PRETRAINED SYSTEMS FOR LOW-DATA MYOCARDIAL SEGMENTATION: DIRECT VERSUS LV-SUPERVISED FINE-TUNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Choosing a pretrained system is important when myocardial annotations are limited. We compare Pixio-L with Mask2Former and EchoFM with UNETR for myocardial (MYO) segmentation on CAMUS apical four-chamber (A4C) echocardiography, evaluated against official provider-supplied reference annotations. Pixio-L is a released student distilled from the web-scale Pixio-5B teacher; EchoFM is pretrained on echocardiographic videos. Each system starts from its released encoder weights and is trained either directly on MYO or first on EchoNet-Dynamic A4C left ventricular (LV) cavity segmentation and then on MYO. Under full fine-tuning, across target-training patient fractions of 5%, 10%, 25%, 50%, and 100% and three training seeds, Pixio-L achieves higher mean MYO Dice under both routes. At 5%, the cross-system differences favoring Pixio-L are 13.38 percentage points for Direct fine-tuning and 6.52 points after LV supervision. The LV stage improves mean Dice in both systems at every fraction, with the largest gains at 5%: 2.29 points for Pixio-L and 9.15 for EchoFM. An additional 60-run frozen-MYO-encoder ablation shows that the effect of encoder updates depends on the system and initialization route. These exploratory results on one fixed patient split identify Pixio-L with Mask2Former as a strong low-data baseline and support intermediate LV supervision for both systems. Differences in architecture, input, and training recipe prevent attributing the ranking to pretraining domain or distillation alone.

1 INTRODUCTION

Echocardiographic segmentation supports measurements of chamber size and ventricular function, but its usefulness depends on delineating the anatomy required by the downstream analysis (Lang et al., 2015). Available supervision does not always match that target. EchoNet-Dynamic provides left ventricular (LV) cavity tracings, whereas CAMUS includes myocardium (MYO) annotations (Ouyang et al., 2020; Leclerc et al., 2019). A cavity mask cannot directly replace a myocardial mask: the latter must identify tissue between the endocardial and epicardial boundaries. When MYO labels are limited, an existing LV dataset offers a possible intermediate source of supervision. Whether to use it depends both on the transfer route and on the system being adapted.

Pretrained representations provide several starting points for this problem. Masked autoencoders demonstrate that reconstructing missing image content can support transferable visual representations (He et al., 2022), and Pixio develops pixel-based pretraining and releases distilled encoders (Yang et al., 2026; FAIR at Meta and HKU, 2025). Echocardiographic models incorporate different domain information: EchoCLIP aligns images with clinical reports, while EchoFM learns spatial and temporal video representations (Christensen et al., 2024; Kim et al., 2025). These approaches differ in supervision, input, and adaptation requirements. Their availability makes model selection consequential, but a domain-specific pretraining objective alone does not determine the final quality of a particular segmentation system.

Transfer benefits depend on the task and optimization: Raghu et al. (2019) found model- and data-dependent gains in medical classification, while He et al. (2019) showed that suitable normalization and sufficient optimization could match ImageNet initialization on detection and instance segmen-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

tation. In echocardiography, low-label LV benefits vary with pretraining data and objective (Saeed et al., 2022). These findings motivate evaluating final MYO performance separately from the paired benefit of intermediate LV supervision.

EchoNet-Dynamic-to-CAMUS transfer changes both the dataset and foreground. Anatomical similarity motivates intermediate LV supervision without establishing its mechanism. Staged adaptation has broader precedents, including ProPETL (Zhou et al., 2025); our comparison evaluates the complete LV-supervised stage, including its additional labels and full-model optimization.

Low-label evaluation and LV-to-MYO transfer have direct precedents: EchoFM evaluates labeled-data fractions, and ETAB includes EchoNet-Dynamic apical four-chamber (A4C) LV-to-CAMUS A4C MYO (Kim et al., 2025; Alaa et al., 2022). We distinguish final performance from LV-route gain: a larger gain need not yield a higher final score. We also test whether updating the encoder during MYO training improves on training only the segmentation modules.

We compare Pixio-L with Mask2Former and EchoFM with UNETR (Cheng et al., 2022; Hatamizadeh et al., 2022). Pixio-L is a released 303-million-parameter student distilled from Pixio-5B; EchoFM supplies an echocardiographic video encoder (Yang et al., 2026; FAIR at Meta and HKU, 2025; Kim et al., 2025). Direct trains on CAMUS A4C MYO; LV-to-MYO first trains on EchoNet-Dynamic A4C LV. We conduct 120 MYO training runs across two systems, two routes, two MYO adaptation modes, five target-training patient fractions, and three training seeds: 60 with full fine-tuning and 60 with the encoder frozen during MYO training. Within each system, both modes reuse the same full-fine-tuned LV checkpoints for the LV-supervised route. Comparisons concern complete systems, including their decoder, input, and recipe.

We pair the evaluation patients across systems and routes, reporting Dice together with boundary, sensitivity, and precision metrics. This matters because a higher cohort-mean overlap can coexist with missed foreground or worse predictions for individual patients (Maier-Hein et al., 2024). Seed repetitions describe fitting variability on the chosen cohorts; patient-paired analyses describe the distribution of outcomes for those fitted models. The main findings are:

1. **Pixio-L with Mask2Former is a strong low-data MYO baseline.** Under full fine-tuning, Pixio-L achieves higher mean MYO Dice than EchoFM under both routes at all five target-training fractions. At 5%, its advantage is 13.38 percentage points with Direct and 6.52 points with LV-to-MYO.
2. **LV supervision improves MYO segmentation across the target-data range.** Under full fine-tuning, both systems achieve higher mean Dice at every fraction, with the largest mean gains at 5%: 2.29 percentage points for Pixio-L and 9.15 for EchoFM. Gains remain positive at full target-training availability.
3. **The effect of encoder freezing depends on the system and route.** The largest observed change is a 9.30-point Dice decrease for Pixio-L Direct at 5%; freezing does not uniformly reduce performance across the evaluated settings.

These exploratory comparisons use one fixed patient split; cohort-level improvements need not benefit every patient.

2 Related Work

Pretrained representations and anatomical supervision. Alongside the pretrained representations discussed above, supervised anatomy provides transferable features: SuPreM evaluates pretraining on anatomical labels and tumor pseudo-labels for downstream segmentation in three-dimensional CT (Li et al., 2024). Its results motivate investigating anatomical supervision, but do not establish LV-to-MYO transfer in ultrasound. We compare two released systems rather than isolating their pretraining objectives or measuring teacher–student retention.

Low-label echocardiography and adaptation protocols. Saeed et al. (2022) report objective- and data-dependent benefits of self-supervised pretraining for low-label LV segmentation. EchoFM evaluates labeled-data fractions and describes a frozen encoder with task-specific adapters (Kim et al., 2025). Our frozen-encoder ablation trains each system's segmentation modules and does

not reproduce that adapter protocol. Chao et al. (2025) compare LV segmentation systems, with SAM-family models using ground-truth-derived boxes; their task and interaction differ from our prompt-free MYO setting. SAMUS adapts SAM to ultrasound, and AutoSAMUS generates prompts automatically (Lin et al., 2024). EchoPilot performs training-free video segmentation using a first-frame point and anatomical category (Xiao et al., 2026). Numerical comparisons require matched anatomy, patient partitions, and interaction protocols; independent clinical evaluation remains beyond our fixed-cohort study (Zhang et al., 2022).

Intermediate adaptation and LV-to-MYO transfer. ProPETL derives an intermediate task from downstream dense annotations and uses parameter-efficient adaptation before semantic segmentation (Zhou et al., 2025). We share the broad question of intermediate-task utility, but do not use its adapter design or claim parameter efficiency. ETAB provides the closer task precedent: its Table 3 includes EchoNet-Dynamic A4C LV-to-CAMUS A4C MYO with ResNet-50, several segmentation heads, and 32 target examples (Alaa et al., 2022). Its Fig. 4 includes a separate CAMUS A4C LV-to-MYO experiment where source pretraining did not improve on target training from scratch. Thus, the task pair is established, while its benefit remains setting-dependent. Our comparison starts from released Pixio-L or EchoFM weights and measures final performance, paired LV-route gains, and the effect of freezing the target-stage encoder across five matched training-patient fractions, using existing architectures and losses.

3 METHOD

3.1 PRETRAINED SEGMENTATION SYSTEMS

We evaluate Pixio-L with Mask2Former (Yang et al., 2026; Cheng et al., 2022) and EchoFM with UNETR (Kim et al., 2025; Hatamizadeh et al., 2022). Pixio-L is a released student distilled from Pixio-5B and processes individual frames; EchoFM uses an echocardiographic video encoder to process multi-frame clips. Both systems load their released encoder weights without additional pretraining or distillation. Input dimensions and optimization settings are specified in Section 4.1.

3.2 DIRECT AND LV-SUPERVISED FINE-TUNING

Both routes target CAMUS A4C MYO segmentation (Figure 1). *Direct* fine-tunes on MYO from the released encoder and newly initialized segmentation modules. The *LV-supervised* route first performs supervised fine-tuning on EchoNet-Dynamic A4C LV cavity segmentation, then downstream fine-tuning on MYO. LV-to-MYO denotes this route, labeled Transfer in the tables.

For the LV-supervised route, source-validation Dice selects the LV checkpoint. Each MYO run loads the complete LV model state for its system and seed, including the encoder, segmentation modules, and compatible binary output head. The foreground changes from LV cavity to MYO, but the head is retained. Each target-training fraction starts independently from this checkpoint, with a new optimizer and no source optimizer state. Both routes produce a single MYO model at inference, without a separate LV prediction or an LV-derived crop.

3.3 TARGET-STAGE ADAPTATION AND MATCHED COMPARISONS

We evaluate both routes under two MYO adaptation modes. *Full fine-tuning* updates all model parameters and constitutes the primary comparison. The *frozen-encoder* ablation keeps the encoder in evaluation mode, with its parameters and buffers unchanged. Pixio-L trains the feature pyramid and Mask2Former; EchoFM trains UNETR and its output head. LV source training always updates the full model, and its checkpoints are reused for both MYO modes. Frozen LV-supervised runs therefore also retain LV-trained segmentation modules; this is not an encoder-only linear probe.

The study comprises **120 MYO training runs**: two systems, two routes, two MYO adaptation modes, five target-training patient fractions, and three training seeds. These comprise 60 full-fine-tuning runs and 60 frozen-encoder runs, excluding LV source training.

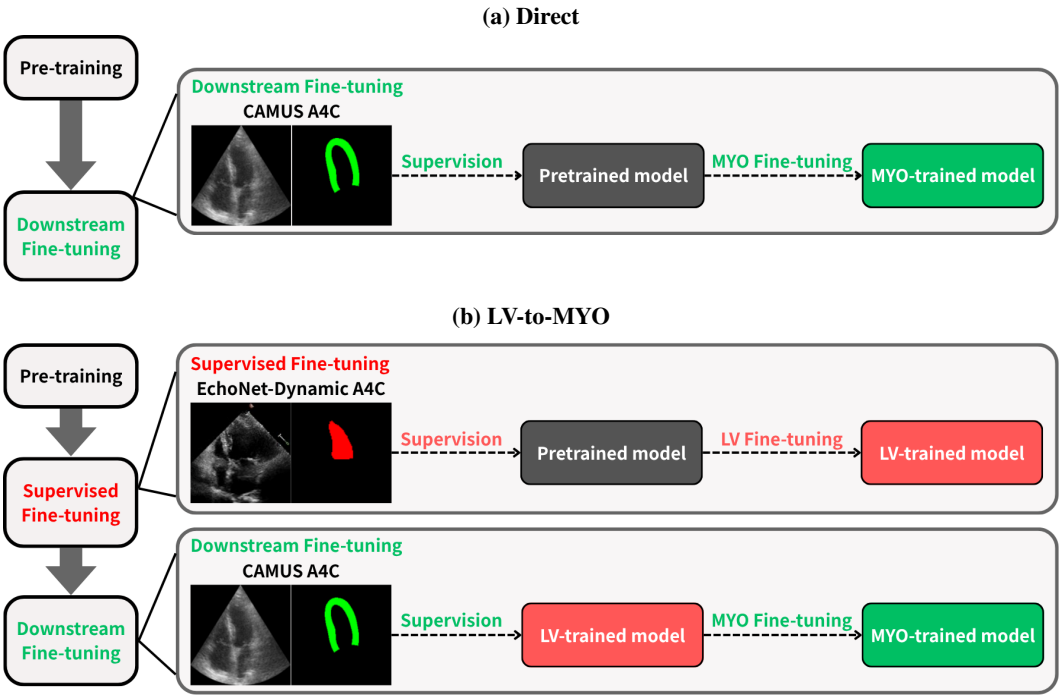


Figure 1: Training routes for Pixio-L with Mask2Former and EchoFM with UNETR. (a) Direct starts MYO fine-tuning on CAMUS A4C from released encoder weights and new segmentation modules. (b) LV-supervised fine-tuning first uses EchoNet-Dynamic A4C LV supervision, then starts MYO fine-tuning from the selected LV checkpoint. Each route is evaluated with full fine-tuning and a frozen encoder during MYO training; LV source training always updates the full model. "Pre-training" denotes released weights, not a training stage performed in this study.

