NORMAL GUIDANCE IS WHAT ATTENTION NEEDS FOR MULTIPLE INSTANCE LOCALIZATION IN CT SCANS

Anonymous authors
Paper under double-blind review

ABSTRACT

We consider training classifiers for 3D medical images using only one binary label for the entire volume rather than a label for each 2D slice. In such weakly supervised settings, can we learn accurate classifiers for slice-level predictions? Attention-based multiple instance learning (MIL) can produce an attention score for every slice. Yet recent work demonstrates that a simple center-focused baseline that ignores image content can outperform attention-based and transformer-based MIL at slice-level classification of 3D brain scans. We show this baseline also outperforms existing MIL at slice-level classification of thoracic and abdominal CT scans. Motivated by this baseline, we propose Normal Guidance, a regularization technique that encourages the learned attention distribution to follow a bell-shaped curve. Across three medical imaging datasets totaling over 4 million 2D slices, we show our Normal Guidance enables attention-based and transformer-based MIL methods to deliver significantly better slice-level localization than the state-of-the-art while remaining competitive at whole-scan classification.

1 INTRODUCTION

Deep learning is now a go-to approach for prediction tasks with 3D medical images (Saab et al., 2019; Wu et al., 2021; Castro-Macías et al., 2024; Xu et al., 2025). The primary task is often to classify the presence or absence of disease in the organ being imaged (e.g., brain or lungs). In practice, clinicians manually interpret a 3D scan by inspecting each 2D slice for signs of disease. Thus, a complementary prediction task is *localization*, where the goal is to indicate which specific 2D slices of the 3D volume show evidence for the overall prediction. Localization models predict a probabilistic binary label for each slice. These slice-level scores can be directly assessed by human experts via their established workflow. Such audits can improve interpretability and trust needed for deployment. Yet in many applied tasks, especially with 3D medical scans, the expense of acquiring expert labels means that model development can only be "weakly" supervised: the training set contains a coarse scan-level binary label for each 3D scan, but not fine-grained labels for individual slices. An unsolved technical challenge is thus to develop effective slice-level localization from weak supervision, while preserving competitive whole-scan classification.

Deep multiple instance learning (MIL) is a strategy for such weakly-supervised learning (Quellec et al., 2017; Saab et al., 2019). MIL methods can take as input a variable-sized set of instances (a "bag"), each with their own feature vector, and produce one predicted binary label. For localization, attention-based MIL (Ilse et al., 2018) has become the foundation of many methods in the last decade due to its by-design interpretability: each instance j in bag i has an attention weight $a_{ij} \in [0, 1]$ indicating its relative importance for the classification task. Overall, localization in MIL remains underexplored (Castro-Macías et al., 2024). Past works have used attention weights for localization primarily through *qualitative* results (Ilse et al., 2018; Lu et al., 2021; Shao et al., 2021; Keshvarikhojasteh et al., 2024) that visually highlight regions of interest with high attention values. A growing body of work also quantitatively evaluates localization (Li et al., 2021; Fourkioti et al., 2024; Castro-Macías et al., 2024; Chraki et al., 2026), though usually not on 3D medical images.

Recently, Harvey et al. (2026) applied MIL to several datasets of 3D brain scans, including computed tomography (CT) and magnetic resonance imaging (MRI). Harvey et al. showed a simple

Code: anonymous.4open.science/r/normal-guidance

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.

Centered Gaussian (CG) baseline that *ignores all image content* can outperform the learned attention from modern MIL for localization. In CG, the attention values for 3D scan i are set as $a_{ij} \propto \text{NormPDF}(j|\frac{S_i}{2}, 1)$ where S_i is the number of 2D axial slices in the 3D scan ("bag") and integer j indexes each 2D slice ("instance") from 1 to S_i in spatial order. This discretized univariate Normal PDF results in bell-shaped attention when plotted as a function of j, with a peak at the middle slice. It reflects the clinical intuition that the center of the brain is typically more relevant than the edges when detecting the presence of lesions or white matter disease.

In this paper, we offer three contributions to the development of interpretable weakly-supervised classifiers of 3D medical images, focusing on CT scans due to large, high-quality available datasets:

- We show that Harvey et al.'s simple CG baseline outperforms all tested modern MIL at localization across body parts, including lung CT and abdominal CT (kidneys, spleen, liver, etc.). This expands Harvey et al.'s brain-specific past results and suggests current MIL methods are not adequate for 3D CT localization even given large labeled data. Better inductive biases are needed.
- We propose Normal Guidance (NG, Sec. 3.2) for regularizing the attention produced by deep MIL. Motivated by CG's success, our NG forms a reference normal distribution over slice position j parameterized by the empirical mean and variance of the learned attention weights, then encourages the learned attention to match this shape. This reference distribution provides 3 helpful inductive biases during training: smoothness, center-focus, and unimodality. Unlike the CG baseline, both our NG and its multi-head extension (Sec. 3.3) can predict attention vectors that are multi-modal, non-centered, and tailored to individual CT scans. Our NG enables attention-based and transformer-based MIL to beat the CG baseline for localization (Tab. 1) while remaining competitive at whole scan classification (Tab. 2).
- We develop methods to quantify a best-in-class ceiling for localization and scan-level classification (Sec. 4.3). These ceilings unfairly use *instance-level* labels to achieve best-case results that serve as approximate practical upper bounds for weakly-supervised MIL. While other MIL methods remain far from the localization ceiling on all tasks, we show that MIL with Normal Guidance comes much closer (within 0.03 AUROC of ceiling) on two CT tasks: Head and Chest.

Our supporting evidence for these contributions is drawn from experiments on three open-access datasets of real medical CT scans totaling over 4 million 2D slices and covering different body parts (head, chest, abdomen). The classification tasks include acute injuries, embolisms, and lesions. Each dataset includes human expert annotations for every slice. All non-ceiling models are trained in strict weak-supervised fashion, using only scan-level labels for training, early stopping, and hyperparameter selection. Slice-level labels are only used to evaluate localization quality.

We hope this work improves weakly-supervised localization for 3D imaging tasks and advances understanding of what inductive biases are needed for deep MIL to move from the lab to the clinic.

2 Background and Related Works

MIL (Dietterich et al., 1997; Maron & Lozano-Pérez, 1997; Quellec et al., 2017) is a branch of weakly supervised learning. In MIL, the training dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$ consists of N bags, where each bag $x_i = \{x_{i,1}, \ldots, x_{i,S_i}\}$ contains a set of S_i instances. Although instance-level labels exist $\{y_{i,1}, \ldots, y_{i,S_i}\}$ where $\forall j : y_{ij} \in \{0, 1\}$, only the bag-level label $y_i \in \{0, 1\}$ is observed during training. The standard MIL assumption (Dietterich et al., 1997; Raff & Holt, 2023) is that a bag is negative if and only if *all* instances are negative $(y_i = 0 \Leftrightarrow \forall j \in \{1, \ldots, S_i\} : y_{ij} = 0)$ and positive if and only if *at least one* instance is positive $(y_i = 1 \Leftrightarrow \exists j \in \{1, \ldots, S_i\} : y_{ij} = 1)$.

MIL architectures typically consist of three parts. First, an encoder $f(\cdot) : \mathbb{R}^{C\times H\times W} \to \mathbb{R}^M$ maps each C-channel 2D image $x_{ij} \in \mathbb{R}^{C\times H\times W}$ to an instance-level embedding $h_{ij} = f(x_{ij}) \in \mathbb{R}^M$. Second, a pooling operation $\sigma(\cdot) : \mathbb{R}^{M\times S_i} \to \mathbb{R}^M$ maps the set of instance-level embeddings $h_i = \{h_{i,1}, \ldots, h_{i,S_i}\}$ to a single bag-level embedding $z_i = \sigma(h_i) \in \mathbb{R}^M$. Finally, a classifier $g(\cdot) : \mathbb{R}^M \to [0, 1]$ maps the bag-level embedding to a predicted probability $\hat{y}_i = g(z_i) \in [0, 1]$.

Permutation-invariant pooling. Early MIL methods (Pinheiro & Collobert, 2015; Zhu et al., 2017; Feng & Zhou, 2017) used simple, non-learnable pooling operations to aggregate instance-level embeddings. Ilse et al. (2018) proposed learnable attention-based MIL (ABMIL) to be both data-driven

and interpretable. Extensions to ABMIL have maintained permutation-invariance (e.g., CLAM (Lu et al., 2021), DSMIL (Li et al., 2021), MAD-MIL (Keshvarikhojasteh et al., 2024)).

Correlated MIL. Permutation-invariance is inappropriate in cases with clear spatial dependencies across instances, such as bags of patches drawn from the same high-resolution 2D image or bags of 2D slices from a 3D volume. To address this, recent work has proposed *correlated MIL* (Shao et al., 2021) where transformer-based MIL methods (e.g., TransMIL (Shao et al., 2021), CAMIL (Fourkioti et al., 2024)) use multi-head self-attention to incorporate dependencies between instances.

Smooth operator. To improve localization while accounting for dependencies, Castro-Macías et al. (2024) proposed a smoothing operator that smoothes the learned embeddings of neighboring instances. Their module achieved state-of-the-art localization results among all tested MIL methods on 3D neuroimage and whole slide imaging (WSI) classification tasks.

Centered Gaussian. Harvey et al. (2026) raised questions about the quality of modern MIL for localization in brain scans. They examined a simple baseline: a discretized Gaussian distribution centered on the middle 2D slice of a 3D scan, $a_{ij} \propto \mathrm{NormPDF}(j|\frac{S_i}{2}, 1)$. This baseline outperformed the smoothing operator and other MIL at localizing in 3D neuroimages.

Other regularization for attention. In MIL for WSI, prior works have proposed regularization techniques for learned attention weights. Zhang et al. (2025)'s *attention entropy maximization (AEM)* subtracts an entropy term on the attention weights to penalize excessive concentration. Similarly, Sharma et al. (2021)'s *cluster-to-conquer* penalizes the Kullback-Leibler (KL) divergence between attention weights of patches in a cluster and a uniform distribution. Chraki et al. (2026)'s *counterfactual intervention for attention (CIA)* encourages differences in bag predictions across learned and random attention. Peled et al. (2026) integrate 2D spatial context into pairwise multi-head attention through a learnable distance-decayed prior. Unlike these works, our Normal Guidance encourages a univariate learned attention distribution to follow a bell-shaped curve.

Alternatives to attention. Instead of the bag-level focus of many MIL methods, Javed et al. (2022)'s *additive MIL* produces instance-level logits which are combined into a bag prediction via an unweighted average. A similar logit construction with a weighted average underlies Early et al. (2024)'s *conjunctive MIL*. Both argue these logits provide direct instance-level interpretability.

For further expanded discussion of related work, see App. H.

3 Methods

3.1 Guided Attention

Our guided attention framework seeks to regularize the learned attention weights a_i over the S_i instances in bag i by penalizing their divergence from a reference distribution r_i. Both a_i and r_i are non-negative and sum to one. The guided objective that we minimize during training is defined as

$$\mathcal{L}_{\text{Guided}} := \frac{1}{N}\sum_{i=1}^{N} \ell^{\text{BCE}}(y_i, \hat{y}_i) + \lambda D(r_i, a_i), \tag{1}$$

where ℓ^{BCE} is the standard binary cross entropy loss, $\lambda \geq 0$ is a hyperparameter that controls regularization strength, and $D(r_i, a_i)$ is a divergence measure of the "distance" between the reference distribution and the attention weights. To construct the reference r_i, we recommend the Normal Guidance procedure in Sec. 3.2, which takes as input the vector a_i. We apply $D(r_i, a_i)$ to all bags (positive and negative). This is mostly for simplicity, but also to avoid shortcut learning from wider attention in negative bags (App. G.4). We leave designing an r_i that depends on i's bag label to future work. Beyond that, we could separately use instance-level labels to inform the reference r_i (see App. G.1), but this requires available labels for all N bags in the train set and breaks the assumptions of MIL.

Among possible choices for divergence $D(\cdot)$, we study three variants with different behavior:

- **Squared error.** $D(r_i, a_i) = \sum_{j=1}^{S_i}(r_{ij} - a_{ij})^2$ encourages a_i to match r_i symmetrically.
- **Forward KL (mass-covering).** The forward KL $D_{\text{KL}}(r_i \| a_i) = \sum_{j=1}^{S_i} r_{ij} \log \frac{r_{ij}}{a_{ij}}$ is zero-avoiding with respect to a_i, resulting in *mass-covering* behavior.

- **Reverse KL (mode-seeking).** The reverse KL $D_{\mathrm{KL}}(a_i \| r_i) = \sum_{j=1}^{S_i} a_{ij} \log \frac{a_{ij}}{r_{ij}}$ is zero-forcing with respect to a_i, resulting in *mode-seeking* behavior. See Murphy (2012) for more details.

These variants are compared side-by-side in Fig. 1. Forward KL asymmetrically penalizes most if a_{ij} is small and r_{ij} is large; while Reverse KL penalizes most when a_{ij} is large and r_{ij} is small.

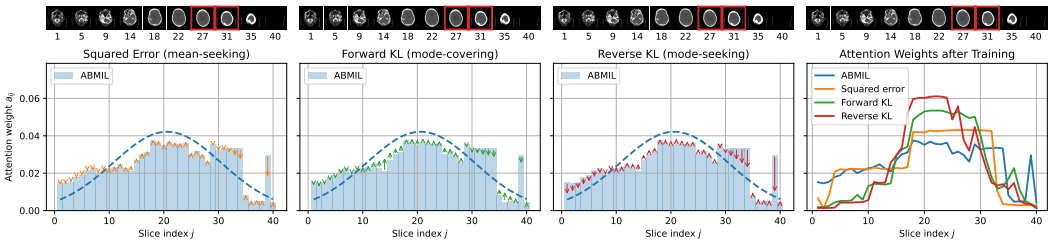

Figure 1: Normal Guidance on one 3D CT scan under different divergences D. *Left panels*: using squared error (orange), forward KL (green), and reverse KL (red). Arrows indicate direction of change for each attention weight a_{ij} to reduce $D(\cdot)$ via a gradient step; line length indicates magnitude of the regularization term for each attention weight. *Right*: Attention weights a_i after training with standard ABMIL or with Normal Guidance. Normal Guidance improves the distribution's unimodality and focus, though recovering the true ROI (red region in top bar) remains imperfect.

