This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

Annotation-Efficient Conditional Diffusion for Subacute Stroke Segmentation in West African Non-Contrast CT: A Calibration-First Evaluation

Joseph D. A. Nti Koduah[1,2] [0009−0001−5910−0693], Prince Ebenezer Adjei[1,2,3] [0000−0001−6244−2582], Belinda Esinam Akpaloo[5], Kojo Obed Otoo[5] [0000−0001−8828−3188], Livingstone E. Ayivor[1,2] [0009−0008−3032−204X], and John Amuasi[2,3,4] [0000−0002−8640−2662]

[1] Department of Computer Engineering, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana.
[2] Global Health and Infectious Disease Group, Kumasi Centre for Collaborative Research in Tropical Medicine, Kumasi, Ghana.
[3] Global One Health Group, Bernhard Notch Institute of Tropical Medicine (BNITM), Hamburg, Germany
[4] Department of Global Health, School of Medical Sciences, Kwame Nkrumah University of Science and Technology, Kumasi, Ghana
[5] Department of Radiology, Komfo Anokye Teaching Hospital, Kumasi, Ghana
peadjei.coe@knust.edu.gh

Abstract. Non-contrast CT (NCCT) is the primary imaging modality for acute stroke in low- and middle-income countries (LMICs). However, in our West African clinical cohort of 316 scans, only 29 carry expert annotations, proving that specialist labeling, not imaging access, is the bottleneck for AI development. We present the first study applying conditional diffusion models to West African stroke data, investigating annotation efficiency using MedSegDiff-V2 across clinical budgets of $N \in \{5, 10, 15, 20, 29\}$ patients. To balance performance and deployment constraints, we adopt a 2.5D conditioning scheme under fixed training compute. Uniquely, we treat the model's native stochastic K-pass sampling as a core evaluation axis, assessing Expected Calibration Error (ECE) and predictive entropy alongside standard overlap metrics. Reporting on a pilot subset (Fold 0, $K = 3$, 2.5D), we analyze high estimator variance and poor calibration. We frame this miscalibration not as a failure, but as a crucial clinical safety signal, arguing that knowing when a model is uncertain is vital in settings lacking routine specialist review.

Keywords: Stroke Segmentation, Conditional Diffusion, Annotation-Efficient Learning, MedSegDiff-v2, West-Africa, Uncertainty Calibration

1 Introduction

Stroke remains a leading cause of mortality and disability in sub-Saharan Africa [2], where non-contrast CT (NCCT) serves as the primary acute neuroimaging modality. While physical scanner availability is often cited as the principal barrier to

AI deployment, clinical workflows in low- and middle-income countries (LMICs) face a fundamentally different bottleneck. In our West African clinical cohort of 316 stroke NCCT scans, only 29 carry expert annotations. In this environment, expert labeling—rather than raw data acquisition—presents the critical limiting factor, driven by severe specialist radiologist constraints relative to scanner throughput [5, 9]. Consequently, automated segmentation models intended for deployment in resource-limited settings must be evaluated explicitly as a function of the annotation budget, rather than assumed to operate under the idealized, high-resource conditions of standard benchmarks [10].

Diffusion-based segmentation architectures, such as MedSegDiff-V2 (a transformer-based model), have achieved state-of-the-art performance on public medical benchmarks including BRATS and ISIC, while yielding calibrated, sample-based uncertainty estimates as a native byproduct of generative sampling [13]. However, standard benchmarks rely on annotated datasets that far exceed what clinical pipelines in LMICs can realistically support. Furthermore, conventional evaluation paradigms neither hold training compute fixed while scaling annotation volume nor prioritize calibration quality as a primary clinical objective. Rigorously characterizing diffusion models under a jointly annotation- and compute-constrained regime requires an experimental framework centered on these constraints from the outset.

