

SAM-NeuroAdapt: an atlas-guided, training-free pipeline for skull-stripping in infant and adult MRI

Charmarke Moussa Abdi

`charmake.moussaabdi@usmba.ac.ma`

Sidi Mohamed Ben Abdellah University

Ali Yahyaouy

Sidi Mohamed Ben Abdellah University

Fatima-Ezzahraa Ben-Bouazza

Université Mohammed VI des Sciences et de la Santé

Zakaria Said

Sidi Mohamed Ben Abdellah University

Research Article

Keywords: skull-stripping, brain extraction, Segment Anything Model, zero-shot segmentation, multi-atlas registration, reproducibility, benchmark bias, pediatric neuroimaging

Posted Date: October 27th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11333521/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

SAM-NeuroAdapt: an atlas-guided, training-free pipeline for skull-stripping in infant and adult MRI

Charmarke Moussa Abdi^{1*}, Ali Yahyaouy^{1*},
Fatima-Ezzahraa Ben-Bouazza^{2,3,4} and Zakaria Said⁵

^{1*}L3IA Laboratory, Faculty of Sciences Dhar El Mahraz, Sidi Mohamed
Ben Abdellah University, Fes, Morocco.

²BRET Lab, Mohammed VI University of Sciences and Health
(UM6SS), Casablanca, Morocco.

³LaMSN — La Maison des Sciences Numériques, Sorbonne Paris Nord
University, Paris, France.

⁴AIRA Lab, Faculty of Science and Technology, Hassan 1st University,
Settat, Morocco.

⁵Mathematical Analysis and Applications Laboratory, Sidi Mohamed
Ben Abdellah University, Fes, Morocco.

*Corresponding author(s). E-mail(s):

charmake.moussaabdi@usmba.ac.ma; ali.yahyaouy@usmba.ac.ma;

Contributing authors: fatimaezzahraa.benbouazza@uhp.ac.ma;
zakaria.said@usmba.ac.ma;

Abstract

Skull-stripping, the removal of non-brain tissue from magnetic resonance images, is an early step of most neuroimaging pipelines whose errors propagate to later measurements. However, current methods such as HD-BET, SynthStrip and LifespanStrip depend on task-specific training on large image collections. To test how much of this benefit survives without training, this study presents SAM-NeuroAdapt, a training-free pipeline. First, an age-matched template is selected from a library of 21 atlases and registered to the image. The registered atlas then supplies the prompts of a frozen Segment Anything Model 2, and the mask is refined by classical boundary correction. The pipeline was compared with HD-BET on three whole-head benchmarks, with multiplicity-adjusted p -values. On infants (Baby Open Brains, 71 visits, 51 participants), it achieved higher Dice (**0.920** versus **0.905**) and a lower average symmetric surface distance (ASSD; **2.07** versus **5.11** mm). On adults, no Dice difference was detected; ASSD was

lower on Mindboggle-101 ($n = 101$) and, under one of two adjustment procedures, on IBSR-20 ($n = 20$). Furthermore, an ablation showed that the age-matched atlas alone reached the infant Dice of the complete pipeline (**0.923** versus **0.920**), while the remaining stages reduced infant surface distances by 36–46%; without atlas guidance, 26–100% of cases failed. These findings indicate that atlas matching, rather than the foundation model, drives overlap accuracy and that a well-matched atlas is a competitive training-free baseline. Two evaluation pitfalls are also documented: a surface-distance definition that rewards over-segmentation and benchmarks distributed already skull-stripped.

Keywords: skull-stripping, brain extraction, Segment Anything Model, zero-shot segmentation, multi-atlas registration, reproducibility, benchmark bias, pediatric neuroimaging

1 Introduction

Skull-stripping, the separation of brain tissue from skull, scalp, dura and other non-brain structures in a magnetic resonance image, is an early step of most neuroimaging pipelines, and its errors propagate to the measurements derived downstream [6, 29, 31]. It is, for example, the entry point of the multi-cohort pipelines used for brain-age modelling [33]. Classical intensity-based tools such as BET [48], distributed with FSL [32], are fast but sensitive to acquisition protocol and population, which limits their robustness across datasets [6, 29].

This limitation motivated a generation of learned methods. HD-BET [31] trains a dedicated convolutional network on curated multi-sequence clinical data and serves as a comparison baseline in subsequent work [53]. SynthStrip [29] instead trains on synthetic images with randomized contrast and geometry, and its developmental variant extends this strategy to infant anatomy [34]. Most recently, LifespanStrip [53] coupled an extraction network with a learned registration module that produces a personalized atlas prior; trained end-to-end on 21,334 scans from 18 sites, it reports large effect sizes (Cohen’s $d > 1$ on 16 of 18 datasets) over six competing methods, including HD-BET and SynthStrip. All of these methods share one requirement: task-specific training on large collections of brain images.

The atlas prior at the centre of LifespanStrip nevertheless has a long history outside deep learning. In multi-atlas segmentation, labelled templates are registered to the subject and their labels are combined by STAPLE [56], joint label fusion [54], spatially varying fusion [12] or patch-based fusion [19]. The result depends on how closely the template matches the subject, which is why dedicated templates were built for the pediatric [22], neonatal [39] and fetal [24, 46] brain instead of reusing a single adult template.

Promptable foundation models supply a complementary ingredient: segmentation without task-specific training. The Segment Anything Model (SAM) [35] and its successor SAM2 [42] are trained on natural images and videos and segment the object designated by point or box prompts. Applied to medical images without adaptation, however, their accuracy varies widely with the imaging modality, the target structure

and the quality of the prompts [8, 9]. Taken together, these observations raise a question that, to our knowledge, has not been tested directly: how much of the benefit of an atlas prior survives *without any training*, when a registered, well-matched atlas supplies the prompts of a frozen foundation model?

This article addresses that question with SAM-NeuroAdapt, a training-free pipeline that selects a template from a lifespan atlas library, registers it to the subject [13], uses it to prompt a frozen SAM2, and refines the boundary with classical intensity-based methods [17]. A preliminary version was presented at a conference [2] and evaluated on adult scans only; the age-matching mechanism, the infant evaluation, the three benchmarks, the statistical analysis and the ablation are new. Our contributions are threefold. First, we show that the pipeline is competitive with a trained network (HD-BET) on three whole-head benchmarks spanning infants and adults (Section 3.1). Second, a stage-wise ablation attributes this result to atlas matching rather than to the foundation model (Section 3.3). Third, we document two evaluation pitfalls that can distort reported accuracy: a surface-distance definition that rewards over-segmentation (Section 4.3) and public benchmarks distributed already skull-stripped (Section 2.4). The benefits depend on dataset and metric, and we do not claim to match the absolute accuracy of LifespanStrip; our claim is that a well-matched atlas prior, applied without training, is a competitive baseline against a trained network.

2 Methods

2.1 Design principle

SAM-NeuroAdapt follows directly from the question above and is built around one constraint: *SAM2 is never fine-tuned, and no learned component is trained on the evaluation datasets*. The pipeline is therefore *training-free*; it is not free of benchmark-informed tuning, since the guard thresholds and three bounds of the normative growth band were set by inspecting evaluation-set behaviour during development (Section 2.6). All adaptation to a subject happens through classical operations: atlas selection, diffeomorphic registration [13], label fusion [56] and intensity-driven boundary correction [17]. Figure 1 summarizes these stages, which the following subsections describe in order. The implementation relies on SAM2 [42], PyTorch [40], ANTsX [51], NumPy [27], SciPy [52] and NiBabel [15]. The source code is publicly available at <https://github.com/nabolitains/SAM-NeuroAdapt>.

2.2 Atlas matching and registration

The first stage selects the atlas. It behaves differently depending on whether a per-subject age is available, because the two regimes use different criteria.

With a recorded age. The recorded age drives selection directly. Each candidate template receives a base score combining registration quality and image-atlas alignment, multiplied by a Gaussian age weight centred on the subject’s age ($\sigma = 1.5$ months below 6 months, 3 months to 18 months, 6 months beyond, floored at 0.60 so that a well-aligned template is never eliminated outright); a second term penalises

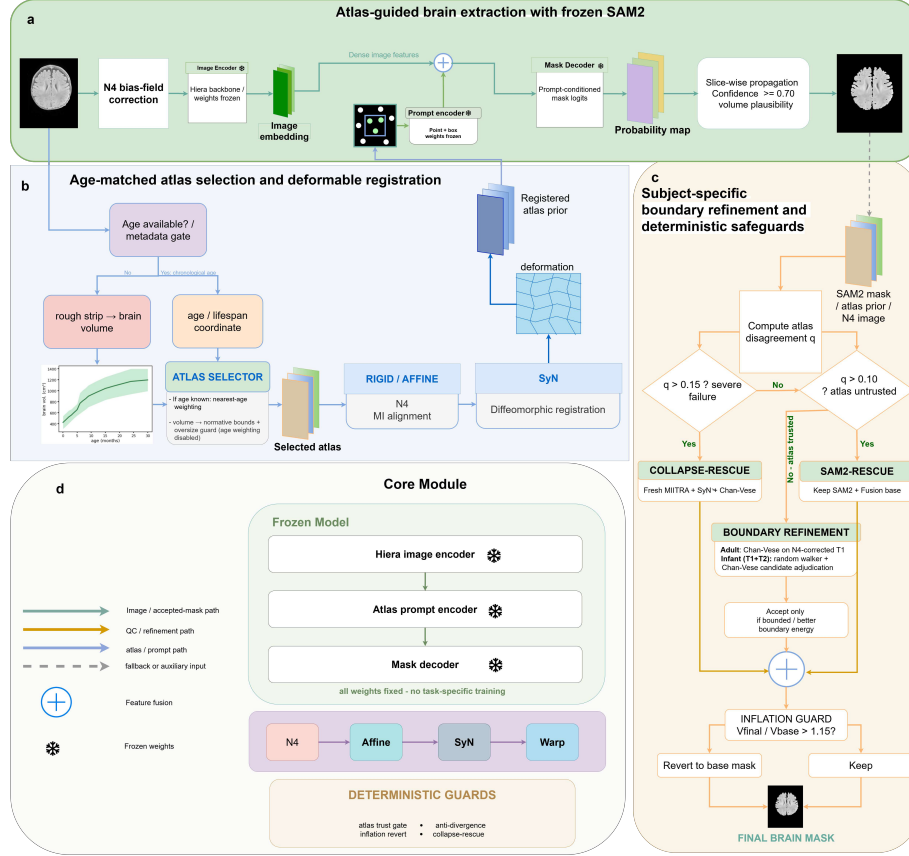


Fig. 1 Complete SAM-NeuroAdapt architecture. (a) Atlas-guided extraction: dense image features from the frozen SAM2 encoder meet point/box prompts derived from the registered atlas in the frozen mask decoder; slice-wise propagation and plausibility checks produce the candidate mask. (b) Atlas-prior construction, which precedes SAM2 — age-weighted template selection where an age is recorded, alignment-driven selection under a volume over-size guard where none is, with no age weighting in that regime (Section 2.2) — followed by rigid/affine alignment and ANTs SyN registration. (c) Boundary refinement and deterministic safeguards, routed by the atlas-disagreement score. (d) Summary of the frozen SAM2 components, the registration stages and the deterministic guards, with the key to arrows and symbols. All adaptation occurs at inference time; no component is fine-tuned on the evaluation datasets.

any template whose registered volume exceeds by more than 15% the upper normative bound for that age. Age is the dominant criterion in this regime.

