SOURCE OR PREFIX? RETAINING WHAT HELPS THE TARGET IN AUTOREGRESSIVE 3D MEDICAL IMAGE TRANSLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Translating between 3D medical imaging modalities requires preserving patient anatomy while recovering target-specific properties. A source-conditioned block-wise autoregressive model uses the source as a fixed anatomical anchor and earlier blocks as an evolving target prefix, and the difference between predictions with and without the prefix defines the prefix correction. The balance between source and prefix evidence varies across regions, while the magnitude of a correction's prediction change (*effect*) does not necessarily reveal whether it helps the true target (*utility*). In this work, we propose Source-Prefix Influence-aware Residual Autoregression (SPIRA), which learns *how much of each prefix correction to retain across regions*. During training, SPIRA finds the largest removable portion of each correction without increasing target loss at the current prediction states and trains a lightweight gate to reproduce the resulting predictive distribution, thus learning how to balance source and prefix during generation. Extensive experiments across multiple modalities and diverse 3D medical translation tasks demonstrate consistent gains in translation quality. Mechanism analyses show that similar prefix effects can have different utilities depending on target alignment, while SPIRA better matches oracle decisions and achieves larger gains when more prefix correction can be safely retracted. Code is available at https://anonymous.4open.science/r/SPIRA-4547.

1 INTRODUCTION

3D medical image translation synthesizes a target-modality volume from a source-modality volume (Wang et al., 2025; Romoli et al., 2026), with potential applications such as reducing gadolinium use through contrast-enhanced MRI synthesis and supporting radiotherapy planning through synthetic CT (Özbey et al., 2023; Zhou et al., 2026). In paired settings, the source already provides rich patient-specific anatomy and spatial structure, making it a strong anchor for target generation (Chen et al., 2025a). However, the same anatomy can appear differently across imaging modalities (Shin et al., 2025; Du et al., 2026). The source alone may therefore leave target-specific properties such as contrast, density, or functional activity ambiguous (Meng et al., 2024; Zhao et al., 2026).

An autoregressive (AR) formulation complements this source anchor with earlier target content (Tian et al., 2024; Zhu et al., 2026), which we call the target *prefix*. Each prediction therefore combines a fixed, spatially aligned source condition with evolving target-side context. During training, the prefix comes from the true target; at inference, it is generated from previous predictions and may carry errors forward (Huang et al., 2025; Mao et al., 2026). The relative contributions of the fixed condition and evolving context therefore need not be uniform across predictions (Jin et al., 2026; Chen et al., 2026). In paired medical translation, this variation is spatially explicit because the source is aligned with the target and already captures much of its anatomy (Dorent et al., 2026; Song et al., 2026a). The prefix then contributes a *correction* beyond the source-only prediction, whose value can vary across regions. Some regions are already well explained by the source, whereas others still benefit from target-side context (Liu et al., 2025) (Fig. 1a). This leads to our central question:

Key Question

How much of the current prefix correction should be retained for each target region?

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.

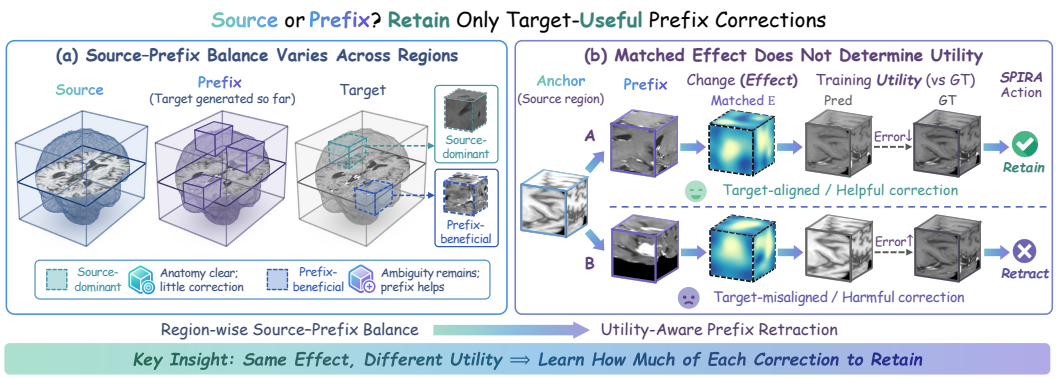

Figure 1: **(a) Challenge:** The balance between source evidence and target-side context varies across target regions. **(b) Insight:** Predictive *effect* measures how much a prefix correction changes the prediction, while correction *utility measures* whether that change favors the true target. Corrections with matched effects can therefore have different utilities, motivating utility-aware prefix retraction.

To answer this question, a natural proxy for prefix need is its predictive *effect*, namely how much the correction changes the predictive distribution (Chuang et al., 2024). We define its *utility* by whether this change favors the true target. Predictive effect alone, however, does not determine utility. In Fig. 1b, corrections with similar effects may increase the probability of the target, leave it almost unchanged, or decrease it. This pattern is also observed on BraTS 2024 (de Verdier et al., 2024), where among regions with similar prediction changes, 29.3% reduce the true-target likelihood while 40.0% increase it (Sec. 2), further showing that similar effects can have different utilities.

Key Insight

Prefix corrections with matched predictive effects can have different utilities, so effect alone cannot determine how much of a correction to retain.

Building upon this insight, we propose **Source-Prefix Influence-aware Residual Autoregression (SPIRA)**, which controls the contribution of generated context through selective prefix retraction to *reduce harmful or unnecessary prefix corrections* while *preserving target-useful prefix information*. Specifically, we instantiate SPIRA for 3D medical image translation in a source-conditioned blockwise AR model that generates 3D latent blocks sequentially. Within this paradigm, a shared AR model produces source-only and full-prefix predictions, whose difference is the current prefix correction. We define target-aware retraction regret as the change in target loss after retraction relative to full-prefix AR, and show that it decomposes exactly into predictive effect and target alignment. During training, a ground-truth (GT) oracle finds the maximal zero-regret retraction, which is the largest portion of the correction that can be retracted without increasing target loss. A lightweight correction-aware gate learns to reproduce the resulting predictive distribution from the source and current prefix. At inference, the gate applies the predicted retraction before each new target unit is generated and appended to the prefix. Our contributions are summarized as follows:

- **Reframing.** We frame source-conditioned AR generation as a balance between a fixed, information-rich condition and evolving target-side context, and formulate this problem region-wise in 3D medical image translation, where predictive effect alone does not determine how much of a prefix correction should be retained.
- **Approach.** We derive an exact decomposition of target-aware retraction regret into predictive effect and target alignment, and introduce SPIRA with a maximal zero-regret oracle and a lightweight gate that directly distills the oracle predictive distribution.
- **Experiments.** Across eight translation tasks on three datasets, SPIRA consistently improves translation quality. Further analyses explain why corrections with similar effects can require different retention and how SPIRA learns these distinctions to improve free-running generation.

2 Motivation Analysis

We conduct the analysis on BraTS 2024 (de Verdier et al., 2024) using the T1n→T1c task. Under teacher forcing, we compare predictions with and without the preceding GT target prefix to isolate clean prefix effects from accumulated generation errors. Further details are provided in Appendix C.

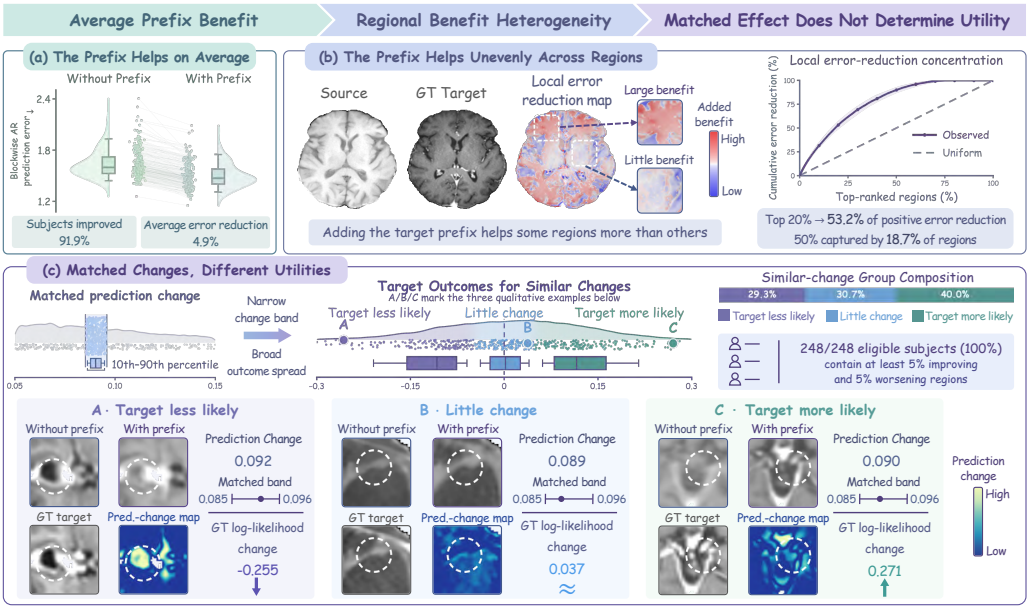

Figure 2: **Motivation analysis. (a)** In token space, the target prefix reduces prediction error overall. **(b)** In decoded image space, its benefit varies substantially across regions. **(c)** In predictive-distribution space, similar prediction changes can yield different true-target outcomes.

The Prefix Helps on Average. As shown in Fig. 2(a), adding the target prefix improves 91.9% of held-out subjects and reduces blockwise AR prediction error by 4.9% on average, confirming that it provides useful information beyond the source.

The Prefix Helps Unevenly Across Regions. We decode both predictions and measure local reconstruction-MAE reduction over fixed 3D regions. Fig. 2(b) shows that some regions benefit substantially while others change little. Across subjects, the top 20% of regions account for 53.2% of the positive error reduction, indicating strongly region-dependent prefix benefit.

Matched Changes, Different Utilities. We next ask whether similarly sized prefix-induced prediction changes yield similar target outcomes. Using factor-averaged total variation distance, we select prediction changes within the 45th–55th percentile range and categorize the corresponding true-target likelihood changes using a small neutral range around zero. As shown in Fig. 2(c), similar prediction changes still produce improving, neutral, and worsening outcomes, and all 248 eligible subjects contain both improving and worsening regions. Because the prefix is GT, these harmful outcomes arise even with correct target-side context rather than accumulated generation errors.

Takeaway

The target prefix provides region-dependent information beyond the source, while similar prediction changes can have different utilities. Thus, predictive effect alone cannot determine the appropriate source–prefix balance, or equivalently how much of each prefix correction to retain.

3 METHODOLOGY

Motivated by the above observations, we introduce SPIRA, which learns a region-wise source–prefix balance by controlling how much of the prefix-induced update is retained at each target region. As illustrated in Fig. 3, the same AR predictor produces predictions with and without the current prefix; their difference defines the prefix correction. SPIRA then learns to retract only the portion that can be removed without increasing target loss. We introduce the AR formulation in Sec. 3.1, characterize prefix correction and retraction in Secs. 3.2–3.3, and summarize training and inference in Sec. 3.4.

3.1 SOURCE-CONDITIONED BLOCKWISE AR

Given a paired 3D source and target volume, a frozen visual tokenizer maps them into spatially aligned discrete latent tokens. For simplicity, we denote the encoded source tokens by S and partition the

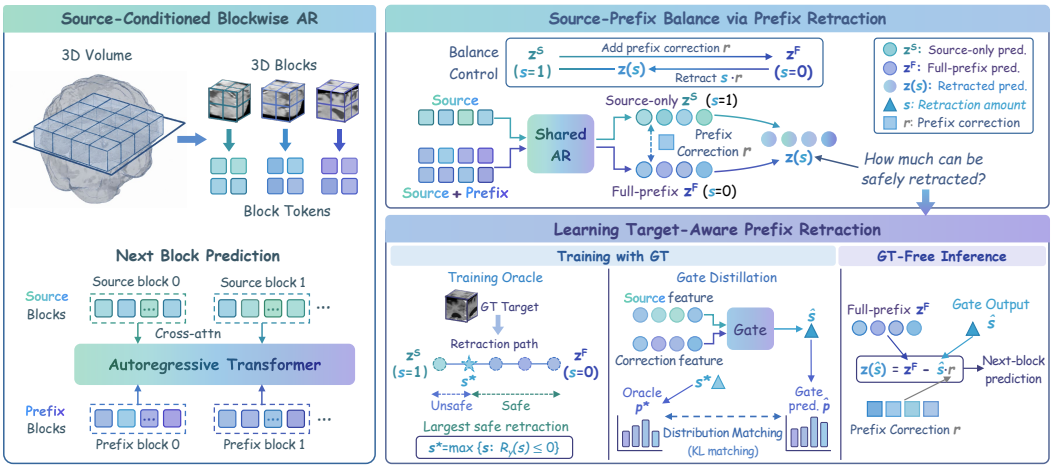

Figure 3: **Overview of SPIRA.** A shared source-conditioned blockwise AR predictor produces source-only and full-prefix logits, whose difference defines a prefix correction and a continuous retraction path. During training, a GT oracle identifies the maximal zero-regret retraction, and a correction-aware gate distills the resulting predictive distribution for GT-free generation.

target latent grid into an ordered sequence of non-overlapping 3D block tokens $Y_{1:N}$. Following Ren et al. (2025a), we factorize the conditional distribution as

$$p_\theta(Y_{1:N} \mid S) = \prod_{i=1}^{N} p_\theta(Y_i \mid S, P_{<i}), \qquad P_{<i} = (Y_1, \ldots, Y_{i-1}), \tag{1}$$

where $P_{<i}$ denotes the target prefix available before predicting block i.

The source is bidirectionally encoded and injected through inter-block cross-attention and aligned source features within each predicted block. At each step, the AR predictor produces next-block logits $\mathbf{z}_i = f_\theta(S, P_{<i}, i)$, where f_θ denotes the shared source-conditioned blockwise AR predictor, conditioned on the source S, causal target prefix $P_{<i}$, and current block index i. Spatial tokens within each block are predicted in parallel. Further architectural details are provided in Appendix B.3.

