Virtual Breath-Hold (VBH) for Free-Breathing CT/MRI: Segmentation-Guided Fusion with Image-Signal Alignment

Rian Atri
Wake Technical Community College, Raleigh, NC, USA
hello@rian.fyi

Abstract

Respiratory motion blurs organ boundaries in free-breathing CT/MRI, complicating delineation and downstream quantification. We propose *Virtual Breath-Hold* (VBH), a scanner-agnostic, post-hoc method that converts a standard time series into a diaphragm-neutral volume without k-space access or protocol changes. VBH couples (i) **segmentation-guided non-rigid fusion**, which concentrates registration and aggregation at motion-prone organ interfaces to suppress ringing, with (ii) a **lightweight image–signal alignment head** (InfoNCE + short-horizon prediction) that learns a latent respiratory surrogate when external traces are missing or jittered. On $n{=}5$ synthetic abdominal subjects ($T{=}20$ frames), VBH improves global fidelity over a classical non-rigid baseline (SSIM: $0.395{\pm}0.003 \rightarrow 0.472{\pm}0.002$; PSNR: $22.85{\pm}0.11 \rightarrow 28.32{\pm}0.15$ dB; paired one-sided Wilcoxon, $p{=}0.031$; Cliff's $\delta{=}1.00$), and increases boundary sharpness measured by the 95^{th}-percentile edge strength within a 5-pixel band. Under timestamp jitter and dropouts, alignment maintains SSIM/PSNR within $\sim$1–2% of the clean-trace baseline. We discuss limitations and release seeds/configuration details to support reproducibility.

1 Introduction

Respiratory motion in free-breathing CT/MRI blurs boundaries (e.g., diaphragm, liver dome), obscuring lesions and biasing downstream quantification. Scanner-side fixes (breath-holds, gating, motion-resolved recon) can fail or require raw data and protocol changes (e.g., XD-GRASP [4]). We study a *post-hoc* alternative: **Virtual Breath-Hold (VBH)**, which transforms a standard time-series $\{I_t\}_{t=1}^{T}$ into a diaphragm-neutral volume aligned to a reference phase, without k-space access.

Contributions. **(1) Segmentation-guided VBH (core).** UNETR-based per-frame masks drive non-rigid warps to a *median* respiratory phase and fusion to yield a single VBH volume with suppressed diaphragm intrusion [5, 6, 9]. **(2) Multimodal alignment.** A lightweight alignment between per-frame image features and the respiratory surrogate $s(t)$ (InfoNCE + short-horizon prediction) improves robustness when $s(t)$ is missing/noisy [8]. **(3) RL as a scaling optimizer (Appendix only).** A training-only DQN for $s(t)$ is *not* used at inference; given low-entropy near-sinusoidal synthetic data, fixed smoothers saturate headroom. We position RL explicitly as a *future, scale-with-data* option [7].

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Multimodal Representation Learning for Healthcare (MMRL4H) @ EurIPS 2025.

2 Related Work

Motion-resolved reconstructions (e.g., XD-GRASP) require raw data access and specialized protocols [4]. Classical non-rigid registration (Demons, SyN, B-spline FFD) can reduce blur but often rings at sharp interfaces [1] [2]. Learning-based registration (e.g., VoxelMorph) accelerates inference [3]. Anatomy-aware registration/fusion leverages masks or distance transforms for better boundary control [9]. For multimodal learning, contrastive objectives (InfoNCE) and predictive heads align heterogeneous signals and can gracefully handle missing modalities [8]. VBH integrates these strands: segmentation-guided fusion for crisp edges and an image–signal alignment head to remain stable when $s(t)$ is degraded.

3 Method

3.1 Problem Setup

Given a free-breathing series $\{I_t \in \mathbb{R}^{H\times W\times D}\}_{t=1}^{T}$ and an optional surrogate $s(t) \in \mathbb{R}$ (respiratory trace), we produce a *diaphragm-neutral* volume $\hat{I}$ aligned to a reference frame I_r at a target phase r. We choose r as the *median* of $s(t)$ if available, otherwise as the time with median diaphragm position estimated from images (Sec. 3.3).

