REGARDING MULTI-INSTANCE LEARNING AS A DENOISING PROCESS FOR SINGLE-CELL DATA

Anonymous authors
Paper under double-blind review

ABSTRACT

Pooling single-cell embeddings into multi-sample bags can improve generalisation to unseen experimental settings. Canonically, bags are formed for convenience, since labelling a bag is cheaper than labelling every instance, and bags are generated the same way during training and at deployment, with instance-level labels free to vary within a bag (Dietterich et al.,1997). Dapello et al. (2025), by contrast, assume bags are constructed with perfect instance level accuracy such that each bag contains samples from the same class during both training and testing. In this work, we show that bag size and bag impurity act through distinct mechanisms. Bag size governs the variance of the mean embedding, with larger bags yielding lower variance, while impurity acts as a bias, shrinking the class-specific bag mean toward a shared centroid. This distinction motivates **Denoising DeepSets (DDS)**, which inserts a noise-conditioned denoiser between mean pooling and classification. By conditioning on the effective noise level implied by bag size, DDS can be trained across bag sizes and can be deployed at any evaluation bag size without retraining. On the datasets from the scGeneScope benchmark, DDS improves on mean-pooling and attention-based MIL baselines with the largest gains at small bag sizes where sampling noise is greatest. These results on both single-cell RNA sequencing and Cell Painting microscopy images suggest that treating bag size as a structured noise variable provides a principled route from bag-level training to robust single-cell prediction.

1 INTRODUCTION

Dapello et al. (2025) show that aggregating multiple single-cell profiles can improve treatment identification under experimental shift. Its multiprofile benchmark samples profiles uniformly without replacement within a treatment category for $k = 32$. This provides a useful controlled benchmark for aggregation, but it does not imply that equally large treatment-consistent groups are available during every deployment setting. In settings where cells are not already grouped by a common indicator or well, or where only few cells are available, a model trained with large bags may need to operate at smaller k, including $k = 1$. Understanding how performance changes with bag size is therefore important for transferring the benefit of aggregation to settings with fewer cells available.

We separate two consequences of bag construction. Under uniform sampling without replacement, the sample mean of a k-cell subsample is an unbiased estimator of the corresponding S-cell population mean, while its sampling variance decreases as k increases (Cochran, 1977). Smaller bags can therefore be viewed as noisier observations of the same underlying full-bag representation, with the noise level determined by k. Secondly, mixing samples across class labels changes the expected mean toward the mean of the contaminating distribution, consistent with the classical contamination framework in robust statistics (Huber, 1964). Under symmetric contamination, analogous to symmetric label noise (Angluin & Laird, 1988; Frénay & Verleysen, 2014), the resulting mixture expectation is correspondingly shifted toward a shared centroid. Bag size can therefore be used as a structured noise coordinate under the subsampling model, whereas conditioning only on bag size cannot in general identify and invert contamination bias.

Motivated by this distinction, we introduce *Denoising DeepSets* (DDS), building on permutation-invariant set representations (Zaheer et al., 2017). DDS applies a denoiser conditioned on the sampling noise implied by bag size, drawing on the broader idea of denoising corrupted representations

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.

(Vincent et al., 2008) and, more specifically, noise-level conditioning in modern denoising models (Song & Ermon, 2019; Ho et al., 2020; Karras et al., 2022), to the pooled representation before classification. The denoiser is trained to map representations constructed from smaller subsamples toward the corresponding full-bag representation, allowing a single model to adapt its correction according to the number of cells available at inference. Across our sequencing and imaging experiments, the DDS–baseline gap is largest at small evaluation bags, where sampling uncertainty is greatest, and generally narrows as averaging reduces this variation. Our contributions are as follows:

I. We derive the finite-population noise schedule for uniformly subsampled bag means, showing how bag size determines the variance of the pooled representation under sampling without replacement, and we distinguish this variance effect from contamination-induced bias.

II. We propose Denoising DeepSets (DDS), a denoiser conditioned on the effective noise level implied by bag size. This allows a single checkpoint to adapt across a range of evaluation bag sizes without retraining.

III. We evaluate DDS on sequencing and Cell Painting datasets under within experiment and held-out experiment shift. Across both modalities, we characterise how performance varies with evaluation bag size and show that the largest numerical gains occur for small bags, where sampling uncertainty is greatest.

2 Background

2.1 Multiple Instance Learning

MIL replaces individual labelled examples with labelled bags, $\mathcal{B} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k\}$ (Dietterich et al., 1997; Maron & Lozano-Pérez, 1997). A common permutation-invariant construction is the DeepSets-style form

$$m(\mathcal{B}) = \rho\left(\bigoplus_{i=1}^{k} \phi(\boldsymbol{x}_i)\right), \tag{1}$$

where ϕ maps instances to latent representations, $\bigoplus$ is a symmetric pooling operator, and ρ is a downstream predictor. The original DeepSets construction uses summation but here we use mean-pooling, which is also permutation invariant (Zaheer et al., 2017). Attention-based MIL instead learns instance weights for constructing a sample-weighted pooled embedding (Ilse et al., 2018), and MIL has also been applied directly to sets of single cells (Arvaniti & Claassen, 2017). In classical MIL, the bag is typically the prediction unit and instance labels are assumed to not be observed. scGeneScope differs from this setting because its multiprofile examples are constructed by sampling profiles within a known treatment category (Dapello et al., 2025). Our setting therefore uses set aggregation to improve treatment representations rather than to infer a bag label from weakly labelled instances.

2.2 Noise-Conditioned Denoising

A denoising model is trained to estimate a clean latent representation $\boldsymbol{z}_0$ from a noisy observation

$$\boldsymbol{z}_\sigma = \boldsymbol{z}_0 + \sigma\boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{2}$$

where σ indexes the noise level. Under squared-error loss, the optimal denoiser is the posterior mean

$$D^*(\boldsymbol{z}_\sigma; \sigma) = \mathbb{E}[\boldsymbol{z}_0 \mid \boldsymbol{z}_\sigma], \tag{3}$$

which is the minimum mean-squared-error estimate of the clean signal (Miyasawa, 1961; Efron, 2011). Modern score-based and diffusion models likewise make the corruption level an explicit conditioning variable (Song et al., 2021; Karras et al., 2022). Because the population-optimal estimator changes with the noise level, we condition the denoiser explicitly on σ and adopt the preconditioning introduced by Karras et al. (2022),

$$D_\theta(\boldsymbol{z}_\sigma; \sigma) = c_{\text{skip}}(\sigma)\boldsymbol{z}_\sigma + c_{\text{out}}(\sigma)F_\theta\big(c_{\text{in}}(\sigma)\boldsymbol{z}_\sigma; c_{\text{noise}}(\sigma)\big), \tag{4}$$

where the coefficients are deterministic functions of σ and the data scale. At low noise, the skip connection dominates and the model remains close to the input; at higher noise, the learned correction contributes more strongly. DDS borrows this noise-conditioning principle, but uses it for a single denoising regression of pooled representations rather than for iterative generation.

2.3 Perturbation Identification under Experimental Shift

scRNA-seq measurements contain technical variation across batches and experimental settings (Hicks et al., 2018; Tung et al., 2017; Luecken et al., 2022). scGeneScope profiles U2-OS cells treated with 28 chemical compounds by scRNA-seq and Cell Painting across two data-generation rounds, and defines within experiment and held-out experiment evaluations from those rounds (Dapello et al., 2025). Its multiprofile inputs are sampled uniformly without replacement within treatment, with $k = 32$ in the main reported setting, and DMSO is collected as a negative control, while the published treatment-identification task predicts the 28 chemical treatments (Dapello et al., 2025). We use the benchmark's within and held-out experiment splits while changing the bag-size schedule. For scRNA-seq we use scVI embeddings (Lopez et al., 2018), and for Cell Painting (Bray et al., 2016), ViT-L embeddings (Dosovitskiy et al., 2021). Recent work has scaled such representations with foundation models for transcriptomics (Cui et al., 2024) and self-supervised models for Cell Painting (Kraus et al., 2024). Modality-specific evaluations are described in Section 5.1.

3 On Noise in Multiple Instance Learning

3.1 Problem Formulation

Consider an encoder $\phi : \mathcal{X} \to \mathbb{R}^{d_z}$ that maps a single-cell profile into a latent representation. For a treatment-specific reference set, let $\mathcal{B} = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_S\}$ contain S cells associated with the same treatment. Its full-bag representation is

$$\bar{\boldsymbol{z}}_S = \frac{1}{S} \sum_{i=1}^{S} \phi(\boldsymbol{x}_i), \tag{5}$$

which is used to infer the treatment label t. At evaluation, only a subset may be available. For $\mathcal{S}_k \subset \{1, \ldots, S\}$ sampled uniformly without replacement with $|\mathcal{S}_k| = k$. The difference $\bar{\boldsymbol{z}}_k - \bar{\boldsymbol{z}}_S$ can therefore be interpreted as sampling-induced corruption of the full-bag representation. This corruption arises naturally from observing fewer cells. It vanishes when $k = S$ and is largest at $k = 1$. Reference bags are constructed within a treatment and data split. The analysis depends only on treating the S observations as a fixed reference population. Modality-specific evaluation choices are deferred to Section 5.1. This formulation exposes two distinct effects. Changing k changes how precisely the fixed reference mean is observed, while changing bag composition changes which population is averaged and can therefore shift the expected representation.

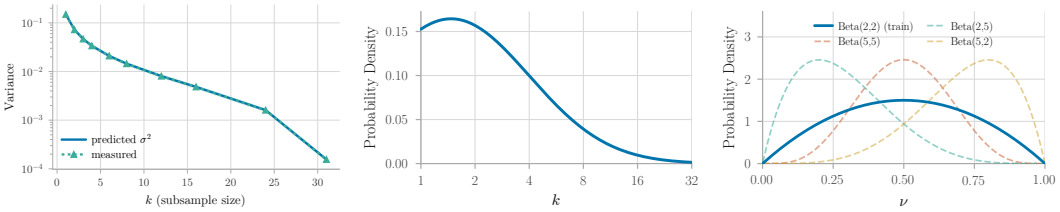
3.2 Empirical Motivation

Bag-size scheduling. The standard multiprofile setting uses large, treatment-consistent bags, but we cannot assume that the same number of cells will always be available at deployment. This train–test mismatch is significant. A DeepSet model trained at $k_{\text{train}} = 32$ and evaluated at $k_{\text{eval}} = 1$ loses approximately 0.06 (0.425 versus 0.484) macro recall relative to a model trained and evaluated at $k = 1$. Training only on large bags therefore leaves the classifier poorly calibrated for the variability of small bags whilst training multiple models is costly.

Table 1: Held-out Macro Recall by training schedule and k_{eval} on DeepSet. Each condition marks the **best** and <u>second-best</u> scores. In each setting k-random gives the highest score, motivating random sampling of k during training across models. k-schedules shown in Appendix Figure 4.

	Held-out Macro Recall		
k-schedule	$k_{\text{eval}} = 1$	$k_{\text{eval}} = 32$	$k_{\text{eval}} \sim \text{lognormal}$
k-random	**0.4420**	**0.5834**	**0.5289**
k-linear	<u>0.4161</u>	0.5712	<u>0.5158</u>
standard	<u>0.4161</u>	<u>0.5716</u>	0.5139

Improved classification with larger bags has already been established (Dapello et al., 2025), but the schedule of bag sizes during training also matters. An obvious choice would be a curriculum that



(a) Finite-population prediction and empirical bag-mean variance.