In this work, we address these challenges through two main contributions:

1. **Annotation-Efficiency Benchmark:** We report a single pilot benchmark for MedSegDiff-V2 stroke lesion segmentation on a previously unstudied West African NCCT cohort. By fixing total training compute across annotation budgets $N \in \{5, 10, 15, 20, 29\}$, we isolate dataset size as the primary independent variable.
2. **Uncertainty-Quantified Evaluation:** We elevate stochastic K-pass diffusion sampling to a co-primary evaluation axis [11]. By reporting Expected Calibration Error (ECE) [3] and predictive entropy alongside traditional spatial overlap metrics, we prioritize clinical safety in settings lacking routine specialist oversight, where a model's ability to signal its own unreliability is vital.

We evaluate this pipeline using a 2.5D axial conditioning scheme, which captures local through-plane spatial context at a substantially reduced computational cost compared to full 3D convolutions [8, 14]. Using $K = 3$ stochastic inference passes, this evaluation provides a baseline for architectural scaling and pseudo-labeling on unlabeled cohort volumes, as detailed in Section 4.3.

2 Method

2.1 Clinical Cohort and Data

We use data from a teaching hospital in West Africa, through an ongoing clinical research collaboration. The cohort comprises 316 non-contrast CT (NCCT)

stroke scans: 29 expert-annotated and 287 unlabeled. This 29:287 imbalance reflects severe resource constraints in regional clinical pipelines and is the central motivation for this work. The annotated subset spans a clinically representative range of lacunar (small, deep) and territorial (large, cortical) infarcts, introducing substantial variability in lesion volume, location, and attenuation contrast. All scans are volumetric NCCT series with 5.0 mm slice thickness. For model input, volumes are resampled in-plane to 1.0×1.0 mm isotropic resolution, center-cropped or zero-padded to 256×256, and intensity-windowed to $[0, 80]$ HU [7]. Preprocessed volumes are cached offline as per-patient HDF5 files; the data loader reads each volume into memory on first access and assembles 2.5D axial slice triplets dynamically via an in-memory sliding window, eliminating redundant disk reads.

2.2 Model Architecture: 2.5D Conditional Diffusion Segmentation

Our architecture builds on MedSegDiff-V2 [12,13], which frames segmentation as conditional generative modeling [1]: the target lesion mask is treated as a noisy channel and iteratively denoised conditioned on the input image via channel-wise concatenation, bypassing dedicated encoder-decoder conditioning branches. MedSegDiff-V2 further modulates intermediate decoder features using a dynamic conditional encoding module ("FF-Parser").

To provide local through-plane context without the cost of full 3D convolutions [8], predictions are conditioned on a three-slice axial triplet, the target slice plus its immediate superior and inferior neighbors, giving a total input depth $C = 4$ (3 image channels + 1 noise/mask channel). The first convolutional layer is sized accordingly within a standard denoising U-Net backbone (128 base channels, 2 residual blocks per resolution, single-head self-attention at 16×16, learned reverse-process variance).

2.3 Training Protocol and Annotation-Budget Sweep

Annotation-Efficiency Sweep. We train independent models across subset sizes $N \in \{5, 10, 15, 20, 29\}$ under 5-fold cross-validation, with held-out validation patients strictly excluded from training at every N within their fold. Subsets at each N are drawn independently within each fold via stratified sampling by lesion category, with sampling variance mitigated through cross-fold aggregation. At $N = 29$, the training subset necessarily comprises the full annotated cohort, including the fold's held-out validation patients: with only 29 total annotated patients and a per-fold training pool of roughly 23–24 after holding out validation cases, no strictly disjoint 29-patient training set exists within a fold. Results at $N = 29$ therefore reflect training-set rather than held-out generalization performance, and should be interpreted accordingly.

Fixed-Compute Regime. To isolate annotation budget as the primary independent variable, we fix the total number of gradient steps across all runs. Models trained on smaller N consequently see individual patient volumes more fre-

quently. Results characterize performance under fixed compute, not fixed epoch count.

Optimization. Networks are optimized with AdamW (learning rate 5×10^{-5}, batch size 16) under automatic mixed precision. The diffusion process uses a linear noise schedule over 1000 timesteps [4] with learned reverse-process variance ("learn-sigma"). EMA weights are tracked and used for all inference evaluations. All experiments were run on two 48GB RTX 8000 GPUs (Python 3.10), with each run bounded by the fixed gradient-step budget above rather than wall-clock time.