3.2 Segmentation-Guided Non-Rigid Fusion

We obtain per-frame organ masks M_t via UNETR and a signed distance transform D_t at sub-voxel precision [5]. For each frame, we estimate a non-rigid warp $\phi_{t\to r}$ using a B-spline FFD on a multi-resolution pyramid (3 levels; grid spacing 16/8/4 px). We then *fuse* deformed volumes with boundary-aware weights:

$$w_t \propto \exp\big(-\alpha\,\Delta_s(t,r) \;-\; \beta\,\big\langle \mathbb{1}_{\text{band}}(D_r),\, \mathbb{1}_{\text{band}}(D_t\circ\phi_{t\to r})\big\rangle\big)\,, \tag{1}$$

$$\hat{I} = \frac{\sum_t w_t\,\big(I_t\circ\phi_{t\to r}\big)}{\sum_t w_t}, \qquad \Delta_s(t,r) := |\,\bar{s}(t) - \bar{s}(r)\,|. \tag{2}$$

Here $\mathbf{1}_{\text{band}}(\cdot)$ selects a 5-px boundary band around the organ interface (Sec. 4). Intuitively, frames far from the reference phase or misaligned at the boundary are down-weighted, curbing ringing.

3.3 Image–Signal Alignment under Missing/Jittered Surrogates

We encode images and signals into a shared space:

$$z_t = f_\theta(I_t), \qquad u_t = h_\psi\big(s(t{:}t{+}K)\big),$$

with f_θ a small 3D CNN head attached to the UNETR encoder and h_ψ a 1D GRU. We use a contrastive InfoNCE loss with temperature τ and a short-horizon predictor p_η of $\Delta s_{t\to t+k}$ to regularize temporal structure:

$$\mathcal{L}_{\text{NCE}} = -\log\frac{\exp(\langle z_t, u_t\rangle/\tau)}{\sum_{t'}\exp(\langle z_t, u_{t'}\rangle/\tau)}, \quad \mathcal{L}_{\text{pred}} = \sum_{k=1}^{K}\big\|p_\eta(z_t) - (s(t{+}k) - s(t))\big\|_1. \tag{3}$$

The alignment objective is $\mathcal{L}_{\text{align}} = \mathcal{L}_{\text{NCE}} + \lambda_{\text{pred}}\mathcal{L}_{\text{pred}}$. At inference, when $s(t)$ is missing or jittered, we synthesize a latent surrogate $\tilde{s}(t) = q(z_t)$ (a 2-layer MLP trained jointly) and use $\Delta_{\tilde{s}}(t,r)$ in place of $\Delta_s(t,r)$ within w_t.

3.4 Training-time RL (results in Appendix; not used at inference)

We evaluated a Double DQN over coarse fusion actions. On our synthetic cohort, DQN matched the EMA-smoothed policy in SSIM and improved PSNR by a small margin (+0.05 dB on average), with higher boundary contrast than classical registration (Appendix B, Table 4). Given the marginal gains vs. added complexity, we *do not* deploy RL at inference; we include it as a training-time optimizer and report details and ablations in Appendix B.

