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

A1 CT Pseudo-Labels and Leakage Boundary

The unlabeled CT directory of MVAA [1] contains 1,040 volumes. Two volumes with undefined orientation entries in their affines are removed, leaving 1,038 inputs. A five-fold 3D nnU-Net [2] teacher ensemble predicts one mask for every valid volume. Candidate masks are filtered and ranked without unlabeled ground truth as follows. For mask y, let $L(y)$ be its largest 26-connected component, $f(y) = |L(y)|/|y|$, and $V(y)$ its physical volume. A candidate is eligible when $f(y) \geq 0.94$ and

$$0.45\, Q_{0.10}(V_{\text{lab}}) \;\leq\; V(y) \;\leq\; 1.75\, Q_{0.90}(V_{\text{lab}}), \tag{A1}$$

which evaluates to 972.38–6,761.54 mm^3; 947 masks are eligible. They are ranked by

$$r(y) = f(y) - 0.05 \left| \log \frac{V(y)}{\text{median}(V_{\text{lab}})} \right|, \tag{A2}$$

and the top 300 are retained after largest-component projection. Each student fold trains on 21 or 22 original training cases plus all 300 pseudo cases and validates only on its 5 or 6 original held-out cases.

What is and is not nested. The surface-transition model and its optimization labels exclude the corresponding outer cases. The pseudo-label teacher, however, is fitted across all 27 labeled cases before labeling the separate unlabeled pool. The outer label is therefore never copied into a pseudo target for the same patient, but it can influence the common teacher and hence every fold's frozen anchor. We accordingly limit the claim to a paired, case-disjoint transition comparison relative to a frozen anchor; the complete CT training chain is not fully nested. A fully nested reproduction would regenerate the 300 pseudo-labels with an outer-specific teacher and retrain all five anchors.

A2 Development Consequences of the Acceptance Rules

The gate conditions are given in Sec. 2.3 of the main text; here we report their development-time consequences and the model-selection procedure.

CT. With identical mixed action logits and an add-only executor, the pure gate-off control ($\theta = 0$) improves pooled Dice by 0.00716 but reduces Dice on fold 2 and therefore fails the $N_+(d_f^D) = 5$ condition. Threshold $\theta = .85$ passes all conditions and retains 27.7% of the gate-off edits. It is the fold-consistent deployment choice, not the maximum-Dice setting.

Video. Both $\alpha = .40$ and $\alpha = .50$ pass the retrospective development rule. The latter is reported for its larger Dice and ASD gains; the former yields the best HD and was used in the later hidden candidate. Its hidden regression (Sec. A9) shows that passing this rule did not guarantee external transfer.

TEE. At strength .10 the case-count condition $N_+(d_k^D) \geq 12$ is met exactly (12/21), but the mean delta is $-0.0000149 < 0.002$, so training stops after the fold-0 audit and deployment selects the null action.

Checkpoint and threshold selection. Gradient training excludes the outer case or video, but model selection is not nested. For CT folds 0–4, seed 3407 selects epochs 40,

Table A1. CT five-fold gate ablation. Both settings use the same mixed action logits, add-only executor, and removal scale 0, differing only in confidence threshold (0 vs. .85). Deltas are candidate minus anchor: positive ΔDSC and negative ΔHD/ΔASD are favorable.

Fold	n	Anchor DSC	Anchor HD	Anchor ASD	Gate-off ($\theta = 0$)			Gate on ($\theta = .85$)		
					ΔDSC	ΔHD	ΔASD	ΔDSC	ΔHD	ΔASD
0	6	.89332	3.35962	.12698	+.01059	−.26149	−.01325	+.00623	−.10199	−.00901
1	6	.88514	3.83331	.13374	+.00707	−.20273	−.01026	+.00459	−.20722	−.00869
2	5	.89461	2.97047	.11430	−.00217	−.00240	+.00282	+.00380	−.18451	−.00785
3	5	.88800	2.88567	.13736	+.00934	+.47045	−.01013	+.00583	+.31821	−.00858
4	5	.89259	4.04647	.12898	+.01032	−.34961	−.00889	+.00686	−.18549	−.00844
All	27	.89062	3.43225	.12843	+.00716	−.08123	−.00822	+.00546	−.07831	−.00854

20, 10, 15, and 60; seed 1701 selects 30, 25, 5, 5, and 55. Video folds 0–5 select epochs 45, 40, 23, 21, 45, and 42, with segmentation thresholds τ of .60, .60, .70, .34, .70, and .34. Pooled CT predictions compare action mixtures and confidence thresholds, including the gate-off control; pooled video predictions compare fusion weights. The gate criteria are thus evaluated once per candidate policy and inference uses no label, but the same outer labels select checkpoints, thresholds, and the reported common operating point. No independent calibration cohort remains.

