Supplementary Material: Generative Adversarial Registration for Improved Conditional Deformable Templates

Neel Dey
New York University
neel.dey@nyu.edu

Mengwei Ren
New York University
mengwei.ren@nyu.edu

Adrian V. Dalca
MIT, MGH
adalca@mit.edu

Guido Gerig
New York University
gerig@nyu.edu

A. Supplementary Results

Supplementary animations illustrating several conditional spatiotemporal experiments are available at `https://www.neeldey.com/deformable-templates`.

Figure 1. FFHQ-Aging age and cohort conditional templates with normal glasses (top) and sunglasses (bottom). For ages 7 and older, all methods produce plausible conditional templates, with `Ours` removing border effects and increasing shape and appearance variability. Significant label noise and highly limited sample sizes are apparent for ages 0-2 within the "glasses" label and for ages 0-6 within the "sunglasses" labels. For example, only two images exist within the training set for the male/with sunglasses/0-2 years old FFHQ-aging class with both images displaying *adults* with sunglasses and not infants (as can be seen from the corresponding linear average). As a result, methods using FiLM [20] (`Ablation/noAdv` and `Ours`) produce more adult-like templates in those age ranges. We speculate the results come from the increased data fitting capacity of FiLM. Interestingly, methods which do not use FiLM (`Ablation/VXM+Adv` and `VXM`) produce more plausible age-conditioned templates when all of the data for a category is mislabeled. This phenomenon arising from significant label noise requires future investigation.

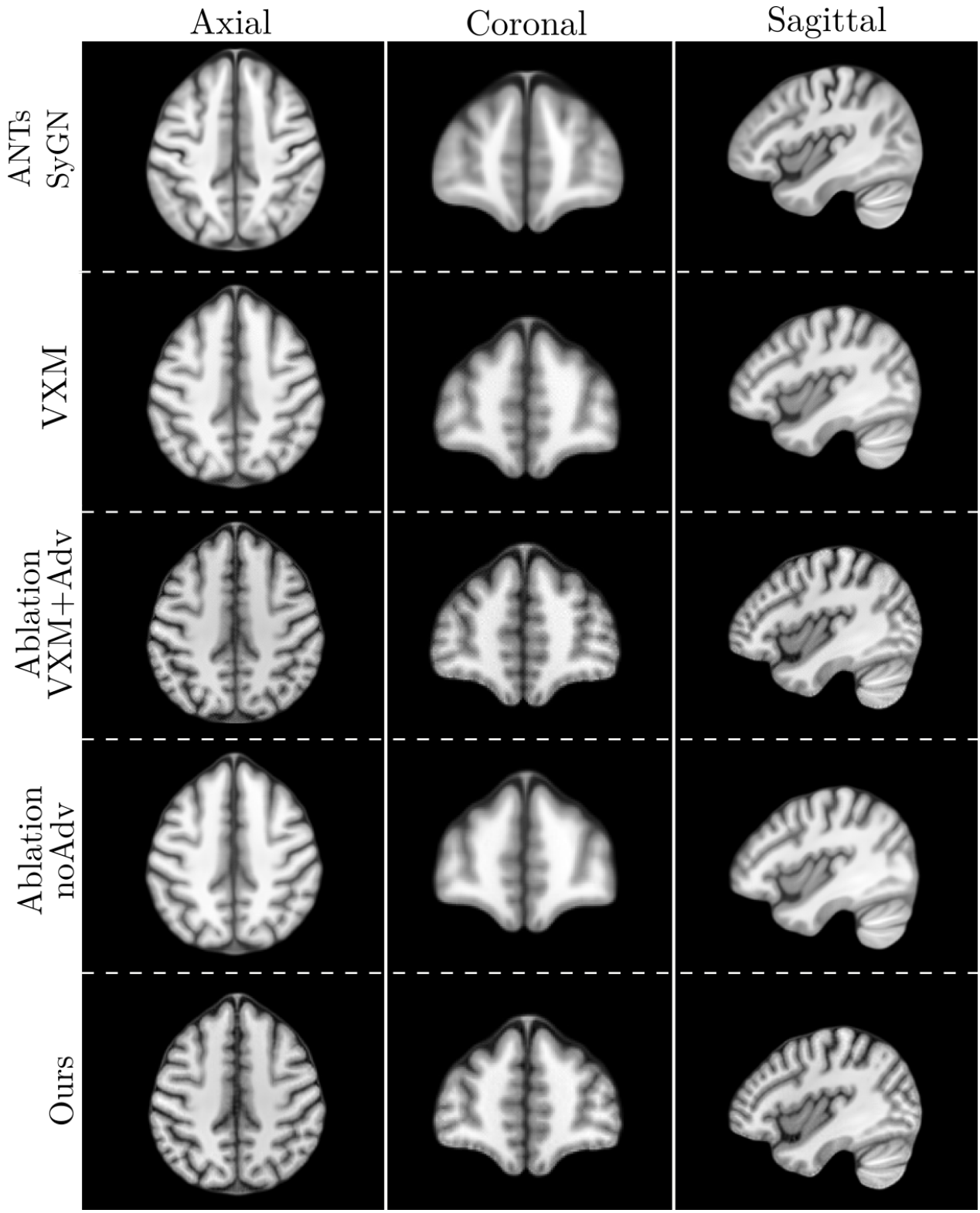

Figure 2. Additional 2D views of unconditional 3D template construction on Predict-HD from all four methods. Methods using a discriminator (`Ablation/VXM+Adv` and `Ours`) exhibit increased sharpness, cortical folding detail, and tissue contrast. `Ours` improves on `Ablation/VXM+Adv` by removing subtle checkerboard artefacts, particularly visible in the coronal view.

