Synthesizing Post-Acetazolamide Cerebral Blood Flow Maps from Baseline MRI in Moyamoya Using 3D Generative AI

Julia Huang[1] JULIH@STANFORD.EDU
Camila Gonzalez[2,3] CAMGONZA@STANFORD.EDU, CAMILA.GONZALEZ@MEDUNIWIEN.AC.AT
Rydham Goyal[1] RYDHAM@STANFORD.EDU
Aja Zou[4] AJAZOU@STANFORD.EDU
Sasha Alexander[4] SASHALEX@STANFORD.EDU
Michael Moseley[2] MOSELEY@STANFORD.EDU
Moss Y. Zhao[4*] MOSSZHAO@STANFORD.EDU
Gary K. Steinberg[4,*] CEREBRAL@STANFORD.EDU

[1] *Department of Computer Science, Stanford School of Engineering, Stanford, CA 94305*
[2] *Department of Radiology, Stanford School of Medicine, Stanford, CA 94305*
[3] *Department of Anesthesia, Intensive Care Medicine, and Pain Medicine, Medical University of Vienna, Vienna, Austria 1090*
[4] *Department of Neurosurgery, Stanford School of Medicine, Stanford, CA 94305*

Abstract

For patients with Moyamoya disease, impaired cerebrovascular reserve (CVR) is an important hemodynamic criterion for recommending extracranial-to-intracranial bypass surgery, making reliable CVR assessment central to treatment planning. The reference protocol used in this cohort requires paired arterial spin labeling (ASL) perfusion MRI acquired before and after administration of the vasodilator acetazolamide (ACZ). ACZ may be contraindicated or avoided in patients with substantial renal dysfunction, relevant hypersensitivity, marked electrolyte derangement, or pregnancy, depending on clinical circumstances and institutional protocol. For affected patients in whom ACZ is not administered, the standard two-scan protocol cannot be completed as intended, and the post-ACZ cerebral blood flow (CBF) data used for hemodynamic assessment and bypass planning are unavailable. We propose **CAE3D**, a deterministic 3D conditional autoencoder that synthesizes post-ACZ CBF maps directly from pre-ACZ ASL input. Among eleven evaluated models (CAE3D, seven directly comparable full-volume in-house baselines spanning deterministic and diffusion-style variants, one middle-slice 2D contextual baseline (CAE_2D), and two frozen-encoder foundation adapters reported separately as contextual references), CAE3D achieves the lowest held-out MAE (MAE 0.066, SSIM 0.80, PSNR 24.0 dB) with near-zero full-brain mean bias, though its regional ΔCBF predictions compress the dynamic range in high-response territories. Its MAE advantage was statistically significant (paired Wilcoxon, Holm-adjusted) over seven of the eight other trained-from-scratch baselines, with the exception of the 2D middle-slice CAE_2D comparator; its SSIM and PSNR advantages were significant over all eight. These results establish the retrospective feasibility of post-ACZ CBF synthesis in patients who completed the standard two-scan protocol; extension to ACZ-contraindicated patients, who were not represented in this cohort, awaits external and prospective validation.

[*] These authors share senior authorship.

1. Introduction

Moyamoya disease is a rare, progressive cerebrovascular disorder in which the major intracranial arteries gradually narrow and occlude, raising the long-term risk of ischemic and hemorrhagic stroke (Scott and Smith, 2009). Cerebrovascular reserve (CVR) reflects the brain's capacity to augment cerebral blood flow (CBF) in response to a vasodilatory stimulus. It is a key hemodynamic marker for disease severity in Moyamoya (Rao et al., 2022), and impaired CVR is an important hemodynamic criterion for recommending extracranial-to-intracranial bypass surgery over conservative management (Rao et al., 2022; Zhao et al., 2022).

In clinical practice, CVR is assessed using an acetazolamide (ACZ) challenge: paired pre- and post-ACZ ASL perfusion maps are an established clinical approach to CVR assessment (Zhao et al., 2022; Rao et al., 2022). However, ACZ may be contraindicated or avoided in patients with substantial renal dysfunction, clinically relevant hypersensitivity (including possible cross-sensitivity with sulfonamides), marked electrolyte imbalance, or pregnancy, depending on clinical circumstances and institutional protocol (Teva Pharmaceuticals, 2019; Rao et al., 2022; Zhao et al., 2022); this study's specific exclusion thresholds are given in §3.1. These contraindication categories may arise in patients undergoing Moyamoya evaluation, although their prevalence cannot be estimated from the present cohort. Moreover, even when no formal contraindication exists, the two-scan protocol may be difficult to complete due to acute illness, scheduling, high cost, or visit-length constraints. In either case, the treating team is left without one of the key hemodynamic inputs used in bypass planning.

To address this clinical gap, synthesizing post-ACZ CBF maps from baseline pre-ACZ ASL input using machine learning could offer a path toward CVR-relevant information without drug administration, pending validation. Encoder-decoder networks, diffusion models, and pretrained 3D encoders have each shown promise for this class of synthesis task in neuroimaging (Kazerouni et al., 2023; Özbey et al., 2023; Guo et al., 2020; Hussein et al., 2024). Goyal et al. (2026) established a 2D conditional synthesis approach for post-ACZ maps from pre-ACZ ASL. We extend that approach to full-volume 3D. We propose **CAE3D**, a 3D conditional autoencoder illustrated in Figures 1–2, and evaluate it against ten comparators under the model-count convention defined in §3 (eleven models evaluated in total, including CAE3D). The contribution of CAE3D lies in its task-specific adaptation and rigorous comparative evaluation of what is, to our knowledge, the first reported full-volume 3D study of this specific pre-ACZ-to-post-ACZ ASL synthesis task, rather than in a novel backbone architecture.

Contributions. This work makes two primary contributions. First, we demonstrate the feasibility of synthesizing post-ACZ CBF maps from pre-ACZ ASL input in patients with Moyamoya disease using 3D conditional image-synthesis models: **CAE3D** achieves the lowest held-out MAE among the nine in-house models in Table 1 (§3), with near-zero Bland-Altman bias and lower territory-level ΔCBF error than the evaluated diffusion baselines. Second, we extend the slice-based approach of Goyal et al. (2026) to full-volume 3D; CAE3D outperformed the middle-slice **CAE_2D** baseline on the reported metrics, though the two operate on different spatial domains and this comparison is supportive rather than

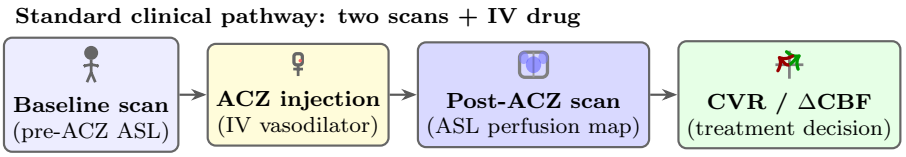


Figure 1: **Established clinical pathway.** The protocol used in this cohort requires two ASL scans (pre- and post-ACZ) bracketing an intravenous acetazolamide administration.

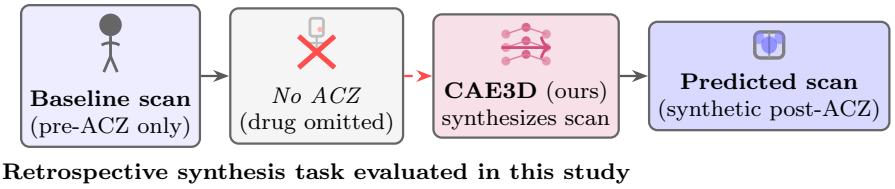


Figure 2: **Retrospective synthesis task evaluated in this study.** CAE3D predicts an independently normalized post-ACZ map from the pre-ACZ input. This diagram illustrates the potential future workflow motivation, not a clinically validated replacement for ACZ imaging.

a controlled dimensionality ablation. Implementation and evaluation scripts are available at https://github.com/TheClassicTechno/cae3d-moyamoya-cbf-synthesis.

Generalizable Insights about Machine Learning in the Context of Healthcare. For clinicians and ML researchers beyond this cohort, the results suggest that baseline-to-challenge image synthesis may warrant investigation in other clinical workflows involving paired baseline and pharmacologic-challenge acquisitions. Such extensions would require task-specific validation, particularly using agreement, regional, and subgroup-level metrics rather than global image-similarity scores alone. This single-center evaluation also offers three further methodological observations that may be relevant to related perfusion-synthesis settings, though not claimed to generalize beyond the evaluated models and protocols. First, deterministic 3D autoencoders performed better than the evaluated diffusion implementations in this dataset and training regime. Second, multimodal evaluation (voxel-level metrics with Bland-Altman agreement, bootstrap CIs, and vascular-territory ΔCBF summaries) exposed limitations not visible in global metrics alone and is offered as a candidate template for other paired perfusion-synthesis tasks. Third, the tested frozen-encoder adaptations did not outperform task-specific training.

2. Related Work

Background: ASL perfusion imaging. Arterial spin labeling (ASL) is a non-invasive MRI technique that uses magnetically labeled arterial blood water as an endogenous tracer to quantify cerebral blood flow (CBF) without exogenous contrast. Each ASL acquisition yields a voxelwise CBF map. The change between pre- and post-ACZ CBF maps, $\Delta\text{CBF} = (\text{CBF}_{\text{post}} - \text{CBF}_{\text{pre}})/\text{CBF}_{\text{pre}} \times 100\,\%$, encodes the regional vasodilatory response used in CVR assessment and may contribute to bypass planning. The synthesis task in this work is

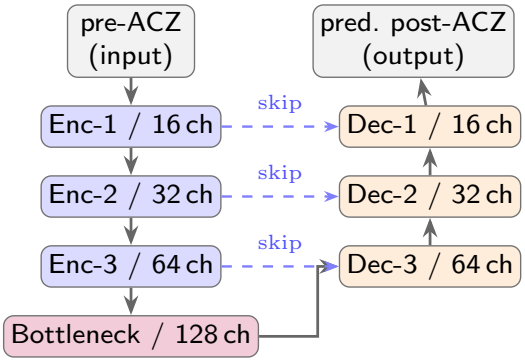


Figure 3: **CAE3D architecture.** Skip-connected 3D encoder-decoder (three downsampling levels).