A3 Intervention Errors and Opportunity Costs

Because g_m is one model-level switch, conventional per-case false-acceptance and false-rejection rates are undefined. The three development decisions agree with their observed mean outcomes (CT and video accepted, TEE rejected), but this resubstitution count is optimistically biased and is not an error estimate. The later official candidate provides one external policy-level observation: video $\alpha = .40$ passed the development rule but decreased hidden Dice from 0.7701 to 0.7567. We count this as one observed false acceptance, not an estimable rate, and disable the video transition in final deployment.

We instead report case-level consequences of each fixed policy. The accepted CT transition lowers Dice in 1/27 cases (3.7%), raises HD in 5/27 (18.5%; 11 ties), and raises ASD in 1/27 (3.7%). The development-accepted video transition ($\alpha = .50$) lowers neither Dice nor HD on any held-out video but raises ASD on 2/6 (33.3%). Conversely, the rejected TEE transition would improve Dice in 12/21 pilot cases (57.1%) despite a negative mean effect. This descriptive opportunity count illustrates the limitation of a coarse cohort-level null decision and motivates a future, independently calibrated per-case risk model.

A4 Gate-Off and Gate-On Ablations

The gate reduces edited voxels from 125,650 to 34,827 and sacrifices part of the pooled Dice gain, but repairs the fold-2 Dice and ASD regressions so that both metrics improve in all five folds. It does not repair the fold-3 HD regression. Its purpose is fold-consistent bounded intervention rather than development-set Dice maximization.

With the gate g_{VID} disabled, the candidate is always applied. With it enabled, the common candidate passes and therefore produces the same six outputs; the difference is

Table A2. Video LOVO development ablation. Each held-out video is excluded from gradient training, but its labels select the checkpoint and threshold; surface metrics use foreground frames only. Improvements are anchor minus candidate for HD/ASD and candidate minus anchor for DSC, so positive is favorable. The pooled $\alpha = .40$ row is included because that setting was used in the later hidden candidate.

Video	n	Anchor			StateMV, $\alpha = .50$			Improvement		
		DSC	HD	ASD	DSC	HD	ASD	ΔDSC	ΔHD	ΔASD
1	21	.78560	139.13	79.79	.79142	130.60	78.65	+.00582	+8.53	+1.14
2	20	.83502	97.50	12.95	.83848	96.50	13.32	+.00346	+1.00	−0.37
3	23	.81845	107.55	15.04	.82147	105.60	14.21	+.00301	+1.95	+0.83
4	19	.81562	87.14	14.90	.81644	84.52	15.08	+.00082	+2.63	−0.18
5	16	.82954	60.21	8.99	.83227	58.19	8.82	+.00273	+2.02	+0.17
6	19	.61022	200.33	44.30	.62360	191.41	42.26	+.01338	+8.91	+2.04
Pooled, $\alpha = .50$	118	.7829	116.70	30.08	.7878	112.50	29.46	+.0049	+4.20	+0.62
Pooled, $\alpha = .40$	118	.7829	116.70	30.08	.7875	112.37	29.55	+.0045	+4.33	+0.52

Table A3. Exploratory paired summaries. Improvements are candidate minus anchor for Dice and anchor minus candidate for distances. CIs use 200,000 bootstrap resamples. Raw one-sided Wilcoxon p-values are followed by Holm adjustment across the six rows.

Task	Metric	Mean improvement	95% CI	Raw p	Holm p	Wins/units
CT	DSC	.005459	[.003770, .007318]	1.49×10^{-8}	8.94×10^{-8}	26/27
CT	HD	.078305	[−.156150, .307121]	.1388	.2775	11/27
CT	ASD	.008538	[.005701, .011530]	6.56×10^{-7}	3.28×10^{-6}	26/27
Video	DSC	.004872	[.002263, .008573]	.0156	.0625	6/6
Video	HD	4.1733	[1.7574, 6.6531]	.0156	.0625	6/6
Video	ASD	.6030	[−.0262, 1.2950]	.1563	.2775	4/6

the availability of an exact anchor fallback if the cohort-level criteria fail. The fixed-area TopK projection is a hard executor constraint, independent of the gate.

