PREDICT-THEN-ADAPT: CROSS-PLANE JOINT-EMBEDDING FOR TEST-TIME MEDICAL SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Test-time medical image segmentation adapts source-trained models to unlabeled target volumes under distribution shifts. Existing methods mainly derive adaptation signals from model predictions or individual slices, leaving the intrinsic three-dimensional structure of test volumes underexploited. In particular, orthogonal views observe the same anatomy with exact voxel-wise correspondence, providing natural within-volume supervision. However, a representation that is predictable across views may not remain compatible with the source-trained segmentation task. We propose **CP-JEPA**, a **C**ross-**P**lane **J**oint-**E**mbedding **P**redictive **A**daptation framework following a **predict-then-adapt** principle. CP-JEPA models cross-plane prediction and task-space translation as two complementary stages. It first predicts masked target-view representations from geometrically corresponding orthogonal views, constructing a JEPA-style objective within each test volume. It then compares the cross-plane prediction with a capacity-matched axial control to isolate geometry-specific counterfactual evidence and translates this evidence into bounded, source-anchored segmentation corrections. CP-JEPA supports both single- and continual-volume adaptation by resetting volume-specific predictive states while selectively retaining reliable task-space evidence across cases. We evaluate CP-JEPA on multi-organ MRI segmentation under single- and continual-volume deployment, with and without labeled target-development data used only for configuration. Across DINOv3- and U-Net-based backbones, CP-JEPA achieves the strongest macro-average performance in three of four evaluation settings and ranks second in the remaining setting, outperforming the strongest TTA baselines by approximately 1.0 Dice point on average.

1 INTRODUCTION

Test-time adaptation (TTA) for medical image segmentation aims to maintain model performance under distribution shifts caused by unseen scanners, acquisition protocols, or imaging contrasts by adapting a source-trained model directly on unlabeled target data Yang et al. (2026); Gao et al. (2026). Unlike conventional domain adaptation, which requires simultaneous access to source and target data, TTA adapts to incoming target cases during deployment and is therefore better suited to clinical settings constrained by data privacy, storage, and transmission. However, clinical cases may arrive either independently or as a continuously evolving stream. An effective adaptation method must therefore extract reliable supervision from the current case while preventing erroneous updates from accumulating across subsequent cases Maharana et al. (2026).

Early works on medical-image TTA mainly rely on model-derived adaptation signals, such as prediction entropy or confidence Dong et al. (2024), normalization statistics Chen et al. (2024), and pseudo-labels Hu et al. (2025). Subsequent methods have incorporated spatial information through three-dimensional shape priors, anatomical atlases, adjacent-slice context, or within-volume prediction consistency Joshi et al. (2025). Despite their effectiveness, these methods still derive supervision primarily from model predictions, source-domain priors, or local slice continuity, which may become unreliable under severe distribution shifts. As illustrated in the upper pathway of Fig. 1, conventional volumetric TTA typically decomposes a 3D volume into a sequence of 2D slices, extracts their representations individually, and subsequently obtains the final segmentation through

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.

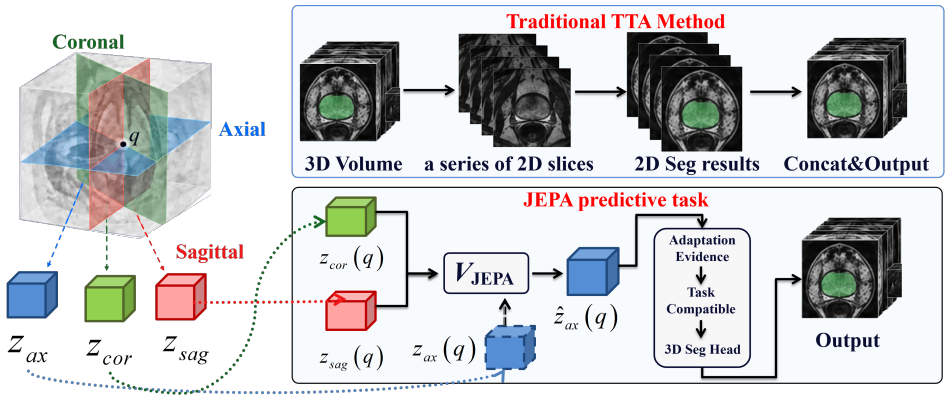


Figure 1: **Comparison of volumetric test-time adaptation paradigms.** Traditional 3D TTA methods usually follow a slice-wise processing strategy, representing a volume as a sequence of 2D slices before volumetric decoding. In contrast, CP-JEPA exploits exact same-voxel correspondence across orthogonal planes to predict masked axial representations from coronal and sagittal context, providing geometry-derived evidence for task-compatible adaptation.

spatial aggregation or volumetric decoding. Although this paradigm incorporates some inter-slice context, three-dimensional information is mainly fused at a downstream stage, and the adaptation process does not explicitly exploit complementary observations of the same voxel from different anatomical planes. In contrast, as shown on the left of Fig. 1, the same anatomical location can be observed from the axial, coronal, and sagittal planes, with exact voxel-wise correspondence across the three views. Despite their different appearances and local contexts, these views describe the same underlying anatomy and therefore provide intrinsic geometric supervision within each unlabeled test volume. Nevertheless, accurate cross-plane prediction alone does not guarantee effective segmentation adaptation: a representation may be predictable from orthogonal views without remaining compatible with the source-trained segmentation task. We refer to this discrepancy as the **predictability–compatibility gap**.

Based on these observations, we propose **CP-JEPA**, a **C**ross-**P**lane **J**oint-**E**mbedding **P**redictive **A**daptation framework. As illustrated in the lower pathway of Fig. 1, CP-JEPA introduces the predictive mechanism of joint-embedding predictive architectures (JEPA) into volumetric TTA. It uses spatially corresponding context from the coronal and sagittal planes to predict masked axial latent representations, rather than reconstructing appearance-dependent pixels Ferdaus et al. (2026). Accordingly, CP-JEPA treats cross-plane predictions as adaptation evidence and translates them into task-compatible corrections. Specifically, CP-JEPA compares the cross-plane prediction with a capacity-matched axial control to isolate counterfactual evidence attributable to exact cross-plane geometry. A source-trained task-compatible module then translates this evidence into bounded segmentation corrections anchored to the frozen source prediction, while conservative gating suppresses updates induced by unreliable evidence and reduces the risk of negative adaptation.

CP-JEPA supports **single-** and **continual-volume** adaptation. In the single-volume setting, each case is adapted using only the current test volume. In the continual-volume setting, volume-specific states are still reset between cases, whereas reliable task-space evidence is retained within the same target stream and gated by its agreement with the current evidence. We further consider no-target-development and target-development evaluation protocols and evaluate CP-JEPA on prostate, cardiac, and liver MRI segmentation using different architectures, including frozen DINOv3 and 2D/3D U-Nets. Across different anatomies, backbones, and deployment protocols, CP-JEPA consistently improves the corresponding source models. Experiments demonstrate the respective contributions of exact cross-plane correspondence, task-compatible translation, and continual evidence memory.

- We identify **cross-plane correspondence** within medical volumes as a natural source of test-time self-supervision, and formulate volumetric adaptation as a JEPA-style latent prediction problem that can be optimized within a single unlabeled test volume.
- We identify the **predictability–compatibility gap**: improving cross-plane latent prediction does not necessarily improve segmentation performance. To address this issue, we intro-

duce cross-plane counterfactual evidence and a task-compatible translation mechanism that converts predictive information into segmentation corrections.

- We develop a framework that supports **single-** and **continual-volume** TTA, and validate it across foundation-model and U-Net backbones. Strong experimental results demonstrate the effectiveness of our framework across multi-organ MRI segmentation tasks.

2 Related Works

Test-Time Adaptation for Medical Image Segmentation. TTA adapts a source-trained model to unlabeled target data during deployment without revisiting source-domain supervision. Early methods such as TENT (Wang et al., 2021) minimize prediction entropy by updating normalization parameters. Subsequent works improve stability and reliability through entropy-weighted integration (Dong et al., 2024), visual prompt learning (Chen et al., 2024), evaluation-guided dynamic adaptation (Wu et al., 2025), progressive energy adaptation (Zhang et al., 2025b), multi-view co-training (Joshi et al., 2025), latent CRF refinement (Chen et al., 2025a), and source-friendly continual adaptation (Wu et al., 2026a). These strategies have been extended to medical image segmentation, where variations in scanners, institutions, acquisition protocols, and imaging sequences can induce substantial domain shifts.

Multi-View and Volumetric Consistency. Medical volumes naturally provide complementary axial, coronal, and sagittal observations of the same anatomy. Existing multi-view methods exploit agreement across anatomical orientations through prediction fusion, feature alignment, or consistency regularization (Joshi et al., 2025). Related volumetric approaches build reusable correspondence spaces from frozen 2D foundation models (Tombak et al., 2026), adapt promptable segmentation models to 3D images and videos (Zhu et al., 2024; Ma et al., 2025), extend 2D biomedical foundation models to volumetric inputs (Zhuang et al., 2025), and perform few-shot adaptation without fine-tuning (He et al., 2025). JEPA models further learn volumetric representations from masked context (Trinh et al., 2026; Kondepudi et al., 2025). These methods show that multi-view information can improve robustness, but they primarily treat different views as complementary inputs to be fused or regularized.

Joint-Embedding Predictive Learning. JEPA learns representations by predicting latent embeddings of masked or unobserved content from contextual observations instead of reconstructing raw pixels (Assran et al., 2023). This principle has been scaled to general visual representations (Siméoni et al., 2025) and extended to promptable image and video segmentation (Ravi et al., 2024). Recent works apply it to volumetric medical data (Trinh et al., 2026; Kondepudi et al., 2025) and to reusable cross-modal correspondence (Tombak et al., 2026). Related test-time and multi-view adaptation methods exploit similar predictive or consistency signals without updating the source model (Joshi et al., 2025; Wu et al., 2026a). However, existing JEPA methods mainly focus on learning general-purpose representations before downstream deployment, rather than using latent prediction itself as an on-the-fly adaptation signal.

3 Methodology

As illustrated in Fig. 2, CP-JEPA follows a **predict-then-adapt** principle. Given an unlabeled target volume, we first exploit exact cross-plane correspondence to construct a volume-specific JEPA task. We then compare cross-plane prediction with a capacity-matched axial control to extract geometry-specific counterfactual evidence, which is translated into a bounded task-space correction. The complete pipeline of our proposed CP-JEPA for single- and continual-volume adaptation is summarized in Algorithm 1. For continual deployment, CP-JEPA retains only reliable task-space evidence from previously observed volumes while resetting all volume-specific predictive states.

3.1 Problem Setup and Cross-Plane Representation

Let $f_\theta : \mathcal{X} \to \mathcal{Y}$ denote a segmentation model trained on a labeled source dataset $\mathcal{D}_s = \{(X_i^s, Y_i^s)\}$. At deployment, the model receives an unlabeled target volume $X_t \in \mathbb{R}^{D\times H\times W}$ from a potentially shifted distribution, while the source model remains frozen and produces $\ell_t^{\mathrm{src}} = f_\theta(X_t)$. Our goal is to improve this prediction without accessing target annotations.

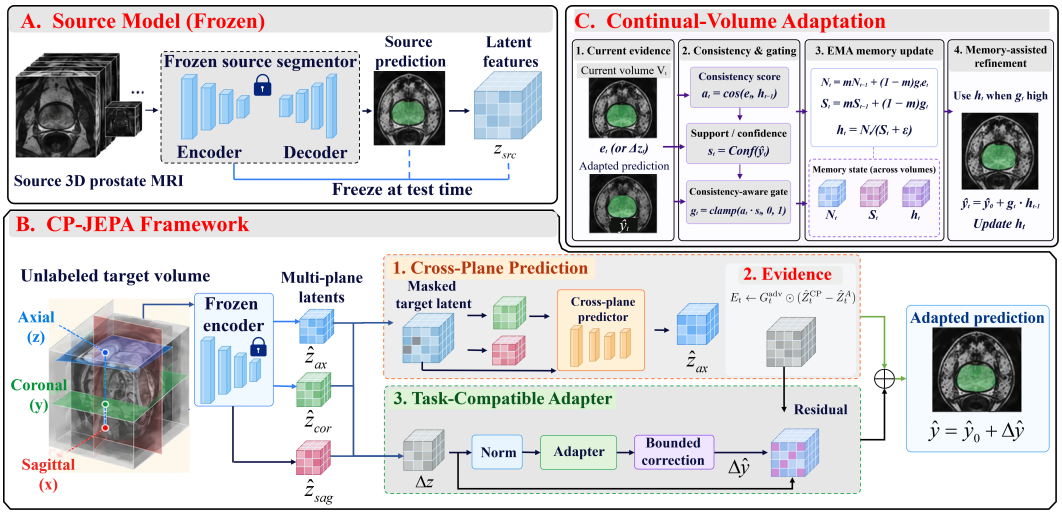


Figure 2: **Framework of CP-JEPA.** (A) The frozen source model produces an initial prediction. (B) Features from three orthogonal views predict masked axial features, and the resulting evidence yields a bounded segmentation correction. (C) Continual adaptation retains reliable evidence across volumes while reinitializing the predictor for each volume.

A volumetric medical image provides three canonical observations of the same anatomy. We therefore reformat X_t into axial, coronal, and sagittal views and extract features $\{z_v^a, z_v^c, z_v^s\}$ using a shared frozen encoder, and associate each voxel with a positional descriptor q_v.

3.2 Volume-Specific Joint-Embedding Prediction

As shown in Fig. 2B, CP-JEPA treats the axial representation as the prediction target and the two orthogonal views as complementary contexts. For each target volume, we initialize lightweight predictors P_c and P_s, producing $\hat{z}_v^c = P_c(z_v^c, q_v)$ and $\hat{z}_v^s = P_s(z_v^s, q_v)$. Their outputs are aggregated into the cross-plane prediction

$$\hat{z}_v^{\mathrm{CP}} = \mathrm{Norm}\left(\frac{\hat{z}_v^c + \hat{z}_v^s}{2}\right). \tag{1}$$

Rather than reconstructing pixels, the predictors directly estimate the latent target z_v^a, allowing each unlabeled test volume to provide its own JEPA-style supervision. To separate information contributed by exact cross-plane geometry from information already predictable from ordinary through-plane context, we simultaneously fit a capacity-matched axial predictor $\hat{z}_v^{\mathrm{A}} = P_a(c_v^a, q_v)$, which belongs to the same predictor family but receives no orthogonal-view information. All predictors are optimized against the detached axial target:

$$\mathcal{L}_{\mathrm{pred}} = \frac{1}{|\mathcal{M}|}\sum_{v\in\mathcal{M}}\sum_{r\in\{c,s,\mathrm{A}\}}\left[1-\cos\left(\hat{z}_v^r, \mathrm{sg}(z_v^a)\right)\right], \tag{2}$$

where $\mathrm{sg}(\cdot)$ denotes stop-gradient and $\mathcal{M}$ contains the masked prediction locations. The predictors are optimized only for the current volume and discarded after inference, preventing predictive parameters from being transferred across target cases.

3.3 Counterfactual Evidence and Task-Compatible Adaptation

Directly replacing z_v^a with $\hat{z}_v^{\mathrm{CP}}$ is not necessarily beneficial for segmentation. A representation can become more predictable from orthogonal views while becoming less compatible with the task space expected by the source-trained segmentor, resulting in the **predictability–compatibility gap**. We therefore use cross-plane prediction as adaptation evidence rather than directly replacing the source representation. For each voxel, we first measure the predictive advantage of the cross-plane predictor over the axial control: $a_v = \cos(\hat{z}_v^{\mathrm{CP}}, \bar{z}_v^a) - \cos(\hat{z}_v^{\mathrm{A}}, \bar{z}_v^a), \bar{z}_v^a = \frac{z_v^a}{\|z_v^a\|_2}$. Only positive advantage is

retained and normalized within the current volume as

$$g_v^{\mathrm{adv}} = \mathrm{clip}([a_v]_+/(Q_\tau + \epsilon), 0, 1), \tag{3}$$

where Q_τ denotes the intra-volume τ-quantile of positive advantages. The **counterfactual evidence** is $e_v = g_v^{\mathrm{adv}} \left(\hat{z}_v^{\mathrm{CP}} - \hat{z}_v^{\mathrm{A}}\right)$. Here, $\hat{z}_v^{\mathrm{A}}$ captures what can already be inferred without orthogonal geometry, while e_v isolates the additional predictive direction attributable to exact cross-plane correspondence. To translate this latent evidence into task-compatible information, we train a lightweight adapter A_ω once on source-domain annotations and keep it frozen during deployment. Given the counterfactual evidence e_t, frozen task bottleneck b_t, and source logits ℓ_t^{src}, the adapter predicts

$$(\Delta\ell_t, g_t) = A_\omega\left(e_t, b_t, \ell_t^{\mathrm{src}}\right), \qquad \Delta\ell_t = \Delta_{\max}\tanh(h_\Delta), \quad g_t = \sigma(h_g). \tag{4}$$

The resulting single-volume prediction is $\ell_t^{\mathrm{single}} = \ell_t^{\mathrm{src}} + g_t \odot \Delta\ell_t$. Thus, the source prediction remains the anchor and the adapter can only apply a bounded residual correction.