(b) Clipped, rounded lognormal bag-size draws.

(c) Training and evaluation purity-shape draws.

Figure 1: Noise and sampling structure in the sequencing experiments. (a) The finite-population prediction closely follows empirical bag-mean variance. (b) The lognormal schedule emphasises small, high-variance bags while retaining a tail through $k = S$. (c) Purity shapes used to study changes in bag composition.

begins with small bags and ends with large ones, however, this does not improve predictive performance relative to fixed-k training (Table 1). Instead, sampling k at random throughout training yields a harder objective and keeps the training loss above both alternatives, which leads to better predictive performance across bag-sizes (Table 1). Performance improves when the classifier is forced to generalise across bag sizes rather than converging on the full-bag objective early. This provides initial evidence that k acts as a nuisance variable rather than defining a different prediction task. Varying it during training is useful when the model must adapt to the noise level that it controls. Figure 1a shows that the noise level at k empirically matches the analytical finite population correction

$$\sigma_{\text{eff}}(k) = \hat{\sigma}\sqrt{\frac{S-k}{k(S-1)}}, \tag{6}$$

where $\hat{\sigma}$ is the variance of a full bag of size S and $k \leq S$ is the subsampled bag-size.

These results suggest that k controls a property to which the classifier should adapt. Averaging k cells estimates the full-bag statistic with lower standard error than using one cell, and increasing k lowers the noise of the summary, whereas decreasing it produces progressively noisier observations of the same underlying quantity (see Appendix E for further details).

Bag purity. Consider a bag labelled c in which an on-label fraction ν is mixed with a contaminant fraction $1-\nu$. Express an instance embedding as $\boldsymbol{z}_i = \boldsymbol{\mu}_{c_i} + \boldsymbol{\epsilon}_i$, where c_i denotes its actual class and $\boldsymbol{\mu}_{c_i} = \mathbb{E}[\boldsymbol{z} \mid c_i]$ is the corresponding centroid. The within-class residual $\boldsymbol{\epsilon}_i$ has zero mean and covariance $\boldsymbol{\Sigma}_\epsilon$. Separating residual variation from uncertainty in the contaminants class assignments provides a decomposition of the pooled mean. For a closed-form starting point, assume that each contaminant is sampled uniformly from the $C-1$ other classes. Averaging over uniformly sampled contaminant labels gives the following conditional expectation for a bag assigned to class c:

$$\mathbb{E}[\bar{\boldsymbol{z}} \mid c] = \alpha\, \boldsymbol{\mu}_c + (1-\alpha)\, \bar{\boldsymbol{\mu}}, \tag{7}$$

using the grand centroid $\bar{\boldsymbol{\mu}} = C^{-1} \sum_{c'=1}^{C} \boldsymbol{\mu}_{c'}$ and coefficient

$$\alpha = \nu - \frac{1-\nu}{C-1}. \tag{8}$$

The coefficient α places the expected representation between the labelled centroid and the grand mean. Pure bags retain the former ($\alpha = 1$ when $\nu = 1$), whereas chance-level target membership yields the latter ($\alpha = 0$ when $\nu = 1/C$). For $\alpha > 0$, this affine mapping remains invertible. Shrinkage by itself therefore preserves separability while reducing every between-class distance by the factor α. Furthermore, reducing k preserves the expected bag mean and changes only the uncertainty with which it is estimated. Contamination instead changes the population being averaged, shifting the expected representation towards the contaminated mixture. Under symmetric contamination, this appears as shrinkage towards a shared centroid.

Bag size and purity therefore describe different failure modes: bag size controls uncertainty around a fixed target and can be associated with an effective noise level $\sigma_{\text{eff}}(k)$, while purity changes the

target itself, so conditioning only on bag-size noise provides no information from which to infer and reverse the corresponding bias. This distinction defines the scope of DDS. We condition the denoiser on $\sigma_{\text{eff}}(k)$ because finite subsampling produces a predictable variance effect. We do not expect the same conditioning variable to remove contamination-induced shifts. Further analysis is provided in Appendix D and Appendix E.

4 METHOD: DENOISING DEEPSETS

Section 3.2 motivates treating bag size as a variance coordinate with noise schedule $\sigma_{\text{eff}}(k)$. Representations from small bags are noisier estimates of the full-bag mean than representations from large bags, while a standard classifier receives no explicit indication of this difference. We address this by applying an EDM-preconditioned denoiser (Karras et al., 2022) between mean pooling and classification, conditioned on $\sigma_{\text{eff}}(k)$. The preconditioning places more weight on the observed representation at large k and more weight on the learned correction at small k.

4.1 NOISE CONDITIONING AND PRECONDITIONING

We obtain a noise scale for each bag size through the finite-population correction in Eq. 6. From this we see that smaller bags correspond to noisier observations of the full-bag perturbation statistic, while $\sigma_{\text{eff}}(S) = 0$.

We use $\sigma_{\text{eff}}(k)$ to condition a denoiser applied to the subsampled perturbation representation $\bar{\boldsymbol{z}}_k$. Following the preconditioning of Karras et al. (2022), we parameterise

$$D_\theta(\bar{\boldsymbol{z}}_k; \sigma_{\text{eff}}) = c_{\text{skip}}(\sigma_{\text{eff}})\bar{\boldsymbol{z}}_k + c_{\text{out}}(\sigma_{\text{eff}})F_\theta\left(c_{\text{in}}(\sigma_{\text{eff}})\bar{\boldsymbol{z}}_k; c_{\text{noise}}\right), \tag{9}$$

where F_θ is a noise-conditioned neural network and the coefficients are deterministic functions of σ_{eff} and the signal scale σ_{data}:

$$c_{\text{skip}} = \frac{\sigma_{\text{data}}^2}{\sigma_{\text{data}}^2 + \sigma_{\text{eff}}^2}, \quad c_{\text{out}} = \frac{\sigma_{\text{eff}}\sigma_{\text{data}}}{\sqrt{\sigma_{\text{data}}^2 + \sigma_{\text{eff}}^2}}, \quad c_{\text{in}} = \frac{1}{\sqrt{\sigma_{\text{data}}^2 + \sigma_{\text{eff}}^2}}, \quad c_{\text{noise}} = \tfrac{1}{4}\log\sigma_{\text{eff}}. \tag{10}$$

This parameterisation gives DDS the desired interpolation across bag sizes. In the full-bag limit $k \to S$ with $\sigma_{\text{eff}} \to 0$, the coefficients satisfy $c_{\text{skip}} \to 1$ and $c_{\text{out}} \to 0$, recovering the unmodified bag mean. Conversely, as $\sigma_{\text{eff}} \to \infty$, the skip contribution vanishes through $c_{\text{skip}} \to 0$ and F_θ determines the estimate. Because c_{noise} diverges as $\sigma_{\text{eff}} \to 0$, we enforce a floor $\sigma_{\text{eff}} \geq \epsilon_\sigma$ for numerical stability. Importantly, DDS uses diffusion-style preconditioning only as a noise-conditioned regression mechanism. It does not perform iterative generation. At inference, the available bag size determines $\sigma_{\text{eff}}(k_{\text{eval}})$ and the denoised representation is obtained in a single forward pass. Further motivation, derivations, and scale-estimation details are provided in Appendix D.

4.2 SCALE ESTIMATION

Computing the preconditioning and $\sigma_{\text{eff}}(k)$ requires the signal scale σ_{data} and within-bag variance σ^2 so we maintain exponential moving estimates of both:

$$\hat{\sigma}_n = (1-\gamma)\,\hat{\sigma}_{n-1} + \gamma\,\hat{\sigma}_{\text{batch}}, \tag{11}$$

Here, $\hat{\sigma}_{\text{batch}}$ is the coordinate-wise standard deviation estimated from complete bags ($k = S$) in the current minibatch, and the signal estimate $\hat{\sigma}_{\text{data}}$ uses the across-batch standard deviation of $\bar{\boldsymbol{z}}_S$. Using full bags for both calculations separates scale estimation from subsampling variability.

4.3 DENOISING OBJECTIVE

The denoiser is trained to recover the full-bag statistic $\bar{\boldsymbol{z}}_S$ from the k-cell subsample $\bar{\boldsymbol{z}}_k$. The full-bag statistic contains $\bar{\boldsymbol{z}}_k$ as a constituent

$$\bar{\boldsymbol{z}}_S = \frac{k}{S}\,\bar{\boldsymbol{z}}_k + \frac{S-k}{S}\,\bar{\boldsymbol{z}}_{S\setminus k}. \tag{12}$$

Part of the target, the fraction k/S, can therefore be matched by copying the input. An alternative that removes this copy term is to regress onto the complement mean

$$\tilde{\boldsymbol{z}}_k = \frac{1}{S-k} \sum_{i \notin \mathcal{S}_k} \phi(\boldsymbol{x}_i), \tag{13}$$

which our implementation supports but which is not used in any reported run. All reported models regress onto the full-bag statistic. The denoising loss is

$$\mathcal{L}_{\text{denoise}} = \min\big(\lambda(\sigma_{\text{eff}}(k)),\, w_{\max}\big)\, \|D_\theta(\bar{\boldsymbol{z}}_k;\, \sigma_{\text{eff}}(k)) - \text{sg}(\bar{\boldsymbol{z}}_S)\|^2, \tag{14}$$

where $\lambda(\sigma_{\text{eff}})$ is the EDM loss weight (Eq. 27), capped at $w_{\max}$ because it diverges as $\sigma_{\text{eff}} \to 0$, and $\text{sg}(\cdot)$ denotes the stop-gradient, which protects only the target maximum-size bag embedding. The input $\bar{\boldsymbol{z}}_k$ to the denoiser is computed from ϕ without a stop-gradient, so a trainable encoder receives gradients from both the classification loss and $\mathcal{L}_{\text{denoise}}$. Consequently, the representation scale used to estimate $\hat{\sigma}$ can itself change during training, creating a potential optimisation instability. The total loss is $\mathcal{L}_{\text{CE}} + \beta\, \mathcal{L}_{\text{denoise}}$, with β a fixed coefficient. Modality-specific stability choices and hyperparameters are reported in the evaluation protocol.

5 Results

5.1 Evaluation Setup

We evaluate DDS using the within- and held-out-experiment splits of the scGeneScope treatment-identification benchmark (Dapello et al., 2025). During development, training DDS with ViT-L embeddings was unstable, which led us to applying the steps in Appendix G.

Baselines. We report two pooling-based baselines alongside DDS. Mean-pooling DeepSet averages instance embeddings before classification. Attention DeepSet replaces mean pooling with gated attention (Ilse et al., 2018). The architectures use comparable input-dependent widths, but optimisation settings and output class sets are tuned specifically for each model (Appendix F).

