WASP: Weakly Aligned Spatiotemporal Pairs for Fetal Brain MRI–Ultrasound Learning

Francesco Correnti[1] **Gabriele Magrini**[1] **Marco Mistretta**[1]
Niccolò Biondi[2] **Pietro Pala**[1] **Alessandro Ramalli**[1]
Simona Fiori[3] **Andrew D. Bagdanov**[1] **Matteo Lenge**[3]

[1]University of Florence, Italy [2]University of Trento, Italy
[3]Meyer Children's Hospital IRCCS, Florence, Italy

{francesco.correnti, gabriele.magrini, marco.mistretta,
pietro.pala, alessandro.ramalli, andrew.bagdanov}@unifi.it
niccolo.biondi@unitn.it {simona.fiori, matteo.lenge}@meyer.it

miccunifi.github.io/WASP · github.com/miccunifi/WASP

Abstract

Magnetic Resonance Imaging (MRI) is widely regarded as the optimal sensor for fetal brain analysis due to its superior soft-tissue contrast and anatomical detail. However, its high cost and operational burden make it invasive and difficult to obtain at scale. Ultrasound (US), in contrast, is cheap, safe, and routinely acquired, and as a result it has produced substantially larger datasets and a growing ecosystem of pretrained models. This asymmetry raises a natural question: *Can we teach a US-only model to understand fetal MRI from only a limited set of examples?* The standard recipe, training a foundation model on subject-to-subject paired MRI–US scans, is not viable since no such paired fetal dataset is publicly available. In this paper we address this gap with Weakly Aligned Spatiotemporal Pairs (WASP), a framework that formulates cross-modal correspondence as an entropic Optimal Transport problem driven by clinical metadata — in particular, Gestational Age (GA) and diagnostic planes — enabling the fitting of a lightweight alignment module that lifts MRI representations into the US latent space, without fine-tuning the backbone. Empirically, WASP yields its largest gains when MRI is unseen by the model during pretraining (on USFM, GA estimation error drops from 21.9 to 17.4 days and standard plane classification accuracy climbs from 61.9% to 69.0%), while providing smaller, backbone-dependent refinements for backbones pretrained on both modalities (e.g., BioMedParse GA estimation error from 6.5 to 6.0 days and SAM-Med2d plane accuracy from 83.3% to 88.1%).

1 Introduction

Modern medical imaging is inherently multimodal, with each sensor offering a different compromise between accessibility, invasiveness, acquisition cost, and diagnostic resolution [1]. Modalities are complementary by design: those that are easier to acquire tend to be coarser, while those that yield finer anatomical detail are typically more expensive, slower, and harder to obtain at scale.

This trade-off is particularly sharp in fetal brain analysis. Magnetic Resonance Imaging (MRI) is the preferred sensor whenever a high-quality view of the developing brain is required, owing to its superior soft-tissue contrast and its ability to resolve subtle morphological structures throughout gestation [2]. Ultrasound (US), in contrast, suffers from speckle noise, acoustic shadowing, operator-dependent views, and offers limited fine-grained detail [3]. Yet US is cheap, safe, real-time, and

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

routinely acquired in standard prenatal screening. This has produced a marked asymmetry in what is publicly available to the research community: large fetal US cohorts and a growing ecosystem of US-pretrained foundation models [4, 5, 6, 7], versus scarce fetal MRI data and comparatively few MRI-specific backbones [8].

This asymmetry creates a concrete need for methods that bridge this cross-domain gap. The practical question is how to make efficient use of the few fetal MRI samples that do exist on top of the much larger pretrained US ecosystem, rather than training MRI-specific models from scratch. The standard recipe for this kind of cross-modal transfer relies on subject-to-subject paired acquisitions in which the same fetus is imaged with both modalities at the same point in time and an alignment objective is supervised on those pairs [9, 10]. However, in fetal imaging this recipe is not viable: no public dataset offers same-subject same-day MRI and US fetal scans, as clinical workflows rarely produce such pairs in the first place.

We thus introduce *Weakly Aligned Spatiotemporal Pairs (WASP)* to establish cross-modal correspondence between fetal MRI and fetal US without the need for same-day same-subject data. The key observation behind WASP is that fetal brain development follows a tightly constrained spatiotemporal trajectory: at a given Gestational Age, on a given standard diagnostic plane, the anatomy is morphologically consistent across subjects [11, 2] (see Appendix D for a detailed biological overview). Indeed, WASP exploits this regularity by using two pieces of metadata that are already routinely recorded in clinical practice, Gestational Age (time) and standard diagnostic plane (space), to define a soft mapping via entropic Optimal Transport [12]. The result is a pairing that is weak by construction, since it never claims subject identity, but at the same time anatomically meaningful since it ties together scans that depict the same developmental stage of the same brain structure.

This soft pairing is enough to learn a lightweight alignment on top of frozen backbones, which in turn enables effective knowledge transfer strategies, including domain adaptation onto previously unseen modalities and teacher-to-student distillation across model families and different modalities, all without fine-tuning the underlying backbone. WASP can teach US-only models how to interpret fetal MRI, and to refine multimodal foundation models that have already been exposed to both modalities.

The main contributions of this work are:

- we introduce WASP, a framework that turns routinely available fetal scan metadata into a soft cross-modal pairing via entropic Optimal Transport, removing the need for subject-to-subject paired MRI and US data;
- we show that WASP can adapt frozen US-only foundation models to fetal MRI without backbone fine-tuning, and that the same recipe further refines multimodal foundation models that have already seen both modalities, with consistent gains across different backbones; and
- we show that WASP improves transfer from teacher to students through a simple intra-modal distillation step.

2 Related Work

Unpaired Multimodal Representation Learning. Aligning representations across modalities without paired observations is a long-standing problem. One family of approaches synthesizes pseudo-paired data through unpaired image-to-image translation, typically regularized by cycle consistency [13, 14, 15, 16], but this pushes the burden onto a generative model that is prone to hallucination when the source modality is noisy. A second family operates directly in feature space, either by matching distributional moments such as covariance and mean [17] or by adversarial alignment of latent statistics [18]. Optimal Transport offers a principled alternative that compares full distributions [19, 20, 21], and entropic regularization [12, 22] makes it tractable at scale. What unifies these methods is that they treat the two domains as exchangeable point clouds and ignore side information about the samples themselves, which in clinical imaging is precisely what is most reliably available.

Prenatal Neuroimaging. Fetal neuroimaging is based on two complementary modalities. Ultrasound is fast, portable, and cheap, but is degraded by speckle noise, acoustic shadowing, and operator-dependent views [23]. MRI yields high soft-tissue contrast across the entire gestational window, at the cost of long acquisitions and limited availability [24]. This asymmetry is reflected in the modeling

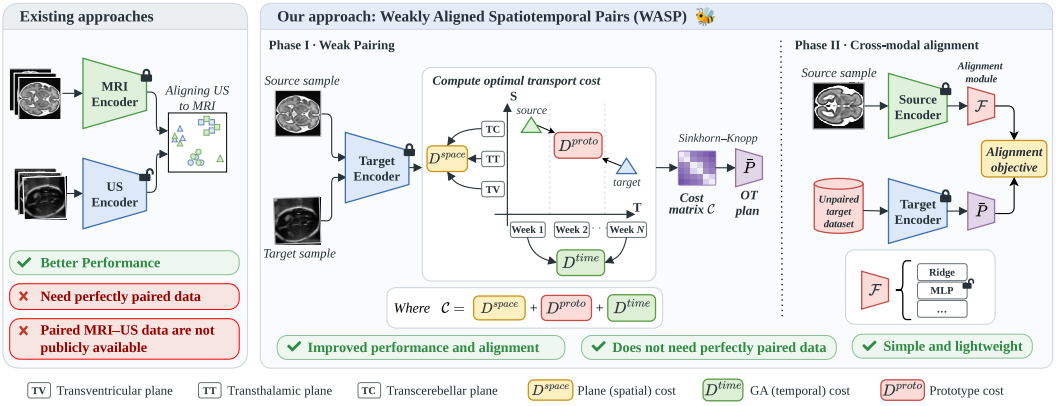

Figure 1: Motivation and overview of WASP. **Left**: Existing MRI–US alignment requires paired fetal scans from the same subject, but such data are not publicly available at scale. **Right**: WASP avoids exact pairing through two phases: *Phase I* builds an entropic Optimal Transport plan P from unpaired cohorts using commonly available metadata such as Gestational Age and diagnostic plane. *Phase II* uses P to project the target encoder space and train a lightweight alignment module $\mathcal{F}$ on the frozen source encoder's features.

landscape: the ultrasound side has seen rapid development of foundation models [4, 5, 6, 7, 25], while the MRI side is comparatively sparse, with MRI-CORE [8] as a recent generalist backbone. Multimodal medical foundation models such as SAM-Med2d [9] and BioMedParse [10] ingest several modalities at once, but treat fetal data as one of many heterogeneous sources with no specific cross-modal alignment mechanism.

Bridging US and MRI. MRI–US registration is technically demanding because the two modalities differ in contrast and artifact profile [26, 27]. Moreover, adult pipelines transfer poorly to fetal data due to motion and rapidly changing anatomy [28]. Recent fetal-specific work has begun to compare US and MRI phenotypes directly [29], but relies on rare near-same-day acquisitions that are not publicly available at scale. Several spatiotemporal atlases [2, 11, 30, 31] show that fetal brain growth follows a stereotyped, age-dependent trajectory with well defined population average anatomy for a given gestational week and standard plane. To the best of our knowledge, no prior work uses these population level regularities to define weak correspondences between fully independent US and MRI cohorts, rather than as registration targets for already paired scans. WASP fills this gap.

3 WASP: Weakly Aligned Spatiotemporal Pairs

In this section we describe how WASP transforms two unpaired unimodal cohorts into a weakly paired multimodal dataset. Let $\mathcal{S} = \{(x_i^{\mathcal{S}}, g_i^{\mathcal{S}}, p_i^{\mathcal{S}})\}_{i=1}^{N_s}$ and $\mathcal{T} = \{(x_j^{\mathcal{T}}, g_j^{\mathcal{T}}, p_j^{\mathcal{T}})\}_{j=1}^{N_t}$ denote the source and target cohorts, where N_s and N_t denote the number of samples (possibly unequal) in the source and target cohorts, respectively. Here, $x \in \mathbb{R}^d$ is a feature embedding extracted from a frozen backbone, $g \in \mathbb{R}$ is the Gestational Age in days, and $p \in \{\text{TV}, \text{TT}, \text{TC}\}$ indicates the standard diagnostic plane—Transventricular (TV), Transthalamic (TT), and Transcerebellar (TC)—as defined by the ISUOG guidelines [3]. The roles of source and target are interchangeable, and we describe the construction symmetrically.

The core intuition behind WASP is grounded in the regularity of fetal brain development. At a given GA, the brain structures visible on a standard diagnostic plane are anatomically consistent across different subjects: size, morphology, and spatial organization are constrained by the developmental stage [11, 2] (see Appendix D and Fig. 3). This inter-subject anatomical regularity makes GA and standard diagnostic plane jointly sufficient to define a meaningful notion of correspondence: two samples acquired at the same age on the same standard diagnostic planes, even from different subjects and different modalities, are expected to depict structurally similar content. Conversely, varying either the GA or the plane breaks this spatiotemporal regularity.

The pair (g, p) is therefore a meaningful spatiotemporal coordinate for cross-subject correspondence, while also being routinely recorded in clinical practice making it a widely available metadata. WASP turns this coordinate into a soft mapping between $\mathcal{S}$ and $\mathcal{T}$ through entropic Optimal Transport, in which a cost matrix encodes temporal distance, prototype distance, and a hard spatial consistency term. Both the Entropic OT, Sinkhorn–Knopp solver and barycentric projection are established tools; the contribution of WASP is a cost design that turns population-level developmental regularities into a definition of cross-modal correspondence when subject-level pairs are unavailable. Crucially, constructing this weak pairing establishes a common reference space, enabling transfer to unseen modalities, cross-modal knowledge distillation from teacher to student models, and improved performance on downstream clinical tasks. Fig. 1 shows the full WASP framework, divided in 2 distinct phases. The first phase represents the Weak Pairing, through which the entropic Optimal Transport plan P for the target space is computed. In the second phase, we use the previously obtained P to learn a cross-modal alignment module $\mathcal{F}$ from the source to the target space.

