Robust semi-supervised segmentation with timestep ensembling diffusion models

Margherita Rosnati margherita.rosnati12@imperial.ac.uk
Mélanie Roschewitz
Ben Glocker
Biomedical Image Analysis Group (BioMedIA)
Department of Computing
Imperial College London

Abstract

Medical image segmentation is a challenging task, made more difficult by many datasets' limited size and annotations. Denoising diffusion probabilistic models (DDPM) have recently shown promise in modelling the distribution of natural images and were successfully applied to various medical imaging tasks. This work focuses on semi-supervised image segmentation using diffusion models, particularly addressing domain generalisation. Firstly, we demonstrate that smaller diffusion steps generate latent representations that are more robust for downstream tasks than larger steps. Secondly, we use this insight to propose an improved ensembling scheme that leverages information-dense small steps and the regularising effect of larger steps to generate predictions. Our model shows significantly better performance in domain-shifted settings while retaining competitive performance in-domain. Overall, this work highlights the potential of DDPMs for semi-supervised medical image segmentation and provides insights into optimising their performance under domain shift.

Keywords: Medial Image Segmentation, Semi-Supervised Learning, Generative Modelling

1. Introduction

Denoising diffusion probabilistic models (DDPM) (Sohl-Dickstein et al., 2015; Ho et al., 2020) have recently emerged as a promising approach for modelling the distribution of natural images, outperforming alternative methods in terms of sample realism and diversity. More recently, DDPM have also been successfully applied to various medical imaging tasks, such as synthetic image generation (Kim and Ye, 2022), image reconstruction (Xie and Li, 2022; Peng et al., 2022), anomaly detection (Wolleb et al., 2022; Pinaya et al., 2022), diagnostics (Aviles-Rivero et al., 2022) and segmentation (Wolleb et al., 2022).

Image segmentation is crucial in medical imaging, where accurate and efficient methods are required to support diagnosis, treatment planning, and disease monitoring. However, medical imaging datasets are often limited in size and may lack sufficient annotations, making it challenging to train accurate segmentation models. Moreover, medical imaging data is characterised by high variability, resulting from differences in acquisition parameters, scanner types, and patient demographics. This phenomenon, also known as domain shift, poses a significant challenge to the generalisation of segmentation models applied to new datasets, leading to potential underperformance in clinical settings.

Recent research in diffusion models has shown promising results for semi-supervised learning (Baranchuk et al., 2021; Deja et al., 2023) based on the discovery that the bottleneck network, tasked to learn the backward process of removing noise from an image, also learns an expressive feature representation that can benefit other downstream analysis tasks. Several techniques have been proposed to leverage intermediate diffusion steps for improved in-domain downstream performance. However, more research is needed on the implications of these design choices regarding model generalisation. Our work focuses on the latter problem.

Specifically, we investigate how to optimally leverage diffusion steps to improve generalisation for semi-supervised image segmentation under domain shift.

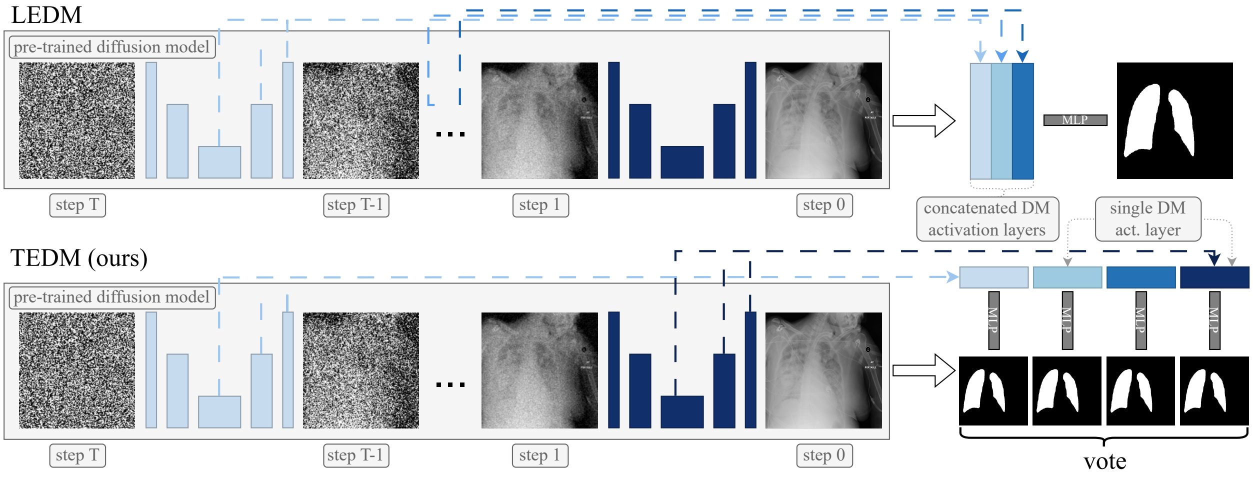


Figure 1: Models diagram. LEDM, the SOTA in semi-supervised segmentation with diffusion models, selects a subset of timesteps and concatenates latent representations extracted from a pretrained diffusion model as features fed to an MLP. Our method (i) selects smaller and more informative timesteps, (ii) predicts through a voting mechanism over our steps selection and (ii) shares the MLP weights across timesteps, resulting in improved segmentation performance.

Based on the analysis of datasets with diverse imaging modalities and domain shifts, our findings demonstrate significant improvements over existing baselines using five different datasets. Our key findings can be summarised as follows:

- Small diffusion steps are crucial for model generalisation;
- Concatenating latent representations over steps to predict segmentation maps can hurt generalisation;
- Instead, generalisation can be significantly improved by (i) optimising which timesteps to use at test time, (ii) ensembling predictions from individual timesteps using a shared predictor and (iii) using these individual predictions for regularisation during training.

2. Background and related work

2.1. Diffusion models

Diffusion models have garnered significant interest in the machine learning community due to their remarkable ability to model complex data distributions efficiently. Diffusion models utilise a series of simple and learnable transformations to diffuse noise iteratively and generate samples from the target distribution. Formally, a DDPM works as follows. Given a data distribution $p(\mathbf{x}_0)$ and forward process:

$$p(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}), \qquad (1)$$

where $\beta_t \in (0,1)$ is the variance schedule and $t \in [0,T]$ is the Markov chain time step, a DDPM aims to learn $\mu_\theta(\mathbf{x}_t, t)$ and $\mathbf{\Sigma}_\theta(\mathbf{x}_t, t)$ which define the backward process:

$$p(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \mathbf{\Sigma}_\theta(\mathbf{x}_t, t)). \qquad (2)$$

In order to do so, Ho et al. (2020) fix the variance $\mathbf{\Sigma}_\theta(\mathbf{x}_t, t)$, reparametrise $\mu_\theta(\mathbf{x}_t, t)$ as a function of the noise $\epsilon_\theta(\mathbf{x}_t, t)$

$$\mu_\theta(\mathbf{x}_t, t) = \frac{1}{\sqrt{\alpha_t}}\big(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\overline{\alpha}_t}}\epsilon_\theta(\mathbf{x}_t, t)\big), \qquad (3)$$

$$\alpha_t = 1-\beta_t, \qquad \overline{\alpha}_t = \prod_{i=1}^{t}\alpha_i \qquad (4)$$

and design a UNet-based (Ronneberger et al., 2015) neural network architecture

$$G_\theta : (\mathbf{x}_t, t) \rightarrow \epsilon_\theta(\mathbf{x}_t, t) \qquad (5)$$

for learning to identify the noise. The UNet is trained through cross-entropy between the injected and predicted noise.

