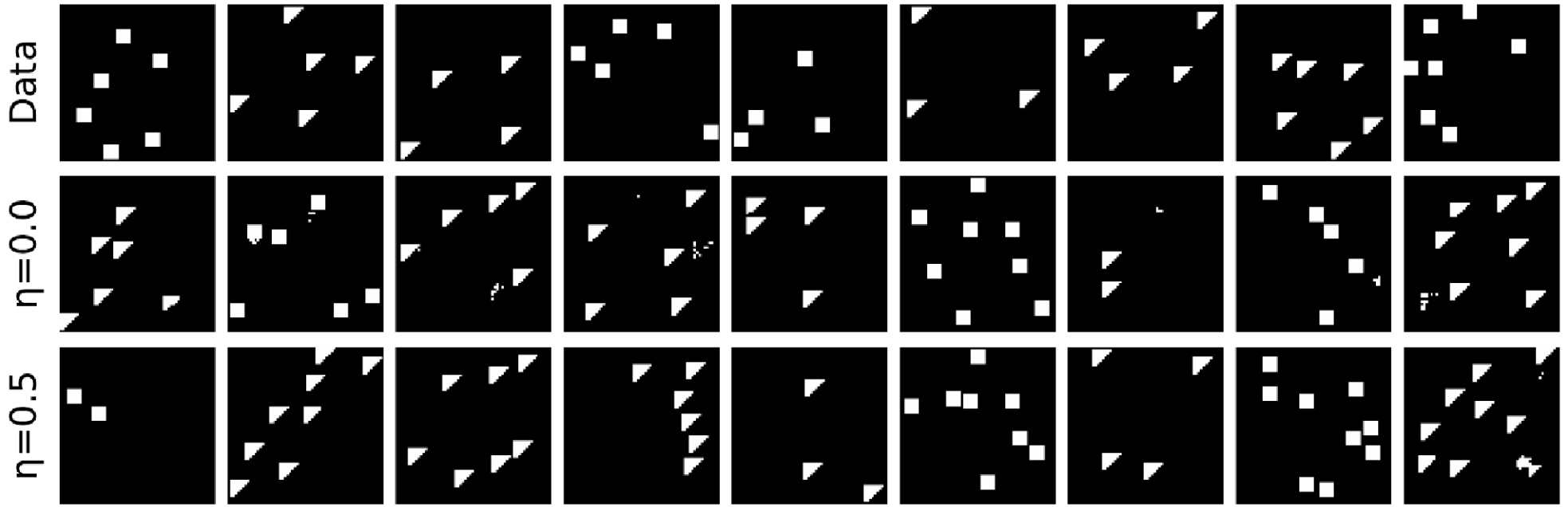
Nearest training image, ℓ_2 ↑

46.02

baseline 42.32 · less memorization

Experiments

Fewer spatial artifacts



792 images with scattered clusters, of 5,000
baseline 3,899

1,014 scattered clusters in total
baseline 11,208

Experiments

Fewer finger-count errors

Baseline



SSD · η = 0.8



Six fingers

26 → 8

Four fingers

9 → 14

Total

35 → 22

Matched pairs from the same noise; red marks a finger-count error · counts over 1,000 samples per model, four- and six-finger hands only

Experiments

DiT: larger gains with limited ImageNet data

	Baseline	SSD, η = 0.98
ImageNet-100K subset 200K steps	25.99	15.58
Full ImageNet 180K steps	4.98	4.83

FID ↓ · DiT-XL/2 · 256 × 256 · CFG = 1.5 · multiple loss masks per sample

Experiments

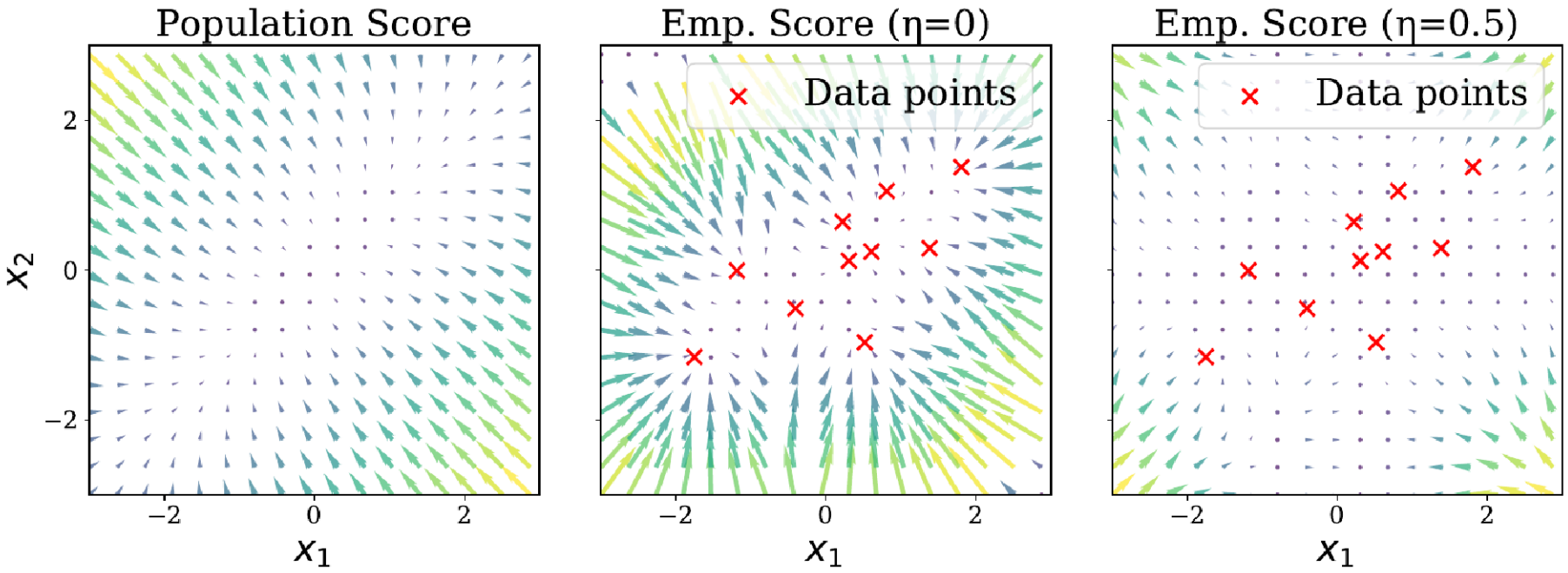
Masking granularity

Pixels or patches: both work at high ratios.

Masking	η	FID ↓
None	0	13.38
Element-wise	0.5	13.29
Element-wise	0.95	6.36
Element-wise	0.98	**5.10**
8 × 8 patch-wise	0.6	7.61
8 × 8 patch-wise	0.9	**5.14**

CelebA-10K · 3,500 epochs · 64 × 64 · ratios differ, so this does not isolate granularity alone

Closer to the population score



2.78

Mean score error, η = 0.5
baseline 3.65

6.18

Maximum score error, η = 0.5
baseline 10.63

Takeaways

Supervise less. Generate consistently.

01 A few lines of code; no architecture or sampler changes

02 Competitive FID, and stable training on small datasets

03 Fewer spatially inconsistent samples



Project page & code