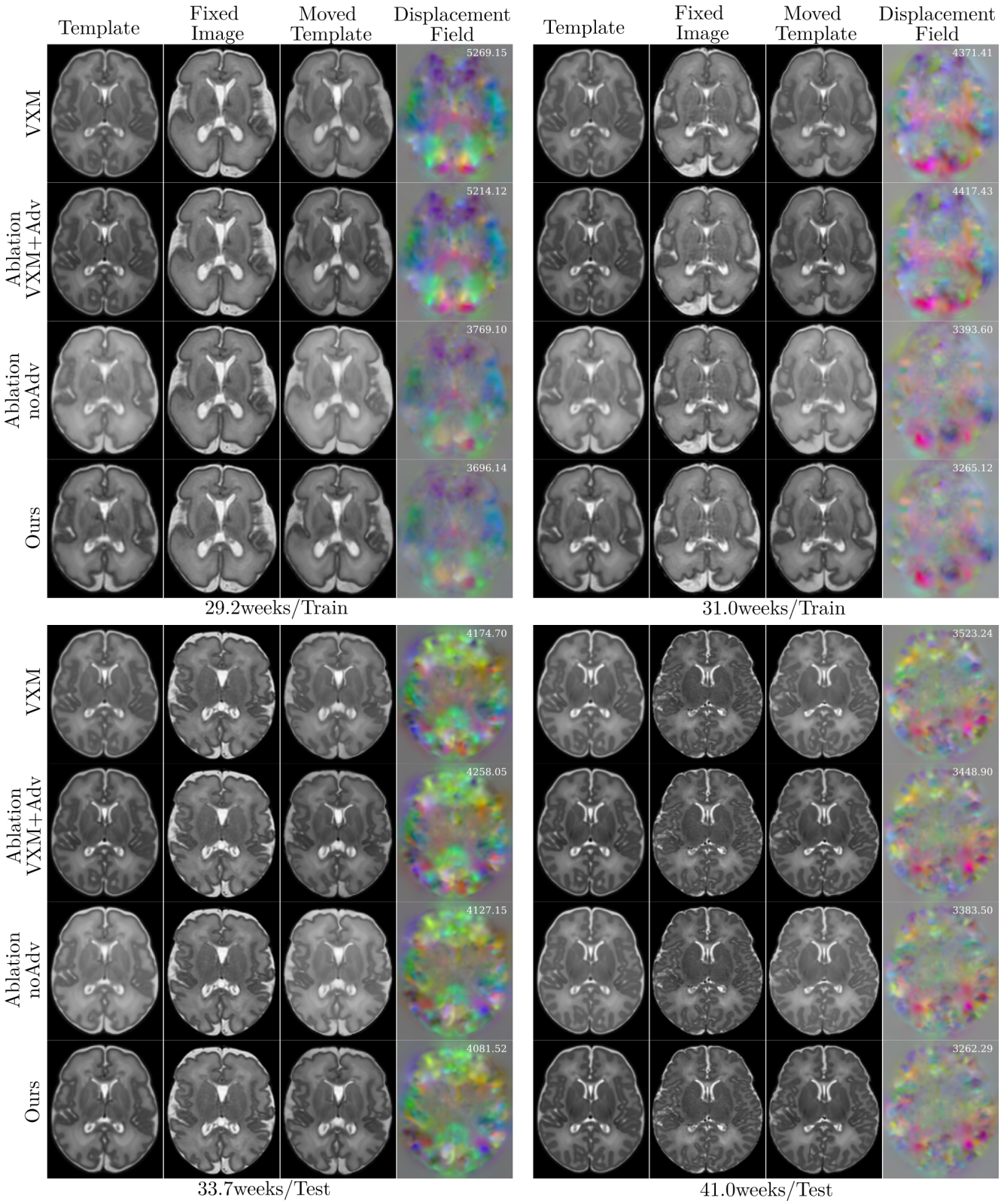

Figure 3. Example dHCP template-to-image registration results for all methods on training data (top subfigures) and held-out test data (bottom subfigures), with varying gestational ages. Deformation norms for the 3D displacement fields are annotated on the top-right. We visualize training set results in addition to test data as a large age range (29-31weeks) of interest is not present in the test set (See Figure B.1). As our templates show higher condition (age) specificity, the deformations are smaller and more anatomically plausible as compared to baselines and ablations.

