Multi-Agent Test-Time Adaptation for Robust Medical Image-to-Image Translation

Irene Iele[1†⋆], Francesco Di Feola[2†], Paolo Soda[1,2], Rosa Sicilia[3§], and Matteo Tortora[4§]

[1] Unit of Artificial Intelligence and Computer Systems, Università Campus Bio-Medico di Roma, Italy
{irene.iele, p.soda}@unicampus.it
[2] Department of Diagnostics and Intervention, Biomedical Engineering and Radiation Physics, Umeå University, Sweden
{francesco.feola, paolo.soda}@umu.se
[3] UniCamillus-Saint Camillus International University of Health Sciences
rosa.sicilia@unicamillus.org
[4] Department of Naval, Electrical, Electronics and Telecommunications Engineering, University of Genoa, Italy
matteo.tortora@unige.it

Abstract. Test-Time Adaptation (TTA) has emerged as a strategy to adapt pretrained models to unseen data during inference without requiring retraining or access to large datasets. Despite recent progress, most existing TTA approaches rely on fixed adaptation strategies, lacking explicit mechanisms to determine whether a test sample requires adaptation and to what extent. As a consequence, adaptation may be unnecessary or suboptimal, especially in scenarios where test samples exhibit heterogeneous degrees of distribution shift, as in the case of medical imaging. Moreover, existing TTA methods are typically developed as by-design solutions that are tightly coupled to specific tasks. This limits modularity and hinders the development of adaptation frameworks that can be flexibly attached to pretrained models. To address these limitations, we propose a novel TTA framework based on a Multi-Agent System formulation that enables sample-specific adaptation and separates adaptation control from the pretrained task model. Our method decomposes TTA into four interacting agents responsible for task execution, distribution shift monitoring, adaptation selection, and model update. We validate our approach by carrying out extensive experiments on two medical image-to-image translation tasks, T_1-to-T_2 MRI translation and low-dose CT denoising, across four publicly available datasets. Results show consistent improvements over the state-of-the-art. Code is available at https://github.com/arco-group/Multi-Agent-TTA.

Keywords: Test-Time Adaptation · Agent System · Medical Imaging

⋆ Corresponding Author.
†These authors contributed equally to this work.
§These authors share senior authorship.

1 Introduction

Test-Time Adaptation (TTA) aims to mitigate performance degradation caused by distribution shifts between training and test data through adaptation mechanisms applied at inference time [12]. TTA is particularly relevant in medical imaging, where distribution shifts arise from variations in acquisition protocols, scanner hardware, and patient populations [8]. Such variability results in test samples that may be in-distribution (ID), i.e., close to the training data, or out-of-distribution (OOD), reflecting distribution shift. Despite the growing interest in TTA, existing approaches in medical imaging remain limited in both scope and formulation. Most methods focus on image segmentation; for instance, [9] proposes a multi-block TTA strategy for breast MRI tumor segmentation based on multi-view consistency and entropy-guided self-training. Similarly, [11] introduces a multi-block input-level adaptation framework that optimizes a shallow normalization network at inference time under a reconstruction prior, whereas [18] uses a domain prior generator to modulate adaptive normalization layers during inference. Beyond segmentation, medical image-to-image translation has attracted increasing interest, supporting key tasks such as noise reduction, cross-modal synthesis, and super-resolution [10,6]. In this context, distribution shifts are particularly critical, as translated images must preserve fine-grained spatial and intensity relationships, and even minor variations can compromise translation fidelity. Although TTA has been extensively studied for segmentation tasks, its application to medical image-to-image translation remains largely unexplored, with only [7] proposing a self-supervised framework based on feature-level adaptation guided by an autoencoder reconstruction loss.

Across the aforementioned methods, TTA approaches exhibit two main limitations: first, they lack selective adaptation mechanisms and remain tightly dependent on the task model. Second, they do not explicitly assess the degree of distribution shift, nor determine whether or how much adaptation is needed. On these grounds, we introduce the Multi-Agent Test-Time Adaptation for medical image-to-image translation, a modular framework that separates adaptation control from the pretrained task model and enables selective, sample-specific updates. It builds on Multi-Agent Systems (MAS) [14], which model complex processes as coordinated interactions among agents with well-defined roles. In summary, our work provides the following contributions: (i) we introduce a modular multi-agent TTA framework for medical image-to-image translation that enables explicit, sample-level control over adaptation; (ii) we structure our TTA approach into four coordinated agents which separate task execution, domain-shift monitoring, adaptation selection, and model update, thereby decoupling adaptation from the pretrained translation model; (iii) we validate our approach on T_1-to-T_2 MRI translation and low-dose CT denoising across four publicly available datasets, achieving improvements over state-of-the-art competitors; (iv) we demonstrate model-agnostic nature by evaluating our approach on both GAN and flow-matching–based pretrained translation models.