3.1 Phase I: Weak Pairing

The Spatiotemporal Cost Matrix. Given $\mathcal{S}$ and $\mathcal{T}$, we define a composite cost matrix $C \in \mathbb{R}^{N_s \times N_t}$ that penalizes associations between samples that are *temporally* or *spatially dissimilar*: the objective is to find a transport plan $P \in \mathbb{R}_+^{N_s \times N_t}$ that minimizes the total cost induced by C.

The *temporal cost* is defined as the absolute GA difference between any source–target pair:

$$D_{i,j}^{\text{time}} = \left| g_i^{\mathcal{S}} - g_j^{\mathcal{T}} \right| \tag{1}$$

For the *spatial cost*, we introduce a hard mismatch term:

$$D_{i,j}^{\text{space}} = \mathbb{1}(p_i^{\mathcal{S}} \neq p_j^{\mathcal{T}}) \tag{2}$$

where $\mathbb{1}(\cdot)$ is the indicator function.

Then, in order to exploit the developmental consistency of fetal brain along with the already well-structured target embedding space, we compute a representative prototype for each unique combination of diagnostic plane and gestational window. The core intuition here is that anatomical correspondence holds only within a narrow developmental window; therefore, pooling samples across all gestation would conflate structurally distinct configurations and yield unreliable references. We therefore partition the target dataset into temporal bins of fixed width (in days) and define the prototype $\mu(g, p)$ as the mean embedding of all target samples within that bin and plane. This aggregation smooths subject-specific variance and grounds the cost in the canonical anatomy of that plane at that developmental stage. Note that prototypes are computed on the target domain, which provides the established reference space into which source representations are aligned. We introduce an additional cost as the Euclidean distance between prototypes, which we call the *prototype cost*. Note that this cost is not metadata-free: prototypes are indexed by GA bin and plane, so this term encodes spatiotemporal information through the target embedding space:

$$D^{\text{proto}}{}_{i,j} = \left\| \mu(g_i^{\mathcal{S}}, p_i^{\mathcal{S}}) - \mu(g_j^{\mathcal{T}}, p_j^{\mathcal{T}}) \right\|_2 \tag{3}$$

Since D^{time} and D^{proto} operate on different scales, they are standardized to zero mean and unit variance (denoted $\mathcal{Z}$) before aggregation into the final cost matrix:

$$C_{i,j} = \mathcal{Z}(D_{i,j}^{\text{time}}) + \mathcal{Z}(D^{\text{proto}}{}_{i,j}) + D_{i,j}^{\text{space}} \tag{4}$$

To ensure non-negativity, we subtract the global minimum: $\tilde{C} = C - \min(C)$.

Soft Optimal Transport Plan. We seek an optimal transport plan $P \in \mathbb{R}_+^{N_s \times N_t}$ that minimizes the total alignment cost defined by the cost matrix $\tilde{C}$. Each source sample carries a developmental signature defined by its gestational age and diagnostic plane, and the goal is to match it to target samples that share an analogous developmental and anatomical context. The transport plan P encodes the strength of correspondence between source and target, where each entry $P_{i,j}$ quantifies the degree to which source scan i is matched with target scan j.

We follow the standard optimal transport formulation [22]. Since no scan is a priori more informative than another, we assign uniform probability weights to both domains via the marginal vectors

$\mathbf{a} = \frac{1}{N_s}\mathbf{1}_{N_s}$ and $\mathbf{b} = \frac{1}{N_t}\mathbf{1}_{N_t}$. These weights act as a fixed budget of correspondence assigned to each scan: every source scan must distribute exactly $1/N_s$ worth of matching across the target set, and every target scan must receive exactly $1/N_t$ worth from the source set. This conservation requirement defines the set of admissible transport plans $\mathcal{U}(\mathbf{a}, \mathbf{b}) = \{\mathcal{P} \in \mathbb{R}_+^{N_s \times N_t} \mid \mathcal{P}\mathbf{1}_{N_t} = \mathbf{a},\ \mathcal{P}^\top \mathbf{1}_{N_s} = \mathbf{b}\}$, which collects all non-negative matrices whose row sums equal $\mathbf{a}$ and whose column sums equal $\mathbf{b}$. Within this set, classical optimal transport seeks the plan that minimizes the linear objective $\langle \mathcal{P}, \tilde{C} \rangle = \sum_{i,j} \mathcal{P}_{i,j} \tilde{C}_{i,j}$. The resulting solution tends to be sparse, producing rigid one-to-one matches between scans. This rigidity is sub-optimal in our setting: biological variability across fetuses means that a single source scan rarely has one perfect counterpart, but rather several plausible matches drawn from neighboring gestational days and the same anatomical plane.

To accommodate this and improve robustness to acquisition noise, we adopt the entropic regularization framework of Cuturi [12], which augments the standard objective with the Shannon entropy of the plan $H(\mathcal{P})$. Maximizing $H(\mathcal{P})$ encourages each source scan to spread its matching budget across several anatomically and temporally compatible target candidates rather than concentrating it on a single counterpart, yielding a smoother routing whose strength is governed by the regularization coefficient $\varepsilon > 0$. Beyond improving robustness, this relaxation admits an efficient GPU-parallel solution via the Sinkhorn–Knopp algorithm (see Sec. B for implementation details and the ε hyperparameter search space). The final entropic Optimal Transport plan is therefore obtained as

$$P = \arg \min_{\mathcal{P} \in \mathcal{U}(\mathbf{a}, \mathbf{b})} \langle \mathcal{P}, \tilde{C} \rangle - \varepsilon H(\mathcal{P}). \tag{5}$$

Barycentric Projection. We then construct for each source embedding a counterpart that lives in the target embedding space, enabling direct cross-modal comparison. Concretely, we map each source embedding to its counterpart via a convex combination of target embeddings, weighted by the optimal transport plan P.

Since the rows of the raw plan P do not in general sum to one, we row-normalize it to obtain a valid stochastic matrix $\bar{P} : \bar{P}_{i,j} = P_{i,j} \,/\, (\sum_{k=1}^{N_t} P_{i,k})$. Restricting the weights to a convex combination ensures that each target embedding computed by the transport plan lies within the convex hull of the target set, preserving its geometric structure: if the raw entries of the transport plan P were used directly as weights, the sum of each row would equal $1/N_s$ rather than 1. Consequently the resulting counterpart embeddings would be scaled down and compressed toward the origin of the vector space.

The counterpart target embedding matrix is $\tilde{X}^{\mathcal{T}} = \bar{P} X^{\mathcal{T}}$, while source embeddings remain unchanged $\tilde{X}^{\mathcal{S}} = X^{\mathcal{S}}$.

3.2 Phase II: Cross-Modal Alignment

The resulting weakly paired embeddings $(\tilde{X}^{\mathcal{S}}, \tilde{X}^{\mathcal{T}})$ provide the necessary correspondences to learn a cross-modal alignment module $\mathcal{F}\colon \mathcal{S} \to \mathcal{T}$ that maps source representations into the target latent space. In effect, WASP converts a metadata coordinate (g, p) recorded routinely in clinical practice, into a soft correspondence dense enough to train such a mapping, without requiring paired acquisitions at any stage. Once $\mathcal{F}$ is learned, source-modality inputs can be projected into the target space, enabling cross-modal transfer and distillation. The specific parametrization of $\mathcal{F}$ and their downstream evaluations are described in Sec. 4.2.

4 Experiments

4.1 Experimental Setup

Dataset. We consolidate our dataset $\mathcal{D}$ as the union of the ultrasound dataset $\mathcal{D}^{\text{US}}$ and the MRI dataset $\mathcal{D}^{\text{MRI}}$, both consisting of 2D images acquired along the standard diagnostic planes TV, TT, and TC. We exclusively include neurotypical subjects in both the US and MRI cohorts.

The ultrasound dataset contains all the brain images from the "Multicentre benchmark dataset for comprehensive landmark-based fetal ultrasound biometry" [32], which aggregates FETAL_PLANES_DB [33], the HC18 head-circumference dataset [34] and an expanded UCLH cohort [35]; we add the remaining TV and TC images of FETAL_PLANES_DB to cover all three

planes. Since US datasets do not provide gestational age directly, g_i^{US} is estimated from whole-body biometric measurements with the Hadlock formulas [36]. We define the ultrasound dataset as $\mathcal{D}^{\text{US}} = \{(z_i^{\text{US}}, g_i^{\text{US}}, p_i^{\text{US}})\}_{i=1}^{N_{\text{US}}}$, where each triplet consists of an ultrasound image z_i^{US}, its corresponding Gestational Age g_i^{US}, and the ground truth anatomical plane label p_i^{US}. For additional details regarding the full composition of the dataset and the estimation of gestational age, see Appendix A.

The MRI dataset combines the Fetal Tissue Annotation and segmentation challenge (FeTa) [37], the Landmark Fetal Cerebellum (LFC) [38], and the Computational Radiology Laboratory (CRL) normative spatiotemporal atlas [2] datasets. We leverage already-reconstructed Super-Resolution Reconstruction (SRR) volumes, from which we extract 2D slices along the TV, TT, and TC planes (for further details, see Appendix A.2). We define the MRI dataset analogously to $\mathcal{D}^{\text{US}}$, i.e. $\mathcal{D}^{\text{MRI}} = \{(z_j^{\text{MRI}}, g_j^{\text{MRI}}, p_j^{\text{MRI}})\}_{j=1}^{N_{\text{MRI}}}$, where z_j^{MRI} denotes a 2D slice sampled from an SRR volume. In total, $\mathcal{D}$ comprises $4,056$ images from $1,973$ unique subjects, with $\mathcal{D}^{\text{US}}$ including $N_{\text{US}} = 3,369$ images from $1,744$ subjects, and $\mathcal{D}^{\text{MRI}}$ consisting of $N_{\text{MRI}} = 687$ images from 229 subjects.

To ensure cross-modal consistency, we apply the data preparation pipeline to all samples in $\mathcal{D} = \{\mathcal{D}^{\text{US}}, \mathcal{D}^{\text{MRI}}\}$ summarized in Fig. 2. Following standard neuroimaging protocols, each sample undergoes brain stripping to remove maternal tissue. This is achieved using a 2-layer CNN decoder head trained on a labeled subsets of the dataset. The resulting brain-stripped images are oriented such that the major and minor axes of the fetal brain ellipsoid align with the horizontal and vertical axes of the image plane, respectively. Finally, samples from each modality are resampled to their respective isotropic resolutions to ensure spatial consistency.

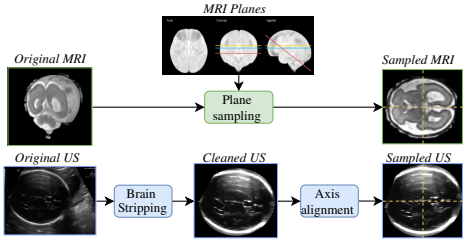

Figure 2: US and MRI data preprocessing. While in MRI we can directly get axis-aligned samples, for US images we also need to fix the orientation of the major and minor axes of the fetal brain.

We partition $\mathcal{D}$ into three disjoint, subject-level subsets: $\mathcal{D}_{\text{pair}}$ is the pairing set used to compute the cross-modal alignment module $\mathcal{F}$. $\mathcal{D}_{\text{train}}$ is the training set used to fit the Ridge regression and the SVM classifier. $\mathcal{D}_{\text{test}}$ is the hold-out evaluation set reserved exclusively for final inference and quantifying the generalization performance of the aligned representations. To tune hyperparameters, we perform subject-level k-fold cross-validation exclusively on $\mathcal{D}_{\text{train}}$ with $k = 3$.