3.2 Normal Guidance

In many *correlated MIL* tasks, instances are spatially or temporally dependent. In 3D medical imaging, if a tumor or lesion is visible in one slice, it is likely also present in adjacent slices. To apply this idea in our guided attention framework, we propose Normal Guidance. First, we get the empirical mean and variance of the attended slice position from the current attention distribution a_i:

$$\mathbb{E}[J] = \sum_{j=1}^{S_i} j \cdot a_{ij}, \qquad \mathrm{Var}(J) = \left(\sum_{j=1}^{S_i} j^2 \cdot a_{ij}\right) - \left(\sum_{j=1}^{S_i} j \cdot a_{ij}\right)^2. \tag{2}$$

Next, these values parameterize a Normal PDF that we evaluate at each position j from 1 to S_i and renormalize to obtain a bell-shaped reference distribution r_i used in the loss in Eq. (1):

$$\hat{r}_{ij} = \mathrm{NormPDF}(j \mid \mathbb{E}[J], \mathrm{Var}(J)), \qquad r_{ij} = \frac{\hat{r}_{ij}}{\sum_{k=1}^{S_i} \hat{r}_{ik}}. \tag{3}$$

During gradient descent, we recompute r_i before each backward pass and apply the stop-gradient operation on r_i so that updates improve the predicted a_i's agreement with r_i.

We hypothesize this construction has 3 useful inductive biases: smoothness, center-focus, and unimodality. In our later CT tasks, positive slices are more likely to occur close to the middle position (see Fig. 3) and most scans have just one contiguous region of interest or ROI (64% of Chest CT, over 84% for Head and Abdomen; see Tab. A.2). Thus, the way NG favors smooth unimodal attention near the center is desirable for our 3D CT tasks. NG's inductive biases are discussed further in App. G.2. We also provide evidence that NG is not outperformed by more flexible guidance that allows multimodal attention via a mixture of Gaussian processes (App. G.3).

3.3 Multi-Head Normal Guidance

While Normal Guidance incorporates local dependencies, it encourages learning *unimodal* attention weights. This can be restrictive in correlated MIL settings where multiple disjoint regions are relevant (e.g., multiple tumors across non-adjacent slices). In such cases, a single normal distribution may over-smooth across regions or focus on only one region.

To overcome this unimodal restriction, we extend our method to Multi-Head Normal Guidance. For multi-head self-attention, we use a separate unimodal reference distribution via Eq. (3) for each head of the final self-attention layer (see Fig. 2). The overall regularization term is computed by averaging the divergence across all heads. Because each head is guided toward a separate bell-shaped curve, the combined effect can be intuitively thought of as guided by a Gaussian mixture model. This extension allows the network to simultaneously attend to multiple spatially separate regions.

4 EXPERIMENTS

4.1 DATASETS

Head CT. The RSNA 2019 Brain CT Hemorrhage Challenge (Flanders et al., 2020) includes 752,803 slices from 21,744 CT scans (20-60 slices per scan) with binary labels for any intracranial hemorrhage. We partition the released training set (no other release has labels) into train, validation, and test sets as described below. Castro-Macías et al. (2024) use a smaller subset (1,150 scans total).

Chest CT. The RSNA Pulmonary Embolism CT dataset (Colak et al., 2021) has 1,790,594 slices from 7,279 CT scans (63-1,083 slices per scan) with labels for presence/absence of any pulmonary embolism. We partition the released train set (no other release has labels) as described below.

Abdomen CT. The RSNA 2023 Abdominal Trauma AI Challenge (Hermans et al., 2024) has 1,500,653 slices from 4,711 CT scans (41-1,727 slices per scan) labeled for presence of any abdominal trauma. We partition the released train set (no other release has labels) as described below.

Partition by patient. For each medical dataset (Head CT, Chest CT, and Abdomen CT), we randomly assign 3D scans into training, validation, and testing sets using a 4:1:1 ratio, ensuring each patient's data belongs to exactly one set to avoid leakage. We stratify by class to ensure comparable class frequencies. We repeat this process across three random seeds, so every learning algorithm is evaluated on the same 3 distinct partitions into training, validation, and test sets.

Expert annotations. All 3 CT scan datasets include a binary label for each whole CT scan *and* an instance-level binary label for each 2D axial slice in every scan. The annotators were 60 neuroradiologists for Head CT; more than 80 thoracic radiologists for Chest CT; and radiologists with subspecialty training or professional experience in abdominal trauma for Abdomen CT. In our work, slice-level labels are only used to evaluate localization and train best-in-class ceilings or post-hoc localization-only methods.

Semi-synthetic. The *Shifted Mean MIL* dataset (Harvey et al., 2025; 2026) was created to stress-test correlated MIL: success requires reasoning across adjacent instances. Using prior work's generative process, we draw $y_i \sim \text{Bern}(0.5)$ so roughly 50% of bags are positive, and draw S_i uniformly so bags have 20 to 60 instances. If a bag is negative, all M features of all S_i instances are drawn from a standard Gaussian (mean=0, var.=1). If a bag is positive, most features come from the same standard Gaussian. We select a contiguous block of R=12 instances uniformly, then draw only the first of M features in this block from a standard Gaussian with mean shifted to 0.5. We call this *semi*-synthetic because the S_i range and R value are set to match the Head CT dataset. The number of features M=768 is set to match the ViT embedding size we use. We sample 10,000/2,500/1,000 bags for training/validation/testing. We repeat this process with three random seeds; each seed samples a different training, validation, and test set. Further details are in App. F.

4.2 EXPERIMENTAL GOALS, METRICS, BASELINES, AND TRAINING DETAILS

We are primarily interested in assessing the *localization performance* of MIL methods trained with weak supervision. Following past work (Castro-Macías et al., 2024), we measure this performance only on positive-labeled bags in the test set. For each bag i such that $y_i = 1$, at all S_i slices indexed by j we assess the predicted attention value a_{ij} versus the ground truth label y_{ij} via the AUROC,

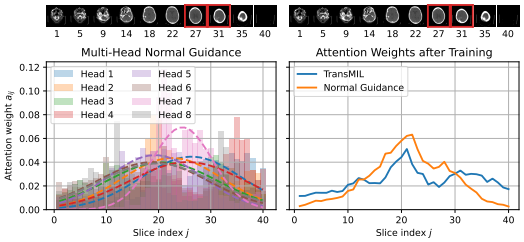

Figure 2: Multi-Head Normal Guidance on 1 example Head CT scan. Left: reference distribution for each attention head in TransMIL (we set $H = 8$ throughout). Right: Mean attention weights a_i (averaged over heads) after training, for standard TransMIL and TransMIL plus Multi-Head NG.

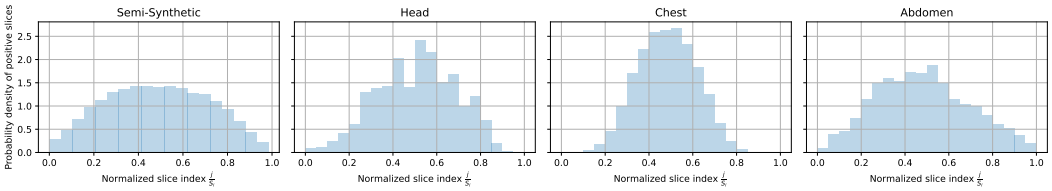

Figure 3: Probability density of positive slices in training set. Because slices in the middle of the scan are more likely to be positive, the centered Gaussian performs well for localization.

using macro averaging over bags. While our focus is localization, we also wish to verify that scan-level classification remains strong. We thus report scan-level ("bag"-level) classification quality in terms of AUROC. For both localization and classification, other metrics like AUPRC are reported in App. B. In all tables, the primary reported number is the *average* over the 3 test partitions of each task. The number after the "±" symbol indicates the empirical standard deviation of the 3 values.

Baselines. Among MIL methods, we consider max pooling MIL, mean pooling MIL, and several variants of attention-based MIL (ABMIL; Ilse et al., 2018) and transformer-based MIL (TransMIL; Shao et al., 2021). We specifically compare to different attention regularization strategies, including Smooth Operator (Castro-Macías et al., 2024, also known as SmAP or SmTP), AEM (Zhang et al., 2025) and CIA (Chraki et al., 2026). We also compare to Additive MIL (Javed et al., 2022). We chose not to try CAMIL (Fourkioti et al., 2024), DSMIL (Li et al., 2021), and CLAM (Lu et al., 2021) because of their poor localization results as reported in Castro-Macías et al.; Smooth Operator outperformed these methods in all tested localization tasks.

Some baselines are specific to localization and cannot produce scan-level predictions. The Centered Gaussian (CG) baseline (Harvey et al., 2026) predicts a centered bell-shaped curve across slices for every scan, uninformed by image content. We further try several post-hoc ways to estimate instance-level scores given a trained ABMIL model. See App. E.2 for more baseline details.

For our main evaluations of Normal Guidance, for simplicity we set D to the forward KL and fix regularization strength $\lambda = 1$; we conduct a sensitivity analysis of these choices later in Fig. 4. We configure baselines following past work, with reproducible details in App. E.

Common arch. For all 3D CT tasks, we use a *frozen* ViT-B/16 (Dosovitskiy et al., 2021) pre-trained on ImageNet (Deng et al., 2009) with embedding size $M{=}768$ as the raw feature encoder. Harvey et al. (2026) found this ViT to be competitive with other frozen encoders (ConvNeXt or MedSAM). Their frozen ViT MIL pipeline was also competitive with fine-tuned 3D CNNs yet faster to train. For all methods and tasks, we feed a final scan-level embedding to a linear binary classifier head.

Training. All methods are optimized using mini-batch stochastic gradient descent (SGD) with a momentum parameter of 0.9 and batch size of 64. We train for 1,000 epochs and use bag-level validation AUROC for early stopping. For all methods, we grid search two key hyperparameters: learning rate in {0.1, 0.01, 0.001, 0.0001} and L1 regularization strength for all pooling and classifier weights in {1.0, 0.1, 0.01, 0.001, 0.0001, 1e-5, 1e-6, 0.0}.

4.3 Best-in-Class Ceilings for Instance-Level and Bag-Level Classification

Here we develop a "best-in-class ceiling" for both localization and classification tasks. Each ceiling is an estimate of what a well-fit MIL model trained with unfair access to slice-level labels is capable of while enforcing other key modeling assumptions (in our case, linear classification head and frozen ViT embeddings). Ceilings are approximate; they are not strict upper bounds that can never be exceeded.

Localization ceiling for CT. We train an instance-level classifier using ground truth instance-level labels y_{ij}. This classifier takes as input for instance index j a local window of R consecutive embedding vectors centered on h_{ij} and outputs an instance-level predicted probability $\hat{y}_{ij}$. We use a 1D convolutional layer with kernel size R over instance representations provided by the frozen ViT, followed by a linear classifier head. For each dataset, we set R equal to the mean number of

consecutive positive instances, allowing gaps of up to 3 negative instances in a "consecutive" block. This sets R=12 for Head CT, R=35 for Chest CT, and R=29 for Abdomen (see Tab. A.2).

This instance-level ceiling consumes a neighborhood of embeddings to account for spatial dependencies in correlated MIL. In contrast, past upper bounds for MIL (Correia et al., 2020; Guillaumin et al., 2010) have handled each instance separately.

Bag-level ceiling for CT. For "best possible" bag-level MIL, we use ground-truth per-slice labels y_{ij} to pool embeddings, instead of learned attention. This of course has an unfair advantage compared to weakly-supervised MIL since it relies on instance-level annotations *even at test time* and not just images alone. Procedurally, for each positive bag, we pool instance-level embeddings using a uniform distribution over that bag's positively labeled instances. For each negative bag, we randomly sample a positive bag and use its instance-level positive labels as the attended ROI, adjusting for differing S_i counts via interpolation. We then pool the negative bag's embeddings uniformly over this ROI. We found enforcing similar attended ROI distributions across positive and negative bags was key to prevent this ceiling classifier from learning "shortcuts" that made the task too easy (App. G.4).

Localization ceiling for semi-synth. Our ceiling method computes $p(y_{ij}{=}1|h_i, S_i, y_i{=}1)$ using the generative process for *Shifted Mean* data. This is the probability that instance j comes from a Gaussian with *shifted mean* given its features. The event $y_{ij}{=}1$ is equivalent to index j falling in the chosen latent block of R consecutive instances. We provide a closed-form expression in App. F. This oracle upper bound for instance-level performance has not been derived in prior work.

Bag ceiling for semi-synthetic. Since the true data-generating process is known, we use the Bayes estimator derived in Eq. (10) of Harvey et al. (2026) as an upper bound on bag-level performance.

5 Results

5.1 Localization performance

Localization results are in Tab. 1, B.1, and B.3. Major findings are summarized in the text below.

The Centered Gaussian baseline outperforms existing MIL for localization. In terms of localization AUROC, the simple CG baseline from Harvey et al. (2026) clearly outperforms all tested MIL methods except NG on all 3 CT tasks (confirmed via bootstrap significance tests, see App. B.4). For Chest CT localization, no existing MIL except CIA scores above 0.69 AUROC while the CG baseline scores 0.85! When we evaluate AUPRC (Tab. B.1), and precision at 10% recall (Tab. B.3), the CG baseline beats or at least statistically ties other MIL methods: CG's wins come from gains in recall that do not sacrifice precision and sometimes improve precision.

Table 1: Localization AUROC on test set, aka slice-level or instance-level classification, using positive bags only. Takeaway: On real CT images, Normal Guidance outperforms all other MIL and is the only MIL to beat or tie the image-free Centered Gaussian baseline.

Method	Semi-Synthetic	Head CT	Chest CT	Abdomen CT
Centered Gaussian	0.688±0.014	0.852±0.001	0.850±0.001	0.642±0.022
Max pooling	0.504±0.002	0.286±0.004	0.461±0.003	0.453±0.011
Mean pooling	0.500±0.000	0.500±0.000	0.500±0.000	0.500±0.000
Additive MIL	0.540±0.057	0.764±0.011	0.725±0.022	0.579±0.044
ABMIL	0.546±0.054	0.736±0.052	0.558±0.086	0.526±0.041
post-hoc Blend a_i with CG	0.692±0.016	0.845±0.016	0.809±0.028	0.631±0.024
post-hoc Each slice as bag	0.640±0.004	0.785±0.004	0.639±0.167	0.581±0.003
post-hoc Gauss. kernel smoothing	0.608±0.122	0.778±0.058	0.566±0.097	0.525±0.036
Attn. Entropy Max. (AEM)	0.542±0.058	0.749±0.031	0.563±0.076	0.525±0.037
Smooth Operator	0.693±0.030	0.831±0.003	0.688±0.043	0.564±0.012
Counterfactual Intervention (CIA)	0.526±0.037	0.767±0.010	0.689±0.097	0.622±0.022
Normal Guidance (ours)	0.541±0.055	**0.871±0.002**	0.866±0.005	**0.663±0.006**
TransMIL	0.593±0.131	0.792±0.031	0.585±0.077	0.579±0.003
Smooth Operator	0.592±0.127	0.823±0.019	0.654±0.035	0.582±0.043
Multi-Head Norm. Guidance (ours)	**0.706±0.076**	0.869±0.002	**0.869±0.002**	0.634±0.007
Best-in-Class Ceiling (see Sec. 4.3)	0.884±0.004	0.896±0.003	0.871±0.004	0.717±0.011

Normal Guidance can outperform this baseline: a naive center bias is not always best. Our Normal Guidance (NG) and Multi-Head Normal Guidance (MHNG) deliver significant gains in localization AUROC over the CG baseline on Head and Chest CT tasks (confirmed via bootstrap tests, App. B.4); on Abdomen CT there is a statistical tie. This reassuringly indicates that learning attention from images with proper steering can outperform the image-agnostic CG baseline.

