Automation and Machine Learning (2026)
Clausius Scientific Press, Canada

DOI: 10.23977/autml.2026.070215
ISSN 2516-5003 Vol. 7 Num. 2

Spatial Attention Consistency and Faithfulness in Whole Slide Classification

Jian Li, Zhitao Wu*, Zhaoguo Jiang*

University of Science and Technology Liaoning, Anshan, 114051, Liaoning, China
3761383069@qq.com

Keywords: Whole Slide Imaging; Attention Consistency; Multiple Instance Learning; Attention Entropy; Tissue Interventions

Abstract: Two stain views of a whole-slide image can have highly similar attention maps even when attention is distributed uniformly over tissue. We evaluated whether coordinate-aligned attention consistency improves the evidence used by a whole-slide classifier when attention concentration is controlled. A shared attention-based multiple instance learning model was evaluated with seven training and control conditions, including coordinate shuffling, entropy matching and forced uniform attention. The joint prediction and evidence model E2 achieved 83.9% case-level AUROC, compared with 82.0% for prediction consistency and 82.1% for the entropy-matched control. E2 and the entropy-matched control had similar attention entropy (2.87 and 2.89) and top-decile attention mass (0.64 and 0.63). At a matched 10% tissue-area budget, their deletion advantages were 0.068 and 0.040, and their retention advantages were 0.061 and 0.037. E2 led the prediction-consistency and entropy-matched arms at all four deletion budgets. Its normalized deletion-area advantage over prediction consistency was 0.0236. The results distinguish spatial agreement and score-based faithfulness from attention concentration in paired-stain whole-slide classification.

1. Introduction

A whole-slide image (WSI) classifier may assign similar attention to two stain views for different reasons. Both views may emphasize the same tissue region, or both may assign nearly equal weight to every tile. In either case, the Jensen–Shannon divergence (JSD) between attention distributions can approach zero. The same concentration of attention can also occur at different tissue coordinates. Evaluating agreement therefore requires knowing where the attention falls and how the selected tissue affects the classifier's score.

Stain augmentation varies the appearance of histological images while preserving tissue organization[1]. When the original and perturbed views share an identical tile grid, the paired views provide a direct test of whether a model attends to the same tissue under a change in appearance. Attention-based multiple instance learning (ABMIL) makes this test accessible through the tile weights used to construct the slide representation[2]. A shared pooling model allows these weights to be compared at corresponding coordinates.

We compare joint prediction and evidence consistency with three controls. Coordinate shuffling

tests whether the alignment depends on the spatial correspondence between views. Entropy matching controls the concentration of attention. Forced uniform pooling supplies a reference with maximal agreement across identical tile sets. We then compare the score effects of deleting or retaining highly weighted tissue with the effects of applying the same intervention to matched tissue.

2. Related Work

Consistency training has been used to improve the stability of model outputs and their spatial attribution. ATCON aligns attribution maps produced by different explanation methods[3]. HASD includes patch-level attention consistency when adapting whole-slide representations across domains[4]. Our experiment uses an identical spatial grid across stain views and compares coordinate alignment with explicit controls for spatial correspondence and attention concentration.

Sanity checks for saliency maps test whether a visualization responds to changes in the trained model[5]. In histopathology, patch-level perturbations have also been used to evaluate the faithfulness of WSI explanations[6]. We use a within-slide comparison, holding tissue area and prespecified tissue properties constant as we change which tiles enter the pooling head. Deletion and retention provide complementary measurements of the classifier's dependence on the selected tissue.

3. Methods

3.1. Paired Views and Shared Attention Pooling

Each slide is represented by a bag of tissue tiles with stored spatial coordinates. An artificial H&E perturbation changes stain concentrations in optical-density space while preserving the tissue mask, tile locations and tile order. The original and perturbed tiles pass through the same frozen image encoder. A shared ABMIL head assigns attention weights across each bag, pools the encoded tiles and returns a class-probability vector. The two attention vectors are indexed by the same tissue coordinates.

The joint objective combines the mean classification loss across the two views with two consistency terms. Prediction consistency compares the class probability distributions using JSD. Evidence consistency uses JSD to compare attention weights at matching tissue coordinates. Each term receives its own training weight.