2.4 Uncertainty-Quantified Inference and Evaluation Metrics

Stochastic K-Pass Inference. At inference, we exploit the diffusion model's native stochasticity to quantify predictive uncertainty: for each target volume, we draw $K = 3$ independent reverse-diffusion trajectories, each initialized from unique Gaussian noise in the mask channel under identical image conditioning. The voxel-wise mean probability across passes gives the point prediction (thresholded at 0.5); the voxel-wise standard deviation gives an empirical uncertainty map. We additionally report a sensitivity comparison against $K = 5$ passes (Section 3.2).

Sampling Implementation. MedSegDiff-V2 supports both standard and DDIM-accelerated sampling, but combining DDIM acceleration with timestep respacing produced index-bounds errors under our conditioning setup. All evaluations therefore use the full conditional sampler on the native 1000-step linear schedule, executing 50 reverse steps per sample (`step=50`).

Evaluation Metrics. Volumetric segmentations are reconstructed by stacking predicted 2.5D axial slice masks along the through-plane axis. Segmentation accuracy is evaluated against expert annotations using Dice Similarity Coefficient (DSC), 95th-percentile Hausdorff Distance (HD95, at $1.0 \times 1.0 \times 5.0$ mm voxel spacing), and mean Intersection-over-Union (mIoU). For predictive uncertainty, we compute mean predictive entropy and Expected Calibration Error (ECE) [3] across all volume voxels using equal-width confidence binning.

3 Results

3.1 Pilot Segmentation Performance

We report pilot segmentation performance from a single cross-validation fold (Fold 0) utilizing the MedSegDiff-V2 2.5D conditioning architecture evaluated under $K = 3$ stochastic inference passes (Table 1).

The observed DSC performance trajectory is non-monotonic across dataset tiers: performance decreases at $N = 10$ relative to $N = 5$ ($0.430 \rightarrow 0.262$)

Table 1: Pilot segmentation performance on Fold 0 (MedSegDiff-V2 2.5D, $K = 3$). Values denote mean ± standard deviation across held-out patient volumes ($n = 6$ per evaluation budget).

N	DSC	HD95 (mm)	mIoU	Entropy	ECE
5	0.430 ± 0.009	71.3 ± 7.1	0.274 ± 0.008	0.446 ± 0.003	0.513 ± 0.10
10	0.262 ± 0.018	74.9 ± 27.4	0.151 ± 0.012	0.343 ± 0.008	0.619 ± 0.09
15	0.335 ± 0.066	73.5 ± 20.1	0.203 ± 0.045	0.420 ± 0.009	0.603 ± 0.05
20	0.510 ± 0.066	65.2 ± 5.2	0.345 ± 0.056	0.475 ± 0.017	0.484 ± 0.10
29*	0.556 ± 0.043	66.3 ± 4.7	0.386 ± 0.041	0.429 ± 0.014	0.485 ± 0.09

*At N=29, the training subset includes held-out fold cases due to cohort size constraints; values reflect training-set evaluation (Section 2.3).

before recovering progressively from $N = 15$ through $N = 20$ ($0.335 \rightarrow 0.510$). As these initial findings represent a single-fold pilot evaluation, we contextualize this trend as an empirical observation warranting full cross-validation verification rather than asserting a definitive scaling law.

Case-level anatomical heterogeneity represents a primary driver of validation variance. For instance, Patient 325 consistently yielded the lowest DSC across all evaluation budgets (ranging from 0.167 to 0.473), whereas the remaining held-out cases spanned 0.222 to 0.607 across the same training regimes. This persistent, budget-independent performance gap suggests underlying clinical confounders, such as subtle attenuation contrast or small lacunar lesion morphology, rather than optimization failure.

Figure 1 illustrates this qualitative contrast at $N = 29$ between Patient 325 (DSC = 0.473) and a higher-performing subject, Patient 341 (DSC = 0.603). In Patient 325, low attenuation contrast leads to incomplete lesion coverage and elevated uncertainty along poorly defined margins. Conversely, Patient 341 exhibits high lesion conspicuity, resulting in strong spatial alignment with expert ground truth and uncertainty tightly localized to the lesion boundary.