Evaluation. We define θ_{GA} as the parameters of a Ridge regressor for GA prediction, and θ_{plane} as the weights of a linear SVM for plane classification. These parameters are fitted on top of the fixed embeddings during the training phase on $\mathcal{D}_{\text{train}}$. GA Regression is measured by Mean Absolute Error (MAE, in days, lower is better), while Diagnostic Plane Classification is evaluated by accuracy (Cls, %, higher is better). To select the optimal hyperparameters, we use a combined metric that balances performance gains across both tasks. For each fold, we compute a normalized score where a positive value indicates improvement over the zero-shot baseline $(\text{MAE}_0, \text{Cls}_0)$: $0.5 \cdot \frac{\text{MAE}_0 - \text{MAE}}{\text{MAE}_0} + 0.5 \cdot \frac{\text{Cls} - \text{Cls}_0}{\text{Cls}_0}$. We select the configuration that maximizes the median score across the three folds.

Models. To measure the generalization capabilities of the proposed approach, we evaluated WASP using five different unimodal and multimodal medical imaging analysis models. In particular, we chose USFM [4] and Echocare [5] as unimodal models for US analysis, MRI-CORE [8] as unimodal model for the MRI domain, with SAM-Med2d [9] and BioMedParse [10] for multimodal US and MRI approaches.

4.2 Experimental Protocol

We compare WASP with other spatiotemporal pairing strategies to benchmark its performance. Subsequently, $\mathcal{D}_{\text{train}}$ is used to fit the task-specific parameters θ_{GA} and θ_{plane}. Downstream models are trained on both raw and aligned features to directly quantify the benefit of the alignment. Finally, $\mathcal{D}_{\text{test}}$ serves as the held-out set for all reported evaluations. CCA and Ridge alignments are deterministic; MLP results are reported as mean $\pm$ std over 3 seeds.

4.2.1 Pairing Strategies

The pairing strategy determines how source samples $X^{\mathcal{S}}$ are matched to target samples $X^{\mathcal{T}}$ *before* learning the alignment module $\mathcal{F}$. We compare WASP against two alternative spatiotemporal approaches:

Hungarian Spatiotemporal Matching. A spatiotemporal pairing baseline that uses the same clinical metadata as WASP but enforces a strict one-to-one assignment. Within each diagnostic plane, a cost matrix is built from the absolute GA discrepancy between source and target samples, and the Hungarian algorithm computes the 1-to-1 bipartite matching that minimizes the total assignment cost.

Greedy Spatiotemporal Matching. A computationally cheaper baseline that uses the same clinical metadata as Hungarian but matches each source sample independently. Within each diagnostic plane, every source sample is paired with the target sample having the smallest absolute GA discrepancy. Unlike Hungarian matching, the same target sample may be assigned to multiple source samples.

4.2.2 Alignment Models

We learn $\mathcal{F}$ with CCA, Ridge, or MLP, and include CORAL as a pairing-free baseline.

Correlation Alignment (CORAL). As a pairing-free baseline, CORAL [17] aligns the second-order statistics of the raw source and target distributions without requiring any explicit point-to-point matching. The transformation $\mathcal{F}_{\text{coral}}$ is defined by $W_{\text{coral}} = \Sigma_{\mathcal{SS}}^{-1/2}\Sigma_{\mathcal{TT}}^{1/2}$ where $\Sigma_{\mathcal{SS}}$ and $\Sigma_{\mathcal{TT}}$ are the regularized covariance matrices of the source and target features, and the aligned representation is obtained via $\mathcal{F}_{\text{coral}}(X^{\mathcal{S}}) = X^{\mathcal{S}}W_{\text{coral}}$.

CCA Alignment. Given paired embeddings $(\tilde{X}^{\mathcal{S}}, \tilde{X}^{\mathcal{T}})$, Canonical Correlation Analysis [39] is used to derive a transformation matrix W that maps the source domain directly into the target embedding space. The algorithm computes the whitened cross-covariance $\mathbf{M} = \Sigma_{\mathcal{SS}}^{-1/2}\Sigma_{\mathcal{ST}}\Sigma_{\mathcal{TT}}^{-1/2}$ and its SVD, $\mathbf{U}\mathbf{\Lambda}\mathbf{V}^{\top}$, to find the optimal canonical directions. The final alignment weights are defined as $W = \Sigma_{\mathcal{SS}}^{-1/2}\mathbf{U}\mathbf{V}^{\top}\Sigma_{\mathcal{TT}}^{1/2}$. The transformation function $\mathcal{F}_{\text{cca}}(X^{\mathcal{S}}) = (X^{\mathcal{S}} - \bar{x}^{\mathcal{S}})W + \bar{x}^{\mathcal{T}}$, where $\bar{x}^{\mathcal{S}}$ and $\bar{x}^{\mathcal{T}}$ are the mean source and target embeddings, aligns source features by first projecting them into the canonical space and then expanding them into the target coordinate system.

Ridge Alignment. An affine transformation $\mathcal{F}_{\text{ridge}}$ is computed by minimizing the ridge-regularized mean squared error between paired embeddings, following the formula $W_{\text{ridge}} = \left((\tilde{X}^{\mathcal{S}})^{\top}\tilde{X}^{\mathcal{S}} + \lambda I\right)^{-1}(\tilde{X}^{\mathcal{S}})^{\top}\tilde{X}^{\mathcal{T}}$, where λ is the regularization hyperparameter. The resulting alignment is $\mathcal{F}_{\text{ridge}}(\tilde{X}^{\mathcal{S}}) = \tilde{X}^{\mathcal{S}}W_{\text{ridge}}$. For numerical stability, this closed-form solution is computed via SVD of the mean-centered source matrix (see Appendix B for the λ search grid).

MLP Alignment. To capture non-linear cross-modal alignments, we implement an MLP architecture $\mathcal{F}_{\text{mlp}}$ featuring a single hidden layer with GELU activations, optimized via AdamW with weight decay using the MSE loss over the paired embeddings (training hyperparameters and search space in Appendix B).

4.3 Experimental Results

WASP can be used in two qualitatively different scenarios. When the source modality is one the backbone has *never seen*, alignment acts as an unlocking mechanism that grafts the unseen modality onto a specialist representation. When the backbone has *already been exposed* to both modalities, alignment instead refines the existing cross-modal geometry. Our primary setting is MRI→US ($\mathcal{S} = \mathcal{D}^{\text{MRI}}$, $\mathcal{T} = \mathcal{D}^{\text{US}}$): it matches the motivating asymmetry, where scarce MRI is interpreted through the much larger US ecosystem, and the US cohort provides the larger target space to anchor the transport plan (961 vs. 480 pairing images). Tab. 1 reports this direction. Distillation and ablation studies below also use this direction. We then reverse source and target and report US→MRI in Tab. 2, which adds the MRI-native MRI-CORE [8].

Unlocking an unseen modality. USFM and Echocare are pretrained exclusively on ultrasound; MRI is, for them, a new modality. WASP delivers the largest relative gains observed in the paper: with WASP +Ridge, USFM's GA MAE drops from 21.9 to 17.4 days and plane classification climbs from 61.9% to 69.0%, while Echocare improves to 9.1 days and 90.5% accuracy. Other approaches, while

Table 1: Cross-modal alignment results in the **MRI→US** direction across four backbones: two unimodal US models (USFM, Echocare) and two multimodal models (SAM-Med2d, BioMedParse). We compare zero-shot transfer, CORAL, and three pairing strategies combined with CCA, Ridge, or MLP alignment. Performance is measured by GA regression error (↓) and diagnostic-plane classification accuracy (↑). CORAL, CCA, and Ridge are deterministic; MLP rows report mean ± std over 3 seeds. Best and second-best results are shown in bold and underlined, respectively.

MRI→US		Unimodal (US)				Multimodal			
		USFM [4]		Echocare [5]		SAM-Med2d [9]		BioMedParse [10]	
Pairing	Alignment	GA Reg.↓	Plane Cls.↑	GA Reg.↓	Plane Cls.↑	GA Reg.↓	Plane Cls.↑	GA Reg.↓	Plane Cls.↑
Zero-Shot	—	21.9	61.9	10.3	81.0	**7.7**	83.3	6.5	**85.7**
—	CORAL [17]	22.4	54.8	29.7	69.0	41.2	50.0	24.9	40.5
Greedy	CCA	25.5	66.7	16.8	76.2	21.0	83.3	14.5	78.6
	Ridge	25.2	66.7	10.4	83.3	10.1	85.7	6.1	**85.7**
	MLP	24.1±0.5	62.7±1.4	10.0±0.2	81.8±1.4	8.1±0.9	82.5±2.7	8.2±0.5	82.5±1.4
Hungarian	CCA	28.0	59.5	17.0	73.8	38.5	59.5	26.1	61.9
	Ridge	21.7	61.9	11.8	83.3	15.6	85.7	11.3	**85.7**
	MLP	20.2±0.1	61.9±2.4	11.3±0.9	80.2±1.4	11.1±0.7	**90.5±2.4**	12.6±1.8	84.1±2.7
WASP (Ours)	CCA	17.7	64.3	11.9	81.0	28.3	50.0	15.2	54.8
	Ridge	**17.4**	**69.0**	**9.1**	**90.5**	9.0	88.1	6.9	**85.7**
	MLP	19.9±0.1	65.1±1.4	9.7±0.5	84.1±1.4	7.9±0.3	89.7±2.7	**6.0±0.6**	84.1±1.4

still moderately improving performance, are less reliable than WASP, and in many instances degrade performance.

Refining a multimodal representation. Differently from the unimodal pretrained models, SAM-Med2d and BioMedParse have already been exposed to both modalities, so the question shifts from unlocking to refinement: can WASP improve on a representation that already crosses the modality gap? In the MRI→US direction, WASP +MLP achieves the best MAE on BioMedParse (6.0 ± 0.6 vs. 6.5 Zero-Shot) and lifts SAM-Med2d plane classification to $89.7 \pm 2.7\%$ (vs. 83.3% Zero-Shot), with Hungarian+MLP reaching $90.5 \pm 2.4\%$, while WASP +Ridge ties Zero-Shot on BioMedParse plane classification at 85.7%. With MLP alignment, WASP is the only pairing that improves one metric on each multimodal backbone while staying close to Zero-Shot on the other: Greedy+MLP degrades both metrics on both backbones, and Hungarian+MLP substantially increases the GA error.

Complementary direction: US→MRI. Reversing the roles maps the native US representations of USFM and Echocare toward the non-native MRI space. MRI-CORE is the mirrored setting, with US as the source modality, although it is already strong on US without any alignment (12.3 days, 98.1%), on par with the multimodal backbones. Starting from high Zero-Shot performance ($\geq 92.8\%$ plane accuracy on every backbone), this direction leaves little headroom: WASP +MLP never loses more than 1.4 days of GA MAE or 0.3 points of plane accuracy with respect to Zero-Shot, and improves plane classification on four of the five backbones (e.g., SAM-Med2d 97.3% vs. 96.2%), while most CCA and Ridge variants degrade performance.

Table 2: Cross-modal alignment results in the complementary **US→MRI** direction across five backbones: two unimodal US models (USFM, Echocare), one unimodal MRI model (MRI-CORE), and two multimodal models (SAM-Med2d, BioMedParse). We report gestational-age regression (GA Reg.; lower is better) and plane classification accuracy (Plane Cls.; higher is better), comparing Zero-Shot performance against different pairing and alignment strategies. CORAL, CCA, and Ridge are deterministic; MLP rows report mean ± std over 3 seeds. Best and second-best results within each backbone and task are shown in bold and underlined, respectively.