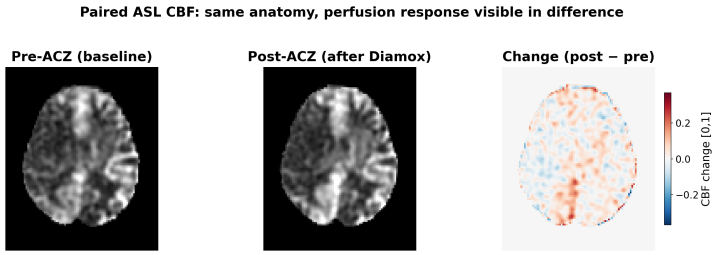


Figure 4: **Image context.** Pre-ACZ, post-ACZ GT, and post-minus-pre change (middle axial slice, one subject). Image panels use intensity windowing for visualization ($v_{\max}$ set to the 99th percentile of non-zero voxels); see the plotting scripts in the supplementary material for exact details.

therefore image-to-image regression: predicting the post-ACZ CBF map from the pre-ACZ CBF map in normalized voxel space, where each volume is independently rescaled to $[0, 1]$ (§3.1); CAE3D synthesizes post-ACZ maps in this independently normalized intensity space rather than in physical CBF units, which bears on how normalized-space ΔCBF relates to physiological CVR (§5).

CVR assessment with acetazolamide administration. Clinicians assess hemodynamic reserve in Moyamoya using pharmacologic vasodilation with CVR imaging (Rao

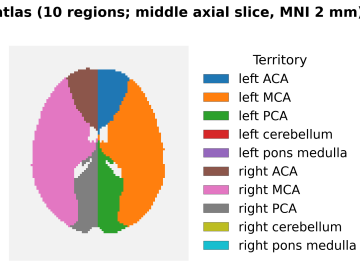


Figure 5: **Vascular territory atlas** (Liu et al., 2023) on a representative pre-ACZ slice; defines regional summaries (§3.4). Windowed for visualization as in Figure 4.

et al., 2022; Zhao et al., 2022); the practical constraints (time, logistics, cost, contraindications) that motivate computational alternatives are detailed in §1.

Post-ACZ perfusion synthesis from baseline ASL perfusion maps. Goyal et al. (2026) established slice-based conditional and diffusion baselines for this clinical problem, focusing on a 2D middle-slice prediction with a smaller model comparison. By contrast, the present study evaluates full-volume 3D prediction, includes side-by-side 2D and 3D baselines within a single protocol, and extends the evaluation with agreement and territory-level analyses. The objective (synthetic post-ACZ from pre-ACZ alone) is the same.

Physiological basis and scope of the learned mapping. CVR is, by definition, the response to a vasodilatory challenge, so it is fair to ask why a pre-ACZ baseline map should carry any information about it. In Moyamoya, chronic large-vessel occlusion drives collateral recruitment and territorial steal well before ACZ is administered, so resting-state (pre-ACZ) perfusion already reflects, in part, how exhausted or preserved a territory's autoregulatory reserve is (Rao et al., 2022; Zhao et al., 2022). This motivates a statistical association between baseline perfusion patterns and vasodilatory response at the cohort level. We do not claim that two patients with visually identical pre-ACZ maps necessarily share identical CVR, nor that the model recovers an individual patient's exhausted-but-compensated reserve from the baseline scan alone; such cases are, by construction, the hardest for any baseline-only predictor. Rather, CAE3D learns the population-level correlation between baseline CBF patterns and average post-ACZ response within this Moyamoya cohort, not subject-specific CVR recovery. We return to the evidence for this distinction, including the modest per-subject R^2 and territory-level ΔCBF compression, in §5.

Broader perfusion and multimodal synthesis. Learning-based synthesis of perfusion-related targets from baseline imaging has also been explored in MRI–PET settings (Hussein et al., 2024, 2022; Dayarathna et al., 2024). Foundation models pretrained on large medical imaging datasets are increasingly applied to synthesis tasks; we include two publicly released 3D encoders as frozen-encoder comparators, **Med3DVLM** (Xin et al., 2025) and **SAM-Med3D** (Wang et al., 2023), with adaptation details in §3.3.

3D modeling, evaluation rigor, and clinical interpretability. Diffusion and cold-diffusion models have shown strong results in medical image synthesis (Ho et al., 2020; Bansal et al., 2023). Here, we provide a unified 3D comparison within a single evaluation framework (§3), including atlas-based regional summaries in addition to global image-quality metrics. Our in-house models build on standard MONAI (Cardoso et al., 2022) encoder-decoder and diffusion components; implementation details are described in §3.1.

3. Methods

Model-count convention. We evaluated 11 models: CAE3D, eight in-house baselines, and two separately evaluated foundation-encoder adapters (Med3DVLM and SAM-Med3D). Of the eight in-house baselines, seven are full-volume 3D models directly comparable to CAE3D; the eighth, **CAE_2D**, is evaluated on the middle slice only and is a descriptive, different-domain comparator (§3.2). Primary held-out comparisons (Table 1) involve all nine in-house models (CAE3D plus the eight baselines), while CAE3D has ten total comparator

rows (the eight in-house baselines plus the two foundation adapters). The three-seed stability table (Table 2) and the K-fold table (Table 7, Appendix C) report different, smaller model subsets tailored to those specific analyses, each listed explicitly in its own caption; *nine*, *ten*, and *eleven* refer to the primary held-out comparison unless stated otherwise.

3.1. Proposed method: 3D conditional autoencoder and diffusion variants

CAE3D is a 3D conditional autoencoder that maps a pre-ACZ perfusion map to a predicted post-ACZ perfusion map in a single forward pass. CAE3D treats the task as deterministic conditional regression, prioritizing stable full-volume reconstruction rather than generation of multiple possible outputs. Given the modest cohort size, a deterministic encoder-decoder also favors training stability and spatial fidelity. CAE3D uses MONAI's volumetric encoder-decoder with skip connections between encoder and decoder stages, configured with three spatial dimensions, channel widths of 16, 32, 64, and 128, strides of 2 at each of the three downsampling levels, and 2 residual units per block, preserving fine-grained spatial detail while learning the pre→post perfusion mapping (Algorithm 1, steps 1-2). The output is a continuous perfusion intensity map; the model is termed a conditional autoencoder to distinguish it from segmentation networks. The training objective is

$$L(\theta) = L_{\mathrm{L1}}(\hat{x}, x) + \big(1 - \mathrm{SSIM}(\hat{x}, x)\big), \tag{1}$$

where $\hat{x}$ is the predicted post-ACZ map, x is the ground-truth post-ACZ map, L_{L1} is the mean absolute error, and SSIM is the structural similarity index (3D volumetric SSIM over the brain with data range 1.0; same settings for loss and evaluation). CAE3D was trained under the same optimizer and schedule as the other in-house models (§3.1), without early stopping (full pseudocode: Algorithm 1, Appendix A). CAE3D loss ablation (L1-only / SSIM-only / L1+SSIM) is in the supplementary material; L1+SSIM matches the main row and yields the best MAE/PSNR. We also implemented an exploratory vascular-territory-weighted loss (Liu et al., 2023); all primary results reported here use the unweighted full-brain objective (in-mask voxels via the MNI brain mask; §3.4).

Architecture details. **CAE3D** and several in-house encoder-decoder baselines are built on MONAI's 3D `UNet` (Cardoso et al., 2022), a standard implementation of the encoder-decoder U-Net design (Ronneberger et al., 2015). The contribution of CAE3D lies in the task-specific adaptation: the choice of spatial dimensions, channel widths, strides, residual units, inputs and targets, padding strategy, pre→post mapping, loss design (the combined L1+SSIM objective, Eq. (1)), and training protocol (full 50-epoch schedule selected by validation PSNR).

Protocol. We studied supervised synthesis of post-ACZ ASL perfusion maps from pre-ACZ maps as the shared prediction task. We compared deterministic encoder-decoder models, diffusion-style variants, a Fourier Neural Operator baseline (FNO_3D) (Li et al., 2021), and adapted pretrained 3D encoders under this shared task. We used a fixed subject-level train/validation/test split (random seed 42, determined before any model was trained) for all models, with three training seeds (42, 123, 456) for in-house models and a held-out test set reserved for final evaluation (never used for tuning). **Secondary K-fold.** Separately from the primary held-out, we ran a rotating-test K-fold (K=5) over the full cohort for

nine trained-from-scratch 3D models, including CAE3D and eight comparison models (§4; replication: Appendix C). Foundation adapters (Med3DVLM, SAM-Med3D) used the same combined split with seed-aggregated metrics (§3.3).

Training. All in-house models were trained with the Adam optimizer (learning rate 10^{-3}), batch size 2, for up to 50 epochs on one NVIDIA GPU, with checkpoints selected by the highest validation PSNR. Most deterministic baselines converged by epoch 25–30; diffusion-style models (DDPM_3D, Cold_3D, Residual_3D) ran the same epoch budget but required substantially more compute per epoch because of the full T=1000-step diffusion pass per training sample (early-stopping details for CAE3D vs. CAE3D-ES are in §3.2). Loss terms were model-specific, but all models shared the same pre→post prediction target and metric definitions; full-volume models were evaluated over the full brain volume, whereas CAE_2D was evaluated on the middle axial slice. CAE-family models optimized L1+SSIM (§3.1). Foundation-adapter training used the settings in §3.3. Three randomized seeds (42, 123, 456) were run per in-house model, with test-set mean±std reported across seeds.

Evaluation. The validation set was used only for model selection (checkpointing by validation PSNR) and early stopping where enabled; the held-out test set ($N = 32$) was used only for final reporting. For in-house models, we reported mean±std of test metrics across R=3 independently trained seeds (Table 2), which characterizes seed-to-seed training variability. Table 1 instead reports, per in-house model, the bootstrap 95% CI of the cohort mean from the predesignated seed-42 instance's per-subject metrics; within that seed, the checkpoint with the highest validation PSNR was retained, with no test-set involvement in selection ($N = 32$ paired observations; the same per-subject values are used for the Wilcoxon tests; the selection rule is summarized in Table 5, Appendix A). Because Table 1 and Table 2 are computed from different evaluation runs (a single instance vs. three independently trained seeds), their point estimates for a given model are not directly interchangeable. Bootstrap 95% CIs for cohort-mean MAE, SSIM, and PSNR used B=2000 subject-level resamples (percentile method), quantifying sampling uncertainty of the cohort means rather than per-voxel variability. Bias CIs in Table 1 used the same bootstrap on per-subject Bland-Altman differences (predicted minus target mean intensity); Limits of Agreement (LoA) are full-sample point estimates (±1.96× Standard Deviation (SD)). We reported paired Wilcoxon tests with Holm adjustment as defined and detailed in §3.4 and the supplementary material.