During source-domain training, we construct four matched evidence branches $k \in \{\mathrm{Exact}, \mathrm{Axial}, \mathrm{Shuffled}, \mathrm{Zero}\}$. The Exact branch is trained to improve segmentation, whereas the control branches are encouraged to remain close to the source prediction. The adapter objective is

$$\begin{aligned}\mathcal{L}_{\mathrm{adapter}} = \mathcal{L}_{\mathrm{seg}}(\ell^{\mathrm{Exact}}, Y^s) + \lambda_{\mathrm{id}}\mathcal{L}_{\mathrm{id}} + \lambda_{\mathrm{anchor}}\|\ell^{\mathrm{Exact}} - \ell^{\mathrm{src}}\|_2^2 \\ + \lambda_{\mathrm{rank}}\mathcal{L}_{\mathrm{rank}} + \lambda_g \overline{g}^{\mathrm{Exact}},\end{aligned} \tag{5}$$

where the identity and ranking terms suppress corrections produced by Axial, Shuffled, and Zero evidence. This encourages the adapter to respond specifically to task-relevant information supported by exact cross-plane geometry rather than to additional model capacity.

3.4 Single- and Continual-Volume Adaptation

CP-JEPA explicitly separates *volume-specific predictive states* from *cross-volume task-space evidence*. In single-volume adaptation, P_c, P_s, and P_a are initialized and optimized independently for each target volume and discarded after producing ℓ_t^{single}. Therefore, the prediction of the current case depends only on information contained within that volume.

For continual adaptation, target volumes arrive as a causal stream $\{X_1, \ldots, X_T\}$. Instead of retaining the volume-specific predictors, we maintain a compact memory of reliable task-space residuals. Let $r_t = \mathcal{C}(\ell_t^{\mathrm{single}} - \ell_t^{\mathrm{src}})$ denote the centered current residual and ρ_t a label-free reliability map. The historical evidence is maintained by

$$\begin{aligned} N_t &= \mu N_{t-1} + (1-\mu)\rho_t r_t, \\ S_t &= \mu S_{t-1} + (1-\mu)\rho_t, \\ h_t &= \frac{N_t}{S_t + \epsilon}. \end{aligned} \tag{6}$$

For the first case, the implementation initializes $N_1 = \rho_1 r_1$ and $S_1 = \rho_1$ without multiplying by $(1-\mu)$; here ρ_t is the write weight for memory update.

Algorithm 1 CP-JEPA for Single- and Continual-Volume TTA

Require: Source model f_θ; task adapter A_ω; target volumes $\{X_t\}_{t=1}^T$
Ensure: Predictions $\{\hat{Y}_t\}_{t=1}^T$
1: Initialize evidence memory $\mathcal{H} \leftarrow \varnothing$
2: **for** $t = 1, \ldots, T$ **do**
3: Extract multi-plane features $\{Z_t^a, Z_t^c, Z_t^s\}$ from X_t
4: Obtain source logits ℓ_t^{src}, bottleneck b_t, context C_t^a, and coordinates Q_t
5: Initialize disposable predictors P_c, P_s, P_a
6: **for** $j = 1, \ldots, K$ **do**
7: Update P_c, P_s, P_a using $\mathcal{L}_{\mathrm{pred}}$
8: **end for**
9: $(\hat{Z}_t^c, \hat{Z}_t^s, \hat{Z}_t^A) \leftarrow \textsc{Predict}(P_c, P_s, P_a, Q_t)$
10: $\hat{Z}_t^{\mathrm{CP}} \leftarrow \mathrm{Norm}((\hat{Z}_t^c + \hat{Z}_t^s)/2)$
11: $E_t \leftarrow G_t^{\mathrm{adv}} \odot (\hat{Z}_t^{\mathrm{CP}} - \hat{Z}_t^A)$
12: $(\Delta\ell_t, g_t) \leftarrow A_\omega(E_t, b_t, \ell_t^{\mathrm{src}})$
13: $\ell_t \leftarrow \ell_t^{\mathrm{single}} \leftarrow \ell_t^{\mathrm{src}} + g_t \odot \Delta\ell_t$
14: **if** mode $=$ Continual and $\mathcal{H} \neq \varnothing$ **then**
15: $(h_{t-1}, \eta_t) \leftarrow \textsc{ReadMemory}(\mathcal{H}, E_t)$
16: $\ell_t \leftarrow \ell_t + \lambda_{\mathrm{mem}}\eta_t h_{t-1}$
17: **end if**
18: $\hat{Y}_t \leftarrow \textsc{SafetyCheck}(\ell_t, \ell_t^{\mathrm{src}})$
19: **if** mode $=$ Continual **then**
20: $(\rho_t, r_t) \leftarrow \textsc{ReliableResidual}(\ell_t^{\mathrm{single}}, \ell_t^{\mathrm{src}})$
21: $\mathcal{H} \leftarrow \textsc{UpdateMemory}(\mathcal{H}, \rho_t, r_t)$ ▷ after prediction
22: **end if**
23: Discard P_c, P_s, P_a and volume-specific states
24: **end for**
25: **return** $\{\hat{Y}_t\}_{t=1}^T$

Before the current case is written into memory, h_{t-1} is compared with the current residual r_t. As illustrated in Algorithm equation 1 (Lines 16-21), the memory is read before it is updated with the

current case, ensuring causal continual adaptation. Their directional agreement, joint support, and prediction confidence determine a consistency-aware gate $\eta_t \in [0, \eta_{\max}]$. The continual prediction is then

$$\ell_t^{\text{cont}} = \ell_t^{\text{single}} + \lambda_{\text{mem}} \eta_t h_{t-1}. \tag{7}$$

As illustrated in Fig. 2C, the memory is read before it is updated with the current case. Consequently, the current volume cannot serve as its own historical evidence, and no future target volume is accessed, preserving causal continual adaptation.

4 Experiments

4.1 Experimental Settings

Prostate T2 MRI. We evaluate cross-dataset prostate segmentation using RegPro Yu et al. (2025) as the labeled source domain and seven unseen T2 MRI cohorts as target domains: PROSTATEx Cuocolo et al. (2021), NCI-ISBI Vacacela & Benalcázar (2019), PROMISE12 Litjens et al. (2014), MSD-Prostate Antonelli et al. (2022), QIN-Prostate Fedorov et al. (2017), Polish Prostate Gibała et al. (2023), and Prostate158 Adams et al. (2022), comprising 493 patients and 508 scans in total. We combine a frozen DINOv3 model with a 3D U-Net segmentation head.

M&Ms Cardiac MRI. We use Domain A of M&Ms Campello et al. (2021) as the source domain and Domains B/C/D as unseen target domains, containing 125, 75, and 50 patients, respectively. The task segments the left ventricle, myocardium, and right ventricle using a 2D U-Net.

CHAOS Liver MRI. We evaluate cross-contrast liver segmentation on CHAOS Kavur et al. (2021) using T1 In-phase, T1 Out-of-phase, and T2-SPIR MRI, forming six directed transfer tasks. Each source sequence is used to train a 3D nnU-Net, while the remaining sequences serve as unlabeled target domains. Since the three sequences contain overlapping subjects, this experiment measures cross-contrast rather than cross-patient generalization.

Compared Methods and Implementation Details. To evaluate the generalizability of CP-JEPA across segmentation architectures, we use different backbone configurations for the three dataset families. For prostate T2 MRI, we adopt a hybrid model consisting of a frozen DINOv3 encoder and a 3D U-Net task network. For M&Ms cardiac MRI, we use a 2D U-Net trained on Domain A to segment the left ventricle, myocardium, and right ventricle. For CHAOS liver MRI, we use a 3D nnU-Net with a separate source checkpoint for each MRI sequence. Within each dataset family, all methods use the same segmentation backbone and source checkpoint.

We compare CP-JEPA with the frozen *Source Only* baseline and generic TTA methods, including TENT Wang et al. (2021), COME Zhang et al. (2025a), StatA Zanella et al. (2025), ADTE Wu et al. (2026b), ASR Fang et al. (2026), CAZO Zhang et al. (2026), GOLD Lai et al. (2026), IMSE Baek et al. (2026), PASLE Hu et al. (2025), PEA Ma et al. (2026), REM Han et al. (2025), and SNAP Cha et al. (2025). We also include methods developed for medical image segmentation, namely TEGDA Zhou et al. (2025), GRATA Chen et al. (2025c), PTTEA Zhang et al. (2025b), SPEGC Du et al. (2026), TACT Chen et al. (2025b), and SicTTA Wu et al. (2026a). All methods are evaluated under both NTD and TD protocols with single- and continual-volume deployment. Labeled target volumes in TD are used only for hyperparameter selection. In the single-volume setting, all adaptation states are reset after each case; in the continual-volume setting, only task-space evidence is retained across cases. All experiments were conducted on a server equipped with 8 NVIDIA RTX 4090 GPUs, and each experiment was repeated using three random seeds.

4.2 Comparison with State-of-the-Art Methods

Table 1 compares CP-JEPA with representative TTA methods using matched backbones. We evaluate two deployment strategies. Under no-target-development (NTD), adaptation is performed directly on unseen target data without labeled target samples. Under target-development (TD), labeled target-domain volumes are used to configure the adaptation procedure before inference. Both strategies are evaluated under single-volume (S) and continual-volume (C) deployment.

Table 1: Comparison with state-of-the-art test-time adaptation methods. Results are reported under no-target-development (NTD) and target-development (TD) protocols with single-volume (S) and continual-volume (C) deployment. Average denotes the equally weighted mean over Prostate T2, M&Ms, and CHAOS. The best and second-best adaptation results are highlighted by **red** and blue, respectively.

Method	Reference	Prostate T2				M&Ms				CHAOS				Average			
		NTD		TD		NTD		TD		NTD		TD		NTD		TD	
		S	C	S	C	S	C	S	C	S	C	S	C	S	C	S	C
Source Only	–	0.6829	0.6829	0.6829	0.6829	0.7861	0.7861	0.7861	0.7861	0.7084	0.7084	0.7084	0.7084	0.7258	0.7258	0.7258	0.7258
TENT	ICLR'21	0.6829	**0.7107**	0.6829	0.7250	0.8114	0.8113	0.8113	0.8113	0.7147	**0.7576**	0.7005	0.7463	0.7363	**0.7599**	0.7316	0.7609
COME	ICLR'25	0.6829	0.6589	0.6829	0.6763	0.8171	0.5042	0.8175	0.8109	0.7088	0.7424	0.6993	0.7407	0.7363	0.6352	0.7332	0.7426
TEGDA	MICCAI'25	0.6843	0.6858	0.6850	0.6819	0.8171	0.8190	0.8174	0.8206	0.7085	0.7086	0.7088	0.7086	0.7366	0.7378	0.7371	0.7370
StatA	CVPR'25	0.5277	0.5277	0.5426	0.5426	0.8075	0.8075	0.8075	0.8075	0.7113	0.7113	0.7081	0.7114	0.6822	0.6822	0.6861	0.6872
ADTE	ICLR'26	0.6947	0.6947	0.6958	0.6958	0.7890	0.7214	0.7902	0.7427	0.7097	0.7097	0.7095	0.7095	0.7311	0.7086	0.7318	0.7160
ASR	AAAI'26	0.6829	0.6829	0.6829	0.6829	0.8174	0.8174	0.8174	0.8174	0.7084	0.7084	0.7085	0.7084	0.7362	0.7362	0.7363	0.7362
CAZO	CVPR'26	0.6829	0.6929	0.6829	0.6930	0.7861	0.7840	0.7861	0.7844	0.7084	0.7085	0.7084	0.7085	0.7258	0.7285	0.7258	0.7286
GOLD	CVPR'26	0.6829	0.6829	0.6829	0.6829	0.7861	0.7862	0.7861	0.8019	0.7084	0.7084	0.7084	0.7084	0.7258	0.7258	0.7258	0.7311
GRATA	AAAI'25	0.6830	0.6911	0.6873	0.6956	0.8174	0.8173	0.8175	0.8174	0.7088	0.7100	0.7068	0.7343	0.7364	0.7395	0.7372	0.7491
IMSE	ICLR'26	0.6829	0.5677	0.6829	0.6680	0.8191	0.2564	0.8191	0.6928	0.7084	0.7085	0.7085	0.7092	0.7368	0.5109	0.7368	0.6900
PASLE	ICLR'25	0.6829	0.6534	0.6829	0.6929	0.8208	0.5677	0.8205	0.8172	0.7087	0.7101	0.7039	0.7090	0.7375	0.6437	0.7358	0.7397
PEA	ICLR'26	0.6710	0.6927	0.6753	0.6939	0.7970	0.8126	0.8041	0.8148	0.7012	0.6904	0.7032	0.7039	0.7231	0.7319	0.7275	0.7375
PTTEA	ICCV'25	0.5214	0.1451	0.6748	0.4216	0.7568	0.0357	0.8214	0.1232	0.5220	0.1458	0.7067	0.3168	0.6001	0.1089	0.7343	0.2872
REM	ICML'25	0.6829	0.6896	0.6829	0.6790	0.8177	0.8177	0.8176	0.8178	0.7063	0.7371	**0.7259**	**0.7508**	0.7356	0.7481	0.7421	0.7492
SNAP	NeurIPS'25	0.6829	0.6829	0.6829	0.6830	0.8278	0.8278	**0.8278**	0.8278	0.7084	0.7084	0.7084	0.7084	0.7397	0.7397	0.7397	0.7397
SPEGC	CVPR'26	0.6829	0.6425	0.6829	0.6826	0.6856	0.0387	0.7820	0.2912	0.7030	0.6654	0.7046	0.7029	0.6905	0.4489	0.7232	0.5589
TACT	ICCV'25	0.6921	0.6613	0.6890	0.6615	0.7726	0.7731	0.7830	0.7823	0.5505	0.5174	0.5499	0.5173	0.6717	0.6506	0.6740	0.6537
SicTTA	MedIA'26	0.6553	0.6949	0.6562	0.7112	0.8239	0.8340	0.8173	0.8333	**0.7166**	0.7161	0.7069	0.7221	0.7319	0.7483	0.7268	0.7555
CP-JEPA	Ours	**0.7030**	0.6979	**0.7164**	**0.7277**	**0.8373**	**0.8404**	0.8195	**0.8462**	0.7091	0.7123	0.7189	0.7475	**0.7498**	0.7502	**0.7516**	**0.7738**

Prostate T2 MRI. Under NTD, CP-JEPA improves over the source model by 2.01 and 1.50 dice points under single- and continual-volume deployment, respectively, although continual deployment performs below single-volume deployment. This behavior reveals a limitation of the consistency-aware gate: historical evidence may be admitted when it agrees with the current evidence but does not benefit the current case. Nevertheless, performance remains above the source model. Under TD, CP-JEPA improves over the source model by 3.35 and 4.48 Dice points under single- and continual-volume deployment, respectively. These results indicate that limited target-domain development can better calibrate the gate and more effectively exploit continual evidence accumulation.

M&Ms Cardiac MRI. On M&Ms, CP-JEPA improves over the frozen source model by 5.12 and 5.43 Dice points under NTD single- and continual-volume deployment, respectively. The NTD single-volume result also exceeds its TD counterpart by 1.78 points, suggesting that cross-plane supervision can extract strong adaptation signals directly from each unlabeled volume, while configuration selection using only a few labeled volumes may introduce variance across heterogeneous cardiac domains. Continual adaptation remains highly effective under both strategies, providing gains of 5.43 points under NTD and 6.01 points under TD.

CHAOS MRI. Under NTD, as shown in Fig. 3a, CP-JEPA improves over the frozen source model by only 0.07 and 0.39 Dice points under single- and continual-volume deployment, respectively. This reflects the intended conservative behavior of the framework: when cross-plane evidence is insufficiently reliable, the gate suppresses the correction and tends to preserve the source prediction. In contrast, several competing methods fall below the frozen baseline. Under TD, CP-JEPA improves over the frozen model by 1.05 and 3.91 Dice points under single- and continual-volume deployment.