3.2 Source-prefix Balance via Prefix Retraction

To make the source–prefix balance explicit, we reformulate each prediction as a source-only prediction plus a controllable prefix-induced correction. To isolate the contribution of the target prefix, we evaluate the same AR predictor at the same target position with and without the prefix:

$$\mathbf{z}_i^S = f_\theta(S, \varnothing, i), \qquad \mathbf{z}_i^F = f_\theta(S, P_{<i}, i), \tag{2}$$

where $\mathbf{z}_i^S$ and $\mathbf{z}_i^F$ denote the *source-only* and *full-prefix* predictions, respectively. In implementation, $\varnothing$ is realized by masking all target-prefix blocks as invalid, so that no preceding target blocks are visible to the predictor. The two paths share all model parameters and differ only in whether the prefix is visible. Their difference, $\mathbf{r}_i = \mathbf{z}_i^F - \mathbf{z}_i^S$, defines the *prefix correction*, i.e., the logit-space update introduced by the prefix beyond the source. The correction therefore isolates the contribution of evolving target-side context beyond the fixed source condition.

We control how much of this correction is retained through

$$\mathbf{z}_i(s) = \mathbf{z}_i^F - s\mathbf{r}_i = \mathbf{z}_i^S + (1-s)\mathbf{r}_i, \qquad s \in [0, 1], \tag{3}$$

where $\mathbf{z}_i(s)$ is the *retracted prediction*, and s is the fraction of the correction being removed, i.e., the *retraction amount*. Thus, $s = 0$ retains the full prefix correction, $s = 1$ returns to the source-only prediction, and intermediate values partially retract it. Within this guidance-style interpolation, the remaining question is *how to determine an appropriate retraction amount s for each target region*.

3.3 Learning Target-aware Prefix Retraction

To determine this retraction amount, we first characterize when retraction is safe, and then learn to reproduce the corresponding target-aware decision without requiring the GT target at inference.

When is retraction safe? Let $\mathbf{p}_i^F = \mathrm{softmax}(\mathbf{z}_i^F)$ and $\mathbf{p}_i(s) = \mathrm{softmax}(\mathbf{z}_i(s))$. We define the predictive *effect* as $E_i(s) = D_{\mathrm{KL}}(\mathbf{p}_i^F \| \mathbf{p}_i(s))$, which measures how much retraction changes the prediction without using the GT target. To determine whether a correction is useful to the target, we measure the loss change caused by retraction, which we call *retraction regret*:

$$R_{i,y}(s) = \ell_y(\mathbf{z}_i(s)) - \ell_y(\mathbf{z}_i^F), \tag{4}$$

where $\ell_y(\mathbf{z}) = -\log \mathrm{softmax}(\mathbf{z})_y$. A positive regret means that retraction removes useful target information, whereas $R_{i,y}(s) \leq 0$ means that the correction can be retracted without increasing target loss, i.e., safe retraction. Importantly, the regret decomposes exactly as

$$R_{i,y}(s) = E_i(s) + sA_i, \qquad A_i = r_{i,y} - \mathbb{E}_{k \sim p_i^F}[r_{i,k}]. \tag{5}$$

Here, $A_i = R'_{i,y}(0)$ captures the target-relative alignment of the correction. Safe retraction therefore requires sufficiently negative alignment to offset the predictive effect, so effect alone cannot determine how much correction to retain.

Training oracle. Using the GT target during training, we define the *maximal zero-regret retraction* as

$$s_i^* = \max\{s \in [0,1] \mid R_{i,y}(s) \leq 0\}. \tag{6}$$

It removes as much correction as possible without increasing the full-prefix target loss at the current prediction state, defining a local zero-regret criterion for the oracle. Since $R_{i,y}(s)$ is convex, s_i^* is efficiently obtained by fixed-step bisection (Appendix B.3). Unlike a minimum-loss oracle, it favors the largest zero-regret retraction rather than the lowest-loss point along the path.

Correction-aware gate. Because the oracle requires the GT target, it is unavailable during generation. We therefore use a lightweight MLP gate to predict $\hat{s}_i = g_\phi(\mathbf{u}_i^S, \mathbf{u}_i^r, w_i, e_i) \in [0,1]$, where $\mathbf{u}_i^S$ is the normalized source-only hidden feature and $\mathbf{u}_i^r$ is the normalized hidden residual between the full-prefix and source-only paths, while w_i and e_i summarize the difference between the two prediction endpoints. Unlike the AR predictor trained for next-block prediction, the gate learns how much of the current source-to-prefix correction should be retained.

Gate supervision. Rather than regressing the oracle scalar s_i^* itself, we supervise the gate by matching the predictive distribution induced by the oracle retraction. Let $\mathbf{p}_i^* = \mathrm{softmax}(\mathbf{z}_i(s_i^*))$ and $\hat{\mathbf{p}}_i = \mathrm{softmax}(\mathbf{z}_i(\hat{s}_i))$. We minimize $\mathcal{L}_{\mathrm{ret}} = \mathbb{E}_{i \in \mathcal{V}}[D_{\mathrm{KL}}(\mathbf{p}_i^* \,\|\, \hat{\mathbf{p}}_i)]$, which we call predictive-action distillation, distilling the oracle distribution along the source–full-prefix retraction path. Here, the KL averages conditional divergences over token factors under a shared teacher-forced history.

3.4 Training and Inference

Training. SPIRA is trained in a single stage with $\mathcal{L} = \frac{1}{2}(\mathcal{L}_{\mathrm{full}} + \mathcal{L}_{\mathrm{source}}) + \mathcal{L}_{\mathrm{ret}}$, where $\mathcal{L}_{\mathrm{full}}$ and $\mathcal{L}_{\mathrm{source}}$ are standard next-block prediction losses for the shared AR predictor, while $\mathcal{L}_{\mathrm{ret}}$ updates only the gate with detached oracle and endpoint quantities.

Inference. At inference, the gate predicts $\hat{s}_i$ from source and prefix-derived features without GT or oracle supervision. Source-only pre-head contexts are precomputed and cached, while factor-conditioned logits are evaluated on demand. Setting $\hat{s}_i = 0$ throughout recovers Full-prefix AR.

4 Experiments

In this section, we conduct experiments to answer the following four research questions (RQs):

(RQ1) Effectiveness: Does SPIRA improve translation performance? §4.2
(RQ2) Design: Are SPIRA's key design choices necessary for its performance gains? §4.3
(RQ3) Utility-Effect: What explains the utility–effect gap? §4.4
(RQ4) Actionability: Can SPIRA translate the utility–effect gap into generation gains? §4.5

4.1 Experimental Setup

Datasets. We evaluate SPIRA on three datasets. BraTS 2024 (de Verdier et al., 2024) provides brain MRI with four modalities: T1c, T1n, T2w, and T2f. SynthRAD 2025 (Thummerer et al., 2025) covers

Table 1: **Quantitative results for 3D MRI translation on BraTS 2024.**

Methods	T1n→T1c				T1c→T1n				T2w→T2f				T2f→T1c			
	PSNR↑	SSIM↑	LPIPS↓	FID↓	PSNR↑	SSIM↑	LPIPS↓	FID↓	PSNR↑	SSIM↑	LPIPS↓	FID↓	PSNR↑	SSIM↑	LPIPS↓	FID↓
GAN-based Methods																
3D-mADUNet (Ha et al., 2025)	21.30	0.680	0.162	47.6	24.68	0.822	0.096	37.0	20.84	0.646	0.177	64.3	20.71	0.653	0.178	61.0
MMHVAE (Dorent et al., 2026)	21.66	0.691	0.157	45.2	24.23	0.819	0.097	39.2	20.67	0.643	0.182	61.9	20.90	0.662	0.170	54.6
Diffusion-based Methods																
M2DN (Meng et al., 2024)	21.15	0.673	0.161	47.8	24.35	0.816	0.099	40.7	20.83	0.648	0.179	62.8	20.64	0.651	0.184	59.7
selfRDB (Arslan et al., 2025)	22.03	0.695	0.139	44.8	25.16	0.834	0.082	34.0	21.42	0.674	0.159	46.9	21.33	0.677	0.160	47.0
CRSG (Song et al., 2026b)	21.85	0.691	0.148	46.1	25.31	0.834	0.085	33.6	21.19	0.670	0.166	54.8	21.02	0.669	0.171	53.9
Source-only	21.59	0.690	0.151	46.1	24.96	0.828	0.091	34.7	21.01	0.670	0.164	57.2	21.09	0.671	0.163	49.6
Full-prefix AR	22.10	0.696	0.135	41.9	25.49	0.834	0.085	32.2	21.30	0.674	0.155	48.6	21.55	0.681	0.154	46.1
SPIRA	**22.72**	**0.705**	**0.126**	**38.3**	**25.93**	**0.842**	**0.077**	**28.6**	**21.77**	**0.680**	**0.150**	**43.8**	**22.08**	**0.687**	**0.149**	**41.5**
Δ vs. Best External	**↑ 0.69**	**↑ 0.010**	**↓ 0.013**	**↓ 6.5**	**↑ 0.62**	**↑ 0.008**	**↓ 0.005**	**↓ 5.0**	**↑ 0.35**	**↑ 0.006**	**↓ 0.009**	**↓ 3.1**	**↑ 0.75**	**↑ 0.010**	**↓ 0.011**	**↓ 5.5**

Table 2: **Quantitative results for 3D medical image translation on SynthRAD 2025 and AutoPET.**

Methods	SynthRAD 2025								AutoPET							
	MRI→CT				CBCT→CT				CT→PET				PET→CT			
	PSNR↑	SSIM↑	LPIPS↓	FID↓	PSNR↑	SSIM↑	LPIPS↓	FID↓	PSNR↑	SSIM↑	LPIPS↓	FID↓	PSNR↑	SSIM↑	LPIPS↓	FID↓
GAN-based Methods																
3D-mADUNet (Ha et al., 2025)	20.78	0.701	0.238	95.7	22.94	0.737	0.128	43.2	24.73	0.902	0.130	56.0	23.49	0.896	0.105	91.3
MMHVAE (Dorent et al., 2026)	20.49	0.697	0.249	101.5	22.73	0.736	0.131	46.6	25.57	0.914	0.106	53.1	23.71	0.899	0.103	89.4
Diffusion-based Methods																
M2DN (Meng et al., 2024)	20.55	0.701	0.242	97.5	22.66	0.734	0.133	43.3	25.69	0.912	0.104	52.4	23.81	0.902	0.095	86.0
selfRDB (Arslan et al., 2025)	20.81	0.708	0.236	93.6	23.19	0.740	0.127	39.2	26.08	0.923	0.094	51.1	23.67	0.901	0.099	84.6
CRSG (Song et al., 2026b)	20.62	0.694	0.248	97.3	22.78	0.732	0.136	42.4	26.85	0.937	0.080	51.7	23.98	0.903	0.095	78.2
Source-only	20.27	0.691	0.252	98.8	23.02	0.739	0.129	40.5	27.30	0.939	0.078	49.5	23.58	0.901	0.097	81.9
Full-prefix AR	20.73	0.705	0.232	93.2	23.35	0.744	0.123	37.6	27.93	0.947	0.066	45.7	24.15	0.905	0.093	75.7
SPIRA	**21.29**	**0.714**	**0.225**	**87.3**	**23.59**	**0.751**	**0.119**	**35.6**	**28.44**	**0.956**	**0.058**	**44.6**	**24.76**	**0.912**	**0.086**	**69.5**
Δ vs. Best External	**↑ 0.48**	**↑ 0.006**	**↓ 0.011**	**↓ 6.3**	**↑ 0.40**	**↑ 0.011**	**↓ 0.008**	**↓ 3.6**	**↑ 1.59**	**↑ 0.019**	**↓ 0.022**	**↓ 6.5**	**↑ 0.78**	**↑ 0.009**	**↓ 0.009**	**↓ 8.7**

head-and-neck and abdominal regions with paired MRI/CT and CBCT/CT data for MRI→CT and CBCT→CT translation. AutoPET (Dexl et al., 2026) provides paired PET/CT scans across the brain, pelvis, and abdomen. Further dataset details are provided in Appendix A.

Implementation Details. We use a fine-tuned non-causal VidTok-FSQ (Tang et al., 2024) as the visual tokenizer, with a compression ratio of $(8 \times 8 \times 4)$. Our AR backbone is adapted from LlamaGen (Sun et al., 2024), with 3D patchification and 3D rotary positional embeddings to support volumetric generation. All models are trained on $(256 \times 256 \times 128)$ 3D crops. During inference, we use either full-volume or sliding-window generation, with KV caching to improve efficiency. Additional training details and hyperparameter settings are presented in Appendix B.4.

Comparison Methods. We compare SPIRA with representative GAN- and diffusion-based baselines, together with two matched internal baselines from the same trained model. Full-prefix AR provides the most controlled AR comparison by sharing the same tokenizer and architecture with SPIRA. ***Source-only*** removes the target prefix ($s = 1$), while ***Full-prefix AR*** disables retraction ($s = 0$) and uses the complete generated prefix. Together, they define the two endpoints of the retraction path and isolate whether SPIRA's gains arise from selective prefix retraction.

4.2 Comparison with State-of-the-Art (RQ1)

Quantitative Results. As reported in Tables 1 and 2, SPIRA consistently achieves the best performance across all tasks and metrics, outperforming both representative GAN- and diffusion-based methods and the matched internal baselines. In particular, SPIRA consistently improves upon Full-prefix AR across all tasks and metrics, demonstrating the effectiveness of our prefix retraction strategy. Please refer to Appendix D for annotation-defined tumor/lesion ROI evaluation, statistical significance, and downstream segmentation experiments.

Table 3: **Ablation studies** on BraTS 2024 T1n→T1c and SynthRAD 2025 MRI→CT.

Variants	BraTS T1n→T1c				SynthRAD MRI→CT			
	PSNR↑	SSIM↑	LPIPS↓	FID↓	PSNR↑	SSIM↑	LPIPS↓	FID↓
Reference Model								
Standard AR	22.16	0.696	0.137	42.5	20.89	0.705	0.233	89.8
Retraction Strategy								
Tuned global scalar s	22.28	0.699	0.133	42.8	21.02	0.709	0.230	91.6
Position-only gate	22.45	0.698	0.131	41.0	21.13	0.706	0.228	89.1
Effect-only gate	22.31	0.700	0.131	40.3	20.82	0.705	0.230	90.7
Gate Design								
Direct s^* regression	22.60	0.701	0.129	39.9	21.04	0.707	0.230	89.3
SPIRA	**22.72**	**0.705**	**0.126**	**38.3**	**21.29**	**0.714**	**0.225**	**87.3**

Qualitative Comparison. Fig. 4 provides illustrative cases in which SPIRA shows lower local residual errors while maintaining target-specific appearance relative to the baselines.