Patient study population. Moyamoya disease is exceptionally rare, with an estimated US incidence of 0.57 per 100,000 person-years (Miller et al., 2025). The rarity of Moyamoya makes large paired cohorts difficult to assemble; nevertheless, the single-center design remains an important limitation (§5). The study enrolled patients with Moyamoya disease who had not undergone prior neurosurgical treatment and were undergoing evaluation for extracranial-to-intracranial bypass surgery at the Stanford School of Medicine, which maintains one of the largest single-center Moyamoya cohorts in the Western Hemisphere. Inclusion required confirmation of Moyamoya disease by catheter cerebral angiography, MR angiography, or CT angiography, and completion of paired pre- and post-acetazolamide ASL perfusion maps. Exclusion criteria for the PET/MRI procedures included kidney function impairment (glomerular filtration rate $< 40\,\mathrm{ml/min/1.73\,m^2}$), pregnancy, history of brain injury, and contraindications to MRI or to acetazolamide. At imaging, subjects had no

acute infarction or hemorrhage. The study was approved by the local institutional review board (IRB), and all participants provided written consent. Data were acquired between 2020 and 2023 under an approved simultaneous PET/MRI ASL protocol; only the ASL perfusion (MRI) channel was used in this analysis, and PET data were not modeled. From the 252 subject-level pairs initially identified, one subject was excluded after failed affine registration (§3.1). The final cohort consists of 251 subjects (188 for training, 31 for validation, and 32 for held-out test), all with Moyamoya disease.

Data and preprocessing. All in-house models used the same subject split, registered and normalized input-target pairs, and metric definitions, preprocessed identically before any model-specific training (exceptions noted above, §3.1–§3.3). Inputs and targets are ASL perfusion maps (pre-ACZ baseline, post-ACZ after ACZ administration), modeled as image-to-image regression in a normalized space. **Registration and normalization.** Volumes were affinely registered to MNI152 2 mm (a standard whole-brain template space that enables cross-subject spatial comparison) with DIPY (center-of-mass, translation, rigid, affine; mutual information) and resampled to $91 \times 109 \times 91$ (Collins et al., 1994; Mazziotta et al., 2001). The pre-ACZ transform was applied to both pre- and post-maps, and one failed case (251/252 success) was excluded. Trilinear interpolation was used for images and nearest-neighbor for masks; an MNI brain mask was applied, and per-volume min–max normalization to $[0, 1]$ was performed independently for each pre- and post-ACZ volume (each volume rescaled to its own $[0, 1]$ range). Because this rescaling is independent per volume, absolute between-scan intensity differences are partly absorbed into it, which bears on how normalized-space ΔCBF relates to physical-unit CVR (§5). No additional subject-level exclusions were applied beyond this single registration failure; the held-out test set of 32 subjects passed all preprocessing steps without further selection.

Model inputs and augmentation. 2D baselines used the middle axial slice (91×109), while 3D models used the full volume $(1, 91, 109, 91)$. After MNI resampling, all volumes share this same $91 \times 109 \times 91$ grid. Some architectures require spatial dimensions divisible by 2^{depth} (here, 8 for three downsampling levels); since 91, 109, and 91 are not divisible by 8, we zero-padded to $96 \times 112 \times 96$ for those architectures and cropped back to $91 \times 109 \times 91$ before computing all metrics. Training used paired random flips along the left-right (LR), anterior-posterior (AP), and superior-inferior (SI) axes along with paired intensity scaling in $[0.9, 1.1]$, with the same flip and scale applied to both volumes in a pair to preserve pre/post alignment. Only the LR flip is anatomically motivated by approximate brain symmetry; the AP and SI flips are not anatomically realistic and were included as generic geometric regularization rather than physiologically motivated augmentations. For the ResNet 3D model, we also evaluated Tied-Augment (Kurtuluş et al., 2023) (scheduled pooled-feature tie across two augmented views). This augmentation system improved validation metrics but did not surpass the best 3D conditional autoencoder on held-out testing.

3.2. Baselines and comparison models

Setup and CAE naming. The in-house models were compared using the shared subject split, preprocessing pipeline, and metric definitions, with the spatial-domain and foundation-adapter exceptions described below. The proposed encoder-decoder is **CAE3D (ours)** in tables. CAE3D and the following in-house baselines and ablations—CAE3D-ES,

ResNet_3D, Cold_3D, Residual_3D, DDPM_3D, FNO_3D, Hybrid_3D, Patch_3D—were implemented and trained under the shared protocol using three random seeds (besides the two foundation model adapters in §3.3); CAE_2D is introduced separately below. **CAE3D** uses the full training schedule **without** early stopping (50 epochs, checkpoint = highest validation PSNR); **CAE3D-ES** is the same encoder-decoder trained **with** early stopping when validation PSNR plateaus for 10 consecutive epochs. Both use the same MONAI implementation and L1+SSIM objective, differing only in the training-stop rule; CAE3D-ES is retained as an ablation to isolate that rule's effect. Its consistently higher MAE relative to CAE3D (Table 2) supports the full 50-epoch schedule for this cohort size; no deterioration in the selected validation metric was observed.

Slicing, ResNet, Diffusion, and FNO. **CAE_2D** uses the same L1+SSIM loss on the middle axial slice only (conservative baseline; primary comparisons are 3D). **ResNet_3D** is a 3D ResNet (MONAI) with the same loss and preprocessing; we also report a Tied-Augment variant (Kurtuluş et al., 2023). **Diffusion models** concatenate pre-ACZ to the noisy or degraded target channel-wise; the denoiser matches the encoder-decoder resolution hierarchy (CAE3D family).
`DDPM_3D`: T=1000, linear β, ε-prediction; full DDPM sampling (no DDIM).
`Cold_3D`: linear interpolation from post-ACZ toward pre-ACZ ($x_t = \alpha_t x_{\text{post}} + (1-\alpha_t)x_{\text{pre}}$) with a cosine schedule; the network reverses this path.
`Residual_3D`: diffuse and predict residual post $-$ pre, then pre $+$ prediction; T=1000, linear β.
`Patch_3D`: a patch-based latent diffusion variant; a VAE first encodes overlapping 24^3 patches (stride 12, 50% overlap) into a compact latent space, a diffusion model is trained to denoise in that latent space conditioned on the pre-ACZ patch, and patch predictions are decoded and re-assembled (overlap-averaged) into the full volume.

Residual_3D_tips applies additional training-stabilization measures (DDIM sampling, patch-based training, a cosine noise schedule) intended to reduce the collapse behavior seen in **Residual_3D**; on this cohort it did not succeed (§3.2) and is not a primary row in Table 1. **FNO_3D** (Li et al., 2021) is a Fourier Neural Operator from pre-ACZ to post-ACZ (12 modes, width 64) with the same L1+SSIM loss and pipeline. Diffusion and FNO models used the same held-out test set and full-brain metrics.

3.3. Foundation model adapters: Med3DVLM and SAM-Med3D

To contextualize trained-from-scratch results against transfer learning, two open-source 3D encoders were adapted to the same pre→post ASL task, split, and full-brain evaluator: **Med3DVLM** (Xin et al., 2025) (frozen DCFormer encoder, 8 seeds) and **SAM-Med3D** (Wang et al., 2023) (frozen SAM 3D image encoder, 3 seeds), each paired with a small trainable regression decoder. "Frozen" here means that the pretrained encoder weights are held fixed; only the lightweight decoder head is trained on the ASL synthesis task. The regression decoder is a stack of `ConvTranspose3d` upsampling blocks (BatchNorm, GELU) that double spatial resolution until the target volume size is reached, followed by a final $1\times1\times1$ convolution to a single output channel; no skip connections are used, since the frozen backbones lack a compatible encoder-decoder feature hierarchy. Both are third-party, pub-

lic architectures, distinct from the in-house MONAI baselines. Full training configurations and aggregated outputs are in Appendix C.

3.4. Metrics

MAE, SSIM, and PSNR use the same metric definitions for every model (Hore and Ziou, 2010), over in-mask voxels (MNI brain mask) on $[0, 1]$ data. For 3D models, metrics are computed over the full volume; **CAE_2D** uses the middle axial slice only (§3). Foundation adapters (†) use the same evaluator but a separately described training/aggregation protocol (§3.3). **MAE** refers to mean absolute error over in-mask voxels; lower MAE means closer voxel-wise intensity reconstruction. **SSIM** measures three-dimensional structural similarity (window size 7, data range 1.0), matching the loss; higher SSIM means better preservation of structural contrast and local spatial patterns. **PSNR** is the standard decibel (dB) measure derived from MSE; higher PSNR indicates lower reconstruction noise. We also report R^2, computed as $1 - \mathrm{SS}_{\mathrm{res}}/\mathrm{SS}_{\mathrm{tot}}$ on per-subject mean in-mask intensities (32 subjects) rather than on voxelwise values. We report all four metrics for eleven models in Table 1.

Regional perfusion and ΔCBF. We additionally evaluated performance across predefined vascular territories. Our regional pipeline covers nine 3D models, evaluated on the same 32-subject held-out set. Atlas masks from Liu et al. (2023) at MNI 2 mm define 11 regions (bilateral ACA, MCA, PCA, cerebellum, pons/medulla, plus a pooled vascular mask). Figure 5 shows the territory color map on a representative slice; Supplementary Figure S1 shows mask unions (left hemisphere, right hem., pooled vascular) on the same slice convention. Per territory, all voxels within the atlas mask are first averaged to obtain territory-level scalars $\bar{I}_{\mathrm{pre}}$, $\bar{I}^{\mathrm{GT}}_{\mathrm{post}}$, and $\bar{I}^{\mathrm{pred}}_{\mathrm{post}}$; percent change is then $\Delta\mathrm{CBF} = (\bar{I}_{\mathrm{post}} - \bar{I}_{\mathrm{pre}})/\bar{I}_{\mathrm{pre}} \times 100\,\%$, and absolute change is $\Delta\mathrm{CBF}_{\mathrm{abs}} = \bar{I}_{\mathrm{post}} - \bar{I}_{\mathrm{pre}}$ (normalized units). Computing on territory means rather than voxel-wise ratios avoids denominator instability caused by individual near-zero-perfusion voxels. Laterality metadata for stratification are unavailable; this analysis is not a formal gas-reactivity CVR (§5).