3.2. Training Conditions and Attention Controls

B0 uses single-view ABMIL. B2 uses paired supervision and prediction consistency. E1 uses paired supervision and same-coordinate evidence consistency. E2 combines both consistency terms. The learned-attention arms share the encoder, tile collection, optimization schedule and model capacity. The paired arms receive the same original and perturbed views.

N1 retains the E2 training objective and permutes the second attention vector within the evidence term. The permutation preserves the set of attention weights while altering their coordinate assignment. N2 retains the B2 objective and uses an entropy-based constraint to match the concentration of E2 attention on development data. U assigns equal weight to every tile. For bags with identical tile support, uniform pooling produces identical attention vectors across stain views. Figure 1 shows the spatial organization of these conditions.

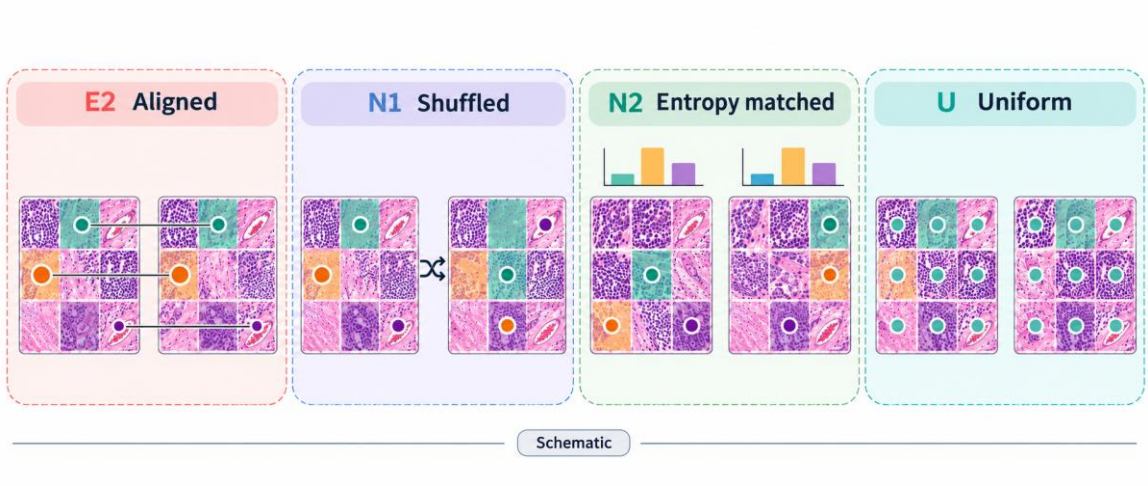


Figure 1 Spatial evidence alignment and its controls. Colored circles represent illustrative attention weights. U assigns identical weights to all tiles

3.3. Attention Concentration

We characterize each attention distribution with Shannon entropy, effective tile count and the fraction of attention mass carried by the top 10% of tiles. The entropy describes how spread out the weights are. The effective count summarizes the number of tiles contributing appreciably to pooling. The top-decile mass measures the weight carried by the highest-ranked tiles. These summaries are invariant to a permutation of tile coordinates, so they provide concentration checks alongside the spatial controls.

3.4. Matched Tissue Interventions

For each full tissue bag, the model ranks tiles by attention. At each budget, we select the highest-ranked tiles that meet the specified tissue area. We draw 20 matched control sets from the same slide, using tissue area, tissue fraction and spatial connectivity as matching criteria. The matching rules are fixed on training data. The trained classifier is frozen for both interventions.

For deletion, we remove the selected tiles and recompute the class score using the remaining embeddings. The deletion advantage measures how much more the score falls after removing the attention-ranked set than after removing a matched set on average. For retention, we recompute the score using only the selected embeddings. The retention advantage measures how much higher the score is for the attention-ranked set than for the matched sets on average. Figure 2 shows both interventions using equal-area selections.

Tissue area is the unit of intervention because equally sized image tiles can contain different amounts of tissue. Matching tissue fraction accounts for how much of each tile is occupied, and matching connectivity accounts for how the selected regions are arranged within the slide. This gives each ranked region a comparison with tissue of comparable area and spatial organisation. The two interventions use the same ranking from the full bag. The deletion score measures the contribution of the selected tissue when the rest of the bag remains available, while the retention score measures the signal carried by the selected tissue on its own.

