EVERY SAMPLE SPEAKS VOLUMES: STEERING 3D MRI GENERATORS UNDER A MEASURED CONTRACT

Anonymous authors
Paper under double-blind review

ABSTRACT

Controlled experiments in neuroimaging often require volumes that real cohorts cannot provide, such as the same brain with and without a specified degree of Alzheimer's disease (AD) atrophy. Pretrained generators of 3D brain MRI could provide such volumes without retraining. However, inference-time guidance fixes only the direction and strength of a change. It neither specifies the magnitude of the change nor bounds its effect on other structures, and it does not verify the output. We propose a closed-loop controller that obtains from a frozen generator a volume satisfying a verifiable specification, which we call a contract. A contract requests a change in the volume of one anatomical structure and bounds the change of every other planned structure. Both are read by an external segmentation tool, the verifier, in units of its test-retest variability. The controller generates pairs of volumes from the same noise with perturbations of opposite sign. After each pair, it updates a local linear model of the verifier's response with a rank-one correction. It returns a volume only when the contract holds within a fixed budget of pairs. On five released generators from four model families, a single configuration satisfies the contract in 84 to 100% of held-out requests. In contrast, gradient-guidance methods with step sizes fixed in advance satisfy it in at most 28%. An AD classifier built on a second, independent segmentation tool scores volumes steered to the full disease profile above their unsteered counterparts. The procedure also transfers to a frozen physics emulator.

1 INTRODUCTION

Suppose you hold a frozen 3D generator of brain MRI that you did not train, and you want a volume in which one defined anatomical structure has changed by a stated amount while the other measured structures stay within a stated bound. Volumes of this kind enable comparisons that real cohorts rarely offer. A classifier can be tested on a specified anatomical change, and a measurement pipeline can be validated against a known ground truth. Unconstrained sampling does not provide them, and retraining a generator for each question is rarely feasible. Brain MRI is costly to acquire and typically shared under access agreements, and fine-tuning requires data and modifies the released weights. This paper therefore addresses the request at inference time.

Alzheimer's disease (AD) is a representative case. On structural MRI it presents atrophy of the medial temporal lobe, with a smaller hippocampus and amygdala, and enlarged ventricles (Frisoni et al., 2010; Nestor et al., 2008). The atrophy accumulates gradually, and its early stage is the most informative. Real cohorts sample this stage only by chance and cannot provide the same subject with and without it. A generator that adds a specified amount of this pattern to a sampled brain would produce such counterfactual volumes. This requires precision: early atrophy changes regional volumes by a few percent. Some released generators are conditioned on regional volumes (Pinaya et al., 2023; Puglisi et al., 2025); to our knowledge, none bounds the change of the others.

Inference-time guidance steers a frozen generator along the gradient of a differentiable score (Chung et al., 2023; Yu et al., 2023; Ye et al., 2024). Its strength, a free parameter, scales a change but does not specify it in measured units. In medical imaging, a request is quantitative: the target structure should change by a stated amount and every other measured structure stay within a stated bound. Following the design-by-contract principle of Meyer (1992), we state such a request as a contract. A contract is a condition on the returned volume, declared in advance and checked by a named verifier.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

The verifier, the segmentation tool SynthSeg (Billot et al., 2023), closes the loop and reads the contract in units of its test-retest variability; a second tool rereads the volumes (Appendix C). The contract bounds the change of every other planned structure by a fixed fraction of the target change, identical for all generators. An unsatisfiable request ends in an abstention, without a volume.

A segmentation tool is not differentiable and cannot be a gradient objective. We therefore propose a closed-loop controller that uses it in two roles. Once per generator, its readings train the surrogate, which proposes control directions. For each request, it measures the effect of an additive perturbation of the sampling trajectory and decides acceptance. The measurement is made on the noise of the request, because this effect is much smaller than the variability between brains. Two volumes are generated from it with perturbations of opposite sign, and half the difference of their readings estimates the response. It then updates a local response model and plans the next perturbation. It returns a volume only when the contract holds, and none when its probe budget runs out.

On five released brain MRI generators, the same code and the same bound satisfy most requests after one or two pairs, without per-generator tuning. Along a fixed direction, the response is close to linear, with a scale that is constant across magnitudes and specific to each generator. One pair of generations measures this scale. Without feedback, a generator produces between a third and the whole of the requested change and changes unrequested structures. Gradient guidance at frozen step sizes also changes unrequested structures. The gain comes from measuring the response and correcting the returned volume, not from the direction: given the verifier's step, guidance reaches the error of the paired variant on MONAI and BrLP profiles. Our contributions are as follows:

- A formulation of inference-time steering as a contract in measured units, verified on the returned volume by an external, non-differentiable instrument.
- A controller that identifies the response of a frozen generator on each request by rank-one updates whose result does not depend on the parameterisation of the directions.
- An evaluation on five released generators, on single-structure requests and on the AD atrophy profile of real patients, against four guidance methods, a mask control and two medical baselines, with acceptance by one segmentation tool and verification by a second, and a transfer to a frozen physics emulator.

2 Related work

A frozen diffusion or flow model can be steered by adding a term to each sampler step; guidance methods differ in how the objective's gradient is computed. DPS (Chung et al., 2023) steps on a measurement residual, and FreeDoM (Yu et al., 2023) differentiates a score, both through the network. MPGD (He et al., 2024) corrects the clean estimate on the data manifold, and FlowChef (Patel et al., 2025) avoids the gradient through the network. Universal Guidance (Bansal et al., 2024) adds recurrence, and TFG (Ye et al., 2024) places these methods in one design space. FlowOpt (Ronai et al., 2025) optimises the source latent through the sampling path, and FMRG (Huang et al., 2026) guides a few-step flow map on a reward. These methods require a differentiable objective, which an external instrument does not provide, and keep a free strength, which scales an intervention without fixing it in measured units. Analyses of flow-matching guidance show the response depends on the sampling trajectory (Feng et al., 2025; Potaptchik et al., 2026). Our controller therefore measures each sample's change. To our knowledge, no method accepts an amount on one quantity with a bound on the others, measures the result externally and withholds it when unmet.

Without a gradient, a non-differentiable instrument can be used in two ways. Sample selection ranks candidates that the frozen model itself produces. Examples are best-of-N, applied by the released TFG code to its guidance score (Ye et al., 2024), and soft value-based decoding (Li et al., 2024), which keeps at each step the continuation scored highest by a non-differentiable reward. These methods return an output of the prior, whose change is drawn from its variability. The alternative acts on the sampler and estimates the action-to-reading map from measurements alone. Each column of that map requires a pair of generations. For one request, we therefore start from a population estimate and refine it with the rank-one correction of Broyden (1965), the minimal change that reproduces each new measurement. These measurements, in the instrument's units, also give the stopping criterion. Noise optimisation adapts the starting noise of one sample to a reward, by gradient ascent in ReNO (Eyring et al., 2024) and by zeroth-order estimates in DNO (Tang et al., 2025). Our pair

of opposite perturbations follows the two-sided measurement of SPSA (Spall, 1992), and correcting each probe by the error of the previous one follows iterative learning control (Bristow et al., 2006).

In brain MRI, anatomy has been controlled by conditioning or guiding a few structures. BrLP (Puglisi et al., 2025) conditions a latent diffusion model on age and regional volumes. AG-LDM (Wan et al., 2026) steers hippocampal and ventricular volume with a guidance term, and reports that its diagnosis covariate moves the anatomy in the unexpected direction, as we observe on BrLP. Atrophy simulators deform a real scan by a specified amount, with a topology-preserving transformation (Karaçali & Davatzikos, 2006), a finite-element model (Camara et al., 2006) or a biophysical model with realistic intensity variation (Khanal et al., 2017); they give known changes on real anatomy, whereas we control a frozen generator with a measured contract. Deep structural causal models give counterfactuals of real brain MRI (Pawlowski et al., 2020), and work on counterfactual validity defines properties of an edit, such as the requested effect, composition and reversibility (Sanchez et al., 2022; Monteiro et al., 2023; Ribeiro et al., 2023). Sun et al. (2025) turn these properties into a six-axis assessment for 3D brain MRI. We adopt it with thresholds from the measured test-retest variability of the verifier and test each volume against its own request. Kondrateva et al. (2025) assess segmentation reproducibility on generated brains with a threshold from prior work, whereas we measure the variability defining the unit of a request. CardioComposer (Kadry et al., 2026) fits target measurements with a loss in cardiac imaging, without constraining what must not change.

3 METHOD

3.1 GENERATOR, VERIFIER AND UNITS

A frozen generator G maps a starting noise z to a 3D volume through the sampler released by its authors, which we leave unmodified. The verifier J, the segmentation tool SynthSeg (Billot et al., 2023), returns the volumes of m anatomical structures and is not differentiable. A differentiable surrogate $\hat{J}$ proposes control directions. The surrogate is a 3D U-Net of 1.46M parameters, trained once per generator on free samples labelled by J, on noises disjoint from the evaluation. It sums label probabilities into volumes, and a per-structure affine calibration maps them to the scale of J. A structure whose calibration slope is below 0.30 is declared unreadable before any run, and requests on it are still counted. Every acceptance decision is made by J, never by the surrogate.

Changes are measured in noise floors. The floor F_j of structure j is the test-retest variability of the verifier pipeline (acquisition, preprocessing and segmentation): the mean absolute difference between the readings of J on two same-session scans of one person, over 92 pairs of real scans, in percent of the volume (Appendix B), 1.19% for the hippocampus and 3.07% for the amygdala. On generated volumes, it is a per-structure unit that makes structures comparable and ties each threshold to a measured property of the verifier. Δ_j floors is the change in percent of j divided by F_j.

3.2 STEERING INTERVENTION

The sampler produces states $z_0, \ldots, z_n$, with z_0 the starting noise, $z_{i+1} = S_i(z_i)$ one released step, and the last state decoded. As in guidance methods (Chung et al., 2023; Yu et al., 2023; Ye et al., 2024), we add a term at each step, here along a fixed direction u and normalised over the trajectory,

$$z_{i+1} \;=\; S_i(z_i) \;+\; \beta\,\frac{w(\tau_i)\,\Delta\tau_i}{Z}\,u, \qquad Z = \sum_i w(\tau_i)\,\Delta\tau_i, \tag{1}$$

where τ_i is the fraction of the trajectory travelled, $\Delta\tau_i$ the length of step i, and w a non-negative weight, uniform ($w = 1$) in all experiments. We call this additive term a *perturbation* of dose β along u. Dividing by Z makes β the total amount injected, independent of whether the sampler uses 25 or 500 steps; the response per unit dose depends on where the dose is spent (late-to-early ratio 0.45 to 1.98, Appendix C). At $\beta = 0$ the loop reduces to the released sampler, which we verify. We write $x = G(z; \beta u)$ for the steered volume and $x_0 = G(z; 0)$ for the unperturbed *free sample*.

3.3 CONTRACTS

A request names a structure k and a change t_k in floors, such as a loss of one floor of hippocampal volume. We write t for the request vector, with entry t_k on the target and zero on the other structures.

With $\Delta \in \mathbb{R}^m$ the change read by J and $\mathrm{leak}(\Delta) = \max_{j \neq k} |\Delta_j| / |\Delta_k|$, the contract

$$|\Delta_k - t_k| \leq \tfrac{1}{4} \qquad \text{and} \qquad \mathrm{leak}(\Delta) \leq \rho\,(1 + \mathrm{tol}), \tag{2}$$

consists of a *target clause* and a *leak clause*, with a tolerance $\mathrm{tol} = 0.1$ on the measured leak. Every single-structure request has $t_k = 1$, so the target clause also bounds the change to within 25% of the request. The value $\rho = 0.125$, used as the cap and in the directions below, is about three times the largest paired placebo noise, 0.038 floors (Section 5.2), and lies inside the plateau of an ablation of ρ in the directions, where the satisfied requests change by at most 3 of 24 per generator (Appendix C). A hippocampus that loses one floor may therefore change no other planned structure by more than an eighth of its floor. The contract and its cap are read by SynthSeg, which closes the loop; a second tool rereads the volumes (Appendix C). The two clauses respond differently.

Observation 1. *Where the reading responds linearly to a perturbation βu, the obtained change is proportional to β and* $\mathrm{leak}(\Delta)$ *does not depend on it: the leak depends only on u.*

The target clause is therefore satisfied by rescaling the dose once a pair measures the response. However, a direction leaking more than ρ cannot satisfy the leak clause at any dose and must change.

Contract-6 applies Eq. 2 to the six planned structures: lateral ventricles, cortex, white matter, hippocampus, amygdala and thalamus. *Contract-21* adds the fifteen non-ventricular labels outside the plan under a looser cap $\rho_{\mathrm{out}} = 0.25$ (first condition below), since no direction holds them. The *profile contract* covers the AD request, a pattern $t \in \mathbb{R}^6$ with no protected structure. The paired reading must lie close to t relative to its largest component (second condition). Since that bound is loose on small components, the returned volume must also satisfy each component within 25% of its request or within one floor (third):

$$\max_{j>6} \frac{|\Delta_j|}{|\Delta_k|} \leq \rho_{\mathrm{out}}(1{+}\mathrm{tol}), \quad \frac{\max_j |\Delta_j - t_j|}{\max_j |t_j|} \leq \rho\,(1{+}\mathrm{tol}), \quad |\Delta_j {-} t_j| \leq \max(\tfrac{1}{4}\,|t_j|,\, 1)\ \forall j. \tag{3}$$

The paired variant stops on the second condition, and the closed loop on the third.

3.4 Probes and the paired reading

A change of one floor is much smaller than the difference between two generated brains. The controller therefore measures every perturbation with a *probe*: two generations from the same noise, $x_\pm = G(z; \pm\beta\, u)$, each read relative to the free sample. Their readings decompose into an odd part $r = \frac{1}{2}\big(J(x_+) - J(x_-)\big)$, which changes sign with the perturbation, and an even part $e = \frac{1}{2}\big(J(x_+) + J(x_-)\big) - J(x_0)$, which does not. Anatomy and any change independent of the sign of the perturbation *cancel* in r, our estimate of the response. Under a linear response, $e = 0$.

The controller returns the positive arm x_+, whose change is $J(x_+) - J(x_0) = r + e$. Every request therefore has two verdicts. The *paired verdict* on r states whether the response was identified, and the *positive-arm verdict* on $r + e$ states whether the returned volume satisfies the request. Neither implies the other. Cancellation requires both volumes to remain close to the free sample, which excludes methods that re-inject noise, such as recurrent TFG.

3.5 Local response model

The controller acts through six fixed directions, one per planned structure. Let $g_1, \ldots, g_6 \in \mathbb{R}^d$ be the gradients of the surrogate's six readings on the free sample with respect to the final state z_n, computed through the authors' decoder only (for the pixel-space MedForj, the volume itself). Let Γ be the matrix with these rows, and $M = \mathrm{diag}(100/(v_j F_j))\,\Gamma\Gamma^\top$, with v_j the free-sample volume. The entry M_{ji} is the predicted change of structure j in floors per unit step along g_i. Along a raw gradient g_i, the leak read by the verifier has a median of 0.18 to 0.49. Direction i is therefore the combination that changes, on the surrogate, structure i by one unit and no other by more than ρ:

$$Mc^{(i)} = s^{(i)}, \qquad s_i^{(i)} = 1, \quad s_j^{(i)} = \mathrm{clip}\big(M_{ji}/M_{ii},\, -\rho,\, \rho\big)\ (j \neq i), \qquad b_i = \Gamma^\top c^{(i)} / \|\Gamma^\top c^{(i)}\|. \tag{4}$$

With the b_i as the rows of $B \in \mathbb{R}^{6 \times d}$, a perturbation is $B^\top a$ for a dose vector $a \in \mathbb{R}^6$. Since the directions are not orthogonal, its squared norm is $a^\top W a$ with $W = BB^\top$.

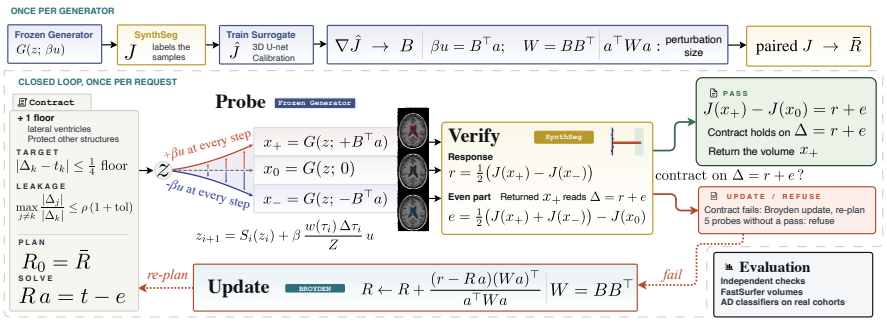


Figure 1: **Controller overview.** Setup once per generator, then the loop of Eq. 5 for each request.

The response matrix $R \in \mathbb{R}^{m\times 6}$ contains the readings of the verifier. Its entry R_{ji} is the change of structure j in floors per unit dose along b_i, estimated from odd parts. Its off-diagonal entries are the cross-responses that produce the leak. Under a linear response, a dose a produces Ra, and a request is planned by solving $Ra = t$. This compensates each cross-response with small doses along the other directions. The quantities derived from it do not depend on how the directions are mixed.

Proposition 1. *Let the directions be linearly independent, which makes W invertible, and let $R = DB^\top$, with D the derivative of the odd part of the reading with respect to the perturbation. For any invertible A, replacing B by AB leaves $K = RW^{-1}R^\top$ unchanged, and $K = D\,\Pi\,D^\top$ with Π the orthogonal projector onto the span of the directions. When K is invertible, the smallest perturbation in that span that produces, to first order, a change t has squared norm $t^\top K^{-1} t$.*

K is the response per unit perturbation within the span of the directions. The quantity $t^\top K^{-1} t$, equal to $a^\top W a$ for the planned dose, is the smallest perturbation a request requires in that span, in the metric W and not in image space, known before generation. It is small when the target moves alone, and large when the request opposes a strong cross-response or a barely reachable structure.

3.6 Closed-loop control

In a setup once per generator, J labels free samples, the surrogate is trained on them, and a population model $\bar{R}$ is measured: each column is the odd part of a probe along one direction divided by the dose, with the median over held-out noises. On each noise, the surrogate's gradients on the free sample give B and W. Each request starts from $R = \bar{R}$ and repeats four steps:

$$\begin{aligned} &\text{plan: } a = R^{-1}(t-e), \qquad &&\text{probe: } x_\pm = G(z; \pm B^\top a), \ \text{ read } r \text{ and } e, \\ &\text{test: } r + e \text{ satisfies the contract}, \qquad &&\text{update: } R \leftarrow R + \frac{(r - Ra)\,(Wa)^\top}{a^\top W a}. \end{aligned} \tag{5}$$

In the plan and the update, R is the 6×6 block of the planned structures. The plan targets the positive arm. Here e is the even part of the previous probe, zero at the first step, and the returned x_+, read as $r + e$, is the planned volume. A new plan obeys $\|a - a_{\text{prev}}\| \le \frac{1}{2}\|a_{\text{prev}}\|$, and no perturbation may exceed three times the first. The test applies both clauses to the positive arm. If they hold, x_+ is returned. Otherwise, the Broyden update fits the model to the measured odd response along the probed dose. Under contract-21 (Appendix A), R has 21 rows, the update is Euclidean, and the plan minimises the largest outside response subject to a unit target response and the planned cap.

Proposition 2. *Let $R^+ = R + (r - Ra)(Wa)^\top/(a^\top W a)$. Then (i) $R^+ a = r$; (ii) R^+ is the matrix closest to R in $\|(\cdot)\,W^{-1/2}\|_F$ among those satisfying (i); (iii) $R^+ b = Rb$ whenever $b^\top W a = 0$, that is, for every perturbation orthogonal to the one probed; (iv) the update commutes with a change of basis $B \to AB$, which the Euclidean update $(r - Ra)\,a^\top/(a^\top a)$ does only for orthogonal A up to a scalar; and (v) if the true response is the predicted one times a nonzero scalar gain, $r/r_k = t/t_k$, the first probe satisfies the leak clause, and one rescaling of the dose satisfies the target.*

By (ii) and (iii), a probe corrects the model only along the measured perturbation, in the geometry defined by W, and preserves it along orthogonal perturbations. By (iv), the update, like K, does not depend on the parameterisation. For a single-structure request, (v) also holds with one gain per

structure, and the leak comes only from errors in the cross-responses (Appendix A). Both propositions are a local analysis, valid where the response is linear, the directions independent and K invertible, not a guarantee of convergence.

The loop ends with *acceptance*, when the positive arm satisfies the contract and is returned, or with *abstention* after five probes, without a volume, counted as a failure. Every returned volume therefore satisfies the contract on the verifier's reading, and an experiment measures the *coverage*, the fraction of requests ending with a returned volume. We call this procedure the *closed loop*. It requires no gradient of the verifier or through the denoiser. FastSurfer (Henschel et al., 2020), a second segmentation tool, and AD classifiers trained on real cohorts re-read its volumes for evaluation only.

Two reference procedures complete the evaluation. The *paired variant*, used to study identification, plans $Ra = t$ with the cross-responses of t clipped to $\pm\rho$. It tests the leak clause on r alone, which does not depend on the dose, then rescales the dose once by t_k/r_k and regenerates the pair. It uses four probes, extended to six while the leak decreases. It abstains when $|e_k| > \frac{1}{4}|r_k|$ or when the target response is below three times that of a placebo perturbation. The *open loop* applies one perturbation along the resolved direction of the target at the surrogate's dose, without probes.

4 Experimental setup

The evaluation addresses three questions. First, whether one controller, without per-generator tuning, satisfies contracts on generators with different architectures. Second, whether it leaks less than guidance methods at the same target change. Third, whether its result holds for an instrument that it never calls. We use five generators as released: MONAI brain LDM (Pinaya et al., 2023), BrLP (Puglisi et al., 2025), MAISI v2 (Zhao et al., 2026), WaveDiT (Danese et al., 2026) and MedForj (Remedios et al., 2025). Trained on different cohorts, some not public, they comprise two latent diffusion models with a variational autoencoder, a rectified flow, a wavelet flow-matching model and a pixel-space diffusion model. All are run at 100 sampler steps. The population model and the hyper-parameters come from noises outside the evaluation. On each starting noise, we request a gain of one floor on each of the six planned structures, and a loss on 78 requests (Appendix C). The closed loop is evaluated on 100 held-out noises per generator, 600 requests. Following the effectiveness and minimality axes of Sun et al. (2025), we report the contract, the leak (leak_6) and the outside leak (leak_{15}, the same ratio over the fifteen non-ventricular labels outside the plan), and the cost. SynthSeg segments 31 labels, hemispheres separate, and the six structures of the contract are sums of them. FastSurfer (Henschel et al., 2020) segments the paired-variant and open-loop probes on the first four noises and closed-loop volumes on 228 requests per generator (Appendix C).

We compare with four guidance methods on the authors' released samplers, FreeDoM (Yu et al., 2023), FlowChef (Patel et al., 2025), MPGD in its latent-diffusion form (He et al., 2024) and TFG (Ye et al., 2024), and a mask control restoring the unrequested region at every step as in RePaint (Lugmayr et al., 2022). TFG is a family; twelve configurations were screened outside the evaluation, and two were confirmed, TFG (mean) and TFG (mean+var.). Each method guides on the surrogate's reading of the target, as designed, and a variant that also penalises the protected readings is reported in the appendix. FreeDoM differentiates through the network, which the released MAISI and WaveDiT code runs without a gradient (tiled decoder, velocity). On these generators, FreeDoM-lin (†), which treats this part as constant, replaces it. The medical baselines are the native conditioning of BrLP and classifier guidance (Dhariwal & Nichol, 2021) on the AD classifier below.

A guidance method receives a step size rather than a target, and the change a step produces depends on the generator, the structure and the noise. We therefore specify how the step is chosen. In the *main comparison*, the step of each baseline per structure is fitted on two of the first four noises and frozen. On the 78 held-out requests, the baseline generates one pair per request, read as ours, as the method would be used without a verifier. The *a posteriori comparison* interpolates, for every request, the step that produces exactly one floor. This gives the baseline, at no cost, the information the closed loop obtains by probing. Failures, wrong signs and missing requests remain in the denominator.

The clinical request is the age-corrected percentage difference between AD patients and controls in the six planned volumes, each divided by the total segmented volume. It is read by SynthSeg on 46 valid scans of 48 BrainLat subjects (Prado et al., 2023) balanced by centre and diagnosis: $+30.6\%$ on the lateral ventricles, -11.3% on the hippocampus, -9.6% on the amygdala, and smaller changes

on the other three. Each generator moves its own free sample toward a fraction f of this profile under the profile contract. This is evaluated at $f = 0.5$ and 1 on 100 held-out noises, on twelve other noises against the conditioning of BrLP and the guidance baselines, and descriptively against classifier guidance on four noises, where the loop runs at $f = 0.1$ to 1. A classifier that we train outside the loop reads the result. It is a logistic regression on the six volumes divided by the total segmented volume, trained on 487 scans from ADNI (Jack et al., 2008) and OASIS-3 (LaMontagne et al., 2019) and validated on 136 BrainLat subjects from other sites (AUC 0.892). We denote its patient-class probability by P(AD). Using only volumes, it cannot exploit image-level synthesis artefacts (Geirhos et al., 2020). It compares the request with cohorts on which it was not measured, rather than diagnosing the sample. The same recipe on FastSurfer volumes gives a second classifier.

5 Results

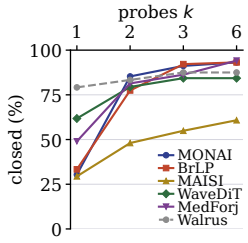
5.1 Single-structure requests

Table 1 reports the share of one-floor requests satisfying the contract and the median leak; Figure 2 shows four returned volumes. The closed loop satisfies 84 to 100% of the requests on every generator, with a median leak of 0.05 to 0.09, below the cap of 0.125, and fewer than three pairs per request on average. The other 135 requests, 98 on MAISI, end after five pairs without a volume; according to the verifier, no returned volume violates the contract. On two further sets of thirteen noises, it satisfies 87 to 100% and 78 to 100% (Appendix C). FastSurfer, which the loop never calls, agrees with SynthSeg on the target (0.89 to 1.17 times on four generators) but segments the protected structures differently: their changes correlate at 0.17 and are read 2 to 6 times larger. Under FastSurfer, the largest protected change stays below one of its floors in 162 to 228 of 228 requests, and below a quarter of the target, the cap of structures outside the plan, in 82 to 179 (Appendix C).