Without a recorded age. The age-weighted selector is *disabled*: with no age to centre it on, the Gaussian weight is identically one and the over-size penalty, which also needs an age, does not apply. Image-atlas alignment quality becomes the primary criterion. The image contributes instead a coarse brain volume (intensity threshold, largest connected component after erosion) used for two narrower purposes: it sets the normative output bounds via the growth curve, and it applies a one-sided penalty

against a registered template whose volume exceeds $1.5\times$ the estimated subject volume — templates that are smaller or comparable are not penalised, which limits the influence of a noisy estimate. Neither adult benchmark carries a per-subject age either; adult images follow a separate procedure that scores each registered candidate by warped-volume proximity to a target, connected-component plausibility and intensity agreement, with no age term whatsoever. Selection is genuinely per subject: four distinct templates across the twenty IBSR-20 subjects, eight across Mindboggle-101.

Only BOBs therefore exercises age-based selection — all 71 sessions carry a recorded age (1–8 months) — while every other dataset here, including *both* adult benchmarks, runs without one. We claim age-based matching as validated end to end on the infant arm alone, and we do *not* claim to select the nearest template by an image-derived age anywhere.

The volume bounds used in both regimes come from a normative growth curve, which interpolates piecewise-linearly over anchors spanning the fetal period (18–40 weeks gestational age [24, 46]) and the postnatal period (0–60 months [25, 38, 39]). The volume it takes as input is measured *independently of any atlas registration*, so that the guard does not depend on the registration it is meant to check.

The candidates themselves are 21 external, publicly available templates spanning the lifespan — dHCP neonatal [39], UNC/BCP and UNC infant [30], the MNI-Infant series [22], and adult/elderly templates including SRI24 [44], MNI152Nlin2009c [23] via TemplateFlow [18] and MIITRA [43]; Figure 2a shows representative templates and Supplementary Table S6 lists the template families. No subject from any evaluation dataset is knowingly used as an atlas (external-templates-only policy); residual participant overlap with the template cohorts was not formally verified (Section 4.4).

Once selected, the template is registered with symmetric diffeomorphic normalization (SyN, [13, 14]) after N4 bias-field correction [50]; the deformed foreground/background masks become SAM2 point and box prompts (Figure 1a,b). Where more than one candidate is registered, SAM2’s response to each is combined by a trust-weighted vote and STAPLE [56]. SAM2 is used only in its promptable, single-image mode: never asked to segment without a prompt, never fine-tuned.

2.3 Boundary refinement and deterministic guards

The mask produced by the prompted model inherits the errors of the registration: it under-covers the true boundary in some regions and over-covers it in others. Three subject-specific refinements correct this without any training. *Adult intensity refinement*: a Chan-Vese active contour [17] evolves the atlas-seeded boundary using the subject’s own N4-corrected intensity; because it can diverge from a poor seed, its output is accepted only if it does not depart from that seed by more than a bounded fraction, and an *atlas-disagreement score* — the residual mismatch between deformed atlas and subject intensity after registration — gates a fallback: above 0.10, the raw SAM2+fusion mask is trusted instead. *Collapse-rescue*: a few subjects show severe disagreement (score > 0.15) that the trust gate does not resolve; for these a second, robust atlas (MIITRA [43]) is registered independently and accepted only if its boundary aligns better with subject intensity. *Pediatric refinement*: for infants with paired T1/T2 acquisitions, multi-modal candidates from a random-walker segmentation [26]

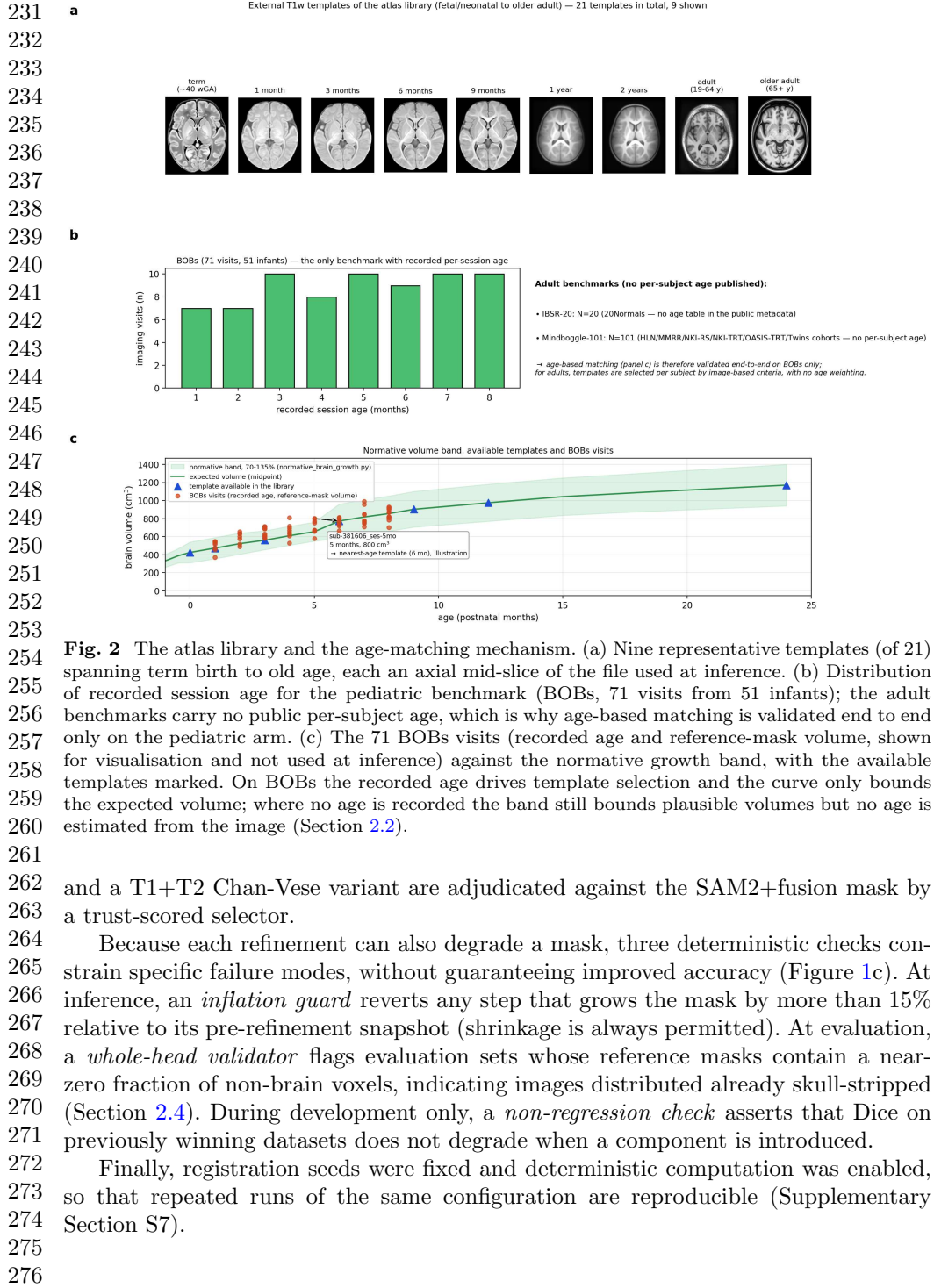


Fig. 2 The atlas library and the age-matching mechanism. (a) Nine representative templates (of 21) spanning term birth to old age, each an axial mid-slice of the file used at inference. (b) Distribution of recorded session age for the pediatric benchmark (BOBs, 71 visits from 51 infants); the adult benchmarks carry no public per-subject age, which is why age-based matching is validated end to end only on the pediatric arm. (c) The 71 BOBs visits (recorded age and reference-mask volume, shown for visualisation and not used at inference) against the normative growth band, with the available templates marked. On BOBs the recorded age drives template selection and the curve only bounds the expected volume; where no age is recorded the band still bounds plausible volumes but no age is estimated from the image (Section 2.2).

and a T1+T2 Chan-Vese variant are adjudicated against the SAM2+fusion mask by a trust-scored selector.

Because each refinement can also degrade a mask, three deterministic checks constrain specific failure modes, without guaranteeing improved accuracy (Figure 1c). At inference, an *inflation guard* reverts any step that grows the mask by more than 15% relative to its pre-refinement snapshot (shrinkage is always permitted). At evaluation, a *whole-head validator* flags evaluation sets whose reference masks contain a near-zero fraction of non-brain voxels, indicating images distributed already skull-stripped (Section 2.4). During development only, a *non-regression check* asserts that Dice on previously winning datasets does not degrade when a component is introduced.

Finally, registration seeds were fixed and deterministic computation was enabled, so that repeated runs of the same configuration are reproducible (Supplementary Section S7).