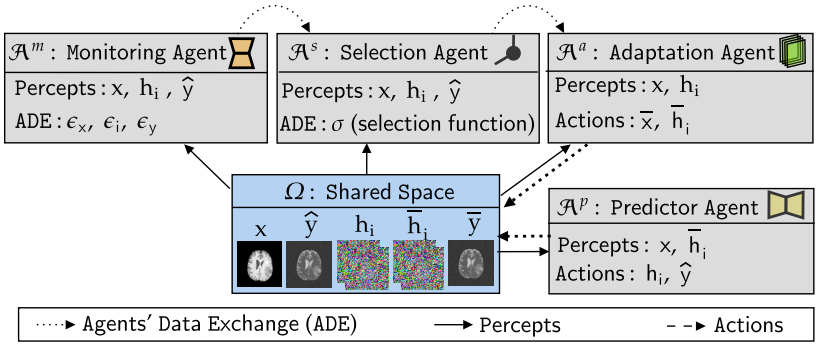

Fig. 1: Multi-Agent TTA. $\mathcal{A}^p$ denotes the predictor agent, $\mathcal{A}^m$ the monitoring agent, $\mathcal{A}^s$ the selection agent and $\mathcal{A}^a$ the adaptation agent. The four agents interact through a shared space Ω. Solid, dashed, and dotted arrows denote percepts, actions, and agent-to-agent communication, respectively.

2 Methods

Let $\boldsymbol{x} \in \mathbb{R}^{w \times h}$ be a source image and $\boldsymbol{y} \in \mathbb{R}^{w \times h}$ its target, with w and h as the image width and height, respectively. Image-to-image translation is defined as $\hat{\boldsymbol{y}} = f_\theta(\boldsymbol{x})$, where f_θ is a learnable mapping and $\hat{\boldsymbol{y}}$ is the output. Under distribution shift, test-time performance may degrade: in this case, instead of retraining f_θ, TTA addresses this by updating model weights at inference time on a per-sample basis. The goal is to reduce the translation error so that $|\bar{\boldsymbol{y}} - \boldsymbol{y}| \leq |\hat{\boldsymbol{y}} - \boldsymbol{y}|$, where $\hat{\boldsymbol{y}}$ corresponds to the model-translated output and $\bar{\boldsymbol{y}}$ the adapted prediction. Our framework, illustrated in Figure 1, is composed of four cooperative agents that interact and share information through a shared space Ω. Interactions within the proposed MAS are mediated by three distinct information flows. The *percepts channel* (solid arrows) communicates information from Ω to the agents, the *action channel* (dashed arrows) enables agents to apply their decisions to Ω, whilst the *agent-to-agent channel (ADE)* (dotted arrows) supports direct communication between agents by exchanging internal feedback signals that coordinate actions on Ω. The Predictor Agent $\mathcal{A}^p$ performs the translation task: given $\boldsymbol{x}$, it produces a prediction $\hat{\boldsymbol{y}}$ and exposes intermediate feature maps $\{\boldsymbol{h}_i\}_{i=1}^n$ from selected internal layers. These representations are shared within Ω and can be modified by the *Adaptation Agent* $\mathcal{A}^a$ before being reintroduced into $\mathcal{A}^p$ during subsequent forward passes. $\mathcal{A}^a$ is composed of a set of trainable feature-level transformation modules $\{a_i\}_{i=1}^n$, referred to as actuators. Each actuator modifies either the input $\boldsymbol{x}$ or a specific feature map $\boldsymbol{h}_i$. Given a selected feature index subset $\mathcal{S} \subseteq \{1, \ldots, n\}$, $\mathcal{A}^a$ activates the corresponding actuators $\{a_i\}_{i \in \mathcal{S}}$, producing adapted representations $\{\bar{\boldsymbol{h}}_i\}_{i=1}^n$ and an adapted input $\bar{\boldsymbol{x}}$. These adapted features are then given back into Ω and made available to $\mathcal{A}^p$, which performs a new forward pass to generate the adapted prediction $\bar{\boldsymbol{y}}$. The *Selection Agent* $\mathcal{A}^s$ in collaboration with the *Monitoring Agent* $\mathcal{A}^m$ deter-