***k*-scheduling.** On sequencing, random-k training outperforms the fixed-$k = 32$ schedule in held-out Macro Recall (Table 1). Thus, exposure to multiple bag sizes improves robustness at both small and full size. We therefore train all three architectures over a distribution of k, rather than fixing k_{train}. In particular, we use a clipped lognormal distribution,

$$k_{\text{train}} \sim \text{Lognormal}(\mu = \ln 4, \sigma = 1), \tag{15}$$

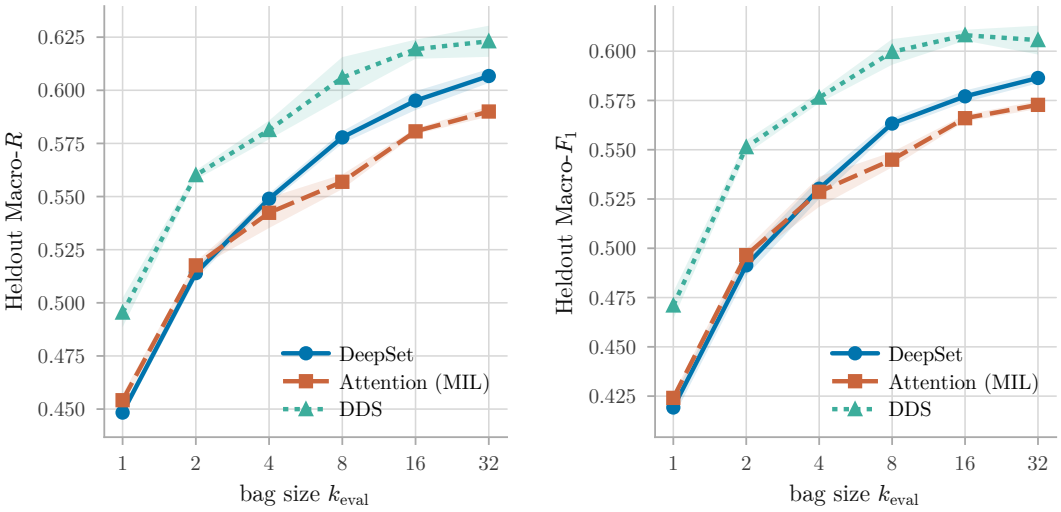
rounded and clipped to $[1, S]$, shown in Figure 1b. Its right-skewed shape deliberately places most training mass on the small, high-noise bags for which noise and size mismatch is most damaging, while retaining a tail of medium and large bags, including the low-noise regime near S. All three models receive the same k-distribution.

Data splits. Models are trained once using the random k protocol and evaluated across $k_{\text{eval}} \in \{1, 2, 4, 8, 16, 32\}$ together with a matched lognormal bag-size condition matching the sampling distribution for k. On imaging the same k_{eval} are drawn from 128-cell evaluation bags (Section 5.1). We use early stopping with the validation split from scGeneScope and the training-time distribution over k. This avoids selecting models using only a specific noise level. Full architecture, optimisation, purity-sweep, and validation details are provided in Appendix F.

Metrics. The primary metrics are balanced accuracy, referred to as Macro Recall, and Macro-F_1, with exact definitions of these in Appendix F.3.

5.2 Single-Profile Classification

At $k_{\text{eval}} = 1$, treatment identification relies on a single profile and averaging provides no variance reduction. In this setting DDS has a clear advantage over the two baselines as evidenced by the scores for both metrics and both splits (Table 2). This is consistent with the hypothesised operating



(a) Macro Recall on held-out experiments.

(b) Macro-F_1 on held-out experiments.

Figure 2: Sequencing held-out performance across evaluation sizes k_{eval} with pure bags and lognormal-k validation. Bands indicate 95% confidence intervals across three seeds. Tabulated results for both within and held-out experiments can be found in Appendix Table 5.

Table 2: Sequencing at $k_{\text{eval}} = 1$ and lognormal evaluation after lognormal-k training and validation ($\nu_{\text{train}} = 1.0$). Mean over 3 seeds. Full results with $\pm$s.e found in Appendix C.

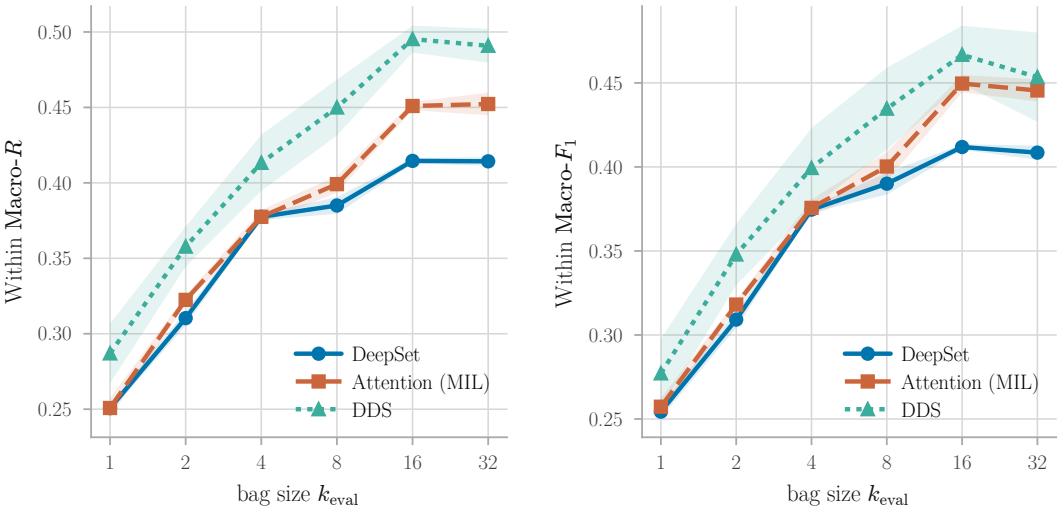
		Within		*Held-out*	
k_{eval}	Model	Macro Recall ↑	Macro-F_1↑	Macro Recall ↑	Macro-F_1↑
1	DeepSet	0.4340	0.4055	0.4484	0.4192
	Attention DeepSet	0.4382	0.4110	0.4542	0.4241
	DDS (ours)	**0.4692**	**0.4555**	**0.4954**	**0.4710**
LogNorm	DeepSet	0.5469	0.5171	0.5504	0.5191
	Attention DeepSet	0.5463	0.5182	0.5373	0.5132
	DDS (ours)	**0.5690**	**0.5565**	**0.5770**	**0.5590**

regime in which the bag-size-conditioned correction is strongest. A possible explanation for the smaller attention-baseline difference is that treatment-consistent bags offer less scope for attention-based instance reweighting (Ilse et al., 2018).

On imaging, using the sequencing model directly leads to the learned noise scale collapsing toward a near pass-through denoiser. This instability motivates the separate ViT-L training recipe used in the evaluation (Appendix G). With this stabilised recipe, DDS has reports higher $k_{\text{eval}} = 1$ scores (Table 3) than the other two methods in all settings. The same processing steps are applied to all of these, ensuring a fair comparison, so the result applies to the complete imaging pipeline. Therefore, there is clear evidence that across both modalities, our denoising approach provides improved performance in the single-profile evaluation case.

5.3 Classification across Bag Sizes

On sequencing, the numerical gap between DDS and the reported baseline scores is largest for the smallest bags and generally contracts as averaging reduces sampling noise (Figure 2), which is consistent with the finite-population view that a uniformly subsampled mean becomes a more precise estimate of the fixed full-bag mean as k grows. Under matched lognormal evaluation, DDS also has the highest reported scores in Table 2. This implies that the benefits of our method are not restricted merely to small bag evaluation settings, and thus, that a choice between models depending on evaluation bag size does not need to be made at deployment.



(a) Macro Recall within experiments.

(b) Macro-F_1 within experiments.

Figure 3: Imaging within-experiment performance across evaluation sizes k_{eval} with pure bags and lognormal-k validation. Bands indicate 95% confidence intervals across three seeds. Tabulated results for both within and held-out experiments can be found in Appendix Table 6.

Table 3: Imaging at $k_{\text{eval}} = 1$ and lognormal evaluation using control-centred PCA-454 inputs ($S = 128$), lognormal-k training and validation, and $\nu_{\text{train}} = 1.0$. Mean over three seeds. Full results with ±s.e found in Appendix C.

		Within		*Held-out*	
k_{eval}	Model	Macro Recall ↑	Macro-F_1↑	Macro Recall ↑	Macro-F_1↑
1	DeepSet	0.2507	0.2544	0.1995	0.2049
	Attention DeepSet	0.2508	0.2573	0.1927	0.1956
	DDS (ours)	**0.2870**	**0.2772**	**0.2334**	**0.2251**
LogNorm	DeepSet	0.3304	0.3274	0.2923	0.2959
	Attention DeepSet	0.3505	0.3513	0.2896	0.2939
	DDS (ours)	**0.4031**	**0.3877**	**0.3268**	**0.3199**

This result also holds for the imaging models as can be seen in Figure 3. However, unlike sequencing, the differences do not decay smoothly with k, which is consistent with bag size alone being an incomplete description of imaging uncertainty. We also see that the DDS and mean-pooling curves approach one another at the largest evaluated size, where variation across seeds also increases. This is to be expected, however, due to the greater batch noise in the imaging data (Appendix G). In any case, the denoising approach does not harm the performance relative to the baselines at any k_{eval} level. Further evidence is provided in Appendix C.

DDS also has higher scores under matched lognormal evaluation (Table 3), so its checkpoint can be used across the evaluated mixture of bag sizes. Both Macro Recall and Macro-F_1 move in the same direction, providing further evidence for this. The contrast between the varying k_{eval} curves is consistent with bag size being a more complete uncertainty coordinate for sequencing than for imaging. On sequencing, the reported gap generally narrows with k, as expected if the sample mean is becoming a more precise estimate of a fixed reference mean. Imaging is less monotonic, suggesting that additional sources of variation remain, despite the additional preprocessing steps. Matched-lognormal evaluation additionally verifies that each DDS checkpoint can be evaluated over the tested mixture of bag sizes without retraining. This highlights the superiority of our method with regards to both raw performance and adaptability.

Purity and ablations. Purity-distribution shape remains consequential even for large bags, which is consistent with contamination changing the expected representation rather than only its variance, and DDS has the highest Macro Recall and Macro-F_1 across the tested shapes (see Section H.1). At $k_{\text{eval}} = 1$, the sequencing ablations identify EDM preconditioning as the most influential component, while the explicit noise embedding has a smaller effect and both analytical shrinkage and learned denoising contribute; see Section H.3. These supplementary results also emphasise the headline claim. The ablations support noise-aware preconditioning most strongly at the noisiest operating point, whereas the purity experiment shows that strong performance does not imply inversion of a contamination-induced mean shift. DDS is therefore best interpreted as targeting structured sampling uncertainty, helping to solve a key problem in multi-instance cell classification.

6 Discussion

The most reproducible pattern in the current experiments is the dependence on bag size. The finite-population analysis predicts increasing uncertainty as k decreases, and the sequencing k_{eval} curves follow the same qualitative trend. DDS reports its largest cross-model gap at small k, where sampling uncertainty is greatest, and narrows as averaging makes the pooled representation a more precise estimate of the full-bag mean. The within-DDS ablations provide more direct evidence for this interpretation than the raw cross-model comparisons. Removing EDM preconditioning produces the largest performance drop among the tested components, particularly in the small-bag regime, which is consistent with its role in adapting the correction to the estimated noise level. The training-schedule experiments support the same view. Models trained only on large bags are poorly calibrated for small-bag evaluation, whereas training over a distribution of k improves robustness across evaluation sizes. This suggests treating bag size as a nuisance variable controlling representation uncertainty rather than as defining a different prediction problem. The lognormal schedule implements this by emphasising small, noisy bags while retaining medium and large bags.