Overall Comparison. CP-JEPA achieves the strongest macro-average performance in three of the four evaluation settings and ranks second in the remaining setting. Relative to the frozen source model, it improves the average Dice by 2.40 and 2.44 points under NTD single- and continual-volume deployment, respectively. Under TD, the corresponding gains are 2.58 and 4.80 points. These results show that CP-JEPA combines direct adaptation from unlabeled target volumes with stable source-anchored correction. They also indicate that labeled target-domain development set can improve the calibration of the gating and memory mechanisms, with the largest benefit observed under continual- deployment. The ablations in Sec. 4.3 further examine the contributions of

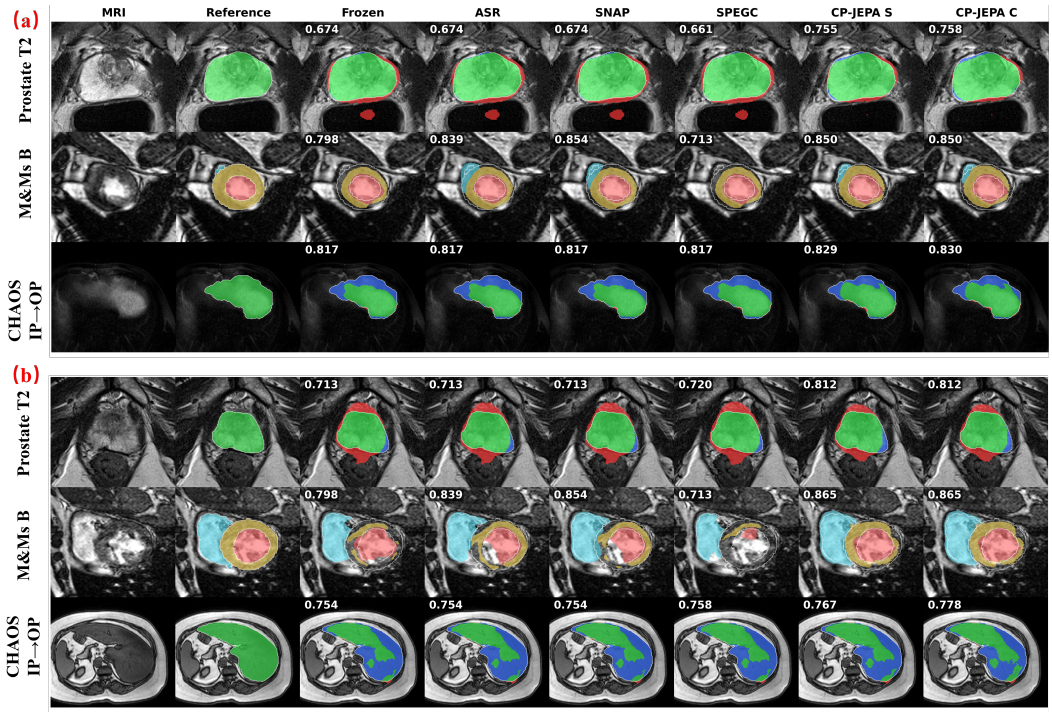


Figure 3: Qualitative comparison on prostate T2, M&Ms, and CHAOS MRI cases under (a) NTD and (b) TD selection. Baselines use continual-volume predictions; CP-JEPA is shown in single- (S) and continual- (C) modes. For binary masks, green indicates correctly segmented foreground, red false positives, and blue missed foreground. Cardiac colors distinguish LV, MYO, and RV.

Table 2: Ablation Studies on Prostate T2 dataset. (a) Effects of the selection protocol under single- and continual-volume deployment. (b) Cross-plane evidence controls under the NTD protocol.

(a) Selection protocol and deployment mode

Protocol	Deployment	Frozen	Direct	CP-JEPA
NTD	Single	0.6829	0.6824	**0.7030**
	Continual	0.6829	0.6824	**0.6979**
TD	Single	0.6829	0.6828	**0.7164**
	Continual	0.6829	0.6828	**0.7277**

(b) Cross-plane evidence controls

Evidence branch	Single	Continual
No evidence (frozen)	0.6829	0.6829
Axial only	0.6830	0.6830
Shuffled cross-plane	0.6832	0.6832
Exact cross-plane	**0.7030**	**0.6979**

cross-plane prediction, task-compatible translation, and continual evidence memory. Please refer to Supplementary Material A.6 for further discussion of detailed results.

4.3 Ablation Studies

We examine the main design choices of CP-JEPA from both representation and task perspectives. All ablation experiments are conducted on the Prostate T2 MRI benchmark. Fig. 4 first evaluates whether orthogonal views contain useful predictive information. Although the frozen coronal, sagittal, and axial representations are not inherently aligned, the cross-plane predictor recovers the structure of the target axial embedding and outperforms a capacity-matched axial-only predictor on 79.4% of held-out voxels. Table 2 then examines whether this predictive information improves segmentation. Part (a) compares direct use of the predicted representation with the complete task-compatible CP-JEPA correction under the NTD and TD protocols. Part (b) isolates the contribution of exact cross-plane correspondence under the NTD protocol.

Cross-Plane Predictability. Fig. 4(a) shows that frozen orthogonal-view features are not naturally aligned with the axial target space. After predictive training, the cross-plane predictions better follow

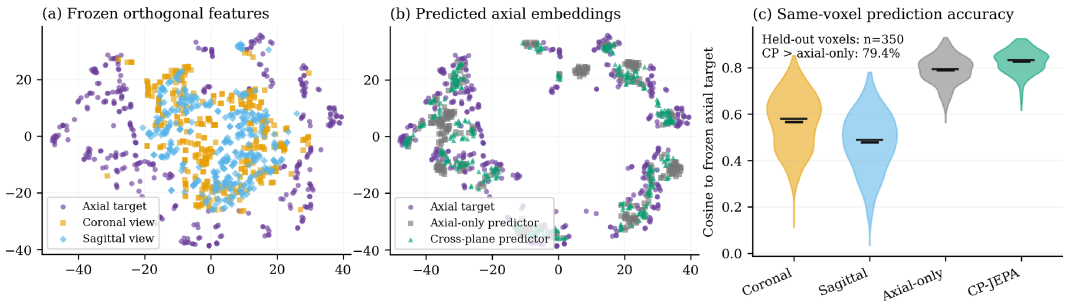


Figure 4: Cross-plane predictability of frozen features. (a) Orthogonal-view features, (b) predicted axial embeddings, and (c) same-voxel cosine similarity, where CP-JEPA outperforms the axial-only predictor on 79.4% of held-out voxels.

the axial target structure than the capacity-matched axial-only predictions, as shown in Fig. 4(b). The same-voxel analysis in Fig. 4(c) further shows that CP-JEPA achieves the highest cosine similarity and outperforms the axial-only predictor on 79.4% of 350 held-out voxels. This confirms that exact cross-plane correspondence provides predictive information beyond additional model capacity.

Task-Compatible Translation. Directly applying the predicted representation provides almost no segmentation benefit. Compared with the frozen Dice of 0.6829, *Direct* obtains 0.6824 under NTD and 0.6828 under TD. In contrast, task-compatible CP-JEPA improves over the frozen model by 1.50–4.48 Dice points across the four evaluation settings. Together with Fig. 4, these results demonstrate the **predictability–compatibility gap**: a representation can be accurately predicted from orthogonal views without being directly compatible with the source-trained segmentation model.

Cross-Plane Correspondence. The evidence controls in Table 2(b) show that the improvement originates from exact volumetric correspondence. Exact cross-plane evidence improves over the frozen baseline by 2.01 and 1.50 Dice points under single- and continual-volume deployment, respectively. In comparison, axial-only evidence yields gains of only 0.01 points, indicating that additional predictive capacity without orthogonal-view information is insufficient. Shuffled cross-plane correspondence similarly improves Dice by only 0.03 points in both deployment modes, even though the predictor still receives features from orthogonal views. The useful adaptation signal therefore arises from anatomically matched cross-plane context rather than merely introducing additional views.

Continual Evidence Memory. Under NTD, continual adaptation performs 0.51 Dice points below single-adaptation, exposing a limitation of the consistency-aware gate: historical evidence may agree with the current evidence without benefiting the current case. Under TD, this trend is reversed, with continual adaptation outperforming single-volume adaptation by 1.13 points. This contrast indicates that hyperparameter selection improves the selective accumulation of task-space evidence, motivating the separation between volume-specific predictive states and cross-volume memory.

5 Conclusion

We presented CP-JEPA, a TTA framework that exploits exact cross-plane correspondence within a single unlabeled medical volume. CP-JEPA predicts masked axial representations from orthogonal views and translates the resulting evidence into bounded, task-compatible corrections anchored to the source prediction. Across prostate, cardiac, and liver MRI segmentation, CP-JEPA achieves the best performance in three of four evaluation settings and ranks second in the remaining setting. Ablations show that these gains depend on exact anatomical correspondence and task-compatible translation, rather than additional views or predictor capacity alone. These results establish cross-plane predictive supervision as a useful signal for both single- and continual-volume adaptation. Future work will develop source-only reliability measures and less conservative memory gating to reduce reliance on labeled target data and improve continual adaptation.

AI USE STATEMENT

In this work, we used generative AI tools for drafting parts of the manuscript and editing the paper to improve readability and language. We have not used generative AI tools for generating synthetic datasets or interpreting results, and the other tasks with required disclosure are not applicable to this work. Additionally, we used generative AI tools for creating or modifying scientific figures or images. We have reviewed all AI-assisted work. Specifically, all AI-generated text was checked for factual accuracy and potential plagiarism through a manual literature survey; all AI-assisted figures were verified and corrected by the authors to ensure they accurately represent the underlying data and methodology; and all AI-generated code was verified and tested for correctness by the authors. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Lisa C Adams, Marcus R Makowski, Gunther Engel, Maximilian Rattunde, Felix Busch, Patrick Asbach, Stefan M Niehues, Shankeeth Vinayahalingam, Bram van Ginneken, Geert Litjens, and Keno K Bressem. Prostate158 - an expert-annotated 3t mri dataset and algorithm for prostate cancer detection. *Computers in Biology and Medicine*, 148:105817, 2022.

Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, et al. The medical segmentation decathlon. *Nature Communications*, 13(1):4128, 2022.

Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15619–15629, 2023.

Sunghyun Baek, Jaemyung Yu, Seunghee Koh, Minsu Kim, Hyeonseong Jeon, and Junmo Kim. Imse: Intrinsic mixture of spectral experts fine-tuning for test-time adaptation. In *International Conference on Learning Representations*, 2026.

Yoshua Bengio and Yann LeCun. Scaling learning algorithms towards AI. In *Large Scale Kernel Machines*. MIT Press, 2007.

Victor M Campello, Polyxeni Gkontra, Cristian Izquierdo, Carlos Martin-Isla, Alireza Sojoudi, Peter M Full, Klaus Maier-Hein, Yao Zhang, Zhiqiang He, Jun Ma, et al. Multi-centre, multi-vendor and multi-disease cardiac segmentation: The m&ms challenge. *IEEE Transactions on Medical Imaging*, 40(12):3543–3554, 2021.

Hyeongheon Cha, Dong Min Kim, Hye Won Chung, Taesik Gong, and Sung-Ju Lee. Snap: Low-latency test-time adaptation with sparse updates. In *Advances in Neural Information Processing Systems*, 2025.

Kecheng Chen, Xinyu Luo, Tiexin Qin, Jie Liu, Hui Liu, Victor Ho Fun Lee, Hong Yan, and Haoliang Li. *Test-time adaptation for foundation medical segmentation model without parametric updates*. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025a.

Kecheng Chen, Xinyu Luo, Tiexin Qin, Jie Liu, Hui Liu, Victor Ho Fun Lee, Hong Yan, and Haoliang Li. Test-time adaptation for foundation medical segmentation model without parametric updates. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025b.

Ziyang Chen, Yongsheng Pan, Yiwen Ye, Ming Lu, and Yong Xia. Each test image deserves a specific prompt: Continual test-time adaptation for 2d medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Ziyang Chen, Yiwen Ye, Yongsheng Pan, and Yong Xia. Gradient alignment improves test-time adaptation for medical image segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 2025c.

Renato Cuocolo, Albert Comelli, Alessandro Stefano, Viviana Benfante, Navdeep Dahiya, Arnaldo Stanzione, Anna Castaldo, Davide Raffaele De Lucia, Anthony Yezzi, and Massimo Imbriaco. Deep learning whole-gland and zonal prostate segmentation on a public mri dataset. *Journal of Magnetic Resonance Imaging*, 54(2):452–459, 2021.

Haoyu Dong, Nicholas Konz, Hanxue Gu, and Maciej A Mazurowski. Medical image segmentation with intent: Integrated entropy weighting for single image test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2024.

Xiaogang Du, Jiawei Zhang, Tongfei Liu, Tao Lei, and Yingbo Wang. Spegc: Continual test-time adaptation via semantic-prompt-enhanced graph clustering for medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Linghan Fang, Tianxin Xie, and Li Liu. Boosting asr robustness via test-time reinforcement learning with audio-text semantic rewards. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 30673–30681, 2026.

Andriy Fedorov, Mark G. Vangel, Clare M. Tempany, and Fiona M. Fennessy. Multiparametric magnetic resonance imaging of the prostate: Repeatability of volume and apparent diffusion coefficient quantification. In *Investigative Radiology*, volume 52, pp. 538–546. Lippincott Williams & Wilkins, 2017.

Md Meftahul Ferdaus, Tanmoy Dam, Md Rasel Sarkar, and Sreenatha S. Anavatti. Joint embedding predictive architecture for autonomous vehicle safety and security: A comprehensive survey. *Advanced Engineering Informatics*, 2026.

Yufei Gao, Zhenyang Li, Jing Xu, Qing Li, Ziyu Li, Yameng Zhang, and Lei Shi. Multi-domain and source-free domain adaptation in medical image analysis: A review. *Neurocomputing*, 2026.

Szymon Gibała, Rafał Obuchowicz, Julia Lasek, Zbigniew Schneider, Adam Piórkowski, Elżbieta Pociask, and Karolina Nurzyńska. Textural features of mr images correlate with an increased risk of clinically significant cancer in patients with high psa levels. *Journal of Clinical Medicine*, 12 (8):2836, 2023.

Ian Goodfellow, Yoshua Bengio, Aaron Courville, and Yoshua Bengio. *Deep learning*, volume 1. MIT Press, 2016.

Jisu Han, Jaemin Na, and Wonjun Hwang. Ranked entropy minimization for continual test-time adaptation. In *International Conference on Machine Learning*, 2025.

Xingxin He, Yifan Hu, Zhaoye Zhou, Mohamed Jarraya, and Fang Liu. Few-shot adaptation of training-free foundation model for 3d medical image segmentation. *arXiv preprint arXiv:2501.09138*, 2025.

Geoffrey E. Hinton, Simon Osindero, and Yee Whye Teh. A fast learning algorithm for deep belief nets. *Neural Computation*, 18:1527–1554, 2006.

Yihao Hu, Congyu Qiao, Xin Geng, and Ning Xu. Selective label enhancement learning for test-time adaptation. In *International Conference on Learning Representations*, 2025.

Hai Huang, Yann LeCun, and Randall Balestriero. Llm-jepa: Large language models meet joint embedding predictive architectures. *arXiv preprint arXiv:2509.14252*, 2025.

Smriti Joshi, Richard Osuala, Lidia Garrucho, Kaisar Kushibar, Dimitri Kessler, Oliver Diaz, and Karim Lekadir. *Single image test-time adaptation via multi-view co-training*. *arXiv preprint arXiv:2506.23705*, 2025.

A Emre Kavur, N Sinem Gezer, M Sinem Barış, Sinem Aslan, Pierre-Henri Conze, Vladimir Groza, Duc Duy Pham, Soumick Chatterjee, Philipp Ernst, Savaş Özkan, et al. Chaos challenge - combined (ct-mr) healthy abdominal organ segmentation. *Medical Image Analysis*, 69:101950, 2021.

Akhil Kondepudi, Akshay Rao, et al. Neurovfm: A visual foundation model for neuroimaging via volumetric joint-embedding predictive architecture. *arXiv preprint arXiv:2511.18640*, 2025.

Guannan Lai, Da-Wei Zhou, Zhenguo Li, and Han-Jia Ye. The golden subspace: Where efficiency meets generalization in continual test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Xiang Li, Huihui Fang, Mingsi Liu, Jinghao Wang, Lixin Duan, Yuqi Fang, and Yanwu Xu. Beyond visual prompt: Unlocking parameter-efficient online adapter for continual test-time adaptation in medical image segmentation. *Information Fusion*, 2026.

Geert Litjens, Robert Toth, Wendy van de Ven, Caroline Hoeks, Sjoerd Kerkstra, Bram van Ginneken, Graham Vincent, Gwenael Guillard, Neil Birbeck, Jindang Zhang, et al. Evaluation of prostate segmentation algorithms for mri: The promise12 challenge. volume 18, pp. 359–373, 2014.

Jun Ma, Zongxin Yang, Sumin Kim, Bihui Chen, Mohammed Baharoon, Adibvafa Fallahpour, Reza Asakereh, Hongwei Lyu, and Bo Wang. Medsam2: Segment anything in 3d medical images and videos. *arXiv preprint arXiv:2504.03600*, 2025.

Xiao Ma, Young D. Kwon, Pan Zhou, and Dong Ma. Architecture-agnostic test-time adaptation via backprop-free embedding alignment. In *International Conference on Learning Representations*, 2026.

Sarthak Kumar Maharana, Shambhavi Mishra, Yunbei Zhang, Shuaicheng Niu, Taki Hasan Rafi, Jihun Hamm, Marco Pedersoli, Jose Dolz, and Yunhui Guo. Continual test-time adaptation: A comprehensive survey. Zenodo, February 2026.