Within each system and adaptation mode, routes share target-patient subsets, training seeds, and the configured target-training recipe. Target-validation Dice selects each checkpoint; validation-based stopping can yield different update counts. All configurations are evaluated on the same test patients. Route contrasts measure the benefit of the complete LV-supervised stage, including its additional source labels and optimization; they do not isolate encoder transfer or establish that retaining the head is preferable to reinitialization.

4 Experiments

4.1 Experimental setup

Cohorts and training fractions. The processed EchoNet-Dynamic cohort has 6,998/1,750/1,276 training/validation/test studies, each with two LV reference frames; these are processed-cohort counts, not the full published dataset (Ouyang et al., 2020). For CAMUS A4C (Leclerc et al., 2019), our fixed Fold 1 partition contains 360/90/50 training/validation/test patients. Fractions of 5/10/25/50/100% contain 18/36/90/180/360 training patients and 334/702/1,768/3,568/7,176 frames. Nested subsets are drawn once with seed 42 and shared across systems, routes and fitting seeds 42, 777 and 2026. Each adaptation mode comprises $2 \times 2 \times 5 \times 3 = 60$ target runs, totaling 120, excluding source training. Seed repeats assess fitting variability on one split, not independent cohort draws.

Data and reference annotations. We use images and reference masks supplied in the official EchoNet-Dynamic and CAMUS releases (Ouyang et al., 2020; Leclerc et al., 2019), without creating new annotations. We evaluate all 976 frame–reference pairs in our processed test manifest from 50 patients, including frames beyond end-diastole (ED) and end-systole (ES).

Table 1: Training configurations for the 60 primary target runs. This table describes full fine-tuning; frozen-encoder runs instead train 20.68 M and 12.53 M parameters, respectively. M denotes 10^6. Batch size counts images for Pixio-L and clips for EchoFM. Early stopping warmup counts validation evaluations, separately from learning-rate warmup. EchoFM's source stopping rule was amended; target runs use the rule shown.

Setting	Pixio-L + Mask2Former	EchoFM + UNETR
Input (frames $\times H \times W$)	$1 \times 256 \times 256$	$32 \times 224 \times 224$
Trainable parameters (M)	324.06	318.20
Physical batch / accumulation	256 / 1	4 / 64
Peak decoder learning rate	10^{-4}	10^{-4}
Peak encoder learning rate	10^{-5}	10^{-5}
Weight decay (base)	10^{-4}	10^{-4}
LR warmup / nominal epoch cap	10 / 100	10 / 100
Early stopping warmup / patience	60 / 20 validations	20 / 10 validations
Minimum Dice improvement	0.01 pp	0.01 pp
Training augmentation	Horizontal flip, $p = 0.5$	Clip-consistent flip, $p = 0.5$

Data separation and model selection. Archived manifests confirm disjoint training, validation and test patients. Test data were excluded from supervised fine-tuning; validation Dice alone selected checkpoints. Original test scores were used for reporting, not to select learning rates, stopping rules, seeds or retained models. The subsequent EchoFM learning-rate/weight-decay revisions and frozen ablation are post-hoc, following availability of those scores. We report the complete revised 30-run EchoFM matrix without per-fraction recipe or seed selection. This is not independent confirmatory evaluation; released-weight provenance is a separate issue (Section 5).

Inputs and prediction. Pixio-L processes individual 256×256 RGB frames. EchoFM processes 32-frame 224×224 clips assembled from numerically ordered frames of the same patient, with endpoint repetition at sequence boundaries. Only anchor slot 16 (zero-based) is supervised and scored; the neighboring frames supply unlabeled temporal context. Missing numeric frame positions are not filled. Positional embeddings index temporal patch positions within each clip rather than measured inter-frame times; temporal gaps are not explicitly modeled. Both systems predict background/MYO masks on a common 224×224 evaluation grid. Images and prediction scores are resized bilinearly, references by nearest neighbor, and foreground predictions follow background/MYO argmax.

Optimization and checkpoint selection. Both systems use AdamW with the target settings in Table 1. Base weight decay is 10^{-4}, with zero decay for one-dimensional parameters and designated tokens/embeddings. Pixio-L uses linear warmup and stepped decay; EchoFM uses linear warmup and cosine decay. Their losses and full schedules are specified in Appendix D.

Validation occurs approximately once per epoch; maximum validation foreground Dice selects the checkpoint. Early stopping follows Table 1. EchoFM's reused LV runs changed from 60/20 to 20/10 validation evaluations during continuation, retaining prior progress; target runs use 20/10. Source and target recipes and recorded costs are distinguished in Appendix D.

4.2 Metrics and paired analysis

Patient-mean Dice (Dice, 1945) is the primary endpoint. IoU (Jaccard, 1901), average symmetric surface distance (ASSD), 95th-percentile Hausdorff distance (HD95), sensitivity, and positive predictive value (PPV) describe complementary errors (Maier-Hein et al., 2024). Distances are pixels on the 224-grid, not millimeters. All 120 validation-selected target checkpoints use the same metric definitions and implementation, preserving each system's inference recipe. HD95 is the maximum of the two directed 95th percentiles for both systems (Appendix D.3). No evaluated target frame had an empty prediction or reference.

Table 2: MYO Dice (%) by target-training fraction and MYO adaptation mode. Values are means ± sample SD across three seeds on the same 50 test patients. Both modes use target base weight decay 10^{-4}; LV-to-MYO reuses full-fine-tuned LV models. Route gains and paired intervals are reported in Appendix B, and frozen-minus-full contrasts in Appendix G.

		Full fine-tuning		Frozen encoder	
Fraction	System	Direct	LV→MYO	Direct	LV→MYO
5%	Pixio-L	79.44 ± 0.39	81.73 ± 0.58	70.14 ± 2.59	82.49 ± 0.71
	EchoFM	66.06 ± 0.10	75.21 ± 1.09	65.79 ± 0.26	74.83 ± 0.74
10%	Pixio-L	81.37 ± 0.28	83.48 ± 0.29	77.38 ± 0.14	83.94 ± 0.13
	EchoFM	71.18 ± 0.75	77.78 ± 0.53	70.87 ± 1.53	77.72 ± 0.31
25%	Pixio-L	83.61 ± 0.14	84.35 ± 0.33	80.70 ± 0.66	84.74 ± 0.18
	EchoFM	77.07 ± 0.11	80.49 ± 0.03	76.94 ± 0.34	80.81 ± 0.24
50%	Pixio-L	84.22 ± 0.30	85.03 ± 0.20	81.90 ± 0.12	85.15 ± 0.07
	EchoFM	78.81 ± 0.47	81.73 ± 0.26	78.85 ± 0.17	81.95 ± 0.16
100%	Pixio-L	84.67 ± 0.14	85.39 ± 0.14	83.88 ± 0.11	85.60 ± 0.03
	EchoFM	80.53 ± 0.29	82.69 ± 0.26	80.42 ± 0.08	82.90 ± 0.07

For patient i with F_i frames, we average frame metrics before weighting the $N = 50$ patients equally:

$$M_{i,q,r,s} = \frac{1}{F_i} \sum_{f=1}^{F_i} M_{i,f,q,r,s}, \qquad \overline{M}_{q,r,s} = \frac{1}{N} \sum_{i=1}^{N} M_{i,q,r,s}, \tag{1}$$

where q, r, and s index route, fraction, and seed; the system and MYO adaptation mode are held fixed. Reported scores are means and sample standard deviations across three fitted seeds, without prediction ensembling. The paired route difference is first averaged across seeds within each patient:

$$d_{i,r} = \frac{1}{3} \sum_{s} \left(M_{i,\mathrm{Transfer},r,s} - M_{i,\mathrm{Direct},r,s} \right), \qquad \Delta_r = \frac{1}{50} \sum_{i} d_{i,r}. \tag{2}$$

We resample the 50 paired patient records with replacement for 10,000 repetitions (bootstrap random-number seed 20260909) and report percentile 95% confidence intervals. Cross-system, cross-fraction, and frozen-versus-full-fine-tuning contrasts use the same procedure with the indicated settings. To compare route gains, each record is $d_{i,r}^{\mathrm{Echo}} - d_{i,r}^{\mathrm{Pixio}}$. Thus, neither frames nor fitted seeds are treated as independent patients. Intervals are conditional on the fitted models, exclude variability from cohort selection, and are not multiplicity-adjusted. Percentage-scale differences are percentage points (pp); an interval spanning zero establishes neither superiority nor equivalence.

We also report medians, quartiles, positive/negative patient counts, and the range of means after omitting each patient once. These descriptive omission checks hold models fixed: they are neither retraining experiments nor confidence intervals. Contrasts and confidence intervals use unrounded patient-level scores; rounding is applied only for presentation.

4.3 System performance and the value of LV supervision

Under full fine-tuning, Pixio-L achieves higher mean MYO Dice under both routes at all five fractions (Table 2). At 5%, its advantage is +13.38 pp for Direct (95% CI $[+11.86, +14.99]$) and +6.52 pp for LV-to-MYO ($[+4.91, +8.26]$). At 100%, these differences are +4.14 and +2.70 pp. All ten same-fraction cross-system intervals are above zero (Table 8).

Incremental route benefit under full fine-tuning. LV supervision improves mean Dice in both systems at every fraction. At 5%, Pixio-L improves from 79.44% to 81.73% (+2.29 pp), and EchoFM from 66.06% to 75.21% (+9.15 pp; CI $[+7.80, +10.47]$). All 30 paired system–fraction–seed gains are positive, with a minimum of +0.52 pp. All ten mean route-gain intervals are above zero (Tables 5 and 6). Gains are largest at 5% and remain positive at 100% (+0.72 pp for Pixio-L; +2.16 pp for EchoFM). EchoFM shows larger route gains, while Pixio-L retains higher absolute mean Dice. Appendix B reports the separate gain figure and all six-metric contrasts.

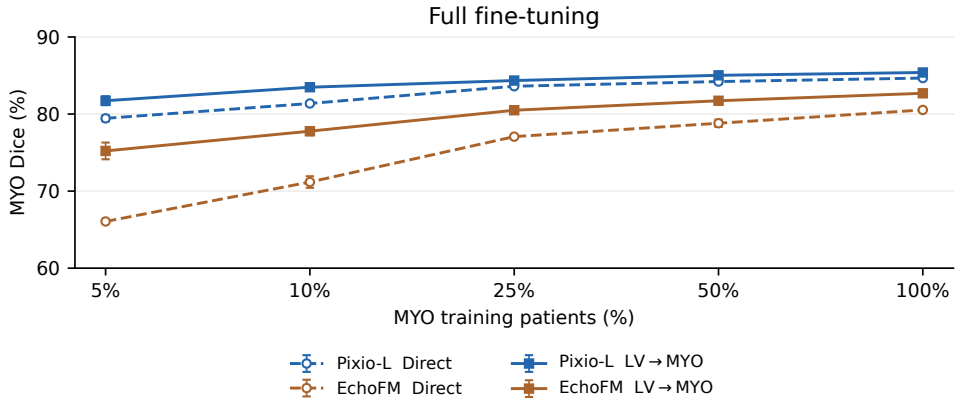


Figure 2: Full fine-tuning during MYO training. Points show mean Dice with sample SD across three seeds for each system and route. Fractions are equally spaced; lines guide the eye.