Purity represents a different statistical challenge. Uniform subsampling changes the variance of the bag mean while preserving its expectation, whereas contamination changes the population being averaged and shifts the expected representation. Under symmetric contamination this appears as shrinkage towards a shared centroid. A denoiser conditioned only on $\sigma_{\text{eff}}(k)$ therefore has information about sampling uncertainty, but not about the amount or direction of contamination, and the purity experiments should consequently be interpreted as tests of sensitivity to bag composition rather than evidence that DDS reconstructs an uncontaminated class mean. Taken together, the finite-population trend, k-scheduling experiment, matched-k evaluation and within-DDS ablations are consistent with a bag-size-aware denoising mechanism. They support the claim that explicitly modelling the sampling uncertainty associated with bag size can improve robustness when the number of available cells changes.

For deployment, the validation distribution should reflect the number of cells expected at inference, and selecting a checkpoint only on full bags evaluates the model where sampling noise is smallest and can conceal poor behaviour on single cells or small bags. Conversely, a model intended for mixed bag sizes should be trained and selected over that mixture. More generally, DDS provides a way to use stable pooled representations during training while remaining robust when fewer cells are available, provided that $\sigma_{\text{eff}}(k)$ remains calibrated in the representation used by the classifier. This flexibility is a clear benefit of this approach and should prove critical in real-world settings.

6.1 Limitations and Future Work

An immediate weakness is that both modalities use only three seeds, and their training regimes are modality-specific because direct transfer to ViT-L was unstable during development. This limits causal cross-modality comparisons even when the evaluation protocol is matched. Moreover, the scalar finite-population model omits anisotropic, class-dependent and correlated noise, while the S-cell target remains finite. Future work should first test broader bag-size distributions, decompose mixture bias from variance, and directly calibrate whether denoising error to the full-bag mean follows $\sigma_{\text{eff}}(k)$ under held-out shift. A natural extension would then be to explore the use of both modalities in tandem through a multimodal model, which would require delicate balancing of the distinct processing used for each. In a different direction, adding a degree of label-missingness and adapting the method to the semi-supervised case could lead to even greater applicability.

THE USE OF LARGE LANGUAGE MODELS

Large language models were used only to proofread, polish the wording of this paper, and generate tables and figures from experimental data. They were not used for research ideation, experimental design or analysis planning. The authors take full responsibility for all content.

REFERENCES

Dana Angluin and Philip Laird. Learning from noisy examples. *Machine Learning*, 2(4):343–370, 1988.

Eirini Arvaniti and Manfred Claassen. Sensitive detection of rare disease-associated cell subsets via representation learning. *Nature Communications*, 8:14825, 2017.

Mark-Anthony Bray, Shantanu Singh, Han Han, Chadwick T. Davis, Blake Borgeson, Cathy Hartland, Maria Kost-Alimova, Sigrun M. Gustafsdottir, Christopher C. Gibson, and Anne E. Carpenter. Cell painting, a high-content image-based assay for morphological profiling using multiplexed fluorescent dyes. *Nature Protocols*, 11(9):1757–1774, 2016.

William G. Cochran. *Sampling Techniques*. John Wiley & Sons, New York, 3 edition, 1977. ISBN 9780471162407.

Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scGPT: Toward building a foundation model for single-cell multi-omics using generative AI. *Nature Methods*, 21(8):1470–1480, 2024.

Joel Dapello, Marcel Nassar, Ridvan Eksi, Ban Wang, Jules Gagnon-Marchand, Kenneth T. Gao, Akram Baharlouei, Kyra Thrush, Nina Riehs, Amy Peterson, Aniket Tolpadi, Abhejit Rajagopal, Henry Miller, Ashley Conard, David Alvarez-Melis, Rory Stark, Simone Bianco, Morgan Levine, Ava Amini, Alex Xijie Lu, Nicolo Fusi, Ravi Pandya, Valentina Pedoia, and Hana El-Samad. scgenescope: A treatment-matched single cell imaging and transcriptomics dataset and benchmark for treatment response modeling. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-4706.

Thomas G Dietterich, Richard H Lathrop, and Tomás Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence*, 89(1-2):31–71, 1997. doi: 10.1016/S0004-3702(96)00034-3.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Bradley Efron. Tweedie's formula and selection bias. *Journal of the American Statistical Association*, 106(496):1602–1614, 2011. doi: 10.1198/jasa.2011.tm11181.

Benoît Frénay and Michel Verleysen. Classification in the presence of label noise: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, 25(5):845–869, 2014.

Stephanie C Hicks, F William Townes, Mingxiang Teng, and Rafael A Irizarry. Missing data and technical variability in single-cell rna-sequencing experiments. *Biostatistics*, 19(4):562–578, 2018. doi: 10.1093/biostatistics/kxx053.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Peter J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964. doi: 10.1214/aoms/1177703732.

Maximilian Ilse, Jakub M. Tomczak, and Max Welling. Attention-based deep multiple instance learning. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 2127–2136. PMLR, 2018.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022. doi: 10.52202/068431-1926.

Oren Kraus, Kian Kenyon-Dean, Saber Saberian, Maryam Fallah, Peter McLean, Jess Leung, Vasudev Sharma, Ayla Khan, Jia Balakrishnan, Safiye Celik, et al. Masked autoencoders for microscopy are scalable learners of cellular biology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11757–11768, 2024.

Romain Lopez, Jeffrey Regier, Michael B. Cole, Michael I. Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15:1053–1058, 2018. doi: 10.1038/s41592-018-0229-2.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Malte D Luecken, Maren Büttner, Kridsadakorn Chaichoompu, Anna Danese, Marta Interlandi, Michaela F Müller, Daniel C Strobl, Luke Zappia, Martin Dugas, Maria Colomé-Tatché, et al. Benchmarking atlas-level data integration in single-cell genomics. *Nature methods*, 19(1):41–50, 2022. doi: 10.1038/s41592-021-01336-8.

Oded Maron and Tomás Lozano-Pérez. A framework for multiple-instance learning. *Advances in neural information processing systems*, 10:570–576, 1997.

Koichi Miyasawa. An empirical bayes estimator of the mean of a normal population. *Bulletin of the International Statistical Institute*, 38(4):181–188, 1961.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

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

Po-Yuan Tung, John D Blischak, Chiaowen Joyce Hsiao, David A Knowles, Jonathan E Burnett, Jonathan K Pritchard, and Yoav Gilad. Batch effects and the effective design of single-cell gene expression studies. *Scientific reports*, 7(1):39921, 2017. doi: 10.1038/srep39921.

Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th International Conference on Machine Learning*, pp. 1096–1103, 2008.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Ruslan Salakhutdinov, and Alexander J. Smola. Deep sets. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

A HYPERPARAMETER SWEEP

The sweep varies optimisation and denoising parameters while retaining the common architecture and three-seed evaluation protocol.

Table 4: Hyperparameter sweep ranges for Denoising DeepSets (DDS).

DDS · HYPERPARAMETERS	Type	Values
`lr`	Log-uniform	$[3\times10^{-6},\ 10^{-4}]$
`weight_decay`	Log-uniform	$[10^{-6},\ 10^{-3}]$
`edm_weight_max`	Categorical	$\{50,\ 100,\ 200\}$
`denoise_weight` (β)	Categorical	$\{0.5,\ 1.0,\ 2.0\}$
`var_reg_lambda`	Categorical	$\{0.0,\ -1.0,\ -5.0\}$
`seed`	Replicate	$\{0,\ 1,\ 2\}$

B ALGORITHMIC IMPLEMENTATION

Algorithm 1: $\hat{\boldsymbol{z}} \leftarrow \text{DDSFORWARD}(\mathcal{B}^t, \mathcal{B}^c, k \mid \phi, \theta)$

Input: $\mathcal{B}^t, \mathcal{B}^c$: bags of S perturbed and control cells; k: subsample size.
Output: $\hat{\boldsymbol{z}}$: denoised perturbation representation.

$\boldsymbol{z}_i^t \leftarrow \phi(\boldsymbol{x}_i^t)$ for all $i \in \{1, \dots, S\}$, $\quad \boldsymbol{z}_i^c \leftarrow \phi(\boldsymbol{x}_i^c)$ for all i
Sample $\mathcal{S}_k \subset \{1, \dots, S\}$ of size k without replacement
$\bar{\boldsymbol{z}}_k \leftarrow k^{-1} \sum_{i \in \mathcal{S}_k} \boldsymbol{z}_i$
$\sigma_{\text{eff}} \leftarrow \hat{\sigma} \sqrt{(S-k)/(k(S-1))}$
$\hat{\boldsymbol{z}} \leftarrow c_{\text{skip}}\, \bar{\boldsymbol{z}}_k + c_{\text{out}}\, F_\theta(c_{\text{in}}\, \bar{\boldsymbol{z}}_k;\ c_{\text{noise}})$
return $\hat{\boldsymbol{z}}$

Algorithm 2: $(\hat{\theta}, \hat{\phi}) \leftarrow \text{DDSTRAINING}(\mathcal{D}, \theta, \phi)$

Input: $\mathcal{D} = \{(\mathcal{B}_i^t, \mathcal{B}_i^c, t_i)\}_{i=1}^N$: training bags with labels.
Input: θ, ϕ: initial denoiser and encoder parameters.
Output: $\hat{\theta}, \hat{\phi}$: trained parameters.
Hyperparameters: η: learning rate; β: denoising weight; $w_{\max}$: EDM weight cap; n_{warm}: warmup steps.

for $n = 1, 2, \dots$ **do**
  Sample minibatch $\mathcal{M} \sim \mathcal{D}$
  foreach $(\mathcal{B}^t, \mathcal{B}^c, t) \in \mathcal{M}$ **do**
    Sample $k \leftarrow \text{clip}(\text{round}(e^u), 1, S)$, $u \sim \mathcal{N}(\ln 4, 1)$
    $\hat{\boldsymbol{z}} \leftarrow \text{DDSFORWARD}(\mathcal{B}^t, \mathcal{B}^c, k \mid \phi, \theta)$
    Compute full-bag target $\bar{\boldsymbol{z}}_S$ via Eq. 5
  Update EMA estimate $\hat{\sigma}$ (σ_{data} is a constant)
  $\mathcal{L}_{\text{CE}} \leftarrow \frac{1}{|\mathcal{M}|} \sum \ell_{\text{CE}}(\rho(\hat{\boldsymbol{z}}),\ t)$
  $\mathcal{L}_{\text{denoise}} \leftarrow \frac{1}{|\mathcal{M}|} \sum \min(\lambda(\sigma_{\text{eff}}), w_{\max}) \|\hat{\boldsymbol{z}} - \text{sg}(\bar{\boldsymbol{z}}_S)\|^2$
  $\beta_n \leftarrow \beta \cdot \mathbb{I}[n > n_{\text{warm}}]$
  $(\theta, \phi) \leftarrow (\theta, \phi) - \eta \nabla_{\theta,\phi} (\mathcal{L}_{\text{CE}} + \beta_n \mathcal{L}_{\text{denoise}})$
return $(\hat{\theta}, \hat{\phi}) = (\theta, \phi)$

The implementation is contained in algorithm 1 and slots directly into the outer training loop algorithm 2 largely unchanged from standard DeepSet. At inference time, k is set to the available bag size and no subsampling is performed. The denoiser adapts its shrinkage via $\sigma_{\text{eff}}(k_{\text{eval}})$.