Alexander Murphy, Michal Danilowski, Soumyajit Chatterjee, and Abhirup Ghosh. Neo: No-optimization test-time adaptation through latent re-centering. *arXiv preprint arXiv:2510.05635*, 2025.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. In *International Conference on Learning Representations*, 2024.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pp. 234–241, 2015.

Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michael Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.

Guney Tombak, Ertunc Erdil, and Ender Konukoglu. *VoxCor: Training-free volumetric feature for multimodel voxel correspondance*. *arXiv preprint arXiv:2605.13798*, 2026.

Quoc-Huy Trinh, Minh-Van Nguyen, and Ulas Bagci. Rad-jepa 3d: Radiology joint-embedding predictive model for 3d computed tomography. *arXiv preprint arXiv:2607.26196*, 2026.

Silvia P. Vacacela and Marco E. Benalcázar. 2d semantic segmentation of the prostate gland in magnetic resonance images using convolutional neural networks. *Escuela Politécnica Nacional, Quito, Ecuador*, 2019.

Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. *Tent: Fully test-time adaptation by entropy minimization*. In *International Conference on Learning Representations (ICLR)*, 2021.

Jianghao Wu, Xinya Liu, Guotai Wang, and Shaoting Zhang. Test-time evaluation-guided dynamic adaptation for medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention*, 2025.

Jianghao Wu, Xinya Liu, Guotai Wang, and Shaoting Zhang. SicTTA: Single image continual test time adaptation for medical image segmentation. *Medical Image Analysis*, 108:103859, 2026a. ISSN 1361-8415.

Xiangyu Wu, Dongming Jiang, Feng Yu, Yueying Tian, Jiaqi Tang, Qing-Guo Chen, Yang Yang, and Jianfeng Lu. Adaptive debiasing tsallis entropy for test-time adaptation. In *International Conference on Learning Representations*, 2026b.

Hancun Yang, Hengjun Guo, Guanlin Liu, Linkuan Zhou, Junlin Xu, Leyi Wei, Ran Su, and Qiangguo Jin. A survey on unsupervised domain adaptation in medical imaging: Methods, datasets, and future outlook. *Applied Soft Computing*, 2026.

Jiabin Yu, Binggang Xiao, Jiayi Wang, Hongmei Mi, Enyu Wang, Ningjie Huang, You Zhou, Dongping Zhang, Furong Luo, Li Yang, Hanzong Lin, Ruigang Huang, Yingying Ding, Xia Li, Chenjie Xu, Guorong Lyu, and Ming Chen. Weakly supervised deep learning for multimodal mri-trus registration: Toward assisting prostate biopsy guidance. *Digital Health*, 2025.

Maxime Zanella, Clément Fuchs, and Ismail Ben Ayed. Realistic test-time adaptation of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Junming Zhang, Shuyu Yin, Peilin Liu, Rendong Ying, and Fei Wen. Curvature-aware zeroth-order optimization for memory-efficient test-time adaptation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2026.

Qingyang Zhang, Yatao Bian, Xinke Kong, Peilin Zhao, and Changqing Zhang. Come: Test-time adaptation by conservatively minimizing entropy. In *International Conference on Learning Representations*, 2025a.

Xiaoran Zhang, James Duncan, and Alex Wong. Progressive test-time energy adaptation for medical image segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025b.

Yubo Zhou, Jianghao Wu, Wenjun Liao, Shichuan Zhang, Shaoting Zhang, and Guotai Wang. Tegda: Test-time evaluation-guided dynamic adaptation for medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention*, 2025.

Jiayuan Zhu, Abdullah Hamdi, Yunli Qi, Yueming Jin, and Junde Wu. Medical sam 2: Segment medical images as video via segment anything model 2. *arXiv preprint arXiv:2408.00874*, 2024.

Jiaxin Zhuang, Linshan Wu, Xi Wang, and Hao Chen. Bio2vol: Adapting 2d biomedical foundation models for 3d medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention*, 2025.

A SUPPLEMENTARY MATERIAL

Supplementary Contents

A.1 Dataset inventory and patient-level partitions 14
A.2 Evaluation units, controls, and result provenance 15
A.3 Implementation of the predictive and task-space modules 16
A.4 Causal memory: diagram-to-code reconciliation 20
A.5 Comparison methods: original objectives and benchmark interfaces 21
A.6 Complete results for all comparison methods 22
A.7 Disaggregated results: seven prostate target cohorts 22
A.8 Disaggregated results: M&Ms B/C/D 23
A.9 Disaggregated results: six CHAOS contrast directions 25
A.10 Prostate ablations and the prediction–compatibility gap 28
A.11 Disaggregated comparison maps 29
A.12 Parameter count, operation count, and measured latency 30
A.13 Limitations and interpretation of the evidence 31

This supplement documents the patient-level splits, implementation, and provenance of the results in Tables 1 and 2, and the feature visualization in Fig. 4. The experiments serve two purposes. First, they test whether the method improves segmentation under a fixed source model. Second, they distinguish the contribution of exact cross-plane prediction from the effects of ordinary axial context, additional parameters, post-processing, and inter-volume state.

A.1 DATASET INVENTORY AND PATIENT-LEVEL PARTITIONS

Prostate T2 MRI. The labeled RegPro source contains 50 distinct patients/volumes: 46 source training and four source validation cases. The source training cache contains 92 entries because it includes augmented versions of the 46 training cases; these entries are not additional patients. The target benchmark contains seven external T2 cohorts, totaling 493 distinct patients and 508 volumes (Table 3). QIN-Prostate contributes two volumes per patient; for every other listed cohort the patient and volume counts coincide. The target-development (TD) and held-out partitions are patient-disjoint *within each target cohort* and kept fixed across methods. The binary target is the whole prostate, not the prostate subzones. Dice is computed per evaluated volume and macro-averaged equally over the seven target cohorts, so a large cohort does not dominate the headline average.

M&Ms cardiac MRI. Domain A provides the labeled source: 95 patients and 190 end-diastolic/end-systolic volumes. The task uses a 2D U-Net and four labels including background; the three foreground classes are left ventricle (LV), myocardium (MYO), and right ventricle (RV). Domains B, C, and D contribute 125/250, 75/150, and 50/100 patients/volumes, respectively. The development and evaluation partitions are patient-disjoint. The displayed Dice first averages the three foreground-class Dice values within a case, then averages cases within a domain, and finally averages B, C, and D equally. This differs from a single foreground-versus-background Dice score. The two cardiac phases from one person remain in the same partition.

CHAOS liver MRI. The three MRI sequences are T1 in-phase (IP), T1 opposed-phase (OP), and T2-SPIR (T2). Each sequence contains 20 volumes from the *same 20 participants*; six directed source-to-target contrasts are evaluated. The development and evaluation partitions contain disjoint participants. The source checkpoint for the current fixed-backbone CHAOS comparison is a sequence-specific 3D nnU-Net v2 trained for 250 epochs, not the 2D U-Net used for M&Ms. The task is binary liver segmentation. The repeated-subject design tests cross-contrast robustness; it cannot by itself establish cross-patient generalization. The six directions are IP→OP, IP→T2, OP→IP, OP→T2, T2→IP, and T2→OP.

Target-development use. NTD means that target labels are unavailable for hyperparameter selection; target volumes remain available without labels at inference. TD permits selection on labeled

Table 3: Total patient and 3D-volume inventory. The six CHAOS directions reuse the same participants under different MRI sequences; sequence counts must therefore not be added to infer the number of unique people.

Family	Cohort or sequence	Patients	Volumes
Prostate	RegPro source Yu et al. (2025)	50	50
	PROSTATEx Cuocolo et al. (2021)	204	204
	NCI-ISBI Vacacela & Benalcázar (2019)	59	59
	PROMISE12 Litjens et al. (2014)	50	50
	MSD-Prostate Antonelli et al. (2022)	32	32
	QIN-Prostate Fedorov et al. (2017)	15	30
	Polish Prostate Gibała et al. (2023)	114	114
	Prostate158 Adams et al. (2022)	19	19
Cardiac	M&Ms A source Campello et al. (2021)	95	190
	M&Ms B target Campello et al. (2021)	125	250
	M&Ms C target Campello et al. (2021)	75	150
	M&Ms D target Campello et al. (2021)	50	100
Liver	CHAOS IP Kavur et al. (2021)	20	20
	CHAOS OP Kavur et al. (2021)	20	20
	CHAOS T2 Kavur et al. (2021)	20	20

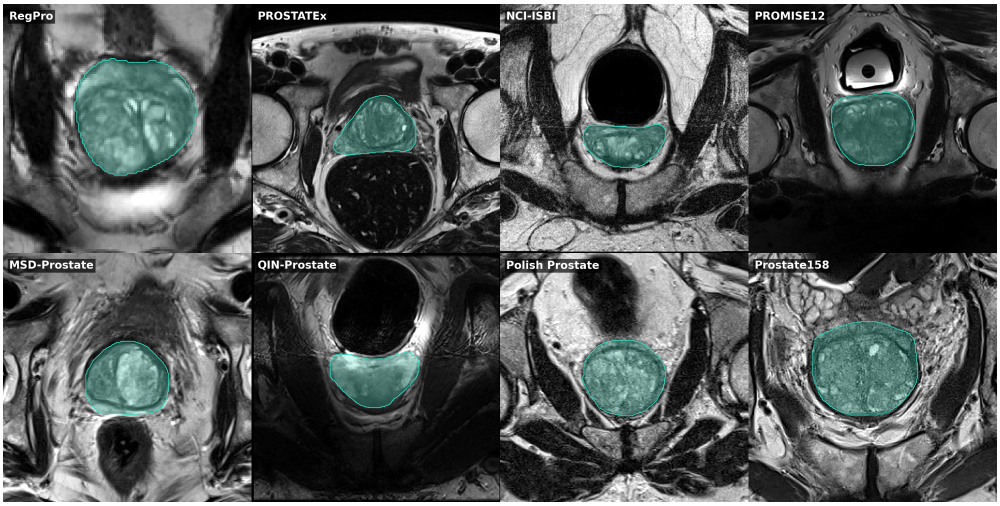


Figure 5: Representative T2 MRI appearances from the RegPro source and seven external prostate cohorts. The panels illustrate visible variation in field of view, intensity, anatomical scale, and acquisition appearance.

development patients, but the reported Dice is computed on disjoint held-out patients. TD is therefore a labeled selection protocol, not an annotation-free choice of hyperparameters. Because the target-development partitions are substantial, TD is more permissive than the NTD protocol. It should not be conflated with the single-volume deployment condition: single/continual describes which test-time state persists, whereas NTD/TD describes how hyperparameters were selected.

A.2 Evaluation units, controls, and result provenance

The source model is trained on labeled source data. In the strict single-volume condition, each test case gets fresh volume-specific predictors and no state from an earlier target case. Continual evaluation processes cases causally; a compact residual memory may pass from case $t-1$ to case t, but it is updated *after* the current prediction. The per-case predictors are not carried forward. Target labels are never used by either prediction procedure, including in TD; TD labels influence the choice of settings before held-out evaluation.

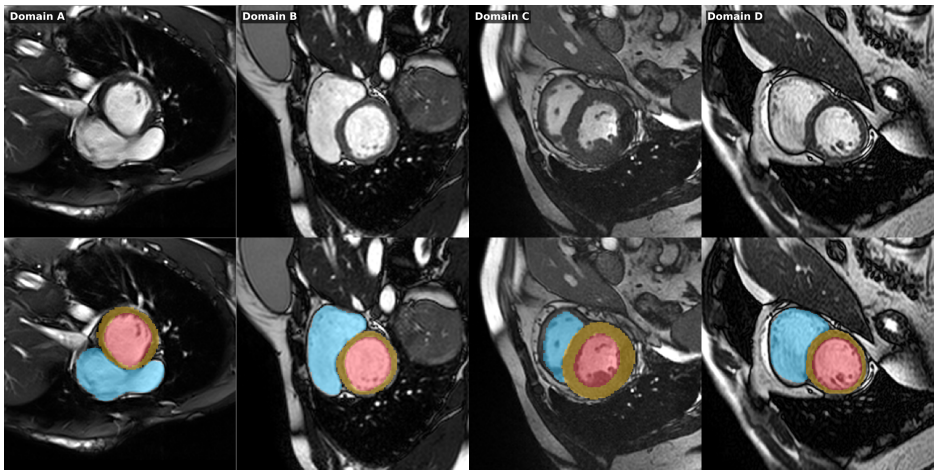


Figure 6: Representative M&Ms images from source Domain A and target Domains B/C/D. Evaluation uses the three separate LV, MYO, and RV masks, not a merged whole-heart foreground.

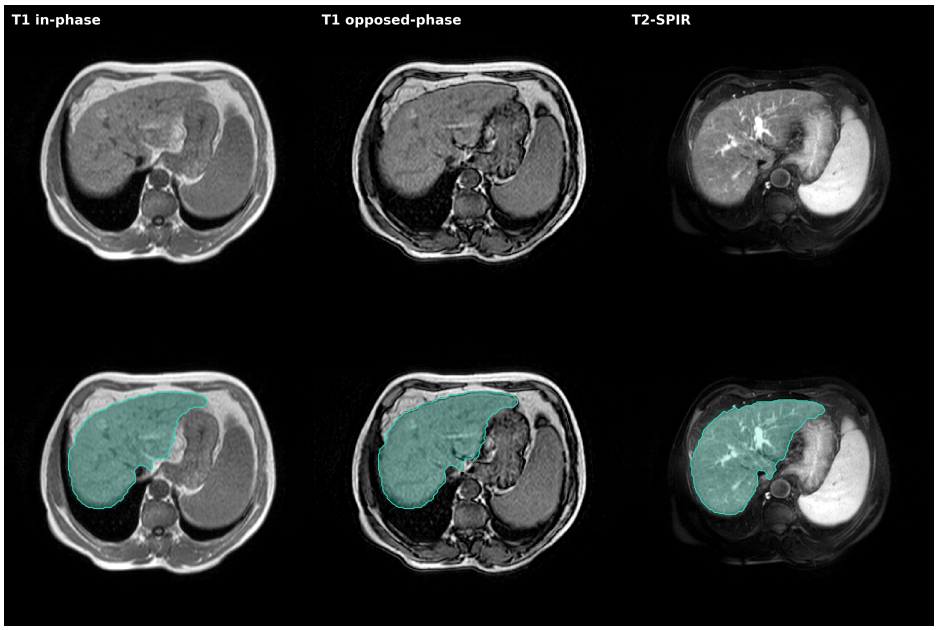


Figure 7: CHAOS T1 in-phase, T1 opposed-phase, and T2-SPIR MRI from the same participant (subject 03). Each lower panel overlays the liver reference on the slice with the largest annotated liver area in that sequence. This illustrates an acquisition-contrast shift, not three independent patients.

Matched comparison. In Table 1, the frozen row is the unadapted source prediction on the same held-out partition. On prostate dataset, the CP-JEPA task model uses frozen DINOv3 dense features with a RegPro-trained 3D U-Net path; M&Ms uses a source-A 2D U-Net; CHAOS uses the sequence-specific 3D nnU-Net checkpoints described above. The table's external method names denote implementations of their adaptation strategies in this benchmark, not wholesale replication of all architectures, preprocessing, and training pipelines from the original papers.

A.3 Implementation of the predictive and task-space modules

Algorithm 1 separates three kinds of parts that can easily be mistaken for one another: the fixed source segmentor, the per-volume predictors, and the optional cross-volume memory. The first

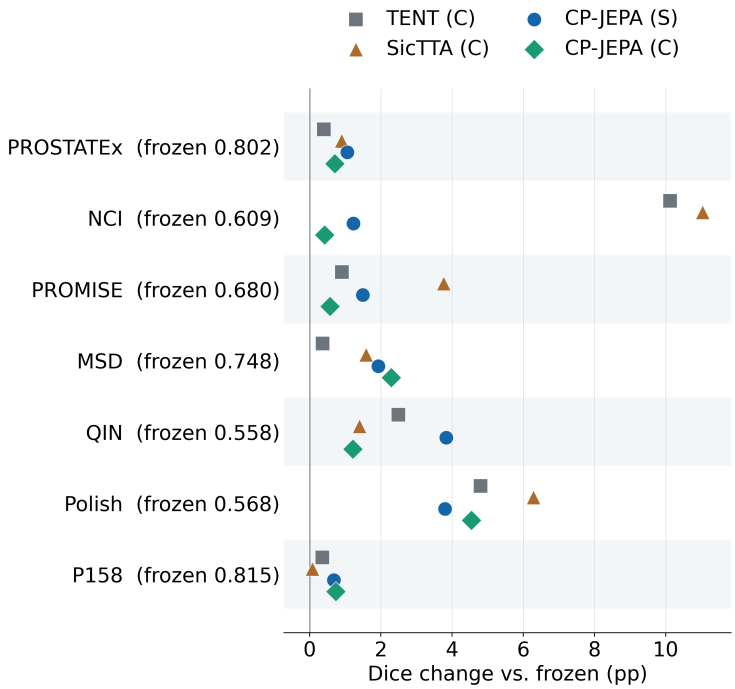


Figure 8: NTD prostate T2 result breakdown: one row per held-out target cohort. Each point is a target-level Dice change from the fixed frozen source model, in percentage points. TENT and SicTTA are shown in continual mode; CP-JEPA is shown in single (S) and continual (C) mode. The frozen Dice for each cohort is printed beside its name.