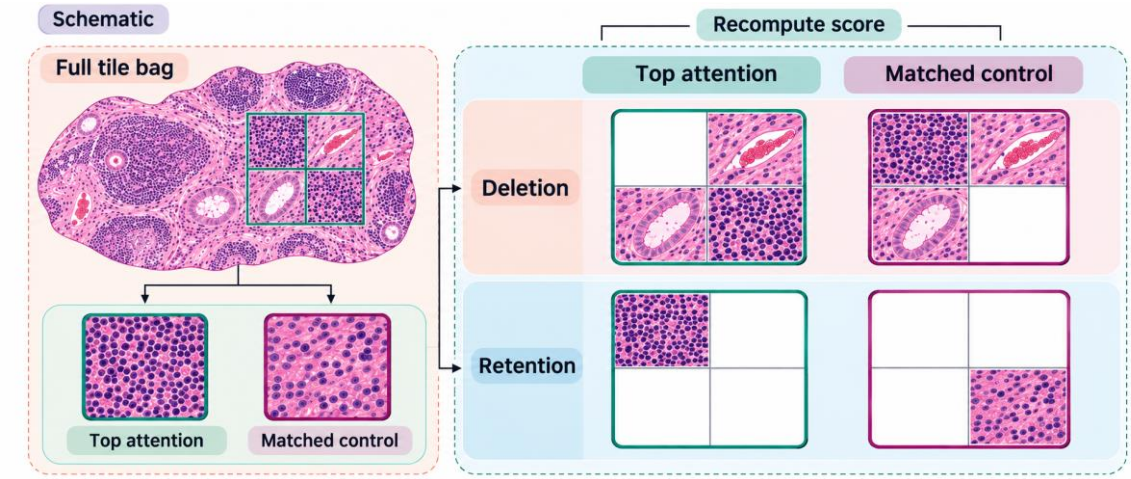


Figure 2 Matched deletion and retention. Each comparison uses equal-area tissue selections and recomputes the score for the resulting bag

4. Experimental Design

We evaluated case-level AUROC, F1 and accuracy over three training seeds. E2 versus B2 estimates the increment from adding evidence consistency to the prediction-consistency objective. E2 versus N1 tests the contribution of coordinate correspondence, and E2 versus N2 tests the contribution of correspondence under comparable attention concentration. U anchors the concentration and intervention comparisons.

The three classification measures capture different properties of the case scores. AUROC measures how well the scores rank cases across decision thresholds. F1 combines precision and recall at the classification operating point, while accuracy measures overall agreement with case labels. Reading these measures together tests whether the ordering seen in AUROC is also reflected in the assigned case labels. The attention concentration and intervention measurements then assess how the model uses tissue within each slide.

We used 10% tissue area as the main matched budget and evaluated deletion at 1%, 5%, 10% and 20% area. The normalized deletion area is the trapezoidal integral over the observed 1%–20% range divided by the width of that range. This gives each interval a weight proportional to its span. We compare the E2 minus B2 normalized deletion-area increment with the 0.02 criterion and assess deletion and retention directions at the matched 10% budget.

5. Results

5.1. Classification across the Seven Conditions

Table 1 shows that E2 achieved the highest AUROC, F1 and accuracy among the seven conditions. Relative to B2, E2 increased AUROC from 82.0% to 83.9%, F1 from 65.0% to 67.4% and accuracy from 66.9% to 69.1%. E1 achieved 81.9% AUROC. N1 and N2 achieved 81.6% and 82.1%, both close to B2, while E2 exceeded them by 2.3 and 1.8 percentage points. The uniform reference achieved 74.3% AUROC, establishing a separate performance point for identical attention across views.