Agreement and statistical significance. Bland-Altman agreement analysis (Bland and Altman, 1986) was used to assess whether predicted and target perfusion summaries agree closely enough for interpretation, by quantifying systematic bias and the spread of paired differences rather than relying solely on correlation. For each model, we computed the mean difference (predicted minus target) and LoA ($\pm 1.96\times$ SD) from per-subject full-brain mean intensities. Agreement should be interpreted together with LoA and reconstruction errors, rather than solely from bias. Table 6 duplicates bias/LoA. Pairwise comparisons used the paired Wilcoxon signed-rank test on per-subject MAE, SSIM, and PSNR, with Holm adjustment to control for multiple comparisons within each metric; full tables are in the supplementary material. Figure 7 illustrates CAE3D guided backpropagation (input-gradient saliency maps highlighting which pre-ACZ voxels most influence the prediction; positive-gradient variant; full set in the supplementary material); representational similarity analysis (RSM: pairwise cosine similarities between bottleneck features across test subjects) is reported in the supplementary material only.

Table 1: Cohort mean MAE ↓, SSIM ↑, PSNR ↑; R^2 and Bland-Altman bias/LoA. Values shown as mean ± half-width of bootstrap 95% CI (B=2000). **Bold**/underline: best/second-best in-house model (unrounded values; displayed ties are a rounding artifact, §4). † Foundation adapter rows (§3.3) use a different protocol; bold/underline compares only Med3DVLM vs. SAM-Med3D (± denotes std over seeds). See Appendix C for foundation R^2/bias/LoA.

Model	MAE ↓ (± half 95% CI)	SSIM ↑ (± half 95% CI)	PSNR ↑ (± half 95% CI)	R^2	Bias (± half 95% CI)	LoA low – LoA high
CAE3D (ours)	**0.066 ± 0.001**	**0.80 ± 0.01**	**24.0 ± 0.1**	0.35	−0.003 ± 0.019	−0.115 – 0.108
FNO_3D	0.072 ± 0.016	0.78 ± 0.06	24.0 ± 1.3	0.47	−0.010 ± 0.018	−0.116 – 0.097
ResNet_3D	0.072 ± 0.008	0.80 ± 0.04	23.2 ± 0.8	0.32	0.045 ± 0.032	−0.119 – 0.208
Patch_3D	0.078 ± 0.016	0.71 ± 0.06	22.9 ± 1.4	0.28	−0.037 ± 0.019	−0.145 – 0.070
CAE_2D	0.081 ± 0.015	0.68 ± 0.09	20.8 ± 1.2	0.58	0.020 ± 0.023	−0.085 – 0.125
Hybrid_3D	0.120 ± 0.013	0.59 ± 0.04	18.8 ± 0.7	−0.03	0.028 ± 0.024	−0.113 – 0.168
Cold_3D	0.179 ± 0.011	0.37 ± 0.02	14.6 ± 0.3	0.31	−0.035 ± 0.019	−0.143 – 0.074
Residual_3D	0.250 ± 0.028	0.31 ± 0.02	13.3 ± 0.95	−7.44	−0.250 ± 0.028	−0.407 – −0.093
DDPM_3D	0.749 ± 0.027	0.04 ± 0.00	1.12 ± 0.11	−68.64	0.748 ± 0.027	0.592 – 0.905
SAM-Med3D†	0.083 ± 0.001	0.701 ± 0.004	22.20 ± 0.03	0.45	0.015	−0.099 – 0.130
Med3DVLM†	0.098 ± 0.003	0.418 ± 0.035	15.98 ± 1.86	0.46	−0.011	−0.126 – 0.105

4. Results

Primary test-set reconstruction. CAE3D achieved the lowest MAE (0.066) among all trained-from-scratch models on the held-out test set, and tied for the highest reported SSIM (0.80, with ResNet_3D) and PSNR (24.0 dB, with FNO_3D) at the precision shown in Table 1. Published pCASL test-retest studies report within-subject coefficients of variation of roughly 3–8% for regional CBF under a fixed acquisition protocol (Lin et al., 2020; Neumann et al., 2021). Because our MAE is computed voxelwise after independent per-volume $[0, 1]$ normalization, it is not directly comparable in units or normalization to these physical-unit regional repeatability estimates; we report the test-retest figures only as qualitative context for the scale of ASL measurement variability, and we do not claim that MAE 0.066 is below, or otherwise directly comparable to, the acquisition's own repeat-scan noise floor. Table 1 reports cohort-mean MAE, SSIM, and PSNR with bootstrap confidence intervals, together with per-subject R^2 and Bland-Altman bias/limits of agreement. For the adapted pretrained rows (†), MAE/SSIM/PSNR are mean±std over seeds, and R^2 and bias/LoA are point estimates from one checkpoint (§3.3, Appendix C). Five-fold cross-validation (Table 7, Appendix C; §4) confirmed CAE3D remained top-performing across alternative partitions.

Seed-to-seed stability. Table 2 summarizes seed-to-seed variability for the primary trained-from-scratch 3D models on the fixed held-out test set. Unlike Table 1, which provides the main bootstrap-based model comparison across all methods, this table focuses on optimization stability across three independent training seeds. **Residual_3D_tips** (DDIM sampling, patch-based training, a cosine noise schedule) collapsed to a near-zero-residual solution across all three seeds despite genuinely distinct training; we report it for transparency only, not as evidence of stability (table caption; Appendix C). **Patch_3D**'s near-zero variance instead reflects a pretrained VAE shared across the three diffusion-stage seeds, not genuine seed-independence (table caption; Appendix C).

Among the trained-from-scratch models, **CAE3D (ours)** achieved the strongest overall held-out performance. Paired Wilcoxon tests with Holm adjustment (family of eight comparisons per metric, N=32 paired subjects) confirmed that CAE3D's MAE advantage

Table 2: Seed-to-seed stability: cohort-mean MAE, SSIM, PSNR as mean±std over three seed-indexed runs for the primary trained-from-scratch 3D models (main comparison: Table 1); runs were independent except Patch_3D, detailed below. ‡Residual_3D_tips's three seeds are genuinely distinct checkpoints, but all collapsed to the same degenerate near-identity solution (predicted residual ≈ 0); not a valid stability/performance result (§4) and excluded from ranking. §For Patch_3D, only the latent diffusion stage was seeded per run; the pretrained VAE was shared across all three fixed-test runs rather than retrained, so this is not a fully independent three-seed result—its near-zero SD reflects the shared VAE, not seed-independence (K-fold replication, Table 7, retrains the VAE per fold and shows nonzero variance).

Model	MAE (mean ± std) ↓	SSIM (mean ± std) ↑	PSNR (mean ± std) ↑
CAE3D (ours)	$\mathbf{0.0663 \pm 0.0008}$	$\mathbf{0.7986 \pm 0.0011}$	$\mathbf{24.00 \pm 0.08}$
Residual_3D_tips‡	0.0675 ± 0.0000	0.7885 ± 0.0000	23.98 ± 0.00
CAE3D-ES	0.0722 ± 0.0013	0.7933 ± 0.0009	23.31 ± 0.15
Patch_3D§	0.0721 ± 0.0000	0.7235 ± 0.0000	23.32 ± 0.00
FNO_3D	0.0725 ± 0.0001	0.7744 ± 0.0003	23.87 ± 0.02
ResNet_3D	0.0767 ± 0.0008	0.7233 ± 0.0056	22.32 ± 0.08
Hybrid_3D	0.1038 ± 0.0035	0.6255 ± 0.0160	19.54 ± 0.26
Cold_3D	0.2476 ± 0.0173	0.2707 ± 0.0390	12.27 ± 0.33

was statistically significant over ResNet_3D, Cold_3D, DDPM_3D, FNO_3D, Hybrid_3D, Patch_3D, and Residual_3D ($p_{\text{Holm}} < 0.05$), but not over the 2D middle-slice **CAE_2D** baseline ($p_{\text{Holm}} = 0.080$); its SSIM and PSNR advantages were statistically significant over all eight comparators, including CAE_2D. Full pairwise tables are in the supplementary material. Because CAE_2D is a different-domain, middle-slice comparator (§3.4), these significance tests are descriptive and supportive rather than evidence that volumetric modeling is statistically superior to slice-based modeling. Among the weaker baselines, **DDPM_3D** failed catastrophically ($R^2 = -68.64$, consistent with near-constant high-intensity output under the unconditioned diffusion schedule), and **Residual_3D** collapsed toward the pre-ACZ image (bias −0.250), producing systematically underestimated post-ACZ intensities. **Cold_3D** and **Hybrid_3D** trained successfully but lagged the deterministic models. Foundation adapters (†) used a different evaluation protocol and are not directly comparable to in-house rows. For example, **SAM-Med3D** outperformed **Med3DVLM**, but neither matched the best deterministic results (Table 1 caption; §3.3). Regional summaries appear in Tables 3 and 4.

Interpreting subject-level R^2. CAE3D's per-subject R^2 was modest (0.35) despite its leading MAE, while **FNO_3D** reached a higher R^2 (0.47) at a comparable MAE (0.072 vs. 0.066). This is not a contradiction: R^2 is computed on per-subject mean in-mask intensities (§3.4) and reflects how well a model reproduces *between-subject* variance in mean post-ACZ signal, whereas MAE and SSIM measure *within-subject*, voxelwise fidelity. A model can reconstruct each subject's volume accurately while still explaining only a fraction of between-subject variance; plausible contributors include heterogeneous vasodilatory response magnitude, per-volume normalization effects, measurement variability, and post-ACZ information not identifiable from the baseline scan alone. We have not formally de-

Table 3: Cohort-mean ΔCBF by vascular territory (fixed held-out, N=32). Percent $\Delta = (\bar{I}_{\text{post}} - \bar{I}_{\text{pre}})/\bar{I}_{\text{pre}} \times 100$; absolute $\Delta_{\text{abs}} = \bar{I}_{\text{post}} - \bar{I}_{\text{pre}}$ (normalized $[0, 1]$ units; both computed from territory-averaged intensities). CAE3D compresses percent Δ in high-dynamic-range territories (cerebellum, pons) but captures a substantial fraction of the true absolute change. ResNet_3D overestimates absolute Δ in most territories.