mines both whether adaptation is needed and its extent. The *Monitoring Agent* $\mathcal{A}^m$ estimates distribution shift through reconstruction errors. It comprises a set of observers $\{O_i\}_{i=1}^n$, independently trained on the same data distribution as $\mathcal{A}^p$ to reconstruct their respective representation levels, and kept frozen during TTA. For a representation $\boldsymbol{z}_i \in \{\boldsymbol{h}_i, \bar{\boldsymbol{h}}_i\}$, where $\boldsymbol{h}_i$ denotes the original feature map at level i and $\bar{\boldsymbol{h}}_i$ its adapted counterpart, the reconstruction error is defined as $\epsilon_i = \|\boldsymbol{z}_i - O_i(\boldsymbol{z}_i)\|_1$. Input and output level errors are computed as $\epsilon_x = \|\boldsymbol{z}_x - O_x(\boldsymbol{z}_x)\|_1$ with $\boldsymbol{z}_x \in \{\boldsymbol{x}, \bar{\boldsymbol{x}}\}$, where $\boldsymbol{x}$ denotes the original input and $\bar{\boldsymbol{x}}$ its adapted version, and $\epsilon_y = \|\hat{\boldsymbol{y}} - O_y(\hat{\boldsymbol{y}})\|_1$. Such reconstruction errors serve as a proxy for distribution shift, with higher values indicating greater deviation from the training distribution. The *Selection Agent* $\mathcal{A}^s$ is based on a selection function σ which identifies the test samples requiring adaptation based on the reconstruction error ϵ_y computed by $\mathcal{A}^m$. To determine the appropriate extent of adaptation, $\mathcal{A}^s$ explores different actuator subsets $\mathcal{S}$ and evaluates each configuration using the reconstruction feedback ϵ_y provided by $\mathcal{A}^m$, which quantifies the output reconstruction error. Since the search space grows exponentially with the number of actuator levels (2^{i-1} possible subsets), we approximate this optimization via random search: for each flagged sample, K actuator subsets $\{\mathcal{S}_k\}_{k=1}^K$ are sampled uniformly at random, and the configuration yielding the lowest reconstruction error ϵ_y is selected as $\mathcal{S}^* = \arg\min_{k\in\{1,\dots,K\}} \epsilon_y(\mathcal{S}_k)$.

3 Experiments and Results

Implementation details. The predictor agent $\mathcal{A}^p$ is implemented with two learning paradigms: (i) $\mathcal{A}_1^p$ is GAN-based, a CycleGAN with a 9-block ResNet generator [21] trained with Adam (2×10^{-4}, batch 8, 100 epochs); (ii) $\mathcal{A}_2^p$ is flow-matching–based, a UNet trained with flow-matching objective [13] using rectified scheduling (1000 steps), Adam (1.5×10^{-5}, batch 16, 300 epochs) and single-step inference. The monitoring agent $\mathcal{A}^m$ consists of multi-level convolutional autoencoders [3,2], denoted as observers O_i, each trained independently on the predictor training set (Adam, batch 16, 50 epochs) with the objective of reconstructing the feature maps $\boldsymbol{h}_i$ at each level of $\mathcal{A}^p$. The adaptation agent $\mathcal{A}^a$ is implemented as lightweight 1×1 convolutional adaptors $\{a_i\}$ operating on Ω, with only selected modules optimized at test time while keeping $\mathcal{A}^p$ frozen. The adaptors optimization is driven by the reconstruction errors computed by $\mathcal{A}^m$. For feature-level selection, we uniformly sample $K = 50$ actuator configurations to bound the search space, as a trade-off between computational cost and search efficiency. For each sampled configuration, adaptation is performed for up to $T = 10$ iterations, with early stopping based on ϵ_y. We investigate three selection functions within $\mathcal{A}^s$ that regulate the activation of $\mathcal{A}^a$, spanning static and dynamic strategies. The static selection function σ_1 defines the decision threshold τ empirically as the 95^{th} percentile of the reconstruction error computed on the available test set used during training of $\mathcal{A}^p$, and triggers adaptation when $\boldsymbol{\epsilon}_y > \tau$. This percentile-based choice allows us to model the upper tail of the ID reconstruction error, such that samples exceeding τ are likely to exhibit atypi-

Table 1: Datasets used for $T_1 \rightarrow T_2$ MRI translation and LD $\rightarrow$ HD CT denoising.

Task	Dataset	ID	#Train	#Test	Notes
$T_1 \rightarrow T_2$	BraTS 2018 [15]	–	5760	768	Brain glioma patients
	OASIS-3 [17]	*A*	–	11054	Alzheimer's disease
	UPENN-GBM [1]	*B*	–	11043	Glioblastoma multiforme
	UCSF-PDGM [4]	*C*	–	3273	Diffuse gliomas, varying grades
LD $\rightarrow$ HD	Mayo Clinic LDCT [16]	–	5936	–	Abdominal scans (used for training)
		D	–	5133	Abdominal scans w/ liver cancer
		E	–	16950	Chest scans w/ lung nodules

cal representations and therefore benefit from adaptation. In order to take into account possible data drift, we include two dynamic selection strategies: (i) the first, σ_2, is based on random resampling and updates τ every j test samples as the 95$^{\text{th}}$ percentile of $k \leq j$ reconstruction errors sampled from recent history $\tau_t = \mathrm{Q}_{0.95}\left(\{\epsilon_y^{(i)}\}_{i=1}^{k}\right)$ with *j=60, k=30.* (ii) The second σ_3 is an exponential moving average (EMA) updating τ after a warm-up of l samples every j samples as $\tau_t = \alpha\,\mathrm{Q}_{0.95}\left(\epsilon_{y_{t-k:t}}\right) + (1-\alpha)\tau_{t-1}$ where *l=60, j=30, k=30* and $\alpha = 0.2$.