2.2. Diffusion models for label-efficient image segmentation

Baranchuk et al. (2021) apply diffusion models to semi-supervised segmentation by using a diffusion model pretrained on unlabelled images, extracting latent representation from the UNet's intermediate layers and using them to train a pixel-wise classifier. More concretely, their Label Efficient Diffusion Model (LEDM) extracts the latent representations generated with a pretrained UNet diffusion model by selecting a set of steps $t \in S \subset \{0, \dots, T\}$, passing a noisy input

$$\mathbf{x}_t = \sqrt{\overline{\alpha}_t} x_0 + \sqrt{1 - \overline{\alpha}_t}\epsilon, \qquad \epsilon \sim \mathcal{N}(0, \mathbf{I}) \tag{6}$$

through the UNet. The resulting activation maps $\mathbf{z}_t \in \mathrm{R}^{c \times h \times w}$ are then upsampled through bilinear interpolation to the input size and concatenated into a feature map $\mathbf{Z} \in \mathrm{R}^{(|S| \times c) \times H \times W}$. Finally, each pointwise prediction is performed independently by an ensemble of lightweight multilayer perceptions

$$C^n_\phi : \mathbf{Z}^{i,j} \rightarrow y^{i,j}; \qquad n \in \{1, .., 10\} \tag{7}$$

trained with a cross-entropy loss. The authors concatenate the diffusion steps $S = \{50, 150, 250\}$ to form the input to these predictors.

Similarly, Deja et al. (2023) also use the latent representations of a pretrained diffusion model for classification tasks. In particular, they propose to use classifier predictions from all intermediate timesteps to regularise the training of the diffusion model. However, at test time, they only use the last diffusion step $t = 1$ to generate predictions.

3. On the importance of the diffusion steps for domain generalisation

Previous findings suggest that latent representations in larger steps contain coarse information, which becomes more granular as the diffusion steps approach the target data distribution (Baranchuk et al., 2021; Deja et al., 2023). Here, we are interested in understanding how the wealth of information in each time step $s \in S$ contributes to model generalisation when the training dataset size varies.

We train a Ridge logistic regression-based pixel-wise classifier over latent representations extracted from specific timesteps $t = \{1,\ 10,\ 25,\ 50,\ 200,\ 400,\ 600 \text{ and } 800\}$ to isolate the predictive power of each timestep. We compare these timestep-wise predictions to LEDM and a fully supervised baseline using the same UNet backbone as the DDPM backbone.

We evaluate our work on the task of chest X-ray lung segmentation. Chest X-rays are among the most frequent radiological examinations in clinical practice, and automatically extracted features from anatomical regions such as the lungs can aid clinical decision-making. Moreover, the availability of several public datasets of chest X-ray images allows us to investigate the methods' generalisation ability in the presence of changes in dataset characteristics.

Following previous work in semi-supervised medical image segmentation (Rosnati et al., 2022), we use the ChestX-ray8 (Wang et al., 2017) (n=108k) as the unlabelled dataset to train the DDPM backbone over $T = 1000$ steps and a subset of the JSRT (Van Ginneken et al., 2006) (n=247) labelled dataset for training (n=197) and validating (n=25) our method. The dataset splits, architecture, and code are available in our code repository[1].

We reserve the remaining JSRT samples (n=25) along with the NIH (Tang et al., 2019) (n=95), and Montgomery (Jaeger et al., 2014) (n=138) labelled datasets for final testing. Notably, the NIH dataset is an annotated subset of the ChestX-ray8 dataset. This setup allows us to test the models on data that is (i) in-domain for the classifier (JSRT), (ii) out-of-domain for the classifier but in-domain for the DDPM (ChesX-ray8/NIH) and (iii) out-of-domain for both (Montgomery).

Figure 2 shows the Dice coefficients[2] from the step-wise experiment when training our segmentation model, the baseline and LEDM on $n = \{197,\ 49,\ 24,\ 12,\ 6,\ 3 \text{ and } 1\}$ JSRT labelled datapoints, corresponding to $\{100,\ 50,\ 25,\ 12,\ 6,\ 3,\ 2 \text{ and } 1\}$ % of the training dataset. Surprisingly, LEDM does not significantly[3] outperform the baseline in the one-shot setting for domain-shifted datasets (NIH, Montgomery). This indicates that LEDM may not fully utilise the latent representation information. Secondly, we find that

[1] Demo: https://huggingface.co/spaces/anonymous2023-21/TEDM-demo

[2] Dice $= 2\frac{|A \cup B|}{|A|+|B|}$

[3] Significance is calculated through a Wilcoxon paired test at level 0.05.

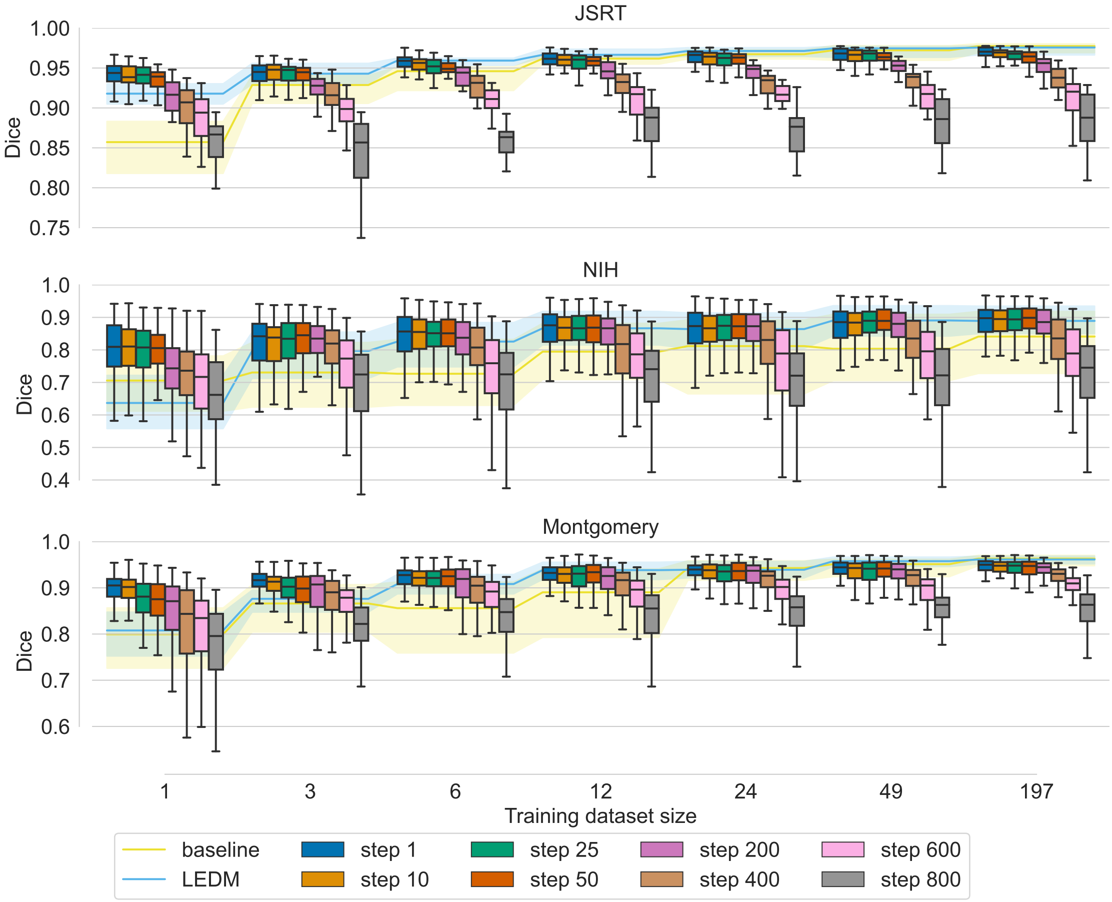


Figure 2: Performance of a logistic regression segmentation model trained on latent features from individual diffusion steps.

the predictor trained on a single step $t = 1$ statistically outperforms both LEDM and the baseline for small training sizes (1, 3, 6 in NIH and Montgomery and for one datapoint in JSRT). In addition, this predictor remains competitive with both the baseline and LEDM across all other training dataset sizes.