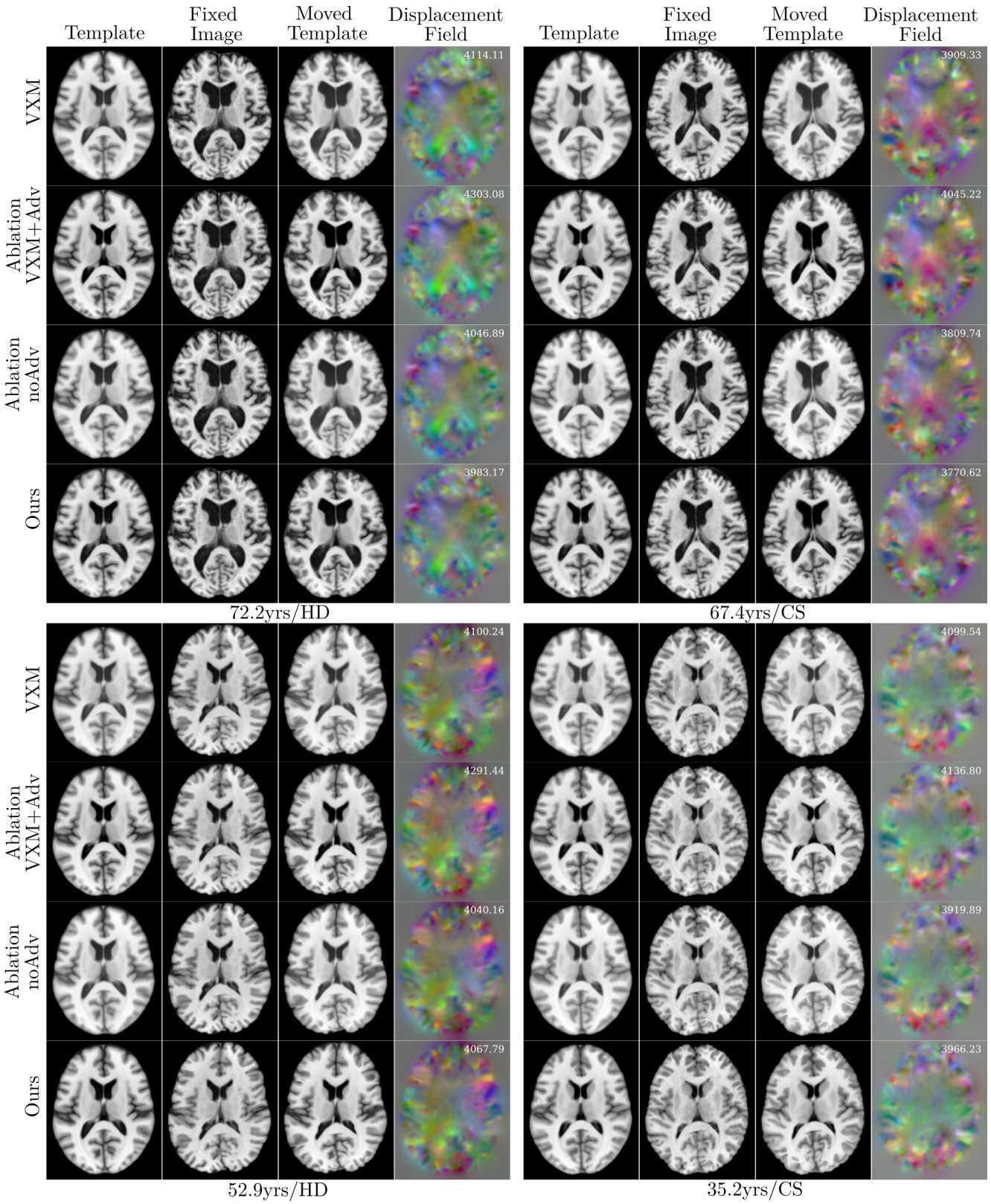

Figure 4. Example Predict-HD template-to-image registration results for all methods on held-out test data, with varying ages and cohorts. Deformation norms for the 3D displacement fields are annotated on the top-right. All methods produce comparable moved templates. However, ours yields smaller deformations as seen from the displacement fields (especially visible in 72.2yrs/HD and 67.4yrs/CS).

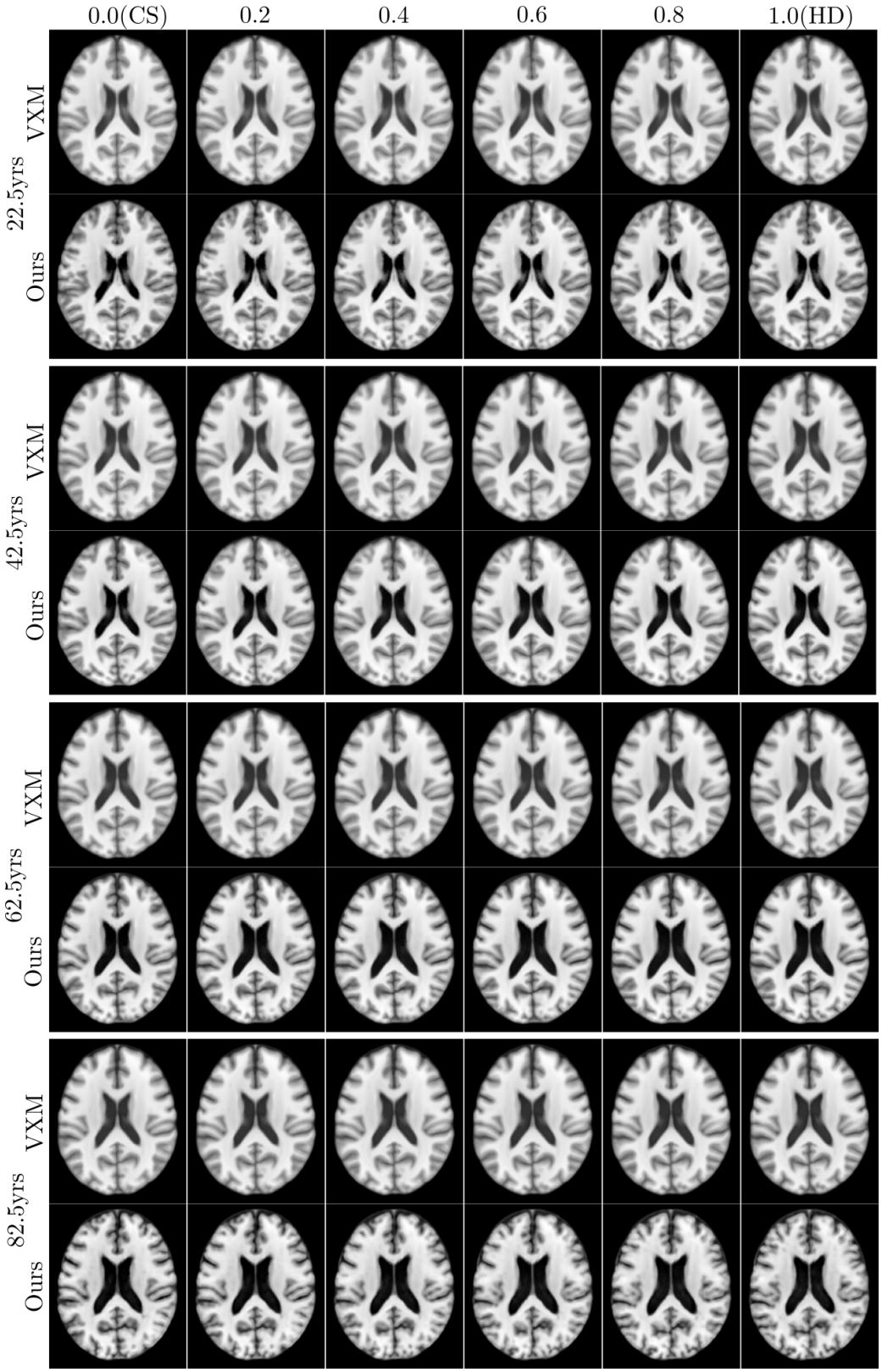

Figure 5. Interpolations between control subjects (CS) and subjects with the Huntington's disease (HD) mutation ($[0, 0.2, 0.4, 0.6, 0.8, 1]$), for fixed ages, obtained by linearly interpolating between one-hot attribute vectors. Both methods (`VXM` and `Ours`) achieve interpolations which match clinical expectations, e.g., with ventricles growing larger as the HD weight increases. `Ours` displays larger differences between CS and HD with correspondingly larger changes visible in the interpolations, as can be seen from the last row of the figure.