The paired variant identifies the response in 84 to 94% of requests on four generators and 61% on MAISI. Its regenerated positive arm satisfies the contract in 48 to 84%, and the closed loop is therefore planned for the returned volume. The two readings differ by the even part e, common to both signs of the perturbation, and a volume planned without it misses the request by e.

At frozen step sizes, the baselines satisfy the contract in at most 28% on the paired reading, and 26% on their returned volume. Their median leaks are 0.14 to 1.34. On the paired reading, the comparison least favourable to us, the paired variant satisfies 51 to 75 contracts against 1 to 22, with at most one request satisfied by the baseline alone. All thirteen comparisons are significant when the requests that share a noise are counted as one (sign permutation by noise, Holm-adjusted $p = 0.0016$). On 25 of the 100 noises, the closed loop satisfies 150 of 150 contracts on the returned volume on MONAI, BrLP and WaveDiT, against 26, 24 and 46 for FlowChef, and 144, 142 and 136 against 26, 25 and 49 on the paired reading ($p = 3.0 \times 10^{-8}$ by noise on both). A penalty on the protected readings raises the paired rate of the baselines to at most 58% on the first four noises, while their returned volumes satisfy the contract in at most 4%. A joint CN score over 31 regional volumes, validated on 806 real

(a)



(b)

method	MONAI		BrLP		MAISI		WaveDiT		MedForj	
open loop	17	0.241	4	0.240	10	0.273	8	0.193	0	0.186
paired variant, paired	**93**	**0.124**	**93**	**0.125**	**61**	**0.124**	**84**	**0.102**	**94**	**0.105**
paired variant, returned	71	0.126	67	0.127	48	0.125	64	0.112	84	0.106
paired variant, held-out	91	0.125	96	0.125	65	0.127	85	0.102	92	0.102
closed loop (ours)	**99**	**0.059**	**100**	**0.046**	84	0.092	99	0.057	**96**	**0.053**
closed loop, held-out	99	0.050	**100**	**0.047**	**87**	**0.081**	**100**	**0.052**	87	0.074
FreeDoM	14	0.509	8	0.508	21	0.33†	28	0.182†	1	1.340
FlowChef	17	0.380	15	0.380	14	0.292	24	0.184	12	0.184
MPGD	18	0.411	17	0.471	14	0.45	15	0.143	19	0.207

Table 1: **Single-structure requests.** (a) Requests closed within k probes by the paired variant, on the paired reading of its final pair: 102 requests per generator, 24 on Walrus. (b) Bars: requests satisfying the contract (%); grey: median leak (cap 0.125, threshold 0.1375). Top: paired variant, 102 requests. Bottom, on 78 held-out requests: baselines at a frozen step and paired variant (paired reading), closed loop; closed loop (ours) on 600 requests from 100 further noises per generator (cohorts in Appendix B). †FreeDoM-lin. Bold: best; underlined: second best.

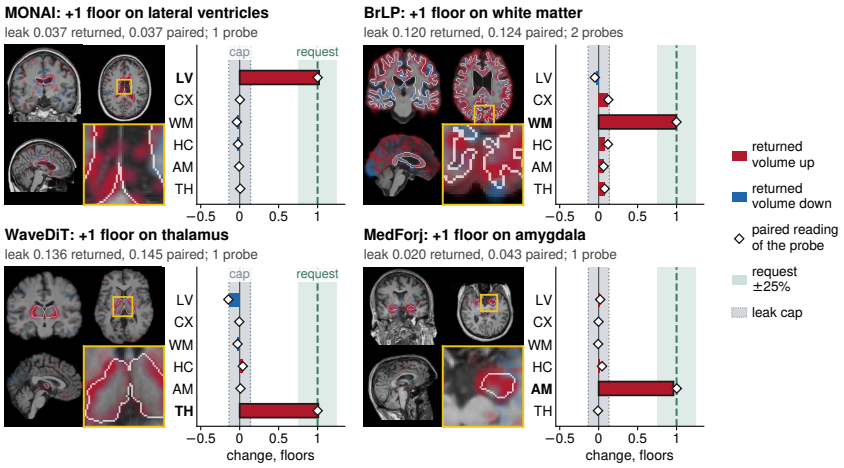


Figure 2: **Examples of single-structure requests.** Four requests of $+1$ floor on one structure. Left: the free sample with the change of the returned volume (red gained, blue lost; yellow box: zoomed region). Right: the six planned structures in floors; bars are the returned volume, diamonds the paired reading of the probe; green band: request $\pm 25\%$; grey band: leak threshold.

scans, accepts the last probe of each request at the rate of its free sample on every generator, from 0.93 on BrLP to 0.01 on MAISI (756 requests each, abstentions included). In 21 of 3,780 requests, only the free sample is accepted (Appendix C). On MONAI, BrLP and WaveDiT, both hold in 54 to 73 of 78 requests for the closed loop, at most 13 for a baseline.

5.2 Probe efficiency and response linearity

A one-floor request must be measured on its own noise. Across brains of one generator, one structure has a standard deviation of 1.3 to 99 floors, and an unpaired test would need 28 to more than 100,000 volumes per group. In contrast, placebo perturbations give paired responses with a standard deviation of at most 0.038 floors (Appendix C). Along a fixed direction, the response is proportional to the dose from 0.5 to 5 floors on every generator, except for the MAISI hippocampus, unreadable by its surrogate. Its gain over the surrogate's prediction ranges from 0.35 on MedForj to about 1 on MAISI, on six noises per generator (Appendix A). Guidance methods use the surrogate's scale as given, whereas one probe measures it (Proposition 2). With two sets of directions, K agrees within 1.2% on three generators and 11% on all five, while their median leak varies up to threefold.

We add the components one at a time on 24 requests. A direction at the predicted dose satisfies the contract in 0 to 17%, and measuring the dose on one pair in 25 to 46%. Planning with the population model reaches 25 to 71%, and the rank-one updates 46 to 100%. On the returned volume of the 78 held-out requests, the closed loop reaches 87 to 100%. A block controller that measures all six columns on the request's noise, without updates, satisfies the paired contract in 97 to 100% of requests on four generators and 81% on MAISI. It requires seven pairs per request, or two when six requests share a noise, against 1.5 to 3 for the rank-one loop, which reaches the same coverage with one generation per probe, also on a profile (Appendix C). Within two probes, the paired variant closes 77 to 85% of 102 requests on four generators (Table 1a), while a baseline uses 200 to 1000 gradient evaluations per request, after a step-size calibration of 24 to 108 pairs per generator and method. On the 78 held-out requests, the closed loop costs 0.02 to 0.11 GPU hours per request served and the baselines 0.4 to 13.2 (Appendix D); a request lasts 58 to 310 s including verification.

5.3 Multi-structure profile requests

The AD profile requests six changes, from 50 floors (lateral ventricles) to 2 (thalamus). At two doses, each component must be within 25% of its request or one floor, the verifier's mean test-retest difference. On 100 held-out noises, the closed loop satisfies all six in 199, 200, 168, 169 and 161 of 200 requests on MONAI, BrLP, MAISI, WaveDiT and MedForj, and unmet requests return none.

The 25% clause alone gives 109, 129, 68, 29 and 81; the floor clause decides only components below four floors, where a quarter of the request is less than one floor (Appendix C). Comparisons use twelve other noises (four for classifier guidance).

BrLP was trained with five of these volumes as conditioning inputs. Its conditioning, used as released, calibrated per structure or corrected by feedback on the same verifier, satisfies all six components in 0 of 24 requests. It changes the thalamus, which is not one of its inputs, against the request in every request. On the five inputs, the three variants satisfy the criterion in 5, 15 and 23 requests, against 24 for the closed loop, and a paired Wilcoxon test on the components satisfied favours the closed loop against each variant ($p < 10^{-5}$). Stopping the same loop on the paired reading satisfies the criterion in 58% of the requests on BrLP, which quantifies the gain of correcting the returned volume. FlowChef, FreeDoM and MPGD at a frozen step satisfy all six components in at most 5 of 24 requests, most often failing on the hippocampus over all generators. Classifier guidance raises P(AD) to 1.00 on MONAI, while the lateral ventricles, which account for most of the request, shrink in 19 of 20 requests. On the twelve comparison noises, the closed loop satisfies all six components in 24 of 24 requests on BrLP and in 19 to 24 of 24 on the other four generators, including the thalamus.

A separate question is whether an instrument that the controller never calls classifies the resulting anatomy as disease. The classifier built on FastSurfer, trained on the same 487 scans, separates 136 held-out BrainLat subjects with an AUC of 0.90, with median scores of 0.27 for patients and 0.01 for controls. At $f = 1$, over two batches of twelve noises per generator, it scores the steered volume above its free sample in 120 of 120 requests, and the opposite volume at the level of controls. In a replication on twelve new noises, the steered medians, 0.08 to 0.20, lie within the interquartile range of real patients. The score does not rank methods. On MONAI, MAISI and MedForj, classifier-guided volumes score highest while their anatomy moves against the profile. On MONAI, BrLP and WaveDiT, the returned profile volumes are accepted as AD by the joint score (Appendix C).

5.4 Transfer to a physics emulator

Neural emulators increasingly replace costly numerical solvers in physics, and steering one toward a specified change in one quantity, with the others held fixed, yields controlled flow variations for sensitivity studies and for testing analyses on known changes. The paired variant, with the same feedback scheme, steers a frozen emulator, Walrus (McCabe et al., 2025), on the shear flow of The Well (Ohana et al., 2024) (Appendix F). The sample is the flow field at the end of a rollout, and a floor is the median error of the emulator against numerical ground truth. Each request asks for one floor on energy, enstrophy or mixing while protecting eight others. Guidance methods act on the steps of a diffusion or flow sampler and do not apply to Walrus, an autoregressive emulator. The paired variant requires only a perturbation added to the rollout and an instrument applied to its output, and runs with the brain settings and an adapter for the new domain. On 24 requests over eight held-out trajectories, the paired variant satisfied 21 contracts and the open loop 12, and on eight further trajectories 23 of 24. All but one successful request on the first cohort required a single probe. Sample selection, the analogue of best-of-N (Section 2), satisfies none of 12 exploratory requests when choosing among up to 7955 initial fields, whose energy and enstrophy decay together.

6 Conclusion

We introduced a closed-loop controller for frozen 3D brain MRI generators. A request specifies the volume change of one structure and bounds those of the other planned structures, measured by an external segmentation tool. In the component analysis, a single direction at the dose predicted by the surrogate satisfies 0 to 17%, and gradient guidance at frozen step sizes at most 28% of the 78 held-out requests. With one configuration on five released generators from four model families, the closed loop satisfies 84 to 100% of 600 held-out single-structure requests per generator, at 1.45 to 2.87 pairs on average, and 161 to 200 of 200 six-component AD profile requests. Unmet requests return no volume; every returned volume satisfies its contract under the verifier that defines it. A second segmentation tool, never called by the loop, measures the target change at 0.89 to 1.17 times the verifier on four generators and every protected change below one of its floors in 162 to 228 of 228 requests. Contracts are defined with one segmentation tool on T1-weighted MRI, and our claims concern the volumes it measures. The response model is local and is identified on the starting noise of each request.

ETHICS STATEMENT

This work uses brain MRI from ADNI (Jack et al., 2008), OASIS-3 (LaMontagne et al., 2019) and BrainLat (Prado et al., 2023), accessed under the data use agreements of each cohort; no new data were collected. The paper reports aggregate measurements only, and no subject identifier appears in the paper or in the released code. The five generators and the two segmentation tools are used as released by their authors, under their licences. The steered volumes are synthetic, and the AD profile is a group difference between patients and controls: a steered volume does not represent a patient.

REPRODUCIBILITY STATEMENT

The supplementary material contains the code of the controller, the baselines and every analysis, with the configuration of each experiment and a map from every table and figure to the command that produces it. The five generators are used as released by their authors, from their repositories and checkpoints, and $\beta = 0$ reproduces each released sampler exactly. Starting noises are drawn from fixed seeds, and the cohorts of noises are listed in Table 4. Appendix B gives every setting and the setup cost. The real scans come from ADNI, OASIS-3 and BrainLat, available under their data use agreements.

AI DISCLOSURE STATEMENT

Generative AI assistants were used only for language polishing. The authors are responsible for all claims, results, and the final manuscript.

REFERENCES

Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. In *ICLR*. OpenReview.net, 2024.

Benjamin Billot, Douglas N. Greve, Oula Puonti, Axel Thielscher, Koen Van Leemput, Bruce Fischl, Adrian V. Dalca, and Juan Eugenio Iglesias. Synthseg: Segmentation of brain MRI scans of any contrast and resolution without retraining. *Medical Image Anal.*, 86:102789, 2023.

Douglas A. Bristow, Marina Tharayil, and Andrew G. Alleyne. A survey of iterative learning control. *IEEE Control Syst. Mag.*, 26(3):96–114, 2006. doi: 10.1109/MCS.2006.1636313.

C. G. Broyden. A class of methods for solving nonlinear simultaneous equations. *Mathematics of Computation*, 19:577–593, 1965.

O. Camara, M. Schweiger, R. I. Scahill, W. R. Crum, B. I. Sneller, J. A. Schnabel, G. R. Ridgway, D. M. Cash, D. L. G. Hill, and N. C. Fox. Phenomenological model of diffuse global and regional atrophy using finite-element methods. *IEEE Trans. Med. Imaging*, 25(11):1417–1430, 2006. doi: 10.1109/TMI.2006.880588.

Sihong Chen, Kai Ma, and Yefeng Zheng. Med3D: Transfer learning for 3D medical image analysis. *arXiv preprint arXiv:1904.00625*, 2019.

Hyungjin Chung, Jeongsol Kim, Michael Thompson McCann, Marc Louis Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *ICLR*. OpenReview.net, 2023.

Danilo Danese, Angela Lombardi, Giuseppe Fasano, Matteo Attimonelli, and Tommaso Di Noia. Wavedit: Distribution-aware wavelet flow matching for efficient 3d brain MRI synthesis. *CoRR*, abs/2606.08670, 2026.

Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion models beat gans on image synthesis. In *NeurIPS*, pp. 8780–8794, 2021.

Luca Eyring, Shyamgopal Karthik, Karsten Roth, Alexey Dosovitskiy, and Zeynep Akata. ReNO: enhancing one-step text-to-image models through reward-based noise optimization. In *NeurIPS*, volume 37, pp. 125487–125519, 2024. doi: 10.52202/079017-3987.

Ruiqi Feng, Chenglei Yu, Wenhao Deng, Peiyan Hu, and Tailin Wu. On the guidance of flow matching. In *ICML*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025.

Giovanni B. Frisoni, Nick C. Fox, Clifford R. Jack, Philip Scheltens, and Paul M. Thompson. The clinical use of structural MRI in Alzheimer disease. *Nature Reviews Neurology*, 6(2):67–77, 2010.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard S. Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nat. Mach. Intell.*, 2(11):665–673, 2020.

Yutong He, Naoki Murata, Chieh-Hsin Lai, Yuhta Takida, Toshimitsu Uesaka, Dongjun Kim, Wei-Hsiang Liao, Yuki Mitsufuji, J. Zico Kolter, Ruslan Salakhutdinov, and Stefano Ermon. Manifold preserving guided diffusion. In *ICLR*. OpenReview.net, 2024.

Leonie Henschel, Sailesh Conjeti, Santiago Estrada, Kersten Diers, Bruce Fischl, and Martin Reuter. FastSurfer - a fast and accurate deep learning based neuroimaging pipeline. *Neuroimage*, 219 (117012):117012, October 2020.

Jerry Huang, Justin Lin, Sheel Shah, Kartik Nair, and Nicholas M. Boffi. How to guide your flow: Few-step alignment via flow map reward guidance. *CoRR*, abs/2604.27147, 2026.

Clifford R. Jack et al. The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. *Journal of Magnetic Resonance Imaging*, 27(4):685–691, 2008.

Karim Kadry, Shoaib A. Goraya, Ajay Manicka, Abdalla Abdelwahed, Naravich Chutisilp, Farhad R. Nezami, and Elazer R Edelman. Cardiocomposer: Leveraging differentiable geometry for compositional control of anatomical diffusion models. In *The Fourteenth International Conference on Learning Representations*, 2026.

B. Karaçali and C. Davatzikos. Simulation of tissue atrophy using a topology preserving transformation model. *IEEE Trans. Med. Imaging*, 25(5):649–652, 2006. doi: 10.1109/TMI.2006.873221.

B. Khanal, N. Ayache, and X. Pennec. Simulating longitudinal brain MRIs with known volume changes and realistic variations in image intensity. *Frontiers in Neuroscience*, 11:132, 2017. doi: 10.3389/fnins.2017.00132.

Ekaterina Kondrateva, Sandzhi Barg, and Mikhail Vasiliev. Benchmarking the reproducibility of brain MRI segmentation across scanners and time. *CoRR*, abs/2504.15931, 2025.

Pamela LaMontagne, Tammie Benzinger, John Morris, Sarah Keefe, Russ Hornbeck, Chengjie Xiong, Elizabeth Grant, Jason Hassenstab, Krista Moulder, Andrei Vlassenko, Marus Raichle, Carlos Cruchaga, and Daniel Marcus. Oasis-3: Longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and alzheimer disease. *medRxiv*, 12 2019. doi: 10.1101/2019.12.13.19014902.

Xiner Li, Yulai Zhao, Chenyu Wang, Gabriele Scalia, Gokcen Eraslan, Surag Nair, Tommaso Biancalani, Shuiwang Ji, Aviv Regev, Sergey Levine, and Masatoshi Uehara. Derivative-free guidance in continuous and discrete diffusion models with soft value-based decoding. *CoRR*, abs/2408.08252, 2024.

Andreas Lugmayr, Martin Danelljan, Andrés Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *CVPR*, pp. 11451–11461. IEEE, 2022.

Michael McCabe, Payel Mukhopadhyay, Tanya Marwah, Bruno Régaldo-Saint Blancard, François Rozet, Cristiana Diaconu, Lucas Meyer, Kaze W. K. Wong, Hadi Sotoudeh, Alberto Bietti, Irina Espejo, Rio Fear, Siavash Golkar, Tom Hehir, Keiya Hirashima, Géraud Krawezik, François Lanusse, Rudy Morel, Ruben Ohana, Liam Holden Parker, Mariel Pettee, Jeff Shen, Kyunghyun Cho, Miles D. Cranmer, and Shirley Ho. Walrus: A cross-domain foundation model for continuum dynamics. *CoRR*, abs/2511.15684, 2025.

Bertrand Meyer. Applying 'design by contract'. *Computer*, 25:40–51, 1992.

Miguel Monteiro, Fabio De Sousa Ribeiro, Nick Pawlowski, Daniel C. Castro, and Ben Glocker. Measuring axiomatic soundness of counterfactual image models. In *ICLR*. OpenReview.net, 2023.

Sean M Nestor, Raul Rupsingh, Michael Borrie, Matthew Smith, Vittorio Accomazzi, Jennie L Wells, Jennifer Fogarty, Robert Bartha, and Alzheimer's Disease Neuroimaging Initiative. Ventricular enlargement as a possible measure of alzheimer's disease progression validated using the alzheimer's disease neuroimaging initiative database. *Brain*, 131(Pt 9):2443–2454, 2008.

Ruben Ohana, Michael McCabe, Lucas Meyer, Rudy Morel, Fruzsina Julia Agocs, Miguel Beneitez, Marsha Berger, Blakesley Burkhart, Stuart B. Dalziel, Drummond B. Fielding, Daniel Fortunato, Jared A. Goldberg, Keiya Hirashima, Yan-Fei Jiang, Rich R. Kerswell, Suryanarayana Maddu, Jonah Miller, Payel Mukhopadhyay, Stefan S. Nixon, Jeff Shen, Romain Watteaux, Bruno Régaldo-Saint Blancard, François Rozet, Liam Holden Parker, Miles D. Cranmer, and Shirley Ho. The well: a large-scale collection of diverse physics simulations for machine learning. In *NeurIPS*, 2024.

Maitreya Patel, Song Wen, Dimitris N. Metaxas, and Yezhou Yang. Steering rectified flow models in the vector field for controlled image generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15308–15318, 2025.

Nick Pawlowski, Daniel Coelho de Castro, and Ben Glocker. Deep structural causal models for tractable counterfactual inference. In *NeurIPS*, volume 33, pp. 857–869, 2020.

Walter H. L. Pinaya, Mark S. Graham, Eric Kerfoot, Petru-Daniel Tudosiu, Jessica Dafflon, Virginia Fernandez, Pedro Sanchez, Julia Wolleb, Pedro F. Da Costa, Ashay Patel, Hyungjin Chung, Can Zhao, Wei Peng, Zelong Liu, XueYan Mei, Oeslle Lucena, Jong Chul Ye, Sotirios A. Tsaftaris, Prerna Dogra, Andrew Feng, Marc Modat, Parashkev Nachev, Sébastien Ourselin, and M. Jorge Cardoso. Generative AI for medical imaging: extending the MONAI framework. *CoRR*, abs/2307.15208, 2023.

Peter Potaptchik, Adhi Saravanan, Abbas Mammadov, Alvaro Prat, Michael S. Albergo, and Yee Whye Teh. Meta flow maps enable scalable reward alignment. *CoRR*, abs/2601.14430, 2026.

Pavel Prado, Vicente Medel, Raul Gonzalez-Gomez, Agustín Sainz-Ballesteros, Victor Vidal, Hernando Santamaría-García, Sebastian Moguilner, Jhony Mejia, Andrea Slachevsky, Maria Isabel Behrens, David Aguillon, Francisco Lopera, Mario A. Parra, Diana Matallana, Marcelo Adrián Maito, Adolfo M. Garcia, Nilton Custodio, Alberto Ávila Funes, Stefanie Piña-Escudero, Agustina Birba, Sol Fittipaldi, Agustina Legaz, and Agustín Ibañez. The brainlat project, a multimodal neuroimaging dataset of neurodegeneration from underrepresented backgrounds. *Scientific Data*, 10(1):889, Dec 2023. ISSN 2052-4463. doi: 10.1038/s41597-023-02806-8. URL https://doi.org/10.1038/s41597-023-02806-8.

Lemuel Puglisi, Daniel C. Alexander, and Daniele Ravì. Brain latent progression: Individual-based spatiotemporal disease progression on 3d brain mris via latent diffusion. *Medical Image Anal.*, 106:103734, 2025.

Samuel W. Remedios, Aaron Carass, Jerry L. Prince, and Blake E. Dewey. Diffusion-driven generation of minimally preprocessed brain MRI. *CoRR*, abs/2510.26834, 2025.

Fabio De Sousa Ribeiro, Tian Xia, Miguel Monteiro, Nick Pawlowski, and Ben Glocker. High fidelity image counterfactuals with probabilistic causal models. In *ICML*, volume 202 of *Proceedings of Machine Learning Research*, pp. 7390–7425. PMLR, 2023.

Or Ronai, Vladimir Kulikov, and Tomer Michaeli. Flowopt: Fast optimization through whole flow processes for training-free editing. *CoRR*, abs/2510.22010, 2025.

Pedro Sanchez, Antanas Kascenas, Xiao Liu, Alison Q. O'Neil, and Sotirios A. Tsaftaris. What is healthy? generative counterfactual diffusion for lesion localization. In *DGM4MICCAI@MICCAI*, volume 13609 of *Lecture Notes in Computer Science*, pp. 34–44. Springer, 2022.

James C. Spall. Multivariate stochastic approximation using a simultaneous perturbation gradient approximation. *IEEE Trans. Autom. Control*, 37(3):332–341, 1992. doi: 10.1109/9.119632.

Pengwei Sun, Wei Peng, Lun Yu Li, Yixin Wang, and Kilian M. Pohl. Evaluation of 3d counterfactual brain MRI generation. In *DGM4MICCAI@MICCAI*, volume 16128 of *Lecture Notes in Computer Science*, pp. 46–56. Springer, 2025.

Zhiwei Tang, Jiangweizhi Peng, Jiasheng Tang, Mingyi Hong, Fan Wang, and Tsung-Hui Chang. Inference-time alignment of diffusion models with direct noise optimization. In *ICML*, volume 267 of *Proceedings of Machine Learning Research*, pp. 58905–58930. PMLR, 2025.

Cheng Wan, Bahram Jafrasteh, Ehsan Adeli, Miaomiao Zhang, and Qingyu Zhao. Anatomically guided latent diffusion for brain MRI progression modeling. *CoRR*, abs/2601.14584, 2026.

Haotian Ye, Haowei Lin, Jiaqi Han, Minkai Xu, Sheng Liu, Yitao Liang, Jianzhu Ma, James Y. Zou, and Stefano Ermon. TFG: unified training-free guidance for diffusion models. In *NeurIPS*, 2024.

Jiwen Yu, Yinhuai Wang, Chen Zhao, Bernard Ghanem, and Jian Zhang. Freedom: Training-free energy-guided conditional diffusion model. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Can Zhao, Pengfei Guo, Dong Yang, Yufan He, Yucheng Tang, Benjamin Simon, Mason Belue, Stephanie A. Harmon, Baris Turkbey, and Daguang Xu. Maisi-v2: Accelerated 3d high-resolution medical image synthesis with rectified flow and region-specific contrastive loss. In *AAAI*, pp. 13088–13098. AAAI Press, 2026.

A METHOD DETAILS

In this section, we prove the two propositions of the method and test the two assumptions on which they rest, namely that the reading responds linearly to the perturbation and that the response does not depend on how the directions are chosen. Every identity was also checked numerically on random instances.

Notation. The leak ratio of Eq. 2 is evaluated only together with the target clause, which keeps $|\Delta_k| \geq 3/4$ whenever the contract holds; its denominator therefore stays away from zero on every accepted volume. The rows of $B \in \mathbb{R}^{6\times d}$ are the six directions, and a perturbation is $v = B^\top a$, with $\|v\|^2 = a^\top W a$ and $W = BB^\top$ for any scaling or mixing of the directions. The verifier is not differentiable, and only the odd part of its reading is assumed to be linear to first order, $\frac{1}{2}\big(\Delta(v) - \Delta(-v)\big) = Dv + o(\|v\|)$ with $D \in \mathbb{R}^{m\times d}$ (tested in Fig. 3); the even part $\frac{1}{2}\big(\Delta(v) + \Delta(-v)\big)$ is not assumed to vanish and is measured by every probe. Column i of $R = DB^\top$, the response to a unit dose along b_i, is estimated by the odd part of a probe, which the even part does not affect.