A5 Paired Statistical Analysis

Outer-label model selection and operating-point reuse mean that these are conditional descriptive summaries even after multiplicity adjustment. For video, the two-sided exact Dice/HD values are .03125, and the one-sided values are the minimum attainable with six favorable nonzero pairs. For CT, treating the five model-selection folds as clusters gives one-sided exact $p = .03125$ and two-sided $p = .0625$ for Dice and ASD; fold-level HD improves in 4/5 folds but does not reach significance. The patient-level and fold-level views together support a stable development effect but not confirmatory inference for a newly selected procedure.

A6 TEE Failure Cases and Null Dynamics

The fold-0 base is 0.846932 Dice. Strength .10 obtains 0.846917 with 12/21 case wins; strength .50 obtains 0.845899 with 9/21 wins; the epoch-10 strength-.50 checkpoint is

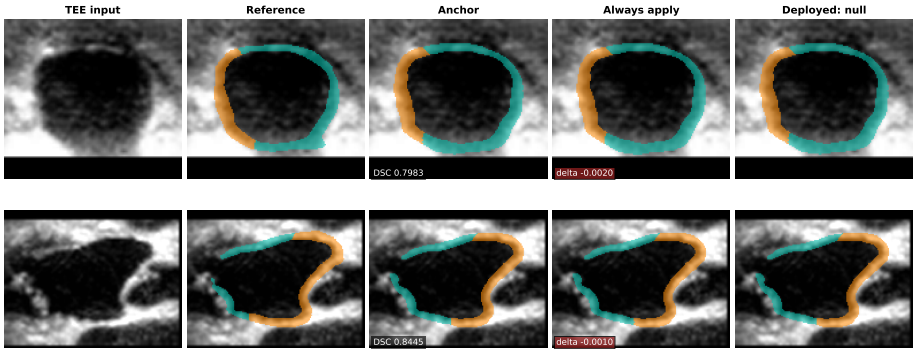


Fig. A1. Two fold-0 TEE failures reconstructed from the archived checkpoint and OOF cache. Cyan and orange denote the two leaflet labels. The least aggressive transition decreases whole-volume class-mean Dice by 0.0020 and 0.0010. Because the cohort mean is also negative, StateMV deploys the exact anchor (right column). Display slices maximize anchor/candidate disagreement and are not used for selection.

0.003787 below base with six wins. The best checkpoint is epoch 1, and the pilot stops after epoch 10. Fig. A1 illustrates how a near-zero mean can conceal coherent local class-boundary changes.

The failure has three plausible, testable causes. First, the multi-plane consensus is already a low-variance posterior, so the residual signal is weak. Second, the anterior and posterior leaflets are adjacent thin sheets: a globally FiLM-conditioned dense action can exchange class identity while remaining geometrically plausible. Third, single-fold training observes only 84 volumes, and the errors requiring correction are sparse. A successor should use two semantic leaflet slots, explicit mutual-exclusion and ordering constraints, and connected surface-patch actions. These are hypotheses, not results of the present study.

A7 Non-deployed Representation Audits

These diagnostics do not alter the reported executor results. They comprise three separately trained components rather than one end-to-end checkpoint. The 6.251M-parameter *unified action model* has modality adapters, a 128-D SDF state, a shared 10-D action-conditioned transition, and modality-specific SDF decoders; CT/TEE use controlled no-wrap spatial actions and video uses observed frame gaps. The separate 0.105M-parameter *semantic observation model* maps posterior geometry and quality through a 128-D belief to 24 valve measurements. A 0.102M-parameter canonical-field model decodes semantic beliefs on a 16^3 grid. Outer task cases and trajectories are excluded from the corresponding representation-training fold.

Positive semantic MAE reduction on all five video audits does not imply calibrated target presence: specificity collapses outside fold 0. CT is dominated by one strong fold and has two negative folds; TEE gains are near zero with two marginally negative folds. Independently, dense transfer from the diagnostic unified-action checkpoint changes CT fold-0 Dice by $+0.000011$ (neutral), TEE fold-0 Dice by -0.000003 with a small surface regression (rejected), and video-6 Dice by -0.002520 despite fewer background false

Table A4. Semantic observation model: outer-fold relative reduction in 24-D state MAE versus the unrefined observation. Positive is favorable. Video always uses held-out video 6 while CT/TEE outer folds rotate; this is not a complete factorial cross-product.