Territory	GT Δ%	**CAE3D** Δ%	ResNet Δ%	GT Δ_{abs}	**CAE3D** Δ_{abs}	ResNet Δ_{abs}
left_ACA	19.1	14.3	33.9	0.048	0.025	0.069
left_MCA	32.4	18.4	40.4	0.051	0.033	0.057
left_PCA	43.4	17.8	45.6	0.059	0.042	0.080
left_cerebellum	187.6	18.8	80.8	0.060	0.051	0.087
left_pons_medulla	115.2	19.2	76.6	0.023	0.041	0.040
right_ACA	18.2	16.0	30.1	0.042	0.027	0.057
right_MCA	23.6	17.6	37.2	0.040	0.030	0.057
right_PCA	43.9	17.7	40.1	0.062	0.041	0.064
right_cerebellum	171.6	18.3	82.5	0.060	0.051	0.079
right_pons_medulla	144.7	20.0	79.7	0.027	0.041	0.036
vascular_territory	25.8	16.6	33.9	0.046	0.030	0.059

composed these contributions, nor analyzed the FNO_3D/CAE3D gap as a bias-variance trade-off. We do not read $R^2 = 0.35$ as evidence that CAE3D recovers each patient's individualized CVR; together with the territory-level ΔCBF compression discussed below, it indicates the model captures population-level structure more reliably than subject-specific vasodilatory variation, a distinction we return to in §5.

Agreement and Bland-Altman. Bias and LoA appear per model in Table 1; Table 6 (Appendix C) additionally reports the SD underlying each LoA. Agreement broadly followed reconstruction quality: stable deterministic models showed near-neutral bias, while poorly performing diffusion models showed large bias and wide limits of agreement. CAE3D achieved the smallest absolute bias (−0.003), with LoA comparable to FNO_3D and CAE_2D (±0.11 for all three); bias should nonetheless be interpreted together with LoA and reconstruction error, since a small mean difference does not guarantee narrow subject-level spread.

Territory-level ΔCBF (exploratory). Table 3 reports cohort-mean ΔCBF (%) per territory for ground truth, CAE3D, and ResNet_3D. CAE3D preserved territorial ordering with substantially smaller deviation from ground truth than the unstable diffusion baselines, while ResNet_3D tended to overestimate vascular responses; consistent with deterministic autoencoders generally (Goyal et al., 2026), CAE3D compressed the predicted ΔCBF dynamic range, most pronounced in the cerebellum and pons (∼10-fold smaller than ground truth). In absolute normalized units (Δ_{abs}), however, CAE3D retained a substantial fraction of the true territory-level change. Absolute ΔCBF magnitudes should not be read as calibrated CVR values; conversion to physical units (ml/100 g/min) requires inversion using the per-subject pre-ACZ CBF scale, which we have not performed in this study (§5). Pooled vascular-mask means for nine 3D models (excluding 2D models, because the primary focus is full-volume 3D synthesis) appear in Table 4.

Table 4: Regional ΔCBF (%): cohort mean on the fixed held-out test set, vascular territory only. GT = ground truth; pred = model prediction.

Model	mean ΔCBF GT (%)	mean ΔCBF pred (%)
Cold_3D	25.8	−77.9
DDPM_3D	25.8	350.7
FNO_3D	25.8	14.7
Hybrid_3D	25.8	38.3
Patch_3D	25.8	−1.9
Residual_3D	25.8	−99.8
ResNet_3D	25.8	33.9
CAE3D (ours)	25.8	16.6

Supplementary Figure S2 plots cohort-mean ground-truth territorial ΔCBF for comparison.

Note: Extreme predicted ΔCBF for DDPM_3D and Residual_3D reflects training instability under this setup (cf. ranking above).

Five-fold cross-validation. To confirm that the fixed held-out ranking was not driven by a particular train/test split, we conducted a rotating K=5 cross-validation over the full cohort (train on $K-2$ folds, validate on one, test on one). Nine trained-from-scratch 3D models were evaluated under the same protocol; foundation adapters used a separate protocol. The fold-level cohort means in Table 7 (Appendix C) confirm CAE3D remained the top-performing model across partitions, though the ordering of the remaining models varied somewhat across folds.

Results on held-out test set. Table 4 shows that CAE3D and FNO_3D were closest to the ground-truth mean ΔCBF among the evaluated 3D models. **Cold_3D** underperformed the deterministic baselines on reconstruction fidelity (MAE/SSIM/PSNR) without collapsing as severely as **DDPM_3D** or **Residual_3D**, though its pooled-vascular ΔCBF (−77.9% vs. GT +25.8%, Table 4) shows comparably poor territory-level agreement. Supplementary Figures S3–S4 show pre/post/change and a five-panel qualitative layout; the supplementary material also reports pipeline-unit MAE, a CAE3D-vs.-Cold_3D permutation test ($p < 0.001$), regional Wilcoxon/FDR, RSM, and failure panels. High PSNR or favorable qualitative appearance (Figure 6) does not by itself establish clinical appropriateness, because conventional reconstruction metrics may not capture errors that alter territory-level hemodynamic interpretation or treatment-related judgments.

5. Discussion

Main findings and clinical interpretation. CAE3D synthesized post-ACZ perfusion maps from pre-ACZ input with reconstruction error that motivates further clinical validation, though not directly comparable to ASL scan-rescan repeatability (§4). Deterministic conditional autoencoders outperformed diffusion-style variants and frozen-encoder adapters for this task. Qualitative inspection suggested prominent errors near mask boundaries and high-flow or low-signal regions, plausibly reflecting partial-volume, registration, and normalization effects not confirmed quantitatively via voxel-level error localization. Territory ΔCBF remains a surrogate measure, not a validated CVR endpoint, complementing global metrics pending prospective work. CAE3D's modest per-subject R^2 (0.35, §4), together

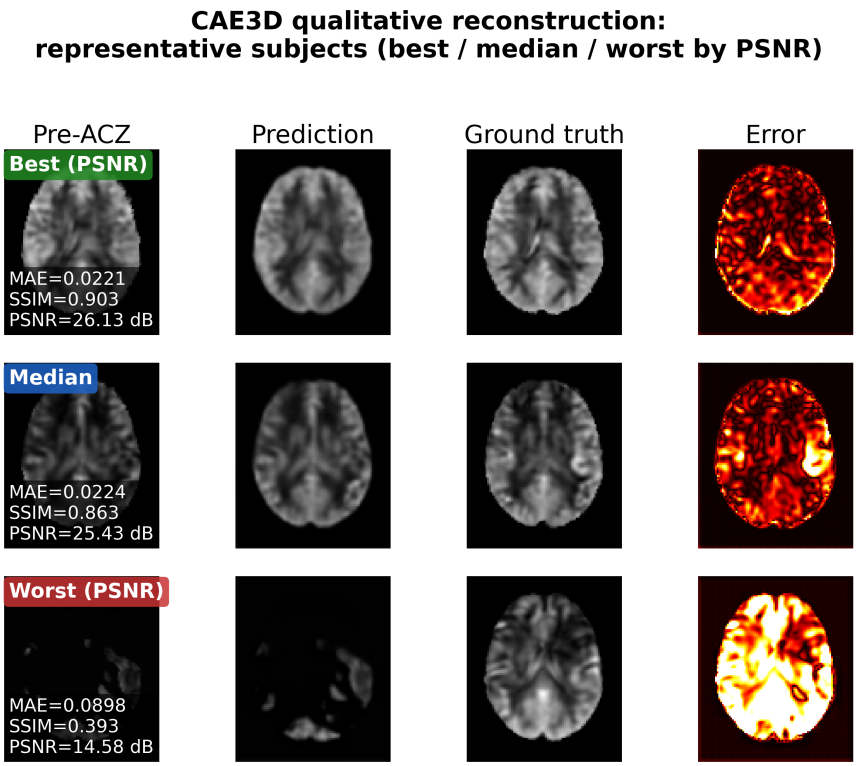


Figure 6: **Qualitative reconstruction** (CAE3D, held-out test set): best, median, worst PSNR subjects (top to bottom). Columns: pre-ACZ, prediction, GT, error (middle axial slice). Per-subject MAE/SSIM/PSNR in row labels.

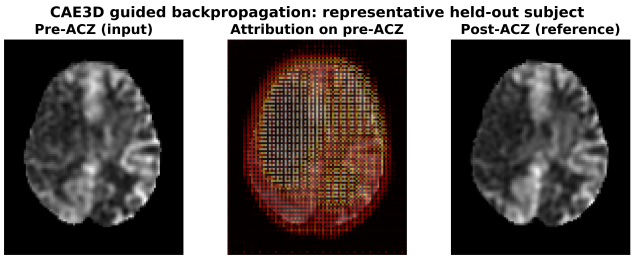


Figure 7: **Guided backpropagation** (CAE3D, one test subject): pre-ACZ, attribution overlay (brighter = stronger influence), GT post-ACZ. Full guided-backpropagation maps are provided in the supplementary material.

with the territory-level ΔCBF compression, indicates that CAE3D captures population-level structure in pre→post mapping more reliably than subject-specific vasodilatory magnitude.

Trustworthiness: uncertainty, agreement, and failure modes. This study reported seed variability, bootstrap CIs, and Bland-Altman bias/LoA, since correlation alone does not imply acceptable bias. Anticipated failure modes include registration/atlas mismatch, unstable ΔCBF at low baseline signal, mask-boundary artifacts, and scanner or cohort shift; future evaluation should pair outputs with predefined quality checks and human review.

Transparency, explainability, and auditing. Because CAE3D may influence downstream CVR interpretation, we provided two explainability analyses: guided backpropagation for one held-out subject (Figure 7), and representational similarity analysis (RSM) of CAE3D bottleneck features across the 32 test subjects (supplementary material). Alignment with clinical subgroups was not tested because laterality and severity metadata were unavailable; curated good/typical/failure cases are more clinically informative than best-case examples. **On ethics and responsible use**, synthesized maps are research outputs for retrospective investigation, not decision-support tools or substitutes for clinical imaging, until prospective validation establishes their intended role. Data sharing must follow PHI/de-identification and drift-monitoring governance policies.