US→MRI		Unimodal (US)				Unimodal (MRI)		Multimodal			
		USFM [4]		Echocare [5]		MRI-CORE [8]		SAM-Med2d [9]		BioMedParse [10]	
Pairing	Alignment	GA Reg.↓	Plane Cls.↑	GA Reg.↓	Plane Cls.↑	GA Reg.↓	Plane Cls.↑	GA Reg.↓	Plane Cls.↑	GA Reg.↓	Plane Cls.↑
Zero-Shot	—	**16.0**	92.8	14.9	95.6	**12.3**	98.1	15.7	96.2	**12.1**	98.7
—	CORAL [17]	23.7	85.0	23.8	85.4	16.7	96.2	19.0	91.1	23.5	88.4
Greedy	CCA	23.2	89.7	21.5	92.2	24.9	93.5	26.1	93.0	18.6	95.8
	Ridge	22.8	90.1	20.6	92.6	25.3	93.5	22.0	96.4	17.9	98.5
	MLP	17.7±0.2	**93.2±0.1**	**14.7±0.1**	**96.2±0.2**	**12.3±0.2**	97.2±0.1	15.7±0.1	97.0±0.2	12.3±0.2	98.5±0.2
Hungarian	CCA	24.3	71.9	17.2	91.1	16.1	92.0	22.0	80.2	17.4	94.3
	Ridge	20.7	80.0	17.2	93.9	14.2	96.8	**14.2**	94.5	12.8	97.9
	MLP	17.6±0.2	90.2±0.1	15.5±0.2	**96.2±0.0**	13.7±0.2	97.4±0.1	15.7±0.1	96.2±0.4	12.8±0.2	98.8±0.1
WASP (Ours)	CCA	22.5	92.2	20.1	92.2	24.6	92.6	25.4	94.9	18.1	97.5
	Ridge	22.0	90.1	20.8	93.5	25.1	93.0	21.9	95.1	17.4	**98.9**
	MLP	17.4±0.3	93.0±0.4	14.9±0.3	**96.2±0.4**	12.8±0.2	**98.2±0.4**	15.3±0.3	**97.3±0.0**	12.4±0.3	98.4±0.4

Propagating gains to student backbones. The previous experiments treat each backbone independently. We now ask if the gains achieved by aligning a teacher's representation can help distillation between different models. Concretely, we train a transfer function $\mathcal{G}_{m\to m}$ that maps a student's raw embeddings in modality $m \in \{\text{US}, \text{MRI}\}$ to the teacher's improved embedding space in the same modality, parameterized as either a Ridge regressor or an MLP. When the teacher is a large multimodal model (BioMedParse or SAM-Med2d) whose representations are inherently richer, $\mathcal{G}_{m\to m}$ provides a lightweight mechanism to boost the student without modifying its backbone; when the teacher has been previously improved via WASP, this boost is amplified: GA regression improves for every teacher–student pair, and plane classification improves in three of the four.

Table 3: **MRI→US.** Teacher-student distillation results on the MRI test set. $\mathcal{G}_{m\to m}$ maps student embeddings to those of a teacher (raw or WASP-improved) using a Ridge regressor; the teacher alignment $\mathcal{F}$ is WASP +MLP. Raw rows are deterministic; + WASP rows report mean ± std over 3 seeds.

Student	Teacher	Setup	GA↓	Cls.↑
USFM [4]	Zero-Shot	-	21.9	61.9
	SAM-Med2d [9]	Raw	12.4	**76.2**
		+ WASP	12.0±0.8	72.2±1.4
	BioMedParse [10]	Raw	11.9	69.0
		+ WASP	**11.1**±0.2	71.4±0.0
Echocare [5]	Zero-Shot	-	10.3	81.0
	SAM-Med2d [9]	Raw	8.9	85.7
		+ WASP	8.8±0.1	89.7±1.4
	BioMedParse [10]	Raw	9.3	88.1
		+ WASP	**8.6**±0.4	**90.5**±2.4

We chain two stages: (i) the teacher is first improved via WASP cross-modal alignment, and (ii) $\mathcal{G}_{m\to m}$ is learned on $\mathcal{D}_{\text{pair}}$ by regressing the student's embeddings onto the teacher's improved representations. At inference time, student features X^m_{test} are passed through $\mathcal{G}_{m\to m}$ before the downstream evaluation. Tab. 3 reports the primary setting: teachers are improved with WASP +MLP in MRI→US, and $m = \text{MRI}$.

4.4 Ablations

We ablate the cost terms and the pairing strategy in the primary MRI→US direction; the MLP hyperparameter sensitivity and the MRI-CORE pairing ablation in both directions are in Appendices C.1 and C.2.

Cost Terms. We ablate the contribution of each term in the composite cost matrix $\tilde{C}$ by evaluating all $2^3 = 8$ subsets of $\{D^{\text{time}}, D^{\text{proto}}, D^{\text{space}}\}$ under Ridge alignment on USFM and MLP alignment on SAM-Med2d. Results are reported in Tab. 4. The results reveal a clear complementarity among the three terms. D^{time} alone drives strong gestational age regression but collapses plane classification, as it matches scans across anatomical planes. Conversely, D^{space} alone preserves plane classification but loses the temporal signal, sharply increasing the GA error. Adding plane information to D^{time} recovers plane accuracy, through D^{proto} on both backbones and through D^{space} on SAM-Med2d. The full cost matrix is the only combination that improves plane classification on both backbones while improving GA on USFM and keeping it within 0.2 days of Zero-Shot on SAM-Med2d, confirming that each component captures a distinct and complementary aspect of cross-modal correspondence.

Table 4: **MRI→US.** Ablation on cost terms. The combination of all three terms results in the most consistent gains. USFM uses Ridge alignment; SAM-Med2d reports one MLP run per subset.

Active cost terms			USFM [4]		SAM-Med2d [9]	
D^{time}	D^{proto}	D^{space}	GA↓	Cls.↑	GA↓	Cls.↑
✗	✗	✗	21.9	61.9	7.7	83.3
✗	✗	✓	35.7 ↑13.8	**76.2** ↑14.3	28.5 ↑20.8	90.5 ↑7.2
✗	✓	✗	20.3 ↓1.6	52.4 ↓9.5	8.9 ↑1.2	90.5 ↑7.2
✗	✓	✓	22.9 ↑1.0	50.0 ↓11.9	8.6 ↑0.9	**95.2** ↑11.9
✓	✗	✗	**12.6** ↓9.3	33.3 ↓28.6	**6.8** ↓0.9	64.3 ↓19.0
✓	✗	✓	12.9 ↓9.0	35.7 ↓26.2	8.6 ↑0.9	83.3
✓	✓	✗	15.4 ↓6.5	66.7 ↑4.8	8.2 ↑0.5	92.9 ↑9.6
✓	✓	✓	17.4 ↓4.5	69.0 ↑7.1	7.9 ↑0.2	88.1 ↑4.8

Pairing Strategies. To isolate the contribution of the spatiotemporal pairing, we replace the WASP transport plan with three metadata-agnostic strategies and evaluate them under the same Ridge, MLP and CCA alignment models used in the main experiments: Random, in which pairing assigns each source sample to a target sample drawn uniformly at random, providing a lower bound on alignment quality; Cosine pairing, which greedily matches each source sample to its nearest neighbor in the target embedding space by cosine similarity, without accounting for the global distribution of assignments; and OT-Cosine, the same entropic OT and barycentric projection as WASP but with the cosine distance between raw embeddings as the only cost, which separates the entropic relaxation

Table 5: **MRI→US.** Ablation of pairing strategies without spatiotemporal metadata. Random pairs each source sample with a target sample drawn uniformly at random, Cosine greedily matches each source to its nearest target by cosine similarity in raw embedding space, and OT-Cosine uses the WASP entropic OT with the cosine distance between raw embeddings as the only cost. All three underperform Zero-Shot in most settings. MLP rows report mean ± std over 3 seeds; for Random, the seed also changes the random pairing. WASP rows are reproduced from Tab. 1.

MRI→US		Unimodal (US)				Multimodal			
		USFM [4]		Echocare [5]		SAM-Med2d [9]		BioMedParse [10]	
Pairing	Alignment	GA Reg.↓	Plane Cls.↑	GA Reg.↓	Plane Cls.↑	GA Reg.↓	Plane Cls.↑	GA Reg.↓	Plane Cls.↑
Zero-Shot	—	21.9	61.9	10.3	81.0	**7.7**	83.3	6.5	**85.7**
Random	CCA	28.9	35.7	36.3	38.1	28.8	40.5	31.3	47.6
	Ridge	32.0	42.9	22.2	66.7	29.2	54.8	20.4	61.9
	MLP	32.3±3.1	45.2±10.4	20.2±0.7	81.0±6.3	21.8±1.6	71.4±2.4	18.6±1.8	63.5±5.0
Cosine	CCA	25.0	35.7	23.3	54.8	29.8	40.5	20.3	50.0
	Ridge	25.0	40.5	31.2	66.7	27.9	64.3	14.9	50.0
	MLP	26.0±0.2	36.5±1.4	20.8±1.4	84.9±2.7	16.2±0.1	84.9±1.4	11.8±0.8	57.9±5.0
OT-Cosine	CCA	25.3	59.5	22.5	57.1	44.1	38.1	22.5	40.5
	Ridge	24.6	64.3	16.6	81.0	22.8	71.4	12.8	59.5
	MLP	22.8±0.1	**69.0±0.0**	15.7±2.2	84.9±2.7	10.5±0.6	84.1±5.0	12.5±1.4	69.8±1.4
WASP (Ours)	CCA	17.7	64.3	11.9	81.0	28.3	50.0	15.2	54.8
	Ridge	**17.4**	**69.0**	**9.1**	**90.5**	9.0	88.1	6.9	**85.7**
	MLP	19.9±0.1	65.1±1.4	9.7±0.5	84.1±1.4	7.9±0.3	**89.7±2.7**	**6.0±0.6**	84.1±1.4

from the metadata-driven cost, resulting in a fully metadata-free baseline. Tab. 5 reports results for all four backbones. The three alternative pairing strategies degrade performance relative to Zero-Shot in most settings, and with Ridge and MLP alignment OT-Cosine has a higher GA error than WASP on every backbone and a lower plane accuracy in all but two settings (USFM and Echocare with MLP), indicating that most of the gain comes from the spatiotemporal cost rather than from the entropic relaxation alone.

5 Conclusions

We introduced WASP, a framework for establishing cross-modal correspondence between fetal MRI and fetal US without requiring subject-level paired acquisitions. By leveraging two routinely available pieces of clinical metadata, gestational age and standard diagnostic plane, as the cost function of an entropic Optimal Transport problem, WASP derives soft correspondences between independent unimodal cohorts. These correspondences are sufficient to learn a lightweight alignment module on top of frozen visual backbones. This alignment enables effective knowledge transfer across modalities. Applied to US-only foundation models, WASP allows representations learned exclusively from ultrasound to generalize to MRI, a previously unseen modality, substantially improving downstream performance. Additionally, WASP further refines representations of multimodal backbones already exposed to both MRI and US, yielding gains on at least one task for every multimodal backbone even in near-saturated settings. In the reverse US→MRI direction, WASP preserves the strong Zero-Shot representations of all five backbones. We additionally show that multimodal improvements can often be distilled from a stronger teacher into a lighter student via a single intra-modal regression.

Our results suggest that routinely collected clinical metadata can serve as a cross-modal supervision signal when subject-level pairs are unavailable, a principle that may apply beyond fetal imaging.

Limitations. Our framework lacks a direct comparison with paired-data methods, as no paired fetal US–MRI dataset is available; matched atlas pairs could serve as a proxy in future work. Using single MRI slices may also be suboptimal, and incorporating neighboring slices could better exploit the 3D context. US gestational ages are Hadlock-estimated from biometric landmarks rather than independently recorded, introducing uncertainty into the GA regression target and potentially sharing anatomical information with the input image. Finally, WASP is evaluated only on neurotypical fetuses, and its behavior in the presence of anatomical anomalies remains an open question.

Acknowledgments and Disclosure of Funding

Niccolò Biondi acknowledges the support of the EU Horizon Europe projects ELIAS (No. 101120237) and ELLIOT (No. 101214398), and of the FIS project GUIDANCE (No. FIS2023-03251).