Proof of Proposition 1. Under $B \to AB$, a perturbation keeps its value when $a \to A^{-\top}a$, while $R \to RA^\top$ and $W \to AWA^\top$. Hence $RA^\top(AWA^\top)^{-1}AR^\top = RW^{-1}R^\top$, and K is unchanged. Substituting $R = DB^\top$ gives $K = D\,B^\top(BB^\top)^{-1}B\,D^\top = D\,\Pi\,D^\top$ with $\Pi = B^\top W^{-1}B$. Π is symmetric, idempotent, with range the row space of B: the orthogonal projector onto the span of the directions. A perturbation $v = B^\top a$ produces Ra; we minimise $a^\top W a$ subject to $Ra = t$. Stationarity of the Lagrangian gives $Wa = R^\top\lambda$, thus $a = W^{-1}R^\top\lambda$, and the constraint gives $K\lambda = t$. For invertible K, $a^\star = W^{-1}R^\top K^{-1}t$ and $a^{\star\top}Wa^\star = t^\top K^{-1}RW^{-1}R^\top K^{-1}t = t^\top K^{-1}t$, the minimum by strict convexity on the affine set. For $m = 6$ and invertible R, $a^\star = R^{-1}t$ is unique, is the plan of Eq. 5, and its size $t^\top K^{-1}t$ is basis-independent. For $m < 6$ it is a genuine minimum over the perturbations producing t. A barely readable structure makes R ill conditioned and $t^\top K^{-1}t$ large.

Proof of Proposition 2. Write $s = r - Ra$ and $E = R^+ - R = s\,(Wa)^\top/(a^\top Wa)$. (i) $R^+a = Ra + s\,(a^\top Wa)/(a^\top Wa) = r$. (ii) Any E' with $(R + E')a = r$ satisfies $E'a = s$. With $F = E'W^{-1/2}$ and $a' = W^{1/2}a$, the condition reads $Fa' = s$. With $P = a'a'^\top/(a'^\top a')$, every solution splits as $F = FP + F(I - P)$, where $FP = s\,a'^\top/(a'^\top a')$ is fixed by the condition. The two terms are Frobenius-orthogonal, and the norm is minimal at $F(I - P) = 0$, giving $E' = FW^{1/2} = s\,(Wa)^\top/(a^\top Wa) = E$. (iii) $Eb = s\,(a^\top Wb)/(a^\top Wa) = 0$ when $b^\top Wa = 0$. (iv) Under $B \to AB$, the same perturbation has dose $\tilde{a} = A^{-\top}a$, with $\tilde{R} = RA^\top$ and $\tilde{W} = AWA^\top$. Hence $\tilde{R}\tilde{a} = Ra$, $\tilde{W}\tilde{a} = AWa$ and $\tilde{a}^\top\tilde{W}\tilde{a} = a^\top Wa$; the updated model is $RA^\top + s\,(Wa)^\top A^\top/(a^\top Wa) = R^+A^\top$, the same map from perturbations to readings. The Euclidean update instead adds $s\,\tilde{a}^\top/(\tilde{a}^\top\tilde{a}) = s\,a^\top A^{-1}/(a^\top A^{-1}A^{-\top}a)$, which equals $s\,a^\top A^\top/(a^\top a)$ for every a only if A^{-1} is a multiple of $A^\top$, that is, for A orthogonal up to a scalar. The trust region of the closed loop, a Euclidean bound on $\|a - a_{\text{prev}}\|$, is its only basis-dependent step. (v) Let the plan solve $Ra = t$ and the true response be $r = g\,Ra$, $g \neq 0$. Then $r = g\,t$ and $r/r_k = t/t_k$: the leak of the reading equals the leak of the request, which satisfies the leak clause, and rescaling the dose by $t_k/r_k = 1/g$ gives $r = t$. If instead $r = G\,Ra$ with G diagonal and $t = t_k e_k$, then $r = G_{kk}\,t_k e_k$: the reading has no leak, and rescaling the dose by $1/G_{kk}$ gives $r = t$. Where the response is linear, $\Delta = \beta\,Du$, and no ratio of two components of Δ contains β (Observation 1).

Clipped protected components. In Eq. 4 the protected components of $s^{(i)}$ keep the surrogate's cross-response up to $\pm\rho$ rather than being set to zero. On three BrLP noises, the surrogate predicts that setting them to zero requires 1.01 to 1.13 times the dose of the raw gradient for one floor, and 1.00 to 1.05 times when protected changes up to 0.5 floors are allowed; the surrogate's median held-out error on BrLP is 0.82% of the volume, against floors of 0.61 to 3.07%. In the ablation of Table 22, under its absolute criterion, $\rho = 0$ satisfies 71 of 96 requests and $\rho = 0.25$ satisfies 77, and on WaveDiT $\rho = 0$ satisfies 21 of 24 against 24 for the raw gradient.

Empirical basis invariance. K changes by at most 11% between the six resolved directions and the six raw surrogate gradients, whereas the median leak of the six directions changes by a factor

of 0.8 to 3 (equal dose, same noises). K differs by at most 1.2% on MONAI, BrLP and MedForj and by 3 to 6% on WaveDiT; the 11% occurs on one starting noise of MAISI. The difference is the relative Frobenius norm $\|K_{\mathrm{res}} - K_{\mathrm{raw}}\|_F / \|K_{\mathrm{res}}\|_F$ of the 6×6 matrix of the planned structures, computed per starting noise on two noises per generator and reported as the maximum over the two, with each R estimated from one probe per direction, at the same dose in the two sets, and each W from its own directions. The two sets span the same subspace, so by Proposition 1 the two matrices coincide where the response is linear.

Constrained planning. Planning the fifteen labels outside the plan as well lowers their leak on four generators, at the cost of one to two probes on average and of a lower rate under contract-6. Under contract-21 the closed-loop plan solves the linear program $\min_{c,\, s} s$ subject to $(Rc)_k = 1$, $|(Rc)_j| \le \rho$ on the five other planned structures and $|(Rc)_j| \le s$ on the fifteen outside structures, with a 21-row R. A sequential quadratic program started from this solution replaces it when it lowers s. The model is updated with the Euclidean rule $(r - Ra)\, a^\top / (a^\top a)$, whereas the contract-6 results use the update in the metric W of Eq. 5. On the first four noises, the paired variant planned under contract-21 lowers the outside leak by 28 to 43% on four generators, to 0.07 to 0.15; on MAISI the outside leak is 0.23 under contract-21 and 0.21 without it (Table 37). It satisfies contract-21 on the paired reading in 21% (MAISI) to 83% (MONAI) of the requests, and on the positive arm in 12% to 50%, against 25% to 58% with the plain plan.

Appendix C.8 reports the response across sampling windows (Fig. 6).

Dose-response linearity. Up to 5 floors the response is linear in the dose: the median prediction error is 0.7 to 3.3% on the five generators, and the gain is constant but differs from one (Fig. 3), 0.83 on MONAI, 0.54 on BrLP, 1.01 on MAISI, 0.92 on WaveDiT and 0.35 on MedForj. Six noises per generator are perturbed in open loop along the resolved target direction at 0.5, 1, 2 and 5 floors (864 pairs; 10 and 20 floors on MONAI and BrLP), and a line fitted on the two lowest doses predicts the target change at 2 and 5 floors. On MONAI, BrLP, WaveDiT and MedForj the median error is 0.7 to 2.5% with 35 or 36 of 36 requests per dose within 15%; on MAISI it is 1.0 and 3.3%, with 33 and 29 of 36. Beyond, the response bends: at 20 floors the gain is 0.74 against 0.83 on MONAI and 0.53 against 0.54 on BrLP. Nine requests lie off the scale of the bottom panel, all on MAISI: the hippocampus on every noise, at 9 to 10 times the request, and the thalamus on three noises, at 2.2 to 2.9 times.

Linearity across combined directions. On the 78 held-out requests, the six columns measured one direction at a time on the noise of the request predict the odd response of every combined perturbation probed by the closed loop with a median relative error of 3.5 to 9.2% over the six structures and 0.8 to 2.1% on the target, against 30 to 40% when only the target column is used. On the 600 requests of the 100 held-out noises, the model of the loop predicts each probe before it is run with a median error of 19 to 45% at the first probe, planned with $\bar{R}$, and 3.9 to 18% from the second probe on, where successive plans stay within a median W-norm ratio of 0.96 to 1.15 of the first. The columns are measured near one floor, and this test concerns the combination of directions, not larger doses. On the same dose scale, the odd response of the five protected structures is also linear in the dose: a line through the origin over 0.5 to 5 floors gives a median R^2 of 0.997, 0.996, 0.952, 0.981 and 0.993 on MONAI, BrLP, MAISI, WaveDiT and MedForj (0.9994 to 0.9999 on the target), and the median leak stays between 0.17 and 0.23 from 0.5 floors to the highest dose on all generators except MAISI, where it rises from 0.24 to 0.32. The even part on the target grows about linearly with the dose (median per-request log-log slope 0.83 to 1.18) and stays at 1 to 3.5% of the odd part. The plan of Eq. 5 uses the even part of the previous probe without rescaling it to the new dose, and successive plans stay within a median W-norm ratio of 0.96 to 1.15 of the first (above). On the thirteen evaluation noises, the even part of a placebo pair, along a unit direction orthogonal to the six gradients at the median one-floor dose, has a median absolute value over the six structures of 0.002, 0.005, 0.009, 0.007 and 0.0001 floors on MONAI, BrLP, MAISI, WaveDiT and MedForj, against 0.005, 0.006, 0.019, 0.011 and 0.002 floors for the open-loop pair at one floor on the same noises.

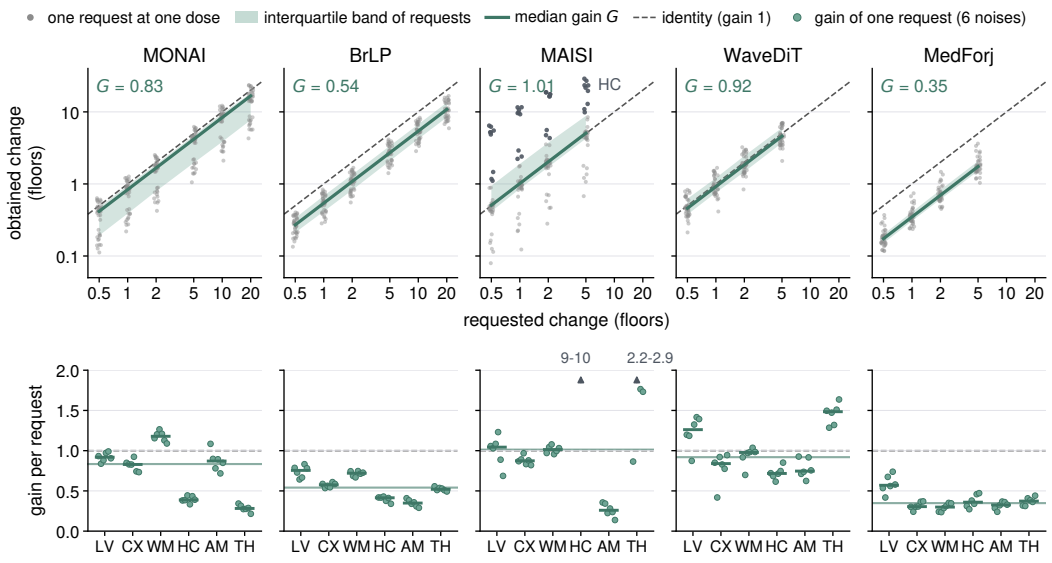


Figure 3: Dose ladder on five generators, six noises each. Top, obtained against requested change on the target: grey points are requests, the green band their interquartile range, the solid line the median gain G and the dashed line the identity; dark points mark the nine MAISI requests off the bottom scale (hippocampus, gain 9 to 10; thalamus, 2.2 to 2.9). Bottom, gain per request by structure; arrows mark gains above 2.

B Data, verifier and settings

In this section, we describe how every volume is measured and in which unit. Since a change of one floor corresponds to about one percent of a regional volume and the five generators differ in grid, field of view and skull, all volumes are measured with a single pipeline.

All five generators are measured by the same pipeline. Each volume is saved as NIfTI with its generator's grid and affine and passed to the verifier unmodified: no skull-stripping, registration, resampling, intensity normalisation or cropping. SynthSeg's internal preprocessing is common to all: it resamples to 1 mm only when an axis departs from that spacing (BrLP, at 1.5 mm), aligns to the identity affine, rescales intensities on the 0.5 and 99.5 percentiles, and pads each axis to the next multiple of 32, with a minimum of 128. WaveDiT volumes carry the affine of the floor corpus. BrLP and MedForj outputs are unclamped, as in their released decoders. FastSurfer reads the same files.

B.1 Cohorts and measurement units

In this subsection, we describe the real scans on which the unit of every change, the disease profile and the disease classifier are measured.

Test-retest floors. The floor of a structure, the unit of every change, is the mean absolute relative difference between two same-session readings over 92 OASIS-3 pairs (LaMontagne et al., 2019), computed on raw volumes (Table 2). One pair per person gives 100 pairs from 961 groups over 667 people, and 8 are discarded because the verifier failed on at least one scan. These scans undergo bias-field correction, nine-parameter registration to MNI152 at 1 mm, skull-stripping and intensity normalisation; generated volumes do not. On MAISI, removing the skull changes the FastSurfer readings by a factor of 0.94 to 1.04. The fifteen structures outside the plan use SynthSeg floors from the same pairs, also under FastSurfer. The floor is the mean absolute difference that the verifier pipeline, acquisition, preprocessing and segmentation together, reads between two scans of the same person; it is not the smallest change detectable in one person. The generated volumes are deterministic given the starting noise, and on them the floor is a unit per structure.

	test-retest floor (% of the structure's volume)					
verifier	lat. ventricles	cortex	white matter	hippocampus	amygdala	thalamus
SynthSeg	0.607	0.933	0.685	1.194	3.067	1.419
FastSurfer	0.546	1.280	1.167	0.946	1.918	1.020

Table 2: **Test-retest floors of the two verifiers.** Mean absolute difference on 92 same-session pairs, in percent of the structure's volume.

Alzheimer's profile and classifier. The profile request measures 51.6 floors, dominated by the ventricles; this length is one patient's dose. It is the patient-control difference in the six planned volumes on BrainLat (Prado et al., 2023), read by the same verifier. We draw 48 subjects with a fixed random seed from scans that pass a quality check on the background fraction, 12 patients and 12 controls from each of two centres, CLB and COB. Of these, 46 are valid; the two excluded have 19 and 20 of 31 structures at zero. Each volume is divided by the total segmented volume, corrected for age with a linear model fitted on the controls, and expressed in raw-volume floors, as are the steered volumes.

We train a disease classifier that the controller never calls. It is a logistic regression with ℓ_2 penalty, $C = 1$ and no class weighting, on the same six volumes, each divided by the total segmented volume without age correction, and standardised. We train it on 487 scans of 145 people (87 scans of patients, 400 of controls) from the Alzheimer's Disease Neuroimaging Initiative (ADNI) (Jack et al., 2008) and OASIS-3. On 136 BrainLat subjects (98 patients, 38 controls) from centres not used by the profile, it reaches an AUC of 0.892 (bootstrap interval 0.832 to 0.943), and 0.756 on the only centre with both classes and at least ten subjects. On every generator it reads through SynthSeg the six quantities the controller steers. It classifies the BrainLat profile as disease, and it exposes the converse case, in which classifier guidance raises P(AD) while the volumes move against the profile.

B.2 Settings

In this subsection, we report the settings, which are identical across generators and therefore involve no per-generator tuning.

The controller settings are the same on every generator, and each has its own surrogate, trained on its free samples and calibrated, with held-out error from 0.4 to 11% per structure. On WaveDiT, the baselines and TFG use a second surrogate, trained on the same samples with a different preprocessing order. The baselines select their surrogate by one rule on every generator, the lowest median held-out error among the surrogates whose calibration slopes all reach 0.30, whereas the controller uses the surrogate that crops the volume before z-scoring its intensities, the order of nnU-Net. On MONAI, BrLP and MedForj the rule selects this surrogate; on WaveDiT it selects the surrogate that z-scores before cropping, with a median error of 1.06% against 1.14%. On WaveDiT, the baselines therefore use the surrogate with the lower median held-out error. Table 3 lists the settings and the exceptions.

The five samplers are deterministic given the starting noise and add no noise at the steps: DDIM with $\eta = 0$ on MONAI, BrLP and MedForj, Euler steps of the rectified flow on MAISI and Heun steps of the flow ODE on WaveDiT. The free sample and both arms of every probe therefore start from the same noise, drawn on the CPU from a fixed seed per noise, and share every step except the added perturbation. Two generations of the free sample in the same process are checked to be identical on every noise. BrLP runs its denoiser under half-precision autocast, as in the authors' sampling script, MAISI under half-precision and WaveDiT under bfloat16 autocast, and MONAI and MedForj in single precision, where the released scripts use half precision for the MONAI decoder and for MedForj.

The sensitivity to ρ in the directions of Eq. 4 is in Table 22 (Appendix C.8).

Setup cost. Setup per generator consists of 160, 160, 160, 240 and 718 free samples labelled by the verifier (MONAI, BrLP, MAISI, WaveDiT, MedForj), 0.42 to 2.88 hours of surrogate training (6.2 hours in total), and 36, 36, 12, 18 and 12 paired probes for the population model. Generating

Contract		
leak cap ρ	0.125, accepted up to 0.1375 (tolerance of 10% on the cap)	success
target	$\lvert\Delta_k - t_k\rvert \le 1/4$ floor, with $t_k = 1$ on every single-structure request, that is, within 25% of the request	success
profile tolerance κ	0.25 of each component, or one floor (third condition of Eq. 3)	success
cap outside the plan ρ_{out}	0.25 on the fifteen non-ventricular labels outside the plan (contract-21)	success
Closed loop		
plan	$a = R^{-1}(t - e)$	method
probes	at most 5, after which the procedure terminates without a returned volume	cost
trust region	$\lVert a - a_{\mathrm{prev}}\rVert \le \frac{1}{2}\lVert a_{\mathrm{prev}}\rVert$, in the Euclidean norm of the dose	cost
largest perturbation	3 times the first	cost
update	Broyden, in the metric W; Euclidean under contract-21	method
Paired variant		
plan	$Ra = t$	method
probes	4, extended to 6 while the leak decreases	cost
trust region	1 in the metric W (the dose vector changes by at most its own size per probe)	cost
abstention	when $\lvert e_k\rvert > \lvert r_k\rvert/4$, or when the target response is below 3 times the placebo response (clipped to 0.05–0.5 floors)	empirical
update	Broyden, in the metric W, skipped when the residual is below the placebo noise	method
Sampler		
steps	100 on every generator	cost
check	$\beta = 0$ reproduces the released sampler exactly (difference $0.000\mathrm{e}{+00}$)	measurement
Surrogate $\hat{J}$		
network	3D U-Net, 3 levels, 16 to 128 channels, GroupNorm and SiLU, 17 classes, 1.46 M parameters	empirical
training	patches of 96^3, one per step, cross-entropy, AdamW (learning rate 3×10^{-4}, weight decay 10^{-4}), cosine schedule, 40 epochs, no augmentation	empirical
training data	free samples labelled by J, 1438 in total over the five generators, a quarter held out	empirical
calibration	per-structure affine fit on the training split (BrLP hippocampus baseline-volume error 122% without it, 3% with it)	measurement
unreadable structure	slope below 0.30: MAISI hippocampus (0.05). Its requests stay in the evaluation. In the profile it is one of six components moved jointly, with a request of 5 to 9 floors	empirical
Baselines		
step, a posteriori	per request, the step size that produces one floor, interpolated on a grid of step sizes; extrapolated requests are kept	method
step, frozen	per structure, the step at which a line through the origin, fitted to the target change on two of the first four noises, predicts one floor; applied unchanged to the 78 held-out requests	method
TFG	12 configurations (c0 to c11) screened on the noises of the population model; retained c5 (mean guidance, four iterations, error 0.49) and c6 (the same with variance guidance, 0.66); recurrent configurations excluded, because their arms leave the paired trajectory	empirical
Compute		
hardware	NVIDIA H100 with 94 GB; verifier on CPU	reported
baseline pair	157 s (WaveDiT) to 2757 s (MAISI, 256^3) for one paired generation of a gradient baseline; TFG 589 to 1089 s	reported
request	closed loop 58 to 310 s; paired variant 224 to 616 s; medians including verifier calls (Table 28)	reported

Table 3: **Experimental settings.** Controller settings are shared; surrogates are per generator. Role: success, a setting that defines when a request is satisfied; method, part of the procedure; cost, a setting that bounds computation and does not change the definition of success; measurement, a check of the instrument; empirical, a choice made on noises outside the evaluation; reported, a measured quantity.

the free samples required 8.0 GPU hours on MedForj and an estimated 0.13 to 0.39 hours on each of the other generators.

Cohorts. Table 4 lists every cohort of starting noises, with counts per generator. The closed loop runs with the same code, settings and population model on the 78 held-out requests, the replication and the 100 noises, and no setting of the closed loop was chosen on the 100-noise cohort.

cohort	noises	requests	methods	used in
population model	6, 6, 2, 3, 2	columns of $\bar{R}$	probes along each direction; TFG screening	Sec. 3.6, Table 3
first four noises	4	24	paired variant, open loop, baselines a posteriori and at a step fitted on two of the four, mask, penalised objectives	Tables 37, 25, 8, 7, 33, 26
held-out	13	78	closed loop, paired variant, block controller, one-sided loop, requests of opposite sign, frozen-step baselines, penalised FlowChef, mask; FastSurfer re-read of the closed loop	Tables 1b, 24, 10, 9, 31, 27, 28, 13
paired variant	17	102	paired variant, on the first four noises and the held-out noises	Tables 1, 5, 27
replication	13	78	closed loop	Table 5
100 noises	100	600	closed loop	Tables 1b, 5, 19
FlowChef	25 of the 100	150	closed loop, frozen-step FlowChef, on MONAI, BrLP and WaveDiT; FastSurfer re-read of the closed loop	Table 24
dose ladder	6	36 per dose	open loop at 0.5 to 5 floors, 10 and 20 on MONAI and BrLP	Fig. 3
five floors	2 of the first four	12	open loop, paired variant	Table 7
MONAI, two verifiers	6	35	open loop, paired variant, read by FastSurfer and SynthSeg	Appendix C.3
profile, fractions	first four	16 (4 fractions)	paired variant stopped on the paired reading, baselines a posteriori, classifier guidance with 2 to 5 step sizes	Tables 16, 38, 30, 32, Fig. 9
profile, comparison	12	24 (2 doses)	closed loop, paired stop, BrLP conditioning, frozen-step baselines	Tables 29, 32, 21, Fig. 8
profile, replication	12	24 (2 doses)	closed loop; six and twelve directions on BrLP and WaveDiT	Tables 21, 20
profile, 100 noises	100	200 (2 doses)	closed loop; on 25 of them, frozen-step FlowChef on BrLP and MONAI and the one-sided loop	Tables 15, 19

Table 4: **Cohorts of starting noises**, per generator. Requests are single-structure requests of one floor unless stated; profile doses are $f = 0.5$ and 1.

C Analysis of the closed loop

In this section, we study our procedure: the reliability of the reported rates, the contribution of each component, the dependence on the verifier, profile requests, the anatomical plausibility and the independent classification of the returned volumes. Results on the first four starting noises, the earlier cohort, are marked as such.

C.1 Reliability of the reported rates

In this subsection, we report the uncertainty of the coverage rates, which are measured on a finite number of noises with correlated requests within each noise, and repeat the evaluation on further noises.

Uncertainty estimates. Table 5 gives 95% bootstrap intervals for the paired variant and the closed loop, resampled over starting noises because requests that share a noise are correlated. The median leak is resampled over requests and computed over the requests whose target moves in the requested direction. A closed-loop request that terminates without a returned volume is read at its last probe. For the paired variant, only requests with a final pair enter.

Replication on further noises. We evaluate the unchanged closed loop on two further sets of noises, disjoint from those of the 78 held-out requests (Table 5). On thirteen noises, it satisfies the contract in 78 of 78 requests on MONAI, BrLP and WaveDiT, 73 on MedForj and 61 on MAISI. On 100 noises, it satisfies 596, 600, 502, 593 and 574 of 600 requests on MONAI, BrLP, MAISI, WaveDiT and MedForj, at 1.45 to 2.87 pairs per request and median leaks of 0.046 to 0.092. On every generator, the coverage stays within 9 points of that on the 78 requests, and the largest drop is on MAISI. All requests that are not served terminate without a returned volume. The closed loop plans only the six structures; on the fifteen non-ventricular labels outside the plan, its median leak over the 600 requests, read as in Table 5, is 0.20, 0.26, 0.38, 0.18 and 0.21 on MONAI, BrLP,

generator	method	noises	requests	contract (%) paired reading ↑	contract (%) returned volume ↑	pairs ↓	median leak ↓
MONAI	paired variant	17	102	93 [87,98]	71 [61,79]	2.82	0.124 [0.122,0.125]
	closed loop	13	78	92 [86,97]	99 [96,100]	1.54	0.050 [0.043,0.068]
	closed loop	13	78	95 [88,100]	100 [75,100]†	1.38	0.049 [0.037,0.068]
	closed loop	100	600	96 [94,97]	99 [98,100]	1.51	0.059 [0.054,0.062]
BrLP	paired variant	17	102	93 [87,98]	67 [58,75]	2.83	0.125 [0.123,0.127]
	closed loop	13	78	92 [86,97]	100 [75,100]†	1.46	0.047 [0.039,0.053]
	closed loop	13	78	100 [75,100]†	100 [75,100]†	1.62	0.040 [0.034,0.048]
	closed loop	100	600	97 [96,98]	100 [96,100]†	1.45	0.046 [0.042,0.049]
MAISI	paired variant	17	102	61 [52,70]	48 [40,56]	3.36	0.124 [0.113,0.130]
	closed loop	13	78	64 [53,74]	87 [73,97]	2.96	0.081 [0.070,0.099]
	closed loop	13	78	68 [50,86]	78 [60,94]	3.01	0.094 [0.079,0.108]
	closed loop	100	600	64 [58,70]	84 [79,88]	2.87	0.092 [0.085,0.096]
WaveDiT	paired variant	17	102	84 [76,91]	64 [53,75]	2.54	0.102 [0.096,0.118]
	closed loop	13	78	92 [79,100]	100 [75,100]†	1.65	0.052 [0.046,0.056]
	closed loop	13	78	94 [87,99]	100 [75,100]†	1.77	0.053 [0.040,0.073]
	closed loop	100	600	92 [88,94]	99 [97,100]	1.62	0.057 [0.053,0.060]
MedForj	paired variant	17	102	94 [90,98]	84 [75,93]	2.86	0.105 [0.098,0.119]
	closed loop	13	78	86 [71,97]	87 [73,97]	2.17	0.074 [0.052,0.081]
	closed loop	13	78	88 [79,96]	94 [85,100]	1.94	0.061 [0.047,0.077]
	closed loop	100	600	92 [89,94]	96 [94,98]	1.79	0.053 [0.048,0.059]

Table 5: **Uncertainty of the paired variant and the closed loop.** Noises: starting noises evaluated. The paired variant regenerates a final pair after its probes. The closed loop stops on the returned volume; its paired reading comes from its last probe, and its three sets of noises are disjoint. The contract is read on the paired reading and on the returned volume, the positive arm; requests without a returned volume count as failures. Pairs: generated pairs per request. Brackets: 95% bootstrap interval, over noises for the percentages and over requests for the median leak. Where every request is satisfied, the bootstrap interval is a point; the bracket marked † gives instead the exact 95% Clopper-Pearson lower bound on the fraction of noises with all requests satisfied, which bounds the request rate from below. Arrows mark the better direction.