Limitations. This study has several limitations. This is internal validation, not deployment-ready evidence: results may not generalize across scanners, ASL sequences, or sites, since we did not stress-test out-of-distribution or external cohorts; affine registration and atlas readouts can also degrade when anatomy departs from template space. $[0, 1]$ normalization aids training but yields unitless MAE; inversion to physical CBF scales for clinical reporting has not been performed in this study (§5). Region-level FDR was limited to one model pair; we did not run voxel/cluster permutation tests, confounder-adjusted analyses, calibrated uncertainty maps, or prospective workflow validation. The territory-level ΔCBF compression noted in §4 may partly reflect the L1+SSIM training objective, which prioritizes average reconstruction fidelity over preservation of extreme vasodilatory responses; territories with inherently high GT ΔCBF (cerebellum, pons/medulla) show the largest percent-scale compression. Finally, because the study cohort enrolled only patients who completed the two-scan protocol, no patient for whom ACZ was contraindicated or withheld under the study protocol is represented in the training or test data. Therefore, the prevalence of ACZ contraindications in the broader Moyamoya population and model performance in that population remain unknown. An independent surrogate for vasodilatory response (e.g., BOLD CVR from a hypercapnic/breath-hold challenge, or a deep-learning-based drug-free CVR estimate (Chen et al., 2020)) would extend evaluation to contraindicated patients without administering ACZ (planned future work). ASL quality is limited by low signal-to-noise ratio and sensitivity to motion, scanner calibration, and acquisition parameters; a systematic study of these effects on synthesis fidelity is planned future work. No consensus CVR threshold for Moyamoya bypass surgery exists (See and Stout, 2023); decisions integrate imaging, anatomy, symptoms, and other factors rather than a single threshold (Nguyen et al., 2022). Consequently, this study establishes synthesis feasibility, not surgical decision support or replacement of the ACZ challenge.

Future directions. Immediate next steps include multicenter external validation and prospective, blinded clinician-in-the-loop evaluation. Broader directions include physical-unit CBF error reporting after inversion, range-aware training for the compression noted above, and full foundation-model fine-tuning. Territory-aware models may improve regional calibration at the cost of less training data per territory and added inference complexity. Flow matching (Lipman et al., 2023) is a relevant alternative given DDPM-family instability. Baseline-to-challenge synthesis warrants separate evaluation in intracranial atherosclerosis and sickle cell cerebrovascular disease.

References

Arpit Bansal, Eitan Borgnia, Hong-Min Chu, Jie S. Li, Hamid Kazemi, Furong Huang, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Cold diffusion: Inverting arbitrary image transforms without noise. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023. URL https://papers.nips.cc/paper_files/paper/2023/hash/80fe51a7d8d0c73ff7439c2a2554ed53-Abstract-Conference.html. arXiv:2208.09392.

J. M. Bland and D. G. Altman. Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet*, 327(8476):307–310, 1986. doi: 10.1016/S0140-6736(86)90837-8.

M. Jorge Cardoso, Wenqi Li, Richard Brown, Nic Ma, Eric Kerfoot, Yiheng Wang, Benjamin Murrey, Andriy Myronenko, Can Zhao, Dong Yang, Vishwesh Nath, Yufan He, Ziyue Xu, Ali Hatamizadeh, et al. MONAI: An open-source framework for deep learning in healthcare. *arXiv preprint arXiv:2211.02701*, 2022.

David Y. T. Chen, Yosuke Ishii, Audrey P. Fan, Jia Guo, Moss Y. Zhao, Gary K. Steinberg, and Greg Zaharchuk. Predicting PET cerebrovascular reserve with deep learning by using baseline MRI: A pilot investigation of a drug-free brain stress test. *Radiology*, 296(3):627–637, 2020. doi: 10.1148/radiol.2020192793.

D. L. Collins, P. Neelin, T. M. Peters, and A. C. Evans. Automatic 3D intersubject registration of MR volumetric data in standardized Talairach space. *Journal of Computer Assisted Tomography*, 18(2):192–205, 1994. URL https://pubmed.ncbi.nlm.nih.gov/8126267/.

Sanuwani Dayarathna, Kh Tohidul Islam, Sergio Uribe, Guang Yang, Munawar Hayat, and Zhaolin Chen. Deep learning based synthesis of MRI, CT and PET: Review and analysis. *Medical Image Analysis*, 92:103046, 2024. doi: 10.1016/j.media.2023.103046.

Rydham Goyal, Camila Gonzalez, Sasha Alexander, Aja Zou, Michael Moseley, Gary K. Steinberg, and Greg Zaharchuk. Generating post-acetazolamide cerebral blood flow MRI for high-risk stroke patients. OpenReview, 2026. URL https://openreview.net/forum?id=WMBUxtRdxB.

Jia Guo, Enhao Gong, Audrey P. Fan, Maged Goubran, Mohammad M. Khalighi, and Greg Zaharchuk. Predicting 15O-water PET cerebral blood flow maps from multi-contrast MRI using a deep convolutional neural network with evaluation of training cohort bias. *Journal of Cerebral Blood Flow & Metabolism*, 40(11):2240–2253, 2020. doi: 10.1177/0271678X19888123.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 6840–6851, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html. arXiv:2006.11239.

A. Hore and D. Ziou. Image quality metrics: PSNR vs. SSIM. In *Proceedings of the 20th International Conference on Pattern Recognition (ICPR)*, pages 2366–2369, 2010. doi: 10.1109/ICPR.2010.579.

Ramy Hussein, David Shin, Moss Zhao, Jia Guo, Guido Davidzon, Michael Moseley, and Greg Zaharchuk. Brain MRI-to-PET synthesis using 3D convolutional attention networks. arXiv preprint arXiv:2211.12082, 2022. URL https://arxiv.org/abs/2211.12082.

Ramy Hussein et al. Turning brain MRI into diagnostic PET: 15O-water PET CBF synthesis from multi-contrast MRI via attention-based encoder–decoder networks. *Medical Image Analysis*, 93:103072, 2024. doi: 10.1016/j.media.2023.103072.

Amirhossein Kazerouni, Ehsan Khodapanah Aghdam, Moein Heidari, Reza Azad, Mohsen Fayyaz, Ilker Hacihaliloglu, and Dorit Merhof. Diffusion models in medical imaging: A comprehensive survey. *Medical Image Analysis*, 88:102846, 2023. doi: 10.1016/j.media.2023.102846.

Emirhan Kurtuluş, Zichao Li, Yann Dauphin, and Ekin Dogus Cubuk. Tied-augment: Controlling representation similarity improves data augmentation. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pages 17994–18007. PMLR, 2023. URL https://proceedings.mlr.press/v202/kurtulus23a.html.

Zongyi Li, Nikola Kovachki, Kamyar Azizzadenesheli, Burigede Liu, Kaushik Bhattacharya, Andrew Stuart, and Anima Anandkumar. Fourier neural operator for parametric partial differential equations. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://openreview.net/forum?id=c8P9NQVtmn0.

Tianye Lin, Jianxun Qu, Zhentao Zuo, Xiaoyuan Fan, Hui You, and Feng Feng. Test-retest reliability and reproducibility of long-label pseudo-continuous arterial spin labeling. *Magnetic Resonance Imaging*, 73:111–117, 2020. doi: 10.1016/j.mri.2020.07.010.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023.

Chin-Fu Liu, Johnny Hsu, Xin Xu, Ganghyun Kim, Shannon M. Sheppard, Erin L. Meier, Michael I. Miller, Argye E. Hillis, and Andreia V. Faria. Digital 3D brain MRI arterial territories atlas. *Scientific Data*, 10:74, 2023. doi: 10.1038/s41597-022-01923-0. URL https://www.nature.com/articles/s41597-022-01923-0. Article 74; digital 3D atlas of ACA, MCA, PCA, and vertebro-basilar territories from 1,298 stroke patients; MNI space.

J. Mazziotta, A. Toga, A. Evans, P. Fox, and J. Lancaster. A probabilistic atlas and reference system for the human brain: International consortium for brain mapping (ICBM). *Philosophical Transactions of the Royal Society of London Series B*, 356(1412):1293–1322, 2001. doi: 10.1098/rstb.2001.0915.

Ronald Miller, Nnabuchi Anikpezie, Haydn Hoffman, Abdulaziz T. Bako, Smit D. Patel, Anurag Sahoo, Ehimen Aneni, Claribel D. Wee, Karen C. Albright, Julius Gene Silva Latorre, James K. Liu, Pankaj K. Agarwalla, Amit Singla, Priyank Khandelwal, and

Fadar O. Otite. Demographic disparities in the incidence of moyamoya angiopathy in the united states. *Neurology*, 105(5):e214013, 2025. doi: 10.1212/WNL.0000000000214013.

Katja Neumann, Martin Schidlowski, Matthias Günther, Tony Stöcker, and Emrah Düzel. Reliability and reproducibility of Hadamard encoded pseudo-continuous arterial spin labeling in healthy elderly. *Frontiers in Neuroscience*, 15:711898, 2021. doi: 10.3389/fnins.2021.711898.

Vincent N. Nguyen, Kara A. Parikh, Mustafa Motiwala, L. Erin Miller, Michael Barats, Christina Milton, and Nickalus R. Khan. Surgical techniques and indications for treatment of adult moyamoya disease. *Frontiers in Surgery*, 9:966430, 2022. doi: 10.3389/fsurg.2022.966430.

Muzaffer Özbey, Onat Dalmaz, Salman U. H. Dar, Hasan A. Bedel, Şaban Öztürk, Alper Güngör, and Tolga Çukur. Unsupervised medical image translation with adversarial diffusion models. *IEEE Transactions on Medical Imaging*, 42(12):3524–3539, 2023. doi: 10.1109/TMI.2023.3290149.

Vaishnavi L. Rao, Laura M. Prolo, Jonathan D. Santoro, Michael Zhang, Jennifer L. Quon, Michael Jin, Aditya Iyer, Vivek Yedavalli, Robert M. Lober, Gary K. Steinberg, Kristen W. Yeom, and Gerald A. Grant. Acetazolamide-challenged arterial spin labeling detects augmented cerebrovascular reserve after surgery for moyamoya. *Stroke*, 53(4):1354–1362, 2022. doi: 10.1161/STROKEAHA.121.036616. URL https://www.ahajournals.org/doi/10.1161/STROKEAHA.121.036616.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 234–241. Springer, 2015.

R. Michael Scott and Edward R. Smith. Moyamoya disease and moyamoya syndrome. *New England Journal of Medicine*, 360(12):1226–1237, 2009. doi: 10.1056/NEJMra0804622.