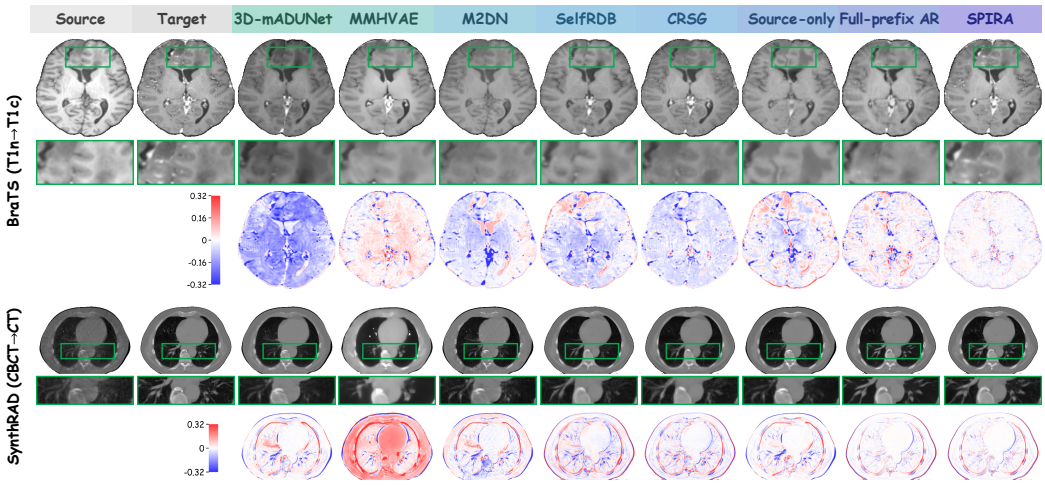

Figure 4: **Qualitative comparison** on BraTS T1n→T1c and SynthRAD CBCT→CT cases.

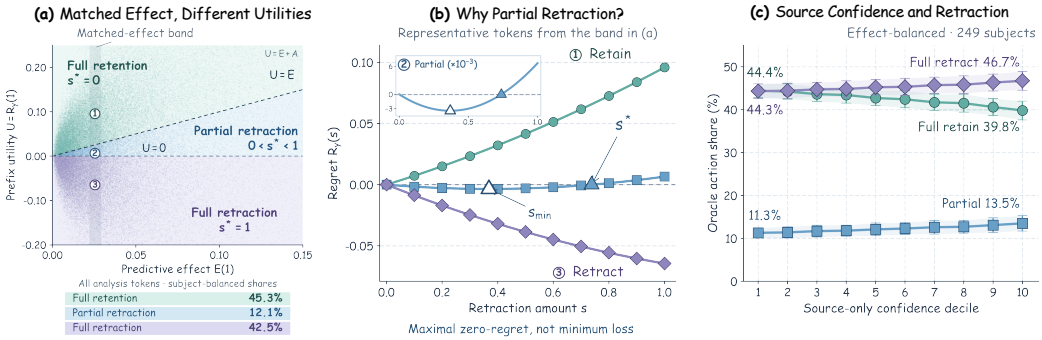

Figure 5: **Explaining the utility–effect gap.** (a) Effect–utility geometry separates three oracle regimes. (b) Matched-effect regret paths distinguish maximal zero-regret from minimum-loss retraction. (c) Source-confidence groups exhibit a modest shift in oracle regime shares after effect balancing. Shares in (c) are effect-standardized, whereas (a) reports marginal subject-balanced shares.

4.3 Ablation Studies (RQ2)

We compare SPIRA with standard AR, global retraction, position-only and effect-only gates, and direct s^* regression (Appendix D.6). Table 3 shows that joint training or effect alone does not explain the gain, while SPIRA performs best. Global amplification also falls below Full-prefix AR on both tasks (Table 15), while token-wise retraction outperforms coarser sharing (Table 16).

4.4 Explaining the Utility–Effect Gap in Source–Prefix Balance (RQ3)

Sec. 2 reveals that corrections with similar predictive effects can have different utilities. We now explain this gap on the frozen dual-endpoint predictor under teacher forcing, using the GT prefix and no gate, with effect measured at the token level by the endpoint KL $E_i(1)$.

Why can matched effects have different utilities? Let $U_i = R_{i,y}(1)$ denote prefix utility, with $U_i > 0$ indicating that the prefix lowers target loss. By Eq. 5, $U_i = E_i(1) + A_i$, where $A_i = r_{i,y} - \mathbb{E}_{k\sim p_i^F}[r_{i,k}]$ measures target alignment; $A_i > 0$ means that the correction favors the true target relative to the full-prefix belief. For $E_i(1) > 0$, convexity of $R_{i,y}$ implies that $A_i \geq 0$ admits no safe positive retraction, whereas $A_i < 0$ reduces target loss. This gives full retention for $U_i \geq E_i(1)$, partial retraction for $0 < U_i < E_i(1)$, and full retraction for $U_i \leq 0$. Fig. 5(a) reports corresponding shares of 45.3%, 12.1%, and 42.5%; the heterogeneity persists across effect levels in Fig. 9(a).

Why partially retract a useful correction? Fig. 5(b) traces corrections from the matched-effect band. In the partial regime, maximal zero-regret preserves the Full-prefix target loss while retaining less correction than minimum-loss. At inference, generated-prefix corrections may carry errors,

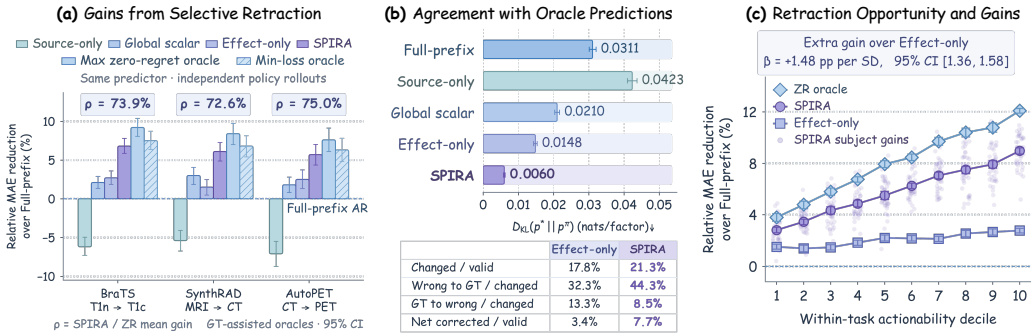

Figure 6: **From selective retraction to generation gains.** (a) Free-running gains of different retraction policies, where ρ denotes the zero-regret oracle gain recovered by SPIRA. (b) Oracle-distribution KL and prediction changes on identical SPIRA-generated states. (c) Generation gains across baseline-actionability deciles, with β measuring SPIRA's extra gain over effect-only.

whose influence is also scaled by $1 - s$ in Eq. 3. This predicts slower error growth with stronger safe retraction, a prediction we test in Sec. 4.5 and Fig. 10(c).

Is this variation associated with source evidence? After effect balancing, higher source-only confidence only slightly reduces the full-retention share, from 44.4% to 39.8% (Fig. 5(c)). Combining the GT-free source and correction features predicts oracle choices much better than confidence or effect alone (macro AUROC 0.80–0.85; Fig. 10(a)).

Takeaway

Target alignment explains why corrections with similar effects can favor different source–prefix balances, and even useful prefix corrections do not always need to be retained in full.

4.5 From Utility–Effect Gap to Performance Gains (RQ4)

Following the teacher-forced analysis, we now examine whether adaptively balancing source and prefix information through selective retraction improves free-running generation on three tasks.

Can Selective Retraction Improve Generation? In Fig. 6(a), SPIRA outperforms source-only, global scalar, and effect-only, reducing MAE over Full-prefix AR by 6.8%, 6.1%, and 5.7% on BraTS, SynthRAD, and AutoPET (6.2% on average), while recovering 72.6%–75.0% of the maximal zero-regret oracle gain. We further compare this oracle with a minimum-loss oracle that chooses the lowest-loss action at each current state (Appendices B.4 and D.6). Although minimum-loss achieves lower local loss, maximal zero-regret yields better free-running generation with slower error growth (Fig. 10(b,c)); its distilled gate also performs better (Table 15).

Does SPIRA Recover Oracle Predictions? On identical rollout states, SPIRA better matches the oracle than effect-only, with a KL divergence of 0.0060 versus 0.0148 nats per factor (Fig. 6(b)). We define prediction changes as factor-wise argmax differences from the unretracted prediction at the same state. Corrected errors minus newly introduced errors account for 7.7% of valid factor decisions for SPIRA, compared with 3.4% for effect-only. Since the gate is trained under teacher forcing, this agreement also suggests that its learned decisions transfer to generated-prefix states.

Do More Actionable Cases Benefit More? We measure baseline actionability on Full-prefix rollout states as the fraction of endpoint predictive effect that can be safely retracted by the GT-assisted oracle. Higher actionability therefore means that the current-state oracle can retract more of the prefix correction without increasing local target loss. In Fig. 6(c), from the lowest to highest actionability decile, SPIRA's gain rises from 2.81% to 8.97%, while effect-only changes only from 1.51% to 2.77%, widening SPIRA's extra gain from 1.30 to 6.20 percentage points. Consistently, this extra gain increases by 1.48 percentage points per standard deviation of actionability.

Fig. 7 illustrates two characteristic source–prefix balancing behaviors. In the retract-dominant case (top), stronger retraction *removes harmful prefix corrections* and reduces local reconstruction errors. In the retain-dominant case (bottom), weaker retraction *preserves target-useful prefix information* where source-only prediction misses target-specific details. Together, the two cases illustrate that SPIRA adaptively retracts or retains prefix corrections according to local prediction needs.

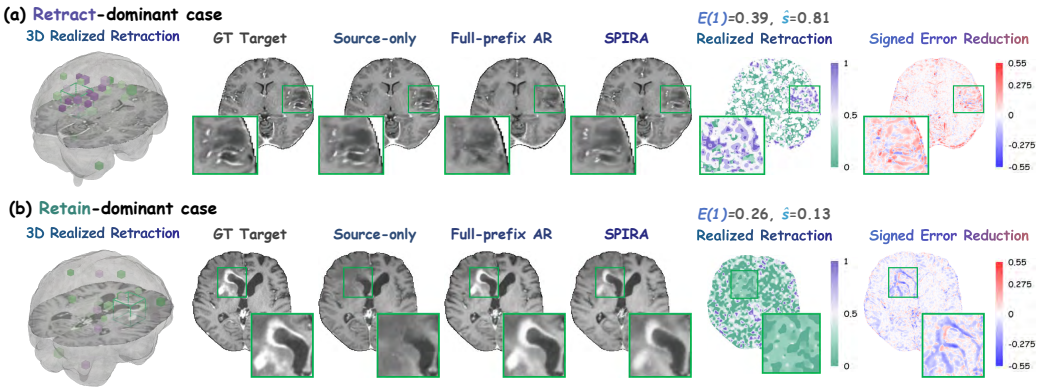

Figure 7: **Visualization of selective prefix retraction.** Retract-dominant (top) and retain-dominant (bottom) cases illustrate how SPIRA *retracts harmful corrections* while *retaining useful ones*. $E(1)$ and $\hat{s}$ are region averages, with *green* indicating stronger retention and *purple* stronger retraction.

Takeaway

SPIRA improves generation by *retaining target-useful prefix information* while *removing harmful corrections*. SPIRA better matches oracle decisions, with larger gains when more correction can be retracted.

5 Related Work

Medical Image Translation. Medical image translation has been widely studied with GANs (Ha et al., 2025; Dorent et al., 2026) and diffusion models (Pang et al., 2026; Liu et al., 2026; Wei et al., 2025). Most methods perform direct or iterative source-conditioned generation (Zhan et al., 2024; Arslan et al., 2025; Zhu et al., 2025; Wu et al., 2026), whereas AR translation exposes an evolving target prefix (Liu et al., 2025; Huang et al., 2025; Jin et al., 2026). SPIRA studies its region-wise interaction with the fixed source condition and learns how much prefix correction to retain.

Autoregressive Visual Generation. Autoregressive visual generation has progressed from pixel-level to token-based modeling (Tian et al., 2024; Ren et al., 2025b; Wang et al., 2024), and has recently expanded to videos and structured visual data (Huang et al., 2025; Yin et al., 2025; Ren et al., 2025a; Yang et al., 2026) as well as 3D settings (He et al., 2026; Feng et al., 2026; Chen et al., 2025b). Most existing work focuses on visual representation and prediction, while the interaction between fixed conditioning and evolving AR context is underexplored (Chen et al., 2026; Zhang et al., 2025; Gu et al., 2025). For conditional translation, where the source and target prefix provide complementary information (Mu et al., 2025), SPIRA learns how much of each prefix correction to retain region-wise.

Guidance and Context Control in Generative Models. Classifier-free guidance contrasts conditional and unconditional predictions (Ho & Salimans, 2022; Shen et al., 2024; Koulischer et al., 2025), while contrastive and context-aware decoding adjust generation through prediction contrasts (Li et al., 2023; Shi et al., 2024; Chuang et al., 2024). Some recent approaches further incorporate confidence-based routing or extrapolation (Tao & Agrawal, 2026; Liao et al., 2026), whereas SPIRA learns continuous token-wise prefix retraction by distilling a target-aware zero-regret oracle.

6 Conclusion

In this work, we study source-conditioned AR translation as a region-wise balance between fixed source evidence and evolving context. We propose SPIRA, which treats the target prefix as a controllable correction and learns to *reduce harmful or unnecessary corrections* while *preserving target-useful prefix information*. Across diverse 3D medical image translation tasks, SPIRA consistently improves performance. Our analyses link different retention needs at similar predictive effects to target alignment, while SPIRA better matches target-aware oracle predictions and gains more when greater retraction is possible. We hope SPIRA motivates broader consideration of how fixed conditioning and AR context should be balanced in conditional generation.