2.4 Datasets and eligibility criteria

The pipeline is evaluated on three benchmarks satisfying two criteria simultaneously: the images retain non-brain tissue (verified with the validator of Section 2.3), and the dataset is *not* part of HD-BET’s own training or evaluation protocol.

BOBs (Baby Open Brains; 71 imaging visits from 51 infants) is a pediatric whole-head T1 dataset with manually verified masks and known session-level age (1–8 months) [21]. **IBSR-20** ($n = 20$) is the “20 Normal Subjects” collection of the Internet Brain Segmentation Repository [16] ($1 \times 3 \times 1$ mm; distinct from the repository’s 18-subject collection), with manually guided adult masks used historically as an independent reference rather than as a training set for any compared method. **Mindboggle-101** ($N=101$) combines the cortical labeling protocol [36, 37] across seven acquisition cohorts, standardized here into a whole-head benchmark; our per-cohort analysis (Supplementary Section S5) reports nine groups because we resolve Extra-18 into the acquisitions it aggregates. We exclude NFBS [41], LPBA40 [47] and CC-359 [49] from the headline comparison because they are the test sets on which HD-BET was itself evaluated and is near-saturated (Dice of about 0.97–0.98 [31]); our losing LPBA40 result and our five-subject NFBS probe are nevertheless reported in Supplementary Section S1.

Pre-extracted benchmarks distort skull-stripping evaluation. The first criterion excludes more datasets than expected. Applying the whole-head validator — the fraction of head signal lying outside the reference mask, substantial (> 0.3) for a genuine whole-head acquisition — showed that iSeg-2017 [55] (non-brain fraction 0.000 ± 0.000 , $n = 10$) and the M-CRIB family [1] (0.022 ± 0.006 , $n = 10$; identical across the MCRIB, MCRIB2 and MCRIB_WM variants, which share images and masks) are distributed already skull-stripped or nearly so, against 0.76–0.88 for the benchmarks we retain. On such data “skull-stripping” measures how well a method reproduces an already-clean boundary rather than how well it removes non-brain tissue, and the Dice it yields (above 0.92 for every method we tested, including FSL BET) is not evidence of generalization. We exclude both families from every accuracy comparison and recommend that future benchmarks report this fraction alongside accuracy. The same validator (Supplementary Section S2) confirms SynthStripInfant (0.824) and FetalFMRI (0.927) are unambiguously whole-head, so the losses we report there are genuine.

Those two datasets carry no per-subject age and therefore stress-test the age-unknown regime: SynthStripInfant, a 16-subject pediatric subset of the collection released with SynthStrip [29], and FetalFMRI, fetal EPI/BOLD volumes with manually traced masks [45]. One caveat bounds the second: we evaluated all 16 SynthStripInfant subjects but only **three** FetalFMRI fetuses, because the twenty available BOLD frames belong to three fetuses and all frames of a fetus share one anatomical reference. FetalFMRI is a *three-fetus pilot*, not a benchmark evaluation, and every figure we report for it is aggregated per fetus. The pipeline is below the best baseline on both (Section 4.4).

2.5 Runtime and hardware

Being training-free removes the training cost but not the inference cost, and on this axis the pipeline is substantially slower than the baselines it is compared against. All results were produced on a single laptop GPU — an NVIDIA RTX 500 Ada with 4 GB of VRAM — with the SAM2.1 Hiera-Large checkpoint [42], so every run is strictly sequential. Median end-to-end time per subject was 510 s on BOBs (182^3 voxels), 662 s on IBSR-20 ($1 \times 3 \times 1$ mm anisotropic) and 1069 s on Mindboggle-101 (256^3); diffeomorphic registration dominates on IBSR-20 (median 350 s, 53% of total). Distributions are right-skewed, because collapse-rescue re-registers a second atlas from scratch: the worst BOBs subject takes 3930 s and the worst IBSR-20 subject 7431 s (Supplementary Table S7). HD-BET and SynthStrip infer in seconds on the same hardware, being single forward passes: we are two to three orders of magnitude slower, and not a drop-in replacement where throughput matters. Its cost is most defensible in the regime the ablation identifies — populations far from a trained model’s distribution, where that model may fail outright and no retraining budget exists.

2.6 Evaluation metrics and statistics

Calibration policy. Before the metrics are defined, the boundary between development and evaluation must be stated, because this pipeline is an assembly of gated heuristics rather than a trained model. The guard thresholds — trust gate 0.10, collapse-rescue 0.15, inflation guard 15% — were neither derived from theory nor calibrated on a held-out split: they were set by inspecting pipeline behaviour on IBSR-20 and Mindboggle-101 during development, i.e. on the same adult benchmarks used for evaluation, and the non-regression check is by construction a selection on benchmark performance. Reported adult numbers should therefore be read as *optimistically biased* to an extent we cannot quantify without a truly held-out cohort. What was *not* tuned: no ground-truth mask ever enters the pipeline at inference, the atlas library is external-templates-only, and the growth curve’s anchors come from independently published cohorts [25, 38]. BOBs was added after the thresholds had been fixed and none was re-tuned for it, which makes the infant result the least contaminated of the three, though not fully independent of development: one upper edge of the normative band was widened after inspecting a BOBs subject (Section 4.4).

Within this policy, accuracy is measured with the Dice similarity coefficient [20], the average symmetric surface distance (ASSD, mm) and the 95th-percentile Hausdorff distance (HD95, mm). Both distances are computed *surface-to-surface* — boundary voxels to boundary voxels, never to a filled volume — symmetrised over both directions, on a Euclidean distance transform sampled at the real voxel spacing so that anisotropic acquisitions such as IBSR-20 ($1 \times 3 \times 1$ mm) are handled correctly (Supplementary Section S8). All methods are scored with the same implementation, in line with published recommendations on segmentation metrics [10, 11]; Section 4.3 shows why this matters. We additionally report relative false-positive and false-negative volumes and a failure rate, the fraction of subjects with Dice < 0.80 . Differences against HD-BET are assessed per subject with a two-sided Wilcoxon signed-rank test [57], effect size $r = |Z|/\sqrt{N}$; because this literature reports Cohen’s d [53], we also give the

paired d_z and a 10,000-replicate bootstrap 95% confidence interval on the mean difference (Supplementary Section S4). Alongside means we report the *worst-case* subject, because a method that raises the average while lowering the floor is not unambiguously better — and our three benchmarks do not agree on this point.

Multiplicity. These tests are numerous: Table 1 reports nine of them (three benchmarks $\times$ three metrics). Under the global null and assuming independent tests, the chance that at least one uncorrected p falls below 0.05 is $1 - 0.95^9 \approx 37\%$, so uncorrected values would not support a claim about any *individual* comparison. We adjust across the family of nine and base every claim of significance on the adjusted value. The primary correction is Benjamini-Hochberg [4] (p_{adj}), which controls the false discovery rate and is the appropriate frame for a panel of related metrics. Because the choice of procedure is itself a degree of freedom, we also report the more conservative Holm-Bonferroni value [5] wherever the two disagree — exactly one of the nine cells (IBSR-20 ASSD: 0.032 under Benjamini-Hochberg, 0.086 under Holm). Raw p -values appear in the same table. The ablation tests of Table 2 address a different question and are not pooled into this family.

Non-independence in BOBs. A second statistical caveat concerns the infant benchmark. The 71 BOBs visits come from only 51 distinct infants (34 contributing one visit, 14 two, 3 three) [21], so a visit-level paired test violates independence [7]. We therefore report, alongside it, a 10,000-replicate bootstrap resampling *participants*, a single-visit-per-child sensitivity analysis and a per-child mean analysis; the reported Dice effect must hold under all three. These participant-level analyses were run for Dice only; the BOBs surface-distance tests are visit-level.

3 Results

3.1 SAM-NeuroAdapt is competitive with HD-BET on three fair benchmarks

Table 1 summarizes Dice, ASSD and HD95 on the three fair, whole-head benchmarks, with p -values adjusted for the nine comparisons the table contains (Section 2.6). One Dice comparison survives that adjustment: SAM-NeuroAdapt outperforms HD-BET on BOBs (Dice 0.920 ± 0.040 vs. 0.905 ± 0.014 , $\Delta = +0.0155$, $p_{\text{adj}} = 0.0024$). On IBSR-20 our Dice is nominally higher (0.939 ± 0.029 vs. 0.922 ± 0.041 , $\Delta = +0.0170$, 14/20 subjects $\geq$ HD-BET) but the difference does not survive adjustment on twenty subjects ($p = 0.048$, $p_{\text{adj}} = 0.062$; Holm 0.145); we therefore do *not* claim a win there, only a nominal advantage consistent with, but not established by, our data. The per-subject differences are consistent with that caution: the mean advantage is $+0.0170$ but the median only $+0.0090$, and three of the twenty subjects supply 87% of it — removing them leaves $+0.0027$. The margin is not a systematic advantage but is concentrated in three subjects on which both trained baselines score lowest (Section 3.2). On Mindboggle-101 there is no significant difference (Dice 0.942 ± 0.022 vs. 0.943 ± 0.010 , $\Delta = -0.0009$, $p_{\text{adj}} = 0.69$). We avoid the word “parity” there: a non-significant p -value is not evidence of equivalence [3], which would require a pre-specified margin and an equivalence test we did not perform.

415 *IBSR-20 is reported multi-run, not single-run.* Because twenty subjects make this
416 benchmark the most sensitive to run-to-run variation, it was run several times. Four
417 independent uninterrupted runs agree to within ± 0.0006 Dice (0.9389, 0.9380, 0.9395,
418 0.9387), and we report their mean rather than any single run: one earlier run had
419 suggested a larger margin (0.9428, $\Delta = +0.021$) that the multi-run estimate does not
420 support. A fifth run of the same configuration (interrupted and resumed) returned
421 0.8886 and was excluded; its cause, an evaluation artefact unrelated to the method, is
422 documented in Supplementary Section S7.