Datasets. Table 1 reports the datasets used for two image-to-image translation tasks: $T_1 \rightarrow T_2$ MRI and low dose to high dose CT (LD $\rightarrow$ HD). For MRI, $\mathcal{A}^p$ is trained on BraTS 2018 and evaluated on three external cohorts (OASIS-3, UPENN-GBM, UCSF-PDGM), which differ in patient populations and pathological distributions. In this setting, the threshold τ is computed exclusively from the reconstruction error distribution on the BraTS 2018 test set. τ is therefore calibrated using ID data and remains fixed during evaluation on all external cohorts. For CT, training is performed on abdominal Mayo Clinic scans, with evaluation on both abdominal and chest data, introducing an anatomical shift. In this case, the distribution shift is primarily anatomical rather than population-based. Accordingly, τ is determined from the full available training set, without requiring an additional population shift. The same fixed τ is then applied during evaluation on both abdominal and chest test sets.

Competitors. We compare the proposed framework against two state-of-the-art methods applicable to the image-to-image translation setting [7] and [11], both applying uniform TTA without explicit ID/OOD discrimination. Both methods were implemented following the original publications and using the publicly available official code repositories. Whenever required, models were retrained on our datasets under the same experimental protocol adopted for our MAS approach to ensure a fair comparison.

Results. Table 2 reports the results of our experiments on T_1-to-T_2 MRI translation (Datasets A–C) and LDCT denoising (Datasets D–E), using predictors $\mathcal{A}_1^p$ and $\mathcal{A}_2^p$. For each dataset, five configurations are evaluated: two corresponding to the competing methods described above and three variants of our approach differing in the selection strategy (σ_1, σ_2, and σ_3). The table is organized by dataset and divided into two blocks corresponding to $\mathcal{A}_1^p$ and $\mathcal{A}_2^p$. Image qual-

Table 2: TTA results on MRI and CT translation. For each dataset, we report PSNR and ΔPSNR (dB), SSIM and ΔSSIM (%), and the percentage of test samples adapted by each method. **Best** and **second-best** results are highlighted.