AI USE STATEMENT

Generative AI tools were used to aid and polish manuscript writing, support literature retrieval and discovery, including identifying relevant related work, and provide assistance during research ideation and experimental development. All AI-assisted suggestions and outputs were independently reviewed and verified by the authors. The research ideas, methodological design, experimental results, and scientific claims were developed and validated by the authors, who take full responsibility for the final content of the paper.

REFERENCES

Fuat Arslan, Bilal Kabas, Onat Dalmaz, Muzaffer Ozbey, and Tolga Çukur. Self-consistent recursive diffusion bridge for medical image translation. *Medical Image Analysis*, 106:103747, 2025.

Junxin Chen, Zhiheng Ye, Renlong Zhang, Hao Li, Bo Fang, Li bo Zhang, and Wei Wang. Medical image translation with deep learning: Advances, datasets and perspectives. *Medical Image Analysis*, 103:103605, 2025a.

Minghao Chen, Jianyuan Wang, Roman Shapovalov, Tom Monnier, Hyunyoung Jung, Dilin Wang, Rakesh Ranjan, Iro Laina, and Andrea Vedaldi. Autopartgen: Autoregressive 3d part generation and discovery. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 153496–153521, 2025b.

Shuo Chen, Cong Wei, Sun Sun, Tiancheng SHEN, Ping Nie, Kai Zou, Ge Zhang, Ming-Hsuan Yang, and Wenhu Chen. Context forcing: Consistent autoregressive video generation with long context. In *Forty-third International Conference on Machine Learning*, 2026.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R Glass, and Pengcheng He. Dola: Decoding by contrasting layers improves factuality in large language models. In *International Conference on Learning Representations*, volume 2024, pp. 54158–54183, 2024.

Maria Correia de Verdier, Rachit Saluja, Louis Gagnon, Dominic LaBella, Ujjwall Baid, Nourel Hoda Tahon, Martha Foltyn-Dumitru, Jikai Zhang, Maram Alafif, Saif Baig, et al. The 2024 brain tumor segmentation (brats) challenge: Glioma segmentation on post-treatment mri. *arXiv preprint arXiv:2405.18368*, 2024.

Jakob Dexl, Katharina Jeblick, Andreas Mittermeier, Balthasar Schachtner, Anna Theresa Stüber, Johanna Topalis, Maximilian Rokuss, Fabian Isensee, Klaus H Maier-Hein, Hamza Kalisch, et al. The autopet3 challenge–automated lesion segmentation in whole-body pet/ct-multitracer multicenter generalization. *arXiv preprint arXiv:2605.05775*, 2026.

Reuben Dorent, Nazim Haouchine, Alexandra Golby, Sarah Frisken, Tina Kapur, and William Wells. Unified cross-modal medical image synthesis with hierarchical mixture of product-of-experts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(2):1641–1656, 2026.

Chenhe Du, Qing Wu, Xuanyu Tian, Jingyi Yu, Hongjiang Wei, and Yuyao Zhang. Improving 2d diffusion models for 3d medical imaging with inter-slice consistent stochasticity. In *The Fourteenth International Conference on Learning Representations*, 2026.

Haoran Feng, Yifan Niu, Zehuan Huang, Yang-Tian Sun, Chunchao Guo, Yuxin Peng, and Lu Sheng. Repurposing 3d generative model for autoregressive layout generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3231–3243, 2026.

Yuchao Gu, Weijia Mao, and Mike Zheng Shou. Long-context autoregressive video modeling with next-frame prediction. *arXiv preprint arXiv:2503.19325*, 2025.

Juhyung Ha, Jong Sung Park, David Crandall, Eleftherios Garyfallidis, and Xuhong Zhang. Multi-resolution guided 3d gans for medical image translation. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 4342–4351, 2025.

Zhicheng He, Yunpeng Zhao, Junde Wu, Ziwei Niu, Zijun Li, Bohan Li, Lanfen Lin, and Yueming Jin. Medvar: Towards scalable and efficient medical image generation via next-scale autoregressive prediction. *arXiv preprint arXiv:2602.14512*, 2026.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 167283–167308, 2025.

Dongyang Jin, Ryan Xu, Jianhao Zeng, Rui Lan, Yancheng Bai, Lei Sun, and Xiangxiang Chu. Semantic context matters: Improving conditioning for autoregressive models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 30402–30413, 2026.

Felix Koulischer, Florian Handke, Johannes Deleu, Thomas Demeester, and Luca Ambrogioni. Feedback guidance of diffusion models. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 80482–80513, 2025.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12286–12312, 2023.

Xinyao Liao, Qiyuan He, Yicong Li, Jiayin Zhu, Xiaoye Qu, Wei Wei, and Angela Yao. Vpg: Visual prefix guidance for autoregressive image and video generation. *arXiv preprint arXiv:2605.30317*, 2026.

Junkai Liu, Nay Aung, Theodoros N. Arvanitis, Joao A. C. Lima, Steffen E. Petersen, and Le Zhang. Exploiting completeness perception with diffusion transformer for unified 3d MRI synthesis. *Transactions on Machine Learning Research*, 2026.

Yi Liu, Shengqian Li, Zuzeng Lin, Feng Wang, and Si Liu. Cyclevar: Repurposing autoregressive model for unsupervised one-step image translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15194–15204, 2025.

Qingyang Mao, Qi Cai, Yehao Li, Yingwei Pan, Mingyue Cheng, Ting Yao, Qi Liu, and Tao Mei. Visual autoregressive modeling for instruction-guided image editing. In *International Conference on Learning Representations*, volume 2026, pp. 85702–85723, 2026.

Xiangxi Meng, Kaicong Sun, Jun Xu, Xuming He, and Dinggang Shen. Multi-modal modality-masked diffusion network for brain mri synthesis with random modality missing. *IEEE Transactions on Medical Imaging*, 43(7):2587–2598, 2024.

Jiteng Mu, Nuno Vasconcelos, and Xiaolong Wang. Editar: Unified conditional generation with autoregressive models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7899–7909, 2025.

Haowen Pang, Xiaoming Hong, Peng Zhang, Pengli Zhu, Shun Yao, Fengping An, Guoyuan Yang, Tiantian Liu, Anqi Qiu, Chuyang Ye, and Tianyi Yan. Cascaded diffusion model and segment anything model for medical image synthesis. *Pattern Recognition*, 180:114148, 2026.

Shuhuai Ren, Shuming Ma, Xu Sun, and Furu Wei. Next block prediction: Video generation via semi-autoregressive modeling. *arXiv preprint arXiv:2502.07737*, 2025a.

Sucheng Ren, Qihang Yu, Ju He, Xiaohui Shen, Alan Yuille, and Liang-Chieh Chen. Beyond next-token: Next-x prediction for autoregressive visual generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 15781–15791, 2025b.

Giulia Romoli, Alessia Capoccia, Filippo Ruffini, Francesco Di Feola, Luca Boldrini, Arturo Chiti, Renato Cuocolo, Tugba Akinci D'Antonoli, Fatemeh Darvizeh, Marcello Di Pumpo, et al. Cross modality image translation in medical imaging using generative frameworks. *arXiv preprint arXiv:2605.13686*, 2026.

Dazhong Shen, Guanglu Song, Zeyue Xue, Fu-Yun Wang, and Yu Liu. Rethinking the spatial inconsistency in classifier-free diffusion guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9370–9379, 2024.

Weijia Shi, Xiaochuang Han, Mike Lewis, Yulia Tsvetkov, Luke Zettlemoyer, and Wen-tau Yih. Trusting your evidence: Hallucinate less with context-aware decoding. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pp. 783–791, 2024.

Yejee Shin, Yeeun Lee, Hanbyol Jang, Geonhui Son, Hyeongyu Kim, and Dosik Hwang. Anatomical consistency and adaptive prior-informed transformation for multi-contrast mr image synthesis via diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 30918–30927, 2025.

Tao Song, Yicheng Wu, Minhao Hu, Xiangde Luo, Linda Wei, Guotai Wang, Yi Guo, Feng Xu, and Shaoting Zhang. Learning modality-aware representations: Adaptive group-wise interaction network for multimodal mri synthesis. *IEEE Transactions on Medical Imaging*, 45(5):2306–2316, 2026a.

Zhiyun Song, Xin Wang, Honglin Xiong, Mengjun Liu, Manman Fei, Xiangyu Zhao, Zhenrong Shen, Sheng Wang, Dongdong Chen, Linlin Yao, Qian Wang, and Lichi Zhang. Enhancing generative models for modality imputation of 3-d mris via consistency-aware refinement and super-resolution guidance. *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–15, 2026b.

Peize Sun, Yi Jiang, Shoufa Chen, Shilong Zhang, Bingyue Peng, Ping Luo, and Zehuan Yuan. Autoregressive model beats diffusion: Llama for scalable image generation. *arXiv preprint arXiv:2406.06525*, 2024.

Anni Tang, Tianyu He, Junliang Guo, Xinle Cheng, Li Song, and Jiang Bian. Vidtok: A versatile and open-source video tokenizer. *arXiv preprint arXiv:2412.13061*, 2024.

Yufei Tao and Ameeta Agrawal. No-worse context-aware decoding: Preventing neutral regression in context-conditioned generation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 9736–9751, 2026.

Adrian Thummerer, Erik van der Bijl, Arthur Jr Galapon, Florian Kamp, Mark Savenije, Christina Muijs, Shafak Aluwini, Roel J. H. M. Steenbakkers, Stephanie Beuel, Martijn PW Intven, Johannes A. Langendijk, Stefan Both, Stefanie Corradini, Viktor Rogowski, Maarten Terpstra, Niklas Wahl, Christopher Kurz, Guillaume Landry, and Matteo Maspero. Synthrad2025 grand challenge dataset: Generating synthetic cts for radiotherapy from head to abdomen. *Medical Physics*, 52(7): e17981, 2025.

Keyu Tian, Yi Jiang, Zehuan Yuan, Bingyue Peng, and Liwei Wang. Visual autoregressive modeling: Scalable image generation via next-scale prediction. In *Advances in Neural Information Processing Systems*, volume 37, pp. 84839–84865, 2024.

Yulin Wang, Honglin Xiong, Kaicong Sun, Jiameng Liu, Xin Lin, Ziyi Chen, Yuanzhe He, Qian Wang, and Dinggang Shen. Unisyn: A generative foundation model for universal medical image synthesis across mri, ct and pet. In *International conference on medical image computing and computer-assisted intervention*, pp. 673–682, 2025.

Zili Wang, Zheng Zhang, Kun Ding, Qi Yang, Fei Li, and Shiming Xiang. Continuous speculative decoding for autoregressive image generation. *arXiv preprint arXiv:2411.11925*, 2024.

Jinbao Wei, Yuhang Chen, Zhijie Wang, Gang Yang, Shimin Tao, Jian Gao, Aiping Liu, and Xun Chen. Rethinking diffusion bridge model with dual alignments for medical image synthesis. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, pp. 1052–1061, 2025.

Yicheng Wu, Tao Song, Zhonghua Wu, Jin Ye, Zongyuan Ge, Wenjia Bai, Zhaolin Chen, and Jianfei Cai. Virtual full-stack scanning of brain mri via imputing any quantised code. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 21026–21035, 2026.

Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Ying-Cong Chen, Yao Lu, Song Han, and Yukang Chen. Longlive: Real-time interactive long video generation. In *The Fourteenth International Conference on Learning Representations*, 2026.

Tianwei Yin, Qiang Zhang, Richard Zhang, William T. Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22963–22974, 2025.

Chenlu Zhan, Yu Lin, Gaoang Wang, Hongwei Wang, and Jian Wu. Medm2g: Unifying medical multi-modal generation via cross-guided diffusion with visual invariant. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11502–11512, 2024.

Lvmin Zhang, Shengqu Cai, Muyang Li, Gordon Wetzstein, and Maneesh Agrawala. Frame context packing and drift prevention in next-frame-prediction video diffusion models. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 30546–30566, 2025. doi: 10.52202/085713-1024.

Jinze Zhao, Anqi Xiao, Jie Tian, Shimao Zhu, and Zhenhua Hu. Spid-dm: Structure-guided diffusion model with proportional-integral-derivative controller for unpaired medical image translation. *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2026.

Langtao Zhou, Xiaoxia Qu, Tianyu Fu, Jiaoyang Wu, Hong Song, Jingfan Fan, Danni Ai, Deqiang Xiao, Junfang Xian, and Jian Yang. Anatomy-aware sketch-guided latent diffusion model for orbital tumor multi-parametric mri missing modalities synthesis. *IEEE Transactions on Medical Imaging*, 45(5):2140–2155, 2026.

Hongzhou Zhu, Min Zhao, Guande He, Hang Su, Chongxuan Li, and Jun Zhu. Causal forcing: Autoregressive diffusion distillation done right for high-quality real-time interactive video generation. In *Forty-third International Conference on Machine Learning*, 2026.

Xiyue Zhu, Dou Hoon Kwark, Ruike Zhu, Kaiwen Hong, Yiqi Tao, Shirui Luo, Yudu Li, Zhi-Pei Liang, and Volodymyr Kindratenko. Introducing 3d representation for dense volume-to-volume translation via score fusion. In *Forty-second International Conference on Machine Learning*, 2025.

Muzaffer Özbey, Onat Dalmaz, Salman U. H. Dar, Hasan A. Bedel, Şaban Özturk, Alper Güngör, and Tolga Çukur. Unsupervised medical image translation with adversarial diffusion models. *IEEE Transactions on Medical Imaging*, 42(12):3524–3539, 2023.

Appendix Contents

A Datasets **15**

B Additional Implementation and Training Details **15**

B.1 Visual Tokenizer Fine-Tuning 15

B.2 Latent Pre-computation and Caching 16

B.3 Blockwise AR Architecture 16

B.4 Training and Inference Details 18

C Motivation Analysis Details **19**

D Additional Experimental Results **20**

D.1 Tumor- and Lesion-ROI Reconstruction Evaluation 20

D.2 Statistical Significance of Main Results 21

D.3 Multi-seed Robustness Analysis 21

D.4 Downstream Applicability to Tumor Segmentation 21

D.5 Additional Qualitative Results 22

D.6 More Ablation Studies 22

E Further Analysis **26**

E.1 Additional Utility–Effect Analysis 26

E.2 Additional Analysis of Selective Retraction 27

E.3 Computational Efficiency Analysis 29

F Discussion **30**

APPENDIX

A DATASETS

BraTS 2024. We use three cohorts from the BraTS 2024 (de Verdier et al., 2024) collection, including post-treatment adult glioma (GLI), brain metastases (MET), and pediatric brain tumors (PED). Each case provides four spatially aligned MRI modalities, T1n, T1c, T2w, and T2f, together with tumor annotations. Tumor annotations are used only for ROI and downstream evaluation. We split the subjects into training and validation sets at a 9:1 ratio.