423 *The BOBs effect survives participant-level correction.* Because the 71 BOBs
424 visits come from 51 infants, we re-tested the infant result under three analyses
425 that respect participant clustering (Section 2.6). All three agree with the visit-
426 level test: participant-resampled bootstrap $\Delta = +0.0155$, 95% CI [0.0062, 0.0245]
427 (excludes zero); single-visit-per-child $\Delta = +0.0179$, $p = 0.0046$; per-child mean
428 $\Delta = +0.0178$, $p = 0.0014$. The naive visit-level $p = 0.0013$ is thus not an artefact of
429 pseudo-replication, though it is mildly anti-conservative.

430 *Surface distances are lower than HD-BET's on all three benchmarks, and four of*
431 *the six comparisons survive adjustment by a wide margin.* On BOBs, ASSD 2.07 vs.
432 5.11 mm and HD95 10.9 vs. 27.2 mm ($p_{\text{adj}} < 0.001$, $r = 0.87$); on Mindboggle-101,
433 2.42 vs. 3.21 mm and 16.2 vs. 22.4 mm ($p_{\text{adj}} < 0.001$, $r = 0.64$ and 0.77). These four
434 are the strongest results in the paper: adjusted p -values between 10^{-13} and 10^{-10} , so
435 they remain significant under either adjustment procedure, with large paired effect
436 sizes on any convention (Cohen's $d_z = -1.89$ and -2.29 on BOBs; -0.78 and -0.95
437 on Mindboggle-101). On IBSR-20 the ASSD advantage sits at the edge of the evidence
438 (3.13 vs. 3.92 mm, $p = 0.021$, $p_{\text{adj}} = 0.032$, not surviving Holm at 0.086) and no HD95
439 difference is detected ($p_{\text{adj}} = 0.73$).

440 The contrast between the two IBSR-20 metrics is instructive, because the per-
441 subject structure explains it: the Dice advantage has a mean twice its median and
442 collapses when three subjects are removed, whereas the ASSD advantage does not —
443 a mean reduction of 0.79 mm against a median reduction of 0.76 mm, in our favour on
444 16 of 20 subjects, still 0.33 mm after removing the three largest gains. A distributed
445 effect survives adjustment where a tail-driven one does not. SynthStrip trails both
446 methods on all nine comparisons by margins far larger than those separating us from
447 HD-BET: Dice 0.881–0.903 against our 0.920–0.942, ASSD 4.96–5.91 mm against our
448 2.07–3.13 mm.

449 *Our floor is not uniformly better than HD-BET's.* Our worst-case Dice is poorer
450 than HD-BET's on two of the three benchmarks (0.804 vs. 0.867 on BOBs, 0.839 vs.
451 0.903 on Mindboggle-101): a higher median but a longer lower tail. IBSR-20 is the
452 instructive exception — our floor is the better of the two (0.861 vs. 0.792), because
453 the worst HD-BET subject is one our atlas prior handles well. A practitioner choosing
454 on worst-case risk alone would not always choose ours (Supplementary Section S4).

455 These numbers depend critically on how the distance is defined: under a volume-
456 based definition, which measures the distance to the other mask's filled volume,
457 our pipeline appeared markedly *worse* than HD-BET on Mindboggle-101 (2.73 vs.
458 0.52 mm), and the ordering reverses once recomputed surface-to-surface identically
459 for all methods (Section 4.3).

460

Table 1 Results on the three fair, whole-head benchmarks (mean $\pm$ s.d.). Δ and the Wilcoxon signed-rank test compare SAM-NeuroAdapt (Ours) against HD-BET (the trained reference method). p_{adj} is Benjamini-Hochberg-adjusted across the nine tests of this table; **claims of significance are based on p_{adj} , not on p** . Significance: *** $p_{\text{adj}} < 0.001$, ** < 0.01 , * < 0.05 , n.s. not significant after adjustment; Holm-Bonferroni agrees on eight of the nine cells, the exception marked $\dagger$. Bold marks a result both better and significant after adjustment. ASSD and HD95 are in millimetres.

Metric	SynthStrip	HD-BET	Ours	Δ	p	p_{adj}	r
<i>BOBs, 71 visits from 51 infants</i>							
Dice	0.903 ± 0.015	0.905 ± 0.014	0.920 ± 0.040	+0.0155	0.0013	0.0024 (**)	0.38
ASSD	4.96 ± 0.46	5.11 ± 0.42	2.07 ± 1.53	-3.03	2×10^{-13}	<0.001 (***)	0.87
HD95	25.98 ± 1.59	27.23 ± 1.42	10.88 ± 6.71	-16.34	2×10^{-13}	<0.001 (***)	0.87
<i>IBSR-20, $N=20$ (adult)¹</i>							
Dice	0.902 ± 0.031	0.922 ± 0.041	0.939 ± 0.029	+0.0170	0.048	0.062 (n.s.)	0.44
ASSD	4.99 ± 1.06	3.92 ± 1.26	3.13 ± 1.24	-0.79	0.021	0.032 (*) [†]	0.51
HD95	26.60 ± 2.97	22.10 ± 3.28	21.87 ± 3.91	-0.23	0.73	0.73 (n.s.)	0.08
<i>Mindboggle-101, $N=101$ (adult)</i>							
Dice	0.881 ± 0.017	0.943 ± 0.010	0.942 ± 0.022	-0.0009	0.61	0.69 (n.s.)	0.05
ASSD	5.91 ± 0.60	3.21 ± 0.47	2.42 ± 0.81	-0.79	1×10^{-10}	<0.001 (***)	0.64
HD95	29.14 ± 2.63	22.39 ± 2.77	16.19 ± 5.73	-6.20	1×10^{-14}	<0.001 (***)	0.77

Δ = Ours - HD-BET; lower is better for ASSD and HD95. Surface distances are computed surface-to-surface at the real (anisotropic) voxel spacing, identically for all three methods; see Section 2.6. [†]IBSR-20 ASSD is the single cell on which the two multiplicity procedures disagree: significant under Benjamini-Hochberg (0.032), not under Holm (0.086). The IBSR-20 Dice margin survives neither, and we do not claim it.

¹IBSR-20 is reported over four independent, uninterrupted runs: each subject's value is its mean across those runs, the $\pm$ value is the between-subject standard deviation of these per-subject means, and the statistics are computed on them. The between-run standard deviation of the benchmark mean is far smaller (0.0006 Dice, 0.02 mm ASSD, 0.06 mm HD95). A fifth run, interrupted and resumed, was excluded (Supplementary Section S7). BOBs and Mindboggle-101 $\pm$ values are between-case standard deviations of a single run; BOBs cases are visits.

Figure 3 shows the full per-subject distribution underlying Table 1. On BOBs, HD-BET and SynthStrip are tightly clustered (median 0.906, s.d. ≤ 0.015) while SAM-NeuroAdapt has both a higher median (0.935) and a wider spread (s.d. 0.040), driven by the atlas-matching mechanism resolving cases that a fixed adult- or single-template baseline handles poorly. On IBSR-20 the lowest-scoring subject is the *same* one for both trained baselines (IBSR.2.4: HD-BET 0.792, SynthStrip 0.796), and our pipeline segments it at 0.945; the case is therefore one that both trained networks handle poorly rather than one that is intrinsically hard, and our own worst subject on this benchmark still scores 0.861 over the four runs. On Mindboggle-101 our Dice median is above HD-BET's (0.949 vs. 0.944) but our lower tail is longer (minimum 0.839 vs. 0.903), which is why no difference in means is detected; the same asymmetry appears in the surface panels, where our distribution is lower but more dispersed. Two of our three worst subjects there belong to the MMRR-21 cohort discussed in Section 4.

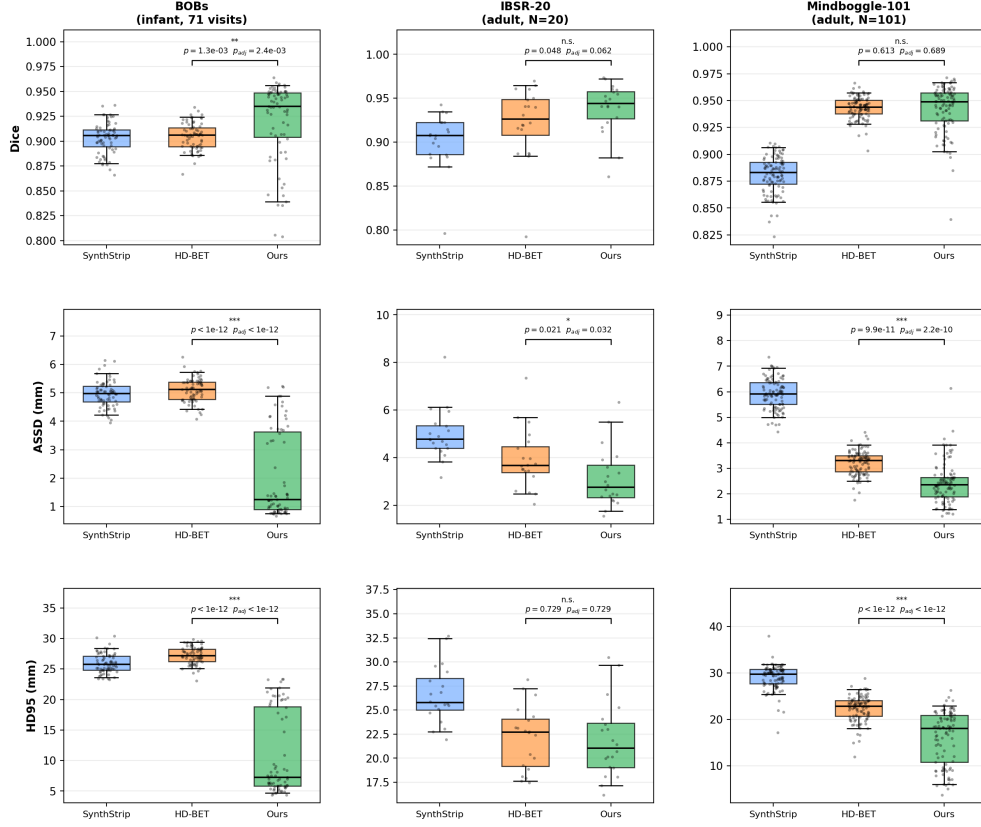


Fig. 3 Dice (top), ASSD (middle) and HD95 (bottom) for SynthStrip, HD-BET and SAM-NeuroAdapt (Ours) on the three benchmarks. Boxes: interquartile range and median; whiskers: 5th and 95th percentiles; points: individual cases (BOBs visits; for Ours on IBSR-20, per-subject means over four runs). Surface distances are computed surface-to-surface at the real voxel spacing, with the same implementation for all three methods. Brackets report the Wilcoxon test for Ours vs. HD-BET with the raw p and the Benjamini-Hochberg p_{adj} of Table 1; **the stars follow p_{adj} , never the raw p** , which is why the IBSR-20 Dice panel is marked n.s. although its raw p falls below 0.05; the IBSR-20 ASSD star does not survive Holm correction (Table 1).