C EXTENDED RESULTS

Table 5: **Sequencing** results vs. evaluation bag size k_{eval}, at $\nu_{\text{train}} = \nu_{\text{eval}} = 1.0$ (pure bags), $k_{\text{val}} \sim \text{lognormal}$, within-test vs. held-out-test. The bottom block is the matched $k_{\text{eval}} \sim \text{Lognormal}(\mu, \sigma)$ condition drawn from the training-time distribution. **Best** and second-best results are highlighted per metric within each k_{eval} block. Mean ± s.e. over 3 seeds.

BAG-SIZE FAN-OUT		*Within*		*Heldout*	
k_{eval}	Model	Macro Recall ↑	Macro-F_1↑	Macro Recall ↑	Macro-F_1↑
1	DeepSet	0.4340 ± 0.0022	0.4055 ± 0.0040	0.4484 ± 0.0021	0.4192 ± 0.0045
	Attention DeepSet	0.4382 ± 0.0015	0.4110 ± 0.0016	0.4542 ± 0.0025	0.4241 ± 0.0016
	DDS (ours)	$\mathbf{0.4692 \pm 0.0017}$	$\mathbf{0.4555 \pm 0.0002}$	$\mathbf{0.4954 \pm 0.0061}$	$\mathbf{0.4710 \pm 0.0060}$
2	DeepSet	0.5089 ± 0.0013	0.4814 ± 0.0046	0.5140 ± 0.0026	0.4913 ± 0.0047
	Attention DeepSet	0.5115 ± 0.0030	0.4869 ± 0.0036	0.5176 ± 0.0016	0.4965 ± 0.0028
	DDS (ours)	$\mathbf{0.5471 \pm 0.0017}$	$\mathbf{0.5342 \pm 0.0021}$	$\mathbf{0.5601 \pm 0.0019}$	$\mathbf{0.5513 \pm 0.0031}$
4	DeepSet	0.5611 ± 0.0022	0.5388 ± 0.0034	0.5490 ± 0.0022	0.5302 ± 0.0046
	Attention DeepSet	0.5633 ± 0.0006	0.5451 ± 0.0003	0.5424 ± 0.0063	0.5287 ± 0.0068
	DDS (ours)	$\mathbf{0.5830 \pm 0.0040}$	$\mathbf{0.5726 \pm 0.0027}$	$\mathbf{0.5813 \pm 0.0039}$	$\mathbf{0.5764 \pm 0.0025}$
8	DeepSet	0.5928 ± 0.0036	0.5688 ± 0.0050	0.5778 ± 0.0026	0.5633 ± 0.0024
	Attention DeepSet	0.5924 ± 0.0029	0.5704 ± 0.0018	0.5569 ± 0.0031	0.5449 ± 0.0034
	DDS (ours)	$\mathbf{0.6138 \pm 0.0014}$	$\mathbf{0.6028 \pm 0.0042}$	$\mathbf{0.6060 \pm 0.0085}$	$\mathbf{0.5996 \pm 0.0057}$
16	DeepSet	0.6108 ± 0.0026	0.5832 ± 0.0038	0.5951 ± 0.0039	0.5771 ± 0.0026
	Attention DeepSet	0.6088 ± 0.0021	0.5809 ± 0.0023	0.5807 ± 0.0012	0.5660 ± 0.0015
	DDS (ours)	$\mathbf{0.6208 \pm 0.0063}$	$\mathbf{0.6088 \pm 0.0069}$	$\mathbf{0.6193 \pm 0.0040}$	$\mathbf{0.6081 \pm 0.0025}$
32	DeepSet	$\mathbf{0.6173 \pm 0.0012}$	0.5887 ± 0.0023	0.6067 ± 0.0029	0.5864 ± 0.0024
	Attention DeepSet	0.6147 ± 0.0003	0.5862 ± 0.0007	0.5900 ± 0.0021	0.5728 ± 0.0018
	DDS (ours)	$\mathbf{0.6173 \pm 0.0044}$	$\mathbf{0.6005 \pm 0.0063}$	$\mathbf{0.6231 \pm 0.0065}$	$\mathbf{0.6056 \pm 0.0065}$
∼ Lognormal	DeepSet	0.5469 ± 0.0016	0.5171 ± 0.0022	0.5504 ± 0.0011	0.5191 ± 0.0007
	Attention DeepSet	0.5463 ± 0.0007	0.5182 ± 0.0002	0.5373 ± 0.0035	0.5132 ± 0.0029
	DDS (ours)	$\mathbf{0.5690 \pm 0.0030}$	$\mathbf{0.5565 \pm 0.0051}$	$\mathbf{0.5770 \pm 0.0069}$	$\mathbf{0.5590 \pm 0.0032}$

Table 6: **ViT-L** results vs. evaluation bag size k_{eval}, at $\nu_{\text{train}} = \nu_{\text{eval}} = 1.0$ (pure bags), $k_{\text{val}} \sim \text{lognormal}$, within-test vs. held-out-test. The bottom block is the matched $k_{\text{eval}} \sim \text{Lognormal}(\mu, \sigma)$ condition drawn from the training-time distribution. **Best** and second-best results are highlighted per metric within each k_{eval} block. Mean ± s.e. over 3 seeds.

BAG-SIZE FAN-OUT		*Within*		*Heldout*	
k_{eval}	Model	Macro Recall ↑	Macro-F_1↑	Macro Recall ↑	Macro-F_1↑
1	DeepSet	0.2300 ± 0.0018	0.2190 ± 0.0055	0.1807 ± 0.0013	0.1557 ± 0.0024
	Attention DeepSet	0.2339 ± 0.0018	0.2323 ± 0.0007	0.1787 ± 0.0046	0.1626 ± 0.0029
	DDS (ours)	$\mathbf{0.2719 \pm 0.0064}$	$\mathbf{0.2527 \pm 0.0075}$	$\mathbf{0.2206 \pm 0.0037}$	$\mathbf{0.2055 \pm 0.0045}$
2	DeepSet	0.2801 ± 0.0047	0.2600 ± 0.0068	0.2157 ± 0.0042	0.1844 ± 0.0073
	Attention DeepSet	0.2778 ± 0.0029	0.2684 ± 0.0021	0.2096 ± 0.0047	0.1853 ± 0.0030
	DDS (ours)	$\mathbf{0.3381 \pm 0.0032}$	$\mathbf{0.3139 \pm 0.0078}$	$\mathbf{0.2869 \pm 0.0010}$	$\mathbf{0.2642 \pm 0.0045}$
4	DeepSet	0.3103 ± 0.0048	0.2818 ± 0.0084	0.2408 ± 0.0056	0.2029 ± 0.0111
	Attention DeepSet	0.3151 ± 0.0015	0.2965 ± 0.0008	0.2327 ± 0.0045	0.2032 ± 0.0022
	DDS (ours)	$\mathbf{0.4104 \pm 0.0033}$	$\mathbf{0.3765 \pm 0.0028}$	$\mathbf{0.3510 \pm 0.0043}$	$\mathbf{0.3161 \pm 0.0075}$
8	DeepSet	0.3322 ± 0.0061	0.2940 ± 0.0085	0.2548 ± 0.0060	0.2118 ± 0.0117
	Attention DeepSet	0.3346 ± 0.0023	0.3069 ± 0.0013	0.2542 ± 0.0046	0.2222 ± 0.0026
	DDS (ours)	$\mathbf{0.4486 \pm 0.0046}$	$\mathbf{0.4008 \pm 0.0036}$	$\mathbf{0.3907 \pm 0.0069}$	$\mathbf{0.3422 \pm 0.0094}$
16	DeepSet	0.3481 ± 0.0060	0.3036 ± 0.0101	0.2691 ± 0.0083	0.2190 ± 0.0125
	Attention DeepSet	0.3477 ± 0.0036	0.3167 ± 0.0023	0.2718 ± 0.0035	0.2335 ± 0.0047
	DDS (ours)	$\mathbf{0.4670 \pm 0.0063}$	$\mathbf{0.4138 \pm 0.0037}$	$\mathbf{0.4181 \pm 0.0101}$	$\mathbf{0.3517 \pm 0.0101}$
32	DeepSet	0.3600 ± 0.0046	0.3100 ± 0.0104	0.2771 ± 0.0071	0.2208 ± 0.0106
	Attention DeepSet	0.3563 ± 0.0033	0.3221 ± 0.0017	0.2774 ± 0.0010	0.2373 ± 0.0076
	DDS (ours)	$\mathbf{0.4735 \pm 0.0152}$	$\mathbf{0.4172 \pm 0.0113}$	$\mathbf{0.4246 \pm 0.0151}$	$\mathbf{0.3387 \pm 0.0147}$
∼ Lognormal	DeepSet	0.3019 ± 0.0037	0.2728 ± 0.0077	0.2299 ± 0.0058	0.1948 ± 0.0102
	Attention DeepSet	0.3061 ± 0.0013	0.2850 ± 0.0009	0.2291 ± 0.0032	0.2013 ± 0.0025
	DDS (ours)	$\mathbf{0.3914 \pm 0.0058}$	$\mathbf{0.3563 \pm 0.0079}$	$\mathbf{0.3386 \pm 0.0040}$	$\mathbf{0.3004 \pm 0.0071}$

D Denoising Interpretation and Preconditioning Details

D.1 Bag Size as a Denoising Problem

For uniform sampling without replacement from a fixed reference bag, the subsampled bag mean is an unbiased estimator of the corresponding full-bag mean. This motivates the approximation

$$\boldsymbol{z}_\sigma = \boldsymbol{z}_0 + \sigma\boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{16}$$

where $\boldsymbol{z}_0$ denotes the full-bag representation and $\sigma = \sigma_{\text{eff}}(k)$ is determined by the bag size.

Under squared-error loss, the optimal estimator of the clean representation from a noisy observation is the posterior mean,

$$D^*(\boldsymbol{z}_\sigma; \sigma) = \mathbb{E}[\boldsymbol{z}_0 \mid \boldsymbol{z}_\sigma], \tag{17}$$

which is the minimum mean-squared-error estimator. This provides the statistical motivation for inserting a denoiser between pooling and classification.

The connection to score-based diffusion models follows from Tweedie's formula (Efron, 2011),

$$\nabla_{\boldsymbol{z}} \log p(\boldsymbol{z}; \sigma) = \frac{D^*(\boldsymbol{z}; \sigma) - \boldsymbol{z}}{\sigma^2}. \tag{18}$$

For the variance-exploding parameterisation, the corresponding probability-flow ODE can be written as (Song et al., 2021)

$$\frac{d\boldsymbol{z}}{d\sigma} = -\sigma \nabla_{\boldsymbol{z}} \log p(\boldsymbol{z}; \sigma). \tag{19}$$

Substituting equation 18 gives

$$\frac{d\boldsymbol{z}}{d\sigma} = \frac{\boldsymbol{z} - D^*(\boldsymbol{z}; \sigma)}{\sigma}. \tag{20}$$

We use this connection only to motivate noise-conditioned denoising. DDS does not numerically integrate this ODE. The squared-error objective makes the posterior mean the population optimum under the assumed corruption model; the trained network only approximates that estimator. Inference therefore consists of one evaluation

$$D_\theta(\bar{\boldsymbol{z}}_k; \sigma_{\text{eff}}(k)), \tag{21}$$

rather than iterative diffusion sampling.