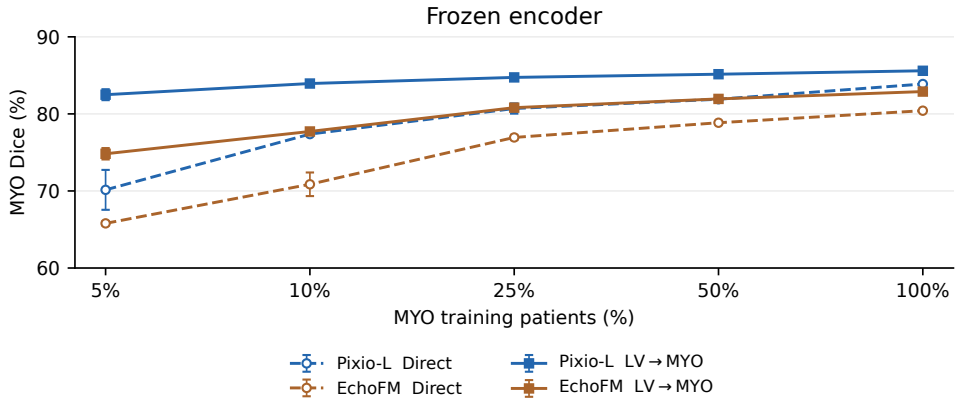


Figure 3: Frozen encoder during MYO training, with trainable segmentation modules. LV-to-MYO reuses the full-fine-tuned LV model; Direct uses released encoder weights. Axes, colors, route styles and three-seed summaries follow Figure 2.

4.4 Patient-level behavior and complementary errors

Under full fine-tuning, after averaging paired differences across seeds, Transfer improves Dice for 34–43 of 50 patients across Pixio-L fractions and 39–49 for EchoFM. Every setting also includes patient-level losses. All ten median gains and all omit-one-patient mean gains are positive; the smallest omission mean is +0.59 pp (Appendix E). These checks hold fitted models fixed and do not imply universal benefit.

Complementary metrics qualify these full-fine-tuning Dice findings. For Pixio-L at 25%, Transfer improves Dice by 0.74 pp and PPV by 2.05 pp, but sensitivity falls by 0.79 pp (95% CI $[-1.38, -0.18]$). All six metrics and their paired intervals are reported in Appendices A and B.

Figure 4 and Appendix F show fixed examples from the current full-fine-tuning checkpoints. They illustrate error structure, not population error prevalence or monotonic case-level improvement.

4.5 Effect of encoder freezing during MYO training

Figure 3 and Appendix G compare the 60 frozen-encoder runs with full fine-tuning. For Pixio-L Direct at 5%, freezing reduces Dice from 79.44% to 70.14% (-9.30 pp, 95% CI $[-10.51, -8.12]$). In contrast, after full-fine-tuned LV initialization, frozen Pixio-L reaches 82.49% versus 81.73% under full target fine-tuning (+0.76 pp, CI $[+0.04, +1.48]$). The other four Pixio-L Transfer intervals span zero. EchoFM differences are smaller (-0.39 to $+0.32$ pp across both routes); nine of ten intervals span zero, without establishing equivalence. The frozen mean-Dice ordering also favors Pixio-L in all ten matched fraction–route settings (Table 2).

Starting from the same LV checkpoint, Pixio-L achieved slightly higher mean MYO Dice with a frozen encoder than with full fine-tuning across all five target-training fractions. One possible explanation is that freezing preserves useful LV-adapted features while the segmentation modules learn to delineate the myocardium. These features may include ventricular localization and endocardial-boundary cues, although their role was not directly examined.

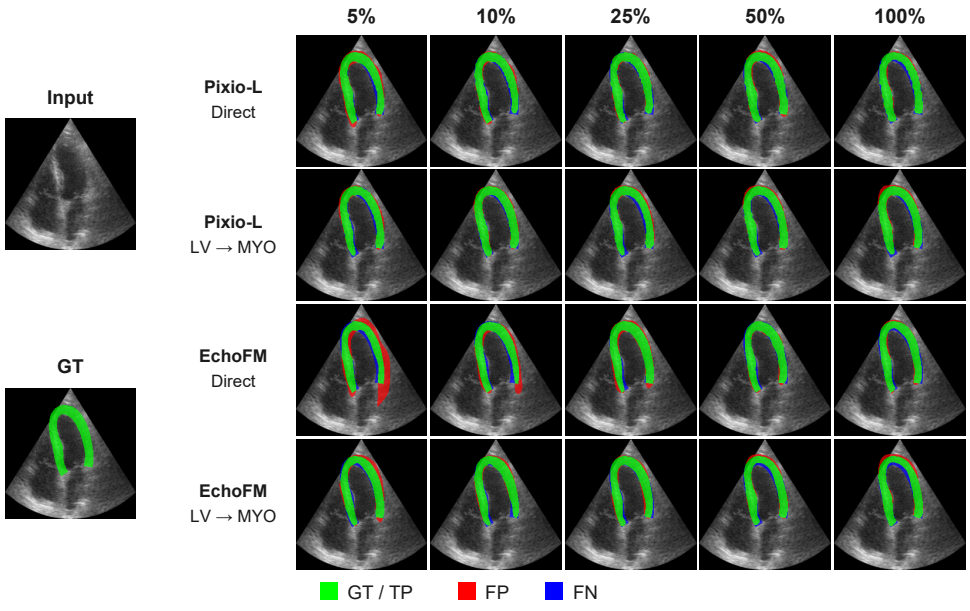


Figure 4: Matched MYO errors under full fine-tuning for patient0316, frame 9, seed 42. Both routes span all five fractions with shared Input/GT. GT/TP is green, FP red and FN blue (70% opacity).

4.6 Data availability and practical interpretation

Under full fine-tuning, Pixio-L LV-to-MYO at 10% reaches 83.48% mean Dice, versus 82.69% for EchoFM LV-to-MYO at 100%. Their post-hoc paired difference is +0.79 pp (CI $[-0.05, +1.67]$; Table 7). An interval containing zero proves neither equivalence nor non-inferiority. Within EchoFM, 50% LV-to-MYO achieves 81.73% versus 80.53% for 100% Direct (+1.19 pp; CI $[+0.41, +2.00]$). These post-hoc contrasts concern target-training cohort size, not total annotation or computational cost; source and validation labels are additional.

We report both adaptation modes without selecting the better mode per test setting. Recorded costs retain different timing scopes and do not support cross-system speed ratios (Appendix D).

5 Conclusion

Across five fractions and three seeds, full-fine-tuned Pixio-L with Mask2Former achieves higher mean MYO Dice under both routes. LV supervision improves both systems at every fraction; EchoFM's larger gains narrow the performance gap. Freezing the encoder substantially reduces Pixio-L Direct performance at 5%, while LV-initialized Pixio-L does not show the same loss. EchoFM shows smaller differences between adaptation modes. Patient-level losses and metric trade-offs qualify these cohort averages.

Limitations and Future Work. Evidence covers one CAMUS A4C split and one nested subset draw. Bootstrap intervals are conditional on fitted models and unadjusted for multiplicity. Irregular clip spacing, endpoint repetition, and the amended source stopping rule may affect EchoFM's results (Section 4.1). EchoFM reports pretraining on EchoNet-Dynamic and CAMUS (Kim et al., 2025), but released-checkpoint membership of our held-out patients is unverified. Pixio-L's web-image teacher documentation does not establish that ultrasound was excluded (FAIR at Meta and HKU, 2025). Neither record establishes test-patient overlap; pretraining provenance is distinct from this study's supervised train/test separation.

The learning-rate and weight-decay revisions and frozen ablation followed availability of the original test results and are not independently confirmatory. Prespecified evaluation on an independent cohort and repeated subset draws would test robustness. Encoder freezing during MYO training is evaluated across both systems and routes; head reinitialization after LV training and freezing during the LV stage remain untested. The frozen ablation is not a complete reproduction of EchoFM's published adaptation protocol. Budget-matched controls and comparable timing scopes are needed for claims about the mechanism of LV-stage gains and relative training costs.

AI Use Statement

Generative AI tools assisted with experimental programming and debugging, study framing, literature search and synthesis, statistical-analysis code and result checking, visualization, manuscript drafting and revision, LaTeX formatting, and simulated manuscript reviews. The authors retain responsibility for the scientific claims, methods, references, and reported results.

Ethics Statement

This study used publicly available, de-identified echocardiographic data and reference annotations obtained directly from the official EchoNet-Dynamic and CAMUS releases (Ouyang et al., 2020; Leclerc et al., 2019). This secondary analysis involved no new participant recruitment, additional clinical data collection, or intervention, and no new annotations were created. Ethics and consent procedures for the original data collections are described in the respective source publications. The reported outputs are research measurements; clinical validation and suitability for deployment have not been established.

Reproducibility Statement

Sections 3 and 4 describe the training and evaluation protocols. The appendices report complete quantitative results, implementation details and matched qualitative examples. The source code for training and evaluation will be made publicly available upon acceptance.

REFERENCES

Ahmed M. Alaa, Anthony Philippakis, and David Sontag. ETAB: A benchmark suite for visual representation learning in echocardiography. In *Advances in Neural Information Processing Systems*, volume 35, pp. 19075–19086, 2022. doi: 10.52202/068431-1386. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/796501434d0dc3a039d5b91261f7f889-Abstract-Datasets_and_Benchmarks.html.

Chieh-Ju Chao, Yunqi Richard Gu, Wasan Kumar, Tiange Xiang, Lalith Appari, Justin Wu, Juan M. Farina, Rachael Wraith, Jiwoon Jeong, Reza Arsanjani, Garvan C. Kane, Jae K. Oh, Curtis P. Langlotz, Imon Banerjee, Li Fei-Fei, and Ehsan Adeli. Foundation versus domain-specific models for left ventricular segmentation on cardiac ultrasound. *npj Digital Medicine*, 8(1):341, 2025. doi: 10.1038/s41746-025-01730-y. URL https://www.nature.com/articles/s41746-025-01730-y.

Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1290–1299, 2022. URL https://arxiv.org/abs/2112.01527.

Matthew Christensen, Milos Vukadinovic, Neal Yuan, and David Ouyang. Vision–language foundation model for echocardiogram interpretation. *Nature Medicine*, 30(5):1481–1488, 2024. doi: 10.1038/s41591-024-02959-y. URL https://www.nature.com/articles/s41591-024-02959-y.

Lee R. Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3): 297–302, 1945. doi: 10.2307/1932409. URL https://doi.org/10.2307/1932409.

FAIR at Meta and HKU. Pixio ViT-L/16 model card. Hugging Face, 2025. URL https://huggingface.co/facebook/pixio-vitl16.

Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R. Roth, and Daguang Xu. UNETR: Transformers for 3D medical image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 574–584, 2022. URL https://arxiv.org/abs/2103.10504.

Kaiming He, Ross Girshick, and Piotr Dollár. Rethinking ImageNet pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4918–4927, 2019. URL https://openaccess.thecvf.com/content_ICCV_2019/html/He_Rethinking_ImageNet_Pre-Training_ICCV_2019_paper.html.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022. URL https://arxiv.org/abs/2111.06377.

Paul Jaccard. Étude comparative de la distribution florale dans une portion des alpes et du jura. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 37(142):547–579, 1901. doi: 10.5169/seals-266450. URL https://doi.org/10.5169/seals-266450.

Sekeun Kim, Pengfei Jin, Sifan Song, Cheng Chen, Yiwei Li, Hui Ren, Xiang Li, Tianming Liu, and Quanzheng Li. EchoFM: Foundation model for generalizable echocardiogram analysis. *IEEE Transactions on Medical Imaging*, 44(10):4049–4062, 2025. doi: 10.1109/TMI.2025.3580713. URL https://arxiv.org/abs/2410.23413.