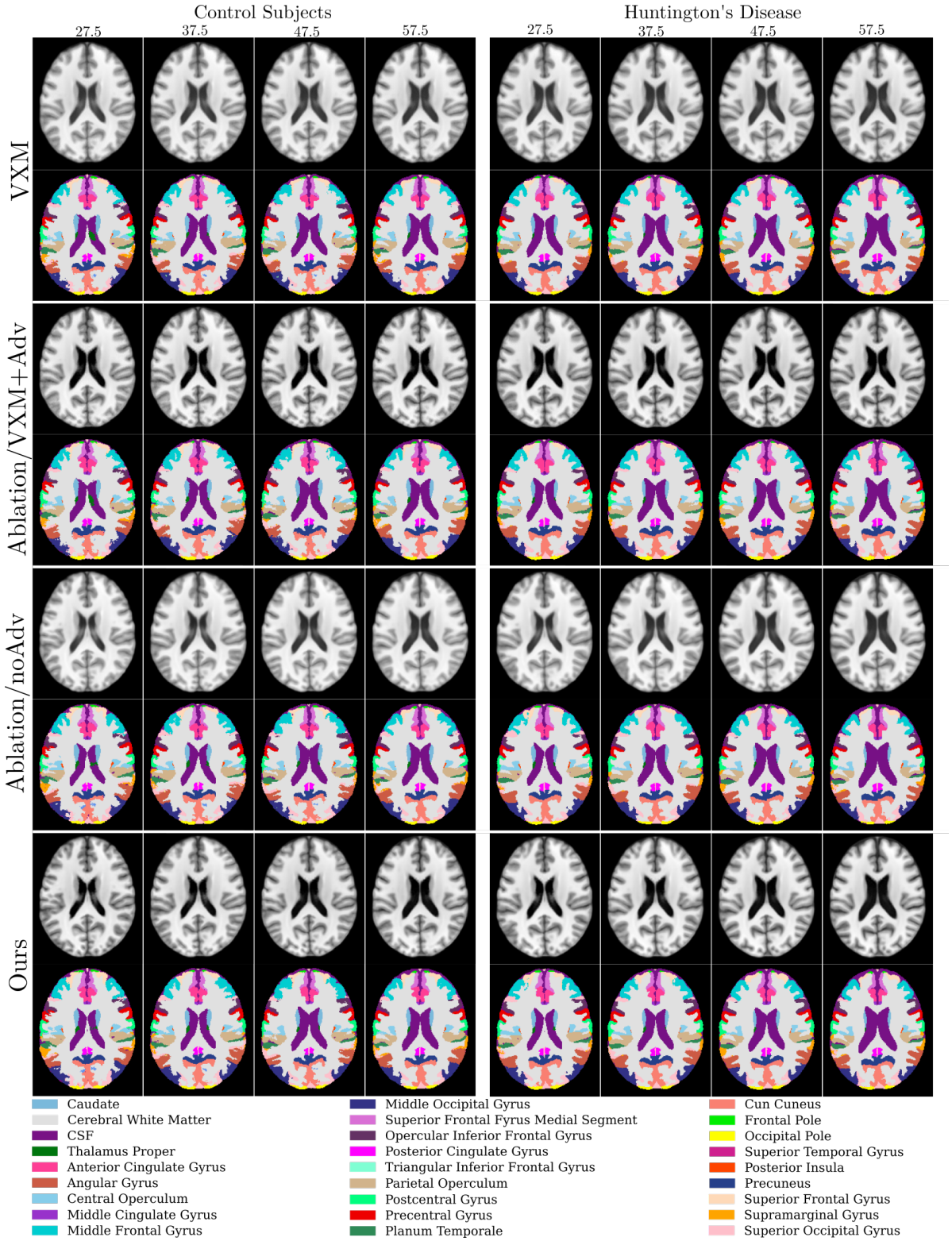

Figure 6. Example template segmentations for all methods generated by majority voting on inverse warped labels of training images. We emphasize that no segmentation labels are used in template construction or registration and that these segmentations are used only for Dice coefficient evaluation and temporal volume trends.

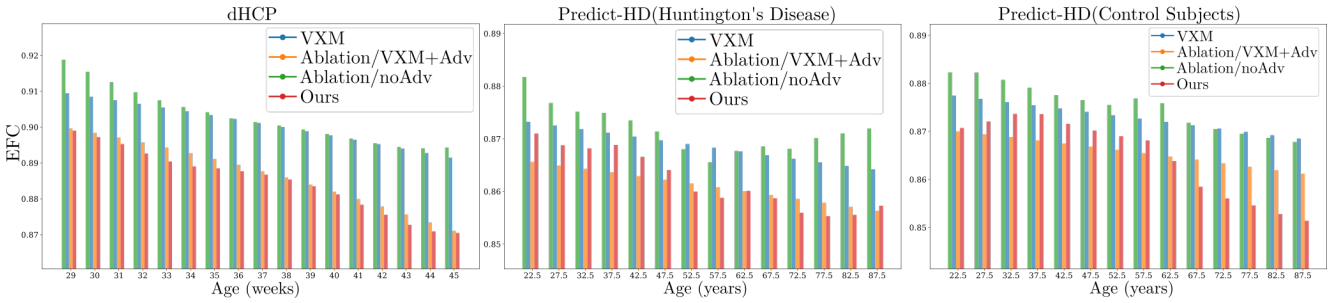

Figure 7. Temporal Entropy Focus Criteria (EFC, lower is better) for conditional templates on the dHCP (left), Predict-HD/Huntington's Disease (center), and Predict-HD/Control Subjects (right). In all cases, methods using a discriminator (`Ablation/VXM+Adv` and `Ours`) achiever lower EFC over non-generative adversarial methods. These results should be interpreted in context as:
(1) While `Ablation/VXM+Adv` and `Ours` achieve equivalent EFC/sharpness, `Ours` displays increased condition-specificity and centrality as shown in the experiments in the primary text.
(2) Although commonly used to evaluate unconditional template sharpness, EFC is a heuristic surrogate for image sharpness and can fluctuate with varying structure. As `Ablation/noAdv` and `Ours` show strong structural changes temporally, their temporal trends show higher variability as compared to techniques which present smaller structural changes (`Ablation/VXM+Adv` and `VXM`). As a result, EFC should be compared across methods at individual timepoints.