SynthRAD 2025. SynthRAD 2025 (Thummerer et al., 2025) is a multi-center radiotherapy dataset containing paired MRI–CT and CBCT–CT scans across multiple anatomical regions. We use the head-and-neck and abdominal subsets. The released image pairs are spatially aligned through dataset preprocessing, providing paired 3D source and target volumes for translation. We follow the official training and validation splits.

AutoPET. AutoPET (Dexl et al., 2026) provides paired whole-body PET/CT examinations with tumor-lesion annotations and broad anatomical coverage. We use the corresponding CT and PET volumes to evaluate CT→PET and PET→CT translation, covering diverse anatomical regions including the brain, pelvis, and abdomen. Lesion annotations are reserved for ROI-based evaluation and are not provided to the translation model. We follow the official training and validation splits.

The dataset splits and the number of training and validation cases for each translation task are summarized in Table 4. For model selection, we reserve a subject-disjoint subset from the training partition, while the validation subjects in Table 4 are used only for final evaluation. All model and hyperparameter choices are fixed before evaluation.

Table 4: **Dataset composition and statistics** used in our experiments. The held-out evaluation subjects are not used for model training or selection.

Dataset	Translation Task	#Train	#Val
BraTS 2024 (de Verdier et al., 2024)	T1n→T1c	2,280	254
	T1c→T1n	2,280	254
	T2w→T2f	2,280	254
	T2f→T1c	2,280	254
SynthRAD 2025 (Thummerer et al., 2025)	MRI→CT	407	106
	CBCT→CT	622	152
AutoPET (Dexl et al., 2026)	CT→PET	1,279	323
	PET→CT	1,279	323

B ADDITIONAL IMPLEMENTATION AND TRAINING DETAILS

B.1 VISUAL TOKENIZER FINE-TUNING

We initialize the tokenizer from the pre-trained non-causal VidTok-FSQ-262144 checkpoint (Tang et al., 2024) and fine-tune one modality-shared tokenizer for each dataset, such that all source and target modalities within the same dataset share a common latent vocabulary.

To adapt VidTok to volumetric data, we treat each 3D volume as a video, with slices along the depth axis regarded as video frames such that the depth dimension corresponds to the temporal dimension. For tokenizer fine-tuning, we follow the same image preprocessing as in the main experiments. Grayscale medical volumes are repeated to three channels before VidTok encoding, while the three-channel decoder output is averaged back to a single-channel reconstruction. We retain the original $8 \times 8 \times 4$ compression ratio and FSQ-262144 quantization configuration, and optimize the original VidTok objective together with grayscale ℓ_1 and 3D SSIM reconstruction losses. We first adapt the decoder while keeping the encoder and FSQ module fixed, and subsequently jointly fine-tune the encoder and decoder with a smaller learning rate for the encoder. The complete fine-tuning configurations are summarized in Table 5.

Table 5: Fine-tuning configurations for the VidTok tokenizer.

Setting	Value
Base tokenizer	VidTok-FSQ-262144, non-causal
Compression ratio	$8 \times 8 \times 4$
Stage I	0–5k steps, decoder only
Stage II	5k–30k steps, encoder + decoder
Encoder learning rate	1×10^{-5}
Decoder learning rate	5×10^{-5}
Optimizer	AdamW
Weight decay	0.01
Batch size	4
Warm-up steps	1,000
Gradient clipping	1.0
Precision	bfloat16

Table 6: **Reconstruction performance** of the pre-trained and fine-tuned VidTok tokenizers. Averages are computed over all evaluated volumes before rounding.

Dataset	Modality	Pre-trained VidTok			Fine-tuned VidTok		
		PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
BraTS 2024	T1n	32.66	0.926	0.077	33.98 ↑ 1.32	0.950 ↑ 0.024	0.069 ↓ 0.008
	T1c	32.98	0.938	0.067	33.61 ↑ 0.63	0.940 ↑ 0.002	0.060 ↓ 0.007
	T2w	33.80	0.951	0.057	34.80 ↑ 1.00	0.960 ↑ 0.009	0.051 ↓ 0.006
	T2f	32.51	0.924	0.074	33.74 ↑ 1.23	0.941 ↑ 0.017	0.070 ↓ 0.004
	Avg	32.99	0.934	0.068	34.03 ↑ 1.04	0.948 ↑ 0.014	0.062 ↓ 0.006
SynthRAD 2025	MRI	28.16	0.790	0.122	34.05 ↑ 5.89	0.943 ↑ 0.153	0.092 ↓ 0.030
	CT	28.49	0.833	0.125	35.52 ↑ 7.03	0.958 ↑ 0.125	0.095 ↓ 0.030
	CBCT	29.38	0.807	0.147	35.72 ↑ 6.34	0.937 ↑ 0.130	0.108 ↓ 0.039
	Avg	28.69	0.816	0.131	35.28 ↑ 6.59	0.949 ↑ 0.133	0.098 ↓ 0.033
AutoPET	CT	30.68	0.885	0.101	35.36 ↑ 4.68	0.959 ↑ 0.074	0.081 ↓ 0.020
	PET	35.25	0.954	0.041	36.28 ↑ 1.03	0.971 ↑ 0.017	0.035 ↓ 0.006
	Avg	32.97	0.920	0.071	35.81 ↑ 2.84	0.965 ↑ 0.045	0.058 ↓ 0.013

After fine-tuning, the tokenizer is frozen throughout autoregressive training and inference, and Table 6 compares the reconstruction quality of the pre-trained and fine-tuned tokenizers for each modality.

B.2 Latent Pre-computation and Caching

Before AR training, we use the fine-tuned and frozen VidTok tokenizer (Tang et al., 2024) to pre-compute full-volume latent representations for all modalities, which are subsequently stored as local caches. Each modality is encoded and cached independently, while all modalities from the same case share identical preprocessing, spatial grids, padding, and patch coordinates to preserve voxel-wise correspondence. For variable-sized volumes, we employ overlapping tiled encoding with center-crop stitching to construct seamless full-volume latents without averaging the discrete FSQ scalar indices. Each cache stores the factorized FSQ scalar indices together with the corresponding latent valid mask, and autoregressive training directly samples fixed-size latent windows from these caches without repeatedly invoking the VidTok encoder.

B.3 Blockwise AR Architecture

We instantiate f_θ as a source-conditioned 3D blockwise Transformer adapted from LlamaGen (Sun et al., 2024). The spatially aligned source and target FSQ grids are partitioned into the same non-overlapping 3D blocks and traversed in a fixed depth–height–width raster order. The architecture is causal across blocks but bidirectional within each block, allowing all spatial tokens in the current target block to be predicted in parallel.

Source and prefix encoding. At each source position, embeddings of the factorized FSQ scalars are summed to form a token representation. A block-local bidirectional Transformer first contextualizes

tokens within each source block, followed by attention pooling into one representation per block. A global bidirectional Transformer with 3D coordinate embeddings then produces source block memory M^S. The resulting global block features are also projected back to the local tokens, yielding spatially aligned dense source features U^S.

Visible target blocks are embedded and encoded using block-local bidirectional attention, and their pooled representations form the inter-block sequence. A Llama-style causal Transformer processes this sequence using 3D rotary positional embeddings. Each layer applies

$$\begin{aligned} H &\leftarrow H + \mathrm{CausalAttn}\big(\mathrm{RMSNorm}(H); \mathrm{RoPE}_{3D}\big), \\ H &\leftarrow H + \mathrm{CrossAttn}\big(\mathrm{RMSNorm}(H), M^S\big), \\ H &\leftarrow H + \mathrm{SwiGLU}\big(\mathrm{RMSNorm}(H)\big), \end{aligned} \tag{7}$$

where source cross-attention uses a factorized relative 3D positional bias.

Next-block prediction. The first target block is predicted from a virtual source-conditioned start block, which is processed through the same intra- and inter-block modules as ordinary target history. For each subsequent block, a local bidirectional decoder combines the causal block context, the encoded preceding block, the aligned source features U_i^S, intra-block positional embeddings, and the 3D transition between consecutive blocks. This produces a context vector $c_{i,v}$ for each spatial position v in block i. The source-only and full-prefix predictions use exactly the same architecture and differ only in whether the preceding target blocks are visible.

Factorized FSQ prediction. Each latent token is represented by d low-cardinality FSQ factors $q_{i,v,1:d}$. Rather than predicting their Cartesian-product code with a single large classifier, the output head factorizes the token distribution as

$$p_\theta(Y_{i,v} \mid S, P_{<i}) = \prod_{j=1}^{d} p_\theta\big(q_{i,v,j} \mid c_{i,v}, q_{i,v,<j}\big). \tag{8}$$

The factors are generated sequentially within each token, whereas all spatial positions in the current block are processed in parallel.

Correction-aware gate inputs. For each valid spatial position, the source-only and full-prefix passes also provide pre-head context features $\mathbf{h}_{i,v}^S$ and $\mathbf{h}_{i,v}^F$. We construct the source and correction features as

$$\mathbf{u}_{i,v}^S = \mathrm{RMSNorm}\big(\mathrm{sg}(\mathbf{h}_{i,v}^S)\big), \qquad \mathbf{u}_{i,v}^r = \mathrm{RMSNorm}\big(\mathrm{sg}(\mathbf{h}_{i,v}^F - \mathbf{h}_{i,v}^S)\big), \tag{9}$$

where sg denotes stop-gradient. Therefore, $\mathbf{u}^S$ represents the source-only prediction state, while $\mathbf{u}^r$ explicitly captures the hidden-space change introduced by the current prefix.

We further provide two lightweight endpoint-difference statistics: $w_{i,v}$, the factor-averaged total-variation distance between the source-only and full-prefix predictive distributions, and $e_{i,v}$, the log-RMS magnitude of their logit residual. The gate concatenates $[\mathbf{u}_{i,v}^S, \mathbf{u}_{i,v}^r, w_{i,v}, e_{i,v}]$ and predicts the token-wise retraction amount $\hat{s}_{i,v}$ with a lightweight MLP. All gate inputs are available at inference and are detached from the shared AR predictor. The gate maps the concatenated $(2H+2)$-dimensional input through a 128-dimensional linear layer, SiLU, LayerNorm, and a scalar linear output layer followed by sigmoid, where H is the context feature dimension.

Oracle construction during training.

Under teacher forcing, the shared predictor produces source-only and full-prefix logits $z_{i,v,j}^S$ and $z_{i,v,j}^F$ for FSQ factor j at each valid position v. For a candidate retraction s, we form

$$z_{i,v,j}(s) = z_{i,v,j}^F - s\left(z_{i,v,j}^F - z_{i,v,j}^S\right), \tag{10}$$

and evaluate the factor-averaged GT token loss

$$\ell_{i,v}(s) = -\frac{1}{d}\sum_{j=1}^{d} \log\left[\mathrm{softmax}\left(z_{i,v,j}(s)\right)\right]_{q_{i,v,j}^{\mathrm{GT}}}. \tag{11}$$

At the full-prefix endpoint, $p_i(0) = p_i^F$, giving $E_i(0) = E_i'(0) = 0$. Differentiating Eq. 5 then yields $R_{i,y}'(0) = A_i$. The oracle target is the maximal zero-regret retraction

$$s_{i,v}^* = \max\left\{s \in [0,1] \mid \ell_{i,v}(s) \le \ell_{i,v}(0)\right\}, \tag{12}$$

Table 7: **Dataset- and modality-specific preprocessing** in our experiments. P_q denotes the qth percentile computed independently for each volume and modality, whereas P_q^+ is computed over positive-valued voxels only.

Dataset	**BraTS 2024**				**SynthRAD 2025**			**AutoPET**	
Modality	T1n	T1c	T2w	T2f	CT	CBCT	MRI	CT	PET
Orientation	RAS				RAS			RAS	
Voxel spacing	Adjusted by volume resizing				Native spacing			Native spacing	
Foreground threshold	$x > 0$				$x > -950$	$x > -950$	$x > 0$	$x > -950$	$x > 0$
Intensity clipping	$[P_{0.5}, P_{99.5}]$				$[-1000, 1000]$	$[-1000, 1000]$	$[P^+_{0.5}, P^+_{99.5}]$	$[-1000, 1000]$	$[P^+_{0.5}, P^+_{99.5}]$
Normalization	$[-1, 1]$				$[-1, 1]$			$[-1, 1]$	
Volume resizing	$256 \times 256 \times 128$				None; variable volume size			None; variable volume size	
Augmentation	Random rotation and flipping				Random rotation and flipping			Random rotation and flipping	

where $\ell_{i,v}(0)$ is the full-prefix loss. The oracle uses the feasibility criterion $\ell_{i,v}(s) \le \ell_{i,v}(0) + 10^{-6}$. We accept $s = 1$ when feasible; otherwise, we perform 10 bisection iterations on $[0, 1]$ and return the feasible lower bound. The resulting factorized distribution at $s^*_{i,v}$ is detached and used as the supervision target for the gate rather than directly regressing $s^*_{i,v}$. The exact architectural dimensions and training configurations are provided in Appendix B.4.

Factor-wise inference and distillation. For the AR FSQ head, token-wise gating is implemented in two passes. We first construct a deterministic greedy factor trajectory using the full-prefix endpoint and evaluate both endpoint heads along this shared trajectory to obtain the gate statistics. The gate then selects one retraction amount per token. Factor generation restarts from an empty history, with both endpoint heads recomputed sequentially using the factors actually generated under the selected action. The action remains fixed within the token, and only the final generated block is appended to the target prefix. Source-only pre-head contexts can be precomputed across blocks, while their conditional factor logits are evaluated using the corresponding within-token history. During teacher-forced training, both endpoints use the same ground-truth factor history when computing gate statistics, oracle actions, and distillation targets. The retraction objective averages $D_{\mathrm{KL}}(p_j^{s^*} \| p_j^s)$ equally over FSQ factors and then over valid tokens. It therefore matches conditional factor distributions at a fixed history, rather than complete autoregressive token distributions. The regret decomposition and maximal zero-regret criterion are defined on the same fixed-history, factor-averaged quantities.