Table 1 Case-level classification performance across three seeds (mean ± SD, %)

Condition	AUROC	F1	Accuracy
B0 Single-view ABMIL	80.1 ± 3.6	62.5 ± 4.8	64.8 ± 4.2
B2 Prediction consistency	82.0 ± 3.0	65.0 ± 4.3	66.9 ± 3.8
E1 Evidence consistency	81.9 ± 3.3	64.8 ± 4.6	66.7 ± 4.1
E2 Joint consistency	**83.9 ± 2.8**	**67.4 ± 4.0**	**69.1 ± 3.6**
N1 Coordinate shuffle	81.6 ± 3.4	64.6 ± 4.7	66.5 ± 4.1
N2 Entropy matching	82.1 ± 3.1	65.2 ± 4.4	67.0 ± 3.9
U Forced uniform attention	74.3 ± 5.2	56.8 ± 6.4	59.6 ± 5.9

The two single-constraint conditions differed by at most 0.2 percentage points across AUROC, F1 and accuracy. Combining the objectives increased these measures by 2.0, 2.6 and 2.4 percentage points, respectively, relative to E1. E2 also exceeded the entropy-matched N2 condition by 2.2 points in F1 and 2.1 points in accuracy. The ordering in AUROC therefore also appears in the two measures computed from case assignments.

5.2. Concentration under Entropy Matching

Table 2 shows that E2 and N2 had closely matched attention concentration. Entropy was 2.87 and 2.89, effective tile count was 12.4 and 12.6, and top-decile mass was 0.64 and 0.63. The 1.8-point AUROC difference between these arms accompanies these closely matched summaries. B2 assigned 0.41 of attention mass to the top decile, and U assigned 0.10. The uniform reference reached 4.61 entropy and an effective count of 100.0, giving the concentration comparison an interpretable reference.

Table 2 Attention concentration across the seven conditions

Condition	Attention entropy	Effective tiles	Top-10% mass
B0 Single-view ABMIL	3.71	41.2	0.33
B2 Prediction consistency	3.42	24.8	0.41
E1 Evidence consistency	3.44	25.3	0.40
E2 Joint consistency	**2.87**	**12.4**	**0.64**
N1 Coordinate shuffle	3.39	23.9	0.42
N2 Entropy matching	2.89	12.6	0.63
U Forced uniform attention	4.61	100.0	0.10

The learned-attention conditions showed three concentration profiles. The single-view baseline assigned one third of its attention mass to the top decile and used 41.2 effective tiles. The B2, E1 and N1 conditions placed 40% to 42% of attention in that decile, with effective tile counts between 23.9 and 25.3. E2 and N2 formed the most concentrated pair, each assigning a majority of attention to its top decile. This grouping places the entropy-matched comparison within the full range of attention concentration observed in the experiment.

5.3. Deletion and Retention at Matched Area

Table 3 shows that E2 achieved the highest deletion and retention advantages at the 10% budget. Its deletion advantage of 0.068 exceeded B2 by 0.027 and N2 by 0.028. Its retention advantage of 0.061 exceeded B2 by 0.023 and N2 by 0.024. The matched-area contrasts therefore favored E2 in both intervention directions. U yielded deletion and retention advantages of 0.006 and 0.004. E2 also achieved the highest reported difference relative to random selection, at 0.034.

Table 3 Matched tissue intervention at 10% area using 20 control sets

Condition	Deletion Δ	Retention Δ	Δ vs. random
B0 Single-view ABMIL	0.028	0.025	0.012
B2 Prediction consistency	0.041	0.038	0.019
E2 Joint consistency	**0.068**	**0.061**	**0.034**
N1 Coordinate shuffle	0.039	0.036	0.018
N2 Entropy matching	0.040	0.037	0.019
U Forced uniform attention	0.006	0.004	0.001

5.4. Deletion across Tissue Budgets

Table 4 and Figure 3 show that E2 led B2 and N2 at all four measured deletion budgets. The E2 minus B2 difference increased from 0.013 at 1% to 0.021 at 5% and 0.027 at 10%, then measured 0.025 at 20%. The B2 and N2 curves were separated by 0.001 at each sampled budget. Integrating the reported points gave normalized deletion areas of 0.0593 for E2, 0.0357 for B2 and 0.0347 for N2. The E2 minus B2 area gain was 0.0236, above the 0.02 criterion. At the primary budget, both the deletion and retention increments favored E2.