5.2 Bag-level classification performance

Bag-level (scan-level) classification is reported in Tab. 2 and B.2. Recall that Normal Guidance is specifically designed to improve localization. Here our goal is to verify how this regularization impacts bag-level classification and interpret how methods compare to our new ceilings.

Normal Guidance keeps scan-level classification competitive. Across all tested CT tasks, our NG and MHNG score better than or within tolerance (within 0.01 AUROC) of their corresponding unregularized or smoothed alternative. On semi-synthetic bag-level classification, ABMIL + Smooth Operator scores highest (0.771) but TransMIL with MHNG is competitive at 0.766.

On two CT tasks, the best MIL methods are near the bag-level ceiling. The best MIL method achieves very good results on Head CT (within 0.002 AUROC of best-in-class ceiling) and good results on Chest CT (within 0.022 AUROC of the best-in-class ceiling). For the semi-synthetic dataset, the best MIL method is within 0.042 AUROC of its ceiling. For the Abdomen CT task, the gap of about 0.1 AUROC may be real. Yet we guess the ceiling is overestimated; this task has much more variability in the organs covered than other tasks, and 6.6% of scans do not fully cover one or more organs. Using instance-level labels from an incomplete positive scan to pool a complete negative scan (or vice versa) could create shortcut features that overestimate performance.

Transformer-based MIL only slightly improves bag-level performance. Transformer-based methods improve AUROC by up to 0.05 on the Abdomen CT task over ABMIL-based counterparts, but otherwise don't seem to add appreciable gain on the other 3 tasks.

Table 2: Bag-level AUROC on test set for whole-scan classification. Takeaway: Normal Guidance is competitive with state-of-the-art MIL alternatives here, while improving localization (Tab. 1).

Method	Semi-Synthetic	Head CT	Chest CT	Abdomen CT
Max pooling	$0.624_{\pm 0.009}$	$0.888_{\pm 0.009}$	$0.656_{\pm 0.004}$	$0.643_{\pm 0.029}$
Mean pooling	$0.752_{\pm 0.002}$	$0.920_{\pm 0.012}$	$0.669_{\pm 0.018}$	$0.625_{\pm 0.027}$
Additive MIL	$0.749_{\pm 0.004}$	$0.921_{\pm 0.009}$	$0.663_{\pm 0.027}$	$0.631_{\pm 0.023}$
ABMIL	$0.751_{\pm 0.003}$	$0.919_{\pm 0.009}$	$0.664_{\pm 0.017}$	$0.639_{\pm 0.044}$
Attn. Entropy Max. (AEM)	$0.751_{\pm 0.003}$	$0.921_{\pm 0.010}$	$0.545_{\pm 0.091}$	$0.642_{\pm 0.046}$
Smooth Operator	$0.771_{\pm 0.008}$	$0.925_{\pm 0.012}$	$0.671_{\pm 0.023}$	$0.648_{\pm 0.043}$
Counterfactual Intervention (CIA)	$0.751_{\pm 0.003}$	$0.920_{\pm 0.007}$	$0.657_{\pm 0.019}$	$0.620_{\pm 0.008}$
Normal Guidance (ours)	$0.751_{\pm 0.003}$	$0.925_{\pm 0.010}$	$0.678_{\pm 0.026}$	$0.651_{\pm 0.026}$
TransMIL	$0.763_{\pm 0.016}$	$0.925_{\pm 0.014}$	$0.663_{\pm 0.023}$	$0.684_{\pm 0.047}$
Smooth Operator	$0.768_{\pm 0.025}$	$0.926_{\pm 0.013}$	$0.670_{\pm 0.015}$	$0.664_{\pm 0.050}$
Multi-Head Normal Guidance (ours)	$0.766_{\pm 0.006}$	$0.926_{\pm 0.012}$	$0.662_{\pm 0.015}$	$0.677_{\pm 0.029}$
Best-in-Class Ceiling (see Sec. 4.3)	$0.810_{\pm 0.006}$	$0.927_{\pm 0.012}$	$0.700_{\pm 0.014}$	$0.776_{\pm 0.033}$

5.3 Sensitivity Analysis

Our Normal Guidance procedure's success could depend on the choice of divergence D and the regularization strength λ. Thus far, these values have been fixed as forward KL and $\lambda = 1$. The two panels of Fig. 4 show how performance on the Head CT task depends on these variables. First, the left panel of Fig. 4 suggests a modest preference for Forward KL for both localization and whole-scan classification. The right panel of Fig. 4 suggests that $\lambda{=}10^0{=}1$ is a reasonable choice for localization, but we might do slightly better by tuning λ. Thus, the localization success of NG and MHNG reported in Sec. 5.1 might be improvable. We further examine alternative designs for the mean and variance parameters for NG in App. D; we find most reasonable choices deliver comparably good performance for both bag-level and localization tasks.

Divergence	Localization	Classification
Squared error	$0.834_{\pm 0.007}$	$0.924_{\pm 0.012}$
Forward KL	$0.871_{\pm 0.002}$	$0.925_{\pm 0.010}$
Reverse KL	$0.869_{\pm 0.003}$	$0.919_{\pm 0.011}$

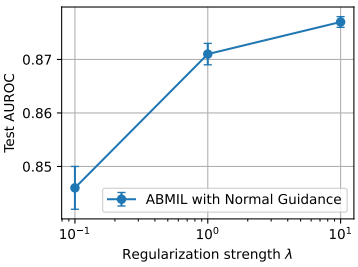

Figure 4: Sensitivity analysis. *Left:* Effect of divergence D on localization and bag-level classification, with $\lambda = 10^0$. *Right:* Effect of regularization strength λ on localization (D: Forward KL). All results here report test set AUROC on the Head CT task, using ABMIL with Normal Guidance.

5.4 Qualitative Side-by-Side Comparisons of Learned Attention

See App. C for visual comparisons of the learned attention distributions from different MIL methods on a handful of randomly-selected CT scans from each of our 3 datasets. In each figure, we either compare ABMIL, ABMIL with Sm, and ABMIL with NG; or TransMIL, TransMIL with Sm, and TransMIL with MHNG. As a primary takeaway, we can visually confirm that Normal Guidance often produces more *focused* and unimodal attention distributions than its counterparts. Some plots, especially the right panels of Fig. C.3, show cases where all methods fail to capture the key block of positive slices, indicating plenty of room for further localization method development.

6 Conclusion

The attention mechanism was originally proposed in MIL as a method that incorporates interpretability (Ilse et al., 2018). Yet this work shows that for localization in 3D medical images, MIL alone cannot outperform a simple baseline due to a lack of inductive bias. We proposed Normal Guidance, a regularization technique for attention-based MIL models that encourages attention to be bell-curve shaped. We investigated three CT datasets totaling over 4 million 2D slices and show that Normal Guidance outperforms all other attention-based MIL methods as well as the simple slice-level baselines, achieving a new state-of-the-art for instance-level MIL results on all datasets while maintaining high bag-level performance.

Limitations. (1) We only use positive bags to evaluate slice-level localization, following past work (Castro-Macías et al., 2024). We do not evaluate the quality of attention in negative bags. This highlights an open problem in the literature: defining what attention should look like for a negative bag. (2) Jain & Wallace (2019) argue that attention weights should not be treated as faithful explanations of classifier decisions, since permuting attention can yield similar predictions. Our results show that Normal Guidance can help learned attention track expert annotations more closely, but this does not by itself enforce that the attended slices are causally responsible for the bag-level prediction. (3) We focus on deep MIL with frozen encoders and linear classifiers, again following past work (Lu et al., 2021; Shao et al., 2021; Fourkioti et al., 2024; Castro-Macías et al., 2024). Parameter-efficient fine-tuning could deliver more effective encoders for this domain, but at prohibitive computational cost. Even with frozen encoders, model training and hyperparameter optimization takes roughly 9 hours for each TransMIL run on the Head CT dataset (not to mention repeating this across 32 hyperparameter configurations and 3 train/test partitions).

Future work. Beyond 3D images, MIL is often used for whole slide images. Alternative inductive biases that account for the 2D spatial dependencies in WSI are critical for understanding localization results in that domain. A second direction is conditioning r_i on the bag label y_i: the right reference (Gaussian, Uniform, etc.) for negative bags is worth investigating. Finally, the same bell-shaped prior (NG) may also suit MIL tasks beyond medical imaging, such as video anomaly detection, where evidence is typically contiguous in time (Sultani et al., 2018).

AI USE STATEMENT

We authors take responsibility for the final content of this work, including all text, all claims, and all artifacts. In addition to an extensive manual effort to review relevant literature, we have used generative AI tools for retrieval and discovery of related work. Suggestions were carefully reviewed by human authors. Citations were managed via a manual process (individually downloading bibtex info from trusted sources like Google Scholar, then manually edited) to ensure up-to-date accurate information.

For the experiments featured in the main paper, we have not used AI tools for coding or experiment execution. For some experiments in App. G.3, we did use AI tools to assist with rapid code prototyping, but high-level ideas and intent came from human authors and the results were always checked by a human author.

The first draft of all text in this paper was directly written by human authors. We have used generative AI tools for minor polishing (rephrasing to balance needs for high-quality description and brevity, etc.). The final text in this manuscript was carefully edited, and approved by human authors.

REPRODUCIBILITY STATEMENT

Anonymous code for repeating our experiments can be found via the URL at the bottom of page 1. Details about experiments, including hyperparameter searches and required hardware, are in App. E. Details about preprocessing of CT datasets are in App. A.1.

REFERENCES

Hassan Baker and Austin J. Brockmeier. Patch2Loc: Learning to Localize Patches for Unsupervised Brain Lesion Detection. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2026.

Finn Behrendt, Debayan Bhattacharya, Julia Krüger, Roland Opfer, and Alexander Schlaefer. Patched Diffusion Models for Unsupervised Anomaly Detection in Brain MRI. In *Medical Imaging with Deep Learning*, 2024.

Francisco M. Castro-Macías, Pablo Morales-Álvarez, Yunan Wu, Rafael Molina, and Aggelos K. Katsaggelos. `Sm`: enhanced localization in Multiple Instance Learning for medical imaging classification. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Imane Chraki, Pierre Marza, Stergios Christodoulidis, and Maria Vakalopoulou. Counterfactual intervention in attention multiple instance learning for digital pathology. In *Medical Imaging with Deep Learning (MIDL)*, 2026.

Errol Colak, Felipe C. Kitamura, Stephen B. Hobbs, Carol C. Wu, Matthew P. Lungren, Luciano M. Prevedello, Jayashree Kalpathy-Cramer, Robyn L. Ball, George Shih, Anouk Stein, Safwan S. Halabi, Emre Altinmakas, Meng Law, Parveen Kumar, Karam A. Manzalawi, Dennis Charles Nelson Rubio, Jacob W. Sechrist, Pauline Germaine, Eva Castro Lopez, Tomas Amerio, Pushpender Gupta, Manoj Jain, Fernando U. Kay, Cheng Ting Lin, Saugata Sen, Jonathan Wesley Revels, Carola C. Brussaard, and John Mongan For the RSNA-STR Annotators and Dataset Curation Contributors. The RSNA pulmonary embolism CT dataset. *Radiology: Artificial Intelligence*, 3 (2):e200254, 2021.

Joana Correia, Isabel Trancoso, and Bhiksha Raj. Automatic in-the-wild dataset annotation with deep generalized multiple instance learning. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 2020.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A Large-Scale Hierarchical Image Database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

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.

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 (ICLR)*, 2021.

Joseph Early, Christine Evers, and Sarvapali Ramchurn. Model agnostic interpretability for multiple instance learning. In *International Conference on Learning Representations (ICLR)*, 2022.

Joseph Early, Gavin Cheung, Kurt Cutajar, Hanting Xie, Jas Kandola, and Niall Twomey. Inherently Interpretable Time Series Classification via Multiple Instance Learning. In *International Conference on Learning Representations*, 2024. URL https://mlanthology.org/iclr/2024/early2024iclr-inherently/.

Ji Feng and Zhi-Hua Zhou. Deep MIML Network. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2017.

Adam E. Flanders, Luciano M. Prevedello, George Shih, Safwan S. Halabi, Jayashree Kalpathy-Cramer, Robyn Ball, John T. Mongan, Anouk Stein, Felipe C. Kitamura, Matthew P. Lungren, Geetika Choudhary, Luciano Cala, Luís Coelho, Mads Mogensen, Fátima Morón, Eric Miller, Ichiro Ikuta, Vahe Zohrabian, Oran McDonnell, Christoph Lincoln, Luciano Shah, Devon Joyner, Ashish Agarwal, Richard K. Lee, and Jayashree Nath. Construction of a Machine Learning Dataset through Collaboration: The RSNA 2019 Brain CT Hemorrhage Challenge. *Radiology: Artificial Intelligence*, 2(3), 2020.

Olga Fourkioti, Matt De Vries, and Chris Bakal. CAMIL: Context-Aware Multiple Instance Learning for Cancer Detection and Subtyping in Whole Slide Images. In *International Conference on Learning Representations (ICLR)*, 2024.

Matthieu Guillaumin, Jakob Verbeek, and Cordelia Schmid. Multiple Instance Metric Learning from Automatically Labeled Bags of Faces. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 2010.

Ethan Harvey, Dennis Johan Loevlie, and Michael C. Hughes. Synthetic Data Reveals Generalization Gaps in Correlated Multiple Instance Learning. *arXiv preprint arXiv:2510.25759*, 2025.

Ethan Harvey, Dennis Johan Loevlie, Amir Ali Satani, Wansu Chen, David M. Kent, and Michael C. Hughes. A Multi-Dataset Benchmark of Multiple Instance Learning for 3D Neuroimage Classification. In *Conference on Health, Inference, and Learning (CHIL)*, 2026.

Sebastiaan Hermans, Zixuan Hu, Robyn L. Ball, Hui Ming Lin, Luciano M. Prevedello, Ferco H. Berger, Ibrahim Yusuf, Jeffrey D. Rudie, Maryam Vazirabad, Adam E. Flanders, George Shih, John Mongan, Savvas Nicolaou, Brett S. Marinelli, Melissa A. Davis, Kirti Magudia, Ervin Sejdić, and Errol Colak. RSNA 2023 Abdominal Trauma AI Challenge: Review and Outcomes. *Radiology: Artificial Intelligence*, 7(1):e240334, 2024.

Zhe Huang, Benjamin S. Wessler, and Michael C. Hughes. Detecting Heart Disease from Multi-View Ultrasound Images via Supervised Attention Multiple Instance Learning. In *Machine Learning for Healthcare Conference (MLHC)*, 2023.

Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based Deep Multiple Instance Learning. In *International Conference on Machine Learning (ICML)*, 2018.