This research was supported by the National Plan for Complementary Investments to the PNRR Hub Life Science-Digital Health (PNC-E3-2022-23683267, DHEAL COM). This publication reflects only the authors' view and the Italian Ministry of Health is not responsible for any use that may be made of the information it contains.

References

[1] Bobby Azad, Reza Azad, Sania Eskandari, Afshin Bozorgpour, Amirhossein Kazerouni, Islem Rekik, and Dorit Merhof. Foundational models in medical imaging: A comprehensive survey and future vision. *arXiv preprint arXiv:2310.18689*, 2023.

[2] Ali Gholipour, Caitlin K Rollins, Clemente Velasco-Annis, Abdelhakim Ouaalam, Alireza Akhondi-Asl, Onur Afacan, Cynthia M Ortinau, Sean Clancy, Catherine Limperopoulos, Edward Yang, et al. A normative spatiotemporal MRI atlas of the fetal brain for automatic segmentation and analysis of early brain growth. *Scientific reports*, 7(1):476, 2017.

[3] G Malinger, D Paladini, K K Haratz, A Monteagudo, GL Pilu, and IE Timor-Tritsch. ISUOG practice guidelines (updated): sonographic examination of the fetal central nervous system. Part 1: performance of screening examination and indications for targeted neurosonography. *Ultrasound in obstetrics & gynecology: the official journal of the International Society of Ultrasound in Obstetrics and Gynecology*, 56(3):476–484, 2020.

[4] Jing Jiao, Jin Zhou, Xiaokang Li, Menghua Xia, Yi Huang, Lihong Huang, Na Wang, Xiaofan Zhang, Shichong Zhou, Yuanyuan Wang, et al. USFM: A universal ultrasound foundation model generalized to tasks and organs towards label efficient image analysis. *Medical image analysis*, 96:103202, 2024.

[5] Hongyuan Zhang, Yuheng Wu, Mingyang Zhao, Zhiwei Chen, Rebecca Li, Fei Zhu, Haohan Zhao, Xiaohua Yuan, Meng Yang, Chunli Qiu, et al. A fully open and generalizable foundation model for ultrasound clinical applications. *arXiv preprint arXiv:2509.11752*, 2025.

[6] Ashwath Radhachandran, Vedrana Ivezić, Shreeram Athreya, Ronit Anilkumar, Corey W Arnold, and William Speier. US-JEPA: A joint embedding predictive architecture for medical ultrasound. *arXiv preprint arXiv:2602.19322*, 2026.

[7] Jakob Ambsdorf, Asbjørn Munk, Sebastian Llambias, Anders N Christensen, Kamil Mikolaj, Randall Balestriero, Martin G Tolsgaard, Aasa Feragen, and Mads Nielsen. General methods make great domain-specific foundation models: A case-study on fetal ultrasound. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 271–281. Springer, 2025.

[8] Haoyu Dong, Yuwen Chen, Hanxue Gu, Nicholas Konz, Yaqian Chen, Qihang Li, and Maciej A Mazurowski. MRI-CORE: a foundation model for magnetic resonance imaging. *arXiv preprint arXiv:2506.12186*, 2025.

[9] Junlong Cheng, Jin Ye, Zhongying Deng, Jianpin Chen, Tianbin Li, Haoyu Wang, Yanzhou Su, Ziyan Huang, Jilong Chen, Lei Jiang, et al. SAM-Med2D. *arXiv preprint arXiv:2308.16184*, 2023.

[10] Theodore Zhao, Yu Gu, Jianwei Yang, Naoto Usuyama, Ho Hin Lee, Sid Kiblawi, Tristan Naumann, Jianfeng Gao, Angela Crabtree, Jacob Abel, et al. A foundation model for joint segmentation, detection and recognition of biomedical objects across nine modalities. *Nature Methods*, 22:166–176, 2025. doi: 10.1038/s41592-024-02499-w.

[11] Ana IL Namburete, Bartłomiej W Papież, Michelle Fernandes, Madeleine K Wyburd, Linde S Hesse, Felipe A Moser, Leila Cheikh Ismail, Robert B Gunier, Waney Squier, Eric O Ohuma, et al. Normative spatiotemporal fetal brain maturation with satisfactory development at 2 years. *Nature*, 623(7985):106–114, 2023.

[12] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26, 2013.

[13] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2223–2232, 2017.

[14] Junhua Chen, Shenlun Chen, Leonard Wee, Andre Dekker, and Inigo Bermejo. Deep learning based unpaired image-to-image translation applications for medical physics: a systematic review. *Physics in Medicine & Biology*, 68(5):05TR01, 2023.

[15] Naomi Silverstein, Efrat Leibowitz, Ron Beloosesky, and Haim Azhari. Translation of fetal brain ultrasound images into pseudo-MRI images using artificial intelligence. *arXiv preprint arXiv:2504.02408*, 2025.

[16] Jianbo Jiao, Ana IL Namburete, Aris T Papageorghiou, and J Alison Noble. Self-supervised ultrasound to MRI fetal brain image synthesis. *IEEE Transactions on Medical Imaging*, 39(12):4413–4424, 2020.

[17] Baochen Sun, Jiashi Feng, and Kate Saenko. Correlation alignment for unsupervised domain adaptation. In *Domain adaptation in computer vision applications*, pages 153–171. Springer, 2017.

[18] Qi Dou, Cheng Ouyang, Cheng Chen, Hao Chen, and Pheng-Ann Heng. Unsupervised cross-modality domain adaptation of ConvNets for biomedical image segmentations with adversarial loss. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI)*, pages 691–697, 2018. doi: 10.24963/ijcai.2018/96.

[19] Nicolas Courty, Rémi Flamary, Devis Tuia, and Alain Rakotomamonjy. Optimal transport for domain adaptation. *IEEE transactions on pattern analysis and machine intelligence*, 39(9):1853–1865, 2017.

[20] David Alvarez-Melis and Tommi Jaakkola. Gromov-Wasserstein alignment of word embedding spaces. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 1881–1890, 2018.

[21] Pinar Demetci, Rebecca Santorella, Björn Sandstede, William Stafford Noble, and Ritambhara Singh. SCOT: single-cell multi-omics alignment with optimal transport. *Journal of computational biology*, 29(1): 3–18, 2022.

[22] Gabriel Peyré and Marco Cuturi. *Computational optimal transport: With applications to data science*. Now Foundations and Trends, 2019.

[23] Maria Chiara Fiorentino, Sara Moccia, Mariachiara Di Cosmo, Emanuele Frontoni, Benedetta Giovanola, and Simona Tiribelli. Uncovering ethical biases in publicly available fetal ultrasound datasets. *npj Digital Medicine*, 8(1):355, 2025.

[24] Farzan Vahedifard, Jubril O Adepoju, Mark Supanich, Hua Asher Ai, Xuchu Liu, Mehmet Kocak, Kranthi K Marathu, and Sharon E Byrd. Review of deep learning and artificial intelligence models in fetal brain magnetic resonance imaging. *World journal of clinical cases*, 11(16):3725, 2023.

[25] Fadillah Maani, Numan Saeed, Tausifa Jan Saleem, Zaid Farooq, Hussain Alasmawi, Werner Diehl, Ameera Mohammad, Gareth Waring, Saudabi Valappil, Leanne Bricker, and Mohammad Yaqub. FetalCLIP: A visual-language foundation model for fetal ultrasound image analysis. *npj Digital Medicine*, 2026. doi: 10.1038/s41746-026-02907-9.

[26] Bernhard Fuerst, Wolfgang Wein, Markus Müller, and Nassir Navab. Automatic ultrasound–MRI registration for neurosurgery using the 2D and 3D LC2 metric. *Medical image analysis*, 18(8):1312–1319, 2014.

[27] Fatemehzahra Darzi and Thomas Bocklitz. A review of medical image registration for different modalities. *Bioengineering*, 11(8):786, 2024.

[28] Bahram Marami, Seyed Sadegh Mohseni Salehi, Onur Afacan, Benoit Scherrer, Caitlin K Rollins, Edward Yang, Judy A Estroff, Simon K Warfield, and Ali Gholipour. Temporal slice registration and robust diffusion-tensor reconstruction for improved fetal brain structural connectivity analysis. *NeuroImage*, 156: 475–488, 2017.

[29] Madeleine K Wyburd, Nicola K Dinsdale, Vanessa Kyriakopoulou, Lorenzo Venturini, Robert Wright, Alena Uus, Jacqueline Matthew, Emily Skelton, Lilla Zöllei, Joseph Hajnal, et al. Cross-modality comparison of fetal brain phenotypes: Insights from short-interval second-trimester MRI and ultrasound imaging. *Human Brain Mapping*, 46(14):e70349, 2025.

[30] Camilo Calixto, Fedel Machado-Rivas, Davood Karimi, Clemente Velasco-Annis, Maria Camila Cortes-Albornoz, Onur Afacan, Simon K Warfield, Ali Gholipour, and Camilo Jaimes. Population atlas analysis of emerging brain structural connections in the human fetus. *Journal of Magnetic Resonance Imaging*, 60 (1):152–160, 2024.

[31] Wietske AP Bastiaansen, Melek Rousian, Anton HJ Koning, Wiro J Niessen, Bernadette S de Bakker, Régine PM Steegers-Theunissen, and Stefan Klein. The 4D human embryonic brain atlas: Spatiotemporal atlas generation for rapid anatomical changes. *Computerized Medical Imaging and Graphics*, page 102702, 2026.

[32] Chiara Di Vece, Zhehua Mao, Netanell Avisdris, Brian Dromey, Raffaele Napolitano, Dafna Ben Bashat, Francisco Vasconcelos, Danail Stoyanov, Leo Joskowicz, and Sophia Bano. A multicentre benchmark dataset for comprehensive landmark-based fetal ultrasound biometry. *Scientific Reports*, 16:17405, 2026. doi: 10.1038/s41598-026-47854-3.

[33] Xavier P Burgos-Artizzu, David Coronado-Gutiérrez, Brenda Valenzuela-Alcaraz, Elisenda Bonet-Carne, Elisenda Eixarch, Fatima Crispi, and Eduard Gratacós. Evaluation of deep convolutional neural networks for automatic classification of common maternal fetal ultrasound planes. *Scientific Reports*, 10(1):10200, 2020.

[34] Thomas LA van den Heuvel, Dagmar de Bruijn, Chris L de Korte, and Bram van Ginneken. Automated measurement of fetal head circumference using 2D ultrasound images. *PloS one*, 13(8):e0200412, 2018.

[35] Sophia Bano, Brian Dromey, Francisco Vasconcelos, Raffaele Napolitano, Anna L David, Donald M Peebles, and Danail Stoyanov. AutoFB: automating fetal biometry estimation from standard ultrasound planes. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 228–238. Springer, 2021.

[36] Frank P Hadlock, Russell L Deter, Ronald B Harrist, and SK Park. Estimating fetal age: computer-assisted analysis of multiple fetal growth parameters. *Radiology*, 152(2):497–501, 1984.

[37] Vladyslav Zalevskyi, Thomas Sanchez, Misha Kaandorp, Margaux Roulet, Diego Fajardo-Rojas, Liu Li, Jana Hutter, Hongwei Bran Li, Matthew J Barkovich, Hui Ji, et al. Advances in automated fetal brain mri segmentation and biometry: insights from the FeTA 2024 challenge. *Medical image analysis*, page 103941, 2026.

[38] Haifan Gong, Huixian Liu, Yitao Wang, Xiaoling Liu, Xiang Wan, Qiao Shi, and Haofeng Li. Fetal cerebellum landmark detection based on 3D MRI: Method and benchmark. *IEEE Journal of Biomedical and Health Informatics*, 2025.

[39] David R Hardoon, Sandor Szedmak, and John Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural computation*, 16(12):2639–2664, 2004.