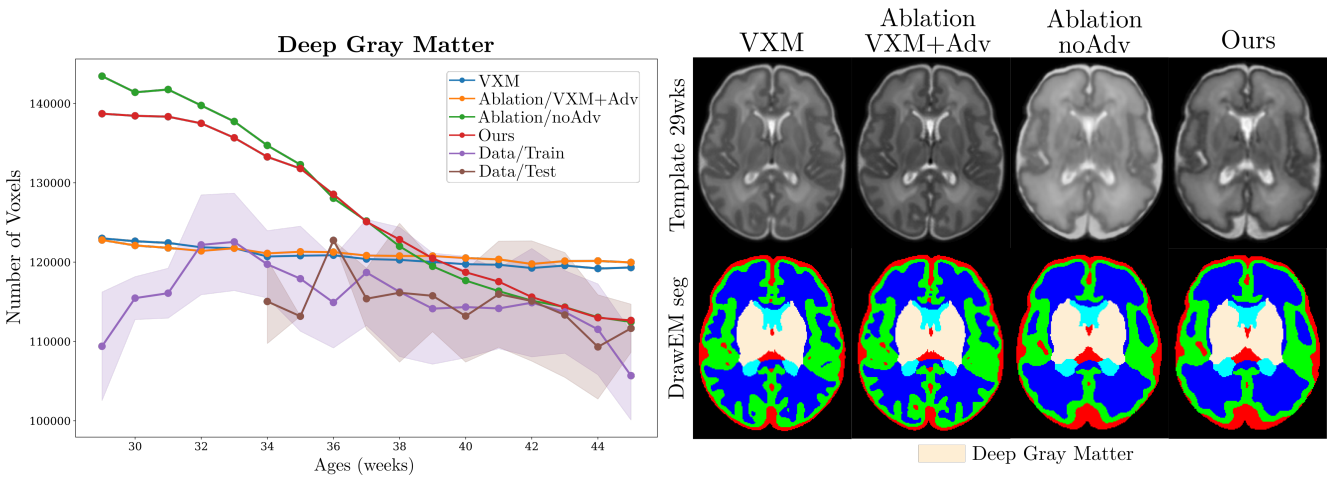

Figure 8. Negative results for dHCP conditional template segmentation for the Deep Gray Matter (dGM) label. `DrawEM` [14] (the tool used for dHCP template segmentation) with its default parameters overestimates dGM volume on the templates sampled at younger timepoints by `Ablation/noAdv` and `Ours`. For example, on the right, we show the generated templates from all methods at 29 weeks gestational age, with their `DrawEM` segmentation results below. While `Ablation/noAdv` and `Ours` produce more anatomically plausible templates compared to `VXM` and `Ablation/VXM+Adv`, the segmentation algorithm overestimates dGM volume. All other labels better match the underlying volume trends on the real data as shown in Figure 4 of the main text. In future work, careful tuning of DrawEM parameters on validation data may resolve this dGM overestimation.

B. Experimental Details

B.1. Data Preparation

All foreground/brain extraction is performed by thresholding provided segmentation labels. All neuroimages are cropped to a central field-of-view of resolution $160 \times 192 \times 160$. We obtain all linear averages required for the neuroimaging experiments using voxel-wise averages of a 100 randomly chosen training scans. All 60,000 training images are used to compute the linear average for FFHQ (while we visualize group-wise L_2 barycenters in row 1 of Figure 1 for comparison, our framework uses the overall L_2 barycenter for training).

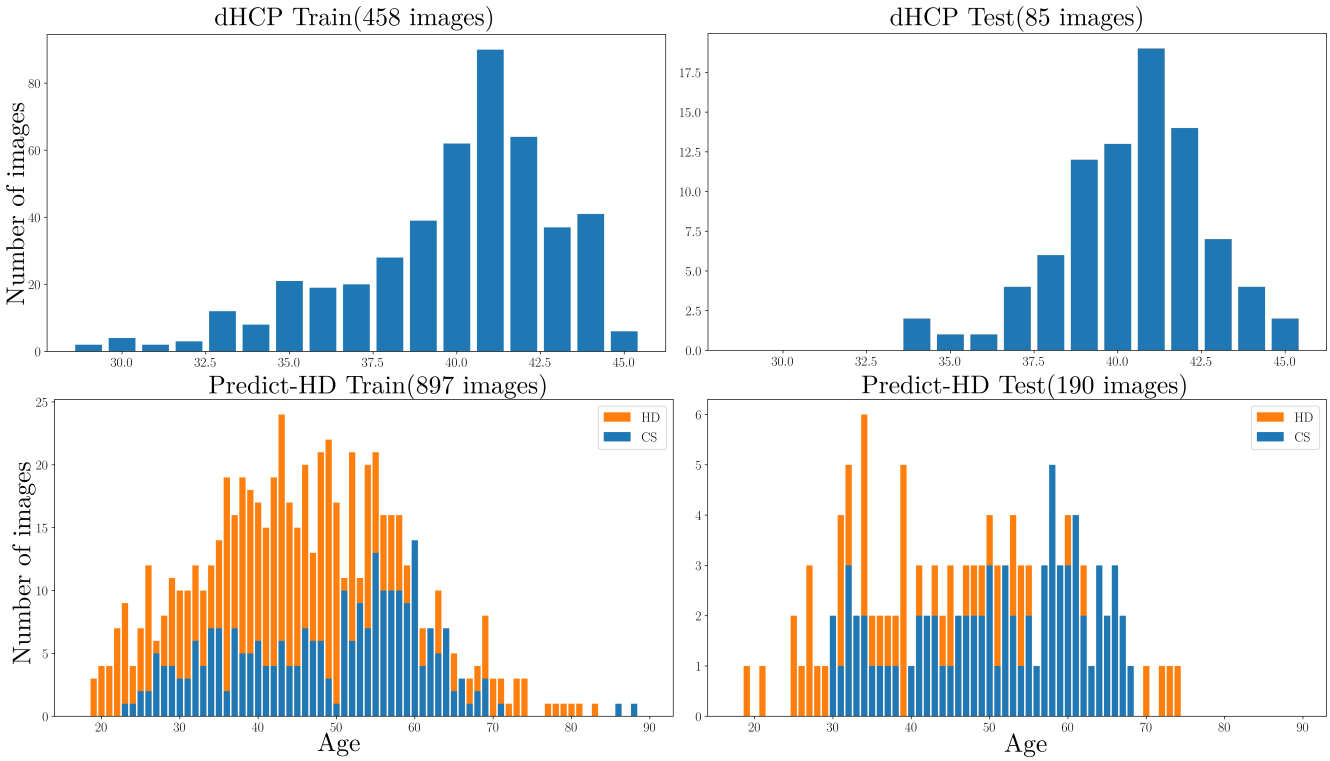

Figure 9. Histograms of sample size vs. scan age for both dHCP (top row) and Predict-HD (bottom row), for the constructed training sets (left column) and test sets (right column). Both datasets present both a significant gap between training and test sets in terms of scan age sampling.