Alfred P. See and Jeffrey N. Stout. Cerebrovascular reserve in moyamoya requires more standardization: Editorial on ASL-MRI guided evaluation of multiple burr hole revascularization surgery in moyamoya disease. *Acta Neurochirurgica*, 165:2071–2072, 2023. doi: 10.1007/s00701-023-05646-y.

Teva Pharmaceuticals. Diamox (acetazolamide) prescribing information, 2019. Available from US FDA DrugsFDA database; contraindications include renal insufficiency and sulfonamide hypersensitivity, with electrolyte imbalance and pregnancy warranting risk-benefit assessment per institutional protocol.

Haoyu Wang, Sizheng Guo, Jin Ye, Zhongying Deng, Junlong Cheng, Tianbin Li, Jianpin Chen, Yanzhou Su, Ziyan Huang, Yiqing Shen, Bin Fu, Shaoting Zhang, Junjun He, and Yu Qiao. SAM-Med3D: Towards general-purpose segmentation models for volumetric medical images. *arXiv preprint arXiv:2310.15161*, 2023.

Yu Xin, Gorkem Can Ates, Kuang Gong, and Wei Shao. Med3DVLM: An efficient vision–language model for 3D medical image analysis. *arXiv preprint arXiv:2503.20047*, 2025.

Moss Y. Zhao, Audrey P. Fan, David Yen-Ting Chen, Yosuke Ishii, Mohammad Mehdi Khalighi, Michael Moseley, Gary K. Steinberg, and Greg Zaharchuk. Using arterial spin labeling to measure cerebrovascular reactivity in Moyamoya disease: Insights from simultaneous PET/MRI. *Journal of Cerebral Blood Flow & Metabolism*, 42(8):1493–1506, 2022. doi: 10.1177/0271678X221083471. PMC9274857.

Appendix A. Reproducibility

Preprocessing, training, and evaluation were implemented in Python using PyTorch and MONAI (Cardoso et al., 2022). In-house encoder-decoder and diffusion denoisers use MONAI's `UNet` / `DiffusionModelUNet` with task-specific configuration and recipes as in §3.2 (architecture details: §3.1). The train/validation/test split is fixed (random seed 42) for all models. As described in §3, Table 1 and Table 6 report the single-instance evaluation for each in-house model, distinct from the three-seed evaluation in Table 2. For every in-house model, seed 42 was predesignated (not selected post hoc from among the three training seeds) as the primary instance; within that seed, the checkpoint with the highest validation PSNR was retained, and the test set was not used in checkpoint selection. Table 5 lists this rule per model. For **CAE3D**, R^2, Bland-Altman bias, and LoA in Tables 1 and 6 were verified by direct re-evaluation of the seed-42 checkpoint. The code repository at https://github.com/TheClassicTechno/cae3d-moyamoya-cbf-synthesis documents the exact checkpoint file and training configuration for **CAE3D** (no early stopping), **CAE3D-ES** (early stopping), and every other baseline.

Table 5: Primary-checkpoint selection rule for each in-house model's rows in Tables 1 and 6. Seed 42 was predesignated for every model before training (not chosen post hoc from among the three seeds); within that seed, the checkpoint with the highest validation PSNR was retained, with no test-set involvement in selection. Exact checkpoint filenames and configurations are documented with the code release.

Model	Primary seed	Checkpoint-selection rule
CAE3D	42 (predesignated)	Highest validation PSNR, no early stopping
CAE3D-ES	42 (predesignated)	Highest validation PSNR, early stopping (10-epoch plateau)
ResNet_3D	42 (predesignated)	Highest validation PSNR
CAE_2D	42 (predesignated)	Highest validation PSNR
FNO_3D	42 (predesignated)	Highest validation PSNR
Patch_3D	42 (predesignated)	Highest validation PSNR
Hybrid_3D	42 (predesignated)	Highest validation PSNR
Cold_3D	42 (predesignated)	Highest validation PSNR
Residual_3D	42 (predesignated)	Highest validation PSNR
DDPM_3D	42 (predesignated)	Highest validation PSNR

Algorithm 1: CAE3D training and inference procedure
Input: paired volumes $(x_{\text{pre}}, x_{\text{post}})$
Output: predicted post-ACZ volume $\hat{x}_{\text{post}}$
Preprocess all pairs: affine registration to MNI, brain masking, normalization to $[0, 1]$, and pad/crop to a valid grid **for** *epoch* $= 1$ *to* 50 **do**
  $\hat{x}_{\text{post}} \leftarrow \text{CAE3D}(x_{\text{pre}})$ $\mathcal{L} \leftarrow \|\hat{x}_{\text{post}} - x_{\text{post}}\|_1 + \big(1 - \text{SSIM}(\hat{x}_{\text{post}}, x_{\text{post}})\big)$ update parameters with Adam (10^{-3}) update the retained checkpoint if validation PSNR improves
end
Inference: load best checkpoint and compute $\hat{x}_{\text{post}} = \text{CAE3D}(x_{\text{pre}})$ Evaluate full-brain MAE/SSIM/PSNR, agreement (Bland-Altman), and regional ΔCBF summaries

Supplementary material includes:

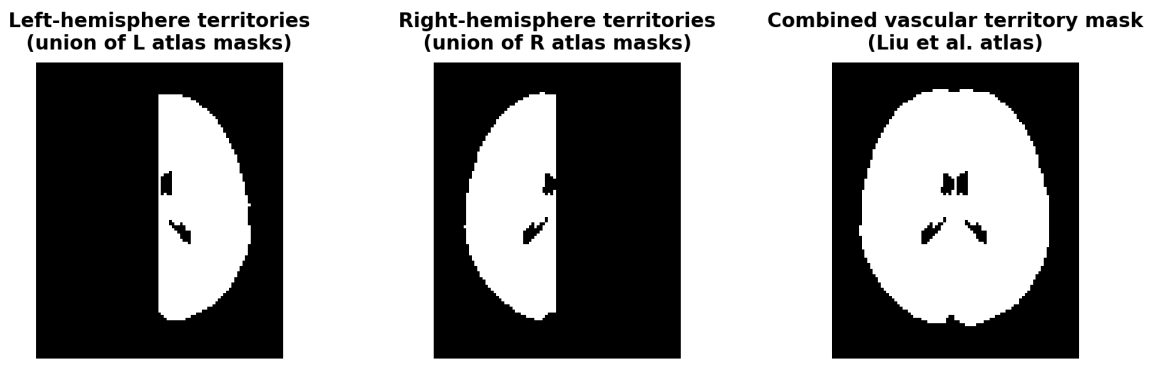
- Holm-adjusted pairwise Wilcoxon tables (MAE/SSIM/PSNR)
- TIPS residual-diffusion note
- Full guided-backprop maps (one panel shown in Figure 7)
- RSM heatmap and table; loss ablation; pipeline-unit MAE
- Permutation and regional FDR analyses; failure-case panels
- **Per-territory MAE/SSIM/PSNR table for CAE3D** (mean±std per atlas territory; Table 9, Appendix C)
- Inputs underlying the five-fold summary (Table 7)
- Foundation-baseline aggregates (Appendix C); per-territory ΔCBF rows for Residual_3D_tips

Training defaults: Adam 10^{-3}, batch size 2, 50 epochs, checkpoint by validation PSNR. Bootstrap resamples B=2000.

Preprocessing: §3.1. Regeneration scripts and file-level provenance are documented in the code repository above.

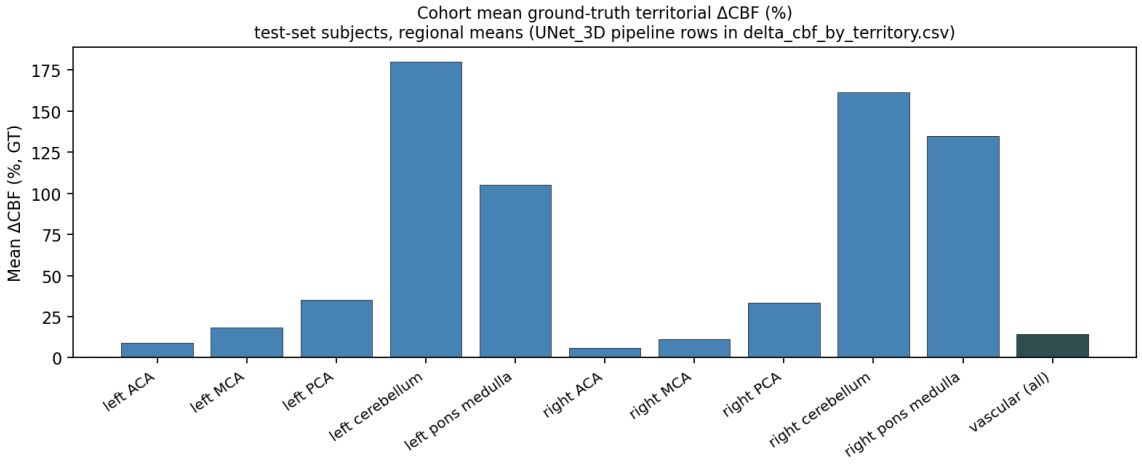
Appendix B. Supplementary figures

Regional and extra qualitative figures match the slide-visual assets used for the main-text figures. Figure S1 below shows binary mask unions on one middle axial slice, using the same slice convention as Figure 5 in the main text.



S1: Atlas mask unions (left hem., right hem., pooled vascular; middle axial slice).

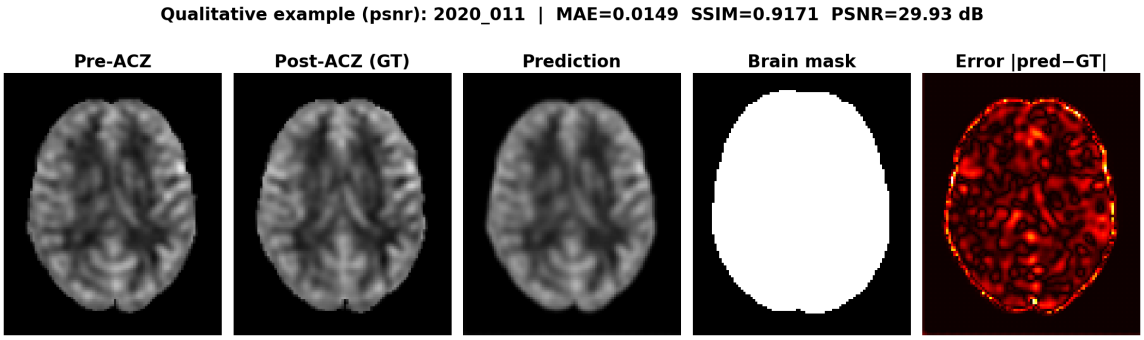
Figure S2 next plots cohort-mean ground-truth territorial ΔCBF (%) for comparison with Table 3 in the main text.