The experiment highlights that latent representations obtained from smaller steps are more powerful predictors than those obtained from larger steps, particularly for domain generalisation. In particular, the LEDM steps 50, 125 and 250 are not the optimal choice for segmentation as single-step approaches with smaller steps perform better on out-of-distribution datasets. In the next section, we investigate whether ensembling different steps can still outperform single-step approaches given the right choice of steps. We investigate several ways of ensembling these steps and their impact on model generalisation.

4. Timestep ensembling diffusion models

In this section, we show that the generalisation of diffusion-based segmentation models in the low data regime can be significantly improved by judiciously combining adequate timesteps both at prediction and training time.

We hypothesise that the lack of generalisation of LEDM observed in the previous section can be mitigated with more model regularisation and reducing the number of parameters that need to be learned.

Table 1: Models performance w.r.t. ground truth segmentations. Reported as mean ± standard deviation over the dataset. Global CL, Global & Local CL and LEDM are a reproduction of Chen et al. (2020), Chaitanya et al. (2020) and Baranchuk et al. (2021) respectively. All statistically comparably best performing models are highlighted in bold. Significance is calculated through a Wilcoxon paired test at level 0.05.

Training size	1 (1%)	3 (2%)	6 (3%)	12 (6%)	197 (100%)
	JSRT (in-domain for classifier)				
Sup. Baseline	84.4 ± 5.4	91.7 ± 3.7	93.3 ± 2.9	95.3 ± 2.3	97.3 ± 1.2
Global CL	88.8 ± 5.9	92.7 ± 1.8	93.6 ± 1.6	95.3 ± 1.1	97.1 ± 1.4
Global & Local CL	89.8 ± 5.2	93.1 ± 1.7	92.9 ± 1.9	94.8 ± 1.49	97.2 ± 1.2
LEDM	90.8 ± 3.5	94.1 ± 1.6	95.5 ± 1.4	96.4 ± 1.4	97.0 ± 1.3
LEDMe	**93.7 ± 2.6**	**95.5 ± 1.5**	**96.7 ± 1.5**	**97.0 ± 1.1**	**97.6 ± 1.2**
TEDM (ours)	**93.1 ± 3.4**	94.8 ± 1.4	95.8 ± 1.2	96.6 ± 1.1	97.3 ± 1.2
	NIH (in-domain for DDPM, OOD for classifier)				
Sup. Baseline	68.5 ± 12.8	71.2 ± 15.1	71.4 ± 15.9	77.8 ± 14.0	81.5 ± 12.7
Global CL	70.7 ± 14.6	80.3 ± 12.2	77.1 ± 16.4	84.6 ± 10.8	86.9 ± 10.8
Global & Local CL	71.1 ± 16.2	79.6 ± 12.7	81.1 ± 14.0	82.2 ± 13.6	87.4 ± 10.8
LEDM	63.3 ± 12.2	78.0 ± 10.1	81.2 ± 9.3	85.9 ± 7.4	88.9 ± 5.9
LEDMe	70.3 ± 11.4	78.3 ± 9.8	83.0 ± 8.6	84.4 ± 8.1	90.1 ± 5.3
TEDM (ours)	**80.3 ± 9.0**	**86.4 ± 6.2**	**89.2 ± 5.5**	**91.3 ± 4.1**	**92.9 ± 3.2**
	Montgomery (OOD for DDPM and classifier)				
Sup. Baseline	77.1 ± 12.0	83.0 ± 12.2	80.9 ± 14.7	83.8 ± 14.9	94.1 ± 6.6
Global CL	76.1 ± 15.0	87.6 ± 9.7	88.8 ± 11.4	90.4 ± 10.4	92.9 ± 10.8
Global & Local CL	77.4 ± 17.4	88.7 ± 9.14	89.9 ± 8.2	90.1 ± 10.9	92.5 ± 11.2
LEDM	79.3 ± 8.1	85.9 ± 7.4	89.4 ± 6.7	92.3 ± 7.2	94.4 ± 7.2
LEDMe	80.7 ± 6.6	86.3 ± 6.5	89.5 ± 5.9	91.2 ± 5.6	**95.3 ± 4.0**
TEDM (ours)	**90.5 ± 5.3**	**91.4 ± 6.1**	**93.3 ± 6.0**	**94.6 ± 6.0**	95.1 ± 6.9

Indeed, the current approach of concatenating features from numerous timesteps to feed into the pixel-wise MLP predictor results in an excessively high-dimensional input, which leads to a complex predictor. To address this concern, we propose using a shared MLP trained to generate a prediction map from each latent representation of the steps considered.

We define our loss function as follows:

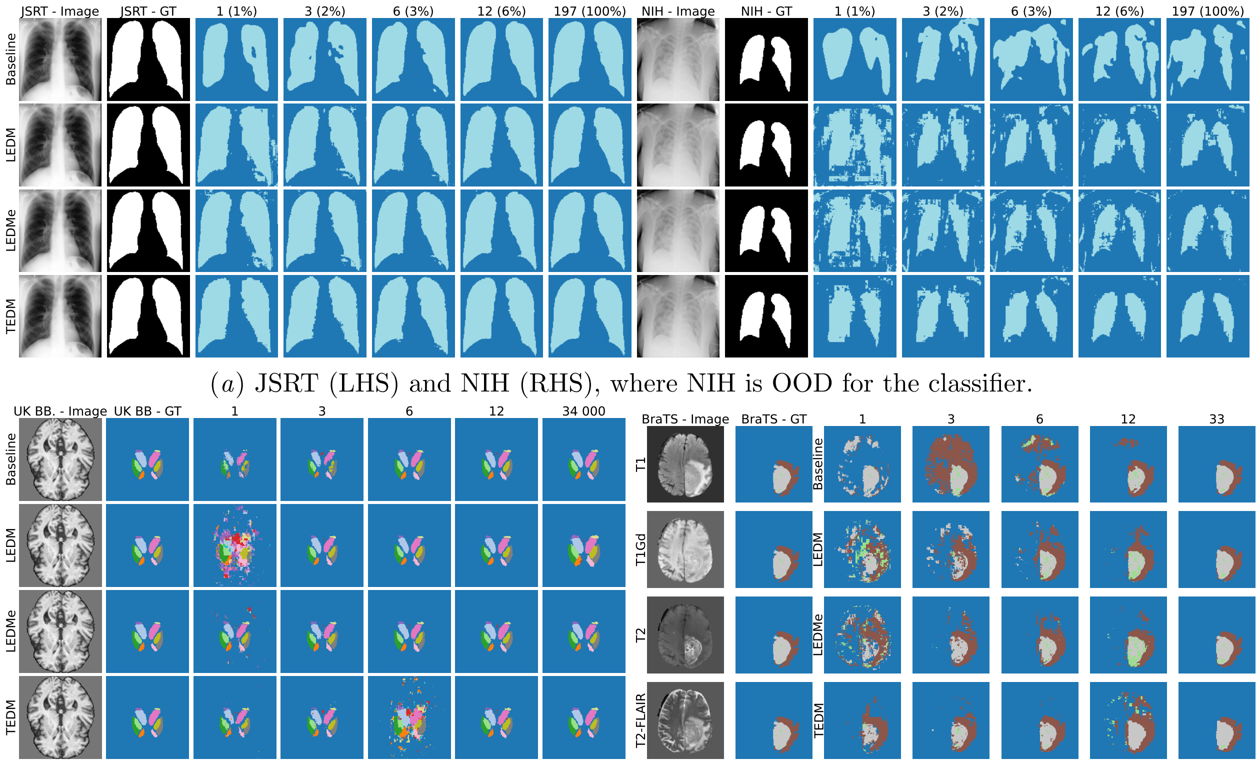
$$\phi = \text{argmin}\ \mathbb{E}_{\mathcal{D}}\mathbb{E}_{i,j}\mathbb{E}_{s\in S}\ \text{CE}\ (C_\phi(\tilde{\mathbf{z}}_s^{i,j}), y^{i,j}), \tag{8}$$

where i, j is the pixel indexing, $y^{i,j}$ is the ground truth class of pixel i, j, $\tilde{\mathbf{z}}_s$ is the upsampled latent representation $\mathbf{z}_s$ of the diffusion model at step s, S is the set of diffusion steps used, C_ϕ is the pixel-wise MLP predictor, CE stands for cross entropy and $\mathcal{D}$ is the training set. At test time, we use a voting mechanism to ensemble the various prediction maps to obtain a final segmentation map. We call this technique "timestep ensembling" and show that it yields superior performance.