MAISI, WaveDiT and MedForj, against 0.213, 0.258, 0.208, 0.109 and 0.203 for the paired variant on the 24-request cohort (Table 37): close on MONAI, BrLP and MedForj, higher on MAISI and WaveDiT.

Requests of opposite sign. With requests of one floor down on the same 13 noises and six structures, and the mirrored contract, the closed loop satisfies 77, 78, 66, 78 and 69 of 78 requests on MONAI, BrLP, MAISI, WaveDiT and MedForj, against 77, 78, 68, 78 and 68 for requests of one floor up (Table 6). The two signs disagree on 2, 0, 10, 0 and 1 paired requests (exact McNemar $p \geq 0.75$), and the median leak and the mean number of pairs differ by at most 0.02 and 0.07. All failures are abstentions. The first plan is linear in the request, so the first probe of a request of one floor down uses the pair of trajectories of the first probe of the opposite request with the arms swapped; the two runs differ from the second probe on.

	requests satisfied ↑		paired requests					median leak ↓		mean pairs ↓		abstentions ↓	
generator	$+1$	-1	both	$+1$ only	-1 only	neither	McNemar p	$+1$	-1	$+1$	-1	$+1$	-1
MONAI	77/78	77/78	76	1	1	0	1.00	0.050	0.052	1.54	1.59	1	1
BrLP	78/78	78/78	78	0	0	0	1.00	0.047	0.042	1.46	1.53	0	0
MAISI	68/78	66/78	62	6	4	6	0.75	0.081	0.096	2.96	2.95	10	12
WaveDiT	78/78	78/78	78	0	0	0	1.00	0.052	0.056	1.65	1.64	0	0
MedForj	68/78	69/78	68	0	1	9	1.00	0.074	0.056	2.17	2.13	10	9

Table 6: **Requests of opposite sign on the same noises**, closed loop on the returned volume, 78 requests per generator on thirteen noises. $+1$: one floor up; -1: one floor down, with the mirrored contract. Paired requests compare the two signs on the same noise and structure; McNemar is exact on the discordant requests. Abstentions count as failures.

Abstentions of the closed loop. Abstentions are 0, 4, 7, 26 and 98 of 600 requests on BrLP, MONAI, WaveDiT, MedForj and MAISI. On MAISI, the 98 abstentions are 3 on the lateral ventri-

cles, 21 on the cortex, 28 on the white matter, 15 on the hippocampus, 16 on the amygdala and 15 on the thalamus. The minimal perturbation size $t^\top K^{-1}t$, known before generation, separates abstained from accepted requests with an AUC of 0.62, 0.50, 0.86 and 0.64 on MONAI, WaveDiT, MedForj and MAISI (0.66 pooled on ranks within each generator), and the abstention rate rises from 1.7% in its lowest quartile to 8.3% in its highest; the held-out error of the surrogate per structure does not separate them (AUC 0.51).

Sensitivity to the contract cap. Rescored a posteriori on the recorded probes of the 600 requests per generator, with the directions unchanged and the loop stopped at the first recorded probe that satisfies the wider contract, coverage is 99.7, 100, 94.8, 98.8 and 98.0% at a cap of 0.25 and 99.8, 100, 97.3, 98.8 and 99.7% at 0.5 on MONAI, BrLP, MAISI, WaveDiT and MedForj, with 1.11 to 2.11 probes per request. At 0.0625, 58.7, 73.3, 33.3, 62.7 and 61.7% of the 600 requests also satisfy the tighter cap, a lower bound for a loop run at that cap.

Dependence on the surrogate. Across the 30 pairs of generator and structure, coverage on the 600 requests per generator is not rank-correlated with the held-out error of the surrogate (Spearman -0.06, $p = 0.76$) or with its calibration slope (-0.02, $p = 0.92$). The MAISI hippocampus, with slope 0.05, reaches 85 of 100 requests, and the lowest coverage, 72 and 79 on MAISI white matter and cortex, occurs where the surrogate error is below 1%.

Population model. $\bar{R}$ is the median, over 6, 6, 2, 3 and 2 noises for MONAI, BrLP, MAISI, WaveDiT and MedForj, of one verifier-read pair along each direction (36, 36, 12, 18 and 12 pairs), on noises disjoint from all evaluation cohorts. Without $\bar{R}$, the block controller, which measures the six columns on the noise of the request, satisfies the contract on the returned volume in 88 to 100% of the 78 held-out requests on four generators and 63% on MAISI, at seven pairs per request.

Probe budget and request magnitude. The paired variant reaches its final rate within four probes on every generator, and at a request of five floors it produces 4.85 to 5.03 floors, where the open loop produces 1.46 to 5.15 (Table 7). Planning under contract-21 is described in Appendix A.

	contract-6 (odd part) within k probes (%)							five floors			
generator	k=1	k=2	k=3	k=4	k=5	k=6	probes ↓	open	paired	leak ↓	accepted (%)
MONAI	38	96	100	100	100	100	1.67	3.66	5.03	0.124	100
BrLP	38	88	88	92	92	92	1.88	2.74	4.92	0.124	100
MAISI	25	42	58	62	62	62	2.79	5.15	4.97	0.113	75
WaveDiT	71	83	96	96	96	96	1.50	4.33	4.85	0.094	92
MedForj	46	88	92	100	100	100	1.75	1.46	4.91	0.120	100

Table 7: **Probe budget and five-floor requests**, paired variant. Left: percentage of the 24 requests of the first four noises in which a probe has satisfied the paired contract within k probes and mean probes used. This is the verdict of the probe that stops the loop, not that of the regenerated pair reported in Table 1; the rates at $k = 6$, on 24 requests, therefore differ from the paired rates there, on 102 requests. Right: at a request of five floors, target change of the open loop and of the paired variant, leak and share of requests accepted, that is, not abstained, after the probe satisfies the leak clause (twelve requests on two noises); the regenerated pair is not tested again against the contract.

The unpaired sample sizes needed to detect a one-floor change are in Appendix C.8 (Table 23).

C.2 Contribution of each component

In this subsection, we add the components of the closed loop one at a time to identify which ones produce the coverage, and compare paired probes with probes of one generation.

Component ablation. With components added one at a time, the paired variant with rank-one updates has the highest contract rate on every generator (Table 8). Intermediate columns are read on a probe rather than on a regenerated pair; by Observation 1, the leak of a probe does not depend on its dose where the response is linear.

generator	requests satisfying contract-6 (%)				
	raw gradient + measured dose	open loop	+ measured dose	+ population model	+ Broyden (paired variant)
MONAI	17 [5, 45]	17 [7, 36]	25 [12, 45]	<u>38</u> [21, 57]	**100** [86, 100]
BrLP	17 [5, 45]	0 [0, 14]	33 [18, 53]	<u>38</u> [21, 57]	**83** [64, 93]
MAISI	25 [9, 53]	8 [2, 26]	<u>33</u> [18, 53]	25 [12, 45]	**46** [28, 65]
WaveDiT	25 [9, 53]	8 [2, 26]	46 [28, 65]	<u>71</u> [51, 85]	**83** [64, 93]
MedForj	17 [5, 45]	0 [0, 14]	29 [15, 49]	<u>46</u> [28, 65]	**100** [86, 100]

Table 8: **Component ablation**: percentage of requests satisfying contract-6, parts added left to right. Raw gradient: guidance with only our dose feedback, on 12 requests from two further noises. Brackets: 95% Wilson interval. Bold: best; underlined: second best.

Component ablation on the returned volume. Only the loop with updates exceeds 54% on every generator, with 87 to 100%, against 0 to 17% for the open loop and 6 to 54% for the first pair planned with the population model (Table 9). All columns read the positive arm of a generated volume.

generator	returned volume satisfying the contract (%)		
	open loop	+ population plan, one pair	+ updates, up to five pairs
MONAI	17 [17,17]	<u>50</u> [40,59]	**99** [96,100]
BrLP	6 [3,12]	<u>54</u> [42,64]	**100** [100,100]
MAISI	<u>10</u> [5,15]	6 [1,13]	**87** [73,97]
WaveDiT	8 [4,12]	<u>47</u> [38,55]	**100** [100,100]
MedForj	0 [0,0]	<u>33</u> [24,42]	**87** [73,97]

Table 9: **Component ablation on the returned volume**: percentage of 78 requests whose returned volume, a generated sample, satisfies the contract. Brackets: 95% bootstrap interval over starting noises. Bold: best; underlined: second best.

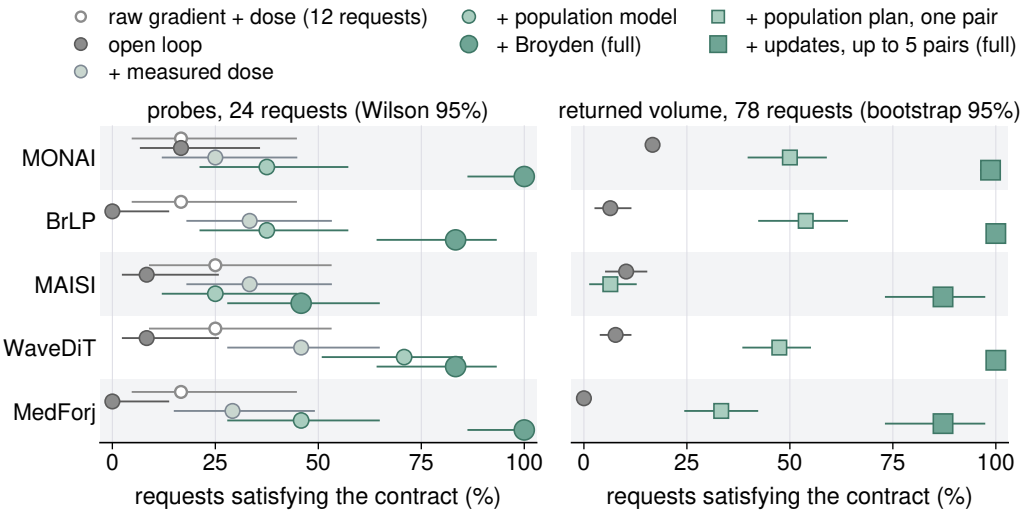


Figure 4: **Component ablation.** Left: requests satisfying the contract, read on probes (24 requests; raw gradient 12, hollow), as components are added. Right: returned volume on 78 requests. Bars: 95% Wilson interval (left), bootstrap over noises (right).

Stopping rules and response estimation. On MONAI, BrLP and WaveDiT, the closed loop covers 77 to 78 of 78 requests at 1.46 to 1.65 pairs, against 64 to 71% at 2.6 to 2.9 pairs for the paired variant, read on its regenerated positive arm (McNemar $p < 10^{-6}$, Table 10). On MAISI it covers

87% against 51% ($p < 10^{-5}$), and on MedForj 87% against 83%, not significant. The closed loop plans for the returned volume, $R\,a = t - e$, accepts the first probe whose positive arm satisfies the contract, regenerates no pair and stops without a volume after five pairs. The block controller, with equal verifier access, measures the six columns of R on the request noise, plans once and generates one pair. On MONAI, BrLP, WaveDiT and MedForj, its paired rate exceeds that of the rank-one update of the paired variant by 9, 4, 13 and 6 points, at seven pairs per request (2.0 amortised over the six requests of a noise, Table 10) instead of 1.5 to 3 for the closed loop, and its returned volume satisfies the contract in 88 to 100%. On MAISI, its paired rate is 81% against 65%, but its returned volume reaches only 63%, below the 87% of the closed loop.

generator	controller	requests	coverage (%)	paired (%)	pairs ↓
MONAI	closed loop (ours)	78	99 [96,100]	92 [86,97]	**1.54**
MONAI	paired variant	78	68 [55,79]	91 [83,97]	2.87
MONAI	block estimate, no updates	78	**100** [100,100]	**100** [100,100]	2.00
BrLP	closed loop (ours)	78	**100** [100,100]	92 [86,97]	**1.46**
BrLP	paired variant	78	71 [63,78]	96 [90,100]	2.85
BrLP	block estimate, no updates	78	99 [96,100]	**100** [100,100]	2.00
MAISI	closed loop (ours)	78	**87** [73,97]	64 [53,74]	2.96
MAISI	paired variant	78	51 [42,60]	65 [56,74]	3.35
MAISI	block estimate, no updates	78	63 [51,74]	**81** [73,88]	**2.00**
WaveDiT	closed loop (ours)	78	**100** [100,100]	92 [79,100]	**1.65**
WaveDiT	paired variant	78	64 [50,78]	85 [76,94]	2.56
WaveDiT	block estimate, no updates	78	88 [77,97]	**97** [94,100]	2.00
MedForj	closed loop (ours)	78	87 [73,97]	86 [71,97]	2.17
MedForj	paired variant	78	83 [72,94]	92 [88,96]	2.90
MedForj	block estimate, no updates	78	**97** [94,100]	**99** [96,100]	**2.00**

Table 10: **Stopping rules and response estimation** on the requests of the closed loop. Brackets: 95% bootstrap interval over starting noises. Block-controller pairs are amortised over the six requests of a noise. New noises: Table 5. Bold: best; underlined: second best.

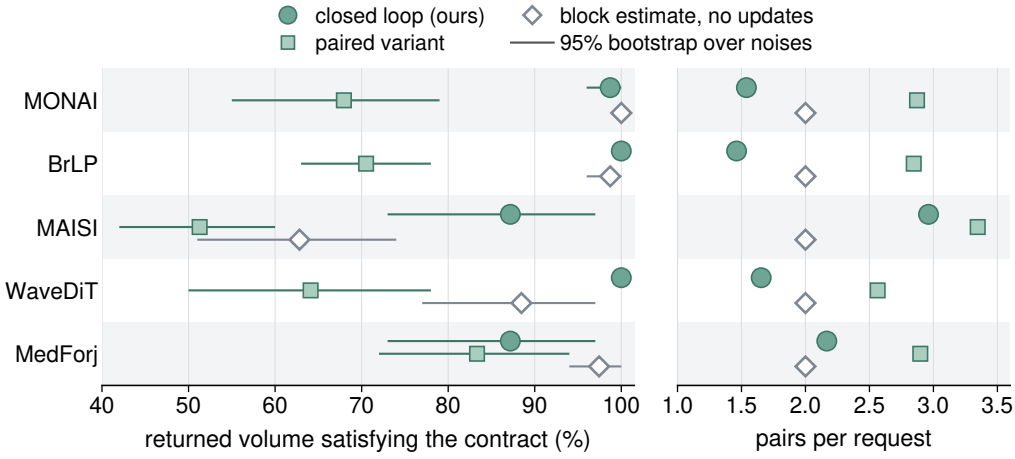


Figure 5: **Stopping rules**: returned-volume coverage against pairs generated per request, per controller. Bars: 95% bootstrap interval over starting noises.

One-sided probes. The closed loop tests the positive arm against the free sample, and a loop that generates only this arm satisfies the same requests with half the generations. This one-sided loop generates x_+ at each probe, reads the total change $J(x_+) - J(x_0)$, updates R by Broyden on that change and plans $a = R^{-1}t$. On the 78 held-out requests, it satisfies the same requests as the paired closed loop on MONAI, BrLP, WaveDiT and MedForj, with no discordant request, at 1.49 to 2.17 generations per request against 2.92 to 4.33 (Table 11); at one floor, the even part that the paired

loop subtracts has a median of 0.03 to 0.045 floors per probe, below the target tolerance of 0.25. On the AD profile, over the first 25 of the 100 held-out noises at $f = 0.5$ and 1, it satisfies all six components in 50 of 50 requests on MONAI and BrLP and in 40 of 50 on MAISI, as the paired loop does, and in 43 of 50 on WaveDiT and MedForj against 41 and 42 (exact McNemar $p = 0.5$ and 1), with 1.46, 1.72, 2.88, 2.80 and 2.78 generations per request against 2.92, 3.48, 6.16, 5.80 and 5.92. Over the five generators, it satisfies 226 of 250 requests against 223, and no request is satisfied by the paired loop alone. There the even part of the paired loop reaches a median of 1.32, 4.08, 7.80, 6.76 and 3.75 floors per probe at $f = 1$, and the Broyden update on the total change absorbs it into R from the second probe on. The mirrored arm separates the linear response from the even part, which grows with the dose (Appendix A), and gives the paired reading on which the paired variant stops and on which the analyses of this section rest.

	requests satisfied ↑		median leak ↓		generations per request ↓		McNemar
generator	paired	one-sided	paired	one-sided	paired	one-sided	p
MONAI	77/78	77/78	0.050	0.049	3.08	1.54	1
BrLP	78/78	78/78	0.047	0.049	2.92	1.49	1
WaveDiT	78/78	78/78	0.052	0.052	3.31	1.58	1
MedForj	68/78	68/78	0.074	0.073	4.33	2.17	1

Table 11: **Paired and one-sided probes**, closed loop on the returned volume, 78 held-out requests per generator on thirteen noises. One-sided: one generation per probe, read against the free sample of the same noise. Generations exclude the free sample. McNemar is exact on the discordant requests.

C.3 Dependence on the verifier

In this subsection, we segment the same volumes with FastSurfer (Henschel et al., 2020), which the controller never calls, to determine whether the conclusions depend on SynthSeg, which defines the contract, and we quantify how local the change is in labels, intensities and features.

Under FastSurfer, the leak decreases from the open loop to the paired variant by about a third on MONAI, BrLP and MAISI (0.32 to 0.19, 0.33 to 0.22, 0.18 to 0.12) and by more than half on MedForj (0.45 to 0.17); on WaveDiT, with the lowest open-loop leak, it changes from 0.11 to 0.14 (Table 14). FastSurfer segments both arms of both methods on the first four noises, with complete measurements for 24 requests on MONAI and MedForj, 23 on WaveDiT, 22 on BrLP and 15 on MAISI, expressed in its own floors. On an earlier MONAI run of 35 requests, the ratio between the changes measured by FastSurfer and by SynthSeg ranges from 0.66 to 2.2 across structures and differs by at most 0.05 between open-loop and paired-variant volumes. The relation between the two instruments is therefore stable under steering on this generator, and each expresses its verdicts in its own floors.

On the returned volumes of the closed loop, 228 requests per generator (the 78 held-out requests and 150 on the first 25 of the 100 noises), the target change measured by FastSurfer is 0.89 to 1.17 times the SynthSeg change on four generators and 1.58 times on MAISI, and the largest change of a protected structure remains below one FastSurfer floor in 162 to 228 of the 228 requests per generator (Table 13). The cap $\rho = 0.125$ is an eighth of a floor, finer than the disagreement between the two instruments: under FastSurfer, 1 to 44 of the 228 returned volumes satisfy it, against 193 to 228 under SynthSeg, which closes the loop.

With the leak bound alone, in FastSurfer floors, 43 to 107 of the 228 returned volumes satisfy a cap of 0.125, 82 to 179 a cap of 0.25, 135 to 225 a cap of 0.5 and 172 to 228 a cap of one floor (Table 14). In percent of volume, on the 78 held-out requests, the two instruments agree on the target change (Pearson 0.71, Spearman 0.76, 390 requests) and not on the changes of the protected structures (Pearson 0.17, Spearman 0.23), whose median absolute change is 2 to 6 times larger under FastSurfer. SynthSeg computes each volume as the sum of its soft posteriors, and 0.125 floors correspond to 11.5 to 15.8 voxels of the bilateral hippocampus or amygdala at 1 mm.

The change is local. Between the free sample and the returned volume, the SynthSeg labels have a median Dice of 0.992 to 0.994 on the requested structure and 0.997 to 0.999 on the five protected structures, and 0.08 to 0.42% of brain voxels change label, against 10.0% between two same-session

scans of the same person (92 pairs). On the requested structure, the voxels that change label are 1.08 to 1.37 times the net volume change (Table 12).

In image space, outside the requested structure dilated by 3 mm, the returned volume differs from the free sample by a median 1-SSIM of 1.7×10^{-6} to 8.8×10^{-5} and a PSNR of 57 to 79 dB, against 0.119 and 23.4 dB between two same-session scans of one person (the same 92 pairs); inside the requested structure the median 1-SSIM is 1.2×10^{-4} to 9.2×10^{-4}. Outside the edit, the change is 1.9 to 15 times that of a placebo pair at the operating dose on four generators, and repeated sampling at $\beta = 0$ is bit-identical. In MedicalNet features (Chen et al., 2019), the median distance between a returned volume and its free sample is 0.16 to 1.3% of the median distance between free samples of different noises on MONAI, BrLP, MAISI and WaveDiT.

	Dice ↑		relabelled voxels		outside the edit		inside
generator	target	protected	% of brain ↓	/ net change	1-SSIM ↓	PSNR (dB) ↑	1-SSIM
MONAI	0.992	0.997	0.33	1.33	8.8×10^{-5}	60	9.2×10^{-4}
BrLP	0.993	0.997	0.42	1.37	6.1×10^{-5}	57	6.0×10^{-4}
MAISI	0.992	0.997	0.23	1.34	5.2×10^{-5}	58	4.6×10^{-4}
WaveDiT	0.994	0.999	0.12	1.08	4.4×10^{-6}	70	1.6×10^{-4}
MedForj	0.993	0.999	0.08	1.09	1.7×10^{-6}	79	1.2×10^{-4}

Table 12: **Locality of the closed-loop change**, 78 held-out requests per generator: the returned volume against the free sample of its noise, medians over requests. Dice: SynthSeg labels of the requested structure (target) and of the five protected structures. Relabelled voxels: fraction of brain voxels whose SynthSeg label changes, and voxels entering or leaving the requested structure over its absolute net change. Outside the edit: brain minus the requested structure dilated by 3 mm, with both volumes mapped to $[0, 1]$ by the 0.5 and 99.5 percentiles of the free sample. Inside: the requested structure. Between two same-session scans of one person (92 pairs), 10.0% of brain voxels change label, and 1-SSIM and PSNR are 0.119 and 23.4 dB.

	contract, $\rho = 0.125$ ↑		target gain	protected below	median leak ↓	
generator	SynthSeg	FastSurfer	FastSurfer / SynthSeg	one FastSurfer floor ↑	SynthSeg	FastSurfer
MONAI	227	19	1.17	228	0.049	0.142
BrLP	228	8	1.07	183	0.047	0.325
MAISI	193	1	1.58	162	0.088	0.417
WaveDiT	228	44	0.91	227	0.051	0.194
MedForj	214	27	0.89	220	0.054	0.256

Table 13: **Closed-loop returned volumes measured by FastSurfer**, 228 requests per generator, the 78 held-out requests and 150 requests on the first 25 of the 100 noises: contracts satisfied under each instrument, in its own floors; median ratio of the target change, FastSurfer over SynthSeg; requests whose largest protected change stays below one FastSurfer floor; median leak. Abstentions count as failures.

	closed loop, leak bound at cap ↑				first four noises, FastSurfer leak		
generator	0.125	0.25	0.5	1	requests	open loop	paired variant ↓
MONAI	107	179	225	228	24	0.32	0.19
BrLP	47	91	144	172	22	0.33	0.22
MAISI	43	82	135	195	15	0.18	0.12
WaveDiT	78	140	206	227	23	0.11	0.14
MedForj	52	119	195	220	24	0.45	0.17

Table 14: **Leak measured by FastSurfer, in its own floors.** Left: returned volumes of the closed loop, out of 228 requests, whose FastSurfer leak is within the cap, with the leak bound alone. Right: requests of the first four noises with complete FastSurfer measurements of both arms of both methods, and median FastSurfer leak of the open loop and the paired variant.

C.4 PROFILE REQUESTS

In this subsection, we test whether our procedure satisfies six-structure requests (Section 5.3). A one-floor profile request is 0.02 of the measured patient profile and leaves the classifier score unchanged, hence the larger fractions of Table 38.

Profile requests on 100 held-out noises. On 100 further noises, with the command, doses and criterion of Table 29, the closed loop satisfies all six components of the AD profile at $f = 0.5$ and at $f = 1$ in 100 and 99 of 100 requests on MONAI, 100 and 100 on BrLP, 84 and 84 on MAISI, 96 and 73 on WaveDiT and 92 and 69 on MedForj, with 1.29 to 2.68 pairs on average at $f = 0.5$ and 1.67 to 3.48 at $f = 1$. Every unmet request ends without a returned volume. Over both doses, coverage is 99.5% on MONAI, 84.0% on MAISI, 84.5% on WaveDiT and 80.5% on MedForj, with 95% bootstrap intervals over noises of 98.5 to 100, 77.0 to 90.5, 79 to 89.5 and 74.0 to 86.5%. At $f = 1$, the joint score of Appendix C.5 accepts as AD 98 of 99 returned volumes on MONAI, 100 of 100 on BrLP and 72 of 73 on WaveDiT, and the 19 labels outside the request change in the direction of the real AD–CN difference in 48, 61 and 60% of the labels on average (Table 19). On MAISI and MedForj, it accepts 2 of 84 and 41 of 69 returned volumes, against 1 and 57 of 100 free samples.

