GUIDE 🧭: Guidance-based Incremental Learning with Diffusion Models

Bartosz Cywiński[1] Kamil Deja[1,2] Tomasz Trzciński [1,2,3] Bartłomiej Twardowski[2,4,5] Łukasz Kuciński [2,6,7]

[1]Warsaw University of Technology [2]IDEAS NCBR [3]Tooploox [4]Computer Vision Center [5]Universitat Autònoma de Barcelona [6]University of Warsaw [7]Polish Academy of Sciences

#TLDR

- We introduce **GUIDE** 🧭 - a generative replay method that benefits from classifier guidance to generate **rehearsal data samples prone to be forgotten**.
- Our sampling approach moves replay examples closer to the classifier's decision boundary, making them highly valuable for class-incremental learning.

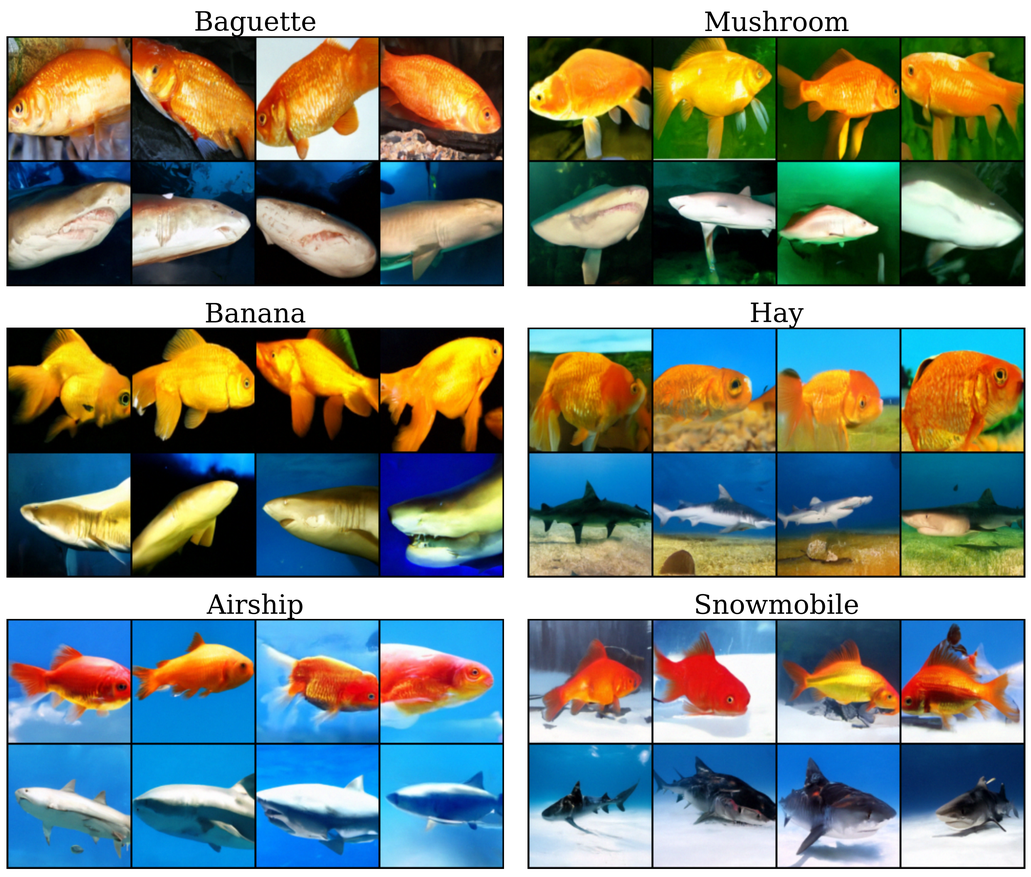
Motivation

Existing generative replay methods randomly sample rehearsal examples from a generative model, which contrasts with buffer-based approaches where **the sampling strategy plays an important role**.

Guidance towards unknown classes

- The combination of guidance signals produces samples that align with the training data distribution of the diffusion model yet exhibit features from classes unknown to the model.
- In GUIDE 🧭, we use this observation in the continual training of a classifier.

(a) Samples generated **without guidance** towards any unknown class.