Roberto M. Lang, Luigi P. Badano, Victor Mor-Avi, Jonathan Afilalo, Anderson Armstrong, Laura Ernande, Frank A. Flachskampf, Elyse Foster, Steven A. Goldstein, Tatiana Kuznetsova, Patrizio Lancellotti, Denisa Muraru, Michael H. Picard, Ernst R. Rietzschel, Lawrence Rudski, Kirk T. Spencer, Wendy Tsang, and Jens-Uwe Voigt. Recommendations for cardiac chamber quantification by echocardiography in adults: An update from the american society of echocardiography and the european association of cardiovascular imaging. *Journal of the American Society of Echocardiography*, 28(1):1–39.e14, 2015. doi: 10.1016/j.echo.2014.10.003. URL https://doi.org/10.1016/j.echo.2014.10.003.

Sarah Leclerc, Erik Smistad, João Pedrosa, Andreas Østvik, Frederic Cervenansky, Florian Espinosa, Torvald Espeland, Erik Andreas Rye Berg, Pierre-Marc Jodoin, Thomas Grenier, Carole Lartizien, Jan D'hooge, Lasse Lovstakken, and Olivier Bernard. Deep learning for segmentation using an open large-scale dataset in 2D echocardiography. *IEEE Transactions on Medical Imaging*, 38(9):2198–2210, 2019. doi: 10.1109/TMI.2019.2900516. URL https://arxiv.org/abs/1908.06948.

Wenxuan Li, Alan Yuille, and Zongwei Zhou. How well do supervised 3D models transfer to medical imaging tasks? In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/47360925c45a166c96f652589265dee9-Abstract-Conference.html.

Xian Lin, Yangyang Xiang, Li Yu, and Zengqiang Yan. Beyond adapting SAM: Towards end-to-end ultrasound image segmentation via auto prompting. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume 15008 of *Lecture Notes in Computer Science*, pp. 24–34. Springer Nature Switzerland, 2024. doi: 10.1007/978-3-031-72111-3_3. URL https://doi.org/10.1007/978-3-031-72111-3_3.

Lena Maier-Hein, Annika Reinke, Patrick Godau, Minu D. Tizabi, Florian Buettner, Evangelia Christodoulou, Ben Glocker, Fabian Isensee, Jens Kleesiek, Michal Kozubek, Mauricio Reyes, Michael A. Riegler, Manuel Wiesenfarth, et al. Metrics reloaded: Recommendations for image analysis validation. *Nature Methods*, 21(2):195–212, 2024. doi: 10.1038/s41592-023-02151-z. URL https://arxiv.org/abs/2206.01653.

David Ouyang, Bryan He, Amirata Ghorbani, Neal Yuan, Joseph Ebinger, Curtis P. Langlotz, Paul A. Heidenreich, Robert A. Harrington, David H. Liang, Euan A. Ashley, and James Y. Zou. Video-based AI for beat-to-beat assessment of cardiac function. *Nature*, 580(7802):252–256, 2020. doi: 10.1038/s41586-020-2145-8. URL https://echonet.github.io/dynamic/.

Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. Transfusion: Understanding transfer learning for medical imaging. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/hash/eb1e78328c46506b46a4ac4a1e378b91-Abstract.html.

Mohamed Saeed, Rand Muhtaseb, and Mohammad Yaqub. Contrastive pretraining for echocardiography segmentation with limited data. In *Medical Image Understanding and Analysis*, volume 13413 of *Lecture Notes in Computer Science*, pp. 680–691. Springer, 2022. doi: 10.1007/978-3-031-12053-4_50. URL https://arxiv.org/abs/2201.07219.

Ruiqiang Xiao, Zhaohu Xing, Yijun Yang, Zhenyan Han, Weiming Wang, Kaishun Wu, and Lei Zhu. EchoPilot: Training-free ultrasound video segmentation via scale-space semantic prompting and reliability-gated memory. *arXiv preprint arXiv:2605.25944*, 2026. URL https://arxiv.org/abs/2605.25944v1.

Lihe Yang, Shang-Wen Li, Yang Li, Xinjie Lei, Dong Wang, Abdelrahman Mohamed, Saining Xie, Hengshuang Zhao, Kaiming He, and Hu Xu. In pursuit of pixel supervision for visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 31974–31984, 2026. URL https://openaccess.thecvf.com/content/CVPR2026/html/Yang_In_Pursuit_of_Pixel_Supervision_for_Visual_Pre-training_CVPR_2026_paper.html.

Xiaoyan Zhang, Alvaro E. Ulloa Cerna, Joshua V. Stough, Yida Chen, Brendan J. Carry, Amro Alsaid, Sushravya Raghunath, David P. vanMaanen, Brandon K. Fornwalt, and Christopher M. Haggerty. Generalizability and quality control of deep learning-based 2D echocardiography segmentation models in a large clinical dataset. *The International Journal of Cardiovascular Imaging*, 38(8):1685–1697, 2022. doi: 10.1007/s10554-022-02554-7. URL https://link.springer.com/article/10.1007/s10554-022-02554-7.

Nan Zhou, Huiqun Wang, Yaoyan Zheng, and Di Huang. Progressive parameter efficient transfer learning for semantic segmentation. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/86c070ce724102ee876d1935590e111a-Abstract-Conference.html.

APPENDIX

A COMPLETE TARGET RESULTS

This appendix reports six metrics for the 60 primary full-fine-tuning target runs. Corresponding results for the 60 frozen-encoder runs appear in Appendix G (Tables 15 and 16). Frame metrics are averaged within each patient, followed by equal weighting of the 50 patients. Means and sample SDs summarize three training seeds, not 150 independent patients. Dice, IoU, sensitivity, and PPV are percentages (higher is better); HD95 and ASSD are pixels on the 224×224 grid (lower is better), not millimeters.

For nonempty foreground sets P and G, $\mathrm{Dice} = 2|P \cap G|/(|P| + |G|)$ and $\mathrm{IoU} = |P \cap G|/|P \cup G|$. Sensitivity is $\mathrm{TP}/(\mathrm{TP} + \mathrm{FN})$ and PPV is $\mathrm{TP}/(\mathrm{TP} + \mathrm{FP})$; these describe segmentation, not diagnosis. Mean IoU cannot be obtained by applying $\mathrm{Dice}/(2 - \mathrm{Dice})$ to mean Dice. Surface-distance conventions are given in Appendix D.3.

Table 3: Pixio-L: complete six-metric results, mean ± sample SD over three seeds. Dice, IoU, sensitivity (Sens.) and PPV are percentages (higher is better). HD95 and ASSD are pixels on the 224×224 grid (lower is better). Both systems use the common maximum-directed HD95 definition (Appendix D.3).

Fraction	Route	Dice	IoU	HD95	ASSD	Sens.	PPV
5%	Direct	79.44 ± 0.39	66.51 ± 0.48	9.55 ± 0.97	3.18 ± 0.11	83.67 ± 1.69	76.59 ± 0.78
	Transfer	81.73 ± 0.58	69.49 ± 0.79	7.35 ± 0.34	2.74 ± 0.13	86.23 ± 1.83	78.51 ± 2.16
10%	Direct	81.37 ± 0.28	69.12 ± 0.40	8.08 ± 0.14	2.81 ± 0.04	84.26 ± 0.49	79.53 ± 0.65
	Transfer	83.48 ± 0.29	72.02 ± 0.41	6.31 ± 0.12	2.44 ± 0.07	87.39 ± 1.17	80.65 ± 1.48
25%	Direct	83.61 ± 0.14	72.31 ± 0.21	6.51 ± 0.41	2.44 ± 0.08	87.59 ± 1.23	80.58 ± 1.24
	Transfer	84.35 ± 0.33	73.34 ± 0.47	5.81 ± 0.19	2.26 ± 0.07	86.80 ± 0.55	82.63 ± 1.03
50%	Direct	84.22 ± 0.30	73.16 ± 0.41	6.20 ± 0.21	2.32 ± 0.03	87.45 ± 1.63	81.70 ± 0.88
	Transfer	85.03 ± 0.20	74.33 ± 0.28	5.61 ± 0.12	2.19 ± 0.04	88.64 ± 1.15	82.17 ± 0.94
100%	Direct	84.67 ± 0.14	73.79 ± 0.17	5.67 ± 0.11	2.21 ± 0.02	87.03 ± 0.40	82.87 ± 0.36
	Transfer	85.39 ± 0.14	74.88 ± 0.21	5.35 ± 0.12	2.11 ± 0.04	87.47 ± 0.54	83.83 ± 0.74

Table 4: EchoFM: encoder peak LR 10^{-5}; complete six-metric results, mean ± sample SD over three seeds. Dice, IoU, sensitivity (Sens.) and PPV are percentages (higher is better). HD95 and ASSD are pixels on the 224×224 grid (lower is better). Both systems use the common maximum-directed HD95 definition (Appendix D.3).

Fraction	Route	Dice	IoU	HD95	ASSD	Sens.	PPV
5%	Direct	66.06 ± 0.10	50.09 ± 0.17	23.13 ± 0.79	5.92 ± 0.24	75.90 ± 4.28	60.55 ± 2.67
	Transfer	75.21 ± 1.09	61.01 ± 1.45	19.31 ± 3.29	4.34 ± 0.44	78.65 ± 1.38	73.55 ± 2.28
10%	Direct	71.18 ± 0.75	56.03 ± 0.90	23.13 ± 3.95	5.25 ± 0.55	78.50 ± 3.13	66.49 ± 2.82
	Transfer	77.78 ± 0.53	64.31 ± 0.76	16.19 ± 1.55	3.80 ± 0.16	80.33 ± 1.03	76.60 ± 0.94
25%	Direct	77.07 ± 0.11	63.32 ± 0.16	14.97 ± 0.84	3.83 ± 0.13	81.77 ± 2.31	73.96 ± 2.01
	Transfer	80.49 ± 0.03	67.83 ± 0.03	11.56 ± 1.59	3.10 ± 0.24	84.33 ± 3.69	77.93 ± 3.07
50%	Direct	78.81 ± 0.47	65.60 ± 0.57	12.89 ± 0.73	3.38 ± 0.06	81.18 ± 2.07	77.58 ± 1.16
	Transfer	81.73 ± 0.26	69.54 ± 0.41	10.51 ± 0.86	2.85 ± 0.03	84.23 ± 0.58	80.09 ± 1.05
100%	Direct	80.53 ± 0.29	67.93 ± 0.43	11.76 ± 0.89	3.07 ± 0.10	83.03 ± 0.18	78.97 ± 0.66
	Transfer	82.69 ± 0.26	70.92 ± 0.32	9.08 ± 0.88	2.64 ± 0.08	84.45 ± 0.27	81.71 ± 0.57

B PAIRED TARGET CONTRASTS

Under full fine-tuning, matched contrasts subtract Direct from Transfer at the same target fraction. Positive overlap differences (percentage points) and negative distance differences (pixels) favor Transfer. For each patient, we average the three seed-specific paired differences, then resample the 50 patients with replacement 10,000 times to obtain percentile 95% intervals. These intervals are conditional on the fitted models, unadjusted for multiplicity, and do not establish equivalence when they contain zero.

Table 5: Pixio-L: paired Transfer minus Direct contrasts. Units are percentage points for overlap, sensitivity and PPV; pixels for HD95 and ASSD. CIs are unadjusted 95% paired-patient bootstrap intervals; negative boundary-distance differences are favorable. Both systems use common HD95 (Appendix D.3).

