This MICCAI paper is the Open Access version, provided by the MICCAI Society. It is identical to the accepted version, except for the format and this watermark; the final published version is available on SpringerLink.

Appendix

A Ablation Studies

We conduct five ablation studies to isolate and validate the individual architectural contributions of our two-stage framework.

Semi-Supervised vs. Fully Supervised Learning (Stage I): To isolate our semi-supervised recipe, we train the Stage I landmark detector strictly on the 50 labeled subjects, completely disabling the Mean Teacher consistency objective and pseudo-labeling. The resulting landmarks are passed through an identical Stage II engine. As shown in Table 3, while local surface fit metrics (P2S, Warp S2S) remain comparable, the supervised baseline exhibits a catastrophic degradation in downstream SSM metrics: Generalization error more than triples (3.402 mm vs. 1.112 mm) and Specificity significantly worsens (16.310 mm vs. 9.604 mm). Furthermore, the fully supervised model displays a $4\times$ increase in LLE variance ($\pm$2.005 mm vs. $\pm$0.542 mm), indicating severe overfitting. This landmark jitter disrupts topological consistency across the population, forcing the linear PCA space to encode alignment noise rather than true morphological anatomy.

Spectral vs. Spatial Deformation (Stage II): To isolate our Stage II Laplace Beltrami spectral framework, we replace it with a standard spatial baseline: Thin Plate Spline (TPS) deformation using the identical Stage I predicted landmarks as control points. As presented in Table 3, full SCALP outperforms the Stage I + TPS pipeline across every surface and statistical metric. Notably, the TPS baseline exhibits a major degradation in Specificity (19.034 mm vs. 9.604 mm) and nearly triples the P2S reconstruction error (4.340 mm vs. 1.478 mm). When confronted with large missing surface regions due to camera occlusion, TPS extrapolates unconstrained deformations outside the sparse landmark set, causing high-frequency wrinkling. Conversely, SCALP's band-limited eigenbasis imposes a hard geometric constraint that prevents high-frequency folding by construction.

Table 3: Ablation analyses isolating Stage I (Semi-supervised vs. Fully Supervised) and Stage II (Spectral vs. TPS Spatial warping).

Metric	SCALP	Stage I Sup.	Stage I + TPS
P2S (mm) ↓	**1.478 ± 0.497**	1.492 ± 0.604	4.340 ± 1.552
Warp S2S (mm) ↓	**1.456 ± 0.492**	1.469 ± 0.583	4.270 ± 1.531
LLE (mm) ↓	**3.520 ± 0.542**	3.929 ± 2.005	4.644 ± 1.113
Compactness ↑	**0.990**	0.984	0.986
Generalization (CD) ↓	**1.112**	3.402	1.487
Specificity (CD) ↓	**9.604**	16.310	19.034

Upstream Prediction vs. Oracle Landmarks: To decouple Stage I detector errors from the Stage II warping engine, we evaluate an "Oracle" configuration by replacing predicted landmarks on the 13 held-out test subjects with expert manual annotations. Note that because a population-wide SSM requires computing a PCA basis over the entire training set (where manual oracle annotations are intentionally omitted), global SSM metrics cannot be computed for this localized test ablation. As shown in Table 4, transitioning to manual landmarks

yields only a modest, incremental improvement (e.g., Warp S2S error decreases by merely ≈ 0.36 mm). This narrow performance margin demonstrates that our upstream Stage I detector has already achieved a high accuracy ceiling and is not a pipeline bottleneck, while confirming that Stage II successfully accommodates minor localized sub-millimeter landmark variances.

Table 4: Ablation analysis decoupling Stage I and Stage II error budgets on test subjects using manual expert Oracle Landmarks.

Metric	SCALP (Predicted)	Oracle Landmarks + Stage II
P2S (mm) ↓	1.478 ± 0.497	$\mathbf{1.106 \pm 0.406}$
Warp S2S (mm) ↓	1.456 ± 0.492	$\mathbf{1.096 \pm 0.410}$
LLE (mm) ↓	3.520 ± 0.542	$\mathbf{3.203 \pm 0.494}$

Tuning of Spectral Basis Dimensionality r: The truncation rank r of the LBO eigenfunctions dictates the deformation space's expressiveness. We evaluate our pipeline across five capacities: $r \in \{64, 100, 120, 150, 200\}$. As presented in Table 5, increasing r from 64 to 200 steadily reduces individual surface fitting errors (P2S and Warp S2S). However, optimizing exclusively for surface alignment causes a severe regression in population-level SSM utility. Beyond $r = 120$, Generalization and Specificity metrics deteriorate significantly (e.g., Generalization increases from 1.112 mm at $r = 120$ to 3.041 mm at $r = 200$). This structural behavior confirms that excessive high-frequency spectral modes enable the model to encode localized acquisition noise into the correspondences, corrupting the downstream linear PCA shape space. Selecting $r = 120$ establishes the optimal equilibrium.

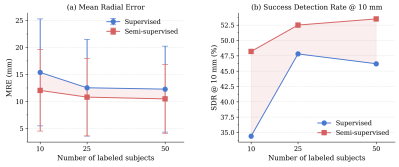


Fig. 5: Data efficiency ablation curves on the test cohort for varying quantities of labeled training subjects ($N \in \{10, 25, 50\}$).

Table 5: Ablation analysis on the LBO spectral mode truncation rank r.

Metric	$r = 64$	$r = 100$	SCALP ($r = 120$)	$r = 150$	$r = 200$
P2S (mm) ↓	1.150 ± 0.408	1.102 ± 0.394	1.478 ± 0.497	1.081 ± 0.386	$\mathbf{1.062 \pm 0.376}$
Warp S2S ↓	1.134 ± 0.406	1.089 ± 0.393	1.456 ± 0.492	1.068 ± 0.385	$\mathbf{1.049 \pm 0.379}$
LLE (mm) ↓	3.469 ± 0.505	3.416 ± 0.518	3.520 ± 0.542	$\mathbf{3.380 \pm 0.496}$	3.384 ± 0.508
Compactness ↑	0.985	0.985	**0.990**	0.985	0.985
Generalization ↓	2.758	2.982	**1.112**	2.969	3.041
Specificity ↓	12.163	11.864	**9.604**	12.254	12.295

Annotation Data Efficiency: We evaluate framework resilience under sparse data constraints by varying the labeled training size ($N \in \{10, 25, 50\}$; Fig. 5). While the supervised baseline degrades steeply—its Success Detection Rate (SDR@10) plummeting to 34.4% at $N = 10$ with high variance, our semi-supervised approach degrades gracefully. At $N = 10$, SCALP achieves an SDR@10 of 48.2%, outperforming the supervised baseline trained on over double the data ($N = 25$) while narrowing error variance. This demonstrates that consistency regularization effectively leverages unlabeled cohorts to minimize manual annotation bottlenecks.