$$\hat{y}_{i,j} = \frac{1}{|S|}\sum_{t\in S} C_\phi(\tilde{\mathbf{z}}_s^{i,j}) \tag{9}$$

Moreover, we leverage the insights from the previous section and combine predictions from the diffusion steps $S = \{1, 10, 25, 50, 200, 400, 600 \text{ and } 800\}$. This approach allows us to benefit from the small steps information content and larger step regularisation effect, unlike LEDM, which only used timesteps {50, 125 and 250. To better understand the distinctions between our model and LEDM, please refer to Figure 1. A discussion on computational complexity can be found in Appendix Section B.

Figure 3: Segmentation examples. Col. 1 and 2 are the image and ground truth segmentation. Subsequent columns correspond to models trained with n training datapoints (see title). Row 1 corresponds to the baseline outcomes, and row 2, 3 and 4 to LEDM, LEDMe and TEDM (our method) respectively.



(a) JSRT (LHS) and NIH (RHS), where NIH is OOD for the classifier.

(b) UK Biobank

(c) BraTS

5. Experiments

We conduct experiments on various percentages of the JSRT training dataset, 12%, 6%, 3%, 2%, and 1%, to fully explore the potential of our semi-supervised method. In addition, we train on 100% of the training set for completeness. To evaluate the performance of our timestep ensembling diffusion model (TEDM), we compare it with the fully supervised baseline (described in Section 3) and LEDM. LEDM and TEDM have the same MLP classifier architecture. In addition, we compare TEDM to two other semi-supervised methods that use contrastive learning (CL): the 'Global CL' (Chen et al., 2020) and the 'Local and Global CL' (Chaitanya et al., 2020). Both these methods are trained with the same backbone architecture as the baseline and the DDPM.

In order to investigate the effect of each component in our TEDM model, we carry out several ablations. Firstly, we compare the original LEDM model with another instance of LEDM, trained with our diffusion steps, which we refer to as LEDMe. This allows us to ablate the effect of our diffusion steps choice. Secondly, we test the voting mechanism by reporting model performance when only steps 1, 10 or 25 are used at test time. We use the same evaluation procedure as in Section 3.

Finally, to test the TEDM method's generalizability, we apply it to two additional datasets: the UK Biobank dataset and the BraTS dataset (Menze et al., 2014; Bakas et al., 2017, 2018). In the UK Biobank dataset, we segment brain structures in 2D slices of brain MRI T1 images. This dataset is particularly challenging due to the low intensity variation between structures and background. The BraTS dataset com-

Table 2: Ablation study on test-time ensembling over timesteps. Each 'Step *i*' experiment only uses predictions from timestep *i* at test time. All statistically comparably best performing models are highlighted in bold. Significance is calculated through a Wilcoxon paired test at level 0.05.

Training size	1 (1%)	3 (2%)	6 (3%)	12 (6%)	197 (100%)
	JSRT (in-domain for classifier)				
Step 1	91.1 ± 5.0	**94.5 ± 2.1**	**96.0 ± 1.4**	**96.8 ± 1.1**	**97.4 ± 1.3**
Step 10	91.6 ± 4.6	**94.6 ± 1.8**	**96.0 ± 1.3**	**96.9 ± 1.0**	**97.4 ± 1.2**
Step 25	91.7 ± 4.2	**94.5 ± 1.6**	95.8 ± 1.2	96.8 ± 1.0	97.3 ± 1.2
TEDM	**93.1 ± 3.4**	**94.8 ± 1.4**	95.8 ± 1.2	96.6 ± 1.1	**97.3 ± 1.2**
	NIH (in-domain for DDPM, OOD for classifier)				
Step 1	70.4 ± 10.9	78.9 ± 9.4	84.2 ± 8.3	87.5 ± 6.5	91.9 ± 3.3
Step 10	73.2 ± 10.3	81.1 ± 8.3	85.8 ± 7.3	88.8 ± 5.6	91.8 ± 3.3
Step 25	75.1 ± 9.8	82.6 ± 7.7	86.5 ± 6.7	89.4 ± 5.2	91.9 ± 3.3
TEDM	**80.3 ± 9.0**	**86.4 ± 6.2**	**89.2 ± 5.5**	**91.3 ± 4.1**	**92.9 ± 3.2**
	Montgomery (OOD for DDPM and classifier)				
Step 1	85.9 ± 4.0	89.3 ± 4.2	92.2 ± 4.2	93.9 ± 3.9	94.9 ± 5.3
Step 10	87.1 ± 4.5	89.3 ± 4.8	92.1 ± 5.2	94.1 ± 5.0	94.8 ± 6.5
Step 25	87.4 ± 5.3	89.1 ± 5.5	91.7 ± 6.2	93.7 ± 6.3	94.6 ± 7.0
TEDM	**90.5 ± 5.3**	**91.4 ± 6.1**	**93.3 ± 6.0**	**94.6 ± 6.0**	**95.1 ± 6.9**

prises brain MRI (T1, T1Gd, T2 and T2-FLAIR) of patients with brain tumours, which we decompose into 2D slices and segment. This dataset is even more difficult as it entails segmenting items of varied shapes and locations. Further details on the experimental process for these two datasets are available in Appendix A.

6. Results

The performance results on chest X-rays and brain MRI are shown quantitatively in Tables 1 and 3, and qualitatively in Figure 3. The ablation results are shown in Table 2. Further results can be found in Appendix C. For all tables, the best-performing model and all statistically equivalent models are highlighted by reporting their results in bold.

Using small step sizes improves performance both in- and out-of-domain.

In Table 1, we observe that in all cases, selecting small diffusion steps generates the best-performing models: LEDMe outperforms LEDM statistically significantly for all experiments but two (Montgomery $n = 12$ and NIH $n = 12$). In addition, LEDMe outperforms LEDM for the UK Biobank and BraTS datasets for training sizes larger than 3 and 1, respectively (see Table 3).

Concatenating latent representations hurts generalisability in the low data regime.

TEDM outperforms LEDMe (and LEDM) for the NIH and Montgomery datasets, except for n=197. We deduce that the concatenation method exploited in LEDM leads to poor generalisation on domains outside the labelled training set. In addition, TEDM performs statistically comparably to LEDM for JSRT, indicating that its generalisation properties come with little to no in-domain performance cost.

Test-time ensembling over timesteps improves generalisation over single-step predictions.

Table 2 shows that using a voting mechanism for prediction (used in TEDM) is more effective than using the smallest step (TEDM outperforms the competing models in OOD cases), implying that different steps

Table 3: Dice scores on the UK Biobank and BraTS datasets. For both datasets, the model was trained on 2D slices, the results are reported on the 3D images. The training size refers to the number of patients in the labelled training set. The number of 2D slices is roughly 100x larger. All statistically comparably best performing models are highlighted in bold. Significance is calculated through a Wilcoxon paired test at level 0.05 with Bonferroni correction to account for multiple classes per patient.