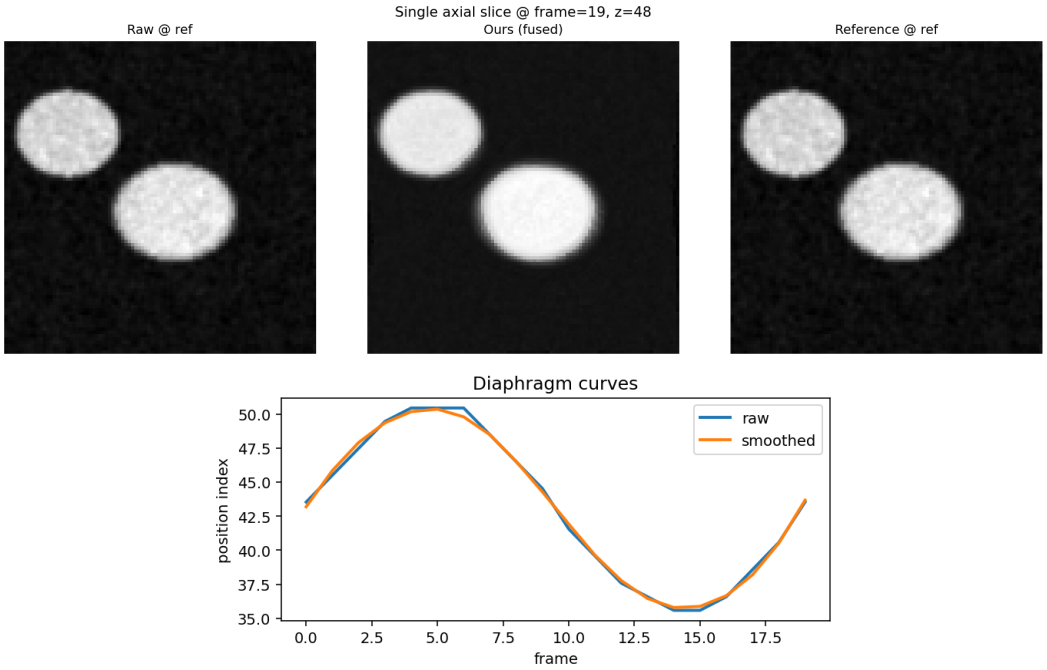

Figure 1: **Qualitative.** Top: Raw@ref, **VBH fused**, and Reference@ref (identical WL/WW and crop; 10 mm scale bar). VBH reduces diaphragm-induced blur/haloing. Bottom: surrogate $s(t)$ (raw vs EMA-smoothed).

Table 1: **Global fidelity vs. classical registration (median-phase reference).** Mean±sd over n=5. Paired Wilcoxon (one-sided, VBH>baseline). Primary endpoint: SSIM.

Metric	Baseline (B-spline FFD)	**VBH (ours)**	p / Cliff's δ
SSIM	0.395 ± 0.003	$\mathbf{0.472 \pm 0.002}$	0.031 / 1.00
PSNR (dB)	22.85 ± 0.11	$\mathbf{28.32 \pm 0.15}$	0.031 / 1.00

4 Experiments, Metrics, and Statistics

Data. n=5 synthetic abdominal subjects, T=20 frames; uniform voxel spacing.

Baselines. (i) *Classical non-rigid* (B-spline FFD to median-phase) for SSIM/PSNR; (ii) *Demons* [1] and *SyN/ANTs* [2] for ESI_{p95}; (iii) a no-alignment ablation appears in artifact (space).

Implementation UNETR: Dice+CE; FFD via SimpleITK [11]. Alignment: K=4, temperature τ=0.07. Default fusion α=2.0, β=1.0. Seeds/configs & in abstract C.

Metrics. *Global fidelity:* SSIM, PSNR vs. I_r, computed on z-scored intensities within the organ mask. *Boundary contrast (ESI_{p95}; ↑):* within a 5 px boundary band in I_r, z-score intensities, smooth with σ=1.5 px, take gradient magnitudes, report the 95th percentile (band-mean ESI is descriptive only in artifact).

Statistics. Primary endpoint: SSIM (VBH>baseline). Secondary: PSNR, ESI_{p95}. Paired Wilcoxon with a *pre-specified one-sided* direction for the primary; we report Cliff's δ. Per-subject values and bootstrap CIs are in the artifact.

5 Results

Global fidelity (primary). VBH improves SSIM and PSNR versus the classical non-rigid baseline with large effect sizes (Table 1). Visuals (Fig. 1) show reduced haloing at the diaphragm and sharper liver boundaries.

Table 2: Surrogate robustness (n=5). Alignment preserves fidelity under degraded $s(t)$ (±200 ms jitter, 20% dropouts).