CT/TEE fold	CT	TEE	Video	Video sensitivity	Video specificity
0	+49.558%	+0.055%	+3.098%	1.00	.091
1	−0.000004%	−0.000112%	+4.839%	1.00	.000
2	−1.308%	+0.181%	+1.814%	1.00	.000
3	+0.135%	+0.160%	+1.444%	1.00	.000
4	+0.046%	−0.137%	+1.283%	1.00	.000

Table A5. Incremental inference cost. Counts exclude the frozen task anchor unless stated.

Component	Parameters	Relative parameter cost	Deployment consequence
CT one transition seed	1.176M	1.33% of 88.21M anchor	one extra 3D pass
CT two-seed transition	2.353M	2.67% of one anchor	two extra 3D passes
TEE null transition	0	0	exact anchor
Video core state module	0.552M	0.93% of 59.64M candidate	encoder/MLP/conditioning
Video full non-backbone addition	0.964M	1.62% of 59.64M candidate	includes topology head
Unified action audit	6.251M	diagnostic only	not deployed
Semantic belief audit	0.105M	diagnostic only	not deployed
Canonical field audit	0.102M	diagnostic only	not deployed

positives (rejected). Canonical-field transfer is likewise retained only as a diagnostic. We therefore preserve the task-specific identity/executor paths. A shared tensor is not, by itself, evidence of a shared causal anatomical factorization.

A8 Computation and Simpler Alternatives

The development-selected video route evaluates six state-transition candidates in addition to four appearance experts; CT adds two compact 3D transition passes. Because no controlled, repeated wall-clock benchmark is available, we report architectural overhead and inference-pass counts rather than a timing claim.

The video anchor already uses identity/horizontal/vertical/both-flip TTA, and CT and TEE use five-fold ensembles, so TTA and ensembling are not omitted simple baselines. We did not retrain every anchor with additional compute matched to the transitions; the present evidence cannot establish that StateMV is more efficient than additional training epochs. The architectural justification is instead falsifiable control: the action vocabulary, feasible set, null action, and exact edit count expose what changed and prevent an unconstrained second decoder from replacing correct regions.

A9 Official Challenge Results and Provenance

The earlier container preceded the CT surface-transition and video predictive-state experiments. The later candidate supplies external container-level evidence: CT improves

Table A6. Official challenge records. The later candidate contains CT gate .85, TEE consensus, and video $\alpha = .40$. Only Dice was archived for its hidden run, and its TEE and video public-validation runs were not archived. Container-level rows are not pooled with OOF/LOVO development results. *Official video HD/ASD follow the challenge server's handling of empty masks; per-case distances were not archived.

System	Split	Task	DSC	HD	ASD
Earlier container	Public validation	CT	.8603	4.5296	.2709
		TEE	.8575	9.6872	.5419
		Video	.8165	73.7989	11.2994
Earlier container	Hidden test	CT	.8383	6.4336	.5796
		TEE	.8558	9.1660	.5418
		Video	.7701	472.3967*	333.8341*
Later candidate	Public validation	CT	.8617	4.3880	.2647
Later candidate	Hidden test	CT	.8393	–	–
		TEE	.8580	–	–
		Video	.7567	–	–

slightly on public and hidden Dice, the TEE consensus improves slightly on hidden Dice, and video loses 0.0134 Dice relative to the earlier stable route. Because the available official record contains neither per-case predictions nor later distance metrics, no paired test or failure rate can be computed. The final deployment configuration disables all tested video world routes and returns to the hidden-stable appearance ensemble.

References

1. Hammad, M., Bai, J., Zhao, J., Jia, T., Zhang, X., Xu, X., Ma, J., Li, S.: Mitral valve anatomy analysis using multimodal imaging data. MICCAI 2026 Challenge (2026). https://doi.org/10.5281/zenodo.19726755
2. Isensee, F., Jaeger, P.F., Kohl, S.A.A., Petersen, J., Maier-Hein, K.H.: nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. Nature Methods **18**(2), 203–211 (2021). https://doi.org/10.1038/s41592-020-01008-z