D.2 EDM Preconditioning

We adopt the preconditioning proposed by Karras et al. (2022). The denoiser is expressed as

$$D_\theta(\bar{\boldsymbol{z}}_k; \sigma_{\text{eff}}) = c_{\text{skip}}(\sigma_{\text{eff}})\bar{\boldsymbol{z}}_k + c_{\text{out}}(\sigma_{\text{eff}}) F_\theta\left(c_{\text{in}}(\sigma_{\text{eff}})\bar{\boldsymbol{z}}_k; c_{\text{noise}}(\sigma_{\text{eff}})\right), \tag{22}$$

with

$$c_{\text{skip}} = \frac{\sigma_{\text{data}}^2}{\sigma_{\text{data}}^2 + \sigma_{\text{eff}}^2}, \tag{23}$$

$$c_{\text{out}} = \frac{\sigma_{\text{eff}}\sigma_{\text{data}}}{\sqrt{\sigma_{\text{data}}^2 + \sigma_{\text{eff}}^2}}, \tag{24}$$

$$c_{\text{in}} = \frac{1}{\sqrt{\sigma_{\text{data}}^2 + \sigma_{\text{eff}}^2}}, \qquad c_{\text{noise}} = \tfrac{1}{4}\log\sigma_{\text{eff}}. \tag{25}$$

Here, σ_{data} characterises the scale of clean perturbation representations, while σ_{eff} characterises the uncertainty introduced by finite-sample pooling.

The limiting behaviour makes the role of the preconditioning explicit. As $k \to S$, $\sigma_{\text{eff}} \to 0$, giving

$$c_{\text{skip}} \to 1, \qquad c_{\text{out}} \to 0. \tag{26}$$

The denoiser therefore approaches the identity map when the observed bag already contains the full set of cells. Conversely, as the noise level increases, the contribution of the direct skip connection decreases and the learned correction becomes increasingly important.

The original EDM formulation additionally uses the noise-dependent loss weighting

$$\lambda(\sigma_{\text{eff}}) = \frac{\sigma_{\text{data}}^2 + \sigma_{\text{eff}}^2}{\sigma_{\text{eff}}^2 \sigma_{\text{data}}^2}. \tag{27}$$

This weight diverges as $\sigma_{\text{eff}} \to 0$, which made the uncapped objective unstable in preliminary experiments. We therefore cap it at $w_{\max}$ (Eq. 14) and scale the capped loss by a fixed coefficient β. On sequencing ($w_{\max} = 100$, $\sigma_{\text{data}} = 0.5$) the uncapped weight exceeds the cap from $k = 3$–6 onwards, so most bag sizes train under a constant weight (Section H.2). Modality-specific values are reported with the evaluation protocol in Appendix F.

D.3 Estimating the Noise and Signal Scales

DDS requires estimates of two scales: the within-bag variability used to construct $\sigma_{\text{eff}}(k)$ and the across-example scale σ_{data} used by the preconditioning.

For a full bag of size S, we estimate the within-bag standard deviation from the instance embeddings and track it during training using an exponential moving average,

$$\hat{\sigma}_n = (1 - \gamma)\hat{\sigma}_{n-1} + \gamma \hat{\sigma}_{\text{batch}}. \tag{28}$$

The effective noise level for a subsample of size k is then

$$\sigma_{\text{eff}}(k) = \hat{\sigma} \sqrt{\frac{S - k}{k(S - 1)}}, \tag{29}$$

following the finite-population correction.

The signal scale σ_{data} is held constant rather than tracked: if it followed the data, a trainable encoder could shrink its own preconditioning denominator. On sequencing we set $\sigma_{\text{data}} = 0.5$, with the output LayerNorm on ϕ fixing the scale of the encoder output. The noise scale $\hat{\sigma}$ is estimated from full bags so that it is not confounded by the subsampling noise it is intended to model. Modality-specific scale choices are reported in the evaluation protocol.

Because $\sigma_{\text{eff}}(S) = 0$ and c_{noise} contains $\log \sigma_{\text{eff}}$, we apply a small numerical floor

$$\sigma_{\text{eff}} \leftarrow \max(\sigma_{\text{eff}}, \epsilon_\sigma) \tag{30}$$

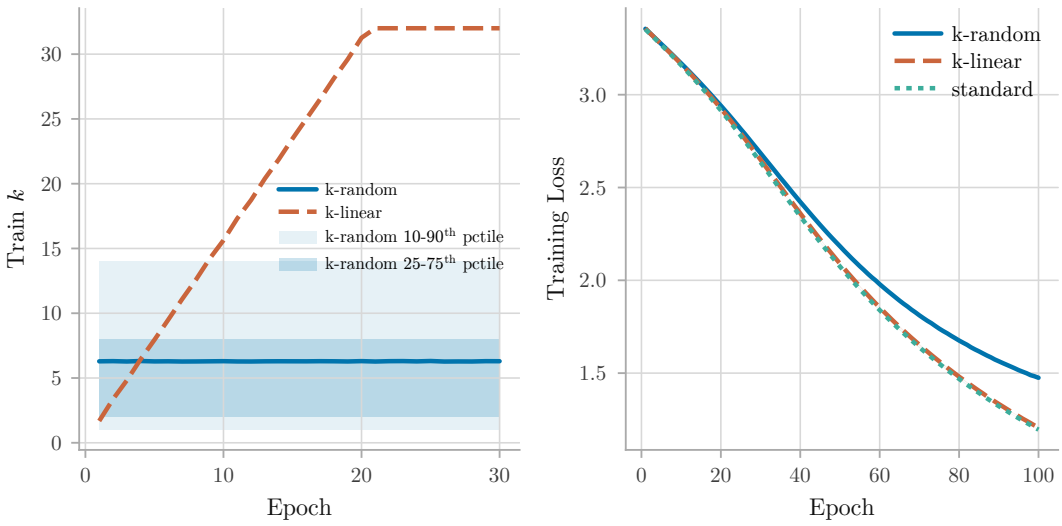
when evaluating the preconditioning coefficients.

E Additional Empirical Motivation

E.1 Bag Size Indexes a Noise Level

scGeneScope reports that treatment-identification performance generally improves as multiprofile cardinality increases, although the trend is not monotonic in every held-out setting (Dapello et al., 2025). Less obvious is whether the training cardinality should itself be fixed. We compared a random bag-size schedule against a curriculum that increases k over training and the standard fixed-$k = 32$ baseline. Random-k training is the hardest optimisation problem in terms of training loss because it continues to present small, noisy bags throughout training, but it generalises better across evaluation sizes. On the held-out split it reaches Macro-$R = 0.4420$ at $k_{\text{eval}} = 1$, compared with 0.4161 for fixed $k = 32$; at $k_{\text{eval}} = 32$ the corresponding scores are 0.5834 and 0.5716; and under matched random-k evaluation they are 0.5289 and 0.5139 (Table 1). The fact that random-k training also improves the $k_{\text{eval}} = 32$ endpoint is important: its benefit is not merely that it trades full-bag performance for robustness to small bags.

These results motivate training over a distribution of k rather than choosing a single training cardinality. We instantiate the random schedule as the clipped lognormal distribution used in the main experiments, $k \sim \text{Lognormal}(\ln 4, 1)$. The distribution is intentionally right-skewed: it repeatedly exposes the model to the small, noisy bags where the fixed-k mismatch is largest, while its upper tail continues to supply medium and large bags so that robustness is not purchased by abandoning the low-noise regime. The scheduling experiment establishes the empirical value of varying k; DDS then makes the associated variation in uncertainty explicit by conditioning the denoiser on $\sigma_{\text{eff}}(k)$.



(a) k-schedule.

(b) Training loss.

Figure 4: Bag-size scheduling changes what a DeepSet classifier learns, not just how fast it trains. Despite k-random incurring a persistently *higher* training loss than k-linear or a fixed $k = 32$ baseline (b) – a direct consequence of training on smaller, noisier bags for the full run rather than converging to the easy full-bag objective within the first 20 epochs (a) – it reaches *higher* test-time Macro-R than both alternatives at every evaluated bag size (Table 1).

E.2 Bag Purity Attenuates Signal and Injects Between-Class Noise

Contamination changes the expected signal, whereas reducing bag size preserves that expectation and increases only its uncertainty. Across the tested contamination levels, centroid shrinkage contributes 4–44$\times$ more to classification degradation than the additional between-class variance. The displacement predicted by Eq. 7 agrees with the observed centroids within 1–3% for both frozen and trained embeddings. This agreement supports the linear decomposition as an account of the principal degradation mechanism, despite the nonlinear mapping used to construct the embeddings.

E.3 Dual Axes for Evaluation

Size and purity change bags in mathematically different ways and represent different practical constraints. This distinction produces a testable prediction for DDS. Along the size axis, σ_{eff} measures the relevant variability, so denoising should improve estimates by shrinking noisy small-bag observations towards a learned class-centred signal at an appropriate strength. Along the purity axis, the same conditioning variable does not describe the principal change in the expected representation. It therefore gives the denoiser no corresponding mechanism for reversing contamination-induced attenuation. We expect robustness to varying bag size without analogous robustness to varying purity. If DDS and mean pooling lose performance at similar rates as contamination rises, that behaviour would be consistent with the stated mechanism. The denoiser corrects the variance represented by its conditioning input but cannot recover information about a bias absent from that input.

F Experimental Protocol Details

F.1 Architecture and Optimisation

Architectures and optimiser settings differ by model and modality; Table 7 lists them. The ViT-L DDS settings are intentionally not copied from sequencing because that transfer was unstable during development: with a trainable encoder, $\hat{\sigma}$ collapsed from 0.43 to 0.07 and the denoiser became nearly a pass-through (Appendix G). On sequencing, the DDS encoder ϕ is a multi-layer perceptron with ReLU activations followed by the output LayerNorm described in Section 4, and the classifier ρ is

a linear map to the perturbation-label space. In the ViT-L evaluation, ϕ is the identity on the 454-dimensional control-centred PCA features and ρ has two hidden layers. Once the unstable trainable encoder is removed, the remaining denoiser/classifier optimisation is retuned rather than inheriting a learning rate calibrated for the sequencing parameterisation.

For DDS, the denoiser F_θ is a separate MLP. The effective noise level σ_{eff} enters through a 64-dimensional Fourier embedding of c_{noise} that is concatenated to the denoiser input (input width $454 + 64 = 518$ on imaging). The sequencing DDS additionally uses inverse-frequency class weighting.

Models are trained with AdamW (Loshchilov & Hutter, 2019). The learning rate is reduced on validation-loss plateaus using `ReduceLROnPlateau` with reduction factor 0.1 and patience 5. Early stopping is applied with patience 12 epochs.