S2: Cohort mean ground-truth territorial ΔCBF (%) (one bar per summary region).

Figures S3 and S4 turn from these territory-level summaries to per-subject qualitative examples. The paired pre-ACZ, post-ACZ, and voxelwise post-minus-pre change maps for this example subject are shown in Figure 4 in the main text; that figure is not reproduced here.

Figure S4 extends this to a five-panel qualitative layout (pre, post, prediction, mask, error) for one subject, complementing the best/median/worst PSNR panels in Figure 6 of the main text.



S4: Five-panel qualitative summary (pre, post, prediction, mask, error).

Appendix C. K-fold sources and foundation baselines

Notes for Table 1 (main text). Rows marked † are frozen-encoder foundation adapters. For these rows, MAE/SSIM/PSNR are reported as mean±std over training seeds (eight for Med3DVLM; three for SAM-Med3D), and bold/underline on those three columns compares only the two adapters. Their R^2 and Bland–Altman bias/LoA are point estimates computed from the seed-42 checkpoint only (sources in Table 8). Bootstrap bias CIs were not computed for these rows. For PSNR, Table 1 bolds only CAE3D: FNO_3D matches the cohort mean to one decimal place (24.0 dB) but is lower in the three-seed summary reported in Table 2.

Bland-Altman detail (Std of paired differences). Table 6 reproduces the per-model Bias and LoA already given in Table 1 (main text) alongside the SD of paired differences underlying each LoA (LoA = Bias $\pm$ 1.96 $\times$ SD).

Table 6: Bland-Altman agreement (predicted − ground truth, normalized $[0, 1]$ scale): mean difference (Bias $\to$ 0), std of differences (Std $\downarrow$), and 95% limits of agreement (LoA = $\pm$1.96$\times$ SD). **Bold**/underline: smallest/second-smallest |Bias|.

Model	Bias $\to$ 0	Std $\downarrow$	LoA_{low}	LoA_{high}
Cold_3D	−0.035	0.055	−0.143	0.074
DDPM_3D	0.748	0.080	0.592	0.905
FNO_3D	−0.010	**0.054**	−0.116	0.097
Hybrid_3D	0.028	0.072	−0.113	0.168
Patch_3D	−0.037	0.055	−0.145	0.070
Residual_3D	−0.250	0.080	−0.407	−0.093
ResNet_3D	0.045	0.083	−0.119	0.208
CAE_2D	0.020	0.054	−0.085	0.125
CAE3D (ours)	**−0.003**	0.057	−0.115	0.108

K-fold replication. Table 7 summarizes fold-level cohort means: each model's five test folds yield one cohort mean per metric, and the table reports mean±std across those five values. Fold-level outputs are stored with the released training scripts. The same protocol can be reproduced from the code release, with optional checks of per-fold tables before exporting LaTeX rows.

Table 7: Five-fold rotation over the full cohort: entries are mean±std of the five fold-level cohort means. **Bold**/underline: best/second-best. ‡Residual_3D_tips collapsed to a near-identity (predicted residual $\approx$ 0) solution on this cohort (§4); its row reflects that collapse, evaluated across five different test folds, and is excluded from best/second-best ranking. The collapse traces to a numerical-stability issue in the shared DDIM/cosine-schedule sampling code that is checkpoint-independent (§5); we directly verified this mechanism against a trivial zero-residual baseline for the fixed-test three-seed instance (Table 2) but did not independently re-verify each of the five K-fold instances against that baseline.

Model	MAE $\downarrow$	SSIM $\uparrow$	PSNR $\uparrow$ (dB)
CAE3D (ours)	0.0746 ± 0.0049	0.7943 ± 0.0115	22.78 ± 0.42
CAE3D-ES	0.0774 ± 0.0056	0.7899 ± 0.0124	22.42 ± 0.43
Residual_3D_tips‡	0.0804 ± 0.0046	0.7774 ± 0.0097	22.74 ± 1.23
ResNet_3D	0.0855 ± 0.0030	0.7237 ± 0.0141	21.20 ± 0.22
FNO_3D	0.0856 ± 0.0047	0.7601 ± 0.0100	22.27 ± 0.32
Patch_3D	0.0856 ± 0.0042	0.7161 ± 0.0122	21.69 ± 0.33
Hybrid_3D	0.1070 ± 0.0041	0.6295 ± 0.0058	19.12 ± 0.24
Cold_3D	0.2506 ± 0.0431	0.2739 ± 0.0409	12.15 ± 1.42
DDPM_3D	0.6150 ± 0.0527	0.0116 ± 0.0032	2.54 ± 0.42

Foundation adapter aggregates. Table 8 reports mean±std MAE/SSIM/PSNR for frozen-encoder adapters and is intended to be read alongside the foundation rows in Table 1.

Table 8: Foundation-style baselines on the combined cohort split. Med3DVLM and SAM-Med3D (frozen image encoder + trained 3D decoder) use pre→post ASL perfusion maps with the same full-brain evaluator as **CAE3D**; mean±std is over training seeds (eight for Med3DVLM; three for SAM-Med3D). Full filenames and seed lists are documented with the code release.

Model	What is scored	MAE	SSIM	PSNR (dB)
Med3DVLM (DCFormer+dec.)	Pred. vs. GT post-ACZ ASL perfusion maps (full-brain)	0.098 ± 0.003	0.418 ± 0.035	15.98 ± 1.86
SAM-Med3D (enc. froz. + dec.)	Pred. vs. GT post-ACZ ASL perfusion maps (full-brain)	0.083 ± 0.001	0.701 ± 0.004	22.20 ± 0.03

Per-territory MAE, SSIM, PSNR for CAE3D. Table 9 reports cohort mean±std per atlas territory for **CAE3D** on the held-out test set (N=32). Territory names follow the atlas of Liu et al. (2023) registered to MNI 2 mm.

Table 9: CAE3D per-territory reconstruction metrics (held-out test set, N=32): cohort mean±std over subjects within each atlas territory. Metrics computed full-brain within each atlas mask on $[0, 1]$ normalized data.

Territory	MAE (mean±std) ↓	SSIM (mean±std) ↑	PSNR (mean±std) ↑
left ACA	0.0827 ± 0.0489	0.9879 ± 0.0120	20.79 ± 4.14
left MCA	0.0891 ± 0.0571	0.9763 ± 0.0248	20.22 ± 4.17
left PCA	0.0939 ± 0.0856	0.9908 ± 0.0142	20.78 ± 4.74
left cerebellum	0.1003 ± 0.1062	0.9943 ± 0.0085	21.02 ± 5.27
left pons/medulla	0.0866 ± 0.0792	0.9982 ± 0.0031	21.82 ± 5.00
right ACA	0.0833 ± 0.0570	0.9879 ± 0.0124	20.87 ± 4.23
right MCA	0.0893 ± 0.0588	0.9755 ± 0.0233	20.24 ± 4.04
right PCA	0.0973 ± 0.0881	0.9911 ± 0.0133	20.52 ± 4.73
right cerebellum	0.1022 ± 0.1058	0.9944 ± 0.0082	20.83 ± 5.41
right pons/medulla	0.0930 ± 0.0875	0.9982 ± 0.0029	21.31 ± 5.03
vascular territory	0.0897 ± 0.0606	0.8910 ± 0.0879	19.96 ± 3.81

Three-seed held-out and K-fold metrics for **Residual_3D_tips** are reported in Tables 2 and 7, respectively, with the collapse caveat noted in both captions; it is not included in the territory-level ΔCBF comparison (Table 4), which uses **Residual_3D** instead.

Residual_3D_tips collapse: checkpoint-level verification. We confirmed via direct re-evaluation of each of the three saved seed-42/123/456 checkpoints that they are genuinely distinct trained weights (distinct checkpoint files, distinct training-loss trajectories), ruling out a simple checkpoint-reuse artifact. Despite this, all three converged to a degenerate solution in which the predicted residual is driven to approximately zero, so the reconstructed output reduces to the clipped pre-ACZ input regardless of checkpoint; independently computing MAE/SSIM/PSNR for a trivial "predict zero residual" baseline on this cohort reproduces the reported Table 2 numbers to within floating-point precision. This is the same collapse failure mode documented for **Residual_3D** (§4); the TIPS stabilization

measures did not prevent it here. The proximate mechanism is a numerical-stability issue in the cosine noise-schedule construction shared by every TIPS checkpoint: the schedule's cumulative-product term can go slightly negative near the tail (a floating-point artifact), producing NaN values that are floored to zero at each DDIM sampling step, driving the predicted residual toward zero regardless of the underlying trained weights. Because this is a property of the shared sampling code rather than of any individual trained instance, it is expected to affect the K-fold instances (Table 7) by the same mechanism; we verified this directly against a trivial zero-residual baseline only for the fixed-test three-seed instance above, not independently for each K-fold checkpoint.

Patch_3D near-zero fixed-test variance: implementation detail. Patch_3D's architecture couples a VAE (encoding overlapping patches into a latent space) with a latent diffusion stage (§3.2). For the fixed-test three-seed experiment, the same pretrained VAE was reused across all three diffusion-stage seeds; only the latent diffusion component varied by seed. Because the shared VAE reconstruction dominates the decoded full-volume output at this patch/stride configuration, genuine per-seed differences confined to the diffusion stage are not resolved at the precision reported in Table 2. This is specific to the fixed-test run rather than the architecture: the K-fold experiment (Table 7) retrained a fold-specific VAE for each fold and shows the expected nonzero seed-to-seed variance for Patch_3D.

Acknowledgments

We thank Prof. Kilian Pohl and Prof. Ehsan Adeli of Stanford University for helpful guidance and discussion during this project. This work is supported by American Heart Association Career Development Award #24CDA1266771.

LLM Use Disclosure

Portions of this manuscript were drafted and revised with the assistance of an AI language model, Claude (Anthropic), consistent with the MLHC LLM Use Policy. AI assistance included drafting and editing prose, and assistance with code used for model implementation, training/evaluation scripts, and figure generation. All experimental results, figures, and scientific claims were reviewed and validated by the authors, who take full responsibility for the accuracy of all content in this manuscript.