Sarthak Jain and Byron C. Wallace. Attention is not Explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019.

Syed Ashar Javed, Dinkar Juyal, Harshith Padigela, Amaro Taylor-Weiner, Limin Yu, and Aaditya Prakash. Additive MIL: Intrinsically interpretable multiple instance learning for pathology. In *Advances in Neural Information Processing Systems*, 2022.

Hassan Keshvarikhojasteh, Josien P. W. Pluim, and Mitko Veta. Multi-head Attention-based Deep Multiple Instance Learning. In *Proceedings of the MICCAI Workshop on Computational Pathology*, 2024.

Bin Li, Yin Li, and Kevin W. Eliceiri. Dual-stream Multiple Instance Learning Network for Whole Slide Image Classification with Self-supervised Contrastive Learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021.

Kechun Liu, Wenjun Wu, Joann G. Elmore, and Linda G. Shapiro. Semantics-Aware Attention Guidance for Diagnosing Whole Slide Images. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2024.

Ming Y. Lu, Drew F. K. Williamson, Tiffany Y. Chen, Richard J. Chen, Matteo Barbieri, and Faisal Mahmood. Data efficient and weakly supervised computational pathology on whole slide images. *Nature Biomedical Engineering*, 5(6):555–570, 2021.

Oded Maron and Tomás Lozano-Pérez. A Framework for Multiple-Instance Learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 1997.

Matthew B McDermott, Haoran Zhang, Lasse H Hansen, Giovanni Angelotti, and Jack Gallifant. A closer look at AUROC and AUPRC under class imbalance. In *Advances in Neural Information Processing Systems*, 2024.

Kevin P. Murphy. *Machine Learning: A Probabilistic Perspective*, chapter 21.2.2 Forward or reverse KL? MIT Press, 2012.

John Muschelli. Recommendations for processing head CT data. *Frontiers in Neuroinformatics*, 13: 61, 2019.

Sharon Peled, Yosef E Maruvka, and Moti Freiman. PSA-MIL: A Probabilistic Spatial Attention-Based Multiple Instance Learning for Whole Slide Image Classification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2026.

Pedro O. Pinheiro and Ronan Collobert. From Image-Level to Pixel-Level Labeling With Convolutional Networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

Gwenolé Quellec, Guy Cazuguel, Béatrice Cochener, and Mathieu Lamard. Multiple-Instance Learning for Medical Image and Video Analysis. *IEEE Reviews in Biomedical Engineering*, 10:213–234, 2017.

Edward Raff and James Holt. Reproducibility in Multiple Instance Learning: A Case For Algorithmic Unit Tests. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Sebastian Raschka. Chapter 25: Confidence Intervals. In *Machine Learning Q and AI*. No Starch Press, 2024. URL https://sebastianraschka.com/books/ml-q-and-ai-chapters/ch25/.

Khaled Saab, Jared Dunnmon, Roger Goldman, Alex Ratner, Hersh Sagreiya, Christopher Ré, and Daniel Rubin. Doubly Weak Supervision of Deep Learning Models for Head CT. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2019.

Sanjai Sangvichien, Komon Boonkaew, Aporn Chuncharunee, Chulaluk Komoltri, Sitha Piyawinitwong, Alongkorn Wongsawut, and Suwalee Namwongsa. Sex determination in Thai skulls by using craniometry: multiple logistic regression analysis. *Siriraj Medical Journal*, 59(5):216–221, 2007.

Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, and Yongbing Zhang. TransMIL: Transformer based Correlated Multiple Instance Learning for Whole Slide Image Classification. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Yash Sharma, Aman Shrivastava, Lubaina Ehsan, Christopher A. Moskaluk, Sana Syed, and Donald Brown. Cluster-to-Conquer: A Framework for End-to-End Multi-Instance Learning for Whole Slide Image Classification. In *Medical Imaging with Deep Learning*, 2021.

Waqas Sultani, Chen Chen, and Mubarak Shah. Real-World Anomaly Detection in Surveillance Videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17, 2020. doi: 10.1038/s41592-019-0686-2.

Andrew Wilson and Ryan Adams. Gaussian process kernels for pattern discovery and extrapolation. In *Proceedings of the International Conference on Machine Learning*, pp. 1067–1075, 2013. URL https://proceedings.mlr.press/v28/wilson13.html.

Yunan Wu, Arne Schmidt, Enrique Hernández-Sánchez, Rafael Molina, and Aggelos K. Katsaggelos. Combining Attention-Based Multiple Instance Learning and Gaussian Processes for CT Hemorrhage Detection. In *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2021.

Tony Xu, Sepehr Hosseini, Chris Anderson, Anthony Rinaldi, Rahul G Krishnan, Anne L Martel, and Maged Goubran. A generalizable 3d framework and model for self-supervised learning in medical imaging. *npj Digital Medicine*, 8(1), 2025.

Shaolei Zhang and Yang Feng. Modeling Concentrated Cross-Attention for Neural Machine Translation with Gaussian Mixture Model. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2021.

Yunlong Zhang, Honglin Li, Yuxuan Sun, Zhongyi Shui, Jingxiong Li, Chenglu Zhu, and Lin Yang. AEM: Attention Entropy Maximization for Multiple Instance Learning based Whole Slide Image Classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, 2025.

Wentao Zhu, Qi Lou, Yeeleng Scott Vang, and Xiaohui Xie. Deep Multi-instance Networks with Sparse Label Assignment for Whole Mammogram Classification. In *International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)*, 2017.

A Dataset Details

A.1 Preprocessing of CT scans

For all medical imaging datasets (Head CT, Chest CT, and Abdomen CT), we convert images into Hounsfield Units (HU) using each image's rescale slope and intercept; resize each 2D slice to 224×224 pixels; and normalize images with the training set mean and standard deviation of each channel. For the Head CT dataset, we use intensity windowing to exclude the skull, other bones, and calcifications (only including -100 to 300 HU) (Muschelli, 2019).

A.2 Descriptive Statistics

Below, we include three tables that summarize the ground truth instance-level labels for our 3 real-world CT datasets:

- the number of contiguous blocks (regions of interest, aka ROIs) in a positive bag,
- the number of instances in a typical ROI block, and
- the fraction of all instances in the bag contained in a typical ROI block.

	1 block	**2 blocks**	**3 blocks**	**4 blocks**	**5 blocks**	**6 blocks**	**7 blocks**
Head CT	8296/8882	569/8882	17/8882	–	–	–	–
Chest CT	1437/2211	497/2211	185/2211	65/2211	16/2211	7/2211	4/2211
Abdomen CT	304/365	53/365	7/365	1/365	–	–	–

Table A.1: Number of ROI blocks in a positive bag, across real-world CT datasets. Given ground-truth instance-level labels for a bag, we define each ROI "block" as the largest possible set of consecutively indexed instances, beginning and ending with a positive label, that contains no more than 3 negative instance labels in a row.

	Mean	**Min**	**5th**	**25th**	**50th**	**75th**	**95th**	**Max**
Head CT	12	1	2	6	11	17	23	43
Chest CT	35	1	1	7	16	44	137	353
Abdomen CT	29	1	2	6	15	40	95	242

Table A.2: Number of positive instances in each ROI block. For a definition of ROI block, see above.

	Mean	**Min**	**5th**	**25th**	**50th**	**75th**	**95th**	**Max**
Head CT	0.350	0.017	0.053	0.172	0.342	0.513	0.688	1.000
Chest CT	0.093	0.001	0.004	0.023	0.059	0.137	0.283	0.499
Abdomen CT	0.119	0.002	0.006	0.027	0.063	0.174	0.401	0.870

Table A.3: Fraction of all instances in each ROI block. For a definition of ROI block, see above.

B Additional Results: AUPRC and Precision-at-Low-Recall

B.1 AUPRC Results

Table B.1: Localization results for the **AUPRC** metric, using the test set of each task. Corresponding AUROC results are in Tab. 1.

Method	Semi-Synthetic	Head CT	Chest CT	Abdomen CT
Centered Gaussian	$0.561_{\pm 0.011}$	$0.711_{\pm 0.007}$	$0.521_{\pm 0.008}$	$0.246_{\pm 0.036}$
Max pooling	$0.384_{\pm 0.003}$	$0.302_{\pm 0.007}$	$0.185_{\pm 0.002}$	$0.109_{\pm 0.002}$
Mean pooling	$0.332_{\pm 0.003}$	$0.359_{\pm 0.009}$	$0.188_{\pm 0.001}$	$0.104_{\pm 0.007}$
Additive MIL	$0.383_{\pm 0.073}$	$0.660_{\pm 0.004}$	$0.345_{\pm 0.027}$	$0.173_{\pm 0.025}$
ABMIL	$0.424_{\pm 0.045}$	$0.634_{\pm 0.049}$	$0.233_{\pm 0.059}$	$0.159_{\pm 0.027}$
post-hoc Blend a_i with CG	$0.577_{\pm 0.013}$	$0.725_{\pm 0.011}$	$0.495_{\pm 0.021}$	$0.245_{\pm 0.036}$
post-hoc Each slice as bag	$0.508_{\pm 0.005}$	$0.684_{\pm 0.011}$	$0.341_{\pm 0.128}$	$0.172_{\pm 0.009}$
post-hoc Gauss. kernel smoothing	$0.518_{\pm 0.122}$	$0.666_{\pm 0.071}$	$0.247_{\pm 0.076}$	$0.177_{\pm 0.037}$
Attn. Entropy Max. (AEM)	$0.404_{\pm 0.064}$	$0.641_{\pm 0.037}$	$0.230_{\pm 0.060}$	$0.151_{\pm 0.030}$
Smooth Operator	$0.574_{\pm 0.035}$	$0.713_{\pm 0.007}$	$0.344_{\pm 0.037}$	$0.167_{\pm 0.020}$
Counterfactual Intervention (CIA)	$0.390_{\pm 0.047}$	$0.669_{\pm 0.002}$	$0.333_{\pm 0.074}$	$0.217_{\pm 0.028}$
Normal Guidance (ours)	$0.403_{\pm 0.062}$	$\mathbf{0.744_{\pm 0.010}}$	$0.522_{\pm 0.008}$	$\mathbf{0.248_{\pm 0.020}}$
TransMIL	$0.487_{\pm 0.140}$	$0.659_{\pm 0.032}$	$0.258_{\pm 0.052}$	$0.191_{\pm 0.021}$
Smooth Operator	$0.467_{\pm 0.147}$	$0.689_{\pm 0.007}$	$0.309_{\pm 0.018}$	$0.195_{\pm 0.039}$
Multi-Head Normal Guidance (ours)	$\mathbf{0.578_{\pm 0.086}}$	$0.740_{\pm 0.010}$	$\mathbf{0.535_{\pm 0.005}}$	$0.227_{\pm 0.010}$
Best-in-Class Ceiling (see Sec. 4.3)	$0.825_{\pm 0.005}$	$0.781_{\pm 0.010}$	$0.528_{\pm 0.004}$	$0.285_{\pm 0.030}$

Table B.2: Bag-level **AUPRC** on test set for whole-scan classification. Takeaway: Normal Guidance is competitive with state-of-the-art MIL alternatives here, while improving localization (Tab. B.1). Corresponding AUROC results are in Tab. 2.

Method	Semi-Synthetic	Head CT	Chest CT	Abdomen CT
Max pooling	$0.616_{\pm 0.005}$	$0.859_{\pm 0.013}$	$0.440_{\pm 0.002}$	$0.133_{\pm 0.016}$
Mean pooling	$0.756_{\pm 0.007}$	$0.903_{\pm 0.011}$	$0.455_{\pm 0.005}$	$0.117_{\pm 0.010}$
Additive MIL	$0.753_{\pm 0.010}$	$0.906_{\pm 0.008}$	$0.463_{\pm 0.013}$	$0.117_{\pm 0.011}$
ABMIL	$0.755_{\pm 0.009}$	$0.903_{\pm 0.007}$	$0.451_{\pm 0.005}$	$0.122_{\pm 0.014}$
Attn. Entropy Max. (AEM)	$0.756_{\pm 0.009}$	$0.921_{\pm 0.010}$	$0.450_{\pm 0.005}$	$0.122_{\pm 0.012}$
Smooth Operator	$0.775_{\pm 0.010}$	$0.910_{\pm 0.011}$	$0.465_{\pm 0.008}$	$0.139_{\pm 0.014}$
Counterfactual Intervention (CIA)	$0.756_{\pm 0.007}$	$0.904_{\pm 0.007}$	$0.454_{\pm 0.018}$	$0.118_{\pm 0.006}$
Normal Guidance (ours)	$0.755_{\pm 0.010}$	$0.911_{\pm 0.008}$	$0.477_{\pm 0.015}$	$0.117_{\pm 0.008}$
TransMIL	$0.766_{\pm 0.019}$	$0.912_{\pm 0.012}$	$0.447_{\pm 0.013}$	$0.139_{\pm 0.025}$
Smooth Operator	$0.770_{\pm 0.025}$	$0.913_{\pm 0.012}$	$0.461_{\pm 0.009}$	$0.135_{\pm 0.007}$
Multi-Head Normal Guidance (ours)	$0.766_{\pm 0.015}$	$0.912_{\pm 0.012}$	$0.451_{\pm 0.010}$	$0.139_{\pm 0.012}$
Best-in-Class Ceiling (see Sec. 4.3)	$0.814_{\pm 0.002}$	$0.913_{\pm 0.010}$	$0.468_{\pm 0.004}$	$0.252_{\pm 0.081}$

Divergence	Localization	Classification
Squared error	$0.702_{\pm 0.011}$	$0.910_{\pm 0.010}$
Forward KL	$0.744_{\pm 0.010}$	$0.911_{\pm 0.008}$
Reverse KL	$0.747_{\pm 0.012}$	$0.905_{\pm 0.011}$

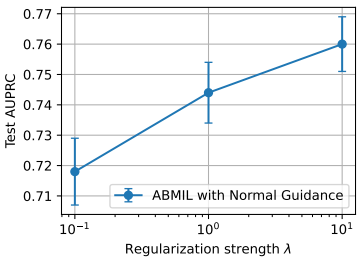

Figure B.1: Sensitivity analysis for the **AUPRC** metric on the test set. *Left:* Effect of divergence D on localization and bag-level classification, with $\lambda = 10^0$. *Right:* Effect of regularization strength λ on localization (D: Forward KL). All results from the Head CT task, using ABMIL with Normal Guidance.

Careful readers may notice that MHNG delivers slightly larger AUPRC numbers than the best-in-class ceiling on Chest CT. We interpret these results as essentially statistically tied, not that MHNG

is truly better. The ceiling is an approximate empirical estimate of an upper bound for simplistic MIL informed by slice-level labels, not a strict upper bound. Future work could consider a ceiling with multi-scale branches for convolutions, or even a supervised transformer, not just one fixed kernel-size convolutional filter.

B.2 Why favor AUROC over AUPRC?