DB		Method	$\mathcal{A}_1^p$ PSNR↑	$\mathcal{A}_1^p$ SSIM↑	$\mathcal{A}_1^p$ ΔPSNR↑	$\mathcal{A}_1^p$ ΔSSIM↑	$\mathcal{A}_1^p$ % adapted	$\mathcal{A}_2^p$ PSNR↑	$\mathcal{A}_2^p$ SSIM↑	$\mathcal{A}_2^p$ ΔPSNR↑	$\mathcal{A}_2^p$ ΔSSIM↑	$\mathcal{A}_2^p$ % adapted
$T_1 \to T_2$	A	[7]	$11.38_{\pm 2.56}$	$46.49_{\pm 20.26}$	-9.46	-39.04	100.00	$13.27_{\pm 2.61}$	$50.23_{\pm 23.32}$	-8.89	-37.22	100.00
		[11]	$20.15_{\pm 2.25}$	$83.30_{\pm 6.48}$	-.82	-1.70	100.00	$23.39_{\pm 2.13}$	$84.17_{\pm 4.66}$	-.36	-1.41	100.00
		σ_1	**$23.16_{\pm 2.28}$**	**$85.53_{\pm 5.24}$**	**+2.29**	**+.72**	96.44	$21.76_{\pm 2.36}$	$83.45_{\pm 5.96}$	+.21	-.16	34.80
		σ_2	$21.06_{\pm 2.67}$	$84.84_{\pm 5.25}$	+2.20	**+.11**	5.23	**$23.78_{\pm 2.28}$**	**$85.58_{\pm 4.92}$**	**+.26**	.00	5.55
		σ_3	**$21.13_{\pm 2.73}$**	**$84.84_{\pm 5.26}$**	**+2.27**	+.04	7.80	**$23.77_{\pm 2.28}$**	**$85.57_{\pm 4.92}$**	**+.25**	**+.08**	6.95
	B	[7]	$11.13_{\pm 1.55}$	$53.34_{\pm 8.70}$	-9.41	-31.98	100.00	$12.59_{\pm 1.71}$	$60.07_{\pm 10.12}$	-8.75	-38.25	100.00
		[11]	$22.21_{\pm 2.44}$	$85.76_{\pm 4.71}$	-.24	-.53	100.00	$23.92_{\pm 2.02}$	$85.39_{\pm 4.67}$	-.14	-.78	100.00
		σ_1	**$23.68_{\pm 2.29}$**	**$86.87_{\pm 4.66}$**	**+1.43**	**+0.61**	91.48	**$24.17_{\pm 2.03}$**	$86.05_{\pm 4.80}$	+.16	-.17	69.37
		σ_2	$22.39_{\pm 2.59}$	$86.32_{\pm 4.64}$	+.34	+.08	5.59	**$24.08_{\pm 2.03}$**	**$86.17_{\pm 4.72}$**	**+.26**	-.17	5.58
		σ_3	**$22.40_{\pm 2.60}$**	**$86.32_{\pm 4.65}$**	**+0.46**	**+0.13**	7.25	**$24.08_{\pm 2.03}$**	**$86.16_{\pm 4.72}$**	**+.27**	**+.10**	7.03
	C	[7]	$10.68_{\pm 1.89}$	$54.74_{\pm 6.64}$	-7.93	-27.67	100.00	$12.02_{\pm 1.95}$	$60.12_{\pm 7.45}$	-7.73	-22.36	100.00
		[11]	$19.96_{\pm 2.44}$	$85.76_{\pm 4.71}$	-.05	-.24	100.00	$21.56_{\pm 2.39}$	$82.83_{\pm 6.22}$	-.13	-.65	100.00
		σ_1	**$21.05_{\pm 2.22}$**	**$83.08_{\pm 6.45}$**	**+1.63**	**+1.56**	71.86	**$21.76_{\pm 2.36}$**	$83.45_{\pm 5.96}$	+.21	.00	34.80
		σ_2	$20.03_{\pm 2.11}$	$82.03_{\pm 6.51}$	**+2.27**	**+1.06**	6.66	$21.70_{\pm 2.35}$	**$83.46_{\pm 5.98}$**	**+.22**	**+.06**	6.48
		σ_3	**$20.08_{\pm 2.12}$**	**$82.04_{\pm 6.51}$**	+2.17	+.97	8.92	**$21.71_{\pm 2.36}$**	**$83.46_{\pm 5.96}$**	**+.23**	**+.12**	7.61
LD→HD	D	[7]	$16.30_{\pm 1.60}$	$50.15_{\pm 8.28}$	-16.86	-42.59	100.00	$20.12_{\pm 2.78}$	$57.89_{\pm 7.17}$	-19.81	-47.01	100.00
		[11]	$33.29_{\pm 6.98}$	$90.60_{\pm 9.41}$	+2.19	+1.20	100.00	$31.29_{\pm 6.98}$	$89.50_{\pm 9.41}$	+.12	+.08	100.00
		σ_1	**$33.49_{\pm 5.59}$**	$90.73_{\pm 9.05}$	**+16.57**	**+14.43**	13.99	**$36.26_{\pm 3.07}$**	**$87.59_{\pm 6.06}$**	**+.24**	**+.91**	13.70
		σ_2	**$33.61_{\pm 5.42}$**	**$90.89_{\pm 8.62}$**	+16.04	+14.35	15.20	$36.26_{\pm 3.08}$	$87.57_{\pm 6.11}$	+.12	**+1.26**	2.32
		σ_3	$33.41_{\pm 5.62}$	**$90.74_{\pm 8.75}$**	**+16.29**	**+14.76**	13.72	**$36.29_{\pm 3.08}$**	**$87.64_{\pm 6.07}$**	**+.25**	+.85	6.55
	E	[7]	$13.28_{\pm 1.08}$	$21.10_{\pm 9.29}$	-10.96	-30.64	100.00	$17.25_{\pm 1.98}$	$27.46_{\pm 10.33}$	-12.06	-32.71	100.00
		[11]	$25.44_{\pm 2.23}$	$53.51_{\pm 9.28}$	-.25	-.42	100.00	$25.70_{\pm 2.23}$	$53.93_{\pm 9.28}$	+.34	.00	100.00
		σ_1	**$25.79_{\pm 2.08}$**	**$56.11_{\pm 9.16}$**	**+4.01**	**+2.78**	2.21	**$28.46_{\pm 2.86}$**	**$68.69_{\pm 8.31}$**	**+.36**	**+.59**	52.89
		σ_2	**$25.81_{\pm 2.08}$**	**$56.11_{\pm 9.18}$**	+2.88	**+1.67**	3.81	$28.29_{\pm 2.87}$	$68.41_{\pm 8.23}$	+.30	**+.42**	7.06
		σ_3	**$25.79_{\pm 2.08}$**	$56.10_{\pm 9.17}$	**+3.21**	+.02	2.00	**$28.31_{\pm 2.86}$**	**$68.42_{\pm 8.23}$**	**+.39**	+.30	11.13