B.1.1 Predict-HD

Predict-HD provides longitudinal scans from 388 individuals with and without the Huntington's Disease (HD) genetic condition. As imaging was performed across several distinct scanning sites, the images present highly heterogeneous appearance. All T1 images were bias corrected and segmented using procedures described in [19]. Prior to learning nonlinear deformable registration, we affinely register all T1 images to MNI [9], thus resampling them to $1 \times 1 \times 1mm^3$. Out of 1121 images, 4 either failed affine alignment or had missing covariates and were discarded. We use 897, 30, and 190 images for training, validation, and testing, respectively, split at the subject level. In the context of this study, we do not currently consider longitudinal subject-specific effects in our conditional template estimation.

To compute template-to-image registration accuracy via Dice coefficient evaluation, we follow the template segmentation protocols outlined in [7]. Briefly, we select training scans within the ages of 25 and 65 years old wherein we have sufficient sample sizes for both cohorts and for both train and test sets. Accordingly, our Dice coefficient evaluation is only conducted on held-out test subjects between the ages of 25 and 65 (176 out of 190). The images are split into 5-year-wide age bins with a single template sampled at the center of the bin (i.e., a 52.5 year old HD template for HD subjects with ages between 50 and 55). For each cohort, all training segmentations within a bin are inverse warped to the bin-specific template, followed by majority voting on the labels to obtain the template segmentation for that age-bin and group. Unconditional template segmentations are performed via the same procedure, without the need for binning time points. In the future, other label fusion methods accounting for local intensities can be incorporated [22].

B.1.2 dHCP

Release 2 of the developing human connectome project (dHCP) was pre-processed by a specialized pipeline for neonatal image analysis [15] including steps such as motion-correction, super-resolution (from $0.8 \times 0.8 \times 1.6mm^3$ to $0.5 \times 0.5 \times 0.5mm^3$), bias correction, brain extraction, and segmentation [14]. For GPU memory, we crop to a central field-of-view and minimally resize images from $0.5 \times 0.5 \times 0.5mm^3$ voxel resolution to $0.6132 \times 0.6257 \times 0.6572mm^3$ for a final image

size of $160 \times 192 \times 160$. For training, validation, and testing, we first assign all twins and repeat scans to the training set to prevent test set leakage and randomly hold-out a 100 scans from the remainder (15 for validation, 85 for testing), resulting in 458 training images. We construct an affine template for the training set with `ANTs` to which every scan is affinely aligned.

To generate segmentations for conditional templates generated by all methods for use in computing registration accuracy via Dice coefficients and analyzing volumetric trends of anatomical regions-of-interest, we use `DrawEM` [14]. Briefly, `DrawEM` is a multi-atlas EM-segmentation pipeline based on the neonatal ALBERT templates [10] using normalized mutual information based image registration. This is in contrast to the majority voting template segmentation procedure for Predict-HD above and [7]. `DrawEM` segmentation was performed instead of majority voting as several time points have very limited sample sizes not suitable for majority voting when using regularized registration, leading to qualitatively inaccurate template segmentations. We find that `DrawEM` produces sufficiently accurate segmentations on templates produced by all methods (as shown in Figure 3 of the main text, see Figure 8 in the supplementary material for a counter-example). Unconditional template segmentation was performed following [7].

B.1.3 FFHQ-Aging

FFHQ-Aging [17] annotates images in the FFHQ [12] human face dataset. These annotations include genders, ages (in 10 age bins), head pose (pitch, roll, and yaw), type of glasses (no glasses, normal glasses, sunglasses), eye occlusion scores, and segmentation labels (obtained by a DeepLabV3 [6] model pre-trained on CelebAMask-HQ). For simplicity, we only use the categorical attributes, leaving head pose and eye-occlusion conditioning to future work. We train all models on the FFHQ training set of 60,000 images (out of 70,000). As is common [3], we restrict ourselves to qualitative evaluations for face templates. Importantly, we note that categorical template conditions for human faces are quite coarse as attributes such as gender are not purely categorical. Further, we note that the data set is skewed towards lighter skin tones (as evidenced by the linear averages of training images visualized in Supplementary Figure 1), which is consequently reflected in the synthesized templates from all methods. In future work, more careful modeling and diverse data collection protocols may work towards ameliorating these issues.

B.2. Additional Implementation Details

Design choices and hyperparameters. Architectures are given in Table 1. All estimated templates for neuroimaging experiments are masked by a foreground mask during training for all methods to suppress commonly occurring background artefacts. The foreground mask was obtained by thresholding a linear average of training images. Reflection padding was used instead of zero-padding for all methods as it led to slightly fewer checkerboard artefacts. LeakyReLU slopes were set to 0.2. A window of 100 updates is used for the moving average deformation penalty $\bar{u}$ for all datasets. For condition vectors z, we encode age as a continuous attribute (for the neuroimaging where we have access to continuous age values) divided by the maximum age in the dataset, and categorical attributes as one-hot vectors. We find that not rescaling continuous attributes in z can lead to discriminator instability. Weight decay was applied on the linear projections from the FiLM embedding to the individual layers with weight 10^{-5} for neuroimages and 10^{-6} for FFHQ-Aging.

We choose the stationary velocity field (SVF [1, 2]) framework primarily for its speed and ease of implementation and note that other frameworks such as LDDMM [4] can also be used. The integration over time $t \in [0, 1]$ is in practice implemented for all methods with five *scaling and squaring* layers which have been shown to produce smooth diffeomorphic displacement fields [1, 8]. While all training is performed on full resolution 3D volumes, velocity and displacement fields are estimated at half-resolution and then linearly scaled up during training as in [7]. This resizing has an implicit smoothing effect. Implementations of spatial transformers and scaling and squaring layers are taken from the `voxelmorph` library at `voxelmorph.mit.edu`.