Though we report both AUROC and AUPRC, we chose to highlight AUROC in the main manuscript. While folk wisdom and common practice today often claims that AUPRC is "better" for tasks with class imbalance, this claim has been recently refuted (McDermott et al., 2024). In McDermott et al.'s framing of evaluation scenario categories, our work is either in "context-independent model evaluation" or "deployment scenarios with elevated false negative costs" (early screenings using CT scans to determine followup). For both scenarios, McDermott et al. recommend AUROC as the primary metric of interest, which justifies our focus in the main paper on AUROC.

B.3 Precision at Low Recall Results

Table B.3: Localization results for the **Precision at 10% Recall** metric, using the test set of each task. Precision is macro-averaged over positive bags. For some bags (e.g., with fewer than 10 positive slices), a threshold achieving exactly 10% recall is not achievable, so we report the interpolated precision at the smallest achievable recall $\geq 10\%$.

Method	Semi-Synthetic	Head CT	Chest CT	Abdomen CT
Centered Gaussian	0.656 ±.013	0.802 ±.008	0.613 ±.014	0.312 ±.036
Max pooling	0.499 ±.007	0.396 ±.008	0.225 ±.001	0.141 ±.008
Mean pooling	0.332 ±.003	0.359 ±.009	0.188 ±.001	0.104 ±.007
Additive MIL	0.443 ±.158	0.805 ±.002	0.455 ±.036	0.257 ±.012
ABMIL	0.579 ±.063	0.784 ±.033	0.276 ±.114	0.225 ±.036
post-hoc Blend a_i with CG	0.679 ±.022	0.834 ±.011	0.611 ±.012	0.304 ±.037
post-hoc Each slice as bag	0.711 ±.005	0.823 ±.008	0.323 ±.051	0.252 ±.020
post-hoc Gauss. kernel smoothing	0.645 ±.115	0.766 ±.057	0.287 ±.128	0.238 ±.037
Attn. Entropy Max. (AEM)	0.508 ±.137	0.812 ±.007	0.273 ±.114	0.211 ±.043
Smooth Operator	**0.766±.027**	0.835 ±.006	0.467 ±.050	0.236 ±.023
Counterfactual Intervention (CIA)	0.494 ±.119	0.810 ±.007	0.441 ±.086	0.298 ±.039
Normal Guidance (ours)	0.511 ±.137	**0.841±.009**	0.622 ±.028	**0.320±.026**
TransMIL	0.635 ±.155	0.790 ±.021	0.351 ±.071	0.274 ±.018
Smooth Operator	0.576 ±.198	0.803 ±.007	0.422 ±.015	0.277 ±.057
Multi-Head Normal Guidance (ours)	0.753 ±.074	0.829 ±.009	**0.644±.005**	0.302 ±.009
Best-in-Class Ceiling (see Sec. 4.3)	0.909 ±.004	0.862 ±.009	0.633 ±.006	0.388 ±.023

B.4 Bootstrap Tests for Significance

To assess whether gains are significant, we perform bootstrap resampling tests (Raschka, 2024, Method 3). For localization experiments, the metric of interest is the positive-bag localization AUROC, reported as the mean over the 3 test splits of each task. To obtain a distribution of plausible values, we resample the data used for computing the cross-split mean 10000 times. This includes resampling which of the 3 splits to include as well as which positive bags within a split to include. To decide if method A is "significantly better at the 95% level", we assess the empirical 5th percentile of $\{\Delta_s\}_{s=1}^{10000}$, where Δ_s is the difference between A's mean AUROC and B's mean AUROC on the s-th resampled set of results. If this 5th percentile is above zero, we know that 95% (or more) of the resamples confirm that A outperforms B.

We conducted the following pairwise tests for localization AUROC:

- is CG better than the best non-NG competitor? This test passed on all 3 CT tasks
- is NG better than CG? This test passed at the 95% level on Head CT and Chest CT, but only at the 80% level on Abdomen CT.
- is NG better than the best non-CG competitor? This test passed at the 95% level on all 3 CT tasks.

C ATTENTION VISUALIZATIONS

Below, you can see visuals of the learned attention weights for key MIL methods on the Head, Chest, and Abdomen CT tasks.

Alongside the predicted attention scores produced by MIL, we also show in shaded blue a "ground truth" attention distribution. This uses the expert-provided slice-level binary labels that come with these CT datasets (not used for training in any weakly supervised setting, only for evaluation). The binary truth labels are illustrated with a red outline (disease) or no outline (healthy) on the thin image strip above the attention graphs. To turn binary ground truth into a distribution comparable to our attention vectors, we normalize the binary values so the vector sums to one over the S_i slices. With only binary values provided, assigning equal attention to all positives seemed reasonable. In reality some slices in a positive region may show more positive signal than others (e.g., signs of more severe disease).

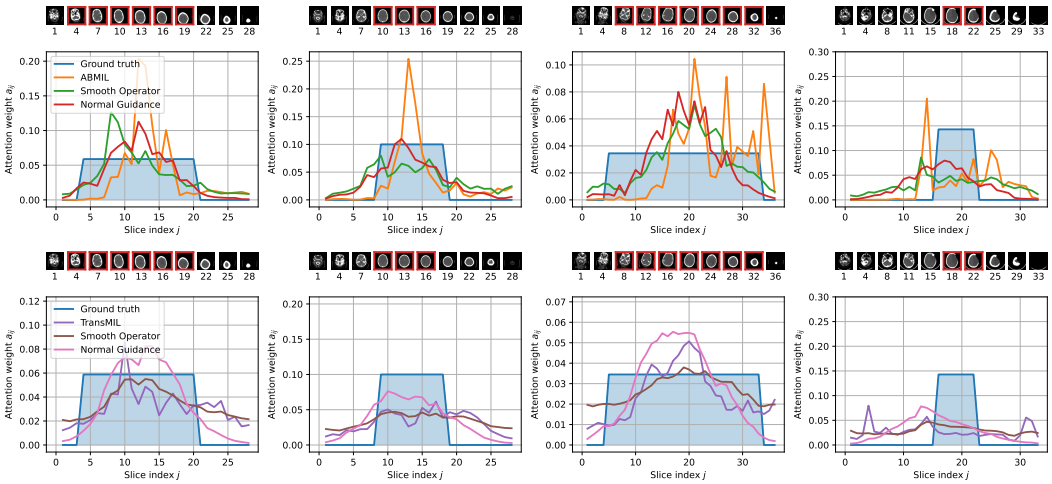

Figure C.1: Learned attention weights for 4 example CT scans (*columns*) from the **Head** CT dataset. *Top row*: methods based on ABMIL. *Bottom row*: transformer-based MIL methods. For "ground truth", the upper y-axis value is $1/P$, where P is the number of positive slices.

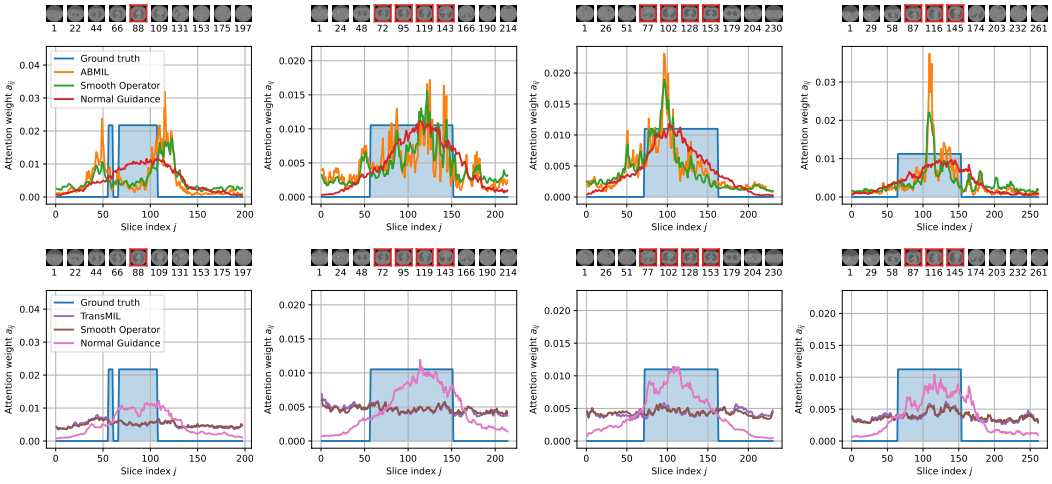

Figure C.2: Learned attention weights for 4 example CT scans (*columns*) from the **Chest** CT dataset. *Top row*: methods based on ABMIL. *Bottom row*: transformer-based MIL methods.

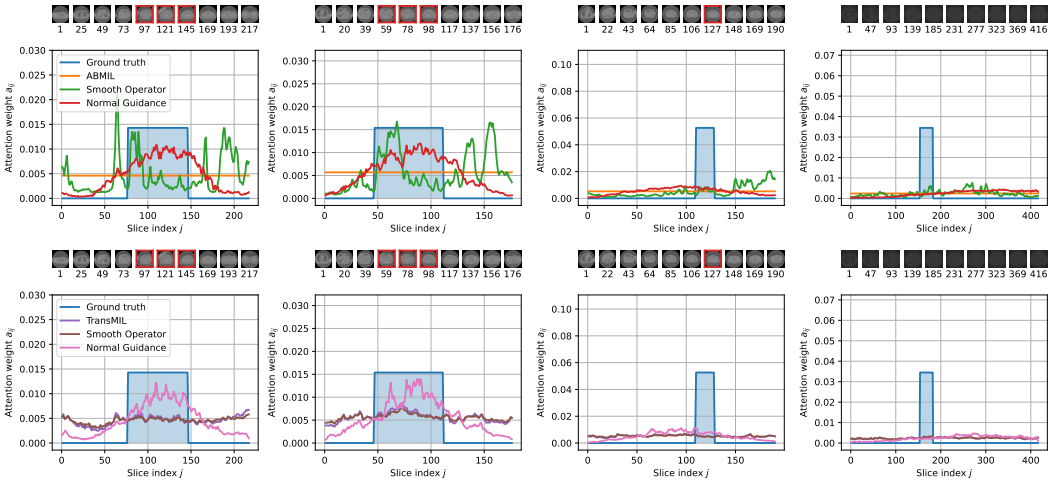

Figure C.3: Learned attention weights for 4 example CT scans (*columns*) from the **Abdomen** CT dataset. *Top row*: methods based on ABMIL. *Bottom row*: transformer-based MIL methods.

C.1 Attention by ROI configuration

Figs. C.1–C.3 show randomly sampled test scans. Figs. C.4–C.15 instead show, for each dataset, test scans with four ROI configurations: one ROI near the center of the scan, one ROI near an end, two or more well-separated ROIs near the center, and two or more well-separated ROIs with one near an end. Within each configuration, scans are drawn at random.

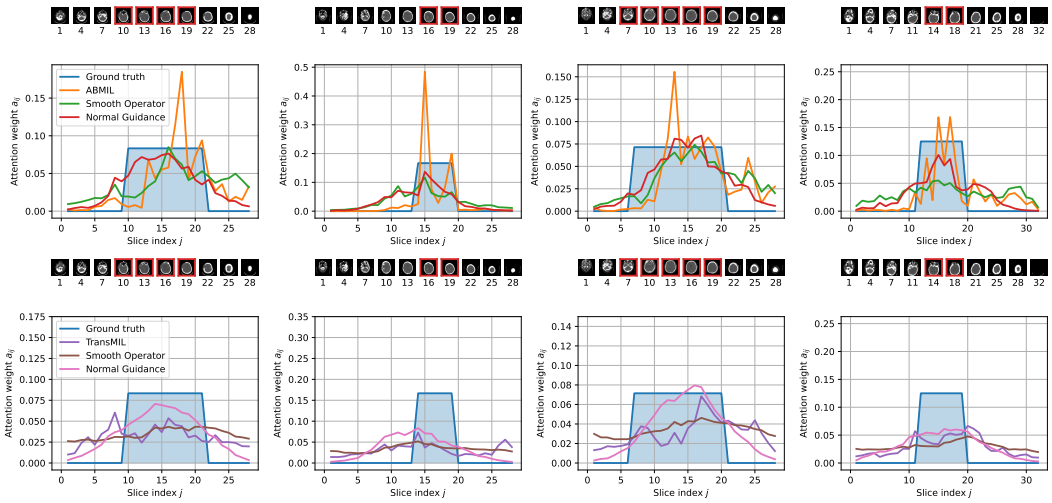

Figure C.4: Learned attention weights for 4 example CT scans (*columns*) from the **Head** CT dataset, each with one ROI near the center. *Top row*: methods based on ABMIL. *Bottom row*: transformer-based MIL methods.

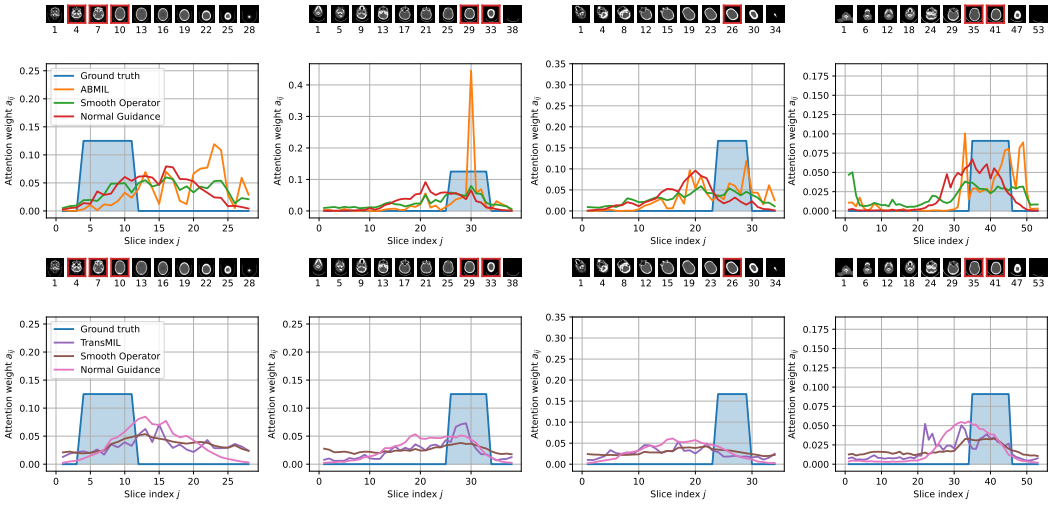

Figure C.5: Learned attention weights for 4 example CT scans (*columns*) from the **Head** CT dataset, each with one ROI near an end. *Top row*: methods based on ABMIL. *Bottom row*: transformer-based MIL methods.