ity is evaluated using PSNR and SSIM [5,19]. PSNR ranges in $[0, +\infty)$, while SSIM ranges in $[0, 1]$ and is reported here as a percentage $[0, 100]$; in both cases, higher values indicate better reconstruction fidelity and structural consistency. Absolute PSNR and SSIM are computed over the entire test set, including both adapted and non-adapted samples, ensuring fair and directly comparable evaluation across methods. ΔPSNR and ΔSSIM denote the difference between post and pre adaptation performance, computed only on the subset of samples selected for TTA; positive values indicate improvement over the baseline prediction without adaptation, whereas negative values denote degradation. The corresponding adaptation rate is reported in the last column of each predictor block. Best and second-best results are highlighted in bold red and blue, respectively. Performance improvements (ΔPSNR, ΔSSIM) are statistically validated using the Wilcoxon signed-rank test [20] ($p < 0.05$). As shown in Table 2, the competing approaches [7,11] adapt 100% of test samples by design. In contrast, the proposed strategies (σ_1, σ_2, and σ_3) restrict adaptation to fewer than 20% of samples in most settings. Across MRI datasets, our approach yields statistically significant improvements for both $\mathcal{A}_1^p$ and $\mathcal{A}_2^p$. While the static strategy σ_1 relies on a fixed threshold calibrated on ID data and therefore triggers adaptation more frequently, the dynamic strategies σ_2 and σ_3 substantially reduce the number of adapted samples while preserving comparable performance gains. Specifically, σ_2 updates the threshold via random resampling of recent reconstruction errors, yielding a conservative and locally adaptive behavior. The EMA strategy σ_3 further stabilizes the threshold through temporal smoothing, resulting in a more consistent adaptation policy. Both dynamic strategies match the performance of σ_1 while operating at markedly lower adaptation rates, indicating improved

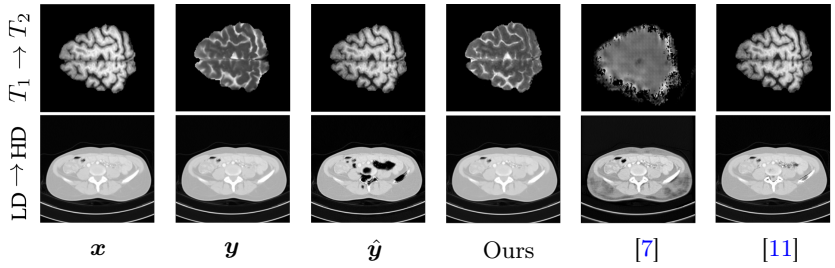

Fig. 2: Qualitative comparison across MRI and CT translation tasks. From left to right: input image ($\boldsymbol{x}$), ground truth ($\boldsymbol{y}$), prediction without TTA ($\hat{\boldsymbol{y}}$), Ours (best selection function per dataset), and competitors [7,11].

adaptation efficiency. A similar trend is observed in LDCT denoising, where the predictor agents are trained on abdominal scans and also evaluated on unseen chest scans, introducing an anatomical shift. Comparing predictors, gains are larger for $\mathcal{A}_1^p$ than for $\mathcal{A}_2^p$. The higher baseline performance of the flow-matching-based predictor $\mathcal{A}_2^p$ reduces the margin for improvement, indicating a lower need for adaptation rather than a limitation of the proposed method. Figure 2 provides a qualitative comparison including both competing approaches and the proposed method. For brevity, we report representative examples on A (MRI) and D (CT) datasets. On MRI, the $\hat{\boldsymbol{y}}$ tends to preserve characteristics of the source domain, resulting in outputs that remain visually closer to $\boldsymbol{x}$ than to the target $\boldsymbol{y}$. Our approach, instead produces translations that are structurally and contrast-wise more consistent with the ground truth. On CT, $\hat{\boldsymbol{y}}$ exhibits visible artifacts and intensity inconsistencies, whereas our approach reduces these degradations and yields cleaner reconstructions. In both tasks, our method provides sharper anatomical details within the highlighted regions compared to competing methods. Figure 3 provides a per-sample analysis of our approach by reporting, for each predictor agent, the percentage of adapted samples that exhibit an improvement ("win") or degradation ("loss") in PSNR and SSIM. For brevity, results are aggregated across MRI and CT datasets using, for each dataset, the best-performing configuration of the proposed method (see bold red values in Table 2). For $\mathcal{A}_1^p$, TTA yields improvements in the large majority of adapted samples across both MRI and CT tasks. The win rate exceeds 80% in most settings, reaching up to 98.6% in CT denoising, while degradations remain limited to at most 27.8%. This indicates that when adaptation is triggered, it is beneficial in over 70% of the cases. For $\mathcal{A}_2^p$, improvements dominate for PSNR, whereas SSIM in MRI exhibits a more balanced distribution between wins and losses, with a win rate as low as 55.3%. This aligns with the smaller average gains observed in Table 2 and reflects the higher robustness of $\mathcal{A}_2^p$, which reduces the margin for improvement. Nonetheless, positive effects remain prevalent.