B.4 Training and Inference Details

Preprocessing. Because the datasets contain different imaging modalities with distinct intensity characteristics and spatial configurations, we use a separate preprocessing pipeline for each dataset and modality, as summarized in Table 7. For BraTS 2024 (de Verdier et al., 2024), all volumes are resized to $256 \times 256 \times 128$. For SynthRAD 2025 (Thummerer et al., 2025) and AutoPET (Dexl et al., 2026), the native volume dimensions vary substantially across subjects, and we therefore do not resize the full volumes to a fixed spatial size. The resulting source and target volumes within each case are processed on the same spatial grid to preserve their correspondence.

Tokenization and AR setup. We use the fine-tuned non-causal VidTok-FSQ (Tang et al., 2024) described in Appendix B.1 as the visual tokenizer. Each 3D volume is treated as a video by regarding the depth axis as the temporal dimension. The tokenizer uses a volumetric compression ratio of $(8, 8, 4)$ in (H, W, D), corresponding to $(4, 8, 8)$ in VidTok's native (T, H, W) ordering. Before autoregressive training, we pre-compute the latent representation of each full volume and cache it locally, as detailed in Appendix B.2, avoiding repeated tokenizer encoding during training. The latent grids are partitioned into non-overlapping $4 \times 4 \times 4$ blocks and traversed using a fixed depth–height–width raster order. Dataset-specific model configurations are summarized in Table 8.

Optimization. We train the AR predictor for 100k optimization steps using AdamW with $(\beta_1, \beta_2) = (0.9, 0.95)$ and weight decay 0.05. The initial learning rate is 1×10^{-4} and follows a cosine decay schedule after warmup. We clip the global gradient norm to 2.0. For the gate parameter group, we use a learning rate three times that of the AR predictor, i.e., 3×10^{-4}, with weight decay 0.01. All experiments are conducted on a single NVIDIA GeForce RTX 5090 GPU. The remaining dataset-specific training configurations are provided in Table 8.

Table 8: **Model, training, and inference configurations** for the datasets used in our experiments.

Setting	BraTS 2024	SynthRAD 2025	AutoPET
Model Configurations			
Layers	12	6	8
Hidden dim	512	384	384
Attention heads	8	8	8
Parameters	77.86M	27.94M	32.66M
Training Settings			
Optimizer	AdamW	AdamW	AdamW
Learning rate	$1e{-4}$	$1e{-4}$	$1e{-4}$
Weight decay	0.05	0.05	0.05
(β_1, β_2)	$(0.9, 0.95)$	$(0.9, 0.95)$	$(0.9, 0.95)$
Learning rate schedule	Warmup cosine	Warmup cosine	Warmup cosine
Batch size	2	4	3
Training steps	100k	100k	100k
Gradient clipping	2.0	2.0	2.0
Precision	bfloat16	bfloat16	bfloat16
Training crop size	$256 \times 256 \times 128$	$256 \times 256 \times 128$	$256 \times 256 \times 128$
Latent window size	$32 \times 32 \times 32$	$32 \times 32 \times 32$	$32 \times 32 \times 32$
Block size	$4 \times 4 \times 4$	$4 \times 4 \times 4$	$4 \times 4 \times 4$
#AR blocks / window	512	512	512
Inference Settings			
Inference protocol	Full volume	Sliding window inference	Sliding window inference
Inference window size	$256 \times 256 \times 128$	$256 \times 256 \times 128$	$256 \times 256 \times 128$
Inference overlap	–	0.25	0.25

Inference. For whole-volume synthesis, we use a $256 \times 256 \times 128$ inference window. This window covers the full resized BraTS volume, whereas SynthRAD 2025 and AutoPET are generated using overlapping sliding windows. We adopt KV caching to avoid recomputing previously generated context, and the source-only endpoint required by SPIRA is also cached and reused during generation. Generation is deterministic, with each FSQ factor decoded by greedy argmax without temperature scaling or top-k/top-p sampling. Each predicted block is appended to the target prefix before generating the next block, following the same raster order used during training.

For all retraction policies, the source-only and full-prefix endpoints are evaluated at the same current generated state before applying the corresponding retraction. For the GT-assisted oracle analyses in Sec. 4.5, candidate retraction amounts are evaluated while keeping the source, generated prefix, and factor-conditioning history fixed; the GT token is used only to score candidate actions and is never appended to the prefix. The completed latent volume is finally decoded by the frozen VidTok decoder to obtain the synthesized 3D volume.

C Motivation Analysis Details

We provide additional details for the three analyses in Fig. 2. All analyses use the BraTS 2024 T1n→T1c held-out set and the same frozen AR predictor. No SPIRA gate or prefix retraction is applied.

Source-only and Full-prefix Predictions. For subject n and target unit i, let S_n denote the source volume and $P_{n,<i}$ the preceding GT target blocks. We construct

$$\mathbf{z}^S_{n,i} = f_\theta(S_n, \varnothing, i), \qquad \mathbf{z}^F_{n,i} = f_\theta(S_n, P_{n,<i}, i), \tag{13}$$

where $\mathbf{z}^S_{n,i}$ and $\mathbf{z}^F_{n,i}$ denote the source-only and full-prefix logits. The two paths use the same predictor and differ only in whether the preceding target prefix is visible.

The full-prefix path uses the GT prefix under teacher forcing to isolate prefix contributions from accumulated generation errors. We exclude padding, the first target unit with an empty prefix, and positions outside the valid nonzero brain support. For each FSQ token, probabilities are computed separately over its d categorical factors and then averaged equally across factors.

Average Prefix Benefit. Let $p^X_{n,i,j}$ be the predicted distribution for factor j under path $X \in \{S, F\}$, and let $y_{n,i,j}$ be its GT category. The subject-level prediction error is

$$\mathrm{Err}^X_n = \frac{1}{|\mathcal{V}_n|} \sum_{i \in \mathcal{V}_n} \left[-\frac{1}{d} \sum_{j=1}^{d} \log p^X_{n,i,j}(y_{n,i,j}) \right], \tag{14}$$

where $\mathcal{V}_n$ contains the valid target positions.

For each subject, we compute the relative error reduction as $100(\mathrm{Err}^S_n - \mathrm{Err}^F_n)/\,\mathrm{Err}^S_n$. The reported average error reduction is the equal-weight mean across subjects. The percentage of improved subjects is the fraction satisfying $\mathrm{Err}^F_n < \mathrm{Err}^S_n$.

Regional Benefit Heterogeneity. We decode the source-only and full-prefix predictions into volumes $\widehat{Y}^S_n$ and $\widehat{Y}^F_n$ using the frozen VidTok decoder. Each volume is divided into non-overlapping 3D regions matching the spatial footprint of one AR block, and regions with at least 50% valid brain voxels are retained.

For region r, we define the local error reduction and positive benefit as

$$d_{n,r} = \mathrm{MAE}^S_{n,r} - \mathrm{MAE}^F_{n,r}, \qquad b_{n,r} = n_{n,r} \max(d_{n,r}, 0), \tag{15}$$

where $n_{n,r}$ is the number of valid voxels. Positive $d_{n,r}$ indicates that the target prefix reduces local reconstruction error. After ranking regions by $d_{n,r}$, let $\mathcal{T}_n(q)$ denote the top fraction q. Their cumulative share of positive benefit is

$$C_n(q) = \frac{\sum_{r \in \mathcal{T}_n(q)} b_{n,r}}{\sum_r b_{n,r}}. \tag{16}$$

We report $C_n(0.20)$ and q_{50}, the smallest region fraction capturing 50% of the positive benefit, and summarize both by the subject-level median. The heatmap visualizes the positive voxel-wise reduction in absolute reconstruction error from source-only to full-prefix prediction.

Matched Prediction Changes and Target Outcomes. For each valid latent token, we measure the prefix-induced prediction change and its effect on the GT target as

$$\Delta^{\mathrm{pred}}_{n,i} = \frac{1}{d} \sum_{j=1}^{d} \frac{1}{2} \sum_c \left| p^F_{n,i,j}(c) - p^S_{n,i,j}(c) \right|, \tag{17}$$

$$\Delta^{\mathrm{GT}}_{n,i} = \frac{1}{d} \sum_{j=1}^{d} \left[\log p^F_{n,i,j}(y_{n,i,j}) - \log p^S_{n,i,j}(y_{n,i,j}) \right]. \tag{18}$$

Here, Δ^{pred} is the factor-averaged total variation distance, while Δ^{GT} measures the change in GT log-likelihood, with positive values indicating a more likely target.

The matched-change interval is defined using Δ^{pred} only. After removing the lowest 20% of near-zero changes, we use the same maximum number of tokens per subject and select the 45th–55th percentiles, yielding the fixed interval $[0.085, 0.096]$.

Using $\tau = 0.05$, matched tokens are categorized as *target more likely* if $\Delta^{\mathrm{GT}} > \tau$, *little change* if $|\Delta^{\mathrm{GT}}| \le \tau$, and *target less likely* if $\Delta^{\mathrm{GT}} < -\tau$. Proportions are computed per subject and then averaged equally across subjects.

For the coexistence analysis, subjects with at least 100 matched tokens are included. A subject is counted as containing both improving and worsening outcomes when each occupies at least 5% of its matched tokens, based on the sign of Δ^{GT}.

D Additional Experimental Results

D.1 Tumor- and Lesion-ROI Reconstruction Evaluation

We further evaluate reconstruction quality within annotation-defined tumor or lesion ROIs on BraTS 2024 (de Verdier et al., 2024) and AutoPET (Dexl et al., 2026), using the provided annotations

Table 9: **Tumor-ROI reconstruction evaluation on BraTS 2024.** We report PSNR and SSIM within annotation-defined tumor regions across four translation tasks.

Methods	T1n→T1c		T1c→T1n		T2w→T2f		T2f→T1c	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑	PSNR↑	SSIM↑
GAN-based Methods								
3D-mADUNet (Ha et al., 2025)	19.47 ±3.05	0.678 ±0.12	20.95 ±3.35	0.794 ±0.11	17.75 ±2.96	0.641 ±0.10	17.97 ±3.44	0.651 ±0.09
MMHVAE (Dorent et al., 2026)	19.73 ±2.76	0.685 ±0.09	21.25 ±2.94	0.791 ±0.12	17.69 ±3.28	0.641 ±0.13	18.00 ±2.79	0.662 ±0.10
Diffusion-based Methods								
M2DN (Meng et al., 2024)	19.59 ±3.12	0.682 ±0.13	21.01 ±2.74	0.792 ±0.10	17.88 ±3.51	0.645 ±0.07	17.92 ±3.22	0.649 ±0.08
selfRDB (Arslan et al., 2025)	19.86 ±3.23	0.692 ±0.10	21.12 ±3.11	0.795 ±0.07	18.15 ±2.98	0.662 ±0.09	18.40 ±2.70	0.670 ±0.11
CRSG (Song et al., 2026b)	20.07 ±2.89	0.692 ±0.08	21.57 ±2.69	0.801 ±0.09	18.03 ±2.72	0.665 ±0.09	18.24 ±2.85	0.667 ±0.11
Source-only	19.79 ±3.20	0.687 ±0.12	21.38 ±3.29	0.799 ±0.10	18.11 ±3.14	0.664 ±0.09	18.03 ±3.04	0.668 ±0.12
Full-prefix AR	20.03 ±3.40	0.690 ±0.09	21.50 ±3.07	0.800 ±0.11	**18.29 ±2.84**	**0.669 ±0.10**	18.21 ±2.90	0.674 ±0.09
SPIRA	**20.12 ±3.06**	**0.696 ±0.10**	**21.71 ±2.82**	**0.803 ±0.09**	18.25 ±3.36	0.667 ±0.12	**18.64 ±2.75**	**0.679 ±0.09**

Table 10: **Lesion-ROI reconstruction evaluation on AutoPET.** We report PSNR and SSIM within annotation-defined tumor-lesion regions.

Methods	CT→PET		PET→CT	
	PSNR↑	SSIM↑	PSNR↑	SSIM↑
GAN-based Methods				
3D-mADUNet (Ha et al., 2025)	16.76 ±1.39	0.887 ±0.04	17.64 ±1.19	0.779 ±0.03
MMHVAE (Dorent et al., 2026)	17.08 ±1.21	0.902 ±0.02	17.80 ±1.33	0.784 ±0.02
Diffusion-based Methods				
M2DN (Meng et al., 2024)	17.53 ±1.52	0.905 ±0.04	17.68 ±1.40	0.785 ±0.03
selfRDB (Arslan et al., 2025)	17.29 ±1.36	0.907 ±0.03	17.52 ±1.60	0.782 ±0.04
CRSG (Song et al., 2026b)	17.51 ±1.34	0.913 ±0.02	17.79 ±1.27	0.788 ±0.02
Source-only	17.26 ±1.50	0.910 ±0.02	17.94 ±1.44	0.791 ±0.03
Full-prefix AR	17.42 ±1.45	0.910 ±0.03	17.85 ±1.52	0.788 ±0.03
SPIRA	**17.79 ±1.28**	**0.917 ±0.02**	**18.22 ±1.30**	**0.792 ±0.04**

only for evaluation. As reported in Tables 9 and 10, SPIRA improves over Full-prefix AR on five of six task-level ROI comparisons and achieves the best results on all AutoPET ROI metrics; the small T2w→T2f difference is within the reported subject-level variability. The smaller ROI gains are consistent with the higher full-retention share in tumor regions (Fig. 9(c)), where less prefix correction can be safely retracted, in line with the actionability analysis in Sec. 4.5.

D.2 Statistical Significance of Main Results

Tables 11 and 12 report subject-level variability and paired statistical comparisons with SPIRA. Across all eight translation tasks, SPIRA consistently outperforms Full-prefix AR on PSNR, SSIM, and LPIPS. The paired tests, together with the improvements over external baselines, support the robustness of the main quantitative results.

D.3 Multi-seed Robustness Analysis

Table 13 reports results over three independent training runs on all four BraTS 2024 (de Verdier et al., 2024) translation tasks. SPIRA consistently achieves favorable three-seed mean performance across all metrics, with modest variation across seeds. The averaged results remain close to those in the main experiments, supporting the stability of selective source–prefix balancing under training randomness.