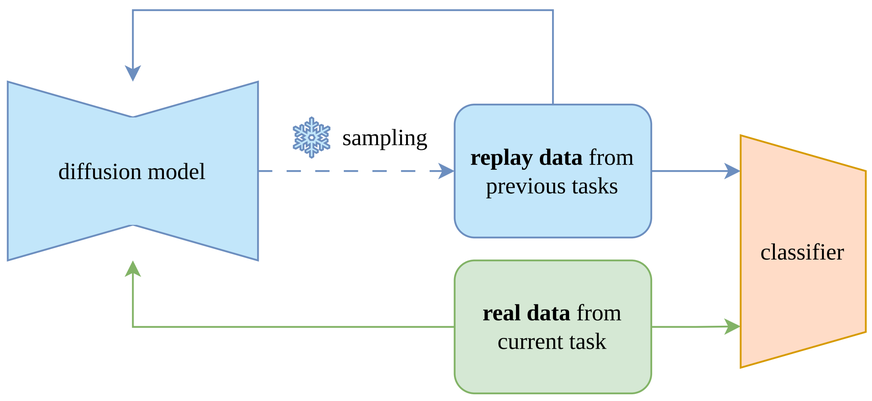


(b) Samples generated **with guidance towards class unknown to diffusion model** depicted above each plot.

Figure 1. Samples from our unconditional diffusion model trained only on *goldfish* and *tiger shark* classes.

Background – generative replay

In each task i, we train a classifier $f_{\phi_i}(y|\mathbf{x})$ and a class-conditional diffusion model $\epsilon_{\theta_i}(\mathbf{x}_t, t, y)$ on currently available data along with synthetic data samples from preceding tasks generated by the previous frozen diffusion model $\epsilon_{\theta_{i-1}}(\mathbf{x}_t, t, y)$.



Results – average accuracy after the final task

Method	CIFAR-10 $T=2$	CIFAR-10 $T=5$	CIFAR-100 $T=5$	CIFAR-100 $T=10$	ImageNet100 $T=5$
Joint	93.14 ± 0.16		72.32 ± 0.24		66.85 ± 2.25
Continual Joint	85.63 ± 0.39	86.41 ± 0.32	73.07 ± 0.01	64.15 ± 0.98	50.59 ± 0.35
Fine-tuning	47.22 ± 0.06	18.95 ± 0.20	16.92 ± 0.03	9.12 ± 0.04	13.49 ± 0.18
DGR VAE	60.24 ± 1.53	28.23 ± 3.84	19.66 ± 0.27	10.04 ± 0.17	9.54 ± 0.26
DGR+distill	52.40 ± 2.58	27.83 ± 1.20	21.38 ± 0.61	13.94 ± 0.13	11.77 ± 0.47
RTF	51.80 ± 2.56	30.36 ± 1.40	17.45 ± 0.28	12.80 ± 0.78	8.03 ± 0.05
MeRGAN	50.54 ± 0.08	51.65 ± 0.40	9.65 ± 0.14	12.34 ± 0.15	-
BIR	53.97 ± 0.97	36.41 ± 0.82	21.75 ± 0.08	15.26 ± 0.49	8.63 ± 0.19
GFR	64.13 ± 0.88	26.70 ± 1.90	34.80 ± 0.26	21.90 ± 0.14	32.95 ± 0.35
DDGR	80.03 ± 0.65	43.69 ± 2.60	28.11 ± 2.58	15.99 ± 1.08	25.59 ± 2.29
DGR diffusion	77.43 ± 0.60	59.00 ± 0.57	28.25 ± 0.22	15.90 ± 1.01	23.92 ± 0.92
GUIDE 🧭	**81.29** ± 0.75	**64.47** ± 0.45	**41.66** ± 0.40	**26.13** ± 0.29	**39.07** ± 1.37

Method

Modified diffusion denoising process during generation of a sample from class y_{i-1}:

$$\hat{\epsilon}_{\theta_{i-1}}(\mathbf{x}_t, t, y_{i-1}) = \epsilon_{\theta_{i-1}}(\mathbf{x}_t, t, y_{i-1}) + \underbrace{s\nabla_{\mathbf{x}_t}\ell\left(f_{\phi_i}(y|\hat{\mathbf{z}}_0(\mathbf{x}_t), y_i\right)}_{\text{guidance towards class from current task } i}$$

- Generated rehearsal samples are positioned near the classifier's decision boundary and are more likely to be forgotten.
- Since we use the previous frozen diffusion model, which is **not trained on the classes from the current task**, we consistently obtain samples from the desired class y_{i-1}.