3.2 Qualitative behaviour and the over-/under-segmentation trade-off

The distributions above do not show where the errors lie. Figure 4 therefore shows representative slices for one infant (BOBs, 3 months) and one adult subject (Mindboggle-101), with circles marking false negatives and false positives against ground truth. SynthStrip and HD-BET both tend to include extraneous dura or neck tissue, more visibly in the infant case, while SAM-NeuroAdapt’s atlas-guided boundary is tighter but occasionally excludes a small true-positive region.

To test whether this impression generalizes, we aggregated per-voxel false-positive and false-negative indicator volumes across all 192 evaluated volumes of the three benchmarks (71 BOBs visits, 20 IBSR-20 and 101 Mindboggle-101 subjects), each

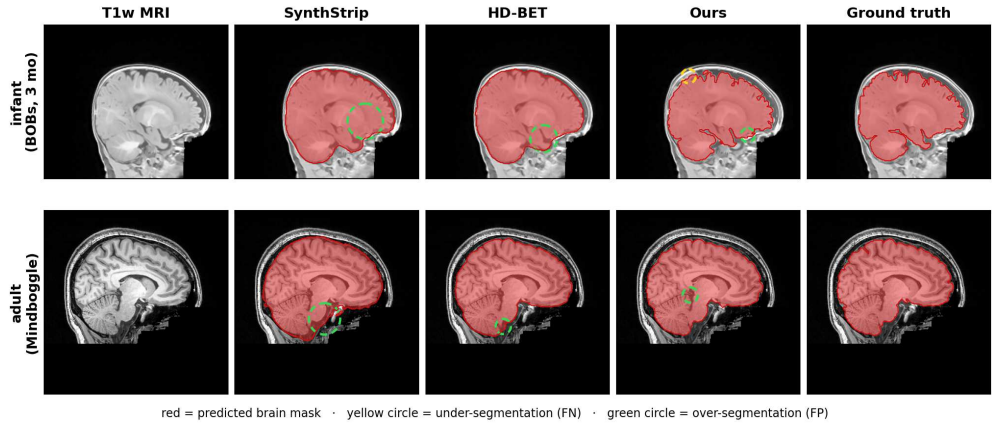


Fig. 4 Qualitative comparison across methods (columns) and ages (rows: infant BOBs 3 months, adult Mindboggle-101). Sagittal slices. The red overlay is the mask of the corresponding column (method output, or the reference mask in the last column); yellow dashed circles mark the largest under-segmented (false-negative) region and green dashed circles the largest over-segmented (false-positive) region relative to the reference.

rigidly recentered on its own ground-truth centroid, and projected the average error rate along three orthogonal axes (Supplementary Figure S2), following the visualization protocol of [53]. The aggregate is consistent with it: SAM-NeuroAdapt trades over-segmentation for under-segmentation. Its over-segmentation map is visibly darker than the baselines', particularly along the mid-sagittal dural/venous-sinus region where both show a bright ridge, while its under-segmentation map carries a diffuse central signal they do not. The same asymmetry holds in native voxel space without any resampling: relative false-positive volume 0.06–0.10 for SAM-NeuroAdapt vs. 0.12–0.27 for SynthStrip/HD-BET, relative false-negative volume 0.03–0.09 vs. near-zero. The two error modes differ in their consequences for surface metrics: the baselines' over-segmentation extends into neck and dura, anatomically distant from the reference boundary, and therefore contributes large distances — which is what drives their higher ASSD and HD95 in Table 1. Our residual under-coverage is thinner and spatially confined, and not caused by topological holes (0/192 volumes show any interior background pocket), so it costs less in surface distance than in overlap.

3.3 Ablation study

Table 1 establishes that the pipeline is competitive with a trained CNN. This section asks which component is responsible, by scoring the mask produced at each stage of the cascade against the same ground truth.

The age-matched atlas alone is not significantly worse than the complete pipeline on infants. Registering the age-matched template and taking its deformed mask directly — no foundation model, no refinement, no guards — yields Dice 0.9230 on BOBs against 0.9201 for the complete pipeline ($\Delta = -0.0029$, $p = 0.34$; the full chain is better on only 30 of 71 visits). The atlas alone therefore already exceeds HD-BET (0.9050) here. On the adult benchmarks the ordering reverses and the rest

of the chain contributes: 0.9007 on IBSR-20 versus 0.9388 complete ($\Delta = +0.0381$, $p = 0.0006$), 0.9145 on Mindboggle-101 versus 0.9418 ($\Delta = +0.0274$, $p < 0.001$). The intermediate stages place this adult gain after SAM2, in the refinement and adjudication steps rather than at the prompting stage; that reading rests on the subjects for which intermediate masks were retained ($n = 3$ on IBSR-20, $n = 17$ on Mindboggle-101, Supplementary Section S3) and is the weakest-supported statement in this section.

Boundary accuracy, unlike overlap, separates the atlas from the complete pipeline everywhere — including on infants. On BOBs, where no Dice difference is detected between the two configurations ($p = 0.34$), the complete pipeline nonetheless cuts both surface distances by more than a third: ASSD 2.07 vs. 3.26 mm ($p = 2 \times 10^{-9}$, $r = 0.71$) and HD95 10.88 vs. 20.12 mm ($p = 4 \times 10^{-12}$, $r = 0.82$). ASSD improves significantly on the adult benchmarks too (IBSR-20 $p = 0.003$, $r = 0.67$; Mindboggle-101 $p = 1.6 \times 10^{-4}$, $r = 0.43$). The interpretation is anatomical: a registered template mask is already in the right place, so Dice — dominated by the interior of a large, roughly convex structure — saturates, while its boundary still follows the template’s anatomy rather than the subject’s. The frozen model and the refinements together pull that boundary onto the subject’s own tissue interface — this comparison does not isolate their separate effects — a contribution a Dice-only ablation would have missed.

The atlas prevents failure rather than improving accuracy, and its necessity is graded by population. Disabling atlas guidance at the SAM2 stage while leaving every downstream step intact does not shift accuracy uniformly — it splits each cohort into subjects that are unaffected and subjects that collapse outright. The failure rate differs sharply across populations: 100% on infants (71/71 below Dice 0.80; without atlas prompts SAM2 over-segments into skull and neck), 45% on IBSR-20 (9/20), 26% on Mindboggle-101 (9/35). We did not measure whether this gradient follows distance from the foundation model’s training distribution, though that is a plausible reading. The decisive observation is what happens to the subjects that do *not* collapse (Figure 5c): compared with the complete pipeline *on those same subjects*, no Dice difference is detected — 0.9436 vs. 0.9481 on IBSR-20 ($n = 11$, Wilcoxon $p = 0.76$) and 0.9399 vs. 0.9388 on Mindboggle-101 ($n = 26$, $p = 0.73$). The subgroups are small and this is not a demonstration of equivalence, but where SAM2 succeeds without atlas prompts we detect no overlap benefit from atlas guidance; where it fails, that machinery is the difference between a usable mask and an unusable one. Combined with the infant result, the picture is consistent across all three benchmarks: *the atlas prior is a failure-containment mechanism, not an accuracy-improvement mechanism*. This is why Table 2 reports failure rate alongside mean Dice — a comparison of means alone would blur a 0% and a 26–100% failure regime.

4 Discussion

4.1 What the ablation establishes

The ablation of Section 3.3 is the most consequential result in this paper, and it is largely negative about our own design. We built the pipeline around the hypothesis

Table 2 Stage-wise ablation. “Fail.” is the fraction of cases with Dice < 0.80 (a mask that failed to be produced counts as a failure). “Dice (all)” averages over every case, a missing mask counting as zero; “Non-fail.” is the mean Dice over the non-failed cases only. “Atlas alone” uses no SAM2; the “SAM2, no atlas” rows disable atlas guidance at the SAM2 stage only, downstream stages remaining active. N differs between Mindboggle-101 rows (footnotes), which therefore do not all describe the same cases. ASSD and HD95 (mm) use the surface-to-surface definition of Table 1. p is an unadjusted Wilcoxon signed-rank test on Dice against the complete pipeline, exploratory and outside the nine-test family of Table 1.

Configuration	N	Dice (all)	Non-fail.	ASSD	HD95	Fail.	p
<i>BOBs (infant)</i>							
Atlas alone	71	0.9230 ± 0.0435	—	3.26	20.12	2.8%	0.34
SAM2, no atlas	71	0.5858	—	19.21 ¹	59.61 ¹	100%	< 0.001
Complete pipeline	71	0.9201 ± 0.0399	—	2.07	10.88	0%	—
<i>IBSR-20 (adult)</i>							
Atlas alone	20	0.9007 ± 0.0554	—	4.34	22.27	10%	0.0006
SAM2, no atlas	20	0.5999	0.9436	12.41 ¹	49.90 ¹	45%	0.021
Complete pipeline	20	0.9388 ± 0.0287^3	—	3.13	21.87	0%	—
<i>Mindboggle-101 (adult)</i>							
Atlas alone	76	0.9145 ± 0.0710	—	3.23	17.18	5.3%	< 0.001
SAM2, no atlas	35 ²	0.7083	0.9399	3.64 ¹	19.27 ¹	26%	< 0.001
Complete pipeline	76	0.9418 ± 0.0230	—	2.42	16.14	0%	—

¹A surface distance is undefined when no mask is produced, so the atlas-free rows average over the subjects that did yield a non-empty mask (BOBs 71/71, IBSR-20 18/20, Mindboggle-101 27/35). These figures therefore *understate* the degradation, the worst cases being the ones excluded — most visibly on Mindboggle-101, where the surviving 27 masks give an ASSD close to the atlas-alone row while the failure rate is 26%.