Denoising-specific hyperparameters include the denoising-loss coefficient β, the EDM weight cap $w_{\max}$, the constant σ_{data}, EMA decay γ, the noise floor ϵ_σ, and the denoising warm-up period. These are selected on the validation split and then fixed across the reported experiments of each modality.

	Sequencing (scVI)	Imaging (ViT-L, centred PCA-454)
DDS ϕ	MLP $200 \to 271 \to 271$, output LayerNorm	identity
DDS F_θ / ρ hidden widths	$[271]$ / none (linear ρ)	$[454]$ / $[454, 454]$
DDS learning rate / weight decay	1.56×10^{-5} / 4.2×10^{-5}	1.3×10^{-3} / 0.01
DDS β / $w_{\max}$ / σ_{data}	1.0 / 100 / 0.5	6.35 / 50 / 0.231
DDS S / batch size	32 / 1024	128 / 1024
Baseline hidden width / head depth	271 / 1	454 / 2
Baseline learning rate / weight decay	1.56×10^{-5} / 4.2×10^{-5}	1.07×10^{-4} / 2.6×10^{-3}
Baseline S / batch size	32 / 3072	128 / 4096

Table 7: Per-modality architecture and optimisation settings. Imaging values are those of the results in Section 5.3; the development runs of Appendix G use the settings stated there.

F.2 Imaging Protocol

DeepSet, Attention DeepSet and DDS all take the centred PCA-454 inputs and use $S = 128$. They are trained with $k_{\text{train}} \sim \text{Lognormal}(\ln 4, 1)$ clipped to $[1, 128]$, pure bags, and lognormal-k validation for early stopping, with three seeds each. Their optimiser settings are model-specific. This is deliberate rather than an attempt to give DDS a different objective: directly transferring the sequencing DDS training recipe was unstable, with the learned noise scale collapsing and the denoiser becoming almost a pass-through (Appendix G). After stabilising the representation, the DDS learning rate is therefore re-tuned for the reduced trainable parameterisation; the baselines retain their earlier ViT-L hyperparameters with only the input dimension changed to 454. DDS uses learning rate 1.3×10^{-3} and weight decay 0.01 (Table 7). Evaluation uses 128-cell bags with $k_{\text{eval}} \in \{1, 2, 4, 8, 16, 32\}$ and a matched lognormal condition, on both within-experiment and held-out-experiment splits.

F.3 Metrics and Checkpoint Selection

Perturbation identification is class-imbalanced, so our primary metric is balanced accuracy, equivalently macro-averaged recall,

$$\text{Macro Recall} = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \frac{\text{TP}_t}{\text{TP}_t + \text{FN}_t}. \tag{31}$$

We additionally report macro-averaged F_1,

$$\text{Macro-}F_1 = \frac{1}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \frac{2\text{TP}_t}{2\text{TP}_t + \text{FP}_t + \text{FN}_t}. \tag{32}$$

For each run we select the epoch with the highest validation Macro Recall and report Macro-F_1 from the same epoch. Each configuration is evaluated over three random seeds, with results reported as mean and standard deviation.

F.4 Compared Models

The reported results compare three architectures. The *Mean DeepSet* averages instance embeddings before classification and is the direct mean-pooling baseline for DDS. *Attention MIL* replaces mean pooling with gated attention following Ilse et al. (2018). *DDS* applies the proposed noise-conditioned denoiser to the pooled representation before classification. The fixed-$k_{\text{train}} = S$ scGeneScope schedule appears only in the bag-size scheduling comparison of Appendix E.

All three models are trained using an identical per-example bag-size distribution,

$$k_{\text{train}} \sim \text{Lognormal}(\mu = \ln 4, \sigma = 1), \tag{33}$$

rounded and clipped to $[1, S]$. This choice follows the scheduling comparison in Appendix E: random-k training improves held-out Macro-R over fixed $k = 32$ at both $k_{\text{eval}} = 1$ and 32, as well as under matched random-k evaluation. The lognormal shape concentrates training on small, high-noise bags while preserving a long tail toward larger k. A separate value of k is sampled for each training example rather than once per minibatch.

F.5 Bag-Size Evaluation

Each trained checkpoint is evaluated at

$$k_{\text{eval}} \in \{1, 2, 4, 8, 16, 32\} \tag{34}$$

on both modalities (drawn from 128-cell evaluation bags on imaging), and under a matched random condition

$$k_{\text{eval}} \sim \text{Lognormal}(\mu, \sigma), \tag{35}$$

using the same distribution as during training. No retraining is performed when changing k_{eval}. Experiments are performed on both within-experiment and held-out-experiment splits, with the held-out split treated as the primary measure of performance under experimental shift.

F.6 Validation Bag-Size Distribution

When k_{train} is stochastic, checkpoint selection can depend on the bag-size distribution used for validation. We therefore consider two validation regimes. The *fixed* regime evaluates validation performance only at $k = S$, matching our fixed-k reproduction of scGeneScope's principal $k = 32$ multiprofile cardinality. The *matched* regime approximates the same lognormal distribution over k used during training by combining fixed-k validation subsets with corresponding weights. The matched regime is used for the headline results so that checkpoint selection reflects the operating distribution encountered during training rather than only the lowest-noise condition $k = S$. The imaging grid runs the matched regime only.

F.7 Purity Experiments

Training-bag purity is varied over

$$\nu_{\text{train}} \in \{1.0, 0.75, 0.625, 0.5, 0.375, 0.25, 0.125, 0.0625\}. \tag{36}$$

This sweep is run on sequencing only. Mean DeepSet, Attention MIL, and DDS are retrained independently at each training purity. A training bag of purity ν contains $\lceil \nu S \rceil$ majority-class cells. For the main purity analysis, evaluation bags remain clean, $\nu_{\text{eval}} = 1$, so training-time contamination is the intended manipulated factor.

Each purity-trained checkpoint is reused across multiple evaluation bag sizes without retraining, including $k_{\text{eval}} = 1$, $k_{\text{eval}} = 32$, and the matched lognormal condition. We additionally evaluate

$$\nu_{\text{eval}} \in \{0.0357, 0.2768, 0.5179, 0.7589, 1.0\}, \tag{37}$$

together with a mixed-purity condition. These results are supplementary to the main analysis of training-time contamination under clean deployment.

F.8 Ablation Protocol

DDS variants on sequencing remove or modify one component at a time: removing the denoiser F_θ (keeping the shrinkage), removing both the denoiser and the shrinkage, removing EDM preconditioning, withholding the noise level, and replacing the noise embedding with a scalar MLP or a discrete lookup embedding. The output LayerNorm and fixed-k_{train} variants are not reported (Section H.3).

G ViT-L Adaptation Details

This appendix documents the training instability that prevented direct transfer of the sequencing DDS configuration to ViT-L and reports the development runs used to stabilise the ViT-L evaluation. The central failure is not merely lower validation accuracy: the learned representation contracts during training, driving the tracked noise scale toward zero and turning the EDM preconditioner into an almost pure skip connection. The ViT-L training differences used in the main evaluation are therefore motivated as stability interventions. Because several interventions are cumulative, these runs do not provide a fully factorial causal decomposition. All numbers here come from development runs that precede the three-seed imaging results of Section 5.3; they are evaluated on 32-cell bags ($k_{\text{eval}} \leq 32$), and their seed counts are given in each caption.

G.1 Diagnosis

The sequencing configuration performs poorly on raw ViT-L. With a learned encoder, DDS reports 0.186 held-out balanced accuracy at $k_{\text{eval}} = 32$, compared with 0.277 for the DeepSet configuration. The output class sets differ, so this numerical comparison is diagnostic rather than a like-for-like architecture comparison. A 40-run Bayesian search and a 10-point learning-rate grid both plateaued at 0.19–0.20 validation accuracy, and the DDS-specific settings (denoising weight, EDM weight cap, variance regulariser) showed no correlation with validation accuracy (Spearman $|\rho| < 0.07$).

The batch offset exceeds the treatment signal. We decomposed the spread of the per-(batch, treatment) mean embeddings, excluding controls, into a treatment term and a batch term, each the mean squared distance of its group centroids from the grand mean. On ViT-L the batch term is $1.74\times$ the treatment term; on sequencing it is $0.12\times$.[1] Subtracting each batch's mean control embedding reduces the ratio to 0.18 on ViT-L. The estimated batch-offset term does not shrink with bag size, whereas the within-group variance of a bag mean falls approximately as $1/k$. In this decomposition, the ViT-L batch term exceeds the sampling term near $k \approx 20$. On sequencing the corresponding crossover is near $k \approx 57$, beyond the largest bag evaluated. A parameter-free nearest-centroid classifier on 32-cell bag means shows a consistent pattern. On ViT-L its held-out accuracy rises from 0.114 to 0.237 when each batch's mean control is subtracted, and on sequencing only from 0.602 to 0.667.

The learned noise scale collapses. $\hat{\sigma}$ is measured in the output space of the trainable encoder, and the encoder receives gradient from the denoising loss through the denoiser's input (Section 4.3), so training can lower $\hat{\sigma}$ by squeezing the within-bag spread of ϕ. On ViT-L it fell from 0.43 at initialisation to 0.07. With $\sigma_{\text{data}} = 0.5$ this gives $c_{\text{skip}} \approx 0.98$ at $k = 1$, so the denoiser became almost a pass-through, and the uncapped EDM weight at $k = 1$ was 345 against a cap of 15, leaving the SNR-based reweighting inactive at every bag size. On sequencing $\hat{\sigma}$ settled at 0.15. A negative variance regulariser ($\lambda = -5$, the setting of earlier ViT-L configurations) did not prevent the collapse: $\hat{\sigma}$ ended at 0.073 against 0.067 without it.

G.2 Effect of each change

Starting from the best configuration of the search above (learning rate 3.87×10^{-6}, weight decay 0.050, $\beta = 6.35$, EDM cap 20.9, batch size 1024, lognormal k with median 4, $S = 32$), the changes of Section G.2 are applied cumulatively (Table 8).

[1] The sequencing embeddings carry no batch labels, so the four splits stand in for batches there.

Control centring. Centring alone raises held-out accuracy at $k_{\text{eval}} = 32$ from 0.186 to 0.350, but barely helps single cells (0.127 to 0.156). The offset removed has norm 42–44 in every batch.

Fixed PCA space. After moving denoising to the fixed space, the $k_{\text{eval}} = 1$ score increases from 0.156 to 0.206. At this step the per-cell noise is $5.9\times$ the signal spread and $c_{\text{skip}}(k{=}1) = 0.03$, so the direct skip path receives little weight; the learning rate was raised to 3×10^{-5}. The held-out $k_{\text{eval}} = 32$ score decreases from 0.350 to 0.286. In this configuration $S = 32$, so a 32-cell bag has $\sigma_{\text{eff}} = 0$ and the preconditioned correction vanishes at that boundary. The observed drop from $k_{\text{eval}} = 16$ to 32 (0.347 to 0.286) is consistent with, but does not by itself establish, that boundary explanation.

Reference bag $S = 128$. After changing to $S = 128$, the $k_{\text{eval}} = 32$ score increases from 0.286 to 0.359, while $k_{\text{eval}} = 1$ is nearly unchanged. A development run at $S = 256$ produced a similar result. In these runs, increasing S without fixing the representation did not produce the same change: with the learned encoder, held-out accuracy at $k_{\text{eval}} = 32$ was 0.341.