D.4 Downstream Applicability to Tumor Segmentation

For each BraTS 2024 (de Verdier et al., 2024) translation task, the target modality is treated as missing and synthesized by the corresponding translation model. A multimodal U-Net is trained using the three available real modalities together with the synthesized modality, and we report the subject-level mean tumor Dice. Table 14 shows that adding a synthesized modality consistently improves over the three-real-modality reference, while SPIRA achieves the best average Dice of 0.801 and leads

Table 11: **Subject-level statistical analysis on BraTS 2024.** We report the mean performance with subject-level standard deviations for four translation tasks. Asterisks indicate statistically significant differences between each competing method and SPIRA based on two-sided paired Wilcoxon signed-rank tests: $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$. Subscripts denote subject-level standard deviations.

Methods	T1n→T1c			T1c→T1n		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
GAN-based Methods						
3D-mADUNet (Ha et al., 2025)	$21.30^{***}_{\pm 2.09}$	$0.680^{***}_{\pm 0.04}$	$0.162^{***}_{\pm 0.01}$	$24.68^{***}_{\pm 1.93}$	$0.822^{***}_{\pm 0.04}$	$0.096^{***}_{\pm 0.02}$
MMHVAE (Dorent et al., 2026)	$21.66^{***}_{\pm 2.40}$	$0.691^{***}_{\pm 0.05}$	$0.157^{***}_{\pm 0.02}$	$24.23^{***}_{\pm 2.32}$	$0.819^{**}_{\pm 0.04}$	$0.097^{***}_{\pm 0.02}$
Diffusion-based Methods						
M2DN (Meng et al., 2024)	$21.15^{***}_{\pm 2.27}$	$0.673^{***}_{\pm 0.04}$	$0.161^{***}_{\pm 0.02}$	$24.35^{***}_{\pm 2.02}$	$0.816^{***}_{\pm 0.03}$	$0.099^{***}_{\pm 0.01}$
selfRDB (Arslan et al., 2025)	$22.03^{***}_{\pm 2.30}$	$0.695^{***}_{\pm 0.06}$	$0.139^{***}_{\pm 0.02}$	$25.16^{***}_{\pm 1.98}$	$0.834^{***}_{\pm 0.05}$	$0.082^{***}_{\pm 0.03}$
CRSG (Song et al., 2026b)	$21.85^{***}_{\pm 2.55}$	$0.691^{***}_{\pm 0.05}$	$0.148^{***}_{\pm 0.03}$	$25.31^{**}_{\pm 2.16}$	$0.834^{**}_{\pm 0.04}$	$0.085^{**}_{\pm 0.02}$
Source-only	$21.59^{**}_{\pm 2.14}$	$0.690^{*}_{\pm 0.05}$	$0.151^{**}_{\pm 0.02}$	$24.96^{***}_{\pm 2.33}$	$0.828^{**}_{\pm 0.04}$	$0.091^{***}_{\pm 0.02}$
Full-prefix AR	$22.10^{**}_{\pm 2.12}$	$0.696^{**}_{\pm 0.06}$	$0.135^{***}_{\pm 0.04}$	$25.49^{***}_{\pm 2.19}$	$0.834^{*}_{\pm 0.04}$	$0.085^{**}_{\pm 0.02}$
SPIRA	**22.72** $_{\pm 2.43}$	**0.705** $_{\pm 0.06}$	**0.126** $_{\pm 0.04}$	**25.93** $_{\pm 2.07}$	**0.842** $_{\pm 0.05}$	**0.077** $_{\pm 0.03}$

Methods	T2w→T2f			T2f→T1c		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
GAN-based Methods						
3D-mADUNet (Ha et al., 2025)	$20.84^{***}_{\pm 1.91}$	$0.646^{***}_{\pm 0.06}$	$0.177^{***}_{\pm 0.03}$	$20.71^{***}_{\pm 1.80}$	$0.653^{***}_{\pm 0.04}$	$0.178^{***}_{\pm 0.04}$
MMHVAE (Dorent et al., 2026)	$20.67^{***}_{\pm 2.20}$	$0.643^{***}_{\pm 0.05}$	$0.182^{***}_{\pm 0.04}$	$20.90^{***}_{\pm 2.11}$	$0.662^{***}_{\pm 0.05}$	$0.170^{***}_{\pm 0.02}$
Diffusion-based Methods						
M2DN (Meng et al., 2024)	$20.83^{**}_{\pm 2.34}$	$0.648^{***}_{\pm 0.07}$	$0.179^{***}_{\pm 0.05}$	$20.64^{***}_{\pm 1.97}$	$0.651^{***}_{\pm 0.06}$	$0.184^{***}_{\pm 0.04}$
selfRDB (Arslan et al., 2025)	$21.42^{***}_{\pm 1.77}$	$0.674^{***}_{\pm 0.04}$	$0.159^{***}_{\pm 0.02}$	$21.33^{**}_{\pm 1.93}$	$0.677^{***}_{\pm 0.05}$	$0.160^{**}_{\pm 0.03}$
CRSG (Song et al., 2026b)	$21.19^{***}_{\pm 2.13}$	$0.670^{**}_{\pm 0.06}$	$0.166^{***}_{\pm 0.03}$	$21.02^{***}_{\pm 2.27}$	$0.669^{***}_{\pm 0.08}$	$0.171^{***}_{\pm 0.03}$
Source-only	$21.01^{**}_{\pm 1.84}$	$0.670^{***}_{\pm 0.07}$	$0.164^{***}_{\pm 0.03}$	$21.09^{**}_{\pm 1.99}$	$0.671^{***}_{\pm 0.06}$	$0.163^{*}_{\pm 0.04}$
Full-prefix AR	$21.30^{**}_{\pm 2.21}$	$0.674^{**}_{\pm 0.06}$	$0.155^{**}_{\pm 0.04}$	$21.55^{***}_{\pm 2.05}$	$0.681^{***}_{\pm 0.06}$	$0.154^{**}_{\pm 0.03}$
SPIRA	**21.77** $_{\pm 2.09}$	**0.680** $_{\pm 0.07}$	**0.150** $_{\pm 0.04}$	**22.08** $_{\pm 1.92}$	**0.687** $_{\pm 0.05}$	**0.149** $_{\pm 0.03}$

on three of four tasks. Although synthesis adds no new patient information, the target-contrast representation can make existing information more useful to the segmentation network, consistent with prior missing-modality synthesis studies (Meng et al., 2024). The gains over Source-only and Full-prefix AR therefore indicate improved downstream segmentation performance in this evaluated BraTS setting, rather than equivalence to an acquired target modality. Fig. 8 provides qualitative examples for T1c→T1n and T2f→T1c, which are consistent with the quantitative segmentation trends.

D.5 Additional Qualitative Results

Figs. 11–14 provide additional illustrative qualitative comparisons across all eight translation tasks on BraTS 2024 (de Verdier et al., 2024), SynthRAD 2025 (Thummerer et al., 2025), and AutoPET (Dexl et al., 2026). For each task, we show axial, coronal, and sagittal views together with the corresponding error maps. Across the selected examples, SPIRA often exhibits lower residual errors while maintaining visible target-specific structures. Together, these examples are consistent with the cross-task quantitative findings in Tables 1–2 and the ROI analyses in Tables 9–10.

D.6 More Ablation Studies

Main ablation settings. We provide additional details for the variants evaluated in Sec. 4.3. Except for Standard AR, all variants use the same dual-endpoint AR predictor as SPIRA and modify only the retraction strategy or gate design. They therefore share the same tokenizer, AR architecture, and training configuration, allowing the comparison to isolate how prefix corrections are controlled.

- **Standard AR.** We separately train the source-conditioned AR predictor using only $\mathcal{L}_{\text{full}}$, without the source-only endpoint, oracle, or retraction gate. Unlike Full-prefix AR in Tables 1 and 2, which disables retraction at $s = 0$ on the dual-endpoint predictor trained within SPIRA, this variant follows standard source-conditioned AR training.

Table 12: **Subject-level statistical analysis on SynthRAD 2025 and AutoPET.** We report the mean performance with subject-level standard deviations for four translation tasks. Asterisks indicate statistically significant differences between each competing method and SPIRA based on two-sided paired Wilcoxon signed-rank tests: $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$. Subscripts denote subject-level standard deviations.

Methods	MRI→CT			CBCT→CT		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
GAN-based Methods						
3D-mADUNet (Ha et al., 2025)	$20.78^{***}_{\pm 1.85}$	$0.701^{***}_{\pm 0.05}$	$0.238^{***}_{\pm 0.03}$	$22.94^{***}_{\pm 1.81}$	$0.737^{***}_{\pm 0.07}$	$0.128^{***}_{\pm 0.04}$
MMHVAE (Dorent et al., 2026)	$20.49^{***}_{\pm 2.11}$	$0.697^{***}_{\pm 0.05}$	$0.249^{***}_{\pm 0.03}$	$22.73^{***}_{\pm 2.08}$	$0.736^{***}_{\pm 0.06}$	$0.131^{***}_{\pm 0.03}$
Diffusion-based Methods						
M2DN (Meng et al., 2024)	$20.55^{***}_{\pm 2.34}$	$0.701^{***}_{\pm 0.05}$	$0.242^{***}_{\pm 0.04}$	$22.66^{***}_{\pm 1.94}$	$0.734^{***}_{\pm 0.07}$	$0.133^{***}_{\pm 0.04}$
selfRDB (Arslan et al., 2025)	$20.81^{***}_{\pm 1.95}$	$0.708^{***}_{\pm 0.08}$	$0.236^{***}_{\pm 0.05}$	$23.19^{**}_{\pm 1.98}$	$0.740^{**}_{\pm 0.08}$	$0.127^{***}_{\pm 0.05}$
CRSG (Song et al., 2026b)	$20.62^{***}_{\pm 2.16}$	$0.694^{***}_{\pm 0.07}$	$0.248^{***}_{\pm 0.05}$	$22.78^{*}_{\pm 1.74}$	$0.732^{***}_{\pm 0.05}$	$0.136^{***}_{\pm 0.03}$
Source-only	$20.27^{***}_{\pm 2.20}$	$0.691^{**}_{\pm 0.05}$	$0.252^{**}_{\pm 0.03}$	$23.02^{***}_{\pm 2.04}$	$0.739^{**}_{\pm 0.06}$	$0.129^{***}_{\pm 0.05}$
Full-prefix AR	$20.73^{***}_{\pm 1.89}$	$0.705^{***}_{\pm 0.06}$	$0.232^{***}_{\pm 0.03}$	$23.35^{*}_{\pm 1.91}$	$0.744^{**}_{\pm 0.06}$	$0.123^{**}_{\pm 0.04}$
SPIRA	**21.29** $_{\pm 2.07}$	**0.714** $_{\pm 0.07}$	**0.225** $_{\pm 0.05}$	**23.59** $_{\pm 1.82}$	**0.751** $_{\pm 0.06}$	**0.119** $_{\pm 0.04}$

Methods	CT→PET			PET→CT		
	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
GAN-based Methods						
3D-mADUNet (Ha et al., 2025)	$24.73^{***}_{\pm 1.39}$	$0.902^{***}_{\pm 0.03}$	$0.130^{***}_{\pm 0.02}$	$23.49^{***}_{\pm 0.91}$	$0.896^{***}_{\pm 0.02}$	$0.105^{***}_{\pm 0.01}$
MMHVAE (Dorent et al., 2026)	$25.57^{***}_{\pm 1.22}$	$0.914^{***}_{\pm 0.02}$	$0.106^{***}_{\pm 0.01}$	$23.71^{***}_{\pm 1.07}$	$0.899^{***}_{\pm 0.03}$	$0.103^{***}_{\pm 0.01}$
Diffusion-based Methods						
M2DN (Meng et al., 2024)	$25.69^{***}_{\pm 1.47}$	$0.912^{***}_{\pm 0.03}$	$0.104^{***}_{\pm 0.02}$	$23.81^{***}_{\pm 0.96}$	$0.902^{***}_{\pm 0.03}$	$0.095^{**}_{\pm 0.02}$
selfRDB (Arslan et al., 2025)	$26.08^{***}_{\pm 1.51}$	$0.923^{***}_{\pm 0.03}$	$0.094^{***}_{\pm 0.02}$	$23.67^{***}_{\pm 1.01}$	$0.901^{***}_{\pm 0.02}$	$0.099^{***}_{\pm 0.01}$
CRSG (Song et al., 2026b)	$26.85^{***}_{\pm 1.27}$	$0.937^{***}_{\pm 0.02}$	$0.080^{**}_{\pm 0.01}$	$23.98^{***}_{\pm 1.12}$	$0.903^{***}_{\pm 0.02}$	$0.095^{***}_{\pm 0.01}$
Source-only	$27.30^{***}_{\pm 1.58}$	$0.939^{***}_{\pm 0.02}$	$0.078^{**}_{\pm 0.01}$	$23.58^{***}_{\pm 1.03}$	$0.901^{***}_{\pm 0.02}$	$0.097^{***}_{\pm 0.01}$
Full-prefix AR	$27.93^{**}_{\pm 1.35}$	$0.947^{**}_{\pm 0.02}$	$0.066^{***}_{\pm 0.02}$	$24.15^{***}_{\pm 0.99}$	$0.905^{***}_{\pm 0.03}$	$0.093^{**}_{\pm 0.02}$
SPIRA	**28.44** $_{\pm 1.43}$	**0.956** $_{\pm 0.02}$	**0.058** $_{\pm 0.01}$	**24.76** $_{\pm 1.07}$	**0.912** $_{\pm 0.02}$	**0.086** $_{\pm 0.02}$

Table 13: **Multi-seed robustness analysis on BraTS 2024.** We report the mean performance over three independent training runs. Source-only, Full-prefix AR, and SPIRA are evaluated from the same trained dual-endpoint predictor, with subscripts denoting standard deviations across seeds.