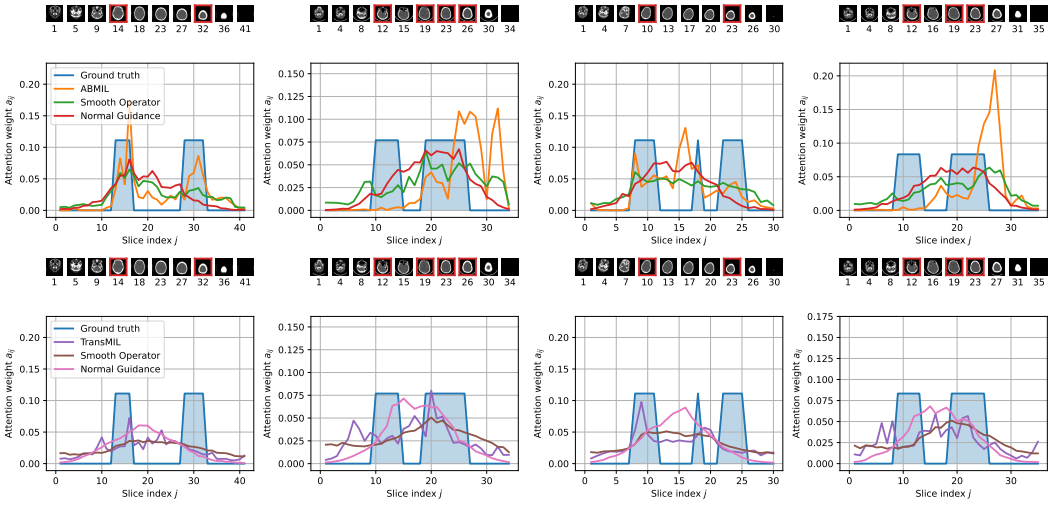

Figure C.6: Learned attention weights for 4 example CT scans (*columns*) from the **Head** CT dataset, each with two or more well-separated ROIs near the center. *Top row*: methods based on ABMIL. *Bottom row*: transformer-based MIL methods.

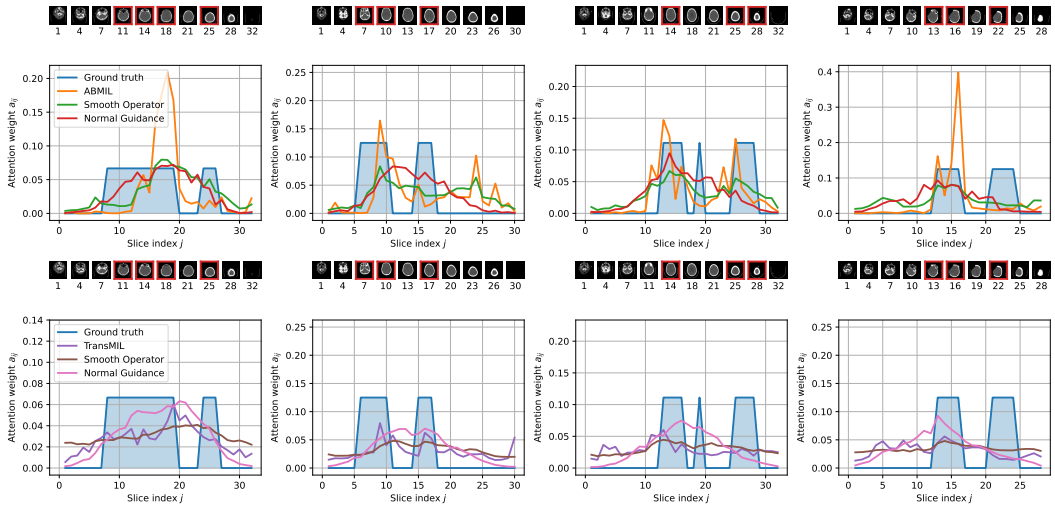

Figure C.7: Learned attention weights for 4 example CT scans (*columns*) from the **Head** CT dataset, each with two or more well-separated ROIs, at least one near an end. *Top row*: methods based on ABMIL. *Bottom row*: transformer-based MIL methods.

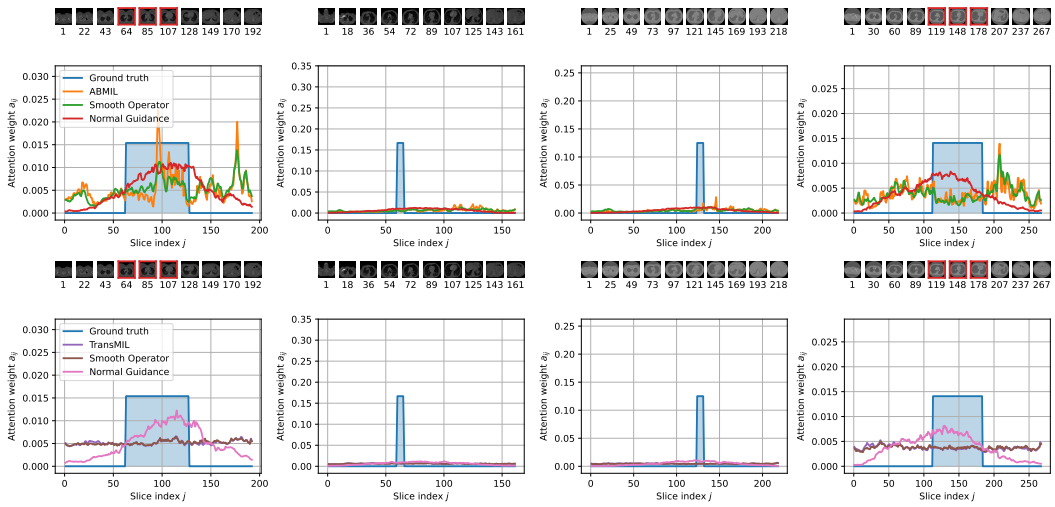

Figure C.8: Learned attention weights for 4 example CT scans (*columns*) from the **Chest** CT dataset, each with one ROI near the center. *Top row*: methods based on ABMIL. *Bottom row*: transformer-based MIL methods.

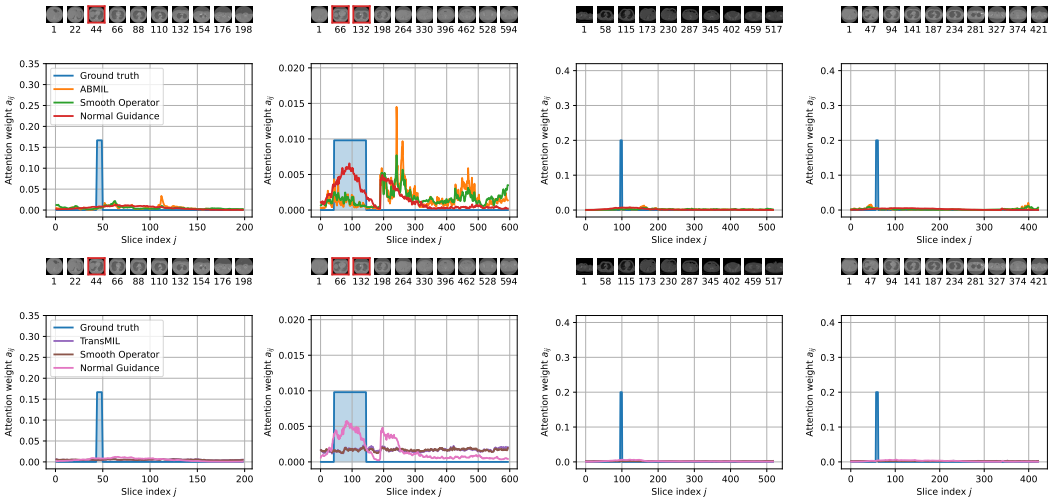

Figure C.9: Learned attention weights for 4 example CT scans (*columns*) from the **Chest** CT dataset, each with one ROI near an end. *Top row*: methods based on ABMIL. *Bottom row*: transformer-based MIL methods.

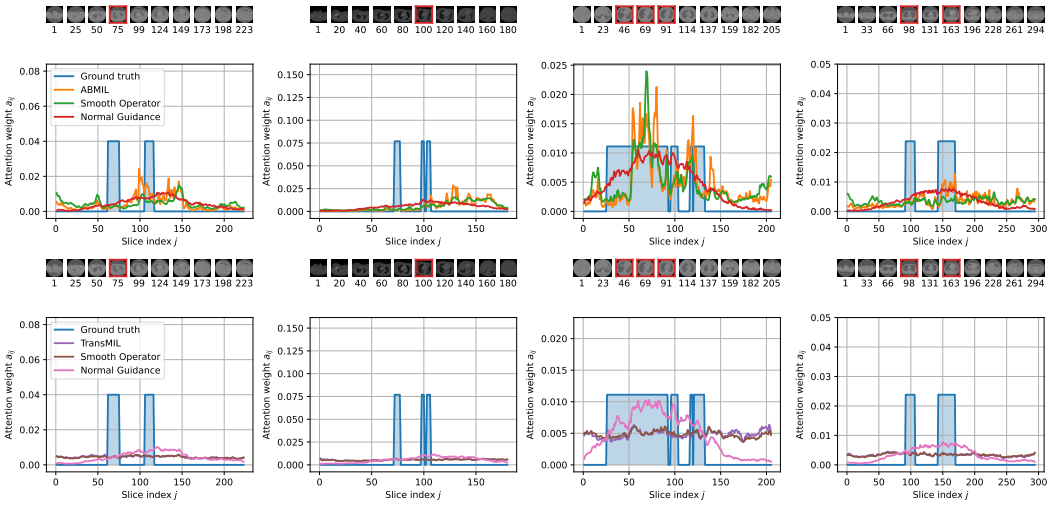

Figure C.10: Learned attention weights for 4 example CT scans (*columns*) from the **Chest** CT dataset, each with two or more well-separated ROIs near the center. *Top row*: methods based on ABMIL. *Bottom row*: transformer-based MIL methods.

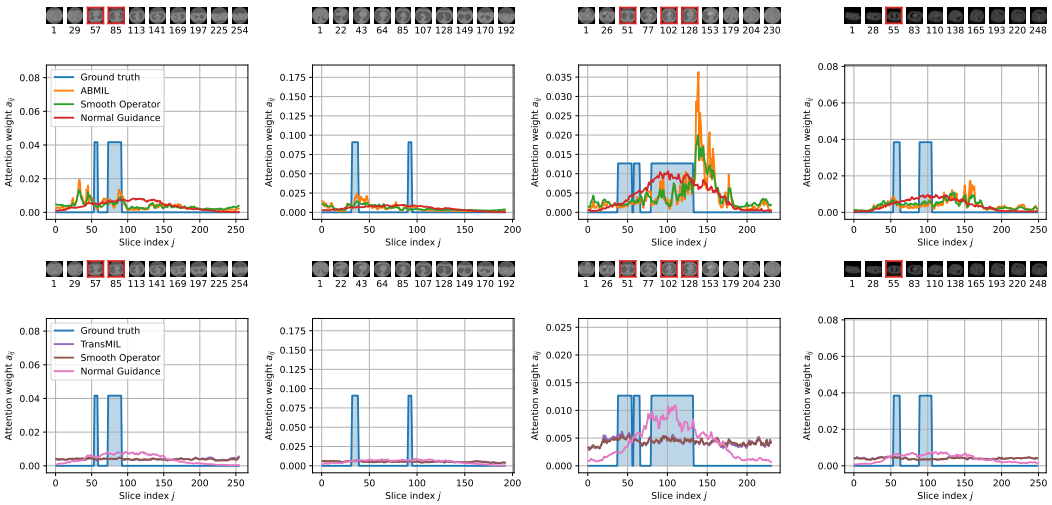

Figure C.11: Learned attention weights for 4 example CT scans (*columns*) from the **Chest** CT dataset, each with two or more well-separated ROIs, at least one near an end. *Top row*: methods based on ABMIL. *Bottom row*: transformer-based MIL methods.

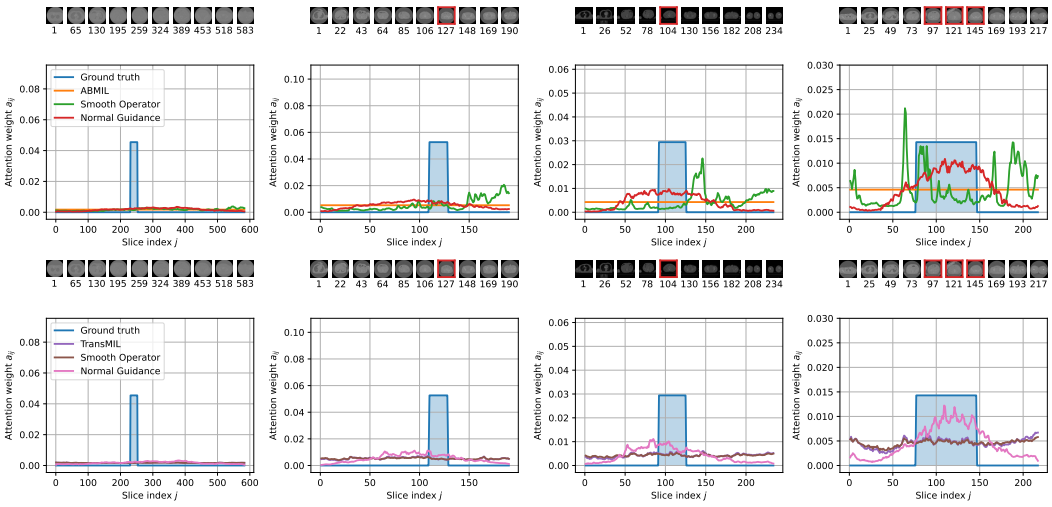

Figure C.12: Learned attention weights for 4 example CT scans (*columns*) from the **Abdomen** CT dataset, each with one ROI near the center. *Top row*: methods based on ABMIL. *Bottom row*: transformer-based MIL methods.

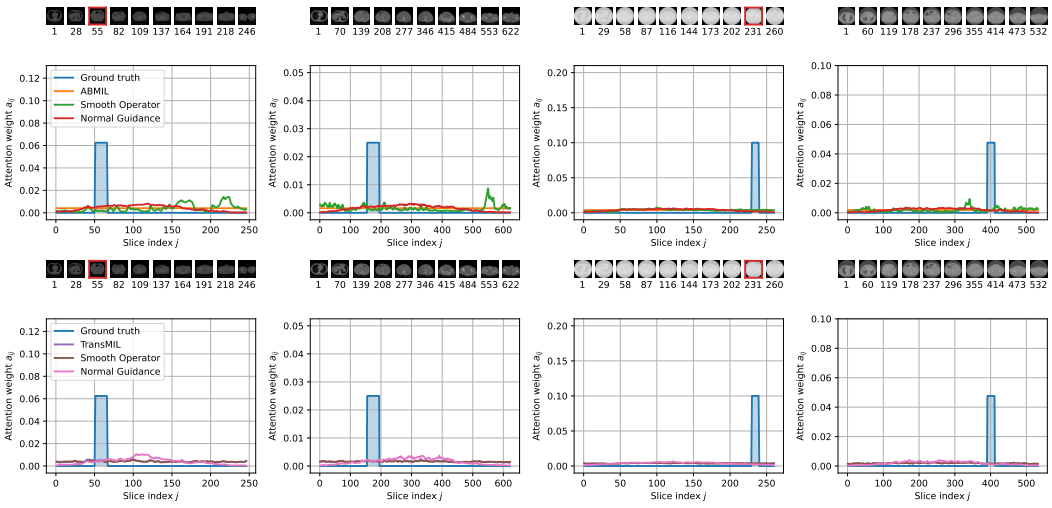

Figure C.13: Learned attention weights for 4 example CT scans (*columns*) from the **Abdomen** CT dataset, each with one ROI near an end. *Top row*: methods based on ABMIL. *Bottom row*: transformer-based MIL methods.