For FFHQ-Aging, we make a few changes from the neuroimaging datasets. FFHQ-Aging provides ages in categorical form and are thus treated as one-hot representations. Linear averages for FFHQ-Aging were computed across the entire training dataset due to the high number of classes. As the dataset has a left-right head pose asymmetry (particularly pronounced in subclasses with few samples), we use horizontal reflection augmentation for all methods when training the template generation and registration sub-networks. We further use a penalty $\|I - I_{LR}\|_2^2$ with unit weight (where I_{LR} indicates a left-right reflection of I) for all methods to encourage symmetric face templates.

Optimization Details. As GAN training involves the joint optimization of two networks, the optimization parameters used in either network impacts training stability. The Adam [13] optimizer is used in all networks. For conditional dHCP and Predict-HD training, we adopt a two-time-scale-update-rule (TTUR [11]), with step size $\eta_G = 10^{-4}$, $\eta_D = 3 \times 10^{-4}$,

using $\beta_1 = 0.0$, $\beta_2 = 0.9$ in both networks as is common in recent GAN works [5, 18]. For conditional FFHQ-Aging, we reduce η_D to 2×10^{-4} as additional stability was needed for highly challenging face registration. Unconditional template optimization was found to be amenable to mild amounts of momentum and was performed with step-size $\eta = 10^{-4}$, $\beta_1 = 0.5$, and $\beta_2 = 0.999$ used in both generator and discriminator for faster convergence. We note that momentum is theoretically contraindicated for R_1 gradient penalty on the discriminator but we did not find this to be an issue in practice. The non-GAN baselines (`VXM` and `Ablation/noAdv`) were trained with the same strategies to enable valid comparisons.

ANTs SyGN parameters. We use the `antsMultivariateTemplateConstruction2.sh` script provided by the ANTsX ecosystem [21] which implements the SyGN algorithm from [3]. We use the default construction parameters, including the squared local normalized cross-correlation objective, four template updates, using a four-level registration pyramid with at $6\times, 4\times, 2\times$ downsampling for the first three resolutions, and $100 \times 100 \times 70 \times 20$ iterations per resolution. We turn off the default bias field correction and linear registration steps as these are performed during data pre-processing. Registrations between the estimated template and held-out test images were performed with the same registration parameters. We leave the default Laplacian sharpening on for all comparisons.

Miscellaneous Experimental Details. All networks are implemented in `TensorFlow 2.2` and trained on a single NVIDIA V100 GPU. As the GAN frameworks (`Ours` and `Ablation/VXM+Adv`) require concurrent optimization of two 3D networks, we found 16 GB vRAM neccessary for training. All entropy focus criteria are calculated within a common brain mask for the dataset.

C. Projection Discriminator

We use the inner product-based framework presented in [16] who observe that the optimum for the standard adversarial loss can be written as (equation 2 of [16]):

$$f^*(x, y) = log\Big(\frac{q(y|x)q(x)}{p(y|x)p(x)}\Big) = log\Big(\frac{q(y|x)}{p(y|x)}\Big) + log\Big(\frac{q(x)}{p(x)}\Big) := r(y|x) + r(x)$$

where x represents unconditional input, y represents conditional information, and q and p are the real and synthesized data distributions, respectively. When we have conditioning $y = [y_{cat}, y_{con}]$ such that y_{cat} is categorical and y_{con} is continuous, assuming that they are conditionally independent given x, we obtain through simple modification:

$$\begin{aligned}
f^*(x, y) &= log\Big(\frac{q(y_{cat}, y_{con}|x)q(x)}{p(y_{cat}, y_{con}|x)p(x)}\Big) \\
&= log\Big(\frac{q(y_{cat}, y_{con}|x)}{p(y_{cat}, y_{con}|x)}\Big) + log\Big(\frac{q(x)}{p(x)}\Big) \\
&= log\Big(\frac{q(y_{cat}|x)q(y_{con}|x)}{p(y_{cat}|x)p(y_{con}|x)}\Big) + log\Big(\frac{q(x)}{p(x)}\Big) \\
&= log\Big(\frac{q(y_{cat}|x)}{p(y_{cat}|x)}\Big) + log\Big(\frac{q(y_{con}|x)}{p(y_{con}|x)}\Big) + log\Big(\frac{q(x)}{p(x)}\Big) := r_{cat}(y|x) + r_{con}(y|x) + r(x),
\end{aligned}$$

with the remaining analysis following [16] leading to the projection-discriminator expression given in the main text.

Template Generator (T)
Inputs: conditions $z \in \mathbb{R}^c$
Embed $z \in \mathbb{R}^c$ into $\hat{z} \in \mathbb{R}^{64}$ using C
Learn Parameters $h \in \mathbb{R}^{80\times96\times80\times8}$
FiLM($\hat{z}$)
ConvSN, $8 \to 32$
$5\times$ ResBlockSN, $32 \to 32$
Upsample $2\times$ trilinearly
ConvSN, $32 \to 8$, FiLM($\hat{z}$), LeakyReLU
ConvSN, $8 \to 8$, FiLM($\hat{z}$), LeakyReLU
ConvSN, $8 \to 8$, FiLM($\hat{z}$)
ConvSN, $8 \to 1$, FiLM($\hat{z}$), tanh
Add to average of training images for $T(\hat{z})$

Registration Network (R)
Inputs: template $T(\hat{z})$; target F
h_0 : Concatenate($T(\hat{z})$, F)
h_1 : Conv, Stride 2, $2 \to 32$, LeakyReLU
h_2 : Conv, Stride 2, $32 \to 32$, LeakyReLU
h_3 : Conv, Stride 2, $32 \to 32$, LeakyReLU
h_4 : Conv, Stride 2, $32 \to 32$, LeakyReLU
h_5 : Conv, $32 \to 32$, LeakyReLU
h_6 : Conv, $32 \to 32$, LeakyReLU, Up $2\times$, concat h_3
h_7 : Conv, $32 \to 32$, LeakyReLU, Up $2\times$, concat h_2
h_8 : Conv, $32 \to 32$, LeakyReLU, Up $2\times$, concat h_1
h_9 : Conv, $32 \to 32$, LeakyReLU
h_{10} : Conv, $32 \to 32$, LeakyReLU
h_{11} : Conv, $32 \to 16$
v : ConvBlock, $16 \to 3$
φ : $5 \times$ Scale and Square(v)
$M(T(\hat{z}))$: STN($T(\hat{z}), \varphi$)