3.2 Uncertainty Calibration and Stochastic Pass Sensitivity

Voxel-wise ECE remained elevated across all evaluated annotation budgets (0.484–0.619), demonstrating pronounced probability miscalibration. In clinical workflows lacking immediate specialist radiological backup, reliable identification of low-confidence predictions is as actionable as the segmentation itself; we therefore report this persistent miscalibration as a primary baseline safety metric.

To analyze the effect of the inference sampling budget on prediction accuracy and calibration stability, we evaluate $K = 3$ versus $K = 5$ stochastic sampling passes at dataset budgets $N = 5$ and $N = 10$ (Table 2).

Expanding the sampling pool from $K = 3$ to $K = 5$ increases mean predictive entropy and slightly depresses thresholded DSC (e.g., $0.430 \rightarrow 0.400$ at $N = 5$), while ECE remains essentially static ($0.513 \rightarrow 0.510$ at $N = 5$; $0.619 \rightarrow 0.617$

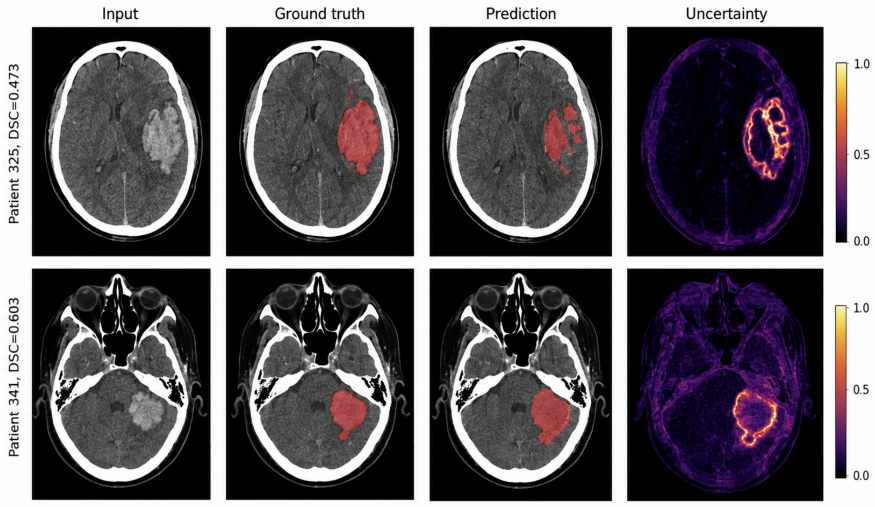

Fig. 1: Qualitative segmentation and uncertainty mapping at $N = 29$ (Fold 0, 2.5D) for representative lower- and higher-performing cases (IDs 325 and 341; DSC of 0.473 and 0.603, respectively). Columns from left to right: input NCCT slice, expert ground truth, mean predictive mask across $K = 3$ stochastic passes, and voxel-wise uncertainty map (predictive standard deviation).

at $N = 10$). Additional stochastic passes inflate predictive variance without correcting underlying model confidence, indicating that inference scaling alone is insufficient to resolve calibration drift.

3.3 Scope of Pilot Evaluation

The findings reported above establish the baseline performance profile on Fold 0 ($K = 3$, 2.5D conditioning). Complete multi-fold cross-validation aggregation, scaling to $K = 10$ sampling passes, direct 2D architectural comparison, and pseudo-labeling across the 287 unlabeled cohort volumes are outlined in Section 4.3.

4 Discussion

The pilot findings (Section 3.1) validate the end-to-end functionality of the proposed pipeline, encompassing clinical data ingestion, conditional diffusion training under extreme data scarcity, and K-pass uncertainty-aware inference on a previously unstudied West African NCCT stroke cohort. At this stage, these metrics demonstrate technical feasibility rather than a fully resolved characterization of annotation efficiency.

The observed non-monotonic DSC trajectory, characterized by a performance drop at $N = 10$ followed by recovery through $N = 20$, represents an empirical

Table 2: Sampling pass sensitivity analysis ($K = 3$ vs. $K = 5$) on Fold 0. Values denote fold-level mean ± standard deviation across held-out cases.