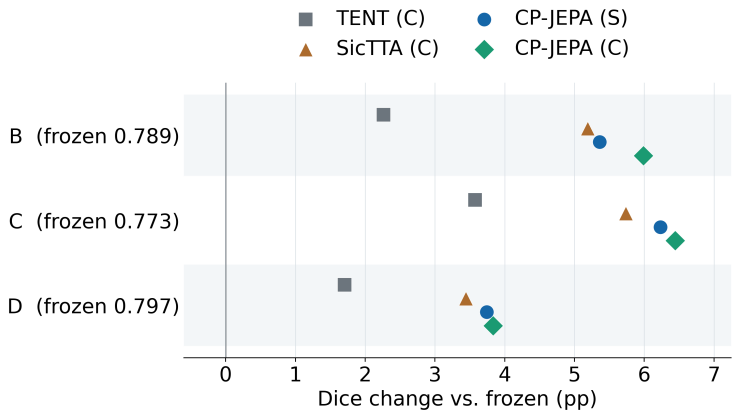


Figure 9: NTD M&Ms cardiac MRI result breakdown, with Domains B, C, and D as separate rows. The metric averages LV, MYO, and RV per patient before domain averaging; points show Dice-point changes from each domain's frozen source result.

provides a source prediction and bottleneck. The second is fitted using only the current unlabeled volume and discarded immediately afterwards. The third contains residual task-space evidence from earlier cases, rather than axial images, segmentation masks, or feature-bank examples.

Cross-plane prediction and its matched control. Features of the same physical voxel are lifted from the axial, coronal, and sagittal planes into one coordinate system. A masked axial embedding serves as a stop-gradient target. The coronal and sagittal predictors estimate it from the two orthogonal observations; a third predictor with the same predictor family estimates it from masked

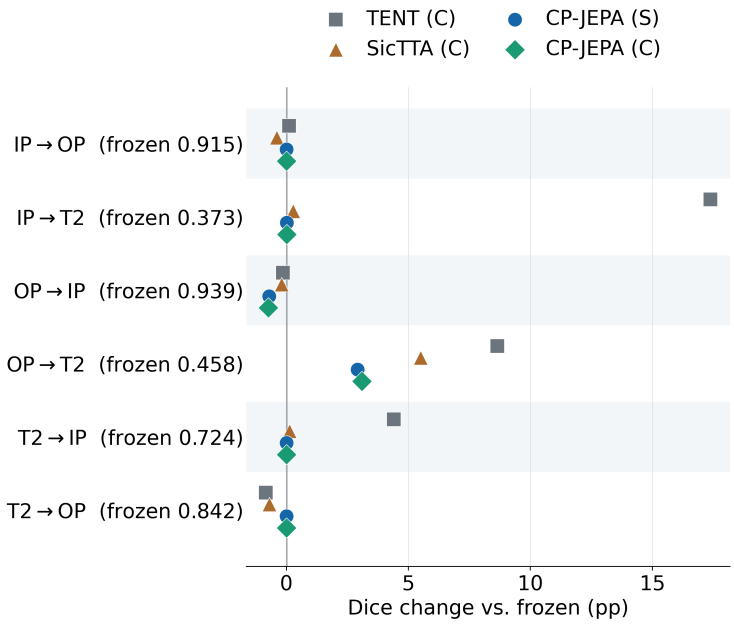


Figure 10: NTD CHAOS liver MRI result breakdown over all six directed sequence transfers. CP-JEPA uses the NTD configuration. The six rows reuse participants across directions.

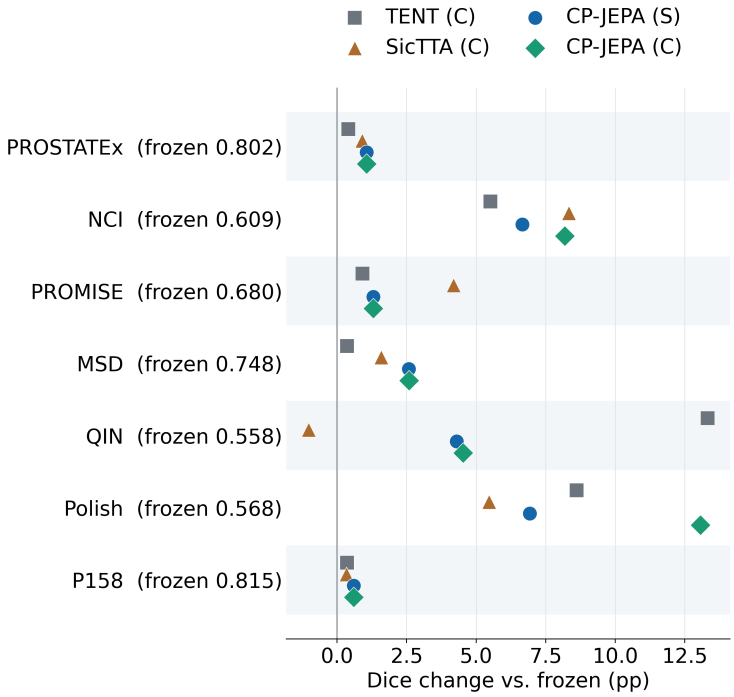


Figure 11: TD prostate T2 breakdown on the seven held-out cohorts. Hyperparameters were selected on disjoint labeled development patients; points are measured on held-out patients.

axial context. This axial branch is a control for prediction that does not require exact orthogonal correspondence. Consequently, the evidence used by the task adapter is the exact cross-plane prediction *minus the axial prediction*, modulated by positive cross-plane predictive advantage. The source encoder's weights are not fitted by this objective. Calling the procedure "JEPA-style" refers

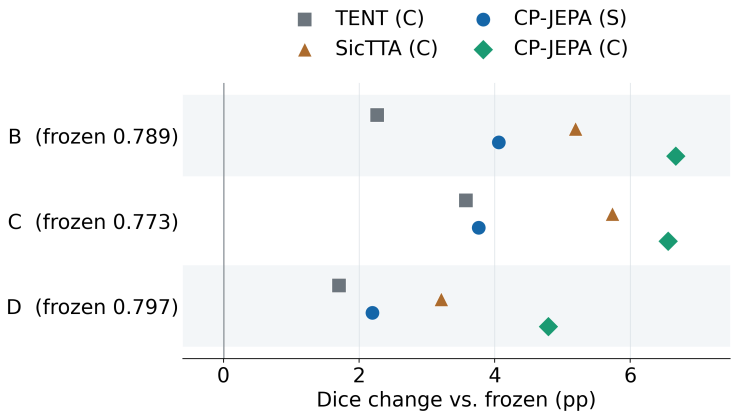


Figure 12: TD M&Ms cardiac MRI breakdown over Domains B/C/D using the same fixed source-A checkpoint. The plotted Dice-point change is relative to each domain's own frozen score.

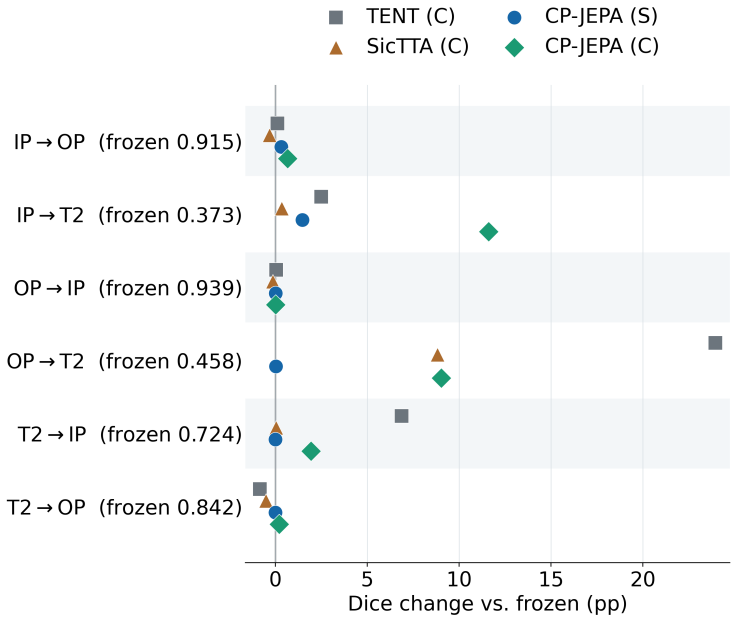


Figure 13: TD CHAOS liver MRI breakdown over all six transfers.

specifically to masked prediction in a latent space with a detached target. It does not imply training an EMA target encoder or replacing the source representation with the predicted one. Equation 2 specifies the fitting objective, and the predictor reset is Algorithm 1's final per-volume operation.

Task-compatible adapter. Fig. 14 expands the translation step in Algorithm 1. The input is a stack of evidence maps from multiple frozen feature layers. For each layer, the implemented module computes a root-mean-square scale over channels and voxels; one $1\times1\times1$ convolution encodes normalized direction, while another encodes the normalized evidence magnitude. The latter is clipped at 10. These per-layer channels are concatenated, joined with a projected frozen task bottleneck, and processed by two $3\times3\times3$ convolution–GroupNorm–GELU blocks. After resizing to the source logit grid, a further task-fusion block receives the softmax of source logits and, in configurations that enable it, three predictive-reliability maps.

Two heads produce a scalar acceptance gate and a class-wise logit update. The update passes through $\tanh$ and is multiplied by a maximum update amplitude; the acceptance gate passes through a sig-

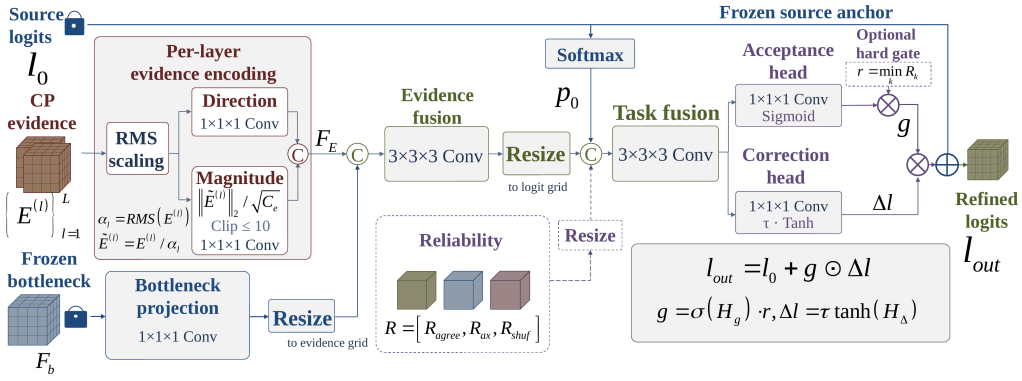


Figure 14: Task-compatible adapter as supplied in the module diagram and checked against the implementation. The three predictive-reliability channels correspond to coronal/sagittal agreement, exact-over-axial advantage, and exact-over-shuffled advantage. Their hard intersection is an optional configuration; the always-present safeguards are the bounded update and the learned acceptance head. Frozen source logits remain an additive anchor.

moid. The output is the original source logits plus their product. The correction-head weights and bias are initialized to zero, and the gate starts from a negative bias. Thus the initial prediction is exactly the source prediction even though the gate itself is nonzero. The adapter is trained once using source annotations; its parameters are fixed during target inference. This is a *learned* task-space translation, not an unsupervised gradient step on target segmentation labels.

What the reliability channels mean. The first channel is high agreement between the two independent orthogonal predictors. The second records where exact correspondence predicts the axial target better than the axial-context control. The third records where it beats a shuffled-correspondence control. Each is computed from the current unlabeled volume, so none requires target segmentation labels.

Four evidence arms during source training. The adapter is exposed to Exact, Axial, Shuffled, and Zero branches. Exact uses the two correctly aligned orthogonal predictions; Axial uses only masked axial context; Shuffled breaks the orthogonal voxel correspondence; Zero supplies no predictive residual. All four pass through the same adapter weights. The source segmentation term teaches the Exact branch to correct the task output; identity and ranking terms penalize a correction that is equally easy to obtain from the controls. This design makes a later Exact-versus-Shuffled comparison more interpretable than a comparison of two separately trained decoders.

A.4 Causal memory: diagram-to-code reconciliation

Fig. 15 depicts the continual module. The sequence arrows are causal: the current case reads h_{t-1}, predicts, and only then writes its own centered class-logit residual. For a new stream, the memory is empty and the first output equals the single-volume proposal. In single-volume mode, resetting memory before every case yields the same property for every output. Different volume shapes are handled by interpolation between the current logit grid and the configured memory grid.

The read operation compares the current and historical centered class-logit residuals with a voxel-wise cosine similarity clamped to $[0, 1]$. A geometric mean of current and historical support, multiplied by current prediction confidence, further limits the memory gate. The historical residual is scaled by this gate and a memory-strength hyperparameter before being added to the single-volume logits. This mechanism is not a generic moving average of predictions: an earlier correction with an opposing class direction is suppressed even if its magnitude is large.

Relation to Algorithm 1. Lines that construct the three plane features and fit disposable predictors implement the latent-prediction stage. The subsequent positive-advantage comparison selects

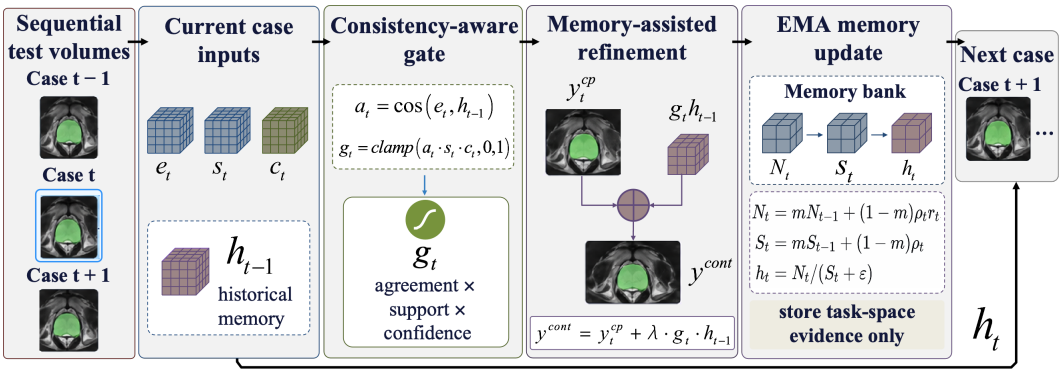


Figure 15: Supplied continual-adaptation diagram, reproduced to explain the state transition. Memory is read before the current case is written, and only task-space evidence is retained.

Table 4: Catalog of methods appearing in Table 1. Mechanism descriptions summarize the cited papers' main ideas.

Method	Original venue	Main adaptation principle
TENT Wang et al. (2021)	ICLR 2021	Entropy minimization; normalization-affine updates
COME Zhang et al. (2025a)	ICLR 2025	Conservative entropy with uncertainty modeling
TEGDA Zhou et al. (2025)	MICCAI 2025	Quality-guided feature bank and teacher adaptation
StatA Zanella et al. (2025)	CVPR 2025	Statistics-based vision–language feature adaptation
ADTE Wu et al. (2026b)	ICLR 2026	Adaptive debiased Tsallis entropy
ASR Fang et al. (2026)	AAAI 2026	Test-time reinforcement/semantic reward idea
CAZO Zhang et al. (2026)	CVPR 2026	Curvature-aware zeroth-order optimization
GOLD Lai et al. (2026)	CVPR 2026	Restricted golden-subspace adaptation
GraTa Chen et al. (2025c)	AAAI 2025	Gradient alignment
IMSE Baek et al. (2026)	ICLR 2026	Spectral expert updates and diversity
PASLE Hu et al. (2025)	ICLR 2025	Candidate pseudo-label sets for uncertain samples
PEA Ma et al. (2026)	ICLR 2026	Backprop-free embedding alignment
PTTEA Zhang et al. (2025b)	ICCV 2025	Progressive test-time energy adaptation
REM Han et al. (2025)	ICML 2025	Ranked entropy minimization
SNAP Cha et al. (2025)	NeurIPS 2025	Sparse low-latency updates
SPEGC Du et al. (2026)	CVPR 2026	Semantic prompts and graph clustering
TACT Chen et al. (2025b)	ICCV 2025	Embedding-level adaptation without parameter update
SicTTA Wu et al. (2026a)	MedIA 2026	Source-friendly retrieval and batch enhancement
CP-JEPA	This work	Cross-plane prediction and task-compatible translation

evidence attributable to orthogonal correspondence. The adapter line converts that evidence into a bounded logit proposal. The continual conditional reads the historical residual before the final update. In practice, the safety decision is exposed by the experiment configuration.