With the 25% clause alone, the closed loop satisfies all six components in 109, 129, 68, 29 and 81 of 200 requests on MONAI, BrLP, MAISI, WaveDiT and MedForj (Table 15), and the one-floor clause decides only the components whose request is below four floors, where 25% of the request is below one floor, the mean test-retest difference of the verifier: the cortex, the white matter, the amygdala and the thalamus, with requests of 1.1 to 1.6 floors at $f = 0.5$ and 2.2 to 3.1 floors at $f = 1$. It decides mostly the thalamus, in 18 to 77% of the requests, the amygdala, in 1 to 35%, and the white matter, in 0 to 18%; the lateral ventricles and the hippocampus, with requests of 25 to 50 and 4.7 to 9.5 floors, never depend on it. The profile itself is uncertain: a bootstrap over the 46 subjects gives a 95% interval of $+2.2$ to $+65.1\%$ on the lateral ventricles, around $+30.6\%$, against a tolerance of 25% of the request.

				within 25% ↑ / floor only (%)						median relative error ↓					
generator	f	criterion satisfied ↑	25% clause only ↑	LV	CX	WM	HC	AM	TH	LV	CX	WM	HC	AM	TH
request (floors), $f = 0.5/1$				25/50	1.2/2.3	1.3/2.6	4.7/9.5	1.6/3.1	1.1/2.2						
MONAI	0.5	100	38	100/0	95/5	88/12	100/0	65/35	59/41	0.04	0.05	0.09	0.06	0.18	0.20
MONAI	1	99	71	99/0	96/3	96/3	99/0	94/5	77/22	0.02	0.03	0.03	0.03	0.07	0.10
BrLP	0.5	100	51	100/0	98/2	94/6	100/0	87/13	58/42	0.03	0.03	0.05	0.06	0.07	0.21
BrLP	1	100	78	100/0	100/0	98/2	100/0	97/3	82/18	0.02	0.01	0.02	0.02	0.03	0.16
MAISI	0.5	84	26	84/0	83/1	84/0	84/0	56/28	37/47	0.07	0.04	0.05	0.08	0.21	0.33
MAISI	1	84	42	84/0	84/0	84/0	84/0	74/10	50/34	0.05	0.03	0.03	0.06	0.15	0.25
WaveDiT	0.5	96	14	96/0	96/0	84/12	96/0	87/9	19/77	0.02	0.03	0.07	0.05	0.05	0.46
WaveDiT	1	73	15	73/0	73/0	67/6	73/0	71/2	19/54	0.03	0.03	0.10	0.03	0.04	0.37
MedForj	0.5	92	43	92/0	91/1	74/18	92/0	81/11	53/39	0.03	0.04	0.13	0.12	0.12	0.22
MedForj	1	69	38	69/0	69/0	58/11	69/0	68/1	47/22	0.06	0.06	0.23	0.14	0.12	0.31

Table 15: **Profile criterion by clause, closed loop**, on 100 held-out noises per generator, returned volume, 100 requests per row: requests satisfying the full criterion and the 25% clause alone (%), given for the baselines on twelve noises in Table 29; per structure, magnitude of the request in floors at $f = 0.5$ and 1 (first row), requests within 25% of the request and requests satisfied by the one-floor clause only (%); median relative error. Requests without a returned volume count as failures.

Paired profile results across fractions. The full profile measures 51.6 floors in Euclidean norm, 50.4 of them on the lateral ventricles, beyond the range of 0.5 to 5 floors over which Appendix A measures a linear response. The paired variant is measured at fractions 0.1, 0.25, 0.5 and 1 on the first four noises (Tables 16 and 38). The paired criterion holds on the six-component request at every fraction from 0.1 to 1, with 1 to 2.75 probes, an outside leak of 0.07 to 0.16 and median residual errors of 0.01 to 0.11 (infinity norm, relative to the largest component). These requests stop on the paired reading; the closed loop that stops on the returned volume is evaluated on 100 held-out noises in Table 15.

The profile, from 46 readings, is reproduced exactly from the stored readings. A bootstrap over subjects keeps the sign of five components in at least 98% of resamples, the thalamus in 89%, with 95% intervals of $+2.2$ to $+65.1\%$ on the lateral ventricles and -18.2 to -3.9% on the hippocampus.

generator	contract (%)	$\text{err}_\infty \downarrow$	$\text{leak}_{15} \downarrow$	probes $\downarrow$
MONAI	100	0.064	0.082	1.25
BrLP	100	0.076	0.099	1.00
MAISI	100	0.054	0.137	1.50
WaveDiT	100	0.085	0.066	2.00
MedForj	100	0.105	0.109	2.75

Table 16: **Profile requests at the full dose** ($f = 1$), stopping on the paired reading. Contract on the paired reading over four requests; err_∞ the residual of Eq. 3; leak_{15} the outside leak; probes per request. Classifier scores of the steered volumes are in Table 21. Fractions 0.1 to 0.5 and the guidance baselines steered to the same profile are in Table 38. On the returned volume at full dose, the positive arm satisfies the second condition of Eq. 3 in 4/4 requests on MONAI, WaveDiT and MedForj, 2/4 on BrLP and 0/4 on MAISI. Across fractions, the components outside 25% of their target are mostly the thalamus (40 of 80 requests) and the amygdala (16), with requests of 2.25 and 3.12 floors at $f = 1$ that fall below one floor at low fractions. All six components have the correct sign in every request but three, which have five, and in median 4.5 to 6 lie within 25% of their target.

C.5 Anatomical plausibility

In this subsection, we test whether the volumes returned by the closed loop resemble real brains, with a joint score over the regional volumes that is first validated on real scans.

Joint anatomical plausibility. A joint score over all 31 regional volumes accepts 95.4% of held-out CN and 95.2% of held-out AD scans and rejects combinations of volumes that a marginal test accepts (Table 17). It tests whether the returned volume is consistent with real scans of the requested class on the regional volumes that SynthSeg measures, including the bilateral labels outside the request. Each volume enters as the logarithm of its fraction of the total segmented volume. The reference, 806 real scans from ADNI1, OASIS-3 and BrainLat (523 cognitively normal, CN; 283 with Alzheimer's disease, AD), models class-specific age trends, site effects and the logarithm of the total segmented volume. The score is the squared Mahalanobis distance from the class mean, with a shrinkage covariance shared by the two classes. A generated volume has no age or site and receives the minimum over the ages and sites of the reference, as do the real calibration scans. Subject-disjoint cross-conformal calibration converts the distance into a p-value; a class accepts a volume when the p-value exceeds 0.05 and the segmentation passes a fixed quality rule.

In repeated subject-disjoint validation, the score accepts 0.1% of CN and 1.0% of AD vectors built by reassigning each regional volume independently across people, which keeps every marginal distribution and breaks the dependence between labels. It rejects more than 90% of scans with the left labels scaled by 1.15, a tripled left lateral ventricle or a halved left hippocampus, of which a marginal test of each of the 45 measures against its own range accepts 85 to 94% for the last two. Real scans changed by up to five floors in one structure are accepted at 0.93 to 0.96, within 0.02 of unmodified scans (0.95); larger changes are rejected once they exceed the spread of the structure across people (Table 18).

Under the CN criterion, single-structure steering keeps acceptance at nearly the rate of the free samples of the same noise on every generator (Table 19). Over 3,780 requests on five generators, read on the last probe of each request, 21 leave the reference and 5 enter it; the median paired change of the squared distance is 0.05 over requests and 0.07 over noises, against a median absolute difference of 9.0 between two held-out real CN scans. Acceptance varies much more across generators: before steering, the free samples of MAISI and MedForj are accepted in 1 and 31% of the noises. MAISI generates 256^3 volumes that include the skull, the neck and the shoulders.

Under the AD criterion at full dose, on the twelve noises shared with the other methods, the returned volume is accepted in 100% of the noises on MONAI and BrLP and in 92% on WaveDiT, and the 19 labels outside the request change in the direction of the real AD–CN difference in 49, 61 and 63%

of the labels on average. The contract constrains only the six planned structures, and the other labels follow the generator; under the conditioning of BrLP, 96 to 97% of them change in that direction on average (Table 32).

	CN reference			AD reference		
test	joint	marginal-45	marginal-31	joint	marginal-45	marginal-31
held-out reals	0.954	0.961	0.961	0.952	0.957	0.971
permuted across people	0.001	0.035	0.658	0.010	0.197	0.773
left hemisphere $\times 1.15$	0.001	0.002	0.897	0.086	0.153	0.872
left lateral ventricle $\times 3$	0.010	0.870	0.891	0.080	0.848	0.838
left hippocampus $\times 0.5$	0.000	0.893	0.944	0.000	0.938	0.943

Table 17: **Validation of the joint anatomical score** on real brains only, before any generated volume was read: fraction accepted as compatible with the CN or AD reference. Joint: the score over the 31 labels; marginal: each of the 45 (31 volumes and 14 asymmetries) or 31 measures against its own range.

		change of k floors, accepted fraction (joint)								
structure	reference	-25	-10	-5	-1	$+1$	$+5$	$+10$	$+25$	SD in floors
lateral ventricles	CN	0.956	0.957	0.956	0.954	0.954	0.953	0.952	0.948	88.7
	AD	0.953	0.954	0.953	0.953	0.952	0.952	0.951	0.946	91.6
cortex	CN	0.225	0.911	0.946	0.953	0.955	0.945	0.908	0.499	3.1
	AD	0.605	0.942	0.948	0.951	0.952	0.951	0.943	0.791	4.4
white matter	CN	0.750	0.941	0.950	0.954	0.954	0.950	0.935	0.857	4.8
	AD	0.875	0.948	0.952	0.953	0.951	0.950	0.946	0.918	4.4
hippocampus	CN	0.655	0.933	0.951	0.954	0.954	0.951	0.937	0.882	9.0
	AD	0.863	0.953	0.954	0.953	0.952	0.949	0.945	0.918	11.4
amygdala	CN	0.000	0.840	0.932	0.952	0.954	0.945	0.897	0.318	3.8
	AD	0.000	0.901	0.939	0.950	0.954	0.955	0.937	0.666	6.9
thalamus	CN	0.002	0.899	0.940	0.953	0.955	0.937	0.894	0.299	5.4
	AD	0.109	0.926	0.952	0.952	0.951	0.949	0.942	0.733	5.9

Table 18: **Sensitivity of the joint score on real scans.** Fraction of held-out real scans accepted when one structure is scaled by k floors, per reference class, in the same nested validation. SD in floors: dispersion of the covariate-adjusted residual of that structure across people.

	free	single structure (CN)					AD profile, $f = 1$ (AD)		
generator	x_0 in CN	requests (noises)	x_+ in CN	flips	contract	contract and joint	x_+ in AD	sign 19	contract and joint
MONAI	0.90	756 (126)	0.90	1/0	751	674	0.99	0.48	98/100
BrLP	0.94	756 (126)	0.93	11/1	756	704	1.00	0.61	100/100
MAISI	0.01	756 (126)	0.01	0/0	631	6	0.02	0.47	2/100
WaveDiT	0.79	756 (126)	0.79	6/4	749	591	0.99	0.60	72/100
MedForj	0.31	756 (126)	0.31	3/0	715	225	0.59	0.64	41/100

Table 19: **Joint anatomical plausibility of generated volumes, closed loop.** Single-structure requests of all cohorts, CN criterion: fraction of free samples x_0 and of the last probe x_+ of each request accepted, abstentions included; flips: last probes that leave the reference while the free sample was accepted, and the reverse; contract and joint: returned volumes satisfying both. AD profile, AD criterion, 100 noises: fraction of returned volumes accepted at $f = 1$; sign 19: mean fraction of the 19 labels outside the request that change in the direction of the real AD–CN difference; contract and joint: over the 100 requests at $f = 1$.

C.6 PLAUSIBILITY-CONSTRAINED PLANNING

In this subsection, we test whether a plausibility objective in the planner brings the returned profile volumes closer to the AD reference.

With twelve directions, the squared distance from the AD reference at the first probe falls by a median of 8.4 on WaveDiT and 6.9 on BrLP, in 11 and 12 of 12 noises, and the contract holds in

the same requests as with the closed loop (Table 20). The surrogate predicts left and right labels of each structure separately, which gives twelve directions without retraining. Among doses predicted to satisfy the contract, up to 1.25 times the closed-loop dose, the planner selects the one whose predicted reading is closest to the AD reference in squared distance; the loop stops when the returned volume satisfies the contract and the joint criterion. Three arms steer twelve noises per generator to the AD profile at $f = 1$: the closed loop, this objective on its six directions, and on twelve. On the returned volume, the distance falls by 10.6 and 8.3, and the median joint p-value rises from 0.17 to 0.34 and from 0.33 to 0.54. The four requests missed on WaveDiT are the same noises in all arms. The six directions alone lower the distance by 1.5 and 1.4 on the returned volume and by 1.6 and 1.2 at the first probe, and most of the gain comes from the separate left and right directions.

Table 20: **Plausibility-constrained planning.** Twelve-direction basis with the joint plausibility objective (AD profile, $f = 1$, twelve noises, paired). A: closed loop; M6: the six directions of the closed loop with the joint objective; M12: left/right basis with the joint objective. Δd^2: paired change of the measured d^2_{AD} at the first probe relative to A, median with bootstrap 95% interval. C6: contract on the returned volume; joint: C6 and $p_{\mathrm{joint}} > 0.05$. Bold: best; underlined: second best.

Generator	Arm	n	d^2 first probe	Δd^2 vs A	C6	Joint	Probes	Dose ratio vs A
WaveDiT	A	12	45.3	ref.	8/12	8/12	3.17	1.00
	M6	12	<u>43.1</u>	-1.6 [-2.3, -0.7]	8/12	8/12	3.50	0.86 [0.84, 0.91]
	M12	12	**35.8**	**-8.4** [-13.7, -4.5]	8/12	8/12	3.83	0.96 [0.92, 1.04]
BrLP	A	12	35.2	ref.	12/12	12/12	1.83	1.00
	M6	12	<u>34.5</u>	-1.2 [-1.5, -0.9]	12/12	12/12	2.00	0.91 [0.91, 0.93]
	M12	12	**28.1**	**-6.9** [-9.4, -4.3]	12/12	12/12	1.92	1.10 [0.95, 1.16]

C.7 Independent classification of the steered volumes

In this subsection, we score the profile-steered volumes of the closed loop with a second classifier that the controller never calls, to test whether the classifier scores depend on the segmenter that defines the contract.

A second classifier, never called by the controller, scores the steered arm above its own free sample in 12 of 12 requests on every generator (Table 21, profile at $f = 1$). It repeats the recipe of Sec. 4 on FastSurfer instead of SynthSeg volumes: the same 487 training scans from ADNI and OASIS-3, six structures normalised by the total segmented volume and 136 held-out BrainLat subjects, all segmented. Its out-of-cohort AUC is 0.900 (bootstrap interval 0.843 to 0.945). Five of six coefficients share the sign of the SynthSeg classifier; the sixth, on the thalamus, weighs -0.01. Held-out patients score 0.273 in median (interquartile range 0.051 to 0.651).

The opposite arm stays at 0.00 to 0.02 in median, the level of real controls, and the median steered score lies within the patient interquartile range on MONAI, BrLP, MAISI and MedForj, and at 0.029 on WaveDiT.

On twelve new noises per generator, the steered arm scores above its free sample in 12 of 12 requests on every generator, and the opposite arm moves by -0.011 to -0.027 in median. At doses of 0.95 to 1.01, the steered medians are 0.18 on MedForj, 0.20 on MAISI, 0.15 on MONAI, 0.10 on BrLP and 0.08 on WaveDiT, all within the patient interquartile range and below its median.

In the first batch, FastSurfer places the steered arms of BrLP and WaveDiT below the patient median (0.11 and 0.03) and those of MONAI, MAISI and MedForj near it (0.24 to 0.31); SynthSeg places WaveDiT below its patient median (0.19), BrLP at it (0.27) and the other three above it (0.35 to 0.88). Both classifiers score BrLP above WaveDiT, they differ in the order of the other three and on the free samples of MAISI (0.39 against 0.01), and SynthSeg scores every steered arm higher. The low FastSurfer score of WaveDiT reflects the logistic form of the classifier: its free samples score 0.007, a logit of -5.0, the farthest from the decision boundary of all generators. The logit shift predicted from the fitted coefficients and the measured FastSurfer fractions gives 0.032, against 0.029 observed, and 0.34 from a free sample at 0.094, the level of MedForj. Most classifier weight per unit of normalised fraction is on the amygdala and hippocampus.

			SynthSeg $P(AD)$			FastSurfer $P(AD)$			steered
generator	plan	dose	free	steered	opposite	free	steered	opposite	above free
Profile at $f = 1$									
MONAI	profile	1.00	0.06	0.48	0.01	0.04	0.24	0.01	12 /12
BrLP	profile	1.01	0.02	0.27	0.00	0.05	0.11	0.02	12 /12
MAISI	profile	0.97	0.39	0.88	0.08	0.01	0.30	0.00	12 /12
WaveDiT	profile	0.99	0.02	0.19	0.00	0.01	0.03	0.00	12 /12
MedForj	profile	0.98	0.11	0.35	0.01	0.09	0.31	0.02	12 /12
real patients ($n = 98$)				median 0.269			median 0.273		
real controls ($n = 38$)				median 0.014			median 0.008		

Table 21: **Disease classification of the profile-steered volumes by two volume-based classifiers.** Each row is one generator and one plan, on twelve noises. Dose is the median measured response in floors on the requested structures divided by the norm of the requested profile on the same structures. The first classifier reads the volumes of SynthSeg, the segmenter the controller closes on. The second is the same form of model rebuilt on FastSurfer volumes, a segmenter that never enters the loop. The last column counts, on FastSurfer and noise by noise, the steered samples scored above their own free sample. It is a classifier score, not a contract verdict. First batch of twelve noises; the replication on twelve new noises is reported in the text. Real BrainLat validation cohorts, read by each classifier and never recomputed here: patients 0.269 and controls 0.014 on SynthSeg, patients 0.273 and controls 0.008 on FastSurfer.

C.8 Supplementary analyses of the closed loop

In this subsection, we characterise the sampler, the leak cap and the sample size that an unpaired design would require.

Response across sampling windows. The five generators show five different patterns, with a late-to-early response ratio from 0.45 to 1.98 across structures and generators (Fig. 6). One dose is spent in one of five equal sampler windows at a time (1188 probes on MONAI and BrLP, 72 on the others). The largest effect is a late gain of the deep structures (thalamus, hippocampus, amygdala) on two generators. All reported experiments keep w uniform.

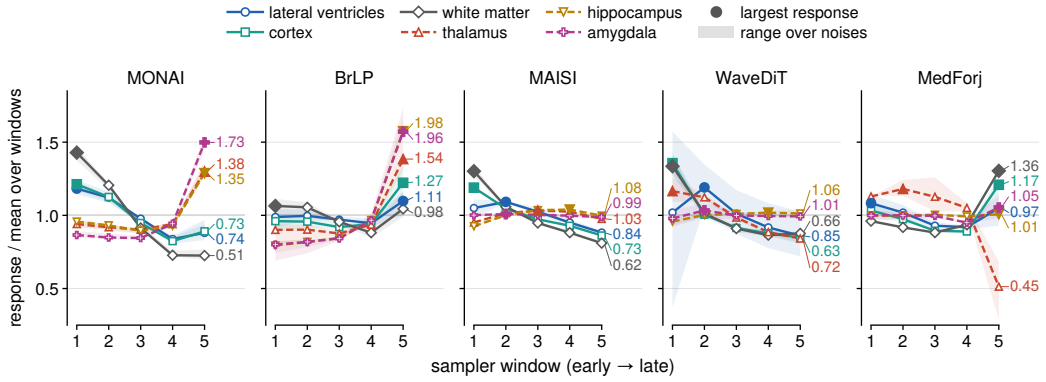


Figure 6: **Response across sampling windows.** Response per unit dose in each window, divided by its mean over the five windows, for the six planned structures. Filled markers: window of largest response. Number at the right: last-window over first-window response. Bands: range over noises (three doses and six noises on MONAI and BrLP, one dose and two noises elsewhere). Odd parts of paired probes, read by the surrogate.

Sensitivity to the leak cap. This ablation varies ρ where it clips the protected components of the directions (Eq. 4); the acceptance criterion does not change with ρ. It is an earlier absolute criterion on the paired reading: the target changes by more than one floor and each of the other five planned

structures by less than one floor. Each request applies one perturbation along the direction of its target, without probes, at a dose chosen per structure among four values as the one that satisfies the most requests on eight noises; the requests are counted on four other noises, six structures each (Table 22). The raw gradient is the normalised surrogate gradient of the target, with the same dose selection. The samplers run at 50 steps on BrLP, 100 on MONAI and WaveDiT and 200 on MAISI, and the raw-gradient run of MAISI at 400. Between $\rho = 0$ and 0.25, the number of requests that satisfy the criterion changes by at most 3 of 24 on every generator. At $\rho = 0.5$, it decreases only on BrLP. $\rho = 0.25$ on the four generators gives 77 of 96 requests, 2 fewer than the highest count per generator over the values run. At $\rho = 0.125$, 0.375 and 0.75, BrLP satisfies 23, 18 and 16 of 24 requests and MONAI 16, 18 and 17.

generator	raw gradient	$\rho = 0$	$\rho = 0.25$	$\rho = 0.5$
BrLP	12	22	22	17
WaveDiT	24	21	24	24
MAISI	14	13	13	14
MONAI LDM	17	15	18	18
MedForj	13	12	13	14

Table 22: **Sensitivity to the leak cap** ρ in the directions of Eq. 4, one perturbation per request without probes: requests, out of 24 (four noises, six structures), that satisfy the earlier absolute criterion on the paired reading (target above one floor, every other planned structure within one floor), which does not depend on ρ. Dose chosen per structure among four values on eight other noises. Sampler steps: 50 on BrLP, 100 on MONAI and WaveDiT, 200 on MAISI, 400 for the MAISI raw gradient. At $\rho = 0.125$, 0.375 and 0.75: BrLP 23, 18 and 16; MONAI 16, 18 and 17. The absolute criterion gives the raw gradient a higher score than in Table 8.

Sample size for unpaired detection. A one-floor request must be measured on the noise of the request (Table 23). At 80% power and two-sided $\alpha = 0.05$, a one-floor change requires $2\,(z_{\alpha/2} + z_{\beta})^2\sigma_b^2$ unpaired samples per group, with σ_b from 1.3 to 99 floors over the free samples, against less than one pair. The paired noise σ_p, the odd part of a placebo perturbation of operating size, is at most 0.038 floors.

	unpaired samples per group						
generator	lat. ventr.	cortex	white m.	hippoc.	amygd.	thalamus	paired noise
MONAI	62 152	166	451	617	107	509	0.029
BrLP	1 536	28	58	235	40	164	0.019
MAISI	9 336	90	91	1 142	510	769	0.038
WaveDiT	29 578	37	222	635	468	454	0.034
MedForj	152 850	177	271	798	287	863	0.002

Table 23: Unpaired samples per group needed to detect a one-floor change, and the largest noise of the paired reading, in floors.

D Comparison with baselines

In this section, we compare the closed loop with the guidance baselines and with the conditioning of BrLP, on single-structure and on profile requests.

D.1 Baseline implementation

In this subsection, we state how each guidance baseline is implemented and how its step is set, because a baseline receives a step size rather than a requested change and its result depends on that step.

We report two comparisons separately, because an interpolated measurement is not a sample generated at the inferred step size. The main comparison evaluates samples generated at frozen step sizes.

The a posteriori comparison reads each method at a matched target change, interpolated on a grid of step sizes.

FreeDoM computes the full gradient through the network and the surrogate reader. FlowChef bypasses the network and differentiates the reader at the current estimate of x_0. MPGD-LDM corrects x_0 on the data manifold and re-noises it. TFG is a family that contains the other methods. Each method needs its own scale: a step of 10^{-4} produces 20 floors on MAISI and 64 on WaveDiT, and 3×10^{-5} produces 1.3 floors on MONAI and 6.6 on MedForj.

FreeDoM differentiates through the network, which the released code of MAISI and WaveDiT runs without a gradient: MAISI decodes its 256^3 volume in tiles under no gradient, and the WaveDiT sampler detaches the velocity. Running it there would entail rewriting released code, while our directions need only the gradient through the authors' decoder, computed tile by tile on MAISI. On those two generators we apply the same step rule to the linear map from the latent to the clean estimate, with the network detached, and name this variant FreeDoM-lin. DPS also differentiates through the network and cannot be used on MAISI and WaveDiT; on MONAI, with the constant step of its released code frozen as for the other baselines, it satisfies the contract in 3 of 78 requests on the paired reading and in none on the returned volume, and it is therefore excluded from the baselines. On MAISI, FreeDoM-lin and FlowChef leak 0.30 and 0.25 at 17% of the requests. The frozen calibration lowers most rates further, with FreeDoM at 0% on MONAI and 8% on BrLP, and raises FreeDoM-lin on MAISI from 17% to 25%. The MAISI entries of FreeDoM and MPGD in Table 1 are estimated from the 24-request cohort with the ratio measured for FlowChef on MAISI. On BrLP, whose released sampler runs the network in half precision, the FreeDoM gradient overflows in the first steps of 8 of 48 profile arms; on those steps only, we rescale the loss by a power of two, which leaves the normalised step unchanged.

The MONAI model is conditioned on sex, age, ventricular volume and brain volume, and every MONAI volume in this paper uses the default values of the released inference configuration. Among the six planned structures, this conditioning addresses the ventricles alone, whereas BrLP conditions on five of them and is the conditioning baseline.

A posteriori target calibration. With the target change calibrated a posteriori for every request, the baselines close 12 to 50% of the requests (Table 37, 24-request cohort). These readings are not generated: of the 432, 264 are interpolated between two points of the grid of step sizes and 168 extrapolated from an edge point (Table 33). This calibration gives each baseline the step that produces one floor on the target of each request, the result that a search over the guidance strength alone can reach with any number of verifier calls; the remaining failures come from the leak, which depends on the direction and not on the strength. The best, TFG (mean) on WaveDiT, leaks 0.206 and closes 50% on both the paired reading and the positive arm, against 83% and 62% for the paired variant. The median leaks of the baselines range from 0.16 for MPGD on MedForj to 2.08 for FreeDoM with the full gradient on MedForj.