Fraction	Metric	Difference	95% CI
5%	Dice	+2.285	[+1.401, +3.325]
	IoU	+2.986	[+1.897, +4.197]
	HD95	-2.201	[-3.214, -1.328]
	ASSD	-0.444	[-0.703, -0.240]
	Sens.	+2.558	[+1.540, +3.731]
	PPV	+1.923	[+0.920, +3.062]
10%	Dice	+2.111	[+1.384, +2.964]
	IoU	+2.901	[+1.977, +3.935]
	HD95	-1.770	[-2.726, -0.932]
	ASSD	-0.374	[-0.600, -0.202]
	Sens.	+3.128	[+2.128, +4.280]
	PPV	+1.125	[+0.298, +2.070]
25%	Dice	+0.740	[+0.297, +1.221]
	IoU	+1.026	[+0.442, +1.629]
	HD95	-0.701	[-1.149, -0.348]
	ASSD	-0.181	[-0.293, -0.097]
	Sens.	-0.790	[-1.378, -0.177]
	PPV	+2.051	[+1.506, +2.585]
50%	Dice	+0.811	[+0.399, +1.214]
	IoU	+1.174	[+0.585, +1.740]
	HD95	-0.597	[-0.992, -0.270]
	ASSD	-0.129	[-0.191, -0.066]
	Sens.	+1.191	[+0.689, +1.713]
	PPV	+0.477	[-0.002, +0.929]
100%	Dice	+0.722	[+0.385, +1.056]
	IoU	+1.099	[+0.600, +1.600]
	HD95	-0.319	[-0.487, -0.157]
	ASSD	-0.108	[-0.155, -0.063]
	Sens.	+0.441	[-0.096, +0.969]
	PPV	+0.959	[+0.453, +1.485]

Table 6: EchoFM: paired Transfer minus Direct contrasts. Units are percentage points for overlap, sensitivity and PPV; pixels for HD95 and ASSD. CIs are unadjusted 95% paired-patient bootstrap intervals; negative boundary-distance differences are favorable. Both systems use common HD95 (Appendix D.3).

Fraction	Metric	Difference	95% CI
5%	Dice	+9.153	[+7.805, +10.468]
	IoU	+10.920	[+9.312, +12.494]
	HD95	-3.823	[-6.609, -1.006]
	ASSD	-1.572	[-1.910, -1.240]
	Sens.	+2.754	[+1.364, +4.139]
	PPV	+12.993	[+11.346, +14.640]
10%	Dice	+6.597	[+5.261, +7.934]
	IoU	+8.272	[+6.663, +9.849]
	HD95	-6.943	[-9.499, -4.489]
	ASSD	-1.444	[-1.768, -1.157]
	Sens.	+1.832	[+0.300, +3.323]
	PPV	+10.112	[+8.909, +11.305]
25%	Dice	+3.426	[+2.537, +4.380]
	IoU	+4.504	[+3.402, +5.674]
	HD95	-3.418	[-5.483, -1.534]
	ASSD	-0.725	[-1.036, -0.467]
	Sens.	+2.560	[+1.505, +3.612]
	PPV	+3.971	[+2.775, +5.239]
50%	Dice	+2.916	[+2.101, +3.770]
	IoU	+3.943	[+2.841, +5.089]
	HD95	-2.385	[-4.117, -0.667]
	ASSD	-0.532	[-0.738, -0.345]
	Sens.	+3.050	[+2.087, +4.036]
	PPV	+2.509	[+1.379, +3.658]
100%	Dice	+2.161	[+1.450, +2.866]
	IoU	+2.991	[+2.015, +3.952]
	HD95	-2.682	[-4.689, -0.712]
	ASSD	-0.431	[-0.625, -0.235]
	Sens.	+1.418	[+0.410, +2.418]
	PPV	+2.740	[+1.745, +3.707]

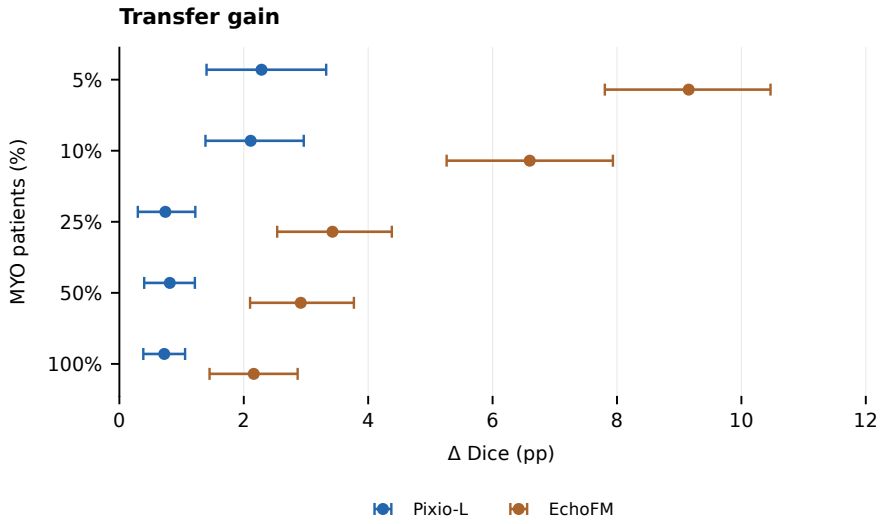


Figure 5: Gain from intermediate LV supervision. Points show LV-to-MYO minus matched Direct Dice in percentage points; error bars are unadjusted 95% paired-patient bootstrap confidence intervals, conditional on the fitted models. Paired differences for the same 50 evaluation patients are averaged across three seeds before 10,000 bootstrap resamples. The legend distinguishes systems; every point is already a within-system route contrast.

Post-hoc cross-fraction comparisons. Table 7 uses the same paired-patient procedure, subtracting the indicated settings within each seed before averaging across seeds. The EchoFM 50% Transfer versus 100% Direct interval lies above zero; the other two contain zero. These post-hoc intervals are conditional on fitted models and unadjusted for multiplicity; they do not establish equivalence or non-inferiority.

Table 7: Post-hoc cross-fraction Dice contrasts (percentage points). The sign is left setting minus right setting. Unadjusted 95% paired-patient bootstrap intervals are conditional on the fitted models. An interval containing zero does not establish equivalence or non-inferiority. Target fractions do not quantify total annotation cost.

Comparison	Δ	95% CI
Pixio-L LV-to-MYO 10% − EchoFM LV-to-MYO 100%	+0.79	[−0.05, 1.67]
Pixio-L LV-to-MYO 50% − Pixio-L Direct 100%	+0.36	[−0.08, 0.78]
EchoFM LV-to-MYO 50% − EchoFM Direct 100%	+1.19	[0.41, 2.00]

Table 8: Post-hoc full-fine-tuning system contrasts: Pixio-L minus EchoFM Dice, in percentage points. Each route compares the same 50 patients after averaging three seeds. Intervals are unadjusted 95% paired-patient bootstrap intervals, conditional on the fitted systems and cohorts; differences do not isolate pretraining domain.

	Direct		LV-to-MYO	
Target fraction	Δ	95% CI	Δ	95% CI
5%	+13.38	[11.86, 14.99]	+6.52	[4.91, 8.26]
10%	+10.19	[8.80, 11.69]	+5.71	[4.31, 7.24]
25%	+6.54	[5.19, 7.95]	+3.85	[2.97, 4.68]
50%	+5.41	[4.28, 6.54]	+3.30	[2.41, 4.21]
100%	+4.14	[2.99, 5.34]	+2.70	[1.83, 3.54]

C SOURCE LV RESULTS

Table 9 reports LV performance before MYO adaptation on 1,276 EchoNet-Dynamic test patients and 2,552 ED/ES frames. Metrics are averaged within patients and then across patients; mean and sample SD summarize three training seeds. Source LV HD95 uses the same definition for both systems. Source and target scores describe different tasks and cohorts and are not directly subtracted.

Table 9: EchoNet-Dynamic LV source test results, before MYO adaptation: 1,276 patients and 2,552 ED/ES frames. Scores average frames within each patient, then weight patients equally. Summary rows report mean ± sample SD across three training seeds, not patient variability. Both systems use the maximum of the two directed 95th percentiles for HD95 on the 224×224 grid.

System	Seed	LV Dice (%)	HD95 (px224)
Pixio-L + Mask2Former	42	92.25	5.77
	777	92.32	5.77
	2026	92.21	5.81
	Mean ± SD	$\mathbf{92.26 \pm 0.05}$	$\mathbf{5.78 \pm 0.02}$
EchoFM + UNETR	42	91.09	7.44
	777	91.22	7.40
	2026	91.16	7.54
	Mean ± SD	$\mathbf{91.16 \pm 0.06}$	$\mathbf{7.46 \pm 0.08}$

D IMPLEMENTATION DETAILS

D.1 OPTIMIZATION

Parameter counts. In the primary full-fine-tuning comparison all parameters are trainable; buffers are excluded. Pixio-L comprises the ViT encoder (303,377,408), feature pyramid (7,805,440), and segmentation head (12,872,707), totaling 324,055,555. EchoFM comprises the encoder (305,669,120) and decoder/output head (12,526,018), totaling 318,195,138.

Pixio-L uses AdamW with $(\beta_1, \beta_2) = (0.9, 0.999)$, $\epsilon = 10^{-8}$, learning rate 10^{-4}, backbone multiplier 0.1, and weight decay 10^{-4}; one-dimensional, class-token, positional, and embedding parameters have zero decay. The feature pyramid shares the backbone learning rate. Mixed precision is enabled, without gradient clipping. `WarmupMultiStepLR` uses a linear warmup factor of 0.001 for 10 nominal epochs, followed by tenfold reductions at 70% and 90% of the update budget. Mask2Former classification, mask, and Dice loss weights are 0.5, 25, and 10; no-object weight is 0.1 and decoder deep supervision is enabled. Point sampling uses 12,544 points, oversampling ratio 3.0, and importance-sampling ratio 0.75.

Primary EchoFM targets use AdamW with $(\beta_1, \beta_2) = (0.9, 0.999)$, $\epsilon = 10^{-8}$, and target base weight decay 10^{-4}, with zero decay for one-dimensional parameters and designated token/embedding parameters. The reused LV source retains weight decay 0.05. Primary target runs use encoder/decoder peak learning rates $10^{-5}/10^{-4}$; the reused LV source models use the same unscaled encoder and decoder schedule, with a peak learning rate of 10^{-4}. For each pixel j of a supervised anchor frame of height H and width W, let $z_j \in \mathbb{R}^2$ denote the background/foreground logits, $p_j = [\mathrm{softmax}(z_j)]_1$ the foreground probability, and $g_j \in \{0, 1\}$ the reference label (0: background; 1: foreground). Foreground denotes LV cavity in source training and MYO in target training. The per-anchor float32 loss is

$$\mathcal{L} = \frac{1}{HW} \sum_j \mathrm{CE}(z_j, g_j) + 1 - \frac{2 \sum_j p_j g_j + 1}{\sum_j p_j + \sum_j g_j + 1}. \tag{3}$$

All sums run over the HW pixels. Only slot 16 of each 32-frame clip is supervised. Microbatch loss sums are divided by effective batch size 256 before accumulation; gradient-norm clipping at 1.0 precedes each update. Forward computation uses bfloat16 autocast.

For n training anchors, the cap is $U = \lceil 100n/256 \rceil$ updates, validation period is $\lceil n/256 \rceil$, and warmup is $U_{\text{warm}} = \lceil 10n/256 \rceil$. At zero-based update $u < U_{\text{warm}}$, the learning rate is $10^{-4}(u + 1)/U_{\text{warm}}$; it then decays cosinusoidally to 10^{-6} over the remaining nominal updates. This defines the decoder schedule. For the primary target encoder, the entire schedule is multiplied by 0.1, including warmup and the endpoint; source training retains the unscaled schedule.

Pixio-L experiments used Python 3.12.13, PyTorch 2.7.1 (CUDA 12.6), torchvision 0.22.1, Detectron2 0.6, NumPy 2.4.4, SciPy 1.17.1, and timm 1.0.24.

Hardware and execution. All 120 MYO experiments used one NVIDIA H200 GPU per run, with PyTorch 2.7.1 and CUDA 12.6. Concurrent allocations distributed independent model, route, fraction and seed runs across GPUs; individual runs did not use distributed data parallelism. Pixio-L used automatic mixed precision; EchoFM used bfloat16 forward autocast and float32 loss computation. Per-device memory capacity and EchoFM's exact Python patch version were not recorded in the archived run metadata.

D.2 Continuation and transfer

EchoFM source training continued after early-stopping warmup/patience changed from 60/20 to 20/10 validation evaluations. Prior validations were replayed under the amended stopping rule while preserving model and optimizer state, sampler position, update count, and learning-rate progress. In both systems, LV-to-MYO transfer instead loads the complete same-seed source validation-best model, starts target optimization afresh, and imports no source optimizer state. Each source checkpoint is reused across all five target fractions, and all model parameters remain trainable in the primary comparison; the ablation freezes only the target encoder.