A.5 Comparison methods: original objectives and benchmark interfaces

Table 4 records the publication venue and original adaptation idea. This distinction is important because the original methods were developed for different tasks and, in several cases, different model families. Method names in Table 1 therefore refer to same-benchmark strategy ports unless a row is explicitly identified as a full author-pipeline reproduction.

Why the baselines matters. TENT, COME, and REM primarily impose confidence or entropy objectives; TEGDA and SicTTA maintain quality- or source-friendliness-aware examples; PASLE uses uncertain candidate pseudo-labels. These mechanisms may update the task model itself or use memories with information different from CP-JEPA's centered logit residual. A higher final Dice can reflect a better source checkpoint, stronger use of target labels for selection, or a more suitable native interface, in addition to a better adaptation rule. We therefore use the frozen score and protocol labels

Table 5: Target-wise held-out Dice for Prostate T2, NTD selection and single-volume deployment. The last column is the unweighted mean across the displayed targets. Rows use the same fixed source checkpoint within the family.

Method	PROSTATEx	NCI	PROMISE	MSD	QIN	Polish	P158	Macro
Source only	0.8023	0.6091	0.6798	0.7482	0.5575	0.5682	0.8154	0.6829
TENT	0.8023	0.6091	0.6798	0.7482	0.5575	0.5682	0.8154	0.6829
COME	0.8023	0.6091	0.6798	0.7482	0.5575	0.5682	0.8154	0.6829
TEGDA	0.8035	0.6085	0.6816	0.7506	0.5599	0.5706	0.8155	0.6843
STATA	0.6605	0.5112	0.5113	0.6077	0.4047	0.3581	0.6406	0.5277
ADTE	0.8073	0.6528	0.6847	0.7545	0.5594	0.5818	0.8222	0.6947
ASR	0.8023	0.6091	0.6798	0.7482	0.5575	0.5682	0.8154	0.6829
CAZO	0.8023	0.6091	0.6798	0.7482	0.5575	0.5682	0.8154	0.6829
GOLD	0.8023	0.6091	0.6798	0.7482	0.5575	0.5682	0.8154	0.6829
GraTa	0.8023	0.6091	0.6798	0.7482	0.5577	0.5683	0.8154	0.6830
IMSE	0.8023	0.6091	0.6798	0.7482	0.5575	0.5682	0.8154	0.6829
PASLE	0.8023	0.6091	0.6798	0.7482	0.5575	0.5682	0.8154	0.6829
PEA	0.7923	0.6010	0.6729	0.7452	0.5414	0.5363	0.8075	0.6710
PTTEA	0.6596	0.5356	0.4532	0.5621	0.4622	0.3720	0.6054	0.5214
REM	0.8023	0.6091	0.6798	0.7482	0.5575	0.5682	0.8154	0.6829
SNAP	0.8023	0.6091	0.6798	0.7482	0.5575	0.5682	0.8154	0.6829
SPEGC	0.8023	0.6091	0.6798	0.7482	0.5575	0.5682	0.8154	0.6829
TACT	0.8044	0.5843	0.7197	0.7766	0.4880	0.6610	0.8109	0.6921
SicTTA	0.7854	0.5920	0.6768	0.7105	0.4059	0.5979	0.8183	0.6553
CP-JEPA	0.8128	0.6214	0.6947	0.7674	0.5959	0.6063	0.8222	0.7030

in the main table, and regard the strategy ports as controlled *mechanism comparisons*, not identical author-repository reproductions.

Observed behavior across families. Under TD continual evaluation, CP-JEPA obtains 0.7277 on prostate and 0.8462 on M&Ms; the strongest listed alternative in those columns is TENT at 0.7250 and SicTTA at 0.8333, respectively. These margins are 0.27 and 1.29 Dice points using the displayed scores. CHAOS differs: TENT reaches 0.7627, above CP-JEPA's exploratory 0.7475, while the frozen source is 0.7084.

A.6 Complete results for all comparison methods

Tables 5–16 disaggregate the complete fixed-backbone comparison into twelve protocol-by-deployment tables: four for each dataset family. Each target cohort or direction has its own column, followed by an equal-target macro. These are held-out results.

A.7 Disaggregated results: seven prostate target cohorts

Tables 7 and 8 give the target-wise values behind the TD prostate macro in Table 1. The source-only row is repeated for S and C because no adaptation state exists in that condition.

The CP-JEPA TD single-volume macro exceeds the frozen macro by about 3.34 Dice points, while the continual macro exceeds it by about 4.47 points. These are *within-row* changes, not a comparison to a different backbone. The largest displayed continual gains appear on Polish (+13.04 points relative to .5684), QIN (+4.53), and NCI (+8.19), whereas PROSTATEx and P158 change by only about one point or less. This heterogeneity matters: the method's average effect is not a uniform shift of all cohorts. On PROMISE12, the displayed S and C values coincide; retaining memory offers no visible additional benefit. On Polish, in contrast, C exceeds S by 6.13 points.

Dataset-size sensitivity. QIN is the smallest target cohort and contributes multiple volumes per patient. Its patient-level estimates are therefore less stable than those of the larger cohorts. The reported equal-cohort macro avoids allowing PROSTATEx, the largest cohort, to dominate a pooled-volume average.

Table 6: Target-wise held-out Dice for Prostate T2, NTD selection and continual-volume deployment. The last column is the unweighted mean across the displayed targets. Rows use the same fixed source checkpoint within the family.

Method	PROSTATEx	NCI	PROMISE	MSD	QIN	Polish	P158	Macro
Source only	0.8023	0.6091	0.6798	0.7482	0.5575	0.5682	0.8154	0.6829
TENT	0.8063	0.7103	0.6889	0.7518	0.5825	0.6163	0.8190	0.7107
COME	0.7392	0.5968	0.6685	0.7419	0.5595	0.4892	0.8169	0.6589
TEGDA	0.8008	0.6353	0.6785	0.7470	0.5572	0.5667	0.8150	0.6858
STATA	0.6605	0.5112	0.5113	0.6077	0.4047	0.3581	0.6406	0.5277
ADTE	0.8073	0.6528	0.6847	0.7545	0.5594	0.5818	0.8222	0.6947
ASR	0.8023	0.6091	0.6798	0.7482	0.5575	0.5682	0.8154	0.6829
CAZO	0.8031	0.6647	0.6837	0.7523	0.5603	0.5700	0.8159	0.6929
GOLD	0.8023	0.6091	0.6798	0.7482	0.5575	0.5682	0.8154	0.6829
GraTa	0.8028	0.6634	0.6800	0.7483	0.5584	0.5693	0.8155	0.6911
IMSE	0.5486	0.5733	0.5761	0.7141	0.4296	0.3505	0.7822	0.5677
PASLE	0.7114	0.6586	0.6329	0.7653	0.5847	0.4443	0.7767	0.6534
PEA	0.7999	0.6686	0.6820	0.7573	0.5787	0.5414	0.8207	0.6927
PTTEA	0.1338	0.1358	0.1209	0.2070	0.1408	0.0868	0.1903	0.1451
REM	0.8054	0.6649	0.6802	0.7481	0.5586	0.5541	0.8155	0.6896
SNAP	0.8023	0.6091	0.6798	0.7482	0.5575	0.5682	0.8154	0.6829
SPEGC	0.7690	0.6074	0.6716	0.7476	0.4705	0.4241	0.8074	0.6425
TACT	0.8009	0.4666	0.6989	0.7720	0.4557	0.6415	0.7936	0.6613
SicTTA	0.7873	0.6954	0.6935	0.7400	0.5485	0.6071	0.7923	0.6949
CP-JEPA	0.8094	0.6133	0.6855	0.7711	0.5697	0.6137	0.8228	0.6979

Table 7: Target-wise held-out Dice for Prostate T2, TD selection and single-volume deployment. The last column is the unweighted mean across the displayed targets. Rows use the same fixed source checkpoint within the family.

Method	PROSTATEx	NCI	PROMISE	MSD	QIN	Polish	P158	Macro
Source only	0.8023	0.6091	0.6798	0.7482	0.5575	0.5682	0.8154	0.6829
TENT	0.8023	0.6091	0.6798	0.7482	0.5575	0.5682	0.8154	0.6829
COME	0.8023	0.6091	0.6798	0.7482	0.5575	0.5682	0.8154	0.6829
TEGDA	0.8116	0.6092	0.6800	0.7419	0.5476	0.5893	0.8154	0.6850
STATA	0.6720	0.5258	0.5319	0.6235	0.4174	0.3702	0.6575	0.5426
ADTE	0.8091	0.6533	0.6855	0.7549	0.5600	0.5843	0.8233	0.6958
ASR	0.8023	0.6091	0.6798	0.7482	0.5575	0.5682	0.8154	0.6829
CAZO	0.8023	0.6091	0.6798	0.7482	0.5576	0.5683	0.8154	0.6829
GOLD	0.8023	0.6091	0.6798	0.7482	0.5575	0.5682	0.8154	0.6829
GraTa	0.8025	0.6366	0.6801	0.7484	0.5592	0.5689	0.8155	0.6873
IMSE	0.8023	0.6091	0.6798	0.7482	0.5575	0.5682	0.8154	0.6829
PASLE	0.8023	0.6091	0.6798	0.7482	0.5575	0.5682	0.8154	0.6829
PEA	0.7983	0.6083	0.6729	0.7452	0.5414	0.5534	0.8075	0.6753
PTTEA	0.7883	0.6630	0.6641	0.7292	0.5389	0.5393	0.8008	0.6748
REM	0.8023	0.6091	0.6798	0.7482	0.5575	0.5682	0.8154	0.6829
SNAP	0.8023	0.6091	0.6798	0.7482	0.5575	0.5682	0.8154	0.6829
SPEGC	0.8023	0.6091	0.6798	0.7482	0.5575	0.5682	0.8154	0.6829
TACT	0.8044	0.5843	0.7197	0.7766	0.4880	0.6634	0.7866	0.6890
SicTTA	0.7780	0.5920	0.6769	0.7098	0.4059	0.6123	0.8183	0.6562
CP-JEPA	0.8128	0.6756	0.6928	0.7739	0.6005	0.6375	0.8214	0.7164

A.8 DISAGGREGATED RESULTS: M&MS B/C/D

Tables 11 and 12 expand the three-domain macro. The frozen source-A model has Dice .7890, .7729, and .7966 on B, C, and D, respectively. CP-JEPA's TD continual values are .8557, .8384, and .8445, improvements of 6.67, 6.55, and 4.79 Dice points against the same frozen model. The single-volume variant also improves each domain, but C adds a further 2.79 points when causal memory is

Table 8: Target-wise held-out Dice for Prostate T2, TD selection and continual-volume deployment. The last column is the unweighted mean across the displayed targets. Rows use the same fixed source checkpoint within the family;.

Method	PROSTATEx	NCI	PROMISE	MSD	QIN	Polish	P158	Macro
Source only	0.8023	0.6091	0.6798	0.7482	0.5575	0.5682	0.8154	0.6829
TENT	0.8063	0.6642	0.6889	0.7518	0.6908	0.6543	0.8190	0.7250
COME	0.7986	0.6054	0.6788	0.7476	0.5282	0.5601	0.8156	0.6763
TEGDA	0.8026	0.6094	0.6802	0.7460	0.5540	0.5653	0.8157	0.6819
STATA	0.6720	0.5258	0.5319	0.6235	0.4174	0.3702	0.6575	0.5426
ADTE	0.8091	0.6533	0.6855	0.7549	0.5600	0.5843	0.8233	0.6958
ASR	0.8023	0.6091	0.6798	0.7482	0.5576	0.5682	0.8154	0.6829
CAZO	0.8031	0.6649	0.6839	0.7524	0.5604	0.5707	0.8159	0.6930
GOLD	0.8023	0.6091	0.6798	0.7482	0.5575	0.5682	0.8154	0.6829
GraTa	0.8071	0.6669	0.6821	0.7497	0.5681	0.5795	0.8158	0.6956
IMSE	0.7649	0.6125	0.6738	0.7465	0.5404	0.5249	0.8128	0.6680
PASLE	0.7996	0.6724	0.6837	0.7605	0.5876	0.5332	0.8134	0.6929
PEA	0.7999	0.6641	0.6802	0.7573	0.5787	0.5564	0.8207	0.6939
PTTEA	0.2694	0.3688	0.3592	0.5857	0.4494	0.2037	0.7152	0.4216
REM	0.8020	0.6363	0.6800	0.7480	0.5586	0.5116	0.8165	0.6790
SNAP	0.8023	0.6091	0.6798	0.7482	0.5576	0.5684	0.8154	0.6830
SPEGC	0.8099	0.6491	0.6794	0.7484	0.4705	0.6079	0.8127	0.6826
TACT	0.8009	0.4666	0.6989	0.7720	0.4557	0.6716	0.7652	0.6615
SicTTA	0.8113	0.6924	0.7217	0.7640	0.5473	0.6229	0.8189	0.7112
CP-JEPA	0.8128	0.6910	0.6928	0.7741	0.6029	0.6988	0.8214	0.7277

Table 9: Target-wise held-out Dice for M&Ms B/C/D, NTD selection and single-volume deployment. The last column is the unweighted mean across the displayed targets. Rows use the same fixed source checkpoint within the family.

Method	B	C	D	Macro
Source only	0.7890	0.7729	0.7966	0.7861
TENT	0.8116	0.8088	0.8137	0.8114
COME	0.8222	0.8164	0.8128	0.8171
TEGDA	0.8218	0.8164	0.8131	0.8171
STATA	0.8132	0.7970	0.8122	0.8075
ADTE	0.7868	0.7869	0.7932	0.7890
ASR	0.8229	0.8162	0.8133	0.8174
CAZO	0.7889	0.7729	0.7966	0.7861
GOLD	0.7890	0.7729	0.7966	0.7861
GraTa	0.8229	0.8162	0.8133	0.8174
IMSE	0.8222	0.8175	0.8177	0.8191
PASLE	0.8196	0.8265	0.8165	0.8208
PEA	0.7960	0.7936	0.8014	0.7970
PTTEA	0.7263	0.7750	0.7690	0.7568
REM	0.8228	0.8169	0.8134	0.8177
SNAP	0.8342	0.8278	0.8212	0.8278
SPEGC	0.6839	0.6780	0.6948	0.6856
TACT	0.7776	0.7497	0.7906	0.7726
SicTTA	0.8269	0.8212	0.8236	0.8239
CP-JEPA	0.8426	0.8352	0.8340	0.8373

retained. SNAP is a strong matched comparator: its .8278 B/C/D TD single macro exceeds several other methods, yet remains 1.83 points below CP-JEPA continual in the displayed TD setting. These Dice values average LV, MYO, and RV.

Table 10: Target-wise held-out Dice for M&Ms B/C/D, NTD selection and continual-volume deployment. The last column is the unweighted mean across the displayed targets. Rows use the same fixed source checkpoint within the family.

Method	B	C	D	Macro
Source only	0.7890	0.7729	0.7966	0.7861
TENT	0.8116	0.8086	0.8136	0.8113
COME	0.3859	0.5049	0.6219	0.5042
TEGDA	0.8246	0.8189	0.8135	0.8190
STATA	0.8132	0.7970	0.8122	0.8075
ADTE	0.7189	0.7076	0.7377	0.7214
ASR	0.8229	0.8162	0.8133	0.8174
CAZO	0.7861	0.7715	0.7944	0.7840
GOLD	0.7890	0.7729	0.7966	0.7862
GraTa	0.8225	0.8162	0.8132	0.8173
IMSE	0.1180	0.2232	0.4279	0.2564
PASLE	0.4334	0.6033	0.6664	0.5677
PEA	0.8163	0.8052	0.8163	0.8126
PTTEA	0.0211	0.0300	0.0561	0.0357
REM	0.8237	0.8169	0.8124	0.8177
SNAP	0.8343	0.8277	0.8215	0.8278
SPEGC	0.0057	0.0414	0.0691	0.0387
TACT	0.7794	0.7493	0.7906	0.7731
SicTTA	0.8409	0.8302	0.8310	0.8340
CP-JEPA	0.8488	0.8373	0.8349	0.8404

Table 11: Target-wise held-out Dice for M&Ms B/C/D, TD selection and single-volume deployment. The last column is the unweighted mean across the displayed targets. Rows use the same fixed source checkpoint within the family.