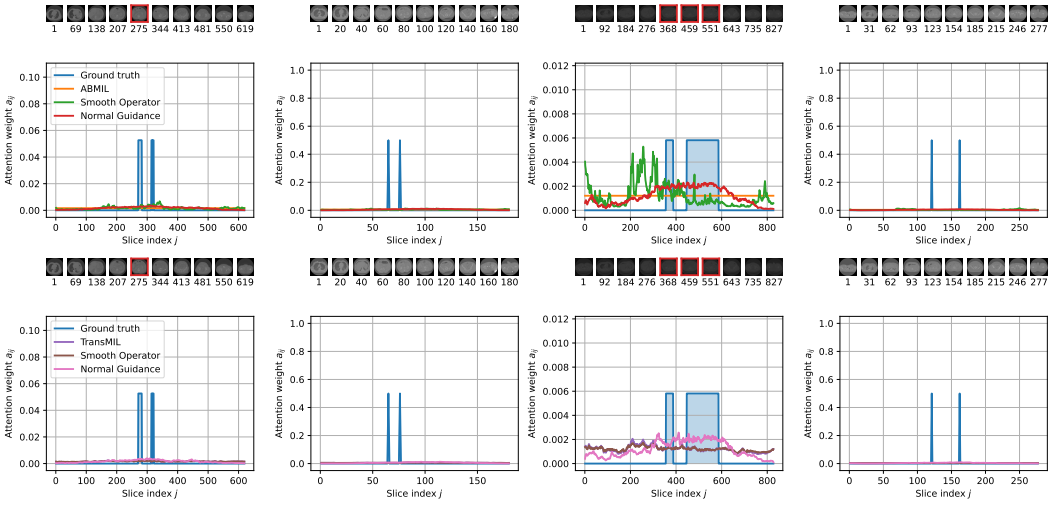

Figure C.14: Learned attention weights for 4 example CT scans (*columns*) from the **Abdomen** CT dataset, each with two or more well-separated ROIs near the center. *Top row*: methods based on ABMIL. *Bottom row*: transformer-based MIL methods.

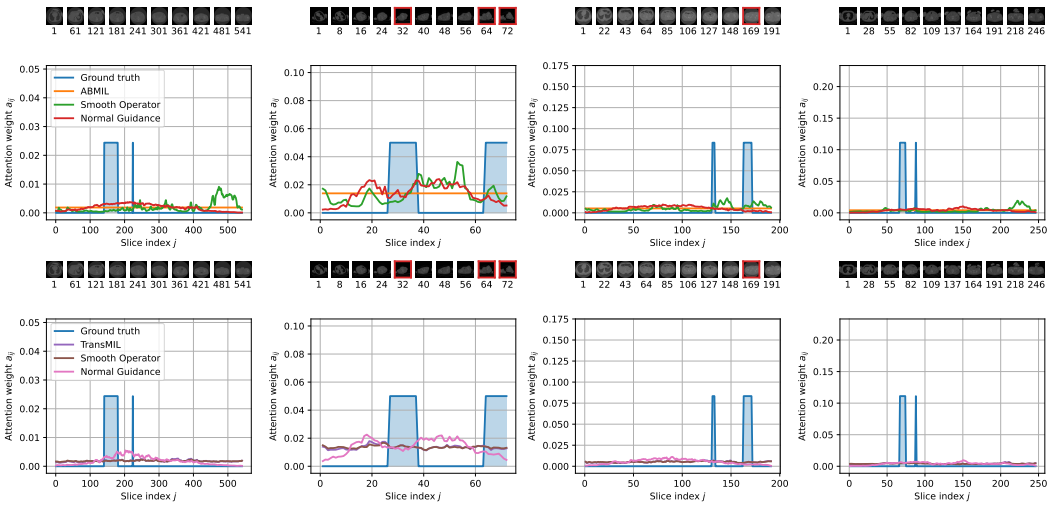

Figure C.15: Learned attention weights for 4 example CT scans (*columns*) from the **Abdomen** CT dataset, each with two or more well-separated ROIs, at least one near an end. *Top row*: methods based on ABMIL. *Bottom row*: transformer-based MIL methods.

D Additional Results: Variations of Normal Guidance

Our proposed Normal Guidance requires determining a mean and a variance for the construction of the guiding reference distribution r_i in Eq. (3).

Several variants are possible. Our recommended version uses the empirical mean and empirical variance of the current attention a_i to set the reference r_i. Instead, we could use a frozen constant value for the mean (e.g. the mean of the across-scan positive slice distribution in Fig. 3), for the variance alone (e.g. the variance of the across-scan positive slice distribution in Fig. 3), or both.

Note: even in the case of "frozen/frozen NG", here we are still *learning* an ABMIL model to produce attention that is pushed toward the provided reference over many epochs. At test time, this version of NG will still produce image-specific attention weights, unlike the CG baseline which just produces one centered Gaussian for every test CT scan regardless of image content.

We compare these options below in Tab. D.1. These variants deliver surprisingly similar values of both bag-level AUROC and localization AUROC, within 0.009 across all cases. This suggests that the precise mean and variance parameterization does not matter so much on the tested Head CT task.

Table D.1: Comparison of variations of Normal Guidance for determining the mean and variance parameters of the normalized Gaussian reference in Eq. (3). Results for test set of Head CT dataset.

mean	var.	bag-level AUROC	loc. AUROC
$\mathbb{E}_a[J]$	$\mathrm{Var}_a[J]$	0.925 $\pm$0.008	0.871 $\pm$0.002
$\mathbb{E}_a[J]$	frozen	0.920 $\pm$0.009	0.865 $\pm$0.007
frozen	$\mathrm{Var}_a[J]$	0.924 $\pm$0.007	0.870 $\pm$0.004
frozen	frozen	0.917 $\pm$0.008	0.872 $\pm$0.002

E Further Details about Baseline Models and Experiments

E.1 Compute Resources

All experiments use either 4 Intel Xeon 6342 CPUs (2.80 GHz) or 4 Intel Xeon 6226R CPUs (2.90 GHz) and either 1 NVIDIA RTX 6000 (24 GB), 1 NVIDIA RTX A6000 (48 GB), or 1 NVIDIA A100 GPU (40 GB).

E.2 Localization Experiment Details

For most MIL methods, the attention score a_{ij} used to evaluate localization is straightforward. For max pooling MIL only, we need to construct a post-hoc attention score for each instance. Because we use element-wise max pooling MIL, we can construct attention weights $\alpha_{ij} = \frac{1}{M}\sum_{k=1}^{M} \mathbb{1}[\mathrm{argmax}_{j'}(h_{ij'k}) = j]$ where $\mathbb{1}[\cdot]$ denotes the indicator function.

For methods like CG, we evaluate attention values in log space to avoid underflow from the Normal PDF on slices j that are many standard deviations away from the mean.

E.3 Post-hoc Baselines for Localization

Given an already trained ABMIL method, we experimented with several post-hoc baselines for alternative instance-level scores. We thank anonymous reviewers for these suggestions.

Post-hoc Blend with CG. Here, we form the final instance-level scores as a weighted blend of the a_i vector produced by ABMIL and the bell-shaped attention a_i^{CG} vector from the Centered Gaussian baseline, whose j-th entry is $a_{ij}^{\mathrm{CG}} \propto \mathcal{N}(j|\frac{S_i}{2}, \tau^2)$. Here, we select the value of τ, the Gaussian's standard deviation, to maximize the blended attention's localization AUROC on the validation set. We fix the blending weights to 0.5, for an even mixture.

Each slice as own bag: We provide each slice's corresponding single image as a standalone "bag" input to the trained ABMIL backbone. The scalar output predicted probability of positive class is

then used as the predicted score of each instance to evaluate localization. This is the "Single Slice" baseline mentioned in Early et al. (2022).

Gaussian kernel smoother. Given the 1D attention vector a_i for all S_i slices produced by the ABMIL backbone, we convolve that signal with a Gaussian kernel using the function `gaussian_filter1d` from the SciPy package (Virtanen et al., 2020) in python. The resulting vector is then renormalized to sum to one. The kernel's "bandwidth" sigma is chosen to maximize localization AUROC on the validation set.

E.4 MIL Method Setup Details

All MIL models use the same frozen ViT-B/16 encoder ($M = 768$); only the pooling head and linear classifier are trained.

Smooth Operator (SM). We use an approximation of SM following Castro-Macías et al. (2024), defined by $T = 10$ iterations of $g^{(t+1)} = (1-\alpha)\, h_i + \alpha\, A\, g^{(t)}$ starting from $g^{(0)} = h_i$. Here A is the normalized adjacency matrix of the chain graph representing the slice ordering of the 3D scan. The mixing scalar is parameterized as $\alpha = \sigma(\tilde{\alpha})$ with $\tilde{\alpha} \in \mathbb{R}$ learnable and initialized so $\alpha_{t=0} = 0.5$. For SmAP, a single SM module is applied to the encoder's output embeddings prior to ABMIL. For SmTP, a separate SM module (each with its own learnable α) is inserted after each MHSA layer.

TransMIL. We follow Shao et al. (2021): a learnable class token is prepended to h_i, followed by two multi-head self-attention blocks separated by a Pyramidal Position Encoding Generator (PPEG) module. The class-token output of the second block is the bag representation. Unlike the reference implementation, which approximates self-attention with Nyström attention, we use full scaled dot-product self-attention (`torch.nn.MultiheadAttention`). Both blocks use $H = 8$ heads at model dim $M = 768$ (per-head dim 96) with no dropout. PPEG is three parallel depthwise 1D convolutions with kernel sizes $\{3, 5, 7\}$, stride 1, and SAME padding, summed and added to the non-class-token positions. Reported slice-level attention $a_{i,j}$ is the class-token row of the second block's attention map, averaged over heads, with the class-token-to-class-token entry removed. We recommend this vector is renormalized to sum to one over the remaining S_i entries before any localization evaluation or regularization.

AEM . We use attention entropy maximization (AEM) during training which encourages attention to have high entropy (penalizing attention sparsity):

$$\mathcal{L}_{\text{AEM}} := \frac{1}{N} \sum_{i=1}^{N} \left(\ell^{\text{BCE}}(y_i, \hat{y}_i) + \lambda \sum_j a_{ij} \log(a_{ij}) \right) + \alpha|\theta|$$

We set $\lambda = 0.001$ following Zhang et al. (MICCAI 2025)'s recommendation and search over the same weight decay strength α and learning rate candidate values as all other methods.

AdditiveMIL: We train the pooling operation and classifier by applying the linear classification layer directly to each attention-weighted instance before adding them to form bag-level logits:

$$\text{logits}_{ij} = g(a_{ij} \cdot h_{ij})$$

$$\text{logits}_i = \sum_j \text{logits}_{ij}$$

. We search over the same weight decay strengths and learning rates as all other methods.

F Semi-Synthetic Data and its Bayes Estimator for Localization

Data-generating process. We set $M = 768$ (num. features) and $R = 12$ (size of the latent block). We sample S_i uniformly between 20 and 60 for any bag regardless of its label. We draw y_i, the bag-level binary label, from a Bernoulli distribution with probability 0.5.

https://docs.scipy.org/doc/scipy/reference/generated/scipy.ndimage.gaussian_filter1d.html

For negative bags, we draw all S_i feature vectors $h_{ij} \in \mathbb{R}^M$ from the same simple distribution: each entry is drawn iid from a Gaussian with mean 0 and variance 1.

For positive bags, we draw the starting index u_i of the modified R adjacent instances from a uniform distribution over the valid range

$$p(u_i|S_i, y_i{=}1) = \text{UnifPMF}(u_i|\{1, \ldots, S_i{-}R{+}1\}). \tag{4}$$

Next, given the u_i for a positive bag we draw only the first feature (of M) of every instance in the chosen block of size R from a Gaussian with *shifted-mean* $\Delta = 0.5$. Otherwise, all features are drawn from the same zero-mean, unit variance Gaussian that defines the negative bags:

$$p(h_i|u_i, S_i, y_i{=}1) = \prod_{j=1}^{S_i} \prod_{k=1}^{M} \begin{cases} \text{NormPDF}(h_{ijk}|0.5, 1), & \text{if } j \in [u_i, u_i{+}R{-}1] \text{ and } k = 1 \\ \text{NormPDF}(h_{ijk}|0, 1), & \text{otherwise.} \end{cases} \tag{5}$$

Relevant set of u_i values for index j. We define the set $\mathcal{U}_{ij}$ as the set of possible u_i values that would make index j fall in the chosen latent segment of R consecutive instances. Formally,

$$\mathcal{U}_{ij} = \{u_i \in \mathbb{Z} : \max(1, j - R + 1) \leq u_i \leq \min(S_i - R + 1, j)\}.$$

We know the size of the set is always at most R: $|\mathcal{U}_{ij}| \leq R$. It will only be smaller than R for indices j within R steps of either edge.

Ceiling for localization. Given the data-generating process above, we can compute the probability that slice j is included in the block of instances with shifted mean. As defined above, let $\mathcal{U}_{ij}$ be the set of possible values of u_i that could make j positive. Adding up the probability of these relevant values, we can write

$$\begin{aligned} p(y_{ij}{=}1|h_i, S_i, y_i{=}1) &= \sum_{u_i \in \mathcal{U}_{ij}} p(u_i|h_i, S_i, y_i{=}1) \\ &= \sum_{u_i \in \mathcal{U}_{ij}} \frac{p(h_i|u_i, S_i, y_i{=}1)p(u_i|S_i, y_i{=}1)}{p(h_i|S_i, y_i{=}1)}. \end{aligned} \tag{6}$$

This uses Bayes' rule. Each term in Eq (6) can be computed in closed form. The numerator is a direct product of expressions in Eq. (5) and (4). The denominator is a marginal likelihood of the observed feature vectors h_i alone via the sum rule, summing over all possible starting indices u_i:

$$p(h_i|S_i, y_i{=}1) = \sum_{u=1}^{S_i - R + 1} p(h_i|u_i = u, S_i, y_i{=}1)p(u_i = u|S_i, y_i{=}1). \tag{7}$$

Again, this right hand side can be evaluated via expressions in Eq. (5) and (4).