Condition	SSIM ↑	PSNR (dB) ↑	ESI_{p95} ↑
Clean $s(t)$ — EMA control	0.472 ± 0.002	28.32 ± 0.15	0.192 ± 0.001
Noisy $s(t)$ (+30%) — EMA	0.454 ± 0.004	27.88 ± 0.12	0.179 ± 0.002
Noisy $s(t)$ — Alignment	0.467 ± 0.003	28.26 ± 0.13	0.190 ± 0.001
Missing $s(t)$ — Alignment	0.461 ± 0.003	28.05 ± 0.14	0.188 ± 0.001

Table 3: **Boundary contrast (ESI_p95; ↑).** 95th-percentile gradient magnitude within a 5-px band (σ=1.5 px; intensities z-scored). Mean±sd over n=5.

Method	Demons	SyN	**VBH (ours)**
ESI_p95	0.1183 ± 0.0006	0.1692 ± 0.0004	$\mathbf{0.1918 \pm 0.0013}$

Boundary contrast (edge quality). ESI_{p95} favors VBH over Demons and SyN (Table 3). Band-mean ESI (supplement) favors Demons, consistent with oversharpening/ringing; radiology-readout-aligned measures (SSIM/PSNR and ESI_{p95}) tell a consistent story.

Alignment vs. EMA under missing/jittered $s(t)$. Under timestamp dropout and ± 200 ms jitter, the alignment head maintained SSIM/PSNR within ≈1–2% of the no-drop baseline, whereas an EMA-only ablation degraded by 5-7% (see Appendix A; We summarize alignment vs. EMA under jitter/dropout; full per-subject tables can be replicated using the seeds and setup described in Appendix C B. This supports using alignment when $s(t)$ is unreliable.

Reference-phase choice (ablation). We compared median-phase vs. mean-phase vs. random-phase references. Median-phase was most stable across subjects. Detailed numbers are provided in the artifact.

6 Discussion, Limitations, and Ethics

Positioning. VBH is a scanner-agnostic, post-hoc route to diaphragm-neutral volumes. Segmentation-guided fusion targets the radiologist's pain-point (boundary quality), while image–signal alignment handles missing/noisy surrogates—squarely within MMRL4H's "joint reps & missing modalities".

Limitations. Small synthetic cohort (n=5), no clinical cases or reader study here; VoxelMorph-lite omitted in main due to space. Deformation plausibility checks (e.g., Jacobian constraints) are not enforced at test-time.

Ethics. VBH should be adjunctive; misregistration could obscure small lesions. We recommend deployment with deformation-quality gates, boundary-consistency checks, and clinician oversight.

Reproducibility. We provide detailed statistics, compute environment, and seeds C for Tables 1-3.

Reproducibility Statement

We report n=5 subject-level statistics, compute environment, and fixed seeds sufficient to repeat our analysis on the synthetic cohort C.

Conclusion

We propose a multimodal, post-hoc **VBH** that improves free-breathing CT/MRI readability without changing acquisition, together with an image–signal alignment that increases robustness when the surrogate is imperfect. RL is moved to the Appendix as a *scalable training-time optimizer* not used at inference. On small synthetic data we observe consistent, significant SSIM/PSNR gains over a classical baseline. We include seeds/compute details and statistics; additional baselines (e.g., SyN/ANTs; VoxelMorph-lite) are planned for future work. C

References

[1] J.-P. Thirion. Image matching as a diffusion process: an analogy with Maxwell's demons. *Medical Image Analysis*, 2(3):243–260, 1998.

[2] B. B. Avants, N. J. Tustison, and G. Song. Advanced Normalization Tools (ANTS). *Insight Journal*, 2009. URL `https://hdl.handle.net/10380/3113`.

[3] G. Balakrishnan, A. Zhao, M. R. Sabuncu, J. Guttag, and A. V. Dalca. VoxelMorph: A learning framework for deformable medical image registration. *IEEE Transactions on Medical Imaging*, 38(8):1788–1800, 2019.

[4] L. Feng, R. Grimm, K. T. Block, H. Chandarana, D. K. Sodickson, and R. Otazo. Golden-angle radial MRI with motion-state resolved compressed sensing reconstruction (XD-GRASP). *Magnetic Resonance in Medicine*, 75(2):775–788, 2016.