UK Biobank ($n_{train}^{unlabelled} = 34\,000$, $n_{test} = 500$)

Training size	1	3	6	12	34 000
Sup. Baseline	54.6 ± 18.6	76.8 ± 12.3	**83.1 ± 8.5**	**85.1 ± 7.6**	**89.6 ± 5.2**
Global CL	42.7 ± 20.4	77.3 ± 11.0	82.0 ± 8.7	85.2 ± 7.4	88.7 ± 5.6
Global & Local CL	44.3 ± 20.3	74.0 ± 11.8	80.6 ± 9.4	82.0 ± 8.9	87.4 ± 6.8
LEDM	60.8 ± 17.1	**81.3 ± 7.9**	82.3 ± 8.9	83.0 ± 9.2	87.7 ± 5.8
LEDMe	54.7 ± 17.8	79.4 ± 10.8	82.5 ± 9.1	83.8 ± 8.6	86.6 ± 7.0
TEDM (ours)	**71.0 ± 14.8**	**81.0 ± 9.0**	**82.8 ± 8.8**	83.2 ± 9.3	85.1 ± 7.4

BraTS ($n_{train}^{unlabelled} = 268$, $n_{test} = 33$)

Training size	1	3	6	12	33
Sup. Baseline	12.5 ± 18.9	**30.9 ± 31.2**	**40.7 ± 33.1**	**47.1 ± 33.8**	**69.5 ± 25.7**
Global CL	4.7 ± 13.6	25.5 ± 29.4	32.3 ± 32.1	40.5 ± 32.0	56.9 ± 28.6
Global & Local CL	11.7 ± 19.1	27.3 ± 30.5	34.1 ± 31.5	38.3 ± 32.2	55.4 ± 30.0
LEDM	24.0 ± 22.9	31.0 ± 31.4	40.8 ± 31.9	**48.0 ± 31.2**	62.6 ± 26.7
LEDMe	21.2 ± 22.7	33.1 ± 31.4	**42.8 ± 32.7**	**49.5 ± 31.7**	63.2 ± 27.6
TEDM (ours)	**27.3 ± 26.1**	**35.6 ± 31.7**	**41.9 ± 32.3**	**47.5 ± 31.7**	59.8 ± 29.0

produce latent representations focusing on slightly different aspects of the image.

TEDM performs robustly for increasingly challenging segmentation tasks.

Table 3 shows that TEDM is statistically superior or equal to its competitors for all cases with less than 12 datapoints, showing that our method remains competitive in more challenging in-domain low labelled data scenarios.

Fully supervised baselines are competitive for in-domain harder segmentation tasks.

Our method TEDM showcases excellent performance on very small dataset sizes (1, 2, 3 and 6 in Table 3). However, for larger datasets (6 patients or more), a well-designed baseline model proves to be more effective than any of the semi-supervised models. This result suggests that although semi-supervised methods with self-supervised pretraining may have their limitations in providing task-specific performance for larger datasets, they present great potential for improving results on small datasets.

7. Conclusions

This study investigated the impact of different diffusion steps on the performance and generalisation of semi-supervised segmentation models. Our comprehensive experiments across multiple datasets revealed that small diffusion steps are crucial for domain generalisation, requiring only a few training samples to become powerful pixel-wise predictors. Furthermore, we found that ensembling segmentation maps over timesteps significantly improves model generalisation in the low data regime while offering competitive performance in-domain. Conversely, concatenating latent representations can hurt the generalisation of the pixel-wise classifier. These findings were demonstrated by the superior performance of our proposed Timestep Ensembling Diffusion Model on chest X-ray lung segmentation and more challenging tasks such as brain structure and tumour segmentation. Our results indicate that latent representations across different steps share semantics and act as a model regulariser, leading to better generalisation than competing methods. This analysis underscores the importance of thoroughly investigating the design decisions for auxiliary tasks in diffusion models, such as timestep selection and ensembling. These decisions can have a significant impact on the model's performance.

Our findings provide important new insights and may inform the development of new approaches leveraging powerful diffusion models for medical imaging tasks. In future work, the performance of TEDM and similar approaches should be compared to the emerging foundation model techniques, where the pre-training is executed at a larger scale than semi-supervised methods. Here, the ability of diffusion models to efficiently capture the data distribution from extensive, unlabelled data holds a promise to overcome the persistent data scarcity problem in medical image segmentation.

Acknowledgments

This research has been conducted using data from the UK Biobank. M. Rosnati is supported by UK Research and Innovation [UKRI Centre for Doctoral Training in AI for Healthcare grant number EP/S023283/1]. M. Roschewitz is supported by an Imperial College President's PhD scholarship.

References

Fidel Alfaro-Almagro, Mark Jenkinson, Neal K Bangerter, Jesper LR Andersson, Ludovica Griffanti, Gwenaëlle Douaud, Stamatios N Sotiropoulos, Saad Jbabdi, Moises Hernandez-Fernandez, Emmanuel Vallee, et al. Image processing and quality control for the first 10,000 brain imaging datasets from uk biobank. *Neuroimage*, 166:400–424, 2018.

Angelica I Aviles-Rivero, Christina Runkel, Nicolas Papadakis, Zoe Kourtzi, and Carola-Bibiane Schönlieb. Multi-modal hypergraph diffusion network with dual prior for alzheimer classification. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part III*, pages 717–727. Springer, 2022.

Spyridon Bakas, Hamed Akbari, Aristeidis Sotiras, Michel Bilello, Martin Rozycki, Justin S Kirby, John B Freymann, Keyvan Farahani, and Christos Davatzikos. Advancing the cancer genome atlas glioma mri collections with expert segmentation labels and radiomic features. *Scientific data*, 4(1): 1–13, 2017.

Spyridon Bakas, Mauricio Reyes, Andras Jakab, Stefan Bauer, Markus Rempfler, Alessandro Crimi, Russell Takeshi Shinohara, Christoph Berger, Sung Min Ha, Martin Rozycki, et al. Identifying the best machine learning algorithms for brain tumor segmentation, progression assessment, and overall survival prediction in the brats challenge. *arXiv preprint arXiv:1811.02629*, 2018.

Dmitry Baranchuk, Andrey Voynov, Ivan Rubachev, Valentin Khrulkov, and Artem Babenko. Label-efficient semantic segmentation with diffusion models. In *International Conference on Learning Representations*, 2021.

Krishna Chaitanya, Ertunc Erdil, Neerav Karani, and Ender Konukoglu. Contrastive learning of global and local features for medical image segmentation with limited annotations. *Advances in Neural Information Processing Systems*, 33:12546–12558, 2020.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

Kamil Deja, Tomasz Trzcinski, and Jakub M Tomczak. Learning data representations with joint diffusion models. *arXiv preprint arXiv:2301.13622*, 2023.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Stefan Jaeger, Sema Candemir, Sameer Antani, Yì-Xiáng J Wáng, Pu-Xuan Lu, and George Thoma. Two public chest X-ray datasets for computer-aided screening of pulmonary diseases. *Quantitative imaging in medicine and surgery*, 4(6):475, 2014.

Boah Kim and Jong Chul Ye. Diffusion deformable model for 4d temporal medical image generation. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part I*, pages 539–548. Springer, 2022.

Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (brats). *IEEE transactions on medical imaging*, 34 (10):1993–2024, 2014.

Brian Patenaude, Stephen M Smith, David N Kennedy, and Mark Jenkinson. A bayesian model of shape and appearance for subcortical brain segmentation. *Neuroimage*, 56(3):907–922, 2011.

Cheng Peng, Pengfei Guo, S Kevin Zhou, Vishal M Patel, and Rama Chellappa. Towards performant and reliable undersampled mr reconstruction via diffusion model sampling. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part VI*, pages 623–633. Springer, 2022.