D.2 Single-structure comparison

In this subsection, we test the differences between the closed loop and the baselines on single-structure requests and compare their cost.

Paired comparisons. On 78 requests, the paired variant satisfies 51 to 75 contracts and a baseline 1 to 22 (Table 24). Each baseline runs as one generated pair per request on new noises, with a step per structure frozen on the earlier requests, and is read on the paired reading. On the positive arm, the volume it would return, a baseline satisfies 1 to 20. On the 78 requests, at most one request per comparison is satisfied by the baseline alone. Requests that share a noise are not independent, and we complement the McNemar test with a sign permutation over noises, which reaches its smallest attainable value, 1.2×10^{-4}, in every comparison on 13 noises. On 25 of the 100 noises, the closed loop satisfies 150 of 150 contracts on the returned volume on MONAI, BrLP and WaveDiT, against 26, 24 and 46 for FlowChef. On the paired reading, it satisfies 144, 142 and 136 against 26, 25 and 49 (Table 24). On both readings, every noise favours the closed loop, and the sign permutation by noise reaches its smallest attainable value, 3.0×10^{-8}. With a posteriori calibration on 24 requests, the setting most favourable to the baselines, $p \leq 0.011$ in all 15 comparisons (Table 25). Against the best baseline of each generator, the paired variant satisfies 53 points more contracts, with a 95%

bootstrap interval of 43 to 63 over the 20 noise-generator clusters and $p = 5 \times 10^{-6}$ for the sign permutation over clusters.

generator	baseline	requests	ours	base.	ours only	base. only	McNemar	by noise	Holm
MONAI	FlowChef	78	**71**	13	58	0	3.5×10^{-18}	1.2×10^{-4}	0.0016
MONAI	FreeDoM	78	**71**	11	60	0	8.7×10^{-19}	1.2×10^{-4}	0.0016
MONAI	MPGD	78	**71**	<u>14</u>	57	0	6.9×10^{-18}	1.2×10^{-4}	0.0016
BrLP	FlowChef	78	**75**	12	63	0	1.1×10^{-19}	1.2×10^{-4}	0.0016
BrLP	FreeDoM	78	**75**	6	69	0	1.7×10^{-21}	1.2×10^{-4}	0.0016
BrLP	MPGD	78	**75**	<u>13</u>	62	0	2.2×10^{-19}	1.2×10^{-4}	0.0016
MAISI	FlowChef	78	**51**	11	40	0	9.1×10^{-13}	1.2×10^{-4}	0.0016
WaveDiT	FlowChef	78	**66**	19	48	1	8.9×10^{-14}	1.2×10^{-4}	0.0016
WaveDiT	FreeDoM-lin	78	**66**	<u>22</u>	45	1	6.7×10^{-13}	1.2×10^{-4}	0.0016
WaveDiT	MPGD	78	**66**	12	55	1	7.9×10^{-16}	1.2×10^{-4}	0.0016
MedForj	FlowChef	78	**72**	9	64	1	1.8×10^{-18}	1.2×10^{-4}	0.0016
MedForj	FreeDoM	78	**72**	1	71	0	4.2×10^{-22}	1.2×10^{-4}	0.0016
MedForj	MPGD	78	**72**	<u>15</u>	58	1	1.0×10^{-16}	1.2×10^{-4}	0.0016
Closed loop against FlowChef on 25 of the 100 noises									
MONAI	FlowChef	150	**144**	26	118	0	3.0×10^{-36}	3.0×10^{-8}	4.8×10^{-7}
BrLP	FlowChef	150	**142**	25	118	1	1.8×10^{-34}	3.0×10^{-8}	4.8×10^{-7}
WaveDiT	FlowChef	150	**136**	49	89	2	1.7×10^{-24}	3.0×10^{-8}	4.8×10^{-7}

Table 24: **Paired tests against frozen-step baselines**, on new starting noises. Columns: contracts satisfied by each method, discordant requests, one-sided exact McNemar test, one-sided sign-permutation test over noises (smallest attainable value 1.2×10^{-4} with 13 noises and 3.0×10^{-8} with 25), and Holm correction over the sixteen rows. The last three rows compare the closed loop with FlowChef on 25 of the 100 noises, on the paired reading; on the returned volume, the counts are 150, 150 and 150 against 26, 24 and 46, with the same by-noise and Holm values. Bold: best; underlined: second best.

		requests satisfying		discordant requests		McNemar	Wilcoxon
generator	baseline	ours	baseline	only ours	only baseline	p	on the leak, p
MONAI	FreeDoM	**24**	4	20	0	$< 10^{-5}$	$< 10^{-5}$
MONAI	FlowChef	**24**	<u>5</u>	19	0	$< 10^{-5}$	$< 10^{-5}$
MONAI	MPGD	**24**	<u>5</u>	19	0	$< 10^{-5}$	$< 10^{-5}$
BrLP	FreeDoM	**20**	<u>4</u>	16	0	1.5e-05	$< 10^{-5}$
BrLP	FlowChef	**20**	<u>4</u>	16	0	1.5e-05	$< 10^{-5}$
BrLP	MPGD	**20**	<u>4</u>	16	0	1.5e-05	$< 10^{-5}$
MAISI	FreeDoM-lin	**11**	<u>4</u>	7	0	0.008	0.001
MAISI	FlowChef	**11**	<u>4</u>	7	0	0.008	0.002
MAISI	MPGD	**11**	<u>4</u>	7	0	0.008	5.8e-04
WaveDiT	FreeDoM-lin	**20**	10	11	1	0.003	3.0e-05
WaveDiT	FlowChef	**20**	11	10	1	0.006	0.002
WaveDiT	MPGD	**20**	<u>12</u>	9	1	0.011	0.005
MedForj	FreeDoM	**24**	3	21	0	$< 10^{-5}$	$< 10^{-5}$
MedForj	FlowChef	**24**	<u>10</u>	14	0	6.1e-05	3.8e-05
MedForj	MPGD	**24**	9	15	0	3.1e-05	4.5e-05

Table 25: **Paired tests with a posteriori baseline calibration** (24 requests), the more generous setting for the baselines. Contract verdict: one-sided exact McNemar test. Leak: one-sided paired Wilcoxon test, on the requests where both methods have one. Bold: best; underlined: second best.

Frozen-step baseline results. On the 78 held-out requests of MONAI, samples generated at the frozen step leak 0.38 for FlowChef and 0.51 for FreeDoM, against 0.39 and 0.38 interpolated (Table 33). We generate these rows wherever the frozen step falls on the grid, because the a posteriori reading is interpolated. On the 24-request cohort, MPGD on BrLP leaks 0.51 and closes 2 of 12 requests on the paired reading and 1 on the positive arm.

Guidance objectives with a penalty on the protected readings, the TFG configurations, the geometric mask and the sources of the baseline measurements are in Appendix D.4.

Generation cost. A calibrated baseline serves a request with one pair, 200 gradient evaluations and 176 to 226 seconds on MONAI and BrLP, after a step-size calibration of 24 to 108 pairs per generator and method, 36 for most (Table 28). The closed loop serves a request with 1.5 to 3.0 pairs and no gradient through the denoiser, in 58 to 310 seconds. Our times are medians between consecutive requests and include the verifier calls, which run on CPU after each pair while generation waits; the two arms of a pair are read in parallel. Baseline times cover guided generation and decoding and exclude the verifier.

Verifier calls and one-time cost. A closed-loop request requires 2.9 to 5.9 verifier calls, or 3.1 to 6.1 with the free sample shared by the requests of one noise, at 20 to 53 seconds each on CPU (Table 26). A frozen-step baseline needs two verifier calls per request after 50 to 218 for calibration, and satisfies the contract less often. Calibrated per request, it needs 5 to 18. On the six-structure profile, a baseline that differentiates through the network needs up to 78 GiB of memory on MONAI, against 11 to 47 GiB for an arm of the closed loop on the same generators. Table 27 divides the GPU hours of each method by the requests it serves.

	Ours, per request					One-time setup		
Generator, controller	probes ↓	verifier calls	+ per-noise	s / verifier call	s / request ↓	labelled vol.	matrix probes	verifier calls
MONAI, closed loop	**1.54**	3.1	3.2	20	**75**	160	36	238
MONAI, paired variant	1.67	5.3	7.8	29	224	160	36	238
BrLP, closed loop	**1.46**	2.9	3.1	30	**58**	160	36	238
BrLP, paired variant	1.88	5.6	8.1	42	224	160	36	238
MAISI, closed loop	2.96	5.9	6.1	53	**249**	160	12	186
MAISI, paired variant	**2.79**	6.8	9.3	78	524	160	12	186
WaveDiT, closed loop	1.65	3.3	3.5	28	**114**	240	18	279
WaveDiT, paired variant	**1.50**	4.9	7.4	40	232	240	18	279
MedForj, closed loop	2.17	4.3	4.5	28	**310**	718	12	744
MedForj, paired variant	**1.75**	5.5	8.0	41	616	718	12	744

	Baselines: step-size grid, every sample judged							
						per request		
Generator, method	grid pts	pairs ↓	grad. / pair	s / pair ↓	verifier calls, grid	(a posteriori)	frozen: calib.	per request
MONAI, FreeDoM	9	**84**	200	225	172	7.0	74	2
MONAI, FlowChef	9	**84**	200	**221**	172	7.0	74	2
MONAI, MPGD	14	94	200	222	192	7.8	74	2
MONAI, TFG	6	144	800 / 1000	868 / 1089	292	12.0	146	2
BrLP, FreeDoM	15	96	200	179	196	8.0	74	2
BrLP, FlowChef	15	96	200	**176**	196	8.0	74	2
BrLP, MPGD	9	**84**	200	177	172	7.0	74	2
MAISI, FreeDoM-lin	8	**60**	200	2736	124	5.0	50	2
MAISI, FlowChef	8	**60**	200	**2731**	124	5.0	50	2
MAISI, MPGD	3	72	200	2748	148	6.0	74	2
WaveDiT, FreeDoM-lin	9	**84**	200	158	172	7.0	74	2
WaveDiT, FlowChef	9	**84**	200	162	172	7.0	74	2
WaveDiT, MPGD	6	144	200	**157**	292	12.0	146	2
WaveDiT, TFG	5	120	792	589	244	10.0	122	2
MedForj, FreeDoM	9	**84**	200	234	172	7.0	74	2
MedForj, FlowChef	9	**84**	200	189	172	7.0	74	2
MedForj, MPGD	9	216	200	**188**	436	18.0	218	2

Table 26: **Verifier calls, runtime and setup cost.** Baselines: 'pairs' is the number of paired generations of the step-size grid, 'grad. / pair' the gradient evaluations of one pair, and 's / pair' the median time of one pair over the grid. TFG: mean and mean with variance, in that order. Closed loop: 78 held-out requests; paired variant: 24-request cohort. The per-noise calls of the closed loop are the free sample only. s / request: median time per request, including the verifier calls; s / verifier call: median duration of one call under concurrent load. Baseline s / pair: guided generation and decoding of one pair, excluding the verifier. Bold: best; underlined: second best.

	GPU hours per request satisfying the contract ↓				
generator	closed loop	paired variant	FreeDoM	FlowChef	MPGD
MONAI	**0.03**	0.11	0.93	0.93	0.82
BrLP	**0.02**	0.11	2.09	1.45	0.62
MAISI	**0.08**	0.32	60.01	13.23	14.30
WaveDiT	**0.04**	0.11	0.44†	1.06	1.05
MedForj	**0.11**	0.22	10.58	1.43	1.10

Table 27: **GPU hours per request satisfying the contract.** Each entry divides the GPU hours of the held-out requests by the requests whose returned volume satisfies the contract: 78 held-out requests for the closed loop and the baselines, and 102 for the paired variant (Table 1a). Closed loop and paired variant include their verifier calls and the population-model probes spread over the cohort; the paired variant also includes its open-loop comparison pair. Baselines include the step-size grid on the first four noises and exclude verifier time, which runs on CPU. Surrogate training and the setup of each noise, shared by its six requests (free sample and six surrogate gradients, and a placebo pair for the paired variant), are excluded throughout. FreeDoM-lin and MPGD have no frozen-step run on the MAISI held-out requests. †FreeDoM-lin. Bold: best; underlined: second best.

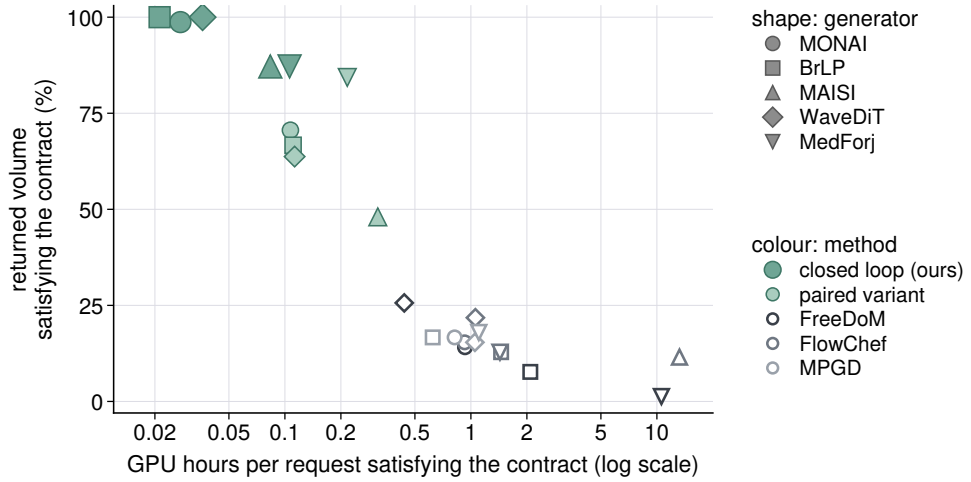


Figure 7: **Computational cost and coverage**: GPU hours per request whose returned volume satisfies the contract, against the share of requests satisfied. Closed loop and paired variant: measured time per request, including the verifier calls inside the loop, plus the population-model probes spread over the cohort. Baselines: measured generation time including the step-size grid on the first four noises, without verifier time. The FreeDoM colour includes FreeDoM-lin on WaveDiT. Surrogate training is excluded throughout.

D.3 Profile comparison

In this subsection, we compare the closed loop with the conditioning of BrLP and with the guidance baselines on the AD profile request, on the returned volume, on the classifier scores and on the joint anatomical score.

Returned-volume comparison with BrLP conditioning. On the five input structures of BrLP, the closed loop satisfies the criterion on the returned volume in 24 of 24 requests, against 5, 15 and 23 for three conditioned arms, which fail on the thalamus, not an input of BrLP, in all 24 (Table 29; paired Wilcoxon test on components satisfied, $p < 10^{-5}$ against each arm). It uses twelve new starting noises and two doses (0.5 and 1). A component is satisfied when its change on the returned volume lies within 25% of the request or within one floor; for requests of 1.1 to 3.1 floors, a quarter of the request is below one floor, the mean test-retest difference of the verifier. The loop corrects the positive arm by solving $Ra = t - e$, with e the even part of the last probe, with at most five

method	∇/pair	seconds ↓ MONAI	BrLP	MAISI	WaveDiT	MedForj	one-time
closed loop (ours)	0	**75** (1.54)	**58** (1.46)	**249** (2.96)	**114** (1.65)	310 (2.17)	$\bar{R}$
paired variant	0	224 (1.67)	224 (1.88)	524 (2.79)	232 (1.50)	616 (1.75)	$\bar{R}$
FreeDoM	200	226	179	2742†	157†	233	24–36
FlowChef	200	222	176	2757	162	**188**	24–36
MPGD	200	222	177	2748	157	**188**	36–108

Table 28: Cost per request: gradient evaluations per pair, seconds per paired generation (median over the test requests), and one-time cost (calibration pairs, or the population model $\bar{R}$). TFG, run on MONAI and WaveDiT, requires 868 and 589 seconds with 800 and 792 evaluations in its mean configuration, and 1089 seconds with 1000 evaluations on MONAI with variance. †FreeDoM-lin, used where the released code does not expose the gradient through the network. Closed loop and paired variant: median seconds per request, including the verifier calls and excluding the setup shared by the requests of a noise (free sample and six surrogate gradients, and a placebo pair for the paired variant), mean probes in parentheses, on the 78 held-out requests and on the 24-request cohort, respectively; the paired variant also includes its open-loop comparison pair. Baselines: seconds per paired generation, excluding the verifier. Bold: best; underlined: second best.

pairs, and stops when the criterion holds. The unmodified released BrLP code runs its conditioning as released, with one gain per input fitted on earlier noises, and with inputs corrected by verifier feedback under the same pair budget. With the 25% clause alone, the conditioned arms satisfy no request. Stopping on the paired reading ("ours, paired stop") gives 58% on BrLP, isolating the effect of correcting and testing the positive arm.

generator	method	requests	all six satisfied (%)	25% clause only	five BrLP inputs	median satisfied (of 6)	median relative error ↓ LV	CX	WM	HC	AM	TH	pairs ↓
MONAI	ours	24	**100** [86,100]	10	**24**	**6.0**	0.06	**0.11**	**0.07**	**0.06**	**0.09**	**0.17**	1.3
MONAI	FlowChef, frozen step	24	21 [9,40]	1	6	5.0	**0.03**	0.32	0.24	0.30	0.24	0.29	**1.0**
MONAI	FreeDoM, frozen step	24	4 [1,20]	0	2	4.0	0.05	0.55	0.35	0.46	0.24	0.75	**1.0**
MONAI	MPGD, frozen step	24	0 [0,14]	0	0	1.0	0.46	1.01	0.42	0.58	0.90	0.88	**1.0**
BrLP	ours	24	**100** [86,100]	16	**24**	**6.0**	0.03	0.03	**0.03**	**0.02**	**0.04**	**0.23**	1.8
BrLP	FlowChef, frozen step	24	0 [0,14]	0	0	3.0	0.09	0.21	0.43	0.64	0.53	0.68	**1.0**
BrLP	FreeDoM, frozen step	24	4 [1,20]	0	1	4.0	0.05	0.45	0.30	0.49	0.35	0.54	**1.0**
BrLP	MPGD, frozen step	24	0 [0,14]	0	0	1.0	0.53	0.59	0.93	0.71	0.78	1.68	**1.0**
BrLP	conditioning, released	24	0 [0,14]	0	5	4.0	0.09	0.16	0.22	0.30	0.37	2.26	**1.0**
BrLP	conditioning, calibrated	24	0 [0,14]	0	15	5.0	0.10	0.18	0.38	0.12	0.23	2.80	**1.0**
BrLP	conditioning + feedback	24	0 [0,14]	0	23	5.0	**0.00**	**0.01**	**0.03**	0.05	0.11	2.59	5.0
BrLP	ours, paired stop	24	58 [39,76]	7	23	**6.0**	0.07	0.05	0.09	0.05	0.06	0.54	1.4
MAISI	ours	24	**96** [80,99]	8	**23**	**6.0**	**0.03**	**0.02**	**0.04**	**0.07**	**0.13**	**0.22**	2.7
MAISI	FlowChef, frozen step	24	0 [0,14]	0	0	3.0	0.08	0.74	0.27	0.45	0.62	0.71	**1.0**
WaveDiT	ours	24	**88** [69,96]	5	**24**	**6.0**	**0.01**	**0.02**	**0.04**	**0.01**	**0.03**	**0.38**	2.7
WaveDiT	FlowChef, frozen step	24	4 [1,20]	0	12	5.0	0.07	0.53	0.15	0.14	0.34	2.26	**1.0**
WaveDiT	FreeDoM-lin, frozen step	24	0 [0,14]	0	0	3.0	0.04	0.46	0.91	0.72	0.33	1.17	**1.0**
WaveDiT	MPGD, frozen step	24	0 [0,14]	0	0	2.0	0.07	0.46	1.13	0.47	0.71	1.25	**1.0**
MedForj	ours	24	**79** [60,91]	7	**22**	**6.0**	**0.05**	**0.05**	**0.11**	**0.14**	**0.15**	**0.24**	3.2
MedForj	FlowChef, frozen step	24	0 [0,14]	0	0	3.0	0.06	0.58	0.36	0.41	0.44	1.27	**1.0**
MedForj	FreeDoM, frozen step	24	0 [0,14]	0	0	2.0	0.16	1.00	0.87	0.41	0.92	2.20	**1.0**
MedForj	MPGD, frozen step	24	0 [0,14]	0	0	1.0	0.28	1.04	1.59	0.76	1.03	5.24	**1.0**

Table 29: **Returned-volume profile accuracy**, 24 requests per row: requests with all six components satisfied, with the 25% clause alone and on the five BrLP inputs (the thalamus is not one); median components satisfied; median relative error per structure; pairs. Brackets: 95% Wilson interval. Baseline steps are set on two calibration noises; on MAISI, FlowChef is the frozen-step baseline. Bold: best; underlined: second best.

Frozen-step FlowChef on 25 of the 100 noises. On the first 25 of the 100 held-out noises of the profile evaluation at $f = 0.5$ and 1, FlowChef at its frozen step satisfies all six components in 0 of 50 requests on BrLP and 9 of 50 on MONAI, and the closed loop in 50 of 50 on both, with no request satisfied by FlowChef alone; with the 25% clause alone, the counts are 0 against 32 on BrLP

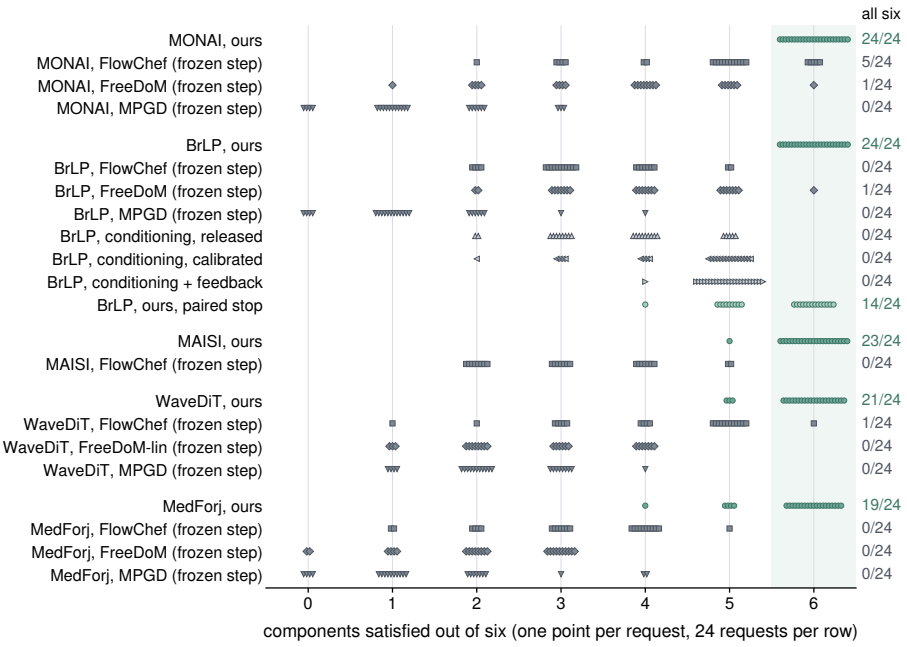


Figure 8: **Profile requests on the returned volume**: components satisfied out of six, read on the positive arm ($r + e$ for our loop), one point per request, 24 requests per row (twelve noises, two doses). Ours, paired stop: the same loop stopped on the paired reading, read on its positive arm. Right: requests with all six satisfied. Missing requests count as zero.

and 0 against 26 on MONAI (exact McNemar $p = 8.9 \times 10^{-16}$ and 4.6×10^{-13}, sign permutation by noise $p = 3.0 \times 10^{-8}$ and 6.0×10^{-8}). On BrLP, FlowChef moves the thalamus against the request in 8 of the 50 requests, the closed loop in none. Each method is compared with its own free sample of the same noise.

Classifier scores and profile-guidance baselines. P(AD) of the volume classifier rises with the dose and falls on the opposite arm on every generator (Table 38). The baseline rows use the step size that produces the requested profile on a grid read by the verifier, a posteriori, and test the direction of a method, not its ability to satisfy a request. With the step size read by the verifier on each request, which is the information the closed loop obtains by probing, FreeDoM and FlowChef reach the error and outside leak of the paired variant on MONAI and in four of the six rows on BrLP, and at a frozen step they satisfy at most 5 of 24 returned-volume requests on each of the two generators, against 24 of 24 for the closed loop (Table 29).

Classifier guidance raises the score while moving the anatomy away from the request. At its largest step, the median P(AD) of the positive arm is 0.62 on BrLP and 1.00 on MONAI, against 0.03 and 0.06 unguided, but the paired change projects negatively on the requested profile in 20 of 20 requests on BrLP and 15 of 20 on MONAI. The lateral ventricles, which carry 50 of the requested floors, shrink in 20 and 19 of 20; the other five structures mostly move as requested. It was run on four noises, with five step sizes on MONAI and BrLP, three on WaveDiT and MedForj and two on MAISI. The diagnosis covariate of BrLP moves the anatomy opposite to Alzheimer's disease, as reported for AG-LDM.

At matched dose, the closed loop is the closest to the request on MONAI and MedForj (Table 30). On BrLP, MAISI and WaveDiT, the returned arm overshoots the requested dose, at doses of 1.13, 1.25 and 1.11, and the baseline arm with the lowest L_2 error, with its step chosen a posteriori, lies at doses of 1.02, 0.92 and 1.06. The score does not order arms by distance from the request, and serves as a check on the resulting anatomy, not a ranking: on every generator at least one baseline arm scores above the closed loop, and on MONAI and MedForj the closest arm scores below arms

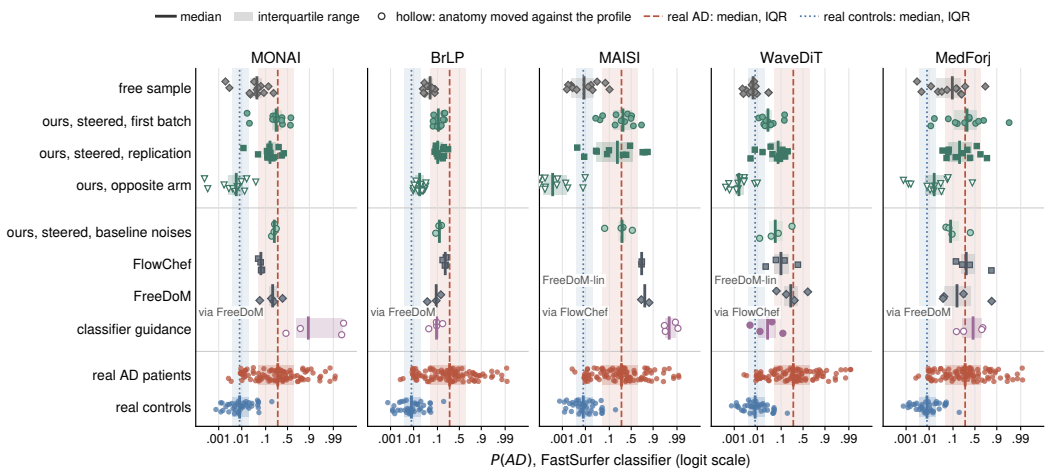


Figure 9: **Classifier scores of the steered volumes relative to real patients**, read by the independent FastSurfer classifier (logit axis). Upper block: twelve noises per row; the replication uses twelve new noises. Lower block: baselines on the four noises of the closed loop in the row above. Each uses the step whose median measured dose is closest to that of our arm on the same noises. Hollow markers: anatomy moved against the profile (the opposite arm by construction, a baseline when its measured response has a negative cosine to the profile). Bands and bottom rows: the held-out real cohort (98 AD, 38 controls) read by the same classifier.