²The atlas-free Mindboggle-101 ablation uses a cohort-stratified subsample of 35 of the 101 subjects (all seven acquisition cohorts represented), for compute reasons; all other rows use the full available set.

³As in Table 1, the IBSR-20 complete-pipeline row uses each subject’s mean over the four uninterrupted runs, and its dispersion is between-subject like every other dispersion here (the between-run standard deviation of the mean is 0.0006). The Mindboggle-101 rows cover the 76 subjects for which ablation intermediates were retained, which is why they differ slightly from the $N = 101$ values of Table 1.

that a frozen foundation model, correctly prompted, would be the main source of accuracy; the data do not support that reading. Our summary is that *atlas selection is the dominant lever for placing the mask, and the refinement chain the dominant lever for placing its boundary* — and the former, the larger effect, is precisely the component requiring no learning at all.

What makes this more than an accounting exercise is that the atlas is needed *in proportion to how far the anatomy sits from what the foundation model has seen*. On infants, SAM2 without atlas guidance collapses on every subject, plausibly because infant contrast and proportions lie far from its natural-image training distribution [9, 42]; yet no Dice difference is detected between the atlas alone and the complete pipeline, so there the atlas is at once necessary and, for overlap, sufficient. On adults it collapses on a large minority, but shows no detectable Dice difference from the complete pipeline on the subjects it does not. That ordering — infant $\gg$ adult

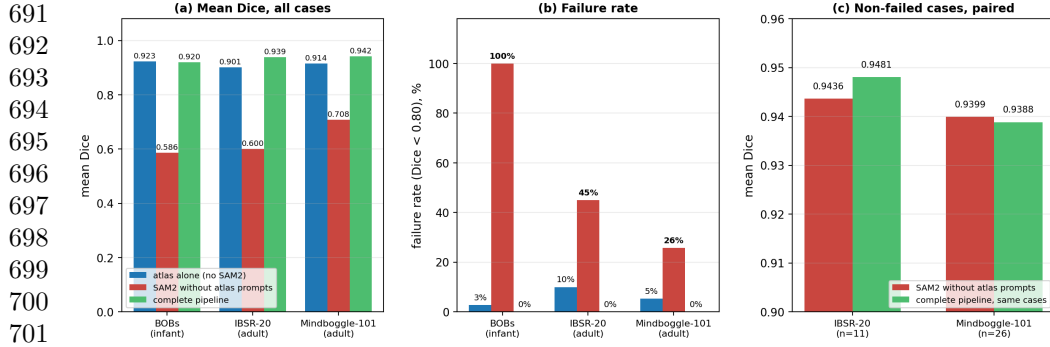


Fig. 5 Overlap and failure rates across ablation configurations. (a) Mean Dice for the atlas alone (no SAM2), SAM2 without atlas prompts and the complete pipeline. (b) Failure rate (Dice < 0.80; an unproduced mask counts as a failure): removing atlas guidance raises it from 0% to 26–100% (c) Adult cases on which SAM2 does *not* fail without atlas prompts, compared with the complete pipeline on the same cases (IBSR-20 $n = 11$, Mindboggle-101 $n = 26$): no Dice difference is detected. Overlap is not the whole story: the complete pipeline still attains significantly lower surface distances than the atlas alone on all three benchmarks (Table 2).

— is a testable prediction for other zero-shot medical segmentation pipelines, and it reframes the atlas as failure containment rather than accuracy. The one contribution a Dice-only ablation would have missed entirely is boundary accuracy: SAM2 and the refinements together reduce ASSD by about 36% and HD95 by about 46% even on infants. We retain the frozen model because it provides the seed those refinements operate on, but we no longer present it as the source of overlap accuracy.

4.2 Positioning relative to LifespanStrip

This reading also clarifies how the present work relates to its closest predecessor. LifespanStrip [53] and this work share the hypothesis that an age-appropriate atlas prior improves lifespan brain extraction, but occupy different points on the supervision-cost/accuracy trade-off. It trains extraction and registration networks jointly on 21,334 scans and reports Dice around 0.98 over six baselines; we reach comparable accuracy to a single trained baseline on the benchmarks we could evaluate fairly, with no task-specific training. We do not claim to match its absolute accuracy and would not expect to. Our contribution is complementary and, in light of the ablation, more specific than we first assumed: a substantial part of the benefit attributed to *learned* atlas-prior modules may be recoverable from the matching mechanism itself. Testing that directly — ablating LifespanStrip’s learned prior against a matched classical registration — is the natural next experiment, and one we cannot run without their training pipeline.

4.3 Pitfalls in skull-stripping evaluation

Beyond the method itself, the evaluation exposed two pitfalls that affect any comparison of skull-stripping methods. The first concerns surface distances. An earlier version of this analysis reported the opposite result on Mindboggle-101 — comparable Dice but a fivefold worse ASSD than HD-BET. That result came from a distance

measured from boundary voxels to the other mask’s filled volume rather than to its surface: whenever one mask contains the other, one direction of the symmetric average collapses to zero, so the metric rewards over-segmentation, the characteristic error mode of the trained baselines (Section 3.2). Recomputed surface-to-surface and identically for every method, our pipeline has lower ASSD and HD95 than HD-BET on all three benchmarks. This sensitivity of boundary metrics to their exact definition is a documented source of error in segmentation validation [10, 11].

The second pitfall concerns the benchmarks themselves: datasets distributed already skull-stripped yield a high Dice for every method and carry no information about the removal of non-brain tissue (Section 2.4). We therefore recommend reporting a surface metric alongside Dice, with its definition stated and computed identically for every method; checking it against an independent view of the errors, such as the error maps used here; validating benchmarks for whole-head content; and reporting multi-run variance whenever a margin is small relative to run-to-run variability.

4.4 Limitations and open problems

These results hold within limits that we state explicitly. *The matching mechanism is validated only where age is recorded.* On BOBs all 71 sessions carry an age, so that benchmark demonstrates atlas selection by chronological age. The age-unknown regime — alignment-driven selection with a volume-based over-size guard, and no age weighting at all — is what would have to carry the method on data without age metadata, and it is exercised only on SynthStripInfant ($n = 16$) and a three-fetus FetalFMRI pilot, where it is below the best baseline on both: Dice 0.798 vs. 0.965 for SynthStrip on the former, 0.455 vs. 0.631 for HD-BET on the latter, where $n = 3$ supports no benchmark-level conclusion and we draw none. It does resolve the pathological case we could verify end to end — a subject whose brain volume was roughly half that of the smallest available atlas, where replacing the over-sized template lifted Dice from 0.535 to 0.752 (+0.217) — which is precisely what the guard exists to catch. But the volume estimate is only moderately reliable ($r \approx 0.73$ with true volume), and the deeper point is structural: without an age there is no age weighting to apply. The *value* of a well-matched atlas is demonstrated; *obtaining* the match without age metadata is not.

Acquisition protocol can defeat a registration prior. A second limitation appears within the adult data. The null Dice result on Mindboggle-101 averages over seven acquisition cohorts that behave differently rather than indicating uniform equivalence: we are ahead by +0.015 on NKI-RS ($n = 22$) and +0.013 on NKI-TRT ($n = 20$), and behind by -0.030 on MMRR-21 ($n = 21$), the aggregate -0.0009 being the sum of these opposing effects (Supplementary Section S5). MMRR-21 — atypical orientation coding, intensity range roughly $\times 2000$ that of the other cohorts — remained below HD-BET’s per-cohort accuracy after the preprocessing and registration modifications we tested, including a contrast-invariant registration descriptor [28], which performed *worse* than plain intensity-based registration here, and an oracle upper bound on refinement-candidate selection, which produced no detected improvement. A CNN trained on varied protocols is more robust to this kind of intensity scaling than any registration-based prior we tested. Two further negative results constrain the design

783 space: multi-orientation SAM2 prompting and STAPLE-based fusion both failed to
784 rescue a mismatched atlas on our hardest case. The converging message is that the
785 dominant lever is atlas *selection*, not the machinery applied once a mismatched atlas
786 has been chosen.

787 *Residual dependencies and scope.* A last set of limitations concerns independence
788 from the evaluation data: the normative growth curve is not entirely independent of it.
789 Its twenty anchor *midpoints* come from independently published cohorts [25, 38, 39]
790 and no curve was fitted to our data, but three anchors’ *upper band edges* were widened
791 during development after inspecting individual subjects from datasets we go on to
792 use — one BOBs subject at month 3, one M-CRIB subject at month 0, the iSeg-
793 2017 cohort at month 6. The band therefore encodes a small amount of benchmark-
794 specific knowledge, the same kind of benchmark-informed tuning declared for the guard
795 thresholds in Section 2.6; its effect on BOBs is bounded, the widened month-3 edge
796 reducing that subject’s over-size penalty from a fraction of a percent to zero. Beyond
797 the external-templates-only policy we have not formally verified the absence of residual
798 participant-level leakage. Finally, our fair-benchmark criterion restricts us to three
799 datasets; larger multi-site validation, ideally against developmental SynthStrip [34],
800 would strengthen the claim.