Walter HL Pinaya, Mark S Graham, Robert Gray, Pedro F Da Costa, Petru-Daniel Tudosiu, Paul Wright, Yee H Mah, Andrew D MacKinnon, James T Teo, Rolf Jager, et al. Fast unsupervised brain anomaly detection and segmentation with diffusion models. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part VIII*, pages 705–714. Springer, 2022.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, pages 234–241. Springer, 2015.

Margherita Rosnati, Fabio De Sousa Ribeiro, Miguel Monteiro, Daniel Coelho de Castro, and Ben Glocker. Analysing the effectiveness of a generative model for semi-supervised medical image segmentation. In *Machine Learning for Health*, pages 290–310. PMLR, 2022.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015.

You-Bao Tang, Yu-Xing Tang, Jing Xiao, and Ronald M Summers. Xlsor: A robust and accurate lung segmentor on chest X-rays using criss-cross attention and customized radiorealistic abnormalities generation. In *International Conference on Medical Imaging with Deep Learning*, pages 457–467. PMLR, 2019.

Bram Van Ginneken, Mikkel B Stegmann, and Marco Loog. Segmentation of anatomical structures in chest radiographs using supervised methods: a comparative study on a public database. *Medical image analysis*, 10(1):19–40, 2006.

Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. ChestX-ray8: Hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2097–2106, 2017.

Julia Wolleb, Florentin Bieder, Robin Sandkühler, and Philippe C Cattin. Diffusion models for medical anomaly detection. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part VIII*, pages 35–45. Springer, 2022.

Yutong Xie and Quanzheng Li. Measurement-conditioned denoising diffusion probabilistic model for under-sampled medical image reconstruction. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part VI*, pages 655–664. Springer, 2022.

Table 4: Methods computational cost

	Theoretical test-time operations	GMAC
Sup. Baseline	N	29.2
Global CL	N	29.2
Global & Local CL	N	29.2
LEDM	$\lvert S_{\text{LEDM}}\rvert \times N + n_{\text{pixels}} \times N_{\text{MLP}}(\lvert S_{\text{LEDM}}\rvert \times n_{\text{latent}}, 1)$	88.0
LEDMe	$\lvert S_{\text{TEDM}}\rvert \times N + n_{\text{pixels}} \times N_{\text{MLP}}(\lvert S_{\text{TEDM}}\rvert \times n_{\text{latent}}, 1)$	234.6
TEDM (ours)	$\lvert S_{\text{TEDM}}\rvert \times N + \lvert S_{\text{TEDM}}\rvert \times n_{\text{pixels}} \times N_{\text{MLP}}(n_{\text{latent}}, 1)$	234.6

Appendix A. Methods details

A.1. UK Biobank data preprocessing

The UK Biobank brains dataset contains 42 791 patients' scans. We initially separate the data in three sets, a training set with $n_{train} = 34\,230$, a validation set with $n_{val} = 4280$ and a test set of $n_{test} = 4280$ patients. After evaluating some methods with $n_{test} = 4280$ and careful consideration of results variance, we reduced the test set to $n_{test} = 500$ without suffering any drops in metrics accuracy.

All scans have voxel size $1mm^3$ and image size $189 \times 233 \times 197$, and are paired with the segmentation of 15 subcortical structures' volumes from FIRST (FMRIB's Integrated Registration and Segmentation Tool Patenaude et al. (2011)) segmentation, and brain masks. For more details on the scan preprocessing, please refer to Alfaro-Almagro et al. (2018).

We preprocess the images by clipping the intensities to [0, 1500] to remove large outliers, then normalise the brain pixels using the brain masks so that the 1^{st} and 99^{th} quantiles correspond to -1 and 1 respectively:

$$x_{norm}[mask \neq 0] = a \cdot x[mask \neq 0] + b \tag{10}$$

$$\text{such that } a = \frac{2}{x^{99\%} - x^{1\%}} \text{ and } b = 1 - a \cdot x^{99\%} \tag{11}$$

where $x^{1\%}$ and $x^{99\%}$ are the 1^{st} and 99^{th} quantiles of $x[mask \neq 0]$.

We then split the image and segmentation in 189 2D slices, and discard all slices where no brain structures are present in the segmentation, resulting in roughly 100 2D slices per brain image.

A.2. BraTS data preprocessing

The BraTS dataset consists of 338 patients' scans. For each patient, four scanner modalities are available, "native T1, post-contrast T1-weighted (T1Gd), T2-weighted (T2), and T2 Fluid Attenuated Inversion Recovery (T2-FLAIR) volumes"[4]. Segmentation maps for GD-enhancing tumour, the peritumoural oedema, and the necrotic and non-enhancing tumour core are provided. In addition, the scans are co-registered, resampled to $1mm^3$ resolution as skull stripped. For more information about the BraTS dataset preprocessing, please refer to Bakas et al. (2018); Menze et al. (2014). We separate the data in three sets, a training set with $n_{train} = 269$, a validation set with $n_{val} = 36$ and a test set of $n_{test} = 33$. For each scan modality, we calculate the mean and variance of the brain pixels across the training set, excluding the background. We use the calculated mean and variance to normalise the data distribution to mean 0 and standard deviation 1.

We then split the images and segmentation in 155 2D slices. For each slice, concatenate the four modalities, and take a centre crop of 176×176.

A.3. Training hyperparameters

We train the DDPM for 100 000 steps with batch size 4 and learning rate $\eta = 0.0001$ on a single NVIDIA TITAN X GPU with 12GB capacity. Similarly, we train the Global CL and Global & Local CL models for 100 000 steps. All downstream models - the supervised baseline, Global CL and Global & Local CL fine-tuning, LEDM, LEDMe and TEDM - are trained for 20 000 steps, with the same learning rate.

[4] https://www.med.upenn.edu/cbica/brats2020/data.html

Appendix B. Computational cost

The backbone UNet used across experiments is of $36m$ parameters. We use the package ptflops to estimate the number of operations to $N = 29.2M$. Therefore, Supervised Baseline, Global CL and Global & Local CL all have a computational cost of N.

LEDM requires $|S_{\text{LEDM}}| = |\{50, 150, 250\}| = 3$ forward passes through the UNet composing the DDPM backbone of the model, one for each used timestep. The latent representations extracted from the UNet has $n_{\text{latent}} = 960$ dimensions. In the case of LEDM, these dimensions are concatenated and passed through a lightweight multilayer perceptron, composed of three linear layers: input_channels ×128, 128 × 32 and 32× out_channels. Here, input_channels=$|S_{\text{LEDM}}| \times n_{\text{latent}}$ and out_channels=1. We denote its size by $N_{\text{MLP}}(\text{in\_c}, \text{out\_c})$, and note that it is executed n_{pixels} times. LEDMe has a similar complexity structure.

Finally, TEDM, like LEDMe, requires $|S_{\text{tEDM}}| = |\{1, 10, 25, 50, 200, 400, 600, 800\}| = 8$ forward passes through the UNet, and requires an MLP of size $N_{\text{MLP}}(n_{\text{latent}}, 1)$ for each latent representation. The final numbers for all models can be found in Table 4.

Note that for LEDM, LEDMe and TEDM, the multiple UNet forward passes are the greatest contributors to computational complexity and can be parallelised provided enough computational power, leading to comparable prediction time to the baseline.