Learning rate. Without a trainable encoder only the denoiser and the classifier head learn. The validation search indicates that the learning rate inherited from the learned-encoder runs is sub-optimal. A 12-run Bayesian search over learning rate ($[3 \times 10^{-6}, 3 \times 10^{-4}]$) and weight decay ($[10^{-3}, 0.3]$) placed the three best runs at 2.4–2.9 $\times 10^{-4}$, with no visible effect of weight decay. Because that optimum lay at the top of the searched range, a follow-up sweep extended the learning rate to 3.5×10^{-3}; validation accuracy was flat across it (0.343–0.352). The imaging results of Section 5.3 use its best validation configuration, learning rate 1.3×10^{-3} and weight decay 0.01.

Table 8: Development runs: Macro Recall of DDS after each cumulative change, evaluated on 32-cell bags. The results are reported as the mean ± std over 3 seeds.

DDS configuration	Held-out, $k_{\text{eval}} = 1$	Held-out, $k_{\text{eval}} = 32$	Within, $k_{\text{eval}} = 32$
Learned space, raw inputs, $S = 32$ (start)	0.127	0.186	0.308
+ per-batch control centring	0.156 ± 0.001	0.350 ± 0.011	0.392 ± 0.021
+ fixed PCA-454 space, $\phi = \text{id}$	0.206 ± 0.002	0.286 ± 0.006	0.438 ± 0.009
+ reference bag $S = 128$	0.204 ± 0.000	0.359 ± 0.013	0.483 ± 0.000
+ learning rate $\approx 2.5 \times 10^{-4}$	$\mathbf{0.238 \pm 0.001}$	$\mathbf{0.416 \pm 0.003}$	$\mathbf{0.497 \pm 0.000}$

G.3 Centring helps every model

Control centring is not specific to DDS. Retrained on the centred 5120-dimensional inputs with their original hyperparameters, the baseline scores increase substantially: at held-out $k_{\text{eval}} = 32$, DeepSet rises from 0.277 to 0.400 and Attention DeepSet from 0.277 to 0.392 (Table 9). This indicates that batch correction accounts for a substantial portion of the earlier numerical gap between the in-model-control DDS and uncentred classifiers; it does not isolate the denoiser's causal contribution. On centred inputs, DDS with the learned encoder reports 0.350 at $k_{\text{eval}} = 32$, while removing its denoiser gives 0.316, a smaller change than that associated with centring.

The fixed-space DDS has higher scores than the centred baselines at small bags in these development runs (for example, 0.238 vs 0.195 at $k_{\text{eval}} = 1$), but the comparisons are not like-for-like: the configurations differ in input dimension, reference size, tuning procedure, seed count, and output class set. Comparisons to the earlier control-subtracting DDS likewise bundle several configuration changes. These development runs are therefore diagnostic rather than evidence of an isolated architecture effect.

Table 9: Development runs: held-out balanced accuracy against evaluation bag size, at $\nu_{\text{train}} = \nu_{\text{eval}} = 1.0$, on 32-cell evaluation bags. The first four columns use control-centred ViT-L inputs (5120-dimensional for DeepSet, Attention DeepSet and learned-ϕ DDS, with $S = 32$; PCA-454 with $S = 128$ for fixed-space DDS). Baselines and learned-ϕ DDS are mean $\pm$ std over 2 seeds; fixed-space DDS is over the 3 best configurations of its first learning-rate search, one seed each. The last column is the earlier ViT-L DDS on raw inputs, over 3 seeds. **Bold** marks the best value per row.

	Control-centred inputs			
k_{eval}	DeepSet	Attention DeepSet	DDS, learned ϕ	DDS, fixed space
1	0.187 ± 0.009	0.195 ± 0.002	0.156 ± 0.001	$\mathbf{0.238 \pm 0.001}$
2	0.246 ± 0.004	0.246 ± 0.001	0.198 ± 0.001	$\mathbf{0.296 \pm 0.004}$
4	0.297 ± 0.004	0.302 ± 0.004	0.243 ± 0.002	$\mathbf{0.346 \pm 0.005}$
8	0.335 ± 0.001	0.337 ± 0.010	0.290 ± 0.008	$\mathbf{0.364 \pm 0.002}$
16	0.372 ± 0.000	0.370 ± 0.015	0.322 ± 0.005	$\mathbf{0.394 \pm 0.001}$
32	0.400 ± 0.007	0.392 ± 0.018	0.350 ± 0.011	$\mathbf{0.416 \pm 0.003}$
$\sim$Lognormal	0.286 ± 0.001	0.291 ± 0.004	0.236 ± 0.000	$\mathbf{0.332 \pm 0.005}$

H Additional Experiments

H.1 Sequencing Purity-Shape Transfer

This experiment trains each architecture with $\nu_{\text{train}} \sim \text{Beta}(2, 2)$ and evaluates pure bags, symmetric Beta distributions with different concentrations, and off-centre distributions with lower or higher mean purity. The effect of shape remains substantial as bag size grows: increasing k reduces sampling variance but leaves the mixture-induced shift in the expected representation. The low-purity $\text{Beta}(2, 5)$ condition is therefore much harder than the high-purity $\text{Beta}(5, 2)$ condition.

At $k_{\text{eval}} = 32$, DDS has the highest reported held-out Macro Recall for every tested shape (Table 10), but the DDS and baseline class sets differ, so this is not a strict architectural ranking. All configurations remain strongly sensitive to bag composition. The experiment tests distributional shape rather than a fixed-severity contamination curve and should not be interpreted as evidence that DDS inverts contamination-induced mean shift.

H.2 Diagnostic Traces

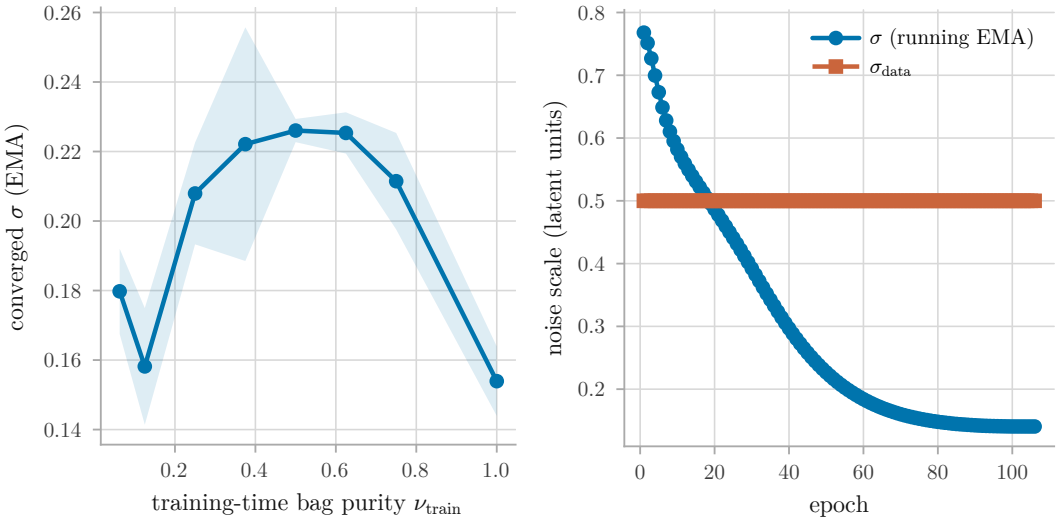
The diagnostic traces test whether numerical failure could explain the sequencing results. In panel (a) of Figure 5, converged $\hat{\sigma}$ has an inverted-U relationship with ν_{train}. Values are low at the extremes, with ≈ 0.16 near $\nu_{\text{train}} = 0.1$ and ≈ 0.15 at $\nu_{\text{train}} = 1.0$, rising to ≈ 0.23 around $\nu_{\text{train}} \approx 0.5$–$0.6$. The epoch trace in panel (b) shows convergence at $\nu_{\text{train}} = 1.0$: $\hat{\sigma}$ falls from ≈ 0.77 to ≈ 0.14 and crosses $\hat{\sigma}_{\text{data}} \approx 0.5$ around epochs 18–20. None of these traces indicates collapse or an unbounded noise estimate.

H.3 Ablations

The ablations are evaluated at $k_{\text{eval}} = 1$, where σ_{eff} is largest and the direct denoising contribution is easiest to distinguish. Among the DDS variants, full DDS has the highest value for every metric and split in Table 11. Removing EDM preconditioning produces the largest loss: held-out Macro Recall falls from 0.4954 to 0.4700. Removing the noise input or replacing the Fourier embedding with a scalar MLP has a smaller effect, while the lookup embedding performs less well. Retaining analytical shrinkage without F_θ reduces held-out Macro Recall to 0.4878; removing both components gives 0.4868. With three seeds, the smaller differences establish an ordering more clearly than they resolve each component's individual effect.

Table 10: Sequencing held-out Macro Recall at $k_{\text{eval}} = 32$ after training with $\nu_{\text{train}} \sim \text{Beta}(2, 2)$. Mean over three seeds. DDS and the classifier baselines use different output class sets, so absolute cross-model scores are not strictly like-for-like.

Evaluation purity shape	DeepSet	Attention DeepSet	DDS
Pure ($\nu = 1$)	.5856	.5852	.6241
$\text{Beta}(20, 20)$	.4246	.4335	.4844
$\text{Beta}(5, 5)$	.3987	.4091	.4674
$\text{Beta}(2, 2)$	.3739	.3863	.4461
$\text{Beta}(1, 1)$	.3700	.3777	.4263
$\text{Beta}(0.5, 0.5)$	.3854	.3918	.4343
$\text{Beta}(2, 5)$	.2133	.2304	.2935
$\text{Beta}(5, 2)$	.4966	.4992	.5476



(a) Noise $\hat{\sigma}$ across ν_{train}.

(b) Epoch-wise $\hat{\sigma}$ at $\nu_{\text{train}} = 1.0$.

Figure 5: DDS scale diagnostics. (a) Estimated noise $\hat{\sigma}$ after convergence across training purity. (b) Evolution of $\hat{\sigma}$ during training at $\nu_{\text{train}} = 1.0$.

Table 11: Component comparisons for sequencing DDS at $k_{\text{eval}} = 1$ and $\nu_{\text{eval}} = 1.0$. Symbols ✓/– indicate whether a component is enabled. Across DDS variants, **bold** marks the best and underlining the second-best value.

Ablations, $k_{\text{eval}} = 1$			*Within*		*Held-out*	
Precond.	Denoiser	Noise embed.	Macro Recall ↑	Macro-F_1↑	Macro Recall ↑	Macro-F_1↑
–	✓	Fourier	0.4410	0.4344	0.4700	0.4479
✓	✓	Lookup	0.4519	0.4408	0.4810	0.4567
✓	✓	None	<u>0.4669</u>	<u>0.4493</u>	<u>0.4901</u>	0.4612
✓	✓	Scalar MLP	0.4637	<u>0.4493</u>	0.4892	0.4615
✓	–	Fourier	0.4589	0.4413	0.4878	0.4571
–	–	Fourier	0.4519	0.4413	0.4868	<u>0.4624</u>
✓	✓	Fourier	**0.4692**	**0.4555**	**0.4954**	**0.4710**
DeepSet (reference)			0.4340	0.4055	0.4484	0.4192