N	K	DSC	Entropy	ECE
5	3	0.430 ± 0.009	0.446 ± 0.003	0.513 ± 0.10
5	5	0.400 ± 0.008	0.492 ± 0.004	0.510 ± 0.11
10	3	0.262 ± 0.018	0.343 ± 0.008	0.619 ± 0.09
10	5	0.231 ± 0.016	0.384 ± 0.011	0.617 ± 0.10

pilot signal. Without multi-fold cross-validation aggregation, it remains unconfirmed whether the dip at $N = 10$ reflects cohort composition variance or residual estimator noise. Furthermore, performance at $N = 29$ represents training-set evaluation due to total cohort size constraints (Section 2.3), serving as an empirical upper bound rather than an out-of-sample generalization metric.

4.1 Calibration as a Primary Clinical Contribution

In resource-constrained clinical settings lacking continuous specialist radiological coverage, an AI system's ability to signal its own unreliability is a vital safety requirement [6]. Clinical workflows can readily defer or flag low-confidence predictions, whereas high-confidence incorrect segmentations present severe diagnostic risk. Across all evaluated budgets, Expected Calibration Error remained persistently elevated (ECE 0.484–0.619), establishing model overconfidence as a primary baseline safety diagnostic rather than a secondary artifact [3].

Sensitivity analysis across inference sampling budgets (Table 2) reveals that expanding from $K = 3$ to $K = 5$ stochastic passes alters mean predictive entropy without correcting calibration drift ($0.513 \rightarrow 0.510$ at $N = 5$; $0.619 \rightarrow 0.617$ at $N = 10$). Drawing additional reverse-diffusion trajectories increases spatial uncertainty (entropy) without re-aligning confidence with accuracy, demonstrating that inference scaling alone is insufficient to resolve underlying miscalibration.

4.2 Limitations

We identify several explicit scope boundaries in this pilot study:

1. **Single-Fold Evaluation Scope:** Benchmarks reflect a single validation fold (Fold 0); establishing aggregate performance trends requires complete multi-fold cross-validation.
2. **Training-Set Scope at $N = 29$:** Due to total cohort size constraints, the $N = 29$ training subset includes held-out validation cases, reflecting training-set evaluation rather than generalization performance.
3. **Inference Sampling Budget:** Primary evaluations utilize $K = 3$ stochastic passes (with $K = 5$ as a sensitivity check); determining whether larger sampling budgets stabilize spatial uncertainty or improve ECE remains unexamined.

4. **Fixed-Compute Normalization:** Normalizing training by total gradient steps means lower-N models process individual volumes with higher frequency, evaluating performance under constant compute rather than equal epoch exposure.

4.3 Future Work

Planned extensions focus on expanding this pilot into a comprehensive benchmark framework: (1) completing full five-fold cross-validation to verify the reproducibility of the observed performance trajectory; (2) conducting case-level diagnostic error analyses to isolate clinical confounders; (3) executing a formal 2D versus 2.5D architectural ablation under edge-deployment compute constraints; (4) evaluating larger stochastic sampling budgets ($K \geq 10$); (5) implementing stability-filtered pseudo-labeling across the cohort's 287 unlabeled scans; and (6) characterizing the algorithmic interactions between timestep respacing and conditional sampling paths.

5 Conclusion

This pilot study evaluates annotation efficiency and uncertainty calibration for diffusion-based stroke lesion segmentation on a previously unstudied West African NCCT cohort. By fixing training compute across annotation budgets and elevating calibration to a co-primary evaluation metric, our protocol provides a resource-normalized baseline on a single validation fold. Two key findings emerge. First, the model exhibits persistent probability miscalibration (ECE 0.484–0.619), which remains largely unchanged when expanding the sampling budget from $K = 3$ to $K = 5$, establishing an actionable safety baseline for low-resource deployment. Second, held-out segmentation performance follows a non-monotonic trajectory, dipping at $N = 10$ before recovering through $N = 20$, while the $N = 29$ result represents training-set evaluation due to cohort size constraints.