1.5 to 6 times farther. On these two generators and MAISI, the arms that score highest are those guided by the classifier itself, with P(AD) of 0.89, 0.45 and 0.98 on MONAI, MedForj and MAISI and a cosine to the profile of -0.24, -0.78 and -0.22. Single-structure arms leave the score at or below that of the free sample on every generator.

The available guidance baselines also keep the acceptance rate of their own free samples, within one request, on 78 requests per generator (Table 31). Contract and CN criterion hold together in 54, 73 and 60 of 78 requests for the closed loop on MONAI, BrLP and WaveDiT, and in at most 9, 12 and 13 for the frozen-step baselines.

On the four noises where classifier guidance was run, the agreement of the 19 labels outside the request with the direction of the real AD–CN difference is 46 and 75% on MONAI and BrLP for the paired variant, stopped on the paired reading, against 13 and 18% for classifier guidance at its strongest step (Table 32), which moves these labels opposite to the profile.

D.4 Supplementary baseline analyses

In this subsection, we characterise the baselines beyond what the main comparisons require: penalised guidance objectives, the TFG configurations, the geometric mask and the sources of the baseline measurements.

Penalised guidance objectives. A penalty on the five protected readings improves the paired reading of the baselines but not their returned volume, which satisfies the contract in at most 4% of the requests. The penalty is added to the target objective of each baseline on the first four noises, over 24 requests calibrated a posteriori (12 on MAISI) and 12 at a step frozen on two noises and applied to the other two. On the paired reading, the penalised baselines satisfy 17 to 58% a posteriori and 0 to 50% at the frozen step, best with FlowChef on WaveDiT and MedForj a posteriori and on WaveDiT and MONAI at the frozen step. The returned arm leaks more (1.27 against 0.29 for FlowChef on BrLP), a change common to both arms that pairing cancels. The paired variant satisfies 50 to 100% of the same requests paired and 33 to 88% returned. On the 78 held-out requests, penalised FlowChef satisfies none on the returned volume on BrLP, WaveDiT and MedForj (17 to 31% on the paired reading), against 78, 78 and 68 for the closed loop (sign permutation by noise, Holm-adjusted $p = 3.7 \times 10^{-4}$).

generator	method	dose	cosine	L_2 error	FastSurfer $P(AD)$ free	FastSurfer $P(AD)$ steered	FastSurfer $P(AD)$ opposite	requests
MONAI	ours	0.975	**+1.000**	**1.4**	0.028	0.210	0.003	4
MONAI	FlowChef, 1e-05	0.031	+0.983	50.0	0.028	0.029	0.028	4
MONAI	FlowChef, 3e-05	0.093	+0.985	46.9	0.028	0.030	0.027	4
MONAI	FlowChef, 0.0001	0.311	+0.986	35.9	0.028	0.034	0.023	4
MONAI	FlowChef, 0.0003	0.934	+0.990	7.8	0.028	0.065	0.013	4
MONAI	FlowChef, 0.001	1.041	+0.999	3.1	0.028	0.280	0.003	4
MONAI	FreeDoM, 1e-05	0.031	+0.980	50.0	0.028	0.029	0.028	4
MONAI	FreeDoM, 3e-05	0.094	+0.980	46.8	0.028	0.029	0.028	4
MONAI	FreeDoM, 0.0001	0.317	+0.982	35.7	0.028	0.032	0.025	4
MONAI	FreeDoM, 0.0003	0.930	+0.986	9.0	0.028	0.056	0.018	4
MONAI	FreeDoM, 0.001	0.997	+0.996	4.9	0.028	0.184	0.005	4
MONAI	FreeDoM (cls.), 1e-05	0.009	-0.360	51.8	0.028	0.029	0.027	4
MONAI	FreeDoM (cls.), 3e-05	0.026	-0.394	52.1	0.028	0.032	0.026	4
MONAI	FreeDoM (cls.), 0.0001	0.089	-0.390	53.5	0.028	0.041	0.020	4
MONAI	FreeDoM (cls.), 0.0003	0.273	-0.367	58.4	0.028	0.111	0.008	4
MONAI	FreeDoM (cls.), 0.001	1.012	-0.243	81.6	0.028	0.890	0.000	4
BrLP	ours	1.130	**+0.999**	6.9	0.049	0.119	0.020	4
BrLP	FlowChef, 3e-05	0.054	+0.986	48.9	0.051	0.052	0.049	4
BrLP	FlowChef, 0.0001	0.179	+0.985	42.5	0.051	0.054	0.048	4
BrLP	FlowChef, 0.0003	0.544	+0.986	24.4	0.051	0.061	0.042	4
BrLP	FlowChef, 0.001	1.064	+0.996	5.4	0.051	0.147	0.019	4
BrLP	FlowChef, 0.003	1.102	+0.995	7.8	0.051	0.195	0.012	4
BrLP	FreeDoM, 3e-05	0.032	+0.987	49.9	0.051	0.060	0.050	2
BrLP	FreeDoM, 0.0001	0.121	+0.988	45.5	0.051	0.056	0.049	3
BrLP	FreeDoM, 0.0003	0.320	+0.987	35.4	0.051	0.062	0.045	2
BrLP	FreeDoM, 0.001	1.015	+0.996	**4.4**	0.051	0.080	0.028	3
BrLP	FreeDoM, 0.003	1.077	+0.997	5.6	0.051	0.089	0.014	3
BrLP	FreeDoM (cls.), 3e-05	0.013	-0.450	51.9	0.051	0.052	0.048	4
BrLP	FreeDoM (cls.), 0.0001	0.044	-0.485	52.7	0.051	0.056	0.045	4
BrLP	FreeDoM (cls.), 0.0003	0.133	-0.494	55.2	0.051	0.071	0.035	4
BrLP	FreeDoM (cls.), 0.001	0.432	-0.479	64.3	0.051	0.130	0.014	4
BrLP	FreeDoM (cls.), 0.003	0.547	-0.522	70.1	0.051	0.092	0.003	4
MAISI	ours	1.248	**+0.999**	13.2	0.014	0.285	0.001	4
MAISI	FlowChef, 0.0003	0.851	**+0.999**	8.0	0.018	0.487	0.001	2
MAISI	FlowChef, 0.001	0.921	+0.995	**6.6**	0.018	0.741	0.001	2
MAISI	FlowChef (cls.), 0.0003	0.239	-0.067	54.2	0.014	0.905	0.000	4
MAISI	FlowChef (cls.), 0.001	0.290	-0.224	56.8	0.014	0.978	0.001	4
MAISI	FreeDoM-lin, 5e-06	0.031	+0.994	50.0	0.018	0.020	0.019	2
MAISI	FreeDoM-lin, 2e-05	0.123	+0.994	45.3	0.018	0.028	0.021	2
MAISI	FreeDoM-lin, 8e-05	0.503	+0.992	26.1	0.018	0.057	0.016	2
MAISI	FreeDoM-lin, 0.0003	0.850	**+0.999**	8.0	0.018	0.454	0.001	2
MAISI	FreeDoM-lin, 0.001	0.916	+0.994	7.2	0.018	0.796	0.000	2
WaveDiT	ours	1.109	**+0.999**	6.3	0.015	0.058	0.006	4
WaveDiT	FlowChef, 3e-05	0.246	+0.973	39.4	0.015	0.029	0.015	4
WaveDiT	FlowChef, 0.0001	0.805	+0.996	10.8	0.015	0.069	0.008	4
WaveDiT	FlowChef, 0.0003	0.925	+0.995	6.4	0.015	0.098	0.006	4
WaveDiT	FlowChef (cls.), 1e-06	0.007	+0.238	51.5	0.015	0.016	0.014	4
WaveDiT	FlowChef (cls.), 3e-06	0.021	+0.233	51.3	0.015	0.018	0.012	4
WaveDiT	FlowChef (cls.), 1e-05	0.070	+0.269	50.7	0.015	0.028	0.008	4
WaveDiT	FreeDoM-lin, 2.77e-05	0.219	+0.971	40.7	0.015	0.027	0.015	4
WaveDiT	FreeDoM-lin, 8.75e-05	0.710	+0.992	15.9	0.015	0.068	0.010	4
WaveDiT	FreeDoM-lin, 0.000277	0.922	+0.993	7.2	0.015	0.095	0.007	4
WaveDiT	FreeDoM-lin, 0.000875	0.955	+0.993	6.2	0.015	0.158	0.008	4
WaveDiT	FreeDoM-lin, 0.00277	1.064	+0.996	**5.7**	0.015	0.226	0.010	4
MedForj	ours	0.999	**+0.998**	**3.3**	0.022	0.080	0.005	4
MedForj	FlowChef, 0.0001	0.652	+0.989	18.8	0.021	0.134	0.004	4
MedForj	FlowChef, 0.0003	0.668	+0.986	18.6	0.021	0.206	0.004	4
MedForj	FlowChef, 0.001	0.927	+0.997	5.4	0.021	0.296	0.005	4
MedForj	FreeDoM, 0.0001	1.075	**+0.998**	5.0	0.021	0.196	0.004	4
MedForj	FreeDoM, 0.0003	1.048	+0.997	5.0	0.021	0.142	0.004	4
MedForj	FreeDoM, 0.001	1.060	+0.997	5.0	0.021	0.119	0.005	4
MedForj	FreeDoM (cls.), 3e-06	0.099	-0.467	52.8	0.021	0.032	0.015	4
MedForj	FreeDoM (cls.), 1e-05	0.372	-0.447	56.3	0.021	0.075	0.006	4
MedForj	FreeDoM (cls.), 3e-05	1.000	-0.782	99.9	0.021	0.454	0.001	4

Table 30: **Per-arm comparison with the guidance baselines on the first four noises (two for some arms).** One row per method and step size, the step size following the method name, for every arm asked for the AD profile or, marked cls., for the classifier score. Baseline step sizes are chosen a posteriori on the grid read by the verifier. Dose, cosine and L_2 error in floors are measured on SynthSeg volumes against the frozen profile, on the component-wise median over the requests. $P(AD)$ is the FastSurfer classifier. Requests: noises on which the steered arm was generated and read. Bold: best; underlined: second best.

generator	method	x_0 in CN	x_+ in CN	flips	contract	contract and joint	ours on the same requests contract / contract and joint
MONAI	FlowChef	0.69	0.69	0/0	12	8	**77/54**
MONAI	FreeDoM	0.69	0.69	0/0	11	7	**77/54**
MONAI	MPGD	0.69	0.69	0/0	13	9	**77/54**
BrLP	FlowChef	0.92	0.92	0/0	10	9	**78/73**
BrLP	FreeDoM	0.92	0.92	0/0	6	5	**78/73**
BrLP	MPGD	0.92	0.92	0/0	13	12	**78/73**
MAISI	FlowChef	0.08	0.08	0/0	9	1	**68/6**
WaveDiT	FlowChef	0.77	**0.78**	0/1	17	13	**78/60**
WaveDiT	FreeDoM-lin	0.77	**0.78**	0/1	20	13	**78/60**
WaveDiT	MPGD	0.77	0.77	1/1	12	9	**78/60**
MedForj	FlowChef	0.08	0.08	0/0	10	1	**68/6**
MedForj	FreeDoM	0.08	0.08	0/0	1	0	**68/6**
MedForj	MPGD	0.08	0.08	0/0	14	1	**68/6**

Table 31: **Joint anatomical plausibility of the guidance baselines** at a frozen step, on the 78 single-structure requests per generator where they were run, CN criterion. Each method is paired with its own free sample x_0. The last column gives the closed loop on the same requests. Bold: best; underlined: second best.

generator	method	noises	x_+ in AD	sign 19
MONAI	closed loop	12	**1.00**	0.49
MONAI	FlowChef, frozen step	12	**1.00**	0.51
MONAI	FreeDoM, frozen step	12	0.75	0.41
MONAI	MPGD, frozen step	12	0.75	0.46
MONAI	classifier guidance, strongest step	4	0.25	0.13
BrLP	closed loop	12	1.00	0.61
BrLP	BrLP conditioning, calibrated	12	1.00	0.97
BrLP	BrLP conditioning + feedback	12	1.00	0.97
BrLP	BrLP conditioning, released	12	1.00	0.96
BrLP	FlowChef, frozen step	12	1.00	0.60
BrLP	FreeDoM, frozen step	12	1.00	0.59
BrLP	MPGD, frozen step	12	1.00	0.55
BrLP	classifier guidance, strongest step	4	1.00	0.18
MAISI	closed loop	12	0.00	0.53
MAISI	FlowChef, frozen step	12	0.00	0.51
MAISI	classifier guidance, strongest step	4	0.00	0.28
WaveDiT	closed loop	12	0.92	0.63
WaveDiT	FlowChef, frozen step	12	0.83	0.60
WaveDiT	FreeDoM-lin, frozen step	12	0.75	0.63
WaveDiT	MPGD, frozen step	12	0.42	0.71
WaveDiT	classifier guidance, strongest step	4	**1.00**	0.46
MedForj	closed loop	12	**0.50**	0.60
MedForj	FlowChef, frozen step	12	0.42	0.69
MedForj	FreeDoM, frozen step	12	0.25	0.63
MedForj	MPGD, frozen step	12	0.00	0.70
MedForj	classifier guidance, strongest step	4	0.25	0.37

Table 32: **Joint anatomical plausibility under the AD profile** at $f = 1$, AD criterion, for the closed loop and the other methods on the same noises. x_+ in AD: fraction of returned volumes accepted; sign 19: mean fraction of the 19 labels outside the request that change in the direction of the real AD–CN difference. Classifier guidance was run on the first four noises, not on the twelve. Bold: best; underlined: second best.

TFG details. TFG is reported on MONAI and WaveDiT, with a per-request dose grid of three scales per configuration on MONAI and five on WaveDiT. It requires 792 to 1000 gradient evaluations per request and leaks 0.50 and 0.54 in its two confirmed configurations on MONAI and 0.206 on WaveDiT.

Geometric mask. The geometric mask, an inpainting control that restores the unrequested region at every step as in RePaint (Lugmayr et al., 2022), rather than a competing guidance method, satisfies the contract in at most 17% of the requests (Table 37). On the 78 held-out requests, it satisfies 12 on MONAI and none on BrLP and MedForj, against 77, 78 and 68 for the closed loop (sign permutation by noise, Holm-adjusted $p = 3.7 \times 10^{-4}$).

Generated and interpolated baseline measurements. No a posteriori baseline request in Table 37 is generated; each is read from a grid of step sizes at the same target change. The frozen-step rows are generated wherever the frozen step falls on the grid (Table 33). On MONAI, FlowChef leaks 0.38 generated and 0.39 interpolated, and FreeDoM 0.51 and 0.38.

		A posteriori step, per request			Frozen step, 24-request cohort		
Generator	Method	requests	interp.	extrap.	requests	generated	not generated
MONAI	FreeDoM	24	20	4	12	0	12
MONAI	FlowChef	24	20	4	12	0	12
MONAI	MPGD	24	16	8	12	12	0
MONAI	TFG	24	3	21	12	0	12
BrLP	FreeDoM	24	20	4	12	12	0
BrLP	FlowChef	24	20	4	12	12	0
BrLP	MPGD	24	20	4	12	12	0
MAISI	FreeDoM-lin	24	4	20	12	12	0
MAISI	FlowChef	24	4	20	12	12	0
MAISI	MPGD	24	17	7	12	0	12
WaveDiT	FreeDoM-lin	24	16	8	12	12	0
WaveDiT	FlowChef	24	16	8	12	12	0
WaveDiT	MPGD	24	24	0	12	0	12
WaveDiT	TFG	24	24	0	12	0	12
MedForj	FreeDoM	24	3	21	12	12	0
MedForj	FlowChef	24	11	13	12	12	0
MedForj	MPGD	24	23	1	12	0	12

Table 33: **Baseline measurement sources.** A posteriori requests are interpolated between two grid points or extrapolated from an edge point; none of the 432 is generated or failed; the MONAI TFG row holds for both configurations, which have the same counts. The frozen-step columns cover the 12 requests of the 24-request cohort not used to fit the step. Every frozen-step run on the 78 held-out requests of Table 1 (13 generator-method pairs) is generated in full.

E LIMITATIONS

In this section, we state the conditions under which the results hold.

Abstentions. Abstentions count as failures throughout. The paired variant abstains when a large even part shows that the response is not linear at the current point; the closed loop plans around the even part and abstains only when its probe budget is spent.

Verifier dependence. The contract verdicts hold for SynthSeg: the cap ρ is defined relative to what SynthSeg resolves, and FastSurfer reads a different leak on the same volumes.

Measurement noise at the acceptance threshold. At one floor, the leak threshold is about 3.6 times the largest noise of the paired reading. The floors are measured on preprocessed real scans and applied to raw generated volumes. They come from 92 same-session pairs of a single cohort, OASIS-3, and may differ with scanner, site and acquisition protocol.

Validity range of the local response model. The response is linear up to five floors on all generators except the MAISI hippocampus. At 20 floors, the gain is 0.74 against 0.83 on MONAI and 0.53 against 0.54 on BrLP. The planner certifies the local model, not the generator, and every plan is therefore verified on the final volume.

Scope of the baseline comparison. At the frozen step, every baseline satisfies at most 22 of 78 contracts on the paired reading, against 51 to 75 for the paired variant (Table 24). The a posteriori calibration gives each baseline the dose the controller finds by probing, mostly by extrapolation for FreeDoM-lin and FlowChef on MAISI, FreeDoM and FlowChef on MedForj and TFG on MONAI. The odd part cancels any change common to both arms, such as an even response or both arms leaving the intended brain (recurrent TFG variants), and the positive-arm reading is therefore reported beside it throughout.

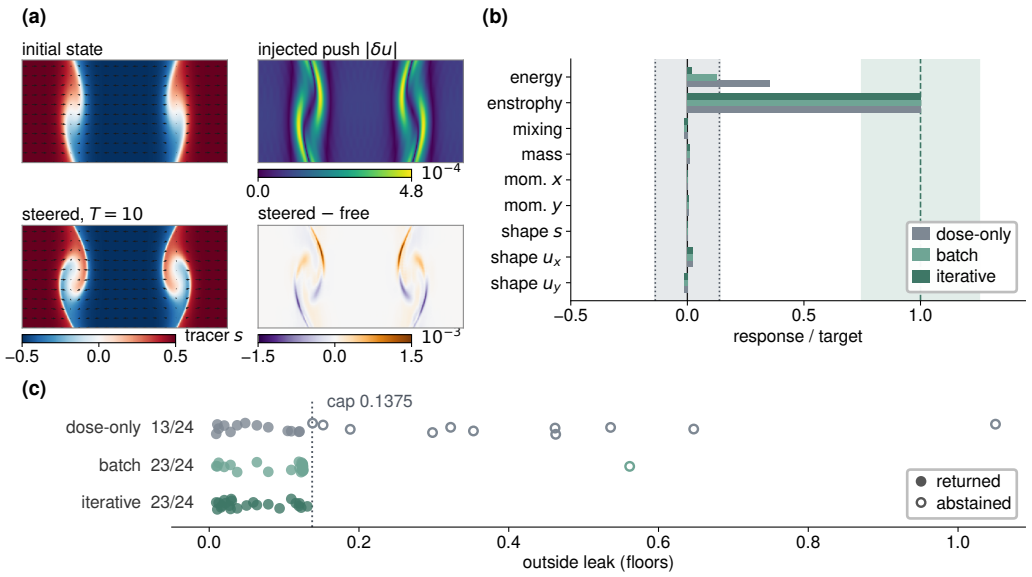


Figure 10: Steering a shear flow on the frozen Walrus emulator. (a) One request: the initial state, the injected velocity perturbation (mean-zero, divergence-free, pressure unchanged), the steered field at $T = 10$, and its difference from the free rollout. The perturbation is 0.034% of the velocity, the tracer changes by at most 0.30% of its range, and enstrophy moves by one floor. (b) A held-out request of $+1$ floor on enstrophy under the three identification schemes. Dose-only adaptation moves energy past the protection cap and abstains; batch and iterative identification satisfy contract-9. (c) Outside leak on the eight protected coordinates for all 24 held-out requests. Filled marks are returned fields, hollow marks abstentions (the iterative abstention stops before its leak is measured and is not drawn), the dotted line is the cap 0.1375, and leaks above 1.05 are drawn at the right edge. Dose-only satisfies 13 of 24 requests, batch and iterative 23 each.

F TRANSFER TO A PHYSICS EMULATOR

In this section, we test whether the same procedure transfers to a frozen physics emulator. Neural emulators predict physical flows, and sensitivity studies require flows in which one physical quantity changes by a specified amount while the others are held fixed.

On a frozen flow emulator, feedback raises the satisfied contracts from 12 to 21 of 24 held-out requests on eight trajectories, with abstentions counted as failures, and every open-loop failure is on a protected quantity. The emulator predicts a fluid flow from a few initial frames; we perturb its input and compute quantities such as energy directly on the final field, without a segmentation tool. The difficulty is selectivity: energy and enstrophy, the squared local rotation of the flow, tend to change together. We write y_j for the change Δ_j in floors.

F.1 FLOW DATA, INTERVENTIONS AND MEASUREMENTS

The experiments use the shear-flow data of The Well (Ohana et al., 2024), where neighbouring fluid layers slide past one another. The simulation solves the incompressible Navier-Stokes equations with a passive tracer, a scalar transported by the flow without changing its dynamics. The four channels are the tracer s, the pressure p and the velocities u_x and u_y, on a periodic 256×512 grid with $dt = 0.1$.

Repeated predictions of the frozen Walrus emulator (McCabe et al., 2025) differ by at most 1.6×10^{-4} floors once we disable the random shift of its patch grid, which the released model applies at every step, evaluation included. Walrus receives 6 context frames and predicts $T = 10$ steps in fp32 with TF32 disabled; measurements use float64.

Of the nine quantities measured on the final field, six describe the whole field: energy $\langle |u|^2 \rangle$, enstrophy $\langle \omega^2 \rangle$, tracer variance $\mathrm{var}(s)$ as the mixing measure, tracer mass $\langle s \rangle$ and the two momentum components, with $\langle \cdot \rangle$ a spatial average and ω the vorticity. The other three compare the means of the two halves of the domain for s, u_x and u_y; they detect changes outside the plan. Only energy, enstrophy and mixing enter the solve; the other six quantities are protected by the contract without a direction of their own. The floor of each quantity is the median absolute error of the free emulator against the numerical ground truth at time T, the emulator's own error rather than the repeatability of an instrument.

A random velocity placebo, projected and normalised alike at the largest amplitude used, moves the target by a median 0.0060 of the intended response at equal norm. A request asks for a one-floor change in energy, enstrophy or mixing, and contract-9 requires $y_k \in [0.75, 1.25]$ on the target and $|y_j| \leq 0.1375\, y_k$ on each of the eight protected measurements. The directions are the gradients of energy and enstrophy with respect to velocity and of mixing with respect to the tracer, projected to zero mean and zero divergence and normalised to unit length. A perturbation enters every context frame with the same coefficient and may not exceed 10% of a channel; violations are recorded, not clamped.

F.2 Open-loop steering and the paired variant

In this subsection, we test whether feedback is needed outside brain MRI.

Feedback raises the satisfied contracts from 6 to 11 of 12 on four held-out trajectories (Table 34). The open loop obtains its direction and cross-responses from a population model, the median response on two calibration contexts, and generates one pair on the request's trajectory only to rescale the dose. Arm B is the paired variant of the brain generators, not the closed loop, unchanged except for a domain adapter, with at most 21 rollouts per request, judged on its regenerated positive arm. The trajectories combine Reynolds numbers $\{5 \times 10^4, 5 \times 10^5\}$, the balance of inertial and viscous effects, with Schmidt numbers $\{0.2, 5\}$, the ratio of momentum to tracer diffusion, and carry three targets each.

Table 34: **Open-loop (A) and paired-variant (B) comparison** on four held-out trajectories, contract-9 on the positive arm. "Worst of A" is the open-loop quantity farthest from the contract, in floors (limit 0.1375).

id	trajectory	target	A	worst of A ↓	B	probes ↓
C-01	Re 5×10^4, Sc 0.2	energy	no	enstrophy 0.297	ok	1
C-02		enstrophy	no	energy 0.668	ok	2
C-03		mixing	ok	energy 0.009	ok	1
C-04	Re 5×10^4, Sc 5	energy	no	enstrophy 0.859	abstained	1
C-05		enstrophy	ok	energy 0.075	ok	1
C-06		mixing	no	enstrophy 0.151	ok	1
C-07	Re 5×10^5, Sc 0.2	energy	no	enstrophy 0.202	ok	1
C-08		enstrophy	no	energy 0.249	ok	1
C-09		mixing	ok	enstrophy 0.047	ok	1
C-10	Re 5×10^5, Sc 5	energy	ok	enstrophy 0.123	ok	1
C-11		enstrophy	ok	mixing 0.053	ok	1
C-12		mixing	ok	shape s 0.007	ok	1

The open-loop failures come from protected quantities, mainly the energy-enstrophy coupling, because the rescaling holds the odd part of the target change at a median 1.000 (range 0.988 to 1.000; 0.975 to 1.001 on the positive arm). The paired variant measures it on each trajectory. For energy at Re 5×10^4 and Sc 5 the target moved with the opposite sign, and the request ended in abstention. A second cohort at Schmidt numbers 0.5 and 2 gives 6 of 12 for the open loop and 10 of 12 for the paired variant, including two abstentions.