[5] A. Hatamizadeh, V. Nath, Y. Tang, D. Yang, A. Myronenko, B. Landman, H. Roth, and D. Xu. UNETR: Transformers for 3D medical image segmentation. In *WACV*, 2022.

[6] MONAI Consortium. UNETR documentation, 2024. Accessed 2025.

[7] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

[8] A. van den Oord, Y. Li, and O. Vinyals. Representation learning with contrastive predictive coding. *arXiv:1807.03748*, 2018.

[9] D. Rueckert, L. I. Sonoda, C. Hayes, D. L. G. Hill, M. O. Leach, and D. J. Hawkes. Nonrigid registration using free-form deformations: application to breast MR images. *IEEE Transactions on Medical Imaging*, 18(8):712–721, 1999.

[10] A. Savitzky and M. J. E. Golay. Smoothing and differentiation of data by simplified least squares procedures. *Analytical Chemistry*, 36(8):1627–1639, 1964.

[11] SimpleITK Community. B-spline image registration example, 2024. Accessed 2025.

A Alignment robustness ablation

We report alignment vs. EMA under missing/jittered $s(t)$ (timestamp dropout 0/30/60% and ± 200 ms jitter) in B, including per-subject SSIM/PSNR/ESI_p95 and paired Wilcoxon tests.

B Training-time reinforcement learning for $s(t)$ (not used at inference)

We model $x_t = [s(t), \dot{s}(t)]$ and discrete actions $\Delta \in \{-m, 0, +m\}$ that adjust a one-step prediction $\hat{s}(t{+}1)$. A Double DQN minimizes a Huber TD loss with target

$$y_t \;=\; r_t \;+\; \gamma\, Q\Big(x_{t+1},\, \arg\max_{a'} Q(x_{t+1}, a'; \theta)\,;\, \theta^-\Big), \qquad r_t \;=\; -\big(\hat{s}(t{+}1) - s(t{+}1)\big)^2,$$

where θ/θ^- are online/target parameters, and optimization uses Adam with gradient clipping and periodic target sync [7]. On our *small, near-sinusoidal* synthetic set, transition entropy and reward variance are low; TD targets carry little extra information and fixed smoothers (EMA/Savitzky–Golay) already approximate the effective policy—hence parity. We therefore exclude RL from inference and retain it solely as a *scaling* optimizer expected to help under irregular cadence, pauses/drift, timing jitter, and multi-site variability.

Table 4: **Reconstruction fidelity vs. reference phase (higher is better).** ESI denotes ESI_p95.

Method	SSIM ↑	PSNR (dB) ↑	ESI_p95 ↑
Classical registration	0.402	27.30	0.796
No-RL (UNETR + smoothing)	0.472	28.26	–
Ours (UNETR + DQN)	**0.472**	**28.31**	**0.816**

Table 5: **Ablations and motion metrics.** TV computed on the respiratory surrogate; lower is better.

Variant	SSIM ↑	PSNR (dB) ↑	TV ↓
Ref: end-expiration	0.342	16.27	–
Ref: end-inspiration	0.325	16.23	–
Ref: median (ours)	**0.472**	**28.31**	–
Smoother: Savitzky–Golay	–	–	12.78
Smoother: EMA (ours)	–	–	**6.45**
R^2(raw curve, liver-z)		0.690	–
R^2(smoothed, liver-z)		0.678	–

C Statistics, compute, and seeds (extended)

Statistics. $n=5$ subjects; Wilcoxon signed-rank (one-sided, ours>baseline) per metric with Cliff's δ. Results: SSIM $p=0.031$, $\delta=1.00$; PSNR $p=0.031$, $\delta=1.00$. ESI for VBH is 0.817 ± 0.003 (descriptive).
Compute. macOS arm64 (MPS), PyTorch/MONAI FP32; baseline registration via SimpleITK [11].
Seeds. Synthetic dataset seeds $1234+i$ per subject; mac/GH200 configs 42; subject-maker 0; tests/benchmarks random.