Appendix C. Further results and visualisations

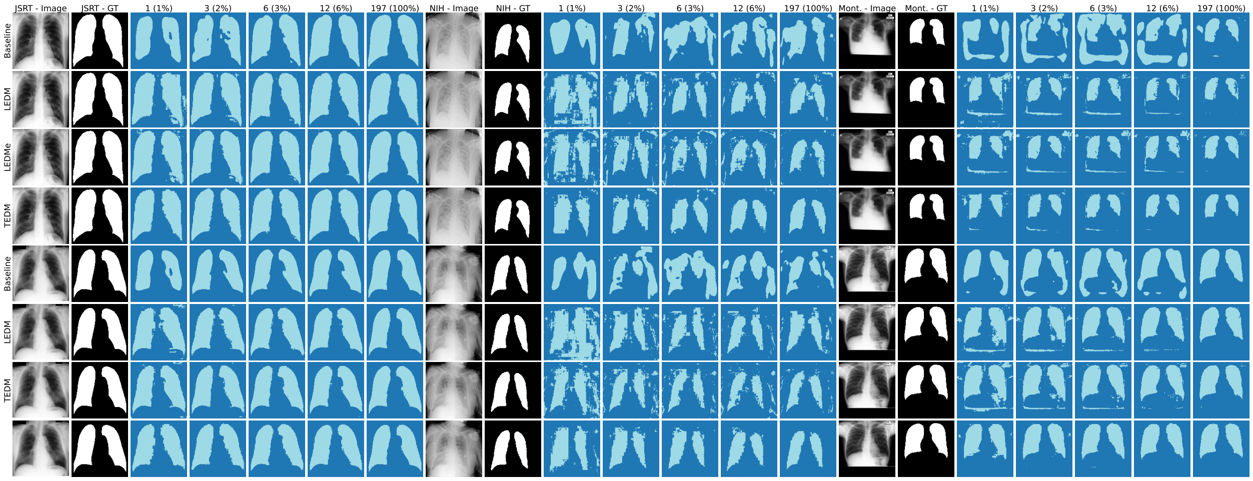


Figure 4: Additional visualisations of segmentations on JSRT, NIH and Montgomery test images as per Figure 3. Please zoom in for better visibility of details.

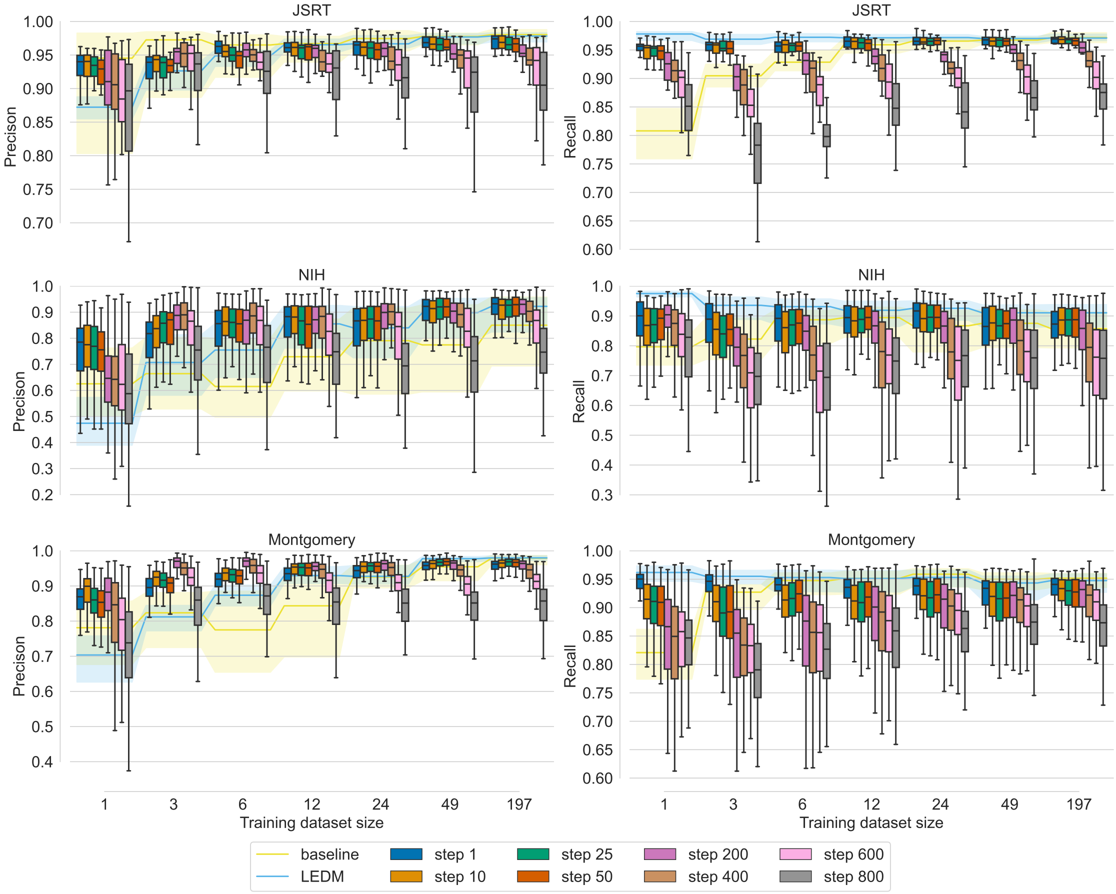


Figure 5: Additional results on the performance of a logistic regression segmentation model trained on latent features from individual diffusion steps.

Table 5: Models precision and recall w.r.t. ground truth segmentations, as per Table 1.