D.3 Evaluation conventions

Both systems score binary foreground on the 224×224 grid with unit pixel spacing; ties favor background. Surfaces are four-connected inner boundaries with outside-image background. Let $D_{P \to G}$ and $D_{G \to P}$ denote directed nearest-boundary Euclidean distances. ASSD averages their concatenation, weighting each boundary point equally. The same implementation computes target MYO HD95 for both systems:

$$\text{HD95} = \max\{Q_{.95}(D_{P \to G}), Q_{.95}(D_{G \to P})\}, \tag{4}$$

where $Q_{.95}$ is the linearly interpolated 95th percentile. The primary 60 checkpoints are evaluated on the same 976 frames from 50 patients. Pixio-L scores come from the archived unified re-evaluation; new EchoFM scores use that common implementation with the revised validation-selected checkpoints. Source LV evaluation uses the same maximum-directed definition and the original reused source checkpoints.

When both masks are empty, overlap scores are one and distances zero. When exactly one is empty, overlap scores are zero and the common distance penalty is $\sqrt{(H-1)^2 + (W-1)^2}$. No empty prediction or reference occurs in the target evaluations. Frames are averaged within each patient before patients are weighted equally; changing the HD95 definition does not change this aggregation.

D.4 Running time

Table 10 distinguishes full fine-tuning and frozen-encoder MYO adaptation. Pixio-L records the completed process, including setup, training, validation, testing and output; frozen runs additionally include encoder-immutability audits. EchoFM sums committed training and validation logs, excluding setup, checkpoint I/O, test, queueing and discarded work. These are observed costs at each run's stopping point, not fixed-update speed comparisons or a complete accounting of interrupted allocations. Source LV training is reused in both modes and costs 3.79 ± 0.03 hours for Pixio-L and 10.99 ± 0.24 for EchoFM under their respective timing scopes; this cost is additional to each target-only Transfer row and is incurred once per source seed.

Table 10: Recorded source LV and target MYO time per seed (hours; mean ± sample SD). Both modes cover all five fractions and three seeds. Pixio-L: process timer; EchoFM: committed training and validation. Different scopes and stopping points preclude treating these as standardized speed benchmarks.

		Pixio-L		EchoFM	
Source LV (full fine-tuning)		3.79 ± 0.03		10.99 ± 0.24	
Fraction	Route	Full FT	Frozen	Full FT	Frozen
5%	Direct	0.80 ± 0.01	0.71 ± 0.00	1.56 ± 0.28	1.60 ± 0.06
	LV→MYO	0.81 ± 0.01	0.71 ± 0.00	1.26 ± 0.17	0.92 ± 0.13
10%	Direct	1.12 ± 0.02	0.99 ± 0.03	1.90 ± 0.42	1.49 ± 0.26
	LV→MYO	1.08 ± 0.01	0.99 ± 0.06	1.38 ± 0.11	0.97 ± 0.15
25%	Direct	1.43 ± 0.01	1.36 ± 0.05	2.59 ± 0.56	2.10 ± 0.12
	LV→MYO	1.42 ± 0.02	1.27 ± 0.00	1.66 ± 0.30	1.22 ± 0.02
50%	Direct	2.08 ± 0.02	1.92 ± 0.13	4.03 ± 1.00	3.02 ± 0.12
	LV→MYO	2.11 ± 0.03	1.84 ± 0.01	2.68 ± 0.36	1.88 ± 0.07
100%	Direct	3.45 ± 0.01	3.41 ± 0.02	5.04 ± 1.29	4.24 ± 0.49
	LV→MYO	3.48 ± 0.03	3.10 ± 0.03	4.13 ± 0.05	3.26 ± 0.00

E Patient Heterogeneity and Sensitivity

Under full fine-tuning, the archived scores pair the same 50 patients across systems, routes and fractions after averaging three fitted seeds. Their means reproduce the full-fine-tuning columns of Table 2; patient IDs also agree with the fixed evaluation manifest. Table 11 summarizes the paired Transfer–Direct Dice differences with linearly interpolated quartiles and strict positive/negative counts, without a clinical-importance threshold.

For each patient i, the omission mean is $(50\Delta_r - d_{i,r})/49$. The reported range holds all fitted models fixed: it is neither a confidence interval nor leave-one-patient-out training or cross-validation.

Table 11: Post-hoc Transfer minus Direct Dice (pp) for 50 patients, averaged over three fitted seeds. Brackets give patient quartiles; $+/0/-$ counts directions of differences. Omit-one ranges keep models fixed and are not confidence intervals.

Fraction	System	Mean	Median [Q1, Q3]	$+/0/-$	Omit-one mean range
5%	Pixio-L	+2.29	+1.41 [0.58, 3.10]	43 / 0 / 7	[1.95, 2.40]
	EchoFM	+9.15	+9.09 [5.52, 13.27]	49 / 0 / 1	[8.97, 9.38]
10%	Pixio-L	+2.11	+1.44 [0.37, 3.00]	41 / 0 / 9	[1.82, 2.17]
	EchoFM	+6.60	+6.57 [4.62, 9.48]	45 / 0 / 5	[6.32, 6.84]
25%	Pixio-L	+0.74	+0.62 [-0.24, 1.36]	34 / 0 / 16	[0.59, 0.83]
	EchoFM	+3.43	+2.58 [1.00, 5.23]	44 / 0 / 6	[3.18, 3.53]
50%	Pixio-L	+0.81	+0.99 [0.05, 1.53]	38 / 0 / 12	[0.73, 0.90]
	EchoFM	+2.92	+2.67 [0.89, 4.08]	42 / 0 / 8	[2.75, 3.03]
100%	Pixio-L	+0.72	+0.74 [0.11, 1.34]	38 / 0 / 12	[0.65, 0.80]
	EchoFM	+2.16	+2.07 [0.14, 3.68]	39 / 0 / 11	[2.04, 2.29]

All medians and omission means are positive, but all ten settings include patient-level losses; EchoFM at 5% has 49 positive and one negative seed-averaged difference. Pixio-L at 25%, for example, improves Dice for 34 patients and reduces it for 16. No single patient is necessary for the positive cohort mean, but broader subgroup failures remain possible. Seed-averaged records cannot establish individual benefit under every seed; these checks do not assess new training subsets or institutions.

E.1 Difference between route gains

We subtract the Pixio-L route gain from the EchoFM gain within each patient, retaining all four scores in each paired record. This equals the reduction in the Pixio-L–EchoFM gap from Direct to Transfer. Table 12 uses 10,000 paired-patient bootstrap resamples (bootstrap random-number seed 20260909). It describes the fitted pipelines, not a causal intervention isolating architecture, pretraining, or LV semantics.

Table 12: Difference in route gain: EchoFM (Transfer − Direct) minus Pixio-L (Transfer − Direct), in Dice pp. Post-hoc 95% paired-patient bootstrap intervals are unadjusted and conditional on the fitted systems and fixed split.

Target fraction	Difference in transfer gain	95% CI
5%	+6.87	[5.44, 8.31]
10%	+4.49	[3.01, 5.91]
25%	+2.69	[1.71, 3.74]
50%	+2.11	[1.15, 3.05]
100%	+1.44	[0.59, 2.29]

F Matched Prediction Errors Across Target-Training Fractions

Two fixed gallery patients, patient0316 and patient0404, are shown under all three training seeds (42, 777 and 2026), one seed per page. Every board uses the five-fraction, four-configuration layout and error colors of Figure 4. Inputs, references and training subsets remain fixed across seeds. Patients and lower-middle test frames were selected independently of model scores; the frames are not designated ED/ES phases. Seeds are repetitions on one split, not folds or an ensemble. Reference masks are the official provider-supplied MYO annotations described in Section 4.1.

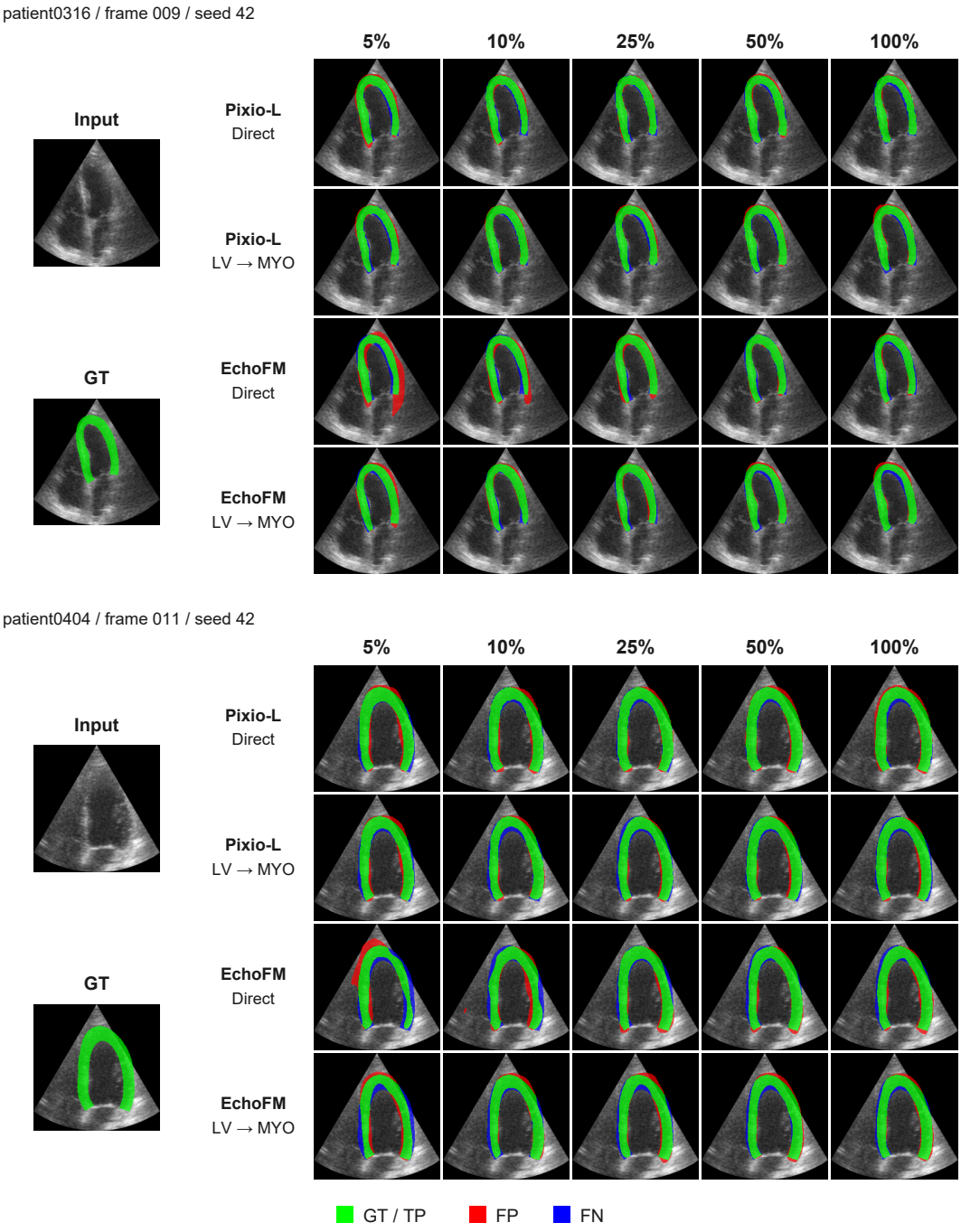


Figure 6: MYO errors at seed 42: patient0316, frame 9 (top, repeated from Figure 4 to complete the seed panels) and patient0404, frame 11 (bottom). All five fractions and four system–route configurations follow Figure 4. GT/TP is green, FP red and FN blue, overlaid on the input at 70% opacity. These fixed examples illustrate errors, not a representative patient distribution.