G Methods: Further Details and Discussion

G.1 Label Guidance

An assumption in MIL is that instance-level labels are unknown during training. However, if instance-level labels $\{y_{i,1}, \ldots, y_{i,S_i}\}$ are known, they can be used to define a discrete reference distribution r_i to regularize the learned attention weights a_i. For example, for a negative bag we could define a uniform distribution over all instances $p(j|y_i{=}0) = \text{UnifPMF}(j|\{1, \ldots, S_i\})$ and for a positive bag we could define a uniform distribution over all positive instances $p(j|y_i{=}1) = \text{UnifPMF}(j|\{k \in \{1, \ldots, S_i\} : y_{i,k}{=}1\})$. These attention weights would achieve a perfect score for instance-level AUROC and AUPRC. Recent work has used instance-level semantic labels to guide attention (Liu et al., 2024). Other works have used instance-level semantic labels to train classifiers to supervise attention in applications to medical ultrasound (Huang et al., 2023).

G.2 Inductive Biases for Normal Guidance

Our NG method described in Sec. 3.2 has 3 inductive biases that we argue are helpful for the 3D CT tasks of interest to us in this work.

First, it encourages the learned attention a_i to be smooth. Smoothness matches the clinical intuition that we shouldn't often rapidly transition between high and low (relevant and irrelevant) among neighboring slices.

The second inductive bias is a center focus. By setting the mean (and thus mode) of the reference r_i to the empirical mean, the reference becomes center-focused and by nature tends to downweight slices near the edge of the 3D volume, since these are rarely positive in the empirical distribution (see Fig 3).

The final inductive bias is unimodality. More often than not in all 3 CT tasks of interest, there is only one contiguous region of interest in these 3D volumes. In this way, unimodality is useful and usually (though not always) preferred.

We are upfront in our stated limitations that unimodality is not a perfect match to the desired reality (sometimes there are several regions of interest). However, overall, we found our NormalPDF based construction was simple, easy to implement and understand, and generally effective due to these inductive biases. We could probably have used another bell-curve like shape, but found this worked well enough. As the saying goes, "All models are wrong, but some are useful."

G.3 Alternative Guidance via a Gaussian Process prior

As an alternative to the NG method, we also explored a more flexible reference distribution built from a mixture of Gaussian processes model we call MixGP. Here, we briefly describe this method and provide evidence that, even though it is well-motivated, it did not outperform our simpler NG method in practice.

Let $g_i = [g_{i1}, \dots g_{ij}, \dots g_{iS_i}]$ denote the pre-softmax logit-valued attention across the slices of scan i. Our MixGP model with R components generates logits for S_i slices via two sampling steps:

$$z_i \sim \text{Cat}(\pi_1, \pi_2, \dots \pi_R), \tag{8}$$

$$g_{i,1:S_i} | z_i{=}r \sim \mathcal{N}(m^r, K^r + (\sigma^r)^2 I_{S_i}), \qquad m^r = [m^r(0), \dots, m^r(\tfrac{j-1}{S_i-1}), \dots m^r(1)]$$

Here, both the mean and covariance functions take univariate position values that are *normalized* between 0 and 1, where 0 means the start position of scan (e.g. top of head for Head CT). We set the mean function m to a piecewise constant function over 25 evenly-spaced bins covering the 0-1 interval. Each bin's value matches the empirical mean of that bin's true slice-level labels in the training set.

For the covariance function, we use a stationary spectral mixture (SM) kernel (Wilson & Adams, 2013), with $Q = 4$ kernel-specific components (distinct from the R outer components of the overall mixture above).

$$K^r_{j,j'} = k^{SM,r}(\tfrac{j-1}{S_i-1} - \tfrac{j'-1}{S_i-1}), \qquad k^{SM,r}(t) = \textstyle\sum_{q=1}^{Q} w_q^r \exp\left(-\frac{1}{2}t^2(\ell_q^r)^{-2}\right) \cos\left(2\pi\omega_q^r t\right). \tag{9}$$

Here, for each component (indexed by r, q), parameter $w_q^r \geq 0$ controls its output scale, $\ell_q^r > 0$ sets the lengthscale, and $\omega_q^r > 0$ defines the frequency (cycles per unit). The SM kernel should have clear advantages over typical stationary kernels like RBF or Matern, such as the ability to model negative correlations at some t values. When $\omega_q^r \to 0$, the q-th component approaches an RBF kernel.

Specially processed Head CT data. To fit and evaluate methods in this subsection, we use a special version Head CT data, preprocessed so that normalized positions 0 and 1 have consistent physical meaning across scans. We set 0 to mean the top of the head, the first available slice where at most 90% of the CT slice is all black. We set the 1 position to mean the last available slice at most 155 millimeters away from position 0. 155 is the maximum observed height of the cranium (basion to bregma) in a recent cohort study of 100 humans (Sangvichien et al., 2007). Slices out of this region are discarded.

Fitting GP parameters on Head CT. Using our specially-processed Head CT dataset, for each positive scan (any ICH lesion) we divide the training data into $R = 3$ cohorts based on the number of contiguous lesion ROIs: unimodal (just one contiguous ROI), bimodal (two separate ROIs), and multi-region (three or more separate ROIs). We model each cohort with a separate mixture component. We directly set π_r and m_r to empirical estimates from that cohort (e.g. the empirical frequency of unimodal scans among all positive scans for π_1, the empirical mean of g values in each bin for

m). Next, we estimate other parameters – $\sigma^r, \{w_q^r, \ell_q^r, \omega_q^r\}_{q=1}^Q$ – by maximizing the likelihood of observed g_i vectors in that cohort (using the PDF in the second line of Eq (8))

Our implementation shows clear advantages to all 3 modeling innovations: (i) using a distinct mixture component for unimodal, bimodal, and multimodal (3+) ROI scans, (ii) using the piecewise constant mean instead of just all zero mean, and (iii) using the SM kernel instead of the RBF kernel. See the table below:

Table G.1: Heldout likelihood evaluation of different components for our MixGP model.

Mixture	mean m	kernel k	tr. NLL	val. NLL
R=1 GP	const	RBF	6.0232	6.0842
R=1 GP	25-bin piecewise	RBF	4.6478	4.7177
R=1 GP	25-bin p.w.	SM with $Q = 4$	4.2684	4.3520
R=2 MixGP	25-bin p.w.	SM with $Q = 4$	3.4147	3.3972
R=3 MixGP	25-bin p.w.	SM with $Q = 4$	3.3475	3.2967

Applying GP regularization to MIL. We now explain how we use our GP model to regularize the attention of MIL. Because Stage-2 attention operates on unbounded pre-softmax logits rather than the $[0, 1]$-scale labels our GP was fit on, we first affinely rescale the frozen prior from $[0, 1]$ onto a logit range of $[-2, 2]$ via $g = 4y - 2$. Under this rescaling, each mixture weight w_q^r and the noise variance $(\sigma^r)^2$ are scaled by 16, while the length-scales ℓ_q^r and frequencies ω_q^r remain unchanged. We then need the ability to evaluate the quality of a given g_i vector under our model. Marginalizing out the cluster indicator variable z_i from Eq. (8) we can compute the marginal likelihood of vector g_i under our MixGP model given parameters $\eta = \{\pi_r, m^r, K^r, \sigma^r\}_{r=1}^R$ as

$$\log p(g_i|\eta) = \text{logsumexp}\begin{pmatrix} \log \pi_1 + \log \mathcal{N}(g_i|m^1, K^1 + (\sigma^1)^2 I_{S_i}) \\ \vdots \\ \log \pi_R + \log \mathcal{N}(g_i|m^R, K^R + (\sigma^R)^2 I_{S_i}) \end{pmatrix} \tag{10}$$

We then offer a logit-attention-guided objective for overall training of MIL models with parameters θ, with loss:

$$L(\theta) = \sum_{i=1}^N \ell^{\text{BCE}}(y_i, \hat{y}_\theta(x_i)) + \lambda_1 \sum_{i:y_i=1} -\log p(g_\theta(x_i)|\eta) + \lambda_0 \sum_{i:y_i=0} D(r_i||a_\theta(x_i)) + \alpha|\theta|_1 \tag{11}$$

The first regularization term steers the attention applied to positive bags to match configurations favored by the MixGP model. The second term applies alternative attention-guidance to negative bags (we usually just set this to be standard Normal Guidance).

We trained two variants of each configuration on our new trimmed Head CT dataset: NG, an ABMIL classifier regularized only by Normal Guidance with no GP term, and GP Regularization, our full approach. Both were trained under two divergence choices for the negative-bag penalty, squared error and forward KL.

Both λ_1 and λ_0 were selected by grid search based on the mean bag-level validation AUROC across seeds, with $\alpha = 10^{-5}$ and $\text{lr} = 0.01$ held fixed throughout. The two weights were swept jointly over $\lambda_1 \in \{10^{-3}, 10^{-2}, 10^{-1}, 1\}$ and $\lambda_0 \in \{0.1, 1, 10\}$. For simplicity, we selected a single pair of λ_0 and λ_1 values for all seeds based on the highest mean validation AUROC. For GP Regularization, $\lambda_1 = 0.01$ yielded the highest mean validation AUROC for both the squared-error and forward-KL variants.

We similarly retrained the ABMIL + Normal Guidance baseline on the specially processed Head CT dataset, sweeping its guidance weight $\lambda \in \{0.1, 1, 10\}$ while holding $\alpha = 10^{-4}$ and $\text{lr} = 0.01$ fixed at their previously established best values in the main paper.

Nevertheless, using our best MixGP model as a regularizer of attention logits, we did not observe a clear improvement over Normal Guidance alone in either classification or localization performance, as shown in Tables G.2 and G.3.

Table G.2: Classification AUROC on the specially processed Head CT dataset for ABMIL under each guidance method as its weight is swept, with λ_1 fixed at 0.01. For the Normal Guidance rows, λ_0 denotes Normal Guidance's own weight λ. Corresponding localization results are reported in Table G.3.

Method	$\lambda_0 = 0.1$	$\lambda_0 = 1$	$\lambda_0 = 10$
Normal Guidance, squared-error	$0.921_{\pm 0.008}$	$0.926_{\pm 0.010}$	$0.924_{\pm 0.008}$
Normal Guidance, forward KL	$0.923_{\pm 0.011}$	$0.923_{\pm 0.008}$	$0.921_{\pm 0.009}$
GP Regularization, squared-error	$0.926_{\pm 0.009}$	$0.926_{\pm 0.008}$	$0.923_{\pm 0.006}$
GP Regularization, forward KL	$0.926_{\pm 0.008}$	$0.922_{\pm 0.005}$	$0.920_{\pm 0.007}$

Table G.3: Localization AUROC on the specially processed Head CT dataset for ABMIL under each guidance method as its weight is swept, with λ_1 fixed at 0.01. For the Normal Guidance rows, λ_0 denotes Normal Guidance's own weight λ. Corresponding classification results are reported in Table G.2.

Method	$\lambda_0 = 0.1$	$\lambda_0 = 1$	$\lambda_0 = 10$
Normal Guidance, squared-error	$0.779_{\pm 0.191}$	$0.805_{\pm 0.162}$	$0.844_{\pm 0.165}$
Normal Guidance, forward KL	$0.832_{\pm 0.163}$	$0.844_{\pm 0.166}$	$0.843_{\pm 0.172}$
GP Regularization, squared-error	$0.826_{\pm 0.165}$	$0.834_{\pm 0.162}$	$0.846_{\pm 0.165}$
GP Regularization, forward KL	$0.837_{\pm 0.162}$	$0.844_{\pm 0.167}$	$0.841_{\pm 0.173}$

G.4 Best-in-Class Ceiling Methods

Avoiding shortcuts in bag-level ceiling for CT. Enforcing similar ROI distributions across positive and negative CT bags was critical to developing MIL classifiers that reflect the task's real difficulty.

As a simpler alternative to enforcing similar ROIs, we tried using positive slice-labels for positive bags while using all slices for negative bags. But we found unintended shortcuts were possible due to position-specific feature distributions. If negative bags pool over the entire volume but positive bags concentrate on specific regions, a simple "how dark is this image?" feature could detect the more dark CT scan slices near the edges at index 1 or S_i in negative bags and yield near-perfect classification that is too optimistic.

H Additional Related Work

Inductive biases in MIL. Other work in MIL has pursued inductive bias in the dependencies across instances. To account for 2D spatial relations among instances in MIL for whole slide imaging (WSI), Peled et al. (2026)'s probabilistic spatial attention (PSA) method modifies self-attention with a learnable distance-decayed prior. This makes nearby instances (patches) likely to attend to each other in the $S_i \times S_i$ pairwise self-attention matrix. Unlike our approach with a regularization term that is agnostic to any pooling strategy, their approach directly modifies the construction of attention values and relies on multi-head self-attention as a specific attention mechanism. PSA also does not necessarily encourage unimodality, which seems to be a beneficial inductive bias in 3D CT datasets we tackle here.

Gaussian and GMM ideas in MIL. Distantly related work (Zhang & Feng, 2021) outside of MIL has used Gaussian mixture ideas to determine how one token should pay attention to other tokens in a text sequence for NLP translation tasks, hoping to encourage concentrated attention. Our work avoids the expense of a distinct GMM for every instance in the sequence, instead forming just one Gaussian-like discrete distribution per head. Our work is also distinct in its purpose to improve localization for MIL in 3D CT settings.

Our work is somewhat related to other efforts in MIL to incorporate probabilistic modeling ideas, such as work using Gaussian Processes to predict instance-level and bag-level labels given permutation-invariant per-instance embeddings (Wu et al., 2021). That work uses Gaussian machinery to model a latent response function which is then transformed into a binary label; they do not regularize attention as we do toward helpful inductive biases such as unimodality and smoothness.

H.1 Discussion of Past work on Upper Bounds using Instance-Level Labels

Correia et al. (2020) measure the performance of a permutation-invariant MIL on in-the-wild speech in Youtube videos to detect Parkinson's or depression. They develop a "fully supervised" upper bound for MIL via a classifier that consumes one instance at a time, somehow comparing this classifier given bags of size 1 to MIL classifiers given bags of larger size in the same figure. Guillaumin et al. (2010) consider a similar instance-level upper bound in a metric-based MIL approach for multi-label MIL problems. Instead, our work develops separate best-in-class ceilings for both instance-level and bag-level tasks, uses a local neighborhood for the 1D CNN instance-level classifier instead of a "one instance in and one label out" design, and always evaluates competitor methods on bags of the same size (much larger than one).

H.2 Alternatives to Weak Supervision

An alternative to weakly-supervised MIL with zero annotation costs is fully-unsupervised anomaly detection. Unsupervised methods for detecting brain lesions from 3D MRI scans have been proposed using diffusion models (Behrendt et al., 2024) or self-supervised learning where the pretext task is predicting the location of a given patch within the 2D axial slice (Baker & Brockmeier, 2026). These approaches generally rely on an expensive bespoke-trained representation rather than frozen encoders that are more easily ported across medical image modalities, and also are likely less accurate at scan-level prediction than the deep MIL studied here due to the lack of supervised labels informing model training.