Training size	1	3	6	12	197
	Precision - JSRT (in-domain for classifier)				
Sup. Baseline	**89.2 ± 12.1**	93.2 ± 7.2	93.8 ± 5.9	95.3 ± 3.7	**97.9 ± 1.1**
Global CL	86.8 ± 10.5	95.5 ± 3.0	**97.3 ± 2.6**	**97.0 ± 2.0**	**97.7 ± 1.5**
Global & Local CL	**90.2 ± 9.2**	**97.1 ± 2.2**	96.8 ± 2.0	96.2 ± 2.0	97.1 ± 1.6
LEDM	85.2 ± 5.8	91.7 ± 2.9	94.1 ± 1.9	96.3 ± 1.5	97.5 ± 1.3
LEDMe	**90.1 ± 4.6**	93.6 ± 2.3	96.2 ± 2.0	96.6 ± 1.6	**97.9 ± 0.9**
TEDM (ours)	**91.3 ± 7.2**	95.4 ± 2.9	95.6 ± 2.0	96.4 ± 1.7	97.5 ± 1.2
	Recall - JSRT (in-domain for classifier)				
Sup. Baseline	81.5 ± 6.6	90.6 ± 3.4	93.2 ± 2.6	95.4 ± 2.5	96.8 ± 2.2
Global CL	91.8 ± 3.2	90.2 ± 3.0	90.3 ± 3.8	93.8 ± 1.9	96.6 ± 2.3
Global & Local CL	90.0 ± 3.2	89.5 ± 3.3	89.5 ± 3.7	93.4 ± 2.4	**97.2 ± 1.8**
LEDM	97.4 ± 1.1	96.7 ± 1.2	**97.0 ± 1.6**	96.6 ± 2.1	96.6 ± 2.1
LEDMe	**97.7 ± 1.0**	**97.5 ± 1.4**	**97.1 ± 1.5**	**97.4 ± 1.3**	**97.3 ± 1.9**
TEDM (ours)	95.4 ± 2.1	94.3 ± 1.6	96.2 ± 1.5	96.9 ± 1.4	**97.2 ± 1.9**
	Precision - NIH (in-domain for DDPM, OOD for classifier)				
Sup. Baseline	63.0 ± 17.0	65.6 ± 18.3	63.6 ± 20.3	72.0 ± 18.3	80.5 ± 17.4
Global CL	60.8 ± 17.9	78.7 ± 15.9	76.0 ± 20.4	83.2 ± 14.4	89.4 ± 13.6
Global & Local CL	65.1 ± 19.1	**81.7 ± 15.2**	**84.5 ± 15.4**	81.7 ± 16.8	88.0 ± 13.9
LEDM	48.4 ± 13.6	69.4 ± 14.8	74.7 ± 14.0	83.0 ± 11.4	88.4 ± 9.2
LEDMe	56.8 ± 14.1	69.3 ± 13.7	77.0 ± 12.9	79.8 ± 12.0	90.8 ± 7.8
TEDM (ours)	**70.5 ± 13.3**	**82.0 ± 10.6**	**86.3 ± 9.3**	**90.4 ± 6.9**	**95.3 ± 3.6**
	Recall - NIH (in-domain for DDPM, OOD for classifier)				
Sup. Baseline	77.7 ± 10.3	80.5 ± 12.0	85.4 ± 10.1	87.4 ± 8.0	84.2 ± 9.9
Global CL	88.6 ± 9.7	83.6 ± 8.1	80.1 ± 13.6	87.4 ± 7.7	85.3 ± 8.9
Global & Local CL	80.9 ± 14.5	78.6 ± 11.7	78.9 ± 14.0	84.0 ± 11.5	87.6 ± 8.5
LEDM	**96.4 ± 4.2**	91.8 ± 5.5	91.1 ± 6.2	90.2 ± 6.5	89.9 ± 5.5
LEDMe	**96.3 ± 3.2**	**92.5 ± 6.2**	**91.8 ± 6.7**	90.9 ± 7.2	89.9 ± 5.7
TEDM (ours)	95.7 ± 4.0	**92.4 ± 4.2**	**92.9 ± 4.1**	**92.7 ± 4.4**	**90.8 ± 5.0**
	Precision - Montgomery (OOD for DDPM and classifier)				
Sup. Baseline	75.1 ± 16.4	77.6 ± 16.1	73.5 ± 18.6	78.1 ± 19.0	94.9 ± 8.9
Global CL	68.3 ± 18.3	86.7 ± 13.7	88.8 ± 15.8	89.2 ± 13.8	93.7 ± 14.1
Global & Local CL	72.2 ± 20.9	**90.1 ± 12.7**	**92.2 ± 11.0**	89.2 ± 14.4	92.9 ± 14.7
LEDM	68.7 ± 10.5	79.4 ± 9.7	85.9 ± 8.8	92.0 ± 6.8	97.5 ± 2.7
LEDMe	69.7 ± 9.2	78.8 ± 9.1	84.8 ± 8.5	88.5 ± 7.3	96.4 ± 3.7
TEDM (ours)	**88.7 ± 5.3**	**90.9 ± 5.9**	**93.5 ± 4.9**	**96.9 ± 2.4**	**98.5 ± 1.0**
	Recall - Montgomery (OOD for DDPM and classifier)				
Sup. Baseline	80.9 ± 7.2	90.9 ± 5.9	93.0 ± 5.6	93.0 ± 5.8	93.6 ± 4.8
Global CL	88.7 ± 7.2	89.9 ± 4.8	90.1 ± 5.5	92.8 ± 5.7	93.0 ± 6.5
Global & Local CL	86.1 ± 10.9	88.3 ± 5.5	88.4 ± 6.4	92.2 ± 5.9	93.2 ± 6.0
LEDM	94.9 ± 4.7	94.5 ± 4.2	93.9 ± 4.8	92.9 ± 8.3	92.0 ± 9.4
LEDMe	**97.0 ± 3.5**	**96.3 ± 3.7**	**95.3 ± 4.3**	**94.3 ± 5.1**	**94.4 ± 5.1**
TEDM (ours)	92.9 ± 6.7	92.4 ± 6.9	93.3 ± 7.1	92.8 ± 7.9	92.6 ± 9.1

Table 6: Precision and recall scores on the UK Biobank and BraTS datasets, as per Table 3

.

UK Biobank ($n_{train}^{unlabelled} = 34\,000$, $n_{test} = 500$)

Training size	1	3	6	12	34 000
			Precision		
Sup. Baseline	67.3 ± 18.9	84.5 ± 11.4	84.0 ± 10.7	85.8 ± 9.5	88.7 ± 9.0
Global CL	59.3 ± 23.3	83.1 ± 11.5	82.9 ± 11.1	85.2 ± 9.7	**89.4 ± 8.6**
Global & Local CL	52.3 ± 22.5	75.1 ± 15.0	80.3 ± 11.5	81.7 ± 10.7	88.6 ± 9.2
LEDM	64.9 ± 21.3	83.2 ± 9.6	84.0 ± 9.6	85.5 ± 9.1	86.9 ± 8.8
LEDMe	51.3 ± 19.5	86.0 ± 8.9	86.4 ± 9.2	85.9 ± 9.0	88.5 ± 8.9
TEDM	**85.9 ± 11.7**	**88.8 ± 8.3**	**86.8 ± 9.1**	**87.8 ± 9.0**	87.7 ± 9.2
			Recall		
Sup. Baseline	41.3 ± 20.5	67.8 ± 16.4	79.7 ± 11.4	**82.5 ± 11.2**	**88.6 ± 6.4**
Global CL	30.6 ± 19.6	70.2 ± 14.9	78.8 ± 11.3	**82.8 ± 10.4**	85.8 ± 9.5
Global & Local CL	39.6 ± 19.4	73.6 ± 11.0	**81.1 ± 9.9**	82.7 ± 10.1	86.6 ± 7.6
LEDM	64.4 ± 17.7	**76.2 ± 13.2**	**81.4 ± 10.2**	81.5 ± 11.2	86.2 ± 8.0
LEDMe	**66.0 ± 18.1**	75.2 ± 12.7	79.4 ± 11.1	82.5 ± 10.7	85.0 ± 8.0
TEDM	58.6 ± 20.3	73.2 ± 13.3	79.7 ± 11.1	80.0 ± 11.9	83.0 ± 8.6

BraTS ($n_{train}^{unlabelled} = 268$, $n_{test} = 33$)

Training size	1	3	6	12	33
			Precision		
Sup. Baseline	25.7 ± 30.0	45.1 ± 37.4	54.6 ± 37.6	62.2 ± 35.1	**74.1 ± 26.8**
Global CL	12.0 ± 25.3	38.6 ± 34.9	48.3 ± 37.1	57.1 ± 34.9	66.6 ± 29.9
Global & Local CL	31.6 ± 35.7	40.5 ± 37.2	49.5 ± 36.3	60.7 ± 35.1	66.3 ± 29.2
LEDM	26.4 ± 28.5	44.5 ± 37.9	56.7 ± 35.8	61.6 ± 35.0	70.6 ± 27.4
LEDMe	27.9 ± 29.4	51.2 ± 37.6	60.8 ± 35.2	61.4 ± 34.8	70.4 ± 27.5
TEDM	**46.2 ± 34.2**	**61.4 ± 35.8**	**67.2 ± 33.6**	**67.4 ± 33.4**	72.4 ± 27.0
			Recall		
Sup. Baseline	18.9 ± 28.4	**43.7 ± 36.4**	**48.1 ± 35.8**	**49.5 ± 35.5**	**71.1 ± 26.2**
Global CL	13.6 ± 29.1	38.9 ± 36.2	33.1 ± 33.6	45.2 ± 33.5	56.9 ± 30.9
Global & Local CL	21.0 ± 31.3	38.3 ± 35.8	40.6 ± 34.9	39.4 ± 33.3	56.8 ± 31.8
LEDM	**35.8 ± 26.7**	37.0 ± 34.3	**45.8 ± 33.6**	51.0 ± 32.0	63.8 ± 26.9
LEDMe	26.8 ± 26.8	36.0 ± 32.9	**46.7 ± 34.6**	**53.1 ± 32.5**	64.7 ± 27.7
TEDM	27.6 ± 28.2	37.3 ± 33.3	42.4 ± 33.3	47.9 ± 32.9	59.3 ± 30.2