Selection among existing initial conditions satisfies none of the 12 exploratory requests under the four null models tested, over up to 7955 candidate fields, because in these candidates energy and enstrophy decay together.

F.3 Response estimation methods

In this subsection, we test how the response matrix should be estimated.

A single batch measurement of the response matrix satisfies 23 of 24 requests, as many as iterative updates, with 152 distinct rollouts against 264 (Table 35). Dose-only control, which keeps the population direction and adapts only the amplitude, satisfies 13 (Fig. 10). The requests use eight further held-out trajectories, with the Reynolds and Schmidt numbers of the first 24 requests and three targets each and a shared ceiling of 21 rollouts per request. Batch estimation measures the full matrix from three paired probes and plans once; the iterative scheme is the paired variant of Table 34, which updates it probe by probe.

Table 35: **Response estimation methods on Walrus**, 24 held-out requests each. Cost counts distinct rollouts; a rollout shared by a context's three targets counts once. The same request fails for all three schemes. Bold: best; underlined: second best.

scheme	contracts satisfied	abstentions	cost
dose only	13/24	11 (final verification)	152
batch matrix	**23**/24	1 (final verification)	152
iterative	**23**/24	1 (controller)	264

F.4 Limitations of the emulator evaluation

The emulator does not enforce the governing equations, and its free outputs already have nonzero divergence. It is also sensitive to representation: a uniform pressure offset changes the measured quantities by up to 2.88 floors, about twenty times the leak allowed for a one-floor request. The reported contracts apply to the pressure convention of the data. The linearity tests cover perturbations between 0.2% and 6.9% of the 10% limit. Requests from the same trajectory are not independent. The rollout counters of these experiments record 1301 Walrus rollouts; the earlier exploratory runs are not counted.

G GENERATED EXAMPLES

In this section, we present the returned volumes, which complement the measured changes reported in the tables with their appearance in the image.

The contract holds on all 30 requests of the test grid (Fig. 11, Table 36). Each generator shows every request issued on one starting noise, the first of the 13 noises of the 78 held-out requests, without omissions: one floor on each of six structures, served by the closed loop. Fig. 13 shows samples steered toward the Alzheimer's profile at its four fractions. Each request appears in three orthogonal planes over the free sample in grey, two requests per row. The overlay is the change of the requested structure from the free sample in Fig. 11, and the paired response, half the difference between the label maps of the two arms, in Fig. 13. Red marks a gain of volume, blue a loss.

generator	lateral ventricles	cortex	white matter	hippocampus	amygdala	thalamus
MONAI	1.09 / 0.024	1.00 / 0.114	1.00 / 0.044	1.01 / 0.016	0.98 / 0.085	1.17 / 0.063
BrLP	0.93 / 0.018	0.94 / 0.107	0.99 / 0.022	0.91 / 0.025	0.80 / 0.069	0.93 / 0.073
MAISI	0.89 / 0.105	0.99 / 0.134	1.01 / 0.059	1.00 / 0.106	1.02 / 0.121	1.00 / 0.054
WaveDiT	0.92 / 0.017	1.00 / 0.093	1.01 / 0.056	0.99 / 0.108	0.93 / 0.053	1.00 / 0.021
MedForj	1.00 / 0.015	0.96 / 0.085	1.00 / 0.052	0.97 / 0.031	0.98 / 0.032	0.77 / 0.070

Table 36: **Measurements of the requests in Fig. 11.** Obtained floors and leak of each returned volume.

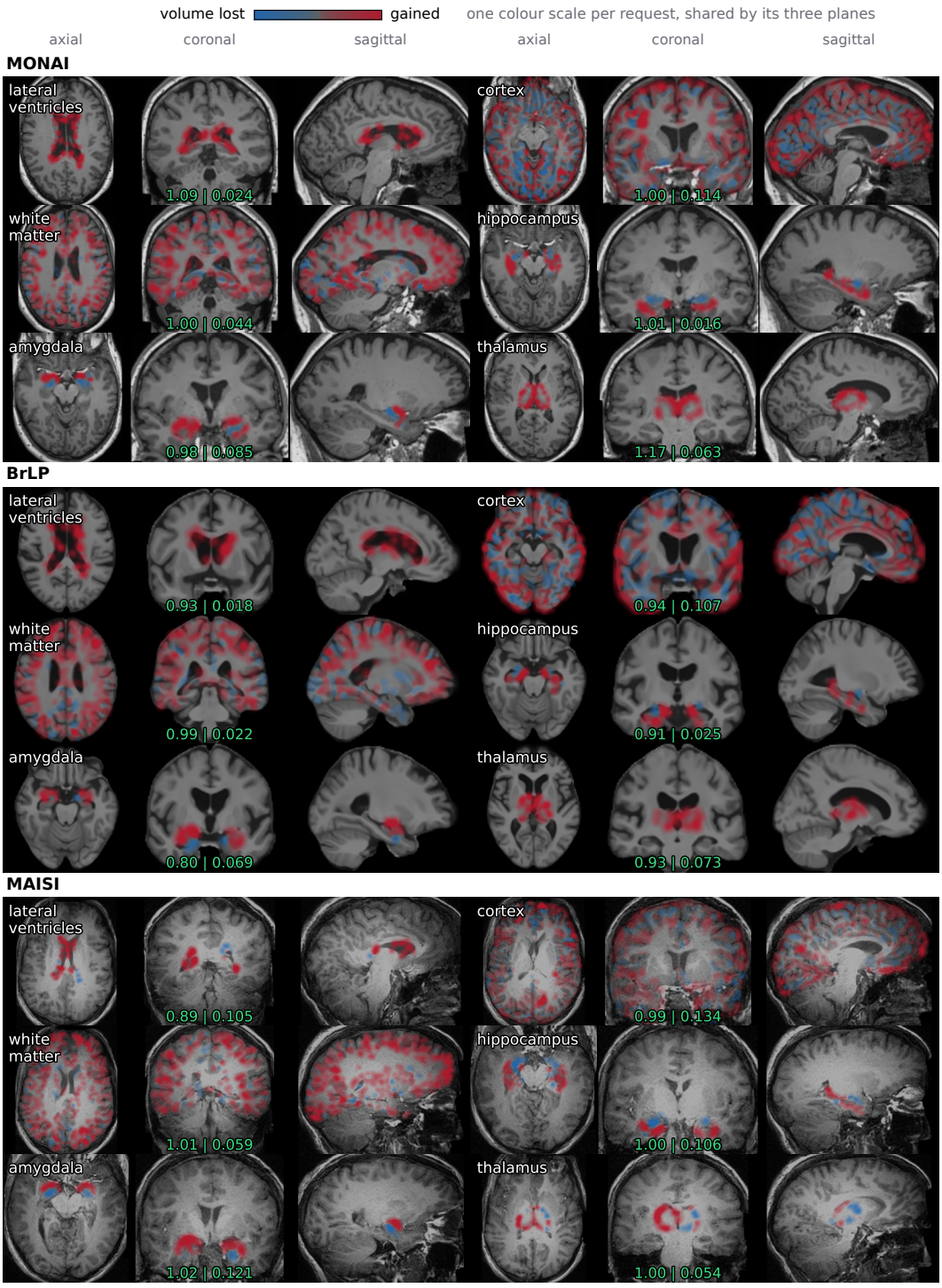


Figure 11: MONAI, BrLP and MAISI samples on the first held-out noise, one floor requested on each structure. Values under the coronal plane: obtained floors and leak, green when the contract holds (Table 36). The three planes of a request share one colour scale.

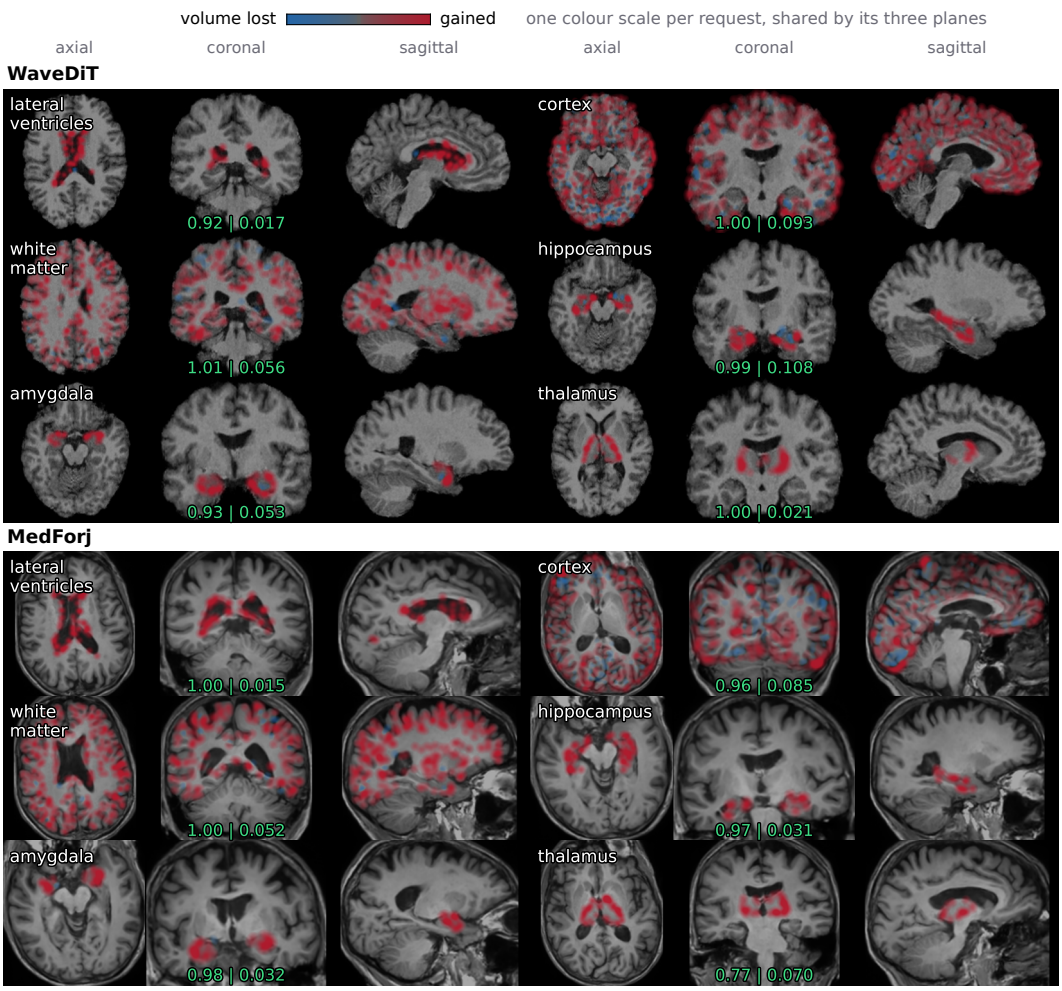


Figure 12: Continuation of Fig. 11: WaveDiT and MedForj.

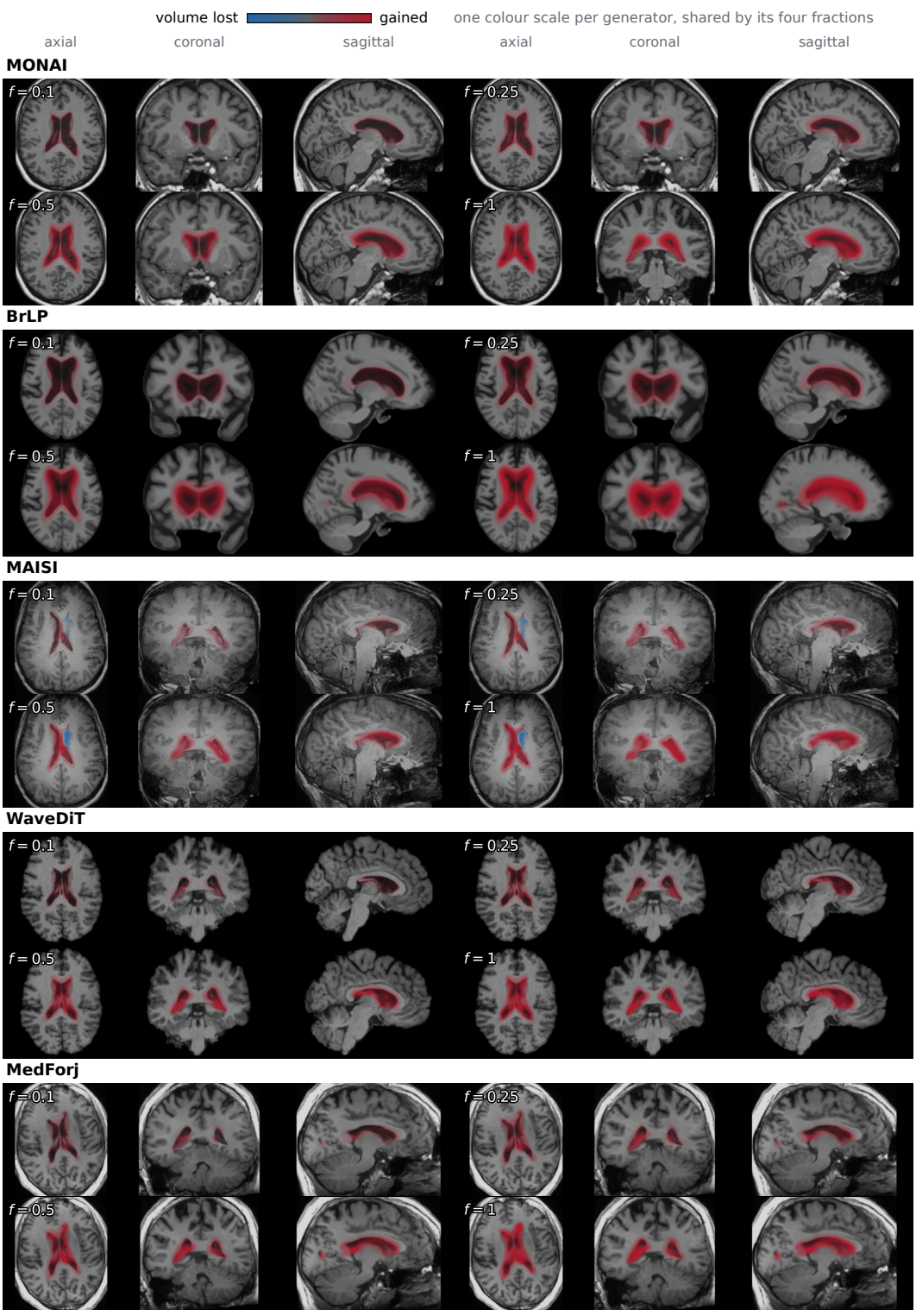


Figure 13: Samples steered toward the Alzheimer's profile at $f = 0.1, 0.25, 0.5, 1$ of a patient's atrophy, on the second of the first four noises. The overlay shows the lateral ventricles, the largest requested change, on the first probe pair of the paired variant. One colour scale per generator across its four fractions.

H COMPLETE COMPARISON TABLES

In this section, we give the complete per-method results of Appendices D.2 and D.3 on the original 24-request single-structure cohort and the profile study, not on the expanded main cohorts, with paired and positive-arm results and the calibration of each baseline row.

generator	method	target change	$\text{leak}_6 \downarrow$	contract-6 (%) paired / returned	$\text{leak}_{15} \downarrow$	contract-21 (%) paired / returned	probes / grad. ↓	s per request ↓
MONAI	open loop	0.76	0.239	17 / 17	0.226	17 / 17	0	61‡
	paired variant (ours)	1.00	**0.124**	**100 / 79**	0.213	79 / **54**	1.67	224
	paired variant, contract-21 plan	1.00	0.126	83 / 46	**0.153**	**83** / 46	2.62	167
	geometric mask (RePaint)	0.65	0.314	12	0.207	12	0	**57**
BrLP	open loop	0.55	0.229	0 / 0	0.282	0 / 0	0	58‡
	paired variant (ours)	1.00	**0.124**	**83 / 54**	0.258	50 / **46**	1.88	224
	paired variant, contract-21 plan	1.00	**0.124**	67 / 42	**0.147**	**67** / 42	2.92	200
	geometric mask (RePaint)	0.07	0.802	0	0.693	0	0	**53**
MAISI	open loop	1.03	0.273	8 / 8	0.243	8 / 8	0	109‡
	paired variant (ours)	1.00	**0.121**	**46 / 38**	**0.208**	**29 / 25**	2.79	524
	paired variant, contract-21 plan	1.01	0.125	29 / 21	0.231	21 / 12	3.83	429
	geometric mask (RePaint)	0.93	0.222	17	0.262	12	0	**107**
WaveDiT	open loop	0.89	0.203	8 / 8	0.134	8 / 8	0	66‡
	paired variant (ours)	1.00	**0.102**	**83 / 62**	0.109	**62 / 54**	1.50	232
	paired variant, contract-21 plan	1.00	0.131	58 / 46	**0.067**	50 / 46	3.25	231
MedForj	open loop	0.27	0.162	0 / 0	0.235	0 / 0	0	164‡
	paired variant (ours)	1.00	**0.115**	**100 / 88**	0.203	**67 / 58**	1.75	616
	paired variant, contract-21 plan	1.00	0.123	62 / 50	**0.126**	58 / 50	2.92	476
	geometric mask (RePaint)	0.17	0.125	0	0.218	0	0	**156**

generator	method	target change	$\text{leak}_6 \downarrow$	contract-6 (%) paired / returned	$\text{leak}_{15} \downarrow$	contract-21 (%) paired / returned	probes / grad. ↓	s per request ↓	frozen: $\text{leak}_6 \downarrow$ / paired / returned (%)
MONAI	FreeDoM, full gradient	1.00	0.382	17 / 17	0.362	17 / 17	200 grad.	225.56	**0.379** / 0 / 0
	FlowChef	1.00	0.396	21 / 21	0.207	21 / 21	200 grad.	221.95	0.387 / 8 / 8
	MPGD	1.00	0.555	21 / 21	0.213	21 / 21	200 grad.	222.28	0.442 / **17** / **17**
	TFG (mean)	1.00	0.502	21 / 21	0.246	21 / 21	800 grad.	867.56	0.419 / 0 / 0
	TFG (mean+var.)	1.00	0.536	17 / 17	0.389	17 / 17	1000 grad.	1088.70	0.822 / 0 / 0
BrLP	FreeDoM, full gradient	1.00	0.560	17 / 17	0.536	17 / 17	200 grad.	178.87	0.520 / 8 / 8
	FlowChef	1.00	0.440	17 / 17	0.300	17 / 17	200 grad.	175.91	**0.446** / **17** / **17**
	MPGD	1.00	0.499	17 / 12	0.284	17 / 12	200 grad.	176.53	0.509 / **17** / 8
MAISI	FreeDoM-lin	1.00	0.300	17 / 17	0.265	17 / 17	200 grad.	2742.29	0.242 / **25** / **25**
	FlowChef	1.00	0.251	17 / 17	0.283	17 / 17	200 grad.	2757.30	**0.214** / 17 / **25**
	MPGD	1.00	0.341	17 / 17	0.261	17 / 17	200 grad.	2747.90	0.329 / 17 / 17
WaveDiT	FreeDoM-lin	1.00	0.259	42 / 38	0.123	42 / 38	200 grad.	157.37	0.269 / **25** / **33**
	FlowChef	1.00	0.260	46 / 46	0.129	46 / 46	200 grad.	161.77	0.308 / **25** / 8
	MPGD	1.00	0.218	50 / 46	0.150	50 / 46	200 grad.	**156.99**	**0.228** / 17 / 17
	TFG (mean)	1.00	0.206	50 / 50	0.152	50 / 50	792 grad.	588.50	0.229 / 8 / 8
MedForj	FreeDoM, full gradient	1.00	2.084	12 / 12	0.654	12 / 12	200 grad.	233.07	1.997 / 8 / 8
	FlowChef	1.00	0.193	42 / 42	0.207	33 / 33	200 grad.	188.49	0.186 / **33** / **42**
	MPGD	1.00	0.162	38 / 38	0.224	29 / 29	200 grad.	188.36	**0.169** / **33** / 33

Table 37: **Single-structure comparison on the original 24-request cohort**, all five generators. Top: the paired variant and the controls that use no gradient. Bottom: the gradient baselines, with the frozen calibration fitted on two of the four noises and applied to the other two. Leak_6 and leak_{15} are medians over the 24 requests, on the six planned and the fifteen outside structures. Each contract column gives the rate on the paired reading and on the returned volume (Sec. 3.4). The geometric mask (Lugmayr et al., 2022) cannot be calibrated and has a single rate. Cost is paired probes for the paired variant and gradient evaluations per pair for the baselines. Seconds per request are measured on one GPU, including generations and verifier calls. ‡Estimated from the paired-variant time per pair. Bold: best; underlined: second best.

generator	method	f	contract (%)	$\text{err}_\infty \downarrow$	$\text{leak}_{15} \downarrow$	P(AD) free	P(AD)+	opposite	probes $\downarrow$
MONAI	ours, paired stop	0.10	100	0.056	0.086	0.056	0.071	0.044	1.25
	ours, paired stop	0.25	100	0.061	0.089	0.056	0.101	0.031	1.25
	FreeDoM, full gradient	0.25	100	0.058	**0.038**	0.056	0.111	0.047	100 ∇
	FlowChef	0.25	100	**0.046**	0.076	0.056	0.096	0.045	100 ∇
	ours, paired stop	0.50	100	0.063	0.092	0.056	0.179	0.018	1.25
	FreeDoM, full gradient	0.50	100	0.055	**0.058**	0.056	0.178	0.028	100 ∇
	FlowChef	0.50	100	**0.041**	0.065	0.056	0.194	0.027	100 ∇
	ours, paired stop	1.00	100	0.064	0.082	0.056	0.440	0.005	1.25
	FreeDoM, full gradient	1.00	100	0.051	**0.050**	0.056	0.569	0.011	100 ∇
	FlowChef	1.00	100	**0.035**	0.052	0.056	0.525	0.005	100 ∇
BrLP	ours, paired stop	0.10	100	0.082	0.109	0.026	0.035	0.020	1.00
	ours, paired stop	0.25	**100**	0.079	0.113	0.026	0.053	0.013	1.00
	FreeDoM, full gradient	0.25	75	**0.058**	0.136	0.028	0.053	0.014	100 ∇
	FlowChef	0.25	**100**	0.064	**0.071**	0.028	0.079	0.011	100 ∇
	ours, paired stop	0.50	100	0.078	0.104	0.026	0.105	0.007	1.00
	FreeDoM, full gradient	0.50	100	**0.061**	0.091	0.028	0.148	0.005	100 ∇
	FlowChef	0.50	100	0.068	**0.090**	0.028	0.117	0.010	100 ∇
	ours, paired stop	1.00	**100**	0.076	0.099	0.026	0.325	0.002	1.00
	FreeDoM, full gradient	1.00	75	**0.049**	**0.073**	0.030	0.156	0.002	100 ∇
	FlowChef	1.00	**100**	0.081	0.095	0.028	0.403	0.002	100 ∇
MAISI	ours, paired stop	0.10	100	0.018	0.160	0.238	0.275	0.198	2.00
	ours, paired stop	0.25	**100**	**0.011**	0.150	0.238	0.343	0.149	2.00
	FreeDoM-lin	0.25	**100**	0.056	0.120	0.238	0.470	0.342	100 ∇
	FlowChef	0.25	50	0.112	**0.117**	0.238	0.457	0.333	100 ∇
	ours, paired stop	0.50	100	**0.019**	0.146	0.238	0.476	0.086	2.00
	FreeDoM-lin	0.50	100	0.074	0.104	0.238	0.517	0.185	100 ∇
	FlowChef	0.50	100	0.082	**0.100**	0.238	0.583	0.219	100 ∇
	ours, paired stop	1.00	**100**	**0.054**	**0.137**	0.238	0.728	0.021	1.50
	FreeDoM-lin	1.00	50	0.097	0.163	0.238	0.790	0.048	100 ∇
	FlowChef	1.00	**100**	0.078	0.159	0.238	0.774	0.073	100 ∇
WaveDiT	ours, paired stop	0.10	100	0.101	0.100	0.018	0.022	0.014	1.25
	ours, paired stop	0.25	100	0.085	0.100	0.018	0.030	0.011	1.25
	FlowChef	0.25	100	**0.066**	**0.079**	0.018	0.038	0.011	100 ∇
	ours, paired stop	0.50	100	**0.033**	0.099	0.018	0.051	0.006	1.50
	FlowChef	0.50	100	0.072	**0.081**	0.018	0.073	0.007	100 ∇
	ours, paired stop	1.00	100	0.085	**0.066**	0.018	0.122	0.003	2.00
	FlowChef	1.00	100	**0.061**	0.080	0.018	0.240	0.003	100 ∇
MedForj	ours, paired stop	0.10	100	0.051	0.152	0.109	0.136	0.088	1.50
	ours, paired stop	0.25	**100**	0.042	0.151	0.109	0.178	0.065	1.75
	FreeDoM, full gradient	0.25	50	0.075	0.269	0.109	0.114	0.035	100 ∇
	FlowChef	0.25	**100**	**0.040**	**0.141**	0.109	0.219	0.057	100 ∇
	ours, paired stop	0.50	**100**	0.071	0.150	0.109	0.265	0.040	1.75
	FreeDoM, full gradient	0.50	50	0.065	0.258	0.109	0.124	0.015	100 ∇
	FlowChef	0.50	**100**	**0.040**	**0.119**	0.109	0.285	0.015	100 ∇
	ours, paired stop	1.00	100	0.105	0.109	0.109	0.370	0.008	2.75
	FreeDoM, full gradient	1.00	100	**0.028**	**0.107**	0.109	0.311	0.008	100 ∇
	FlowChef	1.00	100	0.063	0.118	0.109	0.565	0.005	100 ∇

Table 38: **Alzheimer's profile requests across fractions**, all five generators. Columns as in Table 16, on four noises per row and two for the MAISI baselines, with P(AD) of the volume classifier on the free sample, the steered arm and the opposite arm. The rows of the paired variant stop on the paired reading. Baseline step sizes are selected a posteriori from verifier measurements on a grid; 100 ∇: 100 gradient evaluations per sample. Generated samples at frozen step sizes are compared in Table 29. Bold: best; underlined: second best.