Discriminator (D)
Inputs: image $x \in \mathbb{R}^{160\times192\times160}$; attributes $z \in \mathbb{R}^c$
ConvSN, stride 2, $1 \to 64$, Leaky ReLU
ConvSN, stride 2, $64 \to 128$, Leaky ReLU
ConvSN, stride 2, $128 \to 256$, Leaky ReLU
ConvSN, stride 2, $256 \to 512$, Leaky ReLU
Conv, stride 1, $512 \to 64$ to $D'(x)$
Projection($D'(x)$, z)

Embedding/FiLM Generator (C)
Inputs: attributes $z \in \mathbb{R}^c$
Dense(64), LeakyReLU
Dense(64), LeakyReLU
Dense(64), LeakyReLU
Dense(64), LeakyReLU for $C(z)$

Table 1. Architectures for Conditional Predict-HD and dHCP consisting of a template generator (**top left**), a registration network (**top right**), a discriminator network (**bottom left**) and a FiLM embedding generator (**bottom right**). `Conv` represents a $3 \times 3 \times 3$ convolutional layer (`ConvSN` indicates use of spectral normalization). A `ResBlockSN` consists of two blocks of sequential `ConvSN` and `LeakyReLU` layers with an additive skip connection. For unconditional template estimation, we do not use any FiLM layers. For FFHQ-Aging, we use the same architectures, only replacing the 32 per-layer filters with 64 in the template generator (due to the higher number of classes), using `ConvSN` instead of `Conv` in the generator and the penultimate layer of the discriminator, and reducing the channel multiplier in the discriminator from 64 to 32.

References

[1] Vincent Arsigny, Olivier Commowick, Xavier Pennec, and Nicholas Ayache. A log-euclidean framework for statistics on diffeomorphisms. In Rasmus Larsen, Mads Nielsen, and Jon Sporring, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2006*, pages 924–931, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.

[2] John Ashburner. A fast diffeomorphic image registration algorithm. *NeuroImage*, 38(1):95 – 113, 2007.

[3] Brian B Avants, Paul Yushkevich, John Pluta, David Minkoff, Marc Korczykowski, John Detre, and James C Gee. The optimal template effect in hippocampus studies of diseased populations. *Neuroimage*, 49(3):2457–2466, 2010.

[4] M. Faisal Beg, Michael I. Miller, Alain Trouvé, and Laurent Younes. Computing large deformation metric mappings via geodesic flows of diffeomorphisms. *International Journal of Computer Vision*, 61(2):139–157, Feb 2005.

[5] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018.

[6] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. Rethinking atrous convolution for semantic image segmentation. *arXiv preprint arXiv:1706.05587*, 2017.

[7] Adrian Dalca, Marianne Rakic, John Guttag, and Mert Sabuncu. Learning conditional deformable templates with convolutional networks. In *Advances in neural information processing systems*, pages 806–818, 2019.

[8] Adrian V. Dalca, Guha Balakrishnan, John Guttag, and Mert R. Sabuncu. Unsupervised learning of probabilistic diffeomorphic registration for images and surfaces. *Medical Image Analysis*, 57:226 – 236, 2019.

[9] Vladimir Fonov, Alan C Evans, Kelly Botteron, C Robert Almli, Robert C McKinstry, D Louis Collins, Brain Development Cooperative Group, et al. Unbiased average age-appropriate atlases for pediatric studies. *Neuroimage*, 54(1):313–327, 2011.

[10] Ioannis S Gousias, A David Edwards, Mary A Rutherford, Serena J Counsell, Jo V Hajnal, Daniel Rueckert, and Alexander Hammers. Magnetic resonance imaging of the newborn brain: manual segmentation of labelled atlases in term-born and preterm infants. *Neuroimage*, 62(3):1499–1509, 2012.

[11] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium, 2018.

[12] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4401–4410, 2019.

[13] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[14] Antonios Makropoulos, Ioannis S Gousias, Christian Ledig, Paul Aljabar, Ahmed Serag, Joseph V Hajnal, A David Edwards, Serena J Counsell, and Daniel Rueckert. Automatic whole brain mri segmentation of the developing neonatal brain. *IEEE transactions on medical imaging*, 33(9):1818–1831, 2014.

[15] Antonios Makropoulos, Emma C Robinson, Andreas Schuh, Robert Wright, Sean Fitzgibbon, Jelena Bozek, Serena J Counsell, Johannes Steinweg, Katy Vecchiato, Jonathan Passerat-Palmbach, et al. The developing human connectome project: A minimal processing pipeline for neonatal cortical surface reconstruction. *Neuroimage*, 173:88–112, 2018.

[16] Takeru Miyato and Masanori Koyama. cGANs with projection discriminator. In *International Conference on Learning Representations*, 2018.

[17] Roy Or-El, Soumyadip Sengupta, Ohad Fried, Eli Shechtman, and Ira Kemelmacher-Shlizerman. Lifespan age transformation synthesis, 2020.

[18] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[19] Jane S Paulsen, Jeffrey D Long, Christopher A Ross, Deborah L Harrington, Cheryl J Erwin, Janet K Williams, Holly James Westervelt, Hans J Johnson, Elizabeth H Aylward, Ying Zhang, et al. Prediction of manifest huntington's disease with clinical and imaging measures: a prospective observational study. *The Lancet Neurology*, 13(12):1193–1201, 2014.

[20] Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron C Courville. Film: Visual reasoning with a general conditioning layer. In *AAAI*, 2018.

[21] Nicholas J Tustison, Philip A Cook, Andrew J Holbrook, Hans J Johnson, John Muschelli, Gabriel A Devanyi, Jeffrey T Duda, Sandhitsu R Das, Nicholas C Cullen, Daniel L Gillen, et al. Antsx: A dynamic ecosystem for quantitative biological and medical imaging. *medRxiv*, 2020.

[22] Hongzhi Wang, Jung W Suh, Sandhitsu R Das, John B Pluta, Caryne Craige, and Paul A Yushkevich. Multi-atlas segmentation with joint label fusion. *IEEE transactions on pattern analysis and machine intelligence*, 35(3):611–623, 2012.