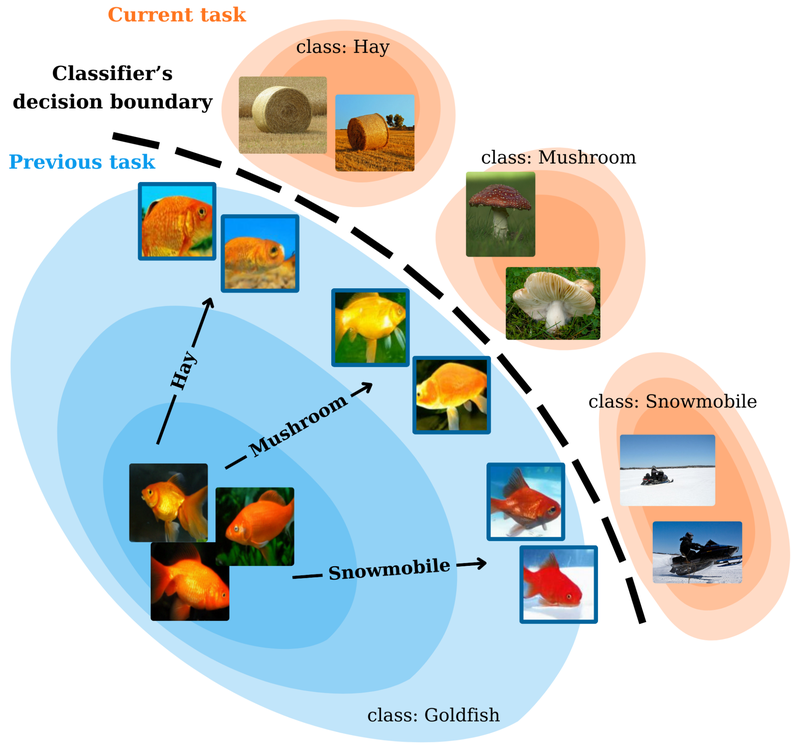
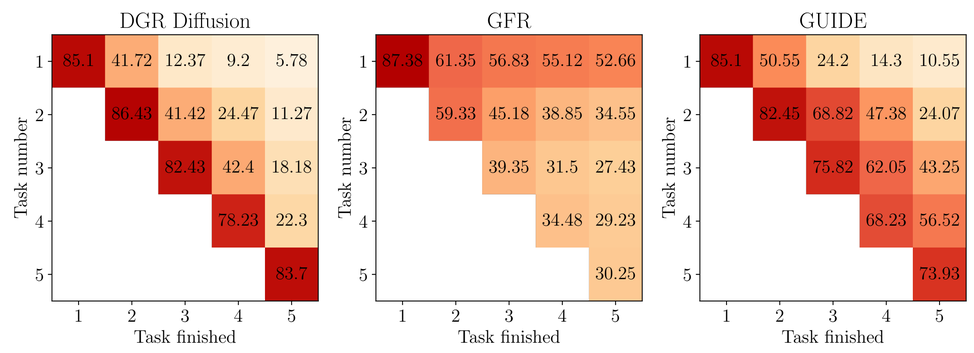


Figure 2. Rehearsal sampling in GUIDE 🧭. The replay samples, highlighted with blue borders, share features with the examples from the current task, which may be related to characteristics such as color or background.

GUIDE 🧭 mitigates forgetting in a classifier



Where are the rehearsal samples located?

We launch an adversarial attack on the generated rehearsal samples to check how easy it is to change their class to the one from the current task.

	Misclassified examples	Prediction confidence	
		Previous classifier	Current classifier
DGR diffusion	55.13%	99.60%	90.03%
GUIDE 🧭	72.66%	86.42%	61.61%

Properties of rehearsal samples in GUIDE 🧭**:**

- It is easy to fool a classifier with them
- Yield lower outputs in the currently trained classifier than in the classifier from the previous task

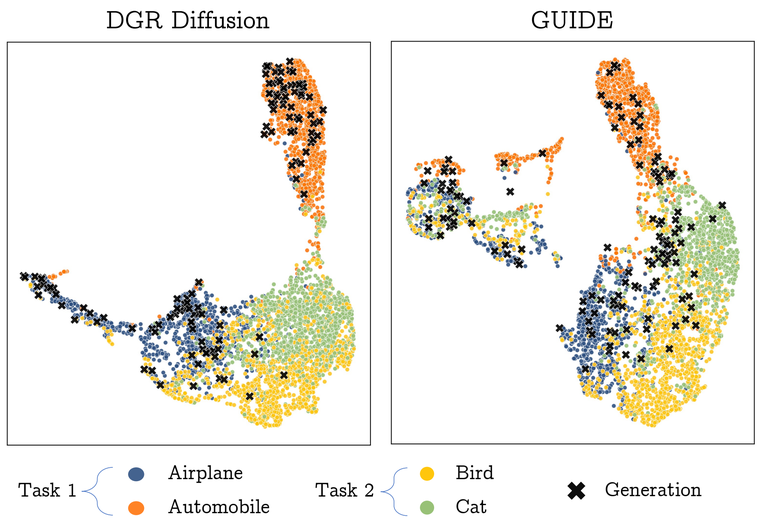


Figure 3. Rehearsal samples in the standard generative replay scenario predominantly originate from high-density regions of class manifolds, while GUIDE 🧭 yields generations that are more similar to the examples from the new task.

Let our paper GUIDE 🧭 you!

contact: bcywinski11@gmail.com