Method	B	C	D	Macro
Source only	0.7890	0.7729	0.7966	0.7861
TENT	0.8116	0.8088	0.8136	0.8113
COME	0.8229	0.8164	0.8133	0.8175
TEGDA	0.8226	0.8164	0.8131	0.8174
STATA	0.8133	0.7970	0.8122	0.8075
ADTE	0.7891	0.7876	0.7939	0.7902
ASR	0.8229	0.8162	0.8133	0.8174
CAZO	0.7890	0.7729	0.7966	0.7861
GOLD	0.7890	0.7729	0.7966	0.7861
GraTa	0.8229	0.8162	0.8133	0.8175
IMSE	0.8222	0.8175	0.8177	0.8191
PASLE	0.8197	0.8253	0.8165	0.8205
PEA	0.8099	0.7905	0.8120	0.8041
PTTEA	0.8236	0.8202	0.8204	0.8214
REM	0.8228	0.8169	0.8133	0.8176
SNAP	0.8342	0.8278	0.8212	0.8278
SPEGC	0.7863	0.7689	0.7908	0.7820
TACT	0.7879	0.7647	0.7964	0.7830
SicTTA	0.8170	0.8113	0.8236	0.8173
CP-JEPA	0.8295	0.8105	0.8185	0.8195

A.9 Disaggregated results: six CHAOS contrast directions

Tables 15 and 16 show why a single CHAOS macro conceals a difficult asymmetric transfer. Frozen IP→OP and OP→IP already score .9151 and .9387; the opposing T1 contrasts are relatively easy for these source checkpoints. IP→T2 and OP→T2 score only .3730 and .4578. The CP-JEPA TD continual gains on these two directions are 11.60 and 9.03 Dice points, but the aggregate remains

Table 12: Target-wise held-out Dice for M&Ms B/C/D, TD selection and continual-volume deployment. The last column is the unweighted mean across the displayed targets. Rows use the same fixed source checkpoint within the family.

Method	B	C	D	Macro
Source only	0.7890	0.7729	0.7966	0.7861
TENT	0.8116	0.8086	0.8136	0.8113
COME	0.8087	0.8117	0.8124	0.8109
TEGDA	0.8272	0.8191	0.8156	0.8206
STATA	0.8133	0.7970	0.8122	0.8075
ADTE	0.7434	0.7310	0.7536	0.7427
ASR	0.8229	0.8162	0.8133	0.8174
CAZO	0.7866	0.7719	0.7947	0.7844
GOLD	0.8126	0.7843	0.8089	0.8019
GraTa	0.8229	0.8162	0.8132	0.8174
IMSE	0.5082	0.7944	0.7757	0.6928
PASLE	0.8053	0.8233	0.8230	0.8172
PEA	0.8170	0.8052	0.8223	0.8148
PTTEA	0.0787	0.1180	0.1729	0.1232
REM	0.8232	0.8169	0.8132	0.8178
SNAP	0.8343	0.8277	0.8215	0.8278
SPEGC	0.0057	0.2839	0.5839	0.2912
TACT	0.7882	0.7625	0.7962	0.7823
SicTTA	0.8409	0.8302	0.8287	0.8333
CP-JEPA	0.8557	0.8384	0.8445	0.8462

Table 13: Target-wise held-out Dice for CHAOS, NTD selection and single-volume deployment. The last column is the unweighted mean across the displayed targets. Rows use the same fixed source checkpoint within the family.

Method	IP→OP	IP→T2	OP→IP	OP→T2	T2→IP	T2→OP	Macro
Source only	0.9151	0.3729	0.9387	0.4578	0.7241	0.8417	0.7084
TENT	0.9004	0.3553	0.9392	0.5266	0.7240	0.8426	0.7147
COME	0.8995	0.3572	0.9356	0.4967	0.7223	0.8416	0.7088
TEGDA	0.9152	0.3732	0.9387	0.4579	0.7241	0.8417	0.7085
STATA	0.9245	0.3864	0.9292	0.4939	0.7005	0.8330	0.7113
ADTE	0.9147	0.3739	0.9386	0.4598	0.7285	0.8424	0.7097
ASR	0.9151	0.3730	0.9387	0.4578	0.7241	0.8417	0.7084
CAZO	0.9151	0.3732	0.9387	0.4578	0.7242	0.8417	0.7084
GOLD	0.9151	0.3729	0.9386	0.4578	0.7241	0.8417	0.7084
GraTa	0.9154	0.3745	0.9387	0.4574	0.7246	0.8419	0.7088
IMSE	0.9152	0.3730	0.9386	0.4578	0.7241	0.8417	0.7084
PASLE	0.9144	0.3744	0.9387	0.4585	0.7241	0.8419	0.7087
PEA	0.9129	0.3356	0.9393	0.4602	0.7249	0.8345	0.7012
PTTEA	0.3765	0.2322	0.6835	0.3296	0.7531	0.7573	0.5220
REM	0.9159	0.3561	0.9387	0.4617	0.7241	0.8416	0.7063
SNAP	0.9151	0.3730	0.9387	0.4578	0.7241	0.8417	0.7084
SPEGC	0.9119	0.3432	0.9386	0.4579	0.7239	0.8423	0.7030
TACT	0.8397	0.0908	0.9356	0.4602	0.2609	0.7159	0.5505
SicTTA	0.9138	0.3687	0.9381	0.5190	0.7205	0.8397	0.7166
CP-JEPA	0.9152	0.3731	0.9317	0.4685	0.7241	0.8417	0.7091

below the TENT strategy port (.7627 versus .7475 in Table 1). On T2→IP and T2→OP, the gains are much smaller. The six displayed per-direction CP-JEPA values average to .7475 when rounded to four decimals.

Fig. 16 places the three disaggregated TD results on a common Dice-point scale. The plotted gains are relative to each target's own frozen score, using the same source checkpoint within each dataset

Table 14: Target-wise held-out Dice for CHAOS, NTD selection and continual-volume deployment. The last column is the unweighted mean across the displayed targets. Rows use the same fixed source checkpoint within the family.

Method	IP→OP	IP→T2	OP→IP	OP→T2	T2→IP	T2→OP	Macro
Source only	0.9151	0.3729	0.9387	0.4578	0.7241	0.8417	0.7084
TENT	0.9162	0.5468	0.9372	0.5442	0.7681	0.8333	0.7576
COME	0.9002	0.5307	0.8959	0.5958	0.7060	0.8256	0.7424
TEGDA	0.9152	0.3734	0.9387	0.4580	0.7243	0.8417	0.7086
STATA	0.9245	0.3864	0.9292	0.4939	0.7005	0.8330	0.7113
ADTE	0.9147	0.3739	0.9386	0.4598	0.7285	0.8424	0.7097
ASR	0.9151	0.3730	0.9387	0.4578	0.7241	0.8417	0.7084
CAZO	0.9151	0.3738	0.9387	0.4578	0.7241	0.8417	0.7085
GOLD	0.9151	0.3729	0.9387	0.4578	0.7241	0.8417	0.7084
GraTa	0.9158	0.3798	0.9386	0.4590	0.7251	0.8415	0.7100
IMSE	0.9154	0.3733	0.9386	0.4579	0.7239	0.8417	0.7085
PASLE	0.9147	0.3840	0.9389	0.4581	0.7227	0.8420	0.7101
PEA	0.9156	0.2976	0.9392	0.4603	0.7125	0.8169	0.6904
PTTEA	0.0594	0.0664	0.1509	0.1749	0.2423	0.1805	0.1458
REM	0.9194	0.4040	0.9375	0.6000	0.7222	0.8392	0.7371
SNAP	0.9151	0.3729	0.9387	0.4578	0.7241	0.8417	0.7084
SPEGC	0.9062	0.1459	0.9398	0.4453	0.7199	0.8356	0.6654
TACT	0.8393	0.0849	0.9354	0.4602	0.0826	0.7017	0.5174
SicTTA	0.9112	0.3757	0.9367	0.5128	0.7254	0.8348	0.7161
CP-JEPA	0.9152	0.3731	0.9312	0.4887	0.7241	0.8417	0.7123

Table 15: Target-wise held-out Dice for CHAOS, TD selection and single-volume deployment. The last column is the unweighted mean across the displayed targets. Rows use the same fixed source checkpoint within the family.

Method	IP→OP	IP→T2	OP→IP	OP→T2	T2→IP	T2→OP	Macro
Source only	0.9151	0.3729	0.9387	0.4578	0.7241	0.8417	0.7084
TENT	0.8824	0.3553	0.9392	0.4586	0.7240	0.8433	0.7005
COME	0.8399	0.3572	0.9384	0.4996	0.7261	0.8348	0.6993
TEGDA	0.9152	0.3744	0.9387	0.4578	0.7245	0.8419	0.7088
STATA	0.9245	0.3764	0.9292	0.4839	0.7012	0.8334	0.7081
ADTE	0.9147	0.3731	0.9387	0.4580	0.7293	0.8431	0.7095
ASR	0.9151	0.3734	0.9387	0.4578	0.7241	0.8417	0.7085
CAZO	0.9151	0.3732	0.9387	0.4578	0.7242	0.8417	0.7084
GOLD	0.9151	0.3729	0.9386	0.4578	0.7241	0.8417	0.7084
GraTa	0.9176	0.3577	0.9387	0.4578	0.7257	0.8431	0.7068
IMSE	0.9154	0.3732	0.9387	0.4578	0.7241	0.8417	0.7085
PASLE	0.9048	0.3553	0.9390	0.4579	0.7241	0.8424	0.7039
PEA	0.9147	0.3356	0.9392	0.4602	0.7249	0.8445	0.7032
PTTEA	0.8386	0.3597	0.9195	0.5143	0.7531	0.8550	0.7067
REM	0.9159	0.3561	0.9386	0.5814	0.7241	0.8395	0.7259
SNAP	0.9151	0.3730	0.9387	0.4578	0.7241	0.8417	0.7084
SPEGC	0.8930	0.3723	0.9386	0.4578	0.7241	0.8418	0.7046
TACT	0.8397	0.0908	0.9351	0.4602	0.2575	0.7159	0.5499
SicTTA	0.9146	0.3694	0.9381	0.5189	0.7205	0.8397	0.7069
CP-JEPA	0.9233	0.3876	0.9387	0.4980	0.7241	0.8419	0.7189

family. Continual memory adds most on the Polish prostate cohort, M&Ms Domains B and C, and the CHAOS T1-to-T2 directions; several other targets change little beyond the single-volume result.

Table 16: Target-wise held-out Dice for CHAOS, TD selection and continual-volume deployment. The last column is the unweighted mean across the displayed targets. Rows use the same fixed source checkpoint within the family.

Method	IP→OP	IP→T2	OP→IP	OP→T2	T2→IP	T2→OP	Macro
Source only	0.9151	0.3729	0.9387	0.4578	0.7241	0.8417	0.7084
TENT	0.9162	0.3978	0.9390	0.6372	0.7545	0.8333	0.7463
COME	0.9002	0.4707	0.9378	0.5864	0.7232	0.8256	0.7407
TEGDA	0.9152	0.3735	0.9388	0.4579	0.7243	0.8418	0.7086
STATA	0.9241	0.3864	0.9292	0.4939	0.7012	0.8334	0.7114
ADTE	0.9147	0.3731	0.9387	0.4580	0.7293	0.8431	0.7095
ASR	0.9151	0.3735	0.9387	0.4576	0.7241	0.8417	0.7084
CAZO	0.9151	0.3738	0.9387	0.4578	0.7241	0.8417	0.7085
GOLD	0.9151	0.3729	0.9387	0.4578	0.7241	0.8417	0.7084
GraTa	0.9205	0.3829	0.9387	0.5808	0.7428	0.8403	0.7343
IMSE	0.9169	0.3752	0.9386	0.4585	0.7241	0.8417	0.7092
PASLE	0.9151	0.3739	0.9387	0.4579	0.7227	0.8455	0.7090
PEA	0.9153	0.3448	0.9392	0.4603	0.7254	0.8385	0.7039
PTTEA	0.2029	0.1439	0.4139	0.3228	0.2423	0.5750	0.3168
REM	0.9160	0.5058	0.9385	0.5832	0.7222	0.8392	0.7508
SNAP	0.9151	0.3729	0.9387	0.4578	0.7241	0.8417	0.7084
SPEGC	0.9154	0.2611	0.9390	0.5415	0.7199	0.8407	0.7029
TACT	0.8393	0.0849	0.9349	0.4602	0.0827	0.7017	0.5173
SicTTA	0.9119	0.3763	0.9372	0.5460	0.7245	0.8365	0.7221
CP-JEPA	0.9217	0.4890	0.9387	0.5481	0.7435	0.8438	0.7475

Table 17: Paired latent-prediction audit underlying Fig. 4. At each sampled voxel of `ProstateX-0072`, the indicated normalized representation is compared by cosine similarity with the *same voxel's* frozen axial DINOv3 embedding. Higher is closer to the held-out axial target. The final row counts voxelwise wins by exact cross-plane prediction over the matched axial-context predictor. These are representation measures, not Dice.

Representation	Mean cosine	Median cosine
Frozen coronal view	.5667	.5803
Frozen sagittal view	.4785	.4889
Axial-context predictor	.7896	.7946
Exact cross-plane predictor	.8272	.8329
Exact better than axial	278/350 sampled voxels (79.4%)	

A.10 Prostate ablations and the prediction–compatibility gap

Table 2 contains two complementary controls. Part (a) holds the source model fixed and compares direct use of the predictive latent with task-compatible translation. Under NTD, Direct is .6824, 0.05 Dice points below the .6829 frozen macro, whereas CP-JEPA reaches .7030 in S and .6979 in C. Under TD, Direct is .6828, approximately unchanged from frozen; the translated correction reaches .7164 in S and .7277 in C. Thus improved latent prediction does not produce a useful task output by simple substitution. The decoder interface determines whether predictive information helps segmentation.

Part (b) changes the evidence source while keeping the adapter family. In the NTD prostate macro, axial-only evidence gives .6830 and shuffled correspondence gives .6832 in both displayed modes, within 0.03 Dice points of frozen. Exact correspondence raises S to .7030 (+1.98 points over shuffled) and C to .6979 (+1.47 points over shuffled). The comparison supports a contribution from anatomically aligned orthogonal observations in this setting.

Figure 4 audit. Fig. 4 uses one fixed PROSTATEx case (`ProstateX-0072`). Of 2,996 eligible held-out voxels, 350 are sampled once and shared by all four rows; no method receives a different set of locations. The frozen coronal and sagittal rows measure how closely unadapted orthogonal

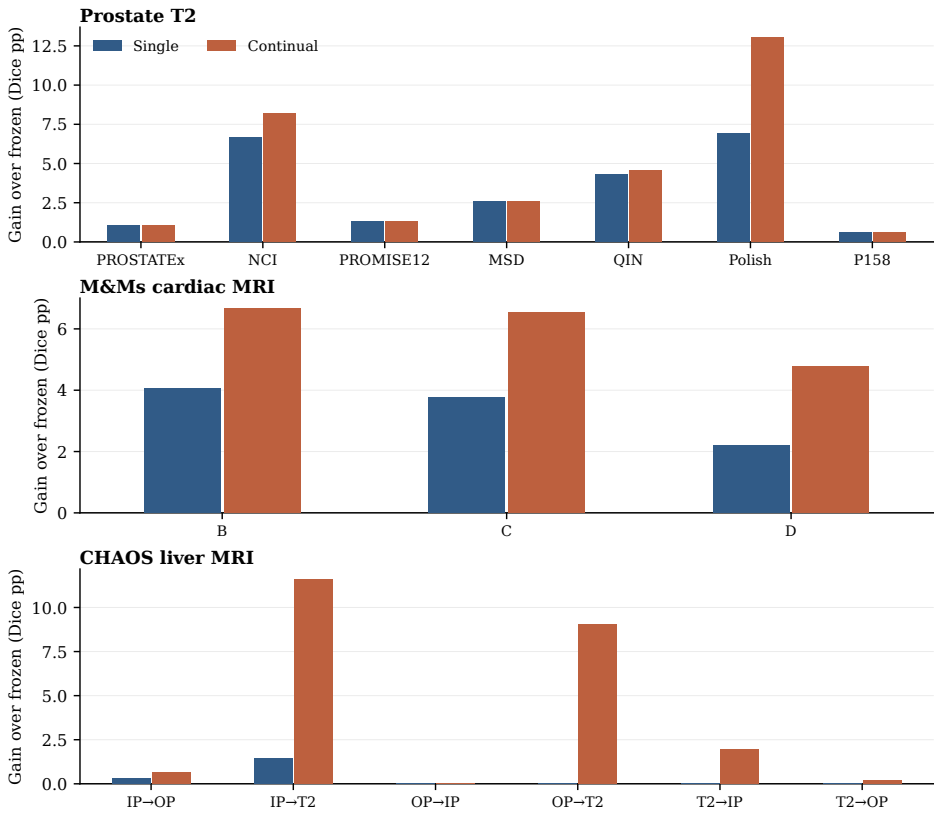


Figure 16: Target-wise CP-JEPA Dice gains over the frozen source model under TD selection. Bars are computed from the displayed rounded values in Tables 8, 12, and 16. Single and continual use the same source checkpoint within each family. The cases and settings were selected under the TD protocol.

observations match the axial target. The axial-context row controls for learning to predict that target without orthogonal correspondence. The exact cross-plane row uses anatomically matched coronal and sagittal observations. Relative to the axial control, its mean cosine rises from .7896 to .8272 (+.0376), its median rises from .7946 to .8329 (+.0383), and it is better at 278/350 locations (79.4%). These paired quantities, rather than visual cluster separation, are the evidence for improved latent prediction. The display first projects features to 50 PCA dimensions, then jointly embeds them by cosine-distance t-SNE (perplexity 35, 1,200 iterations, fixed seed).