[40] Reuben Dorent, Nazim Haouchine, Alexandra Golby, Sarah Frisken, Tina Kapur, and William Wells. Unified cross-modal medical image synthesis with hierarchical mixture of product-of-experts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(2):1641–1656, 2026. doi: 10.1109/TPAMI.2025.3616632.

[41] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2962–2971, 2017.

[42] Paul Govaert, Fabio Triulzi, and Jeroen Dudink. The developing brain by trimester. *Handbook of Clinical Neurology*, 171:245–289, 2020.

[43] Giulia Moltoni, Giacomo Talenti, and Andrea Righini. Brain fetal neuroradiology: a beginner's guide. *Translational pediatrics*, 10(4):1065, 2021.

[44] Pasko Rakic. Mode of cell migration to the superficial layers of fetal monkey neocortex. *Journal of Comparative Neurology*, 145(1):61–83, 1972.

[45] Lakshmi Subramanian, Maria Elisa Calcagnotto, and Mercedes F Paredes. Cortical malformations: lessons in human brain development. *Frontiers in cellular neuroscience*, 13:576, 2020.

[46] Je G Chi, Elizabeth C Dooling, and Floyd H Gilles. Gyral development of the human brain. *Annals of Neurology: Official Journal of the American Neurological Association and the Child Neurology Society*, 1 (1):86–93, 1977.

[47] Catherine Garel, Emmanuel Chantrel, Hervé Brisse, Monique Elmaleh, Dominique Luton, Jean-François Oury, Guy Sebag, and Max Hassan. Fetal cerebral cortex: normal gestational landmarks identified using prenatal MR imaging. *American Journal of Neuroradiology*, 22(1):184–189, 2001.

[48] Jessica Dubois, Julien Lefèvre, Hugo Angleys, François Leroy, Clara Fischer, Jessica Lebenberg, Ghislaine Dehaene-Lambertz, Cristina Borradori-Tolsa, François Lazeyras, Lucie Hertz-Pannier, et al. The dynamics of cortical folding waves and prematurity-related deviations revealed by spatial and spectral analysis of gyrification. *Neuroimage*, 185:934–946, 2019.

[49] Jessica Dubois, M Benders, C Borradori-Tolsa, Arnaud Cachia, F Lazeyras, R Ha-Vinh Leuchter, SV Sizonenko, SK Warfield, JF Mangin, and PS Hüppi. Primary cortical folding in the human newborn: an early marker of later functional development. *Brain*, 131(8):2028–2041, 2008.

[50] Wei-Xi Hu, Xin Zhan, Dan Lu, and Zhi-Qiang Li. Is choroid plexus growth altered in isolated ventriculomegaly on fetal neuro-ultrasound? *European Radiology*, 35(1):463–473, 2025.

[51] Thomas C Winter, Anne M Kennedy, Jan Byrne, and Paula J Woodward. The cavum septi pellucidi: why is it important? *Journal of Ultrasound in Medicine*, 29(3):427–444, 2010.

[52] Usha D Nagaraj, Maria A Calvo-Garcia, and Beth M Kline-Fath. Abnormalities associated with the cavum septi pellucidi on fetal MRI: what radiologists need to know. *American Journal of Roentgenology*, 210(5):989–997, 2018.

[53] Julia A Scott, Kia S Hamzelou, Vidya Rajagopalan, Piotr A Habas, Kio Kim, A James Barkovich, Orit A Glenn, and Colin Studholme. 3D morphometric analysis of human fetal cerebellar development. *The Cerebellum*, 11(3):761–770, 2012.

[54] Fei Liu, Zhonghe Zhang, Xiangtao Lin, Gaojun Teng, Haiwei Meng, Taifei Yu, Fang Fang, Fengchao Zang, Zhenping Li, and Shuwei Liu. Development of the human fetal cerebellum in the second trimester: a post mortem magnetic resonance imaging evaluation. *Journal of anatomy*, 219(5):582–588, 2011.

Appendix A Additional Dataset Details

$\mathcal{D}^{\text{US}}$ contains all the brain images from the "Multicentre benchmark dataset for comprehensive landmark-based fetal ultrasound biometry" [32], a curated collection derived from the Fetal Plane dataset [33], the head circumference dataset [34], and an expanded clinical cohort from University College London Hospital [35]. It is augmented with non-transthalamic brain images from [33] to cover all three standard diagnostic planes. Since gestational age targets are unavailable in the US datasets, g_i^{US} is estimated from whole-body biometric measurements using Hadlock's regression formulas [36].

Primary fetal biometric parameters - Biparietal Diameter (BPD), Occipitofrontal Diameter (OFD), Transverse Abdominal Diameter (TAD), Anteroposterior Abdominal Diameter (APAD), Femur Length (FL), Head Circumference (HC), and Abdominal Circumference (AC) — are extracted from the 2D spatial landmark annotations provided in [32].To maximize the reliability of gestational-age estimation for each image, we evaluate all valid multi-parameter Hadlock equations supported by the available biometrics and select the prediction with the minimum theoretical standard deviation.

$\mathcal{D}^{\text{MRI}}$ is built from three datasets: FeTa [37], LFC [38], and the Computational Radiology Laboratory (CRL) normative spatiotemporal MRI atlas [2]. Natively, fetal MRI is acquired as overlapping stacks of thick 2D T2-weighted slices, which are integrated via Super-Resolution Reconstruction (SRR) into a high-resolution 3D volume. We leverage already-reconstructed SRR volumes from these three sources.

A.1 Dataset Statistics

$\mathcal{D}^{\text{US}}$ comprises 3,369 images from $1{,}744$ unique subjects, covering a gestational age range of 86–284 days ($\approx$12.3–40.6 gestational weeks), with a mean of 163.3 ± 45.7 days and a median of 143 days.

The distribution of anatomical planes is strongly imbalanced: the transthalamic (TT) plane accounts for the large majority of the data ($2{,}756$ images, $\approx 81.8\%$), while the transcerebellar (TC) and transventricular (TV) planes represent 353 ($\approx 10.5\%$) and 260 ($\approx 7.7\%$) images, respectively. This imbalance reflects the clinical prevalence of TT acquisitions in routine fetal neurosonography practice. $\mathcal{D}^{\text{US}}$ is partitioned at the subject level into three disjoint subsets.

The assignment to $\mathcal{D}^{\text{US}}_{\text{pair}}$ is driven by temporal compatibility with the MRI cohort: only US images from 140 days (20 weeks) onward, i.e., close to or within the coverage window of $\mathcal{D}^{\text{MRI}}$ (147–266 days), are eligible to form cross-modal correspondences, since US images at younger gestational ages have no MRI counterpart against which a meaningful pairing can be established. Consequently, $\mathcal{D}^{\text{US}}_{\text{pair}}$ is enriched in older subjects (mean GA 176.7 days, range 140–284 days), whose ages overlap substantially with the MRI coverage window (147–266 days); the younger end of the US spectrum, with no MRI counterpart against which a meaningful transport cost can be computed, is instead directed to $\mathcal{D}^{\text{US}}_{\text{train}}$ and $\mathcal{D}^{\text{US}}_{\text{test}}$, where it contributes exclusively to downstream probe fitting and evaluation. Note that the two pair sets need not have identical GA distributions: the entropic optimal transport plan P is computed over their joint support and naturally up-weights source–target pairs that are close in gestational age, handling any residual distributional asymmetry between the modalities. This selection yields a natural 2:4:1 image ratio (pair:train:test), with the remaining non-pair images split 80/20 between train and test. Each source dataset contributes proportionally to all three splits.

$\mathcal{D}^{\text{US}}_{\text{pair}}$ contains 961 images from 613 subjects (28.5% of images, 35.1% of subjects; mean GA 176.7 ± 39.7 days; TT/TC/TV: 865/57/39), and is used exclusively to compute the alignment module $\mathcal{F}$. $\mathcal{D}^{\text{US}}_{\text{train}}$ comprises $1{,}934$ images from 905 subjects (57.4% of images, 51.9% of subjects; mean GA 157.0 ± 45.7 days; TT/TC/TV: $1{,}514$/239/181), used to fit the downstream task probes θ_{GA} and θ_{plane}. $\mathcal{D}^{\text{US}}_{\text{test}}$ contains 474 images from 226 subjects (14.1% of images, 13.0% of subjects; mean GA 161.8 ± 51.1 days; TT/TC/TV: 377/57/40), and serves as the held-out evaluation set.

$\mathcal{D}^{\text{MRI}}$ comprises 687 images from 229 unique subjects, spanning a gestational age range of 147–266 days ($\approx$21.0–38.0 gestational weeks), with a mean of 214.1 ± 30.6 days and a median of 217 days. Notably, the dataset is perfectly balanced across the three anatomical planes, with exactly 229 images per plane (TT, TC, TV), since one 2D slice per standard diagnostic plane is extracted from each subject's SRR volume via atlas-based registration (see Appendix A.2). The MRI cohort covers a more restricted and older gestational age range compared to $\mathcal{D}^{\text{US}}$, consistent with the clinical context in which fetal MRI is typically acquired in the second and third trimesters. $\mathcal{D}^{\text{MRI}}$ is partitioned by

first assigning ≈70% of subjects to $\mathcal{D}^{\text{MRI}}_{\text{pair}}$, then splitting the remaining 30% 80/20 between $\mathcal{D}^{\text{MRI}}_{\text{train}}$ and $\mathcal{D}^{\text{MRI}}_{\text{test}}$. As with the US cohort, each source dataset contributes proportionally to all three splits. $\mathcal{D}^{\text{MRI}}_{\text{pair}}$ contains 480 images from 160 subjects (69.9% of images and subjects; mean GA 216.1 ± 29.7 days; 160 images per plane), and is used to derive the alignment module $\mathcal{F}$. $\mathcal{D}^{\text{MRI}}_{\text{train}}$ comprises 165 images from 55 subjects (24.0% of images and subjects; mean GA 212.3 ± 31.2 days; 55 images per plane), used to train the downstream probes. $\mathcal{D}^{\text{MRI}}_{\text{test}}$ contains 42 images from 14 subjects (6.1% of images and subjects; mean GA 198.9 ± 35.3 days; 14 images per plane), reserved for final evaluation.

A.2 MRI Slice Extraction

Extracting anatomically consistent 2D slices from 3D fetal MRI volumes requires a principled reference frame, since the standard diagnostic planes (TT, TC, TV) are defined anatomically rather than by a fixed voxel axis. To this end, each subject's SRR volume is registered to its age-matched template from the Computational Radiology Laboratory (CRL) spatiotemporal fetal brain atlas [2], which provides normative 3D references spanning the gestational age range covered by $\mathcal{D}^{\text{MRI}}$. Registration is performed as a rigid-body alignment maximizing normalized mutual information between the subject volume and the corresponding atlas template, yielding a transformation that brings each subject into a common anatomical coordinate system.

The three canonical planes are then identified directly in atlas space and back-projected to the subject volume, from which the corresponding 2D slices are sampled. All extracted slices were subsequently reviewed with the support of an expert radiographer to verify anatomical consistency and manually fix and tune the extraction to comply to the ISUOG guidelines [3].

A.3 Preprocessing and Embedding Extraction

All samples in $\mathcal{D}$ undergo a standardized preprocessing pipeline to ensure cross-modal spatial consistency. Each image is first resampled to the minimum isotropic resolution of its modality: 0.05 mm/px for ultrasound and 0.8 mm/px for MRI. After resampling, the foreground region is cropped and resized to the input resolution expected by each backbone, preserving the original aspect ratio and padding with zeros where necessary. This aspect-ratio-preserving resize ensures that identical anatomical structures subtend the same apparent size in the backbone's patch grid regardless of acquisition parameters. For each backbone, image-level feature vectors are obtained by average-pooling over all patch tokens produced by the Vision Transformer encoder. Formally, given a sequence of N_p patch tokens $\{t_k\}_{k=1}^{N_p}$:

$$z = \frac{1}{N_p} \sum_{k=1}^{N_p} t_k \in \mathbb{R}^d, \tag{6}$$

where d is the hidden dimension of the backbone. The resulting vectors are used directly as input to all alignment and downstream probing experiments, with no further normalization.