Table 4 Deletion advantage at four tissue-area budgets

Condition	1%	5%	10%	20%
B2 Prediction consistency	0.018	0.031	0.041	0.039
N2 Entropy matching	0.017	0.030	0.040	0.038
E2 Joint consistency	**0.031**	**0.052**	**0.068**	**0.064**

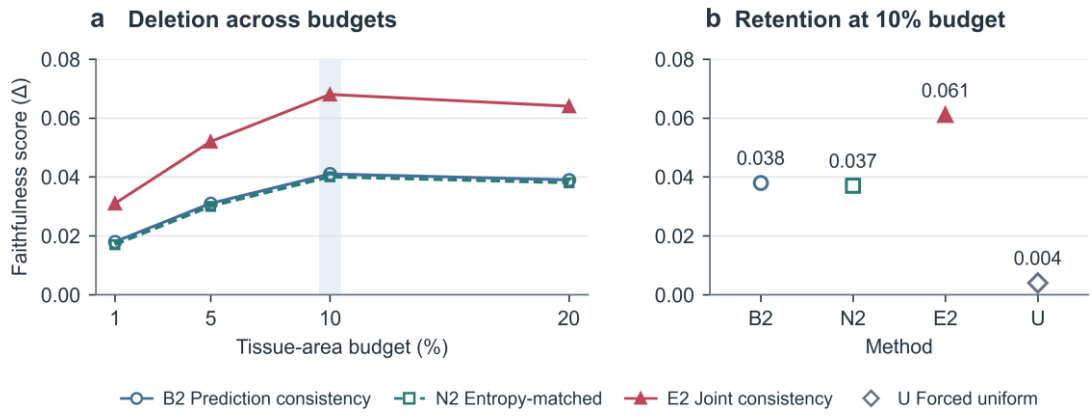


Figure 3 Deletion advantage across four measured tissue-area budgets and retention advantage at 10% area. The shaded band marks the primary budget

At 5% tissue area, E2 reached a deletion advantage of 0.052. This was higher than the largest advantage observed for either B2 or N2 at any of the four budgets. At 20%, the E2 score remained 0.023 above the highest score attained by B2. The two comparisons show that the E2 advantage appears at a smaller tissue selection and continues when the intervention covers a larger region.

The normalized deletion-area comparison also favored E2 over the entropy-matched control by 0.0246. The advantage over B2 was positive on each of the three intervals between measured budgets. The intervals up to 10% contributed about 0.0099 to the normalized gain, and the 10% to 20% interval contributed about 0.0137. The latter interval covers a larger span of tissue area, so its larger contribution reflects both its width and the sustained separation between the observed

deletion scores.

6. Conclusion

Same-coordinate attention consistency improved case-level discrimination and the score-based faithfulness of the top-ranked tissue under matched interventions. The E2 and N2 arms achieved similar entropy, effective tile count and top-decile attention mass, while E2 achieved higher AUROC and stronger deletion and retention effects. The coordinate-shuffle and uniform controls complement these comparisons by testing spatial correspondence and providing a reference for maximal cross-view attention agreement.

Acknowledgement

Liaoning Province College Students Innovation and Entrepreneurship Training Program.

References

[1] Tellez D ,Litjens G ,Bándi P , et al. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology[J].Medical Image Analysis, 2019,58:101544. DOI:10.1016/j.media.2019.101544.

[2] Ilse M ,Tomczak J M ,Welling M .Attention-based deep multiple instance learning[C].Proceedings of the 35th International Conference on Machine Learning,2018,80:2127-2136.

[3] Mirzazadeh A ,Dubost F ,Pike M , et al. ATCON: attention consistency for vision models[C].IEEE/CVF Winter Conference on Applications of Computer Vision,2023:1880-1889.DOI:10.1109/WACV56688.2023.00192.

[4] Liu J ,Li H ,Yang C , et al.HASD: hierarchical adaption for pathology slide-level domain-shift[C].Medical Image Computing and Computer Assisted Intervention,2026,15965:332-342.DOI:10.1007/978-3-032-04978-0_32.

[5] Adebayo J ,Gilmer J ,Muelly M , et al. Sanity checks for saliency maps[C].Advances in Neural Information Processing Systems,2018,31.

[6] Jamshidi Idaji M ,Hense J ,Neuhäuser T , et al. Beyond attention heatmaps: how to get better explanations for multiple instance learning models in histopathology[J]. Medical Image Analysis, 2026, 113:104148. DOI:10.1016/j.media.2026.104148.