Ablation study. Table 3 evaluates the contribution of each agent by progressively enabling monitoring ($\mathcal{A}^m$), adaptation ($\mathcal{A}^a$), and selection ($\mathcal{A}^s$), where selection is instantiated either at sample level (σ_i) or the feature level ($\mathcal{S}_k$). For

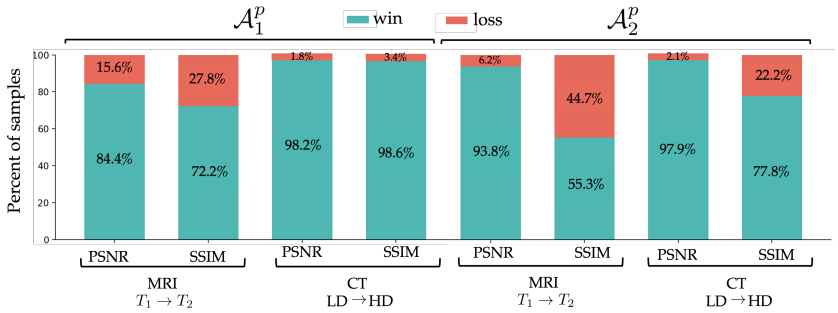

Fig. 3: Win–loss plot. Adapted samples (%) increasing or decreasing PSNR and SSIM across MRI and CT for $\mathcal{A}_1^p$, and $\mathcal{A}_2^p$ using the best strategy per dataset.

Table 3: Ablation study. Each configuration enables (✓) or disables (✗) $\mathcal{A}^m$, $\mathcal{A}^a$, $\mathcal{A}^s$; the full model is highlighted in blue. **Bold** indicates best performance.

DB	$\mathcal{A}_1^p$		$\mathcal{A}^m$	$\mathcal{A}^a$	$\mathcal{A}^s$		$\mathcal{A}_2^p$	
	PSNR↑	SSIM↑			$\mathcal{S}_k$	σ_i	PSNR↑	SSIM↑
$T_1 \rightarrow T_2$	$21.43_{\pm 2.64}$	$.85_{\pm .05}$	✗	✗	✗	✗	$23.62_{\pm 2.29}$	$.86_{\pm .05}$
	$23.22_{\pm 2.04}$	$.86_{\pm .04}$	✓	✓	✗	✗	$23.51_{\pm 2.13}$	$.85_{\pm .04}$
	$23.35_{\pm 2.18}$	$.86_{\pm .05}$	✓	✓	✓	✗	$23.79_{\pm 2.32}$	$.85_{\pm .05}$
	$23.82_{\pm 2.58}$	**$.87_{\pm .04}$**	✓	✓	✓	✓	**$29.17_{\pm 2.42}$**	**$.89_{\pm .04}$**
$LD \rightarrow HD$	$26.97_{\pm 4.54}$	$.64_{\pm .17}$	✗	✗	✗	✗	$30.12_{\pm 4.46}$	$.73_{\pm .11}$
	$28.14_{\pm 5.19}$	$.65_{\pm .18}$	✓	✓	✗	✗	$29.74_{\pm 4.12}$	$.73_{\pm .10}$
	$28.16_{\pm 5.28}$	$.64_{\pm .19}$	✓	✓	✓	✗	$30.54_{\pm 4.49}$	$.73_{\pm .11}$
	$30.36_{\pm 4.49}$	**$.74_{\pm .17}$**	✓	✓	✓	✓	**$31.12_{\pm 4.27}$**	**$.75_{\pm .16}$**

each dataset, we use the best-performing configuration identified in Table 2. As before, we evaluate all configurations on the full test set. However, to better highlight intra-approach differences, the reported metrics are computed over the subset of samples for which adaptation is effectively triggered, and averaged across datasets within each task. Across both tasks, enabling $\mathcal{A}^m$ and $\mathcal{A}^a$ improves metrics over the standalone predictor, supporting the use of reconstruction errors as a proxy for distribution shift. When adaptation is applied without selection, all actuators are activated uniformly, leading to limited gains and, in some cases, performance degradation. Activating $\mathcal{S}_k$ improves the results by allowing the adaptation process to specialize the actuator subset for each individual sample. Finally, combining $\mathcal{S}_k$ with σ_i, which determines which inputs should trigger adaptation, yields the best performance across both tasks and predictor agents. The ablation confirms that the effectiveness of our framework stems from the coordinated interaction between monitoring, adaptation, and selection agents.

4 Conclusion

This work introduces a modular multi-agent framework for selective TTA in medical image-to-image translation, explicitly separating shift monitoring, adaptation selection, and parameter update, enabling sample-specific control under heterogeneous distribution shifts. Experimental results on MRI translation and LDCT denoising demonstrate improvements over state-of-the-art methods, generalizing across different predictor paradigms. These findings demonstrate the effectiveness of controlled, sample-specific TTA. Future work will extend to 3D medical imaging and coordinated interaction among multiple predictor agents, enabling dynamic model selection, ensemble-based adaptation, advanced selection strategies, adaptive actuator policies, and uncertainty-aware monitoring.