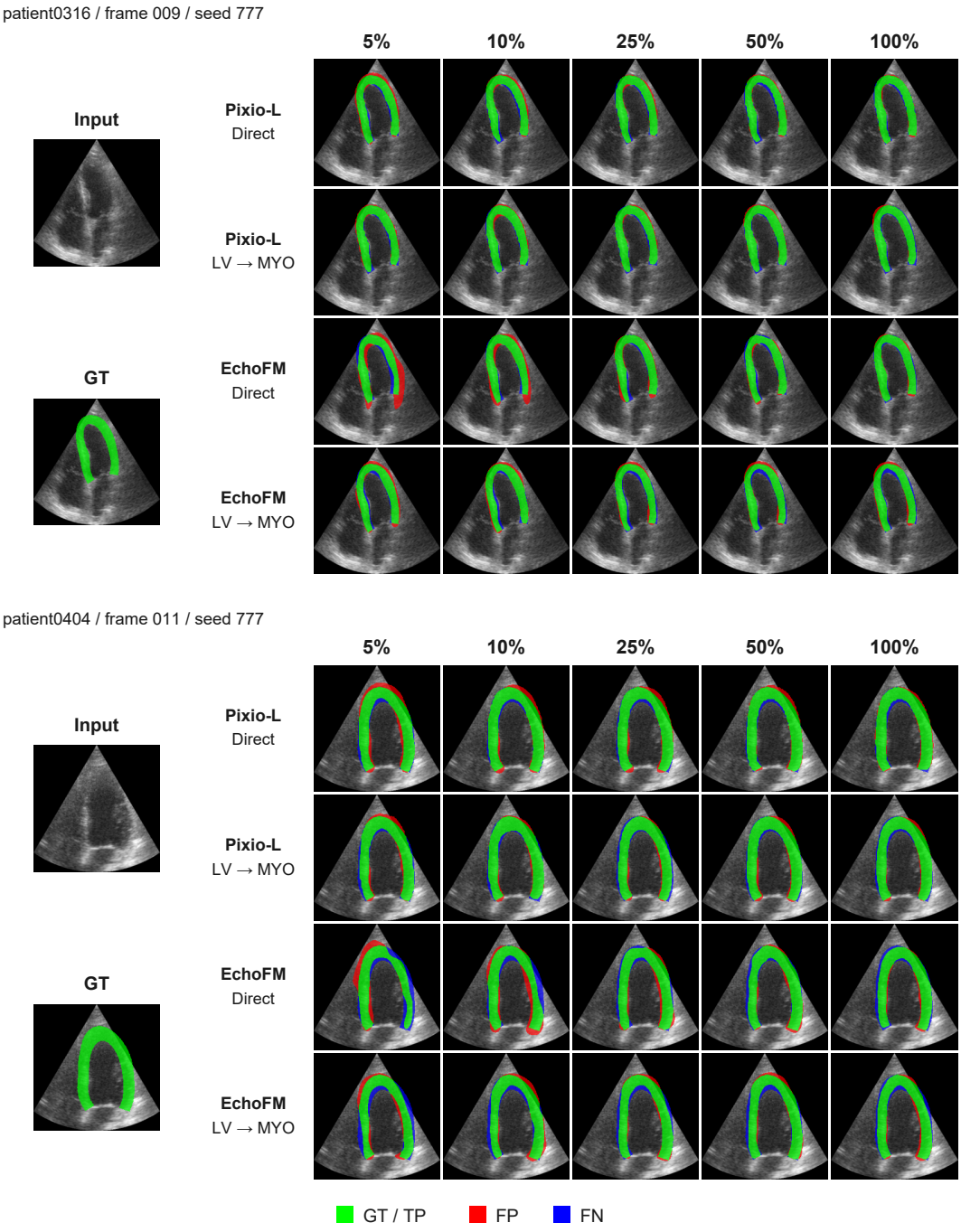


Figure 7: MYO errors at seed 777: patient0316, frame 9 (top) and patient0404, frame 11 (bottom). All five fractions and four system–route configurations follow Figure 4. GT/TP is green, FP red and FN blue, overlaid on the input at 70% opacity. These fixed examples illustrate errors, not a representative patient distribution.

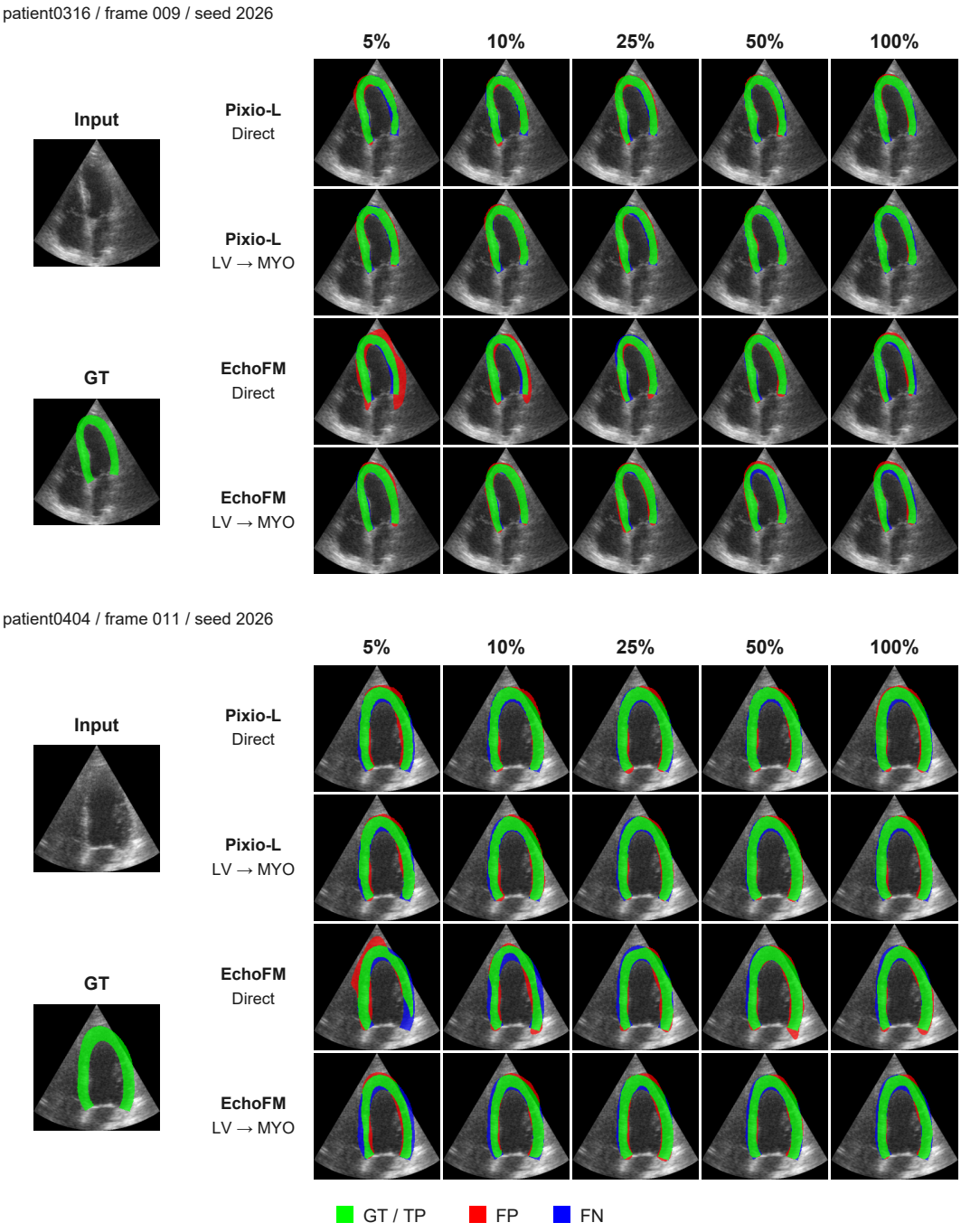


Figure 8: MYO errors at seed 2026: patient0316, frame 9 (top) and patient0404, frame 11 (bottom). All five fractions and four system–route configurations follow Figure 4. GT/TP is green, FP red and FN blue, overlaid on the input at 70% opacity. These fixed examples illustrate errors, not a representative patient distribution.

G FROZEN ENCODER ABLATION

All 60 additional targets use the same 50 evaluation patients and 976 frames as the primary matrix. The source LV models remain fully fine-tuned and are reused, not retrained. Encoders and their buffers remain unchanged at training completion and validation-best selection; segmentation modules change. Frozen parameters are excluded from optimization. Pixio-L freezes 303,377,408 parameters and trains 20,678,147; EchoFM freezes 305,669,120 and trains 12,526,018. Encoder evaluation mode also disables its training-time stochastic layers. Remaining modules retain each system's target recipe; optimization starts afresh. Transfer includes the LV-trained decoder, so freezing is not an encoder-only representation probe. These post-hoc results are exploratory, conditional on the fixed split and fitted seeds. Additional frozen-encoder prediction masks are shown in Appendix G.1; the primary and earlier appendix image panels retain full fine-tuning.

Figures 2 and 3 show the two MYO adaptation modes separately with matched axes and encodings.

Table 13: Pixio-L: full fine-tuning versus frozen MYO encoder. Dice is mean ± sample SD (%). Δ is Frozen minus Full FT in pp; unadjusted 95% paired-patient intervals are conditional on fitted seeds. Contrasts use unrounded scores.

Fraction	Route	Full FT	Frozen	Δ	95% CI
5%	Direct	79.44 ± 0.39	70.14 ± 2.59	-9.30	$[-10.51, -8.12]$
	LV-to-MYO	81.73 ± 0.58	82.49 ± 0.71	+0.76	$[+0.04, +1.48]$
10%	Direct	81.37 ± 0.28	77.38 ± 0.14	-3.99	$[-4.93, -3.08]$
	LV-to-MYO	83.48 ± 0.29	83.94 ± 0.13	+0.46	$[-0.12, +1.08]$
25%	Direct	83.61 ± 0.14	80.70 ± 0.66	-2.91	$[-3.74, -2.16]$
	LV-to-MYO	84.35 ± 0.33	84.74 ± 0.18	+0.39	$[-0.17, +1.00]$
50%	Direct	84.22 ± 0.30	81.90 ± 0.12	-2.31	$[-3.26, -1.53]$
	LV-to-MYO	85.03 ± 0.20	85.15 ± 0.07	+0.12	$[-0.38, +0.65]$
100%	Direct	84.67 ± 0.14	83.88 ± 0.11	-0.79	$[-1.766, +0.003]$
	LV-to-MYO	85.39 ± 0.14	85.60 ± 0.03	+0.20	$[-0.20, +0.63]$

Table 14: EchoFM: full fine-tuning versus frozen MYO encoder. Dice is mean ± sample SD (%). Δ is Frozen minus Full FT in pp; unadjusted 95% paired-patient intervals are conditional on fitted seeds. Contrasts use unrounded scores.

Fraction	Route	Full FT	Frozen	Δ	95% CI
5%	Direct	66.06 ± 0.10	65.79 ± 0.26	-0.27	$[-0.84, +0.27]$
	LV-to-MYO	75.21 ± 1.09	74.83 ± 0.74	-0.39	$[-0.71, -0.07]$
10%	Direct	71.18 ± 0.75	70.87 ± 1.53	-0.31	$[-1.09, +0.42]$
	LV-to-MYO	77.78 ± 0.53	77.72 ± 0.31	-0.06	$[-0.51, +0.41]$
25%	Direct	77.07 ± 0.11	76.94 ± 0.34	-0.12	$[-0.64, +0.41]$
	LV-to-MYO	80.49 ± 0.03	80.81 ± 0.24	+0.32	$[-0.07, +0.70]$
50%	Direct	78.81 ± 0.47	78.85 ± 0.17	+0.04	$[-0.36, +0.45]$
	LV-to-MYO	81.73 ± 0.26	81.95 ± 0.16	+0.22	$[-0.20, +0.63]$
100%	Direct	80.53 ± 0.29	80.42 ± 0.08	-0.12	$[-0.65, +0.43]$
	LV-to-MYO	82.69 ± 0.26	82.90 ± 0.07	+0.20	$[-0.02, +0.42]$

Table 15: Pixio-L: MYO frozen-encoder results. Means ± sample SD over three seeds. HD95 and ASSD are px224; other metrics are percentages.