Extending this framework across all cross-validation folds, scaling inference sampling passes, completing a 2D versus 2.5D architectural ablation, and leveraging the cohort's 287 unlabeled scans through pseudo-labeling represent clear priorities for future work. Beyond the immediate pilot findings, combining fixed-compute annotation sweeps with calibration-focused evaluation offers a practical, resource-aware framework for validating medical imaging models under genuine clinical constraints.

Acknowledgments. The authors thank the Department of Radiology at Komfo Anokye Teaching Hospital for providing access to the clinical data used in this study. This work was supported by the Kwame Nkrumah University of Science and Technology Research Fund (KReF).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Amit, T., Shaharbany, T., Nachmani, E., Wolf, L.: Segdiff: Image segmentation with diffusion probabilistic models. arXiv preprint arXiv:2112.00390 (2021)
2. Attakorah, J., Mensah, K.B., Yamoah, P., Wiafe, E., Bangalee, V., Oosthuizen, F.: A systematic review of the burden of stroke in ghana. BioMed Research International (1), 8298154 (2024). https://doi.org/10.1155/2024/8298154
3. Guo, C., Pleiss, G., Sun, Y., Weinberger, K.Q.: On calibration of modern neural networks. In: Proceedings of the 34th International Conference on Machine Learning. pp. 1321–1330. PMLR (2017)
4. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems **33**, 6840–6851 (2020)
5. Jin, C., Guo, Z., Lin, Y., Luo, L., Chen, H.: Label-efficient deep learning in medical image analysis: Challenges and future directions. arXiv preprint arXiv:2303.12484 (2025)
6. Kohl, S., Romera-Paredes, B., Meyer, C., De Fauw, J., Ledsam, J.R., Maier-Hein, K.H., Eslami, S.M.A., Rezende, D.J., Ronneberger, O.: A probabilistic u-net for segmentation of ambiguous images. Advances in Neural Information Processing Systems **31**, 6965–6975 (2018)
7. Luo, J., Dai, P., He, Z., Huang, Z., Liao, S., Liu, K.: Deep learning models for ischemic stroke lesion segmentation in medical images: A survey. Computers in Biology and Medicine (ScienceDirect) (2024). https://doi.org/10.1016/j.compbiomed.2024.108509
8. Maolin Li, C.Z., Cao, S.: 2d, 2.5d, or 3d? comparing dimensional approaches in deep neural networks for 3d medical image analysis. Journal of Imaging Informatics in Medicine (2025). https://doi.org/10.1007/s10278-025-01827-6, systematic review of 31 comparative studies
9. Ofori-Manteaw, B., Ofori, E., Dzefi-Tettey, K., Ago, J., Angmorterh, S., F, A.: Radiologist workforce challenges and the burden of image interpretation in ghana: Perspectives of frontline doctors and implications for healthcare delivery. PLOS Global Public Health (2026). https://doi.org/10.1371/journal.pgph.0006456
10. Wang, H., Jin, Q., Li, S., Liu, S., Wang, M., Song, Z.: A comprehensive survey on deep active learning in medical image analysis. Medical Image Analysis (ScienceDirect) (2024)
11. Wolleb, J., Sandkühler, R., Bieder, F., Valmaggia, P., Cattin, P.C.: Diffusion models for implicit image segmentation ensembles. In: Proceedings of The 5th International Conference on Medical Imaging with Deep Learning. pp. 1336–1348. PMLR (2022)
12. Wu, J., Fang, H., Zhang, Y., Yang, Y., Xu, Y.: Medsegdiff: Medical image segmentation with diffusion probabilistic model. arXiv preprint arXiv:2211.00611 (2022)
13. Wu, J., Ji, W., Fu, H., Xu, M., Jin, Y., Xu, Y.: Medsegdiff-v2: Diffusion based medical image segmentation with transformer. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 38, pp. 6030–6038 (2024)
14. Yichi Zhang, Q.L., Ding, L., Zhang, J.: Bridging 2d and 3d segmentation networks for computation-efficient volumetric medical image segmentation: An empirical study of 2.5d solutions. Computerized Medical Imaging and Graphics (ScienceDirect) (2022), https://www.sciencedirect.com/science/article/abs/pii/S0895611122000611