801

802 5 Conclusion

803

804 We presented SAM-NeuroAdapt, a training-free brain-extraction pipeline that selects
805 an age-matched template from a lifespan atlas library, uses it to prompt a frozen founda-
806 tion model (SAM2), and refines the result under deterministic guards. On three
807 benchmarks selected to be fair to a trained CNN (HD-BET), with p -values adjusted
808 for multiplicity, it exceeds that baseline in Dice on the infant benchmark and in
809 boundary accuracy on both the infant and the larger adult benchmark, on which
810 no Dice difference was detected — a null result, not a demonstration of equivalence;
811 on the twenty-subject adult benchmark only the ASSD margin survives adjustment,
812 and not under the stricter Holm procedure. All of this with no task-specific training.
813 Our central finding, however, is a stage-wise ablation that reassigns credit within our
814 own architecture: on infants no Dice difference is detected between the age-matched
815 atlas alone and the complete pipeline; atlas guidance acts as failure containment for
816 the frozen foundation model rather than as a source of mean overlap; and the gain
817 over the bare atlas — lower surface distances on all three benchmarks, higher Dice
818 on adults — comes from the stages applied after registration, whose separate con-
819 tributions we do not isolate. Benchmark-informed tuning and slow inference bound
820 these conclusions. The reusable idea is therefore not the foundation model but the
821 atlas-matching mechanism: take the subject’s age, match the nearest available tem-
822 plate. Recovering that age from the image alone, when it is not recorded, remains
823 the open problem. Independently of any ranking, we document two evaluation pit-
824 falls — a surface-distance definition that rewards over-segmentation and benchmarks
825 distributed already skull-stripped — that can distort a reported margin.

826

827 **Supplementary information.** Additional file 1 (PDF) contains the supplementary
828 material: extended per-dataset results, including every dataset on which the pipeline

loses (S1); whole-head validation (S2); the full ablation cascade (S3); effect sizes and
worst cases (S4); Mindboggle-101 by cohort (S5); the atlas library (S6); the deter-
minism protocol and the excluded interrupted run (S7); surface-metric definitions and
error maps (S8); runtime (S9); and reproduction instructions (S10).

Acknowledgements. Not applicable.

Declarations

- **Funding:** The authors received no funding for this work.
- **Competing interests:** The authors declare no competing interests.
- **Relation to a previous publication:** This article extends a conference paper by the first three authors, “SAM-NeuroAdapt: A Robust MRI Preprocessing Pipeline for Atlas-Guided Brain Segmentation” [2], which introduced the atlas-prompted frozen-SAM2 premise and evaluated it on adult T1w scans from ADNI and OASIS. Carried over from that work are the premise itself, the system name and the use of a frozen SAM2 prompted by a registered atlas. New to this article are: age-based template selection and the 21-template lifespan library; the entire pediatric arm; all three evaluation benchmarks (BOBs, IBSR-20, Mindboggle-101 — none shared with the conference paper); the surface-to-surface ASSD and HD95 metrics; all inferential statistics; the stage-wise ablation; the deterministic guards; and the two evaluation pitfalls. No table, figure or numerical result is reproduced from the conference paper.
- **Ethics approval and consent to participate:** Not applicable: this study used only previously collected, publicly available, de-identified neuroimaging datasets cited in Section 2.4; no new human data were collected. Each dataset was acquired under the ethical approval and informed-consent procedures reported by its original investigators and was used in accordance with its data-use terms.
- **Consent for publication:** Not applicable.
- **Data availability:** All datasets used are available from their original sources, subject to their respective access and reuse conditions, cited in Section 2.4: BOBs [21], IBSR-20 [16], Mindboggle-101 [36, 37], SynthStripInfant [29], and FetalFMRI [45]. Atlas templates are drawn from public repositories (TemplateFlow [18], dHCP [39], UNC-BCP [30], MIITRA [43]) cited in Section 2.2.
- **Materials availability:** Not applicable.
- **Code availability:** The source code of the pipeline, of its evaluation and of the statistical analyses is available at <https://github.com/nabolitains/SAM-NeuroAdapt> under the MIT licence; the version used in this article (release v1.0.0) is archived at <https://doi.org/10.5281/zenodo.23287124>. The repository contains no imaging data, atlas template or model weight; these are obtained from their original providers.
- **Author contribution:** C.M.A. conceived the study, implemented the pipeline and its evaluation, designed and ran the experiments, performed the statistical analysis, produced the figures and tables, and wrote the manuscript. A.Y. supervised the work, contributed to the study design, and revised the manuscript. F.-E.B.-B. contributed to the methodology and to the interpretation of the results, and revised the manuscript. Z.S. contributed to the statistical analysis and revised the manuscript. All authors read and approved the final manuscript.

References

- [1] Alexander, B., Murray, A.L., Loh, W.Y., Matthews, L.G., Adamson, C., Beare, R., Chen, J., Kelly, C.E., Rees, S., Warfield, S.K., Anderson, P.J., Doyle, L.W., Spittle, A.J., Cheong, J.L.Y., Seal, M.L., Thompson, D.K.: A new neonatal cortical and subcortical brain atlas: the Melbourne Children’s Regional Infant Brain (M-CRIB) atlas. *NeuroImage*, 147, 841–851 (2017).
- [2] Moussa Abdi, C., Ben-Bouazza, F.-E., Yahyaouy, A.: SAM-NeuroAdapt: a robust MRI preprocessing pipeline for atlas-guided brain segmentation. In: *Artificial Intelligence and Cognitive Sciences for Emerging Technologies. Lecture Notes in Networks and Systems*, pp. 171–184. Springer Nature Switzerland (2025). https://doi.org/10.1007/978-3-032-11411-2_14
- [3] Altman, D.G., Bland, J.M.: Statistics notes: absence of evidence is not evidence of absence. *BMJ*, 311(7003), 485 (1995). <https://doi.org/10.1136/bmj.311.7003.485>
- [4] Benjamini, Y., Hochberg, Y.: Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57(1), 289–300 (1995). <https://doi.org/10.1111/j.2517-6161.1995.tb02031.x>
- [5] Holm, S.: A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2), 65–70 (1979).
- [6] Iglesias, J.E., Liu, C.-Y., Thompson, P.M., Tu, Z.: Robust brain extraction across datasets and comparison with publicly available methods. *IEEE Transactions on Medical Imaging*, 30(9), 1617–1634 (2011). <https://doi.org/10.1109/TMI.2011.2138152>
- [7] Lazic, S.E.: The problem of pseudoreplication in neuroscientific studies: is it affecting your analysis? *BMC Neuroscience*, 11, 5 (2010). <https://doi.org/10.1186/1471-2202-11-5>
- [8] Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. *Nature Communications*, 15, 654 (2024). <https://doi.org/10.1038/s41467-024-44824-z>
- [9] Mazurowski, M.A., Dong, H., Gu, H., Yang, J., Konz, N., Zhang, Y.: Segment anything model for medical image analysis: an experimental study. *Medical Image Analysis*, 89, 102918 (2023). <https://doi.org/10.1016/j.media.2023.102918>
- [10] Reinke, A., Tizabi, M.D., Baumgartner, M., et al.: Understanding metric-related pitfalls in image analysis validation. *Nature Methods*, 21(2), 182–194 (2024). <https://doi.org/10.1038/s41592-023-02150-0>

- [11] Taha, A.A., Hanbury, A.: Metrics for evaluating 3D medical image segmentation: analysis, selection, and tool. *BMC Medical Imaging*, 15, 29 (2015). <https://doi.org/10.1186/s12880-015-0068-x>
- [12] Asman, A.J., Landman, B.A.: Formulating spatially varying performance in the statistical fusion framework. *IEEE Transactions on Medical Imaging*, 31(6), 1326–1336 (2012).
- [13] Avants, B.B., Epstein, C.L., Grossman, M., Gee, J.C.: Symmetric diffeomorphic image registration with cross-correlation: evaluating automated labeling of elderly and neurodegenerative brain. *Medical Image Analysis*, 12(1), 26–41 (2008).
- [14] Avants, B.B., Tustison, N.J., Song, G., Cook, P.A., Klein, A., Gee, J.C.: A reproducible evaluation of ANTs similarity metric performance in brain image registration. *NeuroImage*, 54(3), 2033–2044 (2011).
- [15] Brett, M., Markiewicz, C.J., Hanke, M., Cote, M., Cipollini, B., McCarthy, P., et al.: nibabel. Software, Zenodo, <https://doi.org/10.5281/zenodo.591597> (2020).
- [16] Center for Morphometric Analysis, Massachusetts General Hospital: The Internet Brain Segmentation Repository (IBSR): “20 Normal Subjects” T1-weighted collection with manually-guided expert segmentations. <https://www.nitrc.org/projects/ibsr>. Accessed 1 Oct 2026.
- [17] Chan, T.F., Vese, L.A.: Active contours without edges. *IEEE Transactions on Image Processing*, 10(2), 266–277 (2001).
- [18] Ciric, R., Thompson, W.H., Lorenz, R., et al.: TemplateFlow: FAIR-sharing of multi-scale, multi-species brain models. *Nature Methods*, 19, 1568–1571 (2022). <https://doi.org/10.1038/s41592-022-01681-2>
- [19] Coupé, P., Manjón, E.V., Fonov, V., Pruessner, J., Robles, M., Collins, D.L.: Patch-based segmentation using expert priors: application to hippocampus and ventricle segmentation. *NeuroImage*, 54(2), 940–954 (2011).
- [20] Dice, L.R.: Measures of the amount of ecologic association between species. *Ecology*, 26(3), 297–302 (1945).
- [21] Feczko, E., Stoyell, S.M., et al.: Baby Open Brains: an open-source dataset of infant brain segmentations. *Scientific Data*, 12, 1423 (2025). <https://doi.org/10.1038/s41597-025-05404-y>
- [22] Fonov, V., Evans, A.C., Botteron, K., Almli, C.R., McKinsty, R.C., Collins, D.L.: Unbiased average age-appropriate atlases for pediatric studies. *NeuroImage*, 54(1), 313–327 (2011).