Acknowledgments. Irene Iele is a Ph.D. student enrolled in the National Ph.D. in Artificial Intelligence, XL cycle, course on Health and Life Sciences, organized by Università Campus Bio-Medico di Roma. Resources are provided by the National Academic Infrastructure for Supercomputing in Sweden (NAISS) and the Swedish National Infrastructure for Computing (SNIC) at Alvis @ C3SE. This work was partially supported by: i) Missione 6/componente 2/Investimento: 2.1 "Rafforzamento e potenziamento della ricerca biomedica del SSN", finanziato dall'Unione europea – NextGenerationEU (CUP C83C24000220007) ii) Kempe Foundation project JCSMK24-0094.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bakas, S. et al.(2021). Multi-parametric magnetic resonance imaging (mpMRI) scans for de novo Glioblastoma (GBM) patients from the University of Pennsylvania Health System (UPENN-GBM): The Cancer Imaging Archive https://doi.org/10.7937/TCIA.709X-DN49/
2. Baur, C., et al.: Autoencoders for unsupervised anomaly segmentation in brain MR images: a comparative study. Medical image analysis **69**, 101952 (2021)
3. Bengio, Y., Courville, A., Vincent, P.: Representation learning: A review and new perspectives. IEEE transactions on pattern analysis and machine intelligence **35**(8), 1798–1828 (2013)
4. Calabrese, E., et al. (2022). The University of California San Francisco Preoperative Diffuse Glioma MRI (UCSF-PDGM): The Cancer Imaging Archive https://doi.org/10.7937/tcia.bdgf-8v37/
5. Di Feola, F., Tronchin, L., Soda, P.: A comparative study between paired and unpaired Image Quality Assessment in Low-Dose CT Denoising. In: 2023 IEEE 36th International Symposium on Computer-Based Medical Systems (CBMS). pp. 471–476. IEEE (2023)
6. Guarrasi, V., et al.: Medicine Without Boundaries: Generative AI for Translating Medical Data Across Modalities (2025)
7. He, Y., et al.: Autoencoder based self-supervised test-time adaptation for medical image analysis. Medical image analysis **72**, 102136 (2021)

8. Hognon, C., et al.: Contrastive image adaptation for acquisition shift reduction in medical imaging. Artificial Intelligence in Medicine **148**, 102747 (2024)
9. Joshi, S., et al.: Single Image Test-Time Adaptation via Multi-View Co-Training. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 617–627. Springer (2025)
10. Kaji, S., Kida, S.: Overview of image-to-image translation by use of deep neural networks: denoising, super-resolution, modality conversion, and reconstruction in medical imaging. Radiological physics and technology **12**(3), 235–248 (2019)
11. Karani, N., et al.: Test-time adaptable neural networks for robust medical image segmentation. Medical Image Analysis **68**, 101907 (2021)
12. Liang, J., He, R., Tan, T.: A comprehensive survey on test-time adaptation under distribution shifts. International Journal of Computer Vision **133**(1), 31–64 (2025)
13. Lipman, Y., et al.: Flow matching for generative modeling. arXiv preprint arXiv:2210.02747 (2022)
14. Maldonado, D., et al.: Multi-agent systems: A survey about its components, framework and workflow. IEEE Access **12**, 80950–80975 (2024)
15. Menze, B.H., et al.: The multimodal brain tumor image segmentation benchmark (BRATS). IEEE Transactions on Medical Imaging **34**(10), 1993–2024 (2014)
16. Moen, T.R., et al.: Low-dose CT image and projection dataset. Medical physics **48**(2), 902–911 (2021)
17. T. Benzinger et al.; NIH P30 AG066444, P50 AG00561, P30 NS09857781, P01 AG026276, P01 AG003991, R01 AG043434, UL1 TR000448, R01 EB009352. AV-45 doses were provided by Avid Radiopharmaceuticals, a wholly owned subsidiary of Eli Lilly.: The Cancer Imaging Archive https://sites.wustl.edu/oasisbrains/home/oasis-3/
18. Valanarasu, J.M.J., et al.: On-the-fly test-time adaptation for medical image segmentation. In: Medical Imaging with Deep Learning. pp. 586–598. PMLR (2024)
19. Wang, Z., et al.: Image quality assessment: from error visibility to structural similarity. IEEE transactions on image processing **13**(4), 600–612 (2004)
20. Wilcoxon, F.: Individual comparisons by ranking methods. In: Breakthroughs in statistics: Methodology and distribution, pp. 196–202. Springer (1992)
21. Zhu, J.Y., et al.: Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision. pp. 2223–2232 (2017)