A.4 Cross-Modal Dataset Summary

Table 6 summarises the complete partition statistics for the combined dataset $\mathcal{D} = \mathcal{D}^{\text{US}} \cup \mathcal{D}^{\text{MRI}}$. In total, $\mathcal{D}$ contains $4{,}056$ images from $1{,}973$ unique subjects across both modalities ($N_{\text{US}} = 3{,}369$ from $1{,}744$ subjects; $N_{\text{MRI}} = 687$ from 229 subjects).

Appendix B Implementation Details

B.1 Entropic Optimal Transport

The transport plan P is computed via a log-domain Sinkhorn–Knopp algorithm run for 100 iterations with uniform marginals $\mathbf{a} = \frac{1}{N_s}\mathbf{1}_{N_s}$ and $\mathbf{b} = \frac{1}{N_t}\mathbf{1}_{N_t}$, where N_s and N_t are the number of source and target samples in $\mathcal{D}_{\text{pair}}$, respectively. The implementation uses GPU acceleration. Before computing the soft target anchors $\tilde{X}^{\mathcal{T}} = \bar{P}X^{\mathcal{T}}$, each row of P is normalized to sum to one (giving $\bar{P}$), so that each source sample receives a convex combination of target embeddings.

Table 6: Dataset partition statistics for $\mathcal{D}^{\text{US}}$ and $\mathcal{D}^{\text{MRI}}$. GA is reported in days as mean $\pm$ std. % columns refer to the fraction of the respective modality total. Plane counts are TT / TC / TV.

Modality	Split	#Images	% Img	#Subjects	% Subj	GA (days)	TT / TC / TV
US	Overall	3,369	—	1,744	—	163.3 ± 45.7	2756 / 353 / 260
	$\mathcal{D}_{\text{pair}}$	961	28.5	613	35.1	176.7 ± 39.7	865 / 57 / 39
	$\mathcal{D}_{\text{train}}$	1,934	57.4	905	51.9	157.0 ± 45.7	1514 / 239 / 181
	$\mathcal{D}_{\text{test}}$	474	14.1	226	13.0	161.8 ± 51.1	377 / 57 / 40
MRI	Overall	687	—	229	—	214.1 ± 30.6	229 / 229 / 229
	$\mathcal{D}_{\text{pair}}$	480	69.9	160	69.9	216.1 ± 29.7	160 / 160 / 160
	$\mathcal{D}_{\text{train}}$	165	24.0	55	24.0	212.3 ± 31.2	55 / 55 / 55
	$\mathcal{D}_{\text{test}}$	42	6.1	14	6.1	198.9 ± 35.3	14 / 14 / 14

Cost Matrix Construction. The composite cost matrix $\tilde{C} = C - \min(C)$ is assembled from three terms. The temporal term $D^{\text{time}}_{i,j} = |g^{\mathcal{S}}_i - g^{\mathcal{T}}_j|$ measures the absolute gestational age discrepancy in days. The embedding term $D^{\text{proto}}_{i,j}$ measures the Euclidean distance between the plane-and-age-binned centroids of $X^{\mathcal{T}}$ associated with sample i and sample j: target embeddings are grouped into bins of width Δ days within each anatomical plane, per-bin centroids are computed, and each sample is mapped to the centroid of the nearest available bin within its plane. The spatial term $D^{\text{space}}_{i,j} \in \{0, 1\}$ is a binary indicator that equals 1 whenever source sample i and target sample j belong to different anatomical planes. The two continuous terms are independently z-scored before combining:

$$C = \mathcal{Z}(D^{\text{time}}) + \mathcal{Z}(D^{\text{proto}}) + D^{\text{space}}, \tag{7}$$

where $\mathcal{Z}(\cdot)$ standardizes the entries of a matrix to zero mean and unit variance, as in Eq. (4). The result is shifted by its minimum value to ensure non-negativity before being passed to the Sinkhorn solver.

Hyperparameter Search Space. The entropy regularization ε is searched over $\{10^{-4}, 5 \times 10^{-4}, 10^{-3}, 5 \times 10^{-3}\}$ and the centroid bin width Δ over $\{7, 14, 21\}$ days. All cost term weights are fixed to 1.0 throughout.

B.2 Alignment Models

In the main paper we evaluate the following alignment methods:

- **Ridge.** The regularization coefficient λ is searched over $\{10^{-4}, 5 \times 10^{-4}, 10^{-3}, 5 \times 10^{-3}, 10^{-2}, 5 \times 10^{-2}, 10^{-1}, 5 \times 10^{-1}\}$.
- **CCA.** The full-rank solution is used (no dimensionality reduction). Both within-modality covariance matrices are regularized by adding $\eta\mathbf{I}$ with $\eta = 10^{-6}$ before inversion.
- **CORAL.** The covariance matrices are regularized with $\eta = 10^{-5}$. CORAL is evaluated as a standalone pairing-free alignment baseline.
- **MLP.** The hidden layer dimension is searched over $\{256, 512, 1024\}$. When source and target embedding dimensions coincide, a residual connection is added around the hidden layer. Training uses learning rate 10^{-3} and weight decay searched over $\{10^{-4}, 5\times10^{-4}, 10^{-3}, 5\times 10^{-3}, 10^{-2}, 5 \times 10^{-2}, 10^{-1}, 5 \times 10^{-1}\}$, with a maximum of 600 epochs and early stopping with patience 80 (epochs without improvement on the training MSE). Weights are initialized with Kaiming uniform initialization. Training is performed on a single NVIDIA A100 GPU.

B.3 Downstream Probes

The GA Ridge regressor is fit with regularization coefficient $\alpha = 1.0$. The plane SVM uses balanced class weights to compensate for the plane imbalance in $\mathcal{D}^{\text{US}}$ described in Sec. A.

Appendix C Additional Experiments

All results in this section are on the held-out test sets described in Appendix A. CORAL, CCA and Ridge are deterministic given the pairing; unless stated otherwise, MLP rows report mean $\pm$ std

over 3 seeds, and CCA and Ridge rows with Random pairing are single runs. Sec. C.1 studies the sensitivity of MLP alignment to its hyperparameters, Sec. C.2 reports the MRI-native MRI-CORE in both directions and as a distillation student, and Secs. C.3 and C.4 compare WASP with unpaired alignment baselines and with direct fine-tuning.

C.1 MLP Alignment Hyperparameters

Tab. 7 reports the sensitivity of MLP alignment to the four hyperparameters of WASP: the entropy regularization ε, the centroid bin width Δ, the weight decay, and the hidden dimension of the network. The two dominant factors are Δ and ε, which interact strongly. Larger bin widths produce smoother target-space prototypes that better represent the local geometry of the embedding manifold; combined with a sufficiently large ε (5×10^{-3}), which yields moderately diffuse transport and thus more stable soft anchors, this drives the largest improvements in GA regression ($\Delta = 21$, $\varepsilon = 5 \times 10^{-3}$: MAE 18.3 days). In contrast, the weight decay and the hidden dimension show negligible sensitivity: varying the weight decay across its entire search range changes mean MAE by at most 0.3 days and mean Cls. by 0.5%, while varying the hidden dimension yields differences below 0.2 days and 1.1% respectively. Across all 288 configurations, 96% improve over the Zero-Shot MAE baseline and 83% improve over the Zero-Shot plane classification accuracy, confirming that MLP alignment is robust to hyperparameter choice.

Table 7: MLP alignment hyperparameter sensitivity (USFM, MRI→US, test set). Each cell reports GA Reg. ↓ (days) / Plane Cls. ↑ (%), averaged over weight decay $\in \{10^{-4}, \ldots, 5{\times}10^{-1}\}$ and hidden dim $\in \{256, 512, 1024\}$. **Bold** = column-wise best.

	$\varepsilon = 10^{-4}$		$\varepsilon = 5{\times}10^{-4}$		$\varepsilon = 10^{-3}$		$\varepsilon = 5{\times}10^{-3}$	
Δ (days)	GA ↓	Cls. ↑	GA ↓	Cls. ↑	GA ↓	Cls. ↑	GA ↓	Cls. ↑
7	21.8	64.4	21.4	**66.0**	21.4	**66.5**	20.7	66.6
14	**20.4**	64.3	20.2	60.0	20.4	59.1	20.9	67.7
21	20.8	**65.7**	**20.0**	65.6	**19.8**	65.3	**18.3**	**68.3**

C.2 MRI-CORE

Pairing ablation in both directions. Tab. 8 reports MRI-CORE in both directions. In US→MRI, US is the source and MRI the native target space (the MRI-CORE column of Tab. 2); in MRI→US, the source is MRI-CORE's native modality, mapped toward the US space. In both directions, Random and Cosine pairing underperform WASP across CCA, Ridge, and MLP alignment, with the gap being particularly large for GA regression. In US→MRI, WASP +MLP gives the best plane accuracy ($98.2 \pm 0.4\%$) with a GA MAE of 12.8 ± 0.2 days, close to Zero-Shot (12.3). In MRI→US, Zero-Shot is already strong on GA (6.5 days) and no alignment improves it; WASP +MLP stays closest (7.0 ± 1.0 days), and WASP +CCA gives the best plane accuracy (81.0% vs. 76.2% Zero-Shot).

Teacher and transfer-head ablation. Tab. 9 ablates both the teacher representation (Raw vs. +WASP) and the transfer head (Ridge vs. MLP) when distilling into MRI-CORE. Across both SAM-Med2d and BioMedParse teachers, MLP outperforms Ridge on GA regression. Applying WASP to the teacher is beneficial only with the MLP head: for SAM-Med2d, GA improves from 12.3 to 11.6 days, whereas the corresponding Ridge variant degrades from 13.3 to 14.5 days. A similar interaction appears for BioMedParse, where +WASP with MLP preserves the 12.0-day GA result while +WASP with Ridge degrades to 15.6 days. This ablation shows that the effect of teacher refinement depends strongly on the transfer-head choice.

Table 8: MRI-CORE in both directions. US→MRI is evaluated on the US test set, MRI→US on the MRI test set. Best and second-best results in each metric column are shown in **bold** and underline, respectively.

		US→MRI		MRI→US	
Pairing	Alignment	GA↓	Cls.↑	GA↓	Cls.↑
Zero-Shot	—	**12.3**	98.1	**6.5**	76.2
—	CORAL	16.7	96.2	19.8	57.1
Greedy	CCA	24.9	93.5	13.6	73.8
	Ridge	25.3	93.5	9.1	73.8
	MLP	**12.3±0.2**	97.2±0.1	9.0±0.1	78.6±2.4
Hungarian	CCA	16.1	92.0	16.5	71.4
	Ridge	14.2	96.8	15.8	69.0
	MLP	13.7±0.2	97.4±0.1	8.9±0.7	79.4±1.4
Random	CCA	27.6	81.9	29.5	42.9
	Ridge	27.7	81.2	24.5	47.6
	MLP	20.2±0.7	95.7±0.6	21.6±0.6	55.6±5.5
Cosine	CCA	30.1	75.3	25.4	57.1
	Ridge	30.9	74.7	26.6	54.8
	MLP	20.7±0.7	95.9±0.8	20.8±1.0	73.0±3.6
WASP	CCA	24.6	92.6	15.2	**81.0**
	Ridge	25.1	93.0	13.4	76.2
	MLP	12.8±0.2	**98.2±0.4**	7.0±1.0	77.0±1.4

Table 9: MRI-CORE as student, US→MRI. Teachers are improved with WASP in the US→MRI direction, and $\mathcal{G}_{\text{US}\to\text{US}}$ (Ridge or MLP) maps MRI-CORE's US embeddings to the teacher's; evaluation is on the US test set. Raw rows are deterministic; + WASP rows, including Ridge ones, report mean ± std over the 3 seeds of the teacher alignment. Best and second-best results in each metric column are shown in **bold** and underline, respectively.