Fraction	Route	Dice	IoU	HD95	ASSD	Sens.	PPV
5%	Direct	70.14 ± 2.59	54.86 ± 3.04	17.23 ± 4.64	4.91 ± 0.78	77.27 ± 2.67	66.70 ± 4.04
	LV-to-MYO	82.49 ± 0.71	70.61 ± 0.99	6.52 ± 0.34	2.56 ± 0.13	85.13 ± 0.86	81.08 ± 1.82
10%	Direct	77.38 ± 0.14	63.73 ± 0.14	10.76 ± 0.78	3.38 ± 0.12	79.47 ± 2.84	77.20 ± 2.39
	LV-to-MYO	83.94 ± 0.13	72.67 ± 0.18	6.05 ± 0.11	2.38 ± 0.06	87.86 ± 1.41	81.04 ± 1.34
25%	Direct	80.70 ± 0.66	68.27 ± 0.85	8.57 ± 0.40	2.86 ± 0.13	83.40 ± 4.14	79.22 ± 3.15
	LV-to-MYO	84.74 ± 0.18	73.83 ± 0.27	5.79 ± 0.09	2.24 ± 0.03	87.85 ± 1.00	82.38 ± 0.63
50%	Direct	81.90 ± 0.12	69.95 ± 0.16	8.51 ± 0.25	2.71 ± 0.03	84.23 ± 1.62	80.39 ± 1.31
	LV-to-MYO	85.15 ± 0.07	74.47 ± 0.14	5.51 ± 0.09	2.16 ± 0.03	87.67 ± 1.05	83.25 ± 0.97
100%	Direct	83.88 ± 0.11	72.78 ± 0.16	6.94 ± 0.17	2.40 ± 0.02	86.93 ± 0.59	81.62 ± 0.45
	LV-to-MYO	85.60 ± 0.03	75.16 ± 0.03	5.30 ± 0.04	2.08 ± 0.02	87.06 ± 0.64	84.59 ± 0.59

Table 16: EchoFM: MYO frozen-encoder results. Means ± sample SD over three seeds. HD95 and ASSD are px224; other metrics are percentages.

Fraction	Route	Dice	IoU	HD95	ASSD	Sens.	PPV
5%	Direct	65.79 ± 0.26	49.87 ± 0.27	22.99 ± 0.56	5.85 ± 0.10	73.43 ± 1.65	61.65 ± 1.29
	LV-to-MYO	74.83 ± 0.74	60.56 ± 0.93	17.59 ± 0.98	4.18 ± 0.10	78.02 ± 1.06	73.38 ± 0.89
10%	Direct	70.87 ± 1.53	55.74 ± 1.87	21.89 ± 2.03	5.07 ± 0.23	76.56 ± 1.94	67.48 ± 1.76
	LV-to-MYO	77.72 ± 0.31	64.24 ± 0.53	15.40 ± 0.59	3.86 ± 0.02	83.10 ± 0.98	74.18 ± 1.23
25%	Direct	76.94 ± 0.34	63.12 ± 0.39	15.36 ± 0.78	3.79 ± 0.15	79.92 ± 2.13	75.25 ± 2.47
	LV-to-MYO	80.81 ± 0.24	68.33 ± 0.32	10.46 ± 1.03	2.97 ± 0.08	84.49 ± 0.47	78.22 ± 0.55
50%	Direct	78.85 ± 0.17	65.67 ± 0.28	13.08 ± 0.45	3.41 ± 0.04	82.16 ± 1.34	76.73 ± 1.27
	LV-to-MYO	81.95 ± 0.16	69.90 ± 0.23	10.17 ± 1.41	2.83 ± 0.03	85.11 ± 0.57	79.75 ± 0.62
100%	Direct	80.42 ± 0.08	67.78 ± 0.14	11.45 ± 0.99	3.02 ± 0.07	83.24 ± 0.34	78.51 ± 0.41
	LV-to-MYO	82.90 ± 0.07	71.22 ± 0.13	9.16 ± 1.78	2.64 ± 0.15	85.79 ± 1.55	80.82 ± 1.23

G.1 MATCHED PREDICTION ERRORS WITH A FROZEN ENCODER

Figures 9–11 show patient0316, frame 9, the same example already used in Appendix F, at seeds 42, 777 and 2026. Each figure includes both systems, Direct and LV-to-MYO, and all five target-training fractions. Direct freezes the released pretrained encoder during MYO training; LV-to-MYO freezes the encoder of the full-fine-tuned LV model during MYO training. Segmentation modules are trained in both routes. These fixed qualitative examples are not a representative patient distribution or evidence of consistent improvement for every patient.

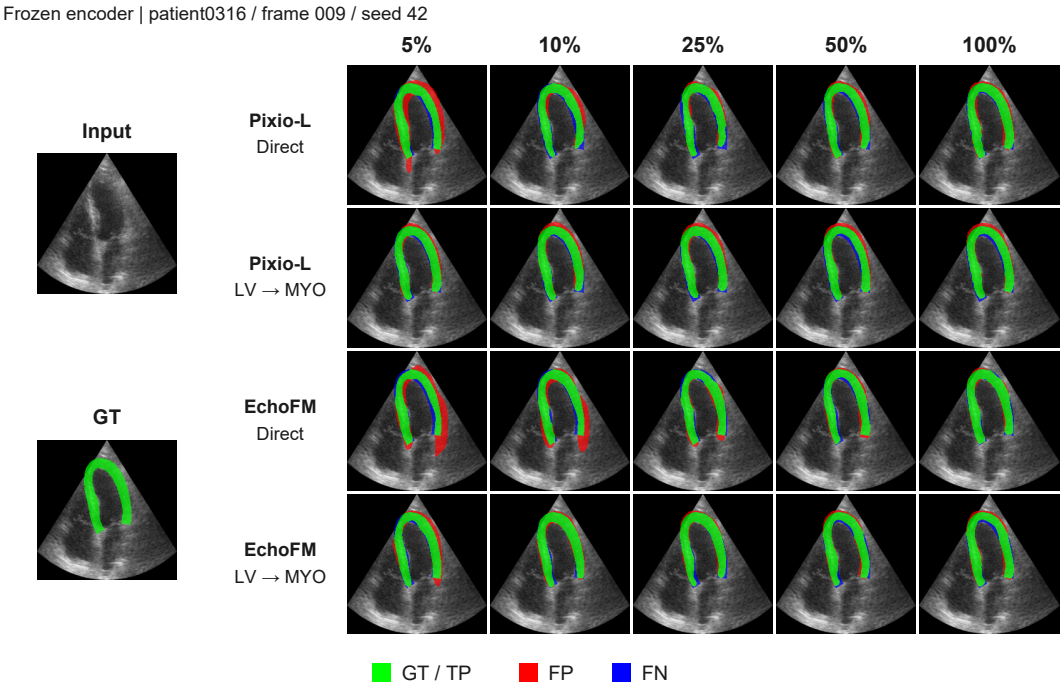


Figure 9: Frozen-encoder MYO errors at seed 42: patient0316, frame 9. Rows show Pixio-L Direct, Pixio-L LV-to-MYO, EchoFM Direct, and EchoFM LV-to-MYO; columns show the five target-training fractions. Input and GT are shared. GT/TP is green, FP red, and FN blue at 70% opacity.

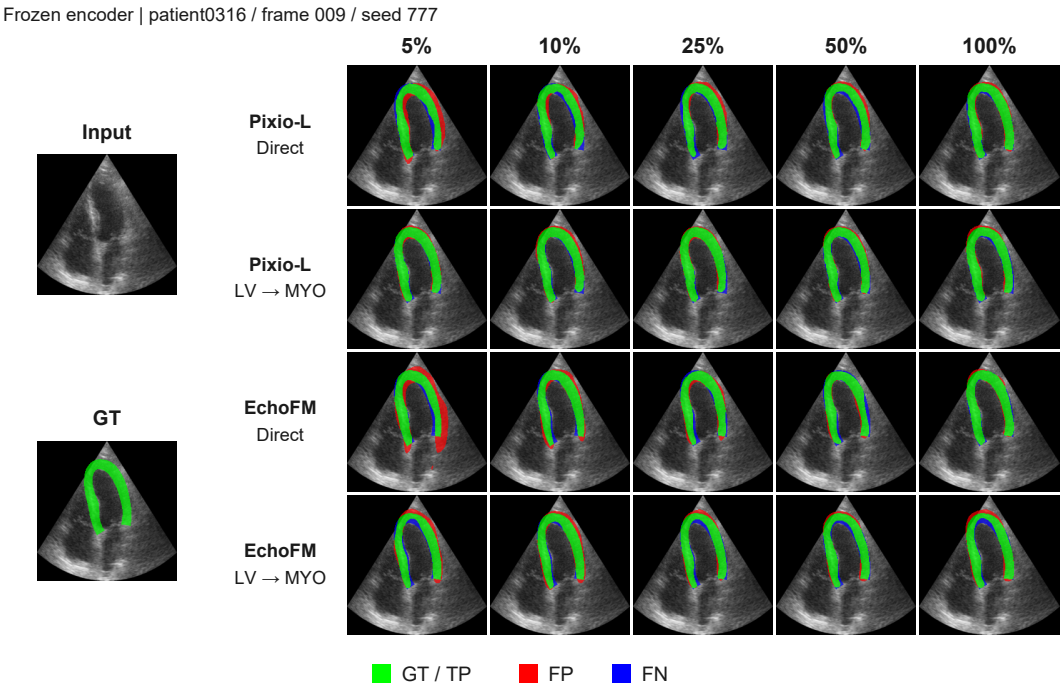


Figure 10: Frozen-encoder MYO errors at seed 777: patient0316, frame 9. Systems, routes, fractions and error colors follow Figure 9.

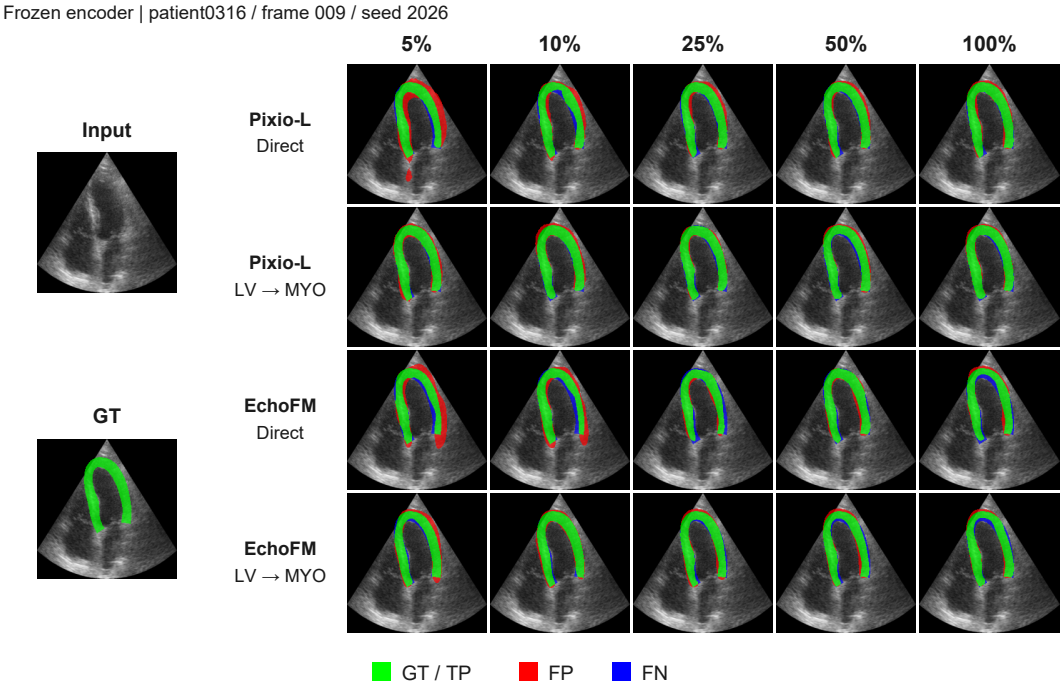


Figure 11: Frozen-encoder MYO errors at seed 2026: patient0316, frame 9. Systems, routes, fractions and error colors follow Figure 9.