- 967 [23] Fonov, V.S., Evans, A.C., McKinstry, R.C., Almli, C.R., Collins, D.L.: Unbi-
968 ased nonlinear average age-appropriate brain templates from birth to adulthood.
969 NeuroImage, 47, S102 (2009).
970
- 971 [24] Gholipour, A., Rollins, C.K., Velasco-Annis, C., Ouaalam, A., Akhondi-Asl, A.,
972 Afacan, O., Ortinau, C.M., Clancy, S., Limperopoulos, C., Yang, E., Estroff,
973 J.A., Warfield, S.K.: A normative spatiotemporal MRI atlas of the fetal brain for
974 automatic segmentation and analysis of early brain growth. Scientific Reports, 7,
975 476 (2017).
976
- 977 [25] Gilmore, J.H., Shi, F., Woolson, S.L., Knickmeyer, R.C., Short, S.J., Lin, W., Zhu,
978 H., Hamer, R.M., Styner, M., Shen, D.: Longitudinal development of cortical and
979 subcortical gray matter from birth to 2 years. Cerebral Cortex, 22(11), 2478–2485
980 (2012).
981
- 982 [26] Grady, L.: Random walks for image segmentation. IEEE Transactions on Pattern
983 Analysis and Machine Intelligence, 28(11), 1768–1783 (2006).
984
- 985 [27] Harris, C.R., Millman, K.J., van der Walt, E.J., Gommers, R., Virtanen, P.,
986 Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N.J., et al.: Array pro-
987 gramming with NumPy. Nature, 585, 357–362 (2020). [https://doi.org/10.1038/
988 s41586-020-2649-2](https://doi.org/10.1038/s41586-020-2649-2)
- 989 [28] Heinrich, M.P., Jenkinson, M., Bhushan, M., Matin, T., Gleeson, F.V., Brady,
990 M., Schnabel, J.A.: MIND: Modality independent neighbourhood descriptor for
991 multi-modal deformable registration. Medical Image Analysis, 16(7), 1423–1435
992 (2012).
993
- 994 [29] Hoopes, A., Mora, J.S., Dalca, A.V., Fischl, B., Hoffmann, M.: SynthStrip: skull-
995 stripping for any brain image. NeuroImage, 268, 119431 (2022). [https://doi.org/
996 10.1016/j.neuroimage.2022.119431](https://doi.org/10.1016/j.neuroimage.2022.119431)
- 997 [30] Howell, B.R., Styner, M.A., Gao, W., Yap, P., Wang, L., Baluyot, K., Yacoub,
998 E., Chen, G., Potts, T., Salzwedel, A., et al.: The UNC/UMN Baby Connec-
999 tome Project (BCP): an overview of the study design and protocol development.
1000 NeuroImage, 185, 891–905 (2019).
1001
- 1002 [31] Isensee, F., Schell, M., Pflueger, I., Brugnara, G., Bonekamp, D., Neuberger,
1003 U., Wick, A., Schlemmer, H., Heiland, S., Wick, W., Bendszus, M., Maier-Hein,
1004 K.H., Kickingereder, P.: Automated brain extraction of multisequence MRI using
1005 artificial neural networks. Human Brain Mapping, 40(17), 4952–4964 (2019).
1006
- 1007 [32] Jenkinson, M., Beckmann, C.F., Behrens, T.E.J., Woolrich, M.W., Smith, S.M.:
1008 FSL. NeuroImage, 62(2), 782–790 (2012).
1009
- 1010 [33] Juglan, R., Ligerio, M., Carrero, Z.I., Rabasco Meneghetti, A., Lenz, T., Misera,
1011 L., Veldhuizen, G.P., Kuntke, P., Kitzler, H.H., Nebelung, S., Truhn, D., Kather,
1012

- J.N.: Generalizable and explainable deep learning for brain MRI: a multi-cohort evaluation of 3D architectures for age and sex prediction. *Brain Informatics*, 13, 31 (2026). <https://doi.org/10.1186/s40708-026-00316-y>
- [34] Kelley, W., Ngo, N., Dalca, A.V., Fischl, B., Zöllei, L., Hoffmann, M.: Boosting skull-stripping performance for pediatric brain images (developmental Synth-Strip). arXiv preprint arXiv:2402.16634 (2024).
- [35] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W., Dollár, P., Girshick, R.: Segment Anything. In: *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4015–4026 (2023).
- [36] Klein, A., Tourville, J.: 101 labeled brain images and a consistent human cortical labeling protocol. *Frontiers in Neuroscience*, 6, 171 (2012).
- [37] Klein, A., Ghosh, S.S., Bao, F.S., Giard, J., Häme, O., Stavsky, E., Lee, N., Rossa, B., Reuter, M., Chaibub Neto, E., Keshavan, A.: Mindboggling morphometry of human brains. *PLoS Computational Biology*, 13(2), e1005350 (2017).
- [38] Knickmeyer, R.C., Gouttard, S., Kang, C., Evans, D., Wilber, K., Smith, J.K., Hamer, R.M., Lin, W., Gerig, G., Gilmore, J.H.: A structural MRI study of human brain development from birth to 2 years. *Journal of Neuroscience*, 28(47), 12176–12182 (2008).
- [39] Makropoulos, A., Robinson, E.C., Schuh, A., Wright, R., Fitzgibbon, S., Bozek, J., Counsell, S.J., Steinweg, J., Vecchiato, K., Passerat-Palmbach, J., et al.: The developing human connectome project: a minimal processing pipeline for neonatal cortical surface reconstruction. *NeuroImage*, 173, 88–112 (2018).
- [40] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al.: PyTorch: an imperative style, high-performance deep learning library. In: *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32 (2019).
- [41] Puccio, B., Pooley, J.P., Pellman, J.S., Taverna, E.C., Craddock, R.C.: The pre-processed connectomes project repository of manually corrected skull-stripped T1-weighted anatomical MRI data. *GigaScience*, 5(1), s13742-016-0150-5 (2016). <https://doi.org/10.1186/s13742-016-0150-5>
- [42] Ravi, N., Gabeur, V., Hu, Y., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., et al.: SAM 2: Segment Anything in Images and Videos. arXiv preprint arXiv:2408.00714 (2024).
- [43] Ridwan, A.R., Niaz, M.R., Wu, Y., Bennett, D.A., Arfanakis, K.: Development and evaluation of a high performance T1-weighted brain template for use in studies on older adults. *Human Brain Mapping*, 42(6), 1758–1776 (2021).

1059 <https://doi.org/10.1002/hbm.25327>
1060
1061 [44] Rohlfing, T., Zahr, N.M., Sullivan, E.V., Pfefferbaum, A.: The SRI24 multichan-
1062 nel atlas of normal adult human brain structure. *Human Brain Mapping*, 31(5),
1063 798–819 (2010).
1064 [45] Rutherford, S., Sturmfels, P., Angstadt, M., Hect, J., Wiens, J., van den Heuvel,
1065 M.I., Scheinost, D., Sripada, C., Thomason, M.: Automated brain masking of
1066 fetal functional MRI with open data. *Neuroinformatics*, 20, 173–185 (2022). <https://doi.org/10.1007/s12021-021-09528-5>
1067
1068
1069 [46] Serag, A., Aljabar, P., Ball, G., Counsell, S.J., Boardman, J.P., Rutherford, M.A.,
1070 Edwards, A.D., Hajnal, J.V., Rueckert, D.: Construction of a consistent high-
1071 definition spatio-temporal atlas of the developing brain using adaptive kernel
1072 regression. *NeuroImage*, 59(3), 2255–2265 (2012).
1073
1074 [47] Shattuck, D.W., Mirza, M., Adisetiyo, V., Hojatkashani, C., Salamon, G.,
1075 Narr, K.L., Poldrack, R.A., Bilder, R.M., Toga, A.W.: Construction of a 3D
1076 probabilistic atlas of human cortical structures. *NeuroImage*, 39(3), 1064–1080
1077 (2008).
1078
1079 [48] Smith, S.M.: Fast robust automated brain extraction. *Human Brain Mapping*,
1080 17(3), 143–155 (2002).
1081
1082 [49] Souza, R., Lucena, O., Garrafa, J., Gobbi, D., Saluzzi, M., Appenzeller, S., Rit-
1083 tner, I., Frayne, R., Lotufo, R.: An open, multi-vendor, multi-field-strength brain
1084 MR dataset and analysis of publicly available skull stripping methods agreement.
1085 *NeuroImage*, 170, 482–494 (2018).
1086
1087 [50] Tustison, N.J., Avants, B.B., Cook, P.A., Zheng, Y., Egan, A., Yushkevich, P.A.,
1088 Gee, J.C.: N4ITK: improved N3 bias correction. *IEEE Transactions on Medical*
1089 *Imaging*, 29(6), 1310–1320 (2010).
1090
1091 [51] Tustison, N.J., Cook, P.A., Holbrook, A.J., Johnson, H.J., Muschelli, J., Devenyi,
1092 G.A., Duda, J.T., Das, S.R., Cullen, N.C., Gillen, D.L., et al.: The ANTsX ecosys-
1093 tem for quantitative biological and medical imaging. *Scientific Reports*, 11, 9068
1094 (2021). <https://doi.org/10.1038/s41598-021-87564-6>
1095
1096 [52] Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Courn-
1097 peau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., et al.: SciPy 1.0:
1098 fundamental algorithms for scientific computing in Python. *Nature Methods*, 17,
1099 261–272 (2020). <https://doi.org/10.1038/s41592-019-0686-2>
1100
1101 [53] Wang, L., Sun, Y., Seidlitz, J., et al.: A lifespan-generalizable skull-stripping
1102 model for magnetic resonance images that leverages prior knowledge from brain
1103 atlases. *Nature Biomedical Engineering*, 9(5), 700–715 (2025). <https://doi.org/10.1038/s41551-024-01337-w>
1104

[54] Wang, H., Suh, J.W., Das, S.R., Pluta, J.B., Craige, C., Yushkevich, P.A.: Multi-atlas segmentation with joint label fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(3), 611–623 (2013). 1105
1106
1107
1108

[55] Wang, L., Nie, D., Li, G., Puybureau, E., Dolz, J., Zhang, Q., Wang, F., Xia, J., Wu, Z., Chen, J., et al.: Benchmark on automatic six-month-old infant brain segmentation algorithms: the iSeg-2017 challenge. *IEEE Transactions on Medical Imaging*, 38(9), 2219–2230 (2019). 1109
1110
1111
1112
1113

[56] Warfield, S.K., Zou, K.H., Wells, W.M.: Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. *IEEE Transactions on Medical Imaging*, 23(7), 903–921 (2004). 1114
1115
1116
1117

[57] Wilcoxon, F.: Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6), 80–83 (1945). 1118
1119
1120
1121
1122
1123
1124
1125
1126
1127
1128
1129
1130
1131
1132
1133
1134
1135
1136
1137
1138
1139
1140
1141
1142
1143
1144
1145
1146
1147
1148
1149
1150