Teacher	Setup	$\mathcal{G}$	GA↓	Cls.↑
Zero-Shot	—	—	12.3	**98.1**
SAM-Med2d	Raw	Ridge	13.3	97.3
	Raw	MLP	12.3±0.1	97.5±0.2
	+ WASP	Ridge	14.5±0.4	95.2±0.7
	+ WASP	MLP	**11.6±0.1**	98.0±0.1
BioMedParse	Raw	Ridge	13.2	97.9
	Raw	MLP	12.0±0.3	97.6±0.2
	+ WASP	Ridge	15.6±0.2	96.5±0.1
	+ WASP	MLP	12.0±0.1	97.7±0.4

C.3 Unpaired Alignment Baselines

We compare against two representative unpaired approaches under the same frozen-backbone protocol (Tab. 10). *MRI→US synthesis*: a multimodal hierarchical VAE [40] translates fetal MRI slices into synthetic US images, which go through the same preprocessing as all other images (Appendix A) and are embedded by each frozen backbone; the probes are fit on the synthetic-US training split and evaluated on the same 14 held-out MRI test subjects. *Adversarial alignment*: a feature-level variant of ADDA [41], where a trainable mapper pushes MRI embeddings toward independently sampled US embeddings to fool a discriminator (two 500-unit ReLU layers, batch size 64, 600 alternating updates, AdamW, multiple seeds), without metadata or subject pairing. Training did not collapse, but domain confusion was never reached: the discriminator stayed 73–98% accurate, and a held-out classifier separated transformed MRI from real US embeddings with >99% accuracy. The mapper is not the identity, yet the downstream metrics stay at Zero-Shot level. The synthesis baseline improves USFM plane accuracy but degrades Echocare, SAM-Med2d and BioMedParse on both tasks.

Table 10: **MRI→US.** Unpaired alignment baselines on the MRI test set. WASP results are in Tab. 1. Best and second-best results within each backbone and metric are shown in **bold** and underline, respectively.

	Zero-Shot		MRI→US synthesis		Adversarial alignment	
Backbone	GA↓	Cls.↑	GA↓	Cls.↑	GA↓	Cls.↑
USFM	21.9	61.9	**20.8**	**71.4**	21.8	64.3
Echocare	**10.3**	**81.0**	10.7	71.4	**10.3**	**81.0**
SAM-Med2d	**7.7**	**83.3**	9.5	81.0	**7.7**	**83.3**
BioMedParse	**6.5**	**85.7**	8.6	81.0	6.6	**85.7**

C.4 Fine-Tuning on MRI

As a reference outside our setting, we fine-tune each backbone directly on the MRI training split (Tab. 11, mean ± std over 3 seeds). Fine-tuning lowers the MRI GA error further than WASP, at the price of modifying the backbone: it increases the GA error on the original US domain for every backbone and reduces Echocare's US plane accuracy. WASP keeps the backbone frozen and preserves the source-domain representation by construction.

Table 11: Direct fine-tuning on MRI. Left: MRI test set. Right: US test set before and after fine-tuning. Best and second-best results within each backbone, test domain, and metric are shown in **bold** and underline, respectively.

	MRI test set				US test set			
	Zero-Shot		Fine-tuned		Before		After fine-tuning	
Backbone	GA↓	Cls.↑	GA↓	Cls.↑	GA↓	Cls.↑	GA↓	Cls.↑
USFM	21.9	61.9	**8.4±1.4**	**81.7±11.3**	**16.0**	92.8	18.3±0.3	**95.1±1.2**
Echocare	10.3	81.0	**7.9±2.0**	**86.5±3.6**	**14.9**	**95.6**	26.0±3.0	88.2±4.2
SAM-Med2d	7.7	83.3	**5.3±0.5**	**88.1±2.4**	**15.7**	96.2	16.8±0.5	**98.6±0.5**
BioMedParse	6.5	85.7	**4.3±0.2**	**92.1±1.4**	**12.1**	98.7	15.0±1.5	**98.9±0.9**

Appendix D Fetal Brain Development

The Weakly Aligned Spatiotemporal Pairs (WASP) framework is fundamentally anchored in the predictable, stereotyped trajectory of fetal brain development. In neurotypical gestations, the structural maturation of the fetal brain produces highly consistent macroscopic morphologies at any given gestational age (GA) [2]. This biological rigidity across subjects allows WASP to establish valid cross-modal correspondences using only time (GA) and space (diagnostic plane) as metadata.

D.1 The Spatiotemporal Trajectory of Neural Development

Fetal brain maturation during the second and third trimesters is characterized by explosive volumetric growth, neuronal migration, and the initiation of complex sulcation and gyration [42, 43]. Early in the second trimester (approx. 14–16 weeks GA), the telencephalic vesicles are relatively smooth (lissencephalic), and the cortical plate is just beginning to organize via the migration of neuroblasts along radial glia [44, 45]. As gestation progresses into the 20–28 week window—the primary overlap period for both the US and MRI cohorts in this study—the brain transitions from a smooth state to a highly convoluted structure. The primary sulci (such as the lateral/Sylvian fissure, parieto-occipital sulcus, and central sulcus) appear in a strict temporal sequence [46, 47]. Because the timing of these invaginations is genetically and developmentally constrained, the sulcal pattern serves as an accurate biological clock. Consequently, the anatomical variance between two neurotypical subjects at the exact same GA is lower than the variance within a single subject observed just weeks apart [48, 49]. This constraint is what minimizes the D^{time} and D^{proto} costs in the WASP Optimal Transport formulation.

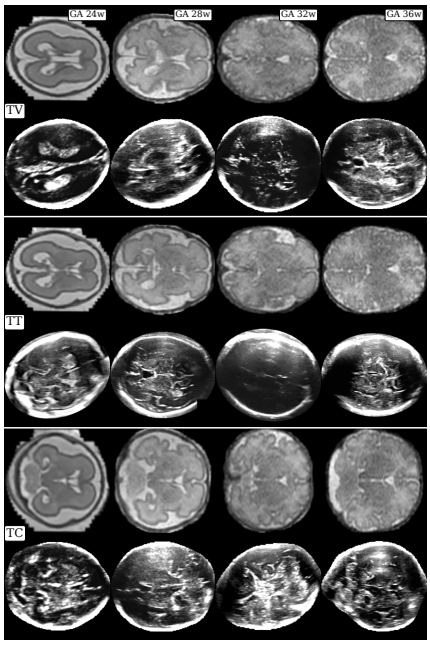

Figure 3: Spatiotemporal Evolution of Fetal Brain Anatomy in MRI and US. Rows represent the three standard diagnostic planes (TV, TT, TC). Columns represent different Gestational Ages.

D.2 Anatomical Grounding of the Diagnostic Planes

Standardized diagnostic planes capture distinct axial cross-sections of this developing 3D anatomy. The spatial cost term (D^{space}) in WASP enforces strict plane matching because each plane intercepts entirely different neurobiological distinct structures [3]:

- **Transventricular Plane (TV):** Intercepts the superior aspect of the lateral ventricles. Early in gestation, the choroid plexus is highly prominent and fills the lateral ventricles. As the cortical mantle thickens due to arborization and synaptogenesis, the relative volume of the lateral ventricles decreases [50, 3].
- **Transthalamic Plane (TT):** A slightly lower axial cut that captures the thalami, the midline falx, and the cavum septi pellucidi (CSP). The CSP is a critical fluid-filled space between the leaflets of the septum pellucidum; its typical development and subsequent obliteration later in life are key markers of normal midline forebrain fusion [51, 52, 3].
- **Transcerebellar Plane (TC):** Angled posteriorly to capture the posterior fossa, including the cerebellar hemispheres, the vermis, and the cisterna magna. The cerebellum undergoes dramatic foliation and exponential growth in the late second and third trimesters, which is highly visible in both US and MRI [53, 54].

Appendix E Broader Impact

Although our study focuses on the alignment of fetal MRI and ultrasound, the core principles of WASP extend far beyond this specific use case. By demonstrating that routine clinical metadata can successfully substitute for subject-level paired scans, our framework provides a versatile blueprint for cross-modal representation learning across diverse medical imaging domains.

Concrete candidates for the same construction include pediatric MRI/CT/US studies indexed by chronological age and acquisition protocol, adult MRI/PET cohorts indexed by anatomical landmarks, and histopathology–radiology pairings indexed by tumor stage. In each case the alignment cost is built from coordinates that are already part of routine clinical reporting, so no new annotation effort is needed beyond what hospitals already record.

Two caveats specific to these broader uses are worth stating. First, WASP establishes anatomical, not subject-level, correspondence: matched scans share a developmental coordinate but not an identity, and must not be interpreted as same-subject evidence in an individual workup. Second, because the underlying backbones are reused as-is, biases present in their pretraining cohorts (scanner vendor, geography, age range) propagate into the aligned representations and should be audited rather than assumed to be neutralized by the alignment step.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The three claims of WASP (soft cross-modal OT pairing, unlocking/refinement of frozen US-backbones MRI representation, and teacher-to-student MRI representation-distillation) map directly to Sec 4, Tab. 1 (e.g., USFM 21.9→17.4 days GA, 61.9%→69.0% plane accuracy), and Tab. 3 respectively (e.g., Echocare 10.3→8.6 days GA, 81.0%→90.5% plane accuracy under cross-model distillation). Scope is bounded to neurotypical cohorts and linear-probe proxy tasks, as restated in Sec. 5.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: A dedicated Limitations paragraph in Sec. 5 discusses the restriction to neurotypical fetuses, the Hadlock-estimated gestational ages, the heterogeneity of the gains across backbones and alignment modules, the scope of the baselines, and the loss of volumetric context from 2D slicing.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: The paper does not introduce new theoretical results.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Sec. 3 fully specifies the WASP pipeline, Secs. 4.1,4.2 detail datasets, splits, probes, and evaluation, and Appendices A and B report preprocessing, per-split statistics, and all hyperparameter search spaces.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: All datasets used are publicly available and cited. Code and precomputed embeddings are released at the repository URL given below the author list, with instructions to reproduce the main experiments.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Splits, probe configurations, the 3-fold subject-level CV protocol, and the full hyperparameter search spaces for Optimal Transport, Ridge, CCA, CORAL, and MLP alignment are reported in Sec. 4.1 and Appendix B.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

Justification: CORAL, CCA and Ridge alignments are deterministic given the pairing; for the MLP alignment, the only stochastic component, results report mean $\pm$ standard deviation over 3 seeds, except for the cost-term ablation (Tab. 4), which reports single runs, as stated in each table caption.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: As reported in Sec. B, all experiments operate on frozen backbones with lightweight alignment modules (closed-form Ridge or single-hidden-layer MLP), so the footprint is modest and dominated by one-time embedding extraction; MLP training uses a single NVIDIA A100 GPU.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work uses only publicly released, de-identified fetal imaging datasets under their original ethics approvals, and collects no new human-subjects data.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.

- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: A dedicated Broader Impact section in Appendix E details the positive potential of extending high-quality fetal MRI analysis to US-only settings, the wider applicability of the metadata-driven alignment recipe to other modality pairs, multi-site harmonization, and low-resource clinical contexts, while cautioning that claims are restricted to neurotypical cohorts and proxy tasks and do not support clinical deployment without prospective validation.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper releases no generative models, LLMs, or scraped datasets; WASP is a lightweight alignment framework over already-released frozen backbones and public de-identified imaging data.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All datasets and backbones are cited at the point of use and used under their respective public licenses.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We release the code and the precomputed backbone embeddings with their metadata and splits, documented in the repository; the embeddings are derived from the public datasets cited in the paper and are subject to their terms.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: No crowdsourcing or new human-subjects data collection is involved; all imaging data come from previously released, de-identified public datasets.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: No new human-subjects research was conducted; all datasets used were released under the IRB or equivalent approvals obtained by their original creators.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not a component of the methodology, which relies solely on entropic OT, closed-form linear alignment, and a single-hidden-layer MLP.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.