Why each control is needed. Shuffling breaks voxel correspondence while preserving the availability of orthogonal-view features. Axial-only replaces orthogonal information with same-axis context. Zero evidence tests whether the adapter changes the output without predictive information. Direct correction instead tests the decoder interface by bypassing task-space translation. Exact-versus-Shuffled identifies an effect of geometry; Translated- versus-Direct identifies the need to convert that information into the source decoder's task space. Neither comparison alone establishes the full causal story.

A.11 Disaggregated comparison maps

Figs. 17– 22 replace the earlier single-case illustrations with complete target-wise comparison maps. Each cell is held-out Dice for the indicated target cohort or transfer direction. The heatmaps show a subset of representative comparison mechanisms for legibility, while the twelve tables in Section A.6 retain every evaluated method. Source-only and adapted rows use the same fixed checkpoint within a dataset family. The two CP-JEPA rows distinguish single from continual deployment.

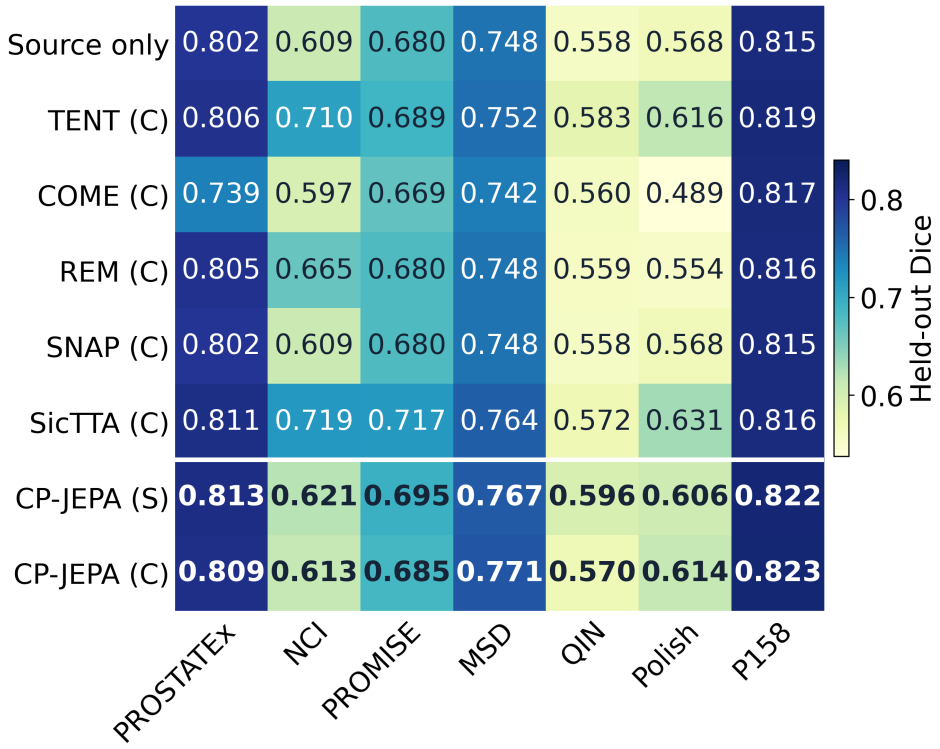


Figure 17: NTD prostate T2 target-wise Dice for source only, five continual-mode comparisons, and CP-JEPA in both deployment modes. Seven external cohorts are shown separately.

A.12 Parameter count, operation count, and measured latency

Trainable parameters at deployment must be distinguished from total parameters loaded for inference. The prostate source backbone contains the frozen DINOv3 encoder and the source-trained segmentation network; neither is updated by the volume-specific prediction loss. The task-compatible adapter is trained on source labels and likewise remains fixed during target inference. For the audited prostate setting, one pure low-rank predictor with feature width 256, rank 128, hidden width 256, and six positional coordinates has 165,248 parameters. The deployed coronal, sagittal, and axial-control predictors therefore contain 495,744 test-time optimizable parameters in total. An earlier diagnostic implementation, which fitted extra residual and shuffled control branches as well, had 1,914,628 predictor parameters. The default task adapter has 111,059 parameters without external reliability channels and 112,355 with the three-channel reliability input; these are *source-trained*, not target-updated. The per-volume predictors are discarded after each patient, so the test-time parameter count is an adaptation footprint rather than permanent model growth.

The matrix multiplications of one pure predictor require $256 \cdot 128 + 6 \cdot 128 + (2 \cdot 128) \cdot 256 + 256 \cdot 256 = 164{,}608$ multiply–accumulates per queried voxel. Counting one multiply and one addition as two operations yields 329,216 arithmetic FLOPs per predictor forward or 987,648 FLOPs for the three deployed predictors per voxel, *excluding* nonlinearities, feature normalization, source DINOv3 extraction, task-network inference, adapter convolutions, backpropagation, optimizer updates, and memory interpolation. It is consequently a local operation count, not an end-to-end per-volume FLOPs result. The number of voxel queries and optimization steps changes the total cost, as do input spacing and volume size. A single universal end-to-end FLOPs number for the three different task backbones would obscure rather than clarify those costs.

The cached-feature measurement in Table 18 shows a parameter–runtime tradeoff rather than a blanket efficiency win. The compact CP-JEPA predictor updates 29.2% fewer parameters than the TENT, COME, and MuVi ports, and 64.0% fewer than the IGTTA port. It also uses less peak allocated GPU memory than all four alternatives in this run. However, its measured latency is roughly 0.19–0.31

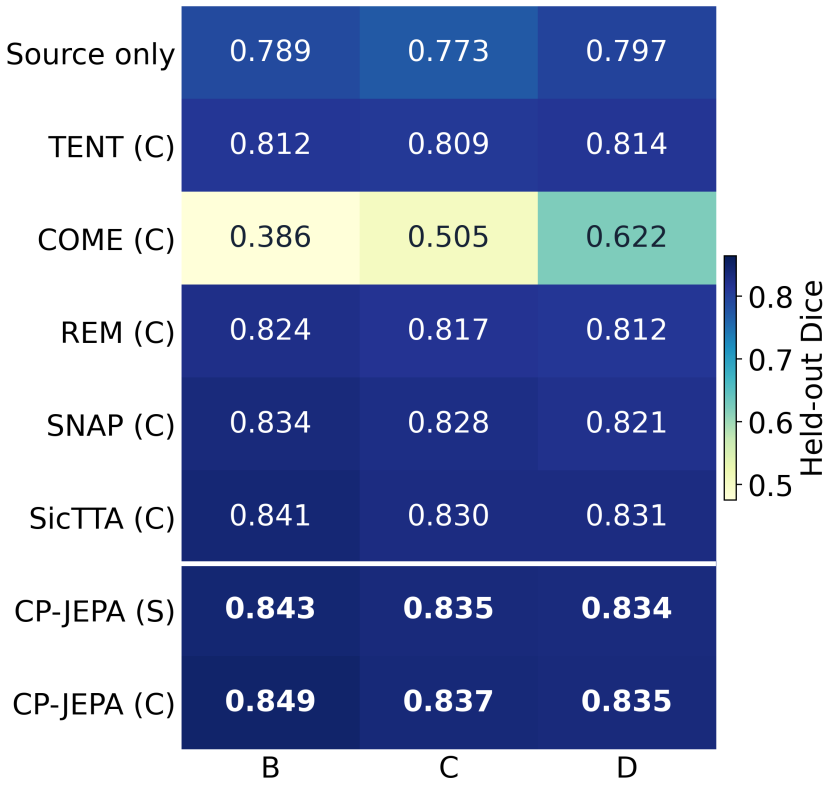


Figure 18: NTD M&Ms B/C/D held-out Dice. Each domain cell averages three foreground classes and patient-level ED/ES volumes under the fixed source-A checkpoint.

Table 18: Measured *cached-feature* efficiency on the same ten NCI volumes, on RTX 4090 GPU with CUDA synchronization. Each row was run as a separate process; values are mean ± standard deviation over nine timed volumes. MuVi and IGTTA rows are same-backbone objective ports. Feature extraction by DINOv3 is excluded.

Method	Updated parameters	s/volume	Peak MiB
COME port	700,258	$.651 \pm .049$	1408.4 ± 5.0
TENT port	700,258	$.644 \pm .033$	1410.7 ± 4.5
MuVi port	700,258	$.652 \pm .034$	1424.3 ± 5.0
IGTTA port	1,378,578	$.534 \pm .037$	991.3 ± 5.3
CP-JEPA compact	495,744	$.840 \pm .067$	815.9 ± 23.2

seconds longer per volume. Thus CP-JEPA is *not* the fastest method here, and a smaller trainable footprint does not imply fewer test-time operations: it fits three predictors for each volume. The comparison excludes the frozen encoder, uses only ten volumes, and does not match every target family or complete test-time computation path.

A.13 Limitations and interpretation of the evidence

CHAOS is not a uniform gain. The CHAOS TD single-volume average is 0.7114 versus 0.7084 frozen, a difference of only 0.30 Dice points. While TENT reaches 0.7627, and the best direction-level improvements concentrate on T1→T2 transfer. The data support a conditional cross-contrast benefit, not a broad claim of superiority on CHAOS. The full six-direction NTD panel in Fig. 10 shows that those limited aggregate gains coexist with direction-level failures.

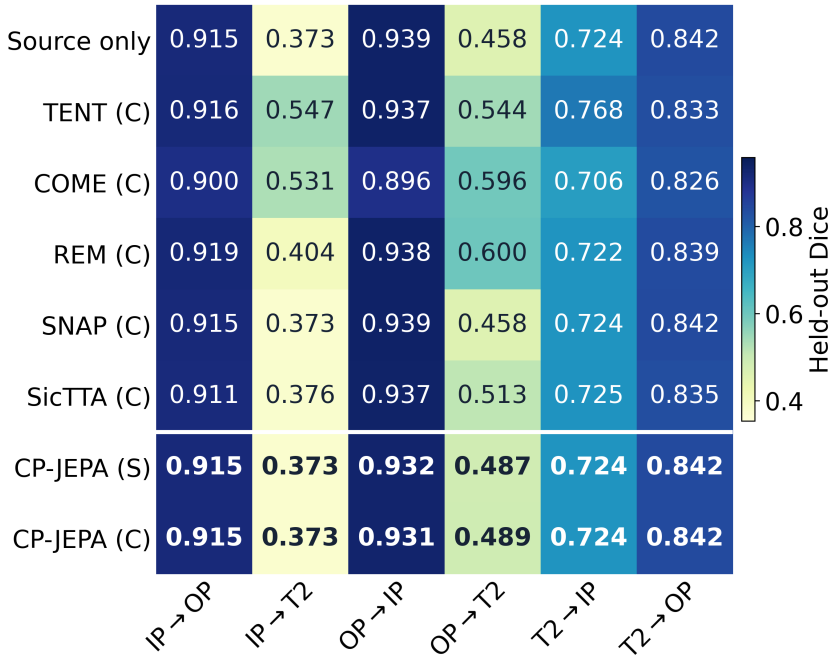


Figure 19: NTD CHAOS Dice on all six sequence-transfer directions.

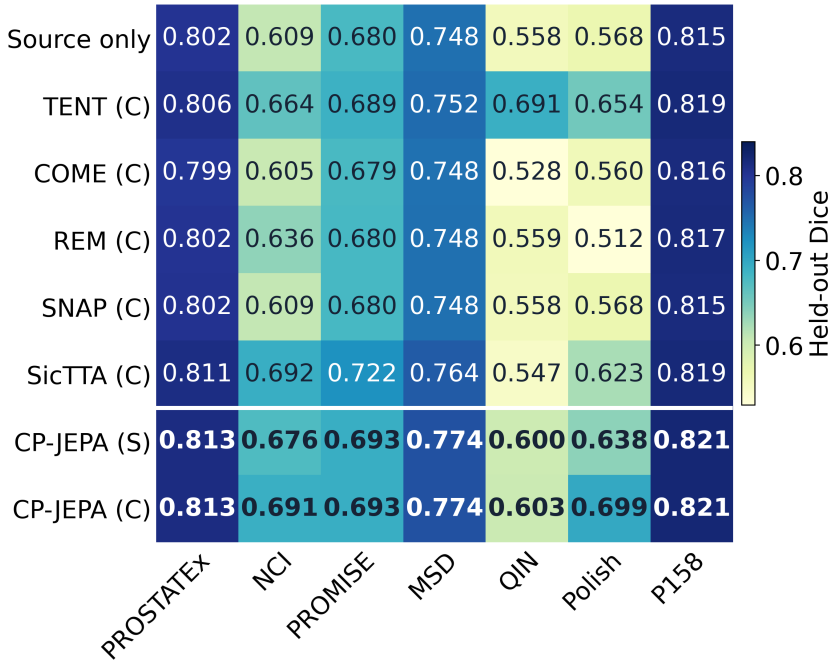


Figure 20: TD prostate T2 target-wise Dice after development-set hyperparameter selection, measured only on disjoint held-out patients.

Target labels and exploratory search. The TD partitions contain a labeled development set in each dataset family. These labels are not used for test-time gradient updates or per-case predictions, but are used to choose hyperparameters. This is a much more permissive selection setting than an entirely label-free deployment. Those rows estimate what the investigated implementation achieved under exploration, not an untouched independent test.

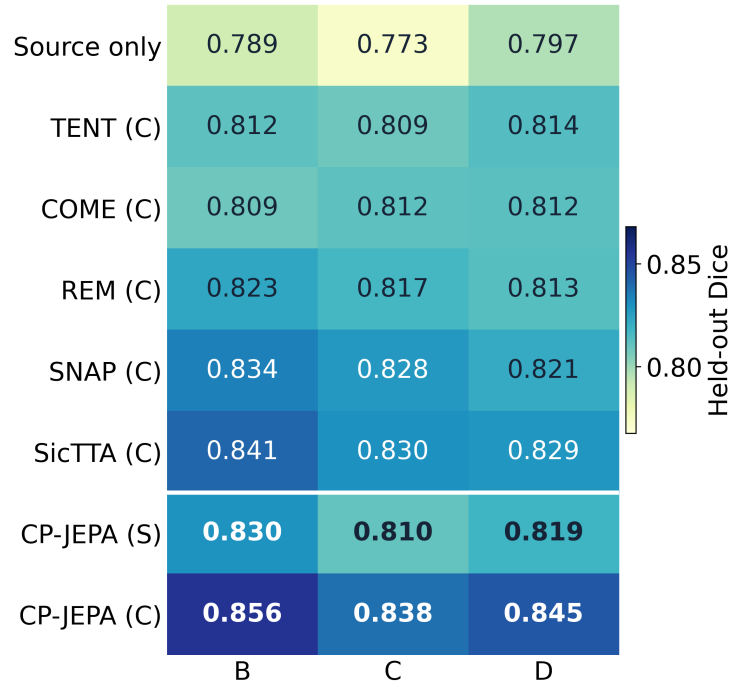


Figure 21: TD M&Ms B/C/D held-out three-class Dice. The distinct response of CP-JEPA single and continual modes is visible in Domains B and C.

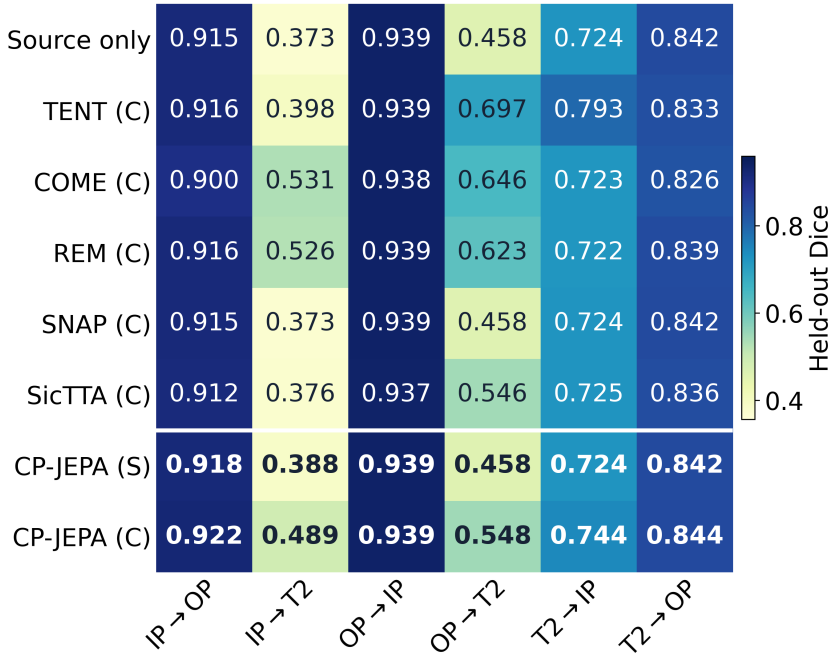


Figure 22: TD CHAOS Dice across all six contrast transfers.