Methods	T1n→T1c				T1c→T1n			
	PSNR↑	SSIM↑	LPIPS↓	FID↓	PSNR↑	SSIM↑	LPIPS↓	FID↓
Source-only	21.55 $_{\pm 0.19}$	0.690 $_{\pm 0.004}$	0.152 $_{\pm 0.002}$	47.0 $_{\pm 1.0}$	24.99 $_{\pm 0.21}$	0.828 $_{\pm 0.005}$	0.090 $_{\pm 0.004}$	34.2 $_{\pm 0.8}$
Full-prefix AR	22.14 $_{\pm 0.22}$	0.697 $_{\pm 0.005}$	0.133 $_{\pm 0.003}$	42.8 $_{\pm 0.9}$	25.51 $_{\pm 0.15}$	0.834 $_{\pm 0.004}$	0.085 $_{\pm 0.002}$	32.5 $_{\pm 1.1}$
SPIRA	**22.74** $_{\pm 0.19}$	**0.706** $_{\pm 0.005}$	**0.126** $_{\pm 0.002}$	**38.9** $_{\pm 1.3}$	**25.90** $_{\pm 0.17}$	**0.841** $_{\pm 0.004}$	**0.079** $_{\pm 0.002}$	**29.1** $_{\pm 1.2}$

Methods	T2w→T2f				T2f→T1c			
	PSNR↑	SSIM↑	LPIPS↓	FID↓	PSNR↑	SSIM↑	LPIPS↓	FID↓
Source-only	21.05 $_{\pm 0.12}$	0.671 $_{\pm 0.004}$	0.162 $_{\pm 0.003}$	55.8 $_{\pm 1.2}$	21.03 $_{\pm 0.20}$	0.670 $_{\pm 0.005}$	0.163 $_{\pm 0.003}$	49.2 $_{\pm 1.0}$
Full-prefix AR	21.31 $_{\pm 0.18}$	0.674 $_{\pm 0.004}$	0.154 $_{\pm 0.004}$	49.1 $_{\pm 0.8}$	21.57 $_{\pm 0.15}$	0.680 $_{\pm 0.003}$	0.152 $_{\pm 0.002}$	46.4 $_{\pm 1.1}$
SPIRA	**21.75** $_{\pm 0.14}$	**0.681** $_{\pm 0.004}$	**0.152** $_{\pm 0.003}$	**43.1** $_{\pm 0.9}$	**22.03** $_{\pm 0.17}$	**0.685** $_{\pm 0.005}$	**0.151** $_{\pm 0.004}$	**42.7** $_{\pm 1.1}$

- **Tuned global retraction.** We remove the adaptive gate and use a single retraction amount $s_i = s_{\text{global}}$ for every prediction. The scalar is selected on the training-side model-selection subset from $s \in [0, 1]$ and then fixed for all held-out evaluation subjects and spatial locations. This variant investigates whether the improvement can be obtained by globally reducing prefix influence rather than adapting the retained correction to the current region.
- **Position-only gate.** We retain the same oracle supervision but provide the gate only with the spatial position of the current prediction, removing all content-dependent inputs. This tests whether SPIRA merely learns a fixed position-dependent retraction pattern rather than adapting to the current source–prefix interaction.
- **Effect-only gate.** We keep the same gate architecture and oracle supervision but provide only the GT-free endpoint-difference statistics, removing the explicit source and correction representations. This evaluates whether predictive effect alone is sufficient to determine how much of a correction to retain.

Table 14: **Downstream applicability to tumor segmentation on BraTS 2024.** Performance is measured by Dice. Asterisks indicate statistically significant differences between each competing method and SPIRA based on paired Wilcoxon signed-rank tests: $^{*}p < 0.05$, $^{**}p < 0.01$, and $^{***}p < 0.001$. Subscripts denote subject-level standard deviations.

Methods	T1n→T1c	T1c→T1n	T2w→T2f	T2f→T1c	Avg↑
Reference Settings					
3 Real Modalities	0.753 $_{\pm 0.09}$	0.782 $_{\pm 0.11}$	0.718 $_{\pm 0.13}$	0.753 $_{\pm 0.09}$	0.752 $_{\pm 0.11}$
GAN-based Methods					
3D-mADUNet (Ha et al., 2025)	0.769 $^{***}_{\pm 0.12}$	0.808 $^{**}_{\pm 0.09}$	0.740 $^{***}_{\pm 0.13}$	0.760 $^{**}_{\pm 0.10}$	0.769 $_{\pm 0.11}$
MMHVAE (Dorent et al., 2026)	0.791 $^{**}_{\pm 0.11}$	0.816 $^{**}_{\pm 0.08}$	0.734 $^{***}_{\pm 0.11}$	0.762 $^{***}_{\pm 0.13}$	0.776 $_{\pm 0.11}$
Diffusion-based Methods					
M2DN (Meng et al., 2024)	0.776 $^{**}_{\pm 0.08}$	0.810 $^{***}_{\pm 0.12}$	0.742 $^{***}_{\pm 0.07}$	0.762 $^{**}_{\pm 0.08}$	0.773 $_{\pm 0.09}$
selfRDB (Arslan et al., 2025)	0.788 $^{***}_{\pm 0.10}$	0.825 $^{***}_{\pm 0.09}$	**0.761** $^{*}_{\pm \mathbf{0.11}}$	0.782 $^{**}_{\pm 0.09}$	0.789 $_{\pm 0.10}$
CRSG (Song et al., 2026b)	0.782 $^{**}_{\pm 0.07}$	0.837 $^{**}_{\pm 0.11}$	0.745 $^{***}_{\pm 0.08}$	0.770 $^{***}_{\pm 0.10}$	0.784 $_{\pm 0.09}$
Source-only	0.793 $^{**}_{\pm 0.10}$	0.835 $^{***}_{\pm 0.14}$	0.748 $^{***}_{\pm 0.12}$	0.778 $^{**}_{\pm 0.09}$	0.789 $_{\pm 0.11}$
Full-prefix AR	0.802 $^{***}_{\pm 0.13}$	0.832 $^{**}_{\pm 0.12}$	0.750 $^{***}_{\pm 0.09}$	0.780 $^{***}_{\pm 0.14}$	0.791 $_{\pm 0.12}$
SPIRA	**0.809** $_{\pm \mathbf{0.12}}$	**0.844** $_{\pm \mathbf{0.10}}$	0.758 $_{\pm 0.14}$	**0.791** $_{\pm \mathbf{0.08}}$	**0.801** $_{\pm \mathbf{0.11}}$

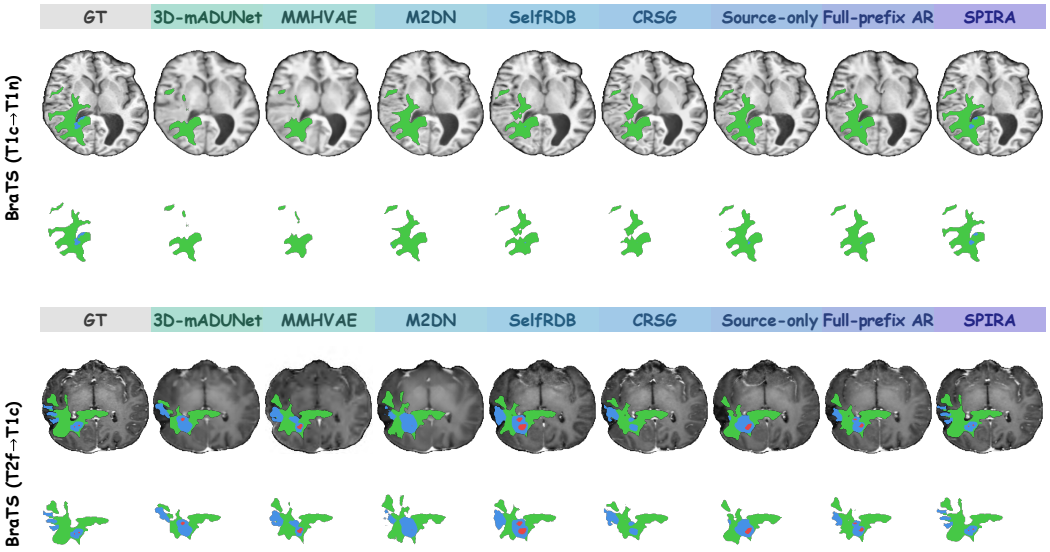

Figure 8: **Qualitative comparison of downstream tumor segmentation on BraTS 2024.** Results are shown for T1c→T1n (upper) and T2f→T1c (lower), with segmentation overlays and the corresponding masks shown for each method.

- **Direct s^* regression.** We use the complete SPIRA gate inputs and the same maximal zero-regret oracle, but replace predictive-action distillation with direct L1 regression, $\mathcal{L}_s = \mathbb{E}_{i \in \mathcal{V}}[|\hat{s}_i - s_i^*|]$. This comparison isolates whether matching the oracle predictive distribution is preferable to directly reproducing its scalar retraction amount.

Impact of retraction strategy and supervision strategy. We compare global prefix amplification with alternative oracle targets to assess how the retraction strategy and supervision affect generation.

- **Global amplification ($s < 0$).** We tune a single global scalar over $s \in [-0.5, 0)$, extending the source–prefix path beyond Full-prefix AR to amplify the prefix correction. This tests whether stronger prefix influence is beneficial.
- **Endpoint zero-regret supervision.** The oracle restricts the retraction action to $s \in \{0, 1\}$, allowing only full retention or full retraction. This variant tests whether continuous partial retraction is necessary.

Table 15: **Additional ablation studies** on BraTS 2024 T1n→T1c and SynthRAD 2025 MRI→CT. We evaluate the impact of retraction strategy, oracle supervision, and gate input.

Variants	BraTS T1n→T1c				SynthRAD MRI→CT			
	PSNR↑	SSIM↑	LPIPS↓	FID↓	PSNR↑	SSIM↑	LPIPS↓	FID↓
Retraction Strategy								
Global amplification ($s < 0$)	21.98	0.696	0.136	44.1	20.68	0.704	0.233	92.7
Supervision Strategy								
Endpoint zero-regret oracle	22.53	0.699	0.131	41.7	21.12	0.708	0.231	90.1
Minimum-loss oracle	22.60	0.700	0.130	40.2	21.13	0.710	0.229	89.7
Direct target-NLL gate	22.56	0.700	0.129	39.6	21.07	0.710	0.229	91.3
Gate Input								
w/o source feature $\mathbf{u}^S$	22.49	0.700	0.130	42.4	21.05	0.706	0.232	90.2
w/o correction feature $\mathbf{u}^r$	22.41	0.697	0.129	41.5	21.00	0.704	0.230	92.4
w/o endpoint statistics (w, e)	22.60	0.702	0.129	39.8	21.17	0.709	0.228	89.5
replace $\mathbf{u}^r$ with full-prefix feature $\mathbf{u}^F$	22.54	0.701	0.131	42.2	21.12	0.710	0.229	91.0
SPIRA (Maximal zero-regret)	**22.72**	**0.705**	**0.126**	**38.3**	**21.29**	**0.714**	**0.225**	**87.3**

Table 16: **Ablation studies of the retraction granularity** on BraTS 2024 T1n→T1c and SynthRAD 2025 MRI→CT. The AR block size is fixed at $(4, 4, 4)$, and only the spatial granularity of the shared retraction action is varied.

Retraction Granularity	Tokens/ Action	Actions/ Block	BraTS T1n→T1c				SynthRAD MRI→CT			
			PSNR↑	SSIM↑	LPIPS↓	FID↓	PSNR↑	SSIM↑	LPIPS↓	FID↓
Block-wise $(4, 4, 4)$	64	1	22.08	0.693	0.137	45.2	20.65	0.695	0.242	95.9
Sub-block-wise $(2, 2, 2)$	8	8	22.39	0.697	0.132	40.9	20.98	0.705	0.231	91.6
Token-wise $(1, 1, 1)$	**1**	**64**	**22.72**	**0.705**	**0.126**	**38.3**	**21.29**	**0.714**	**0.225**	**87.3**

- **Minimum-loss supervision.** The oracle selects the retraction amount that minimizes the current target loss along the retraction path. This tests whether directly optimizing the local prediction loss is a better supervision target.
- **Direct target-NLL gate.** We retain the complete SPIRA gate inputs, architecture, and source-only/full-prefix retraction path, but remove oracle supervision. Instead, the gate directly minimizes the GT factor NLL induced by its predicted retraction:

$$\mathcal{L}_{\text{gate-NLL}} = \mathbb{E}_{i,v}\left[-\frac{1}{d}\sum_{j=1}^{d}\log \operatorname{softmax}\left(z^F_{i,v,j} - \hat{s}_{i,v} r_{i,v,j}\right)_{q^{\text{GT}}_{i,v,j}}\right].$$

  As in SPIRA, all endpoint logits and gate input features are stop-gradient quantities, so this objective updates only the gate and does not alter the AR predictor objective. This matched baseline tests whether directly optimizing target loss under the gate's GT-free inputs and capacity is sufficient, without constructing or distilling a per-example target-aware oracle.

Impact of gate input. We remove or replace individual gate inputs to assess which source–prefix cues are important for retraction decisions.

- **w/o source feature $\mathbf{u}^S$.** We remove the source-only feature from the gate input. This tests whether source evidence is important for deciding how much prefix correction to retain.
- **w/o correction feature $\mathbf{u}^r$.** We remove the feature describing the hidden-space change introduced by the prefix. This tests whether explicitly modeling the prefix correction benefits the retraction decision.
- **w/o endpoint statistics (w, e).** We remove the lightweight statistics describing the difference between the source-only and full-prefix predictions. This tests whether these endpoint cues provide complementary information to the hidden features.
- **replace $\mathbf{u}^r$ with full-prefix feature $\mathbf{u}^F$.** We replace the correction feature with the full-prefix feature. This tests whether explicitly representing the source-to-prefix change is more useful than using the full-prefix state directly.

Table 17: **Ablation studies of the block size** on BraTS 2024 T1n→T1c and SynthRAD 2025 MRI→CT. The number of AR blocks is computed on the 32^3 latent grid.

Block Size	Tokens/Block	# AR Blocks	BraTS T1n→T1c				SynthRAD MRI→CT			
			PSNR↑	SSIM↑	LPIPS↓	FID↓	PSNR↑	SSIM↑	LPIPS↓	FID↓
$2 \times 4 \times 4$	32	1024	22.67	0.705	0.128	39.2	21.21	0.712	0.227	87.9
$4 \times 4 \times 2$	32	1024	22.69	**0.707**	**0.125**	38.8	21.17	0.709	0.229	88.7
$4 \times 4 \times 4$	**64**	**512**	**22.72**	0.705	0.126	**38.3**	**21.29**	**0.714**	**0.225**	**87.3**
$8 \times 4 \times 4$	128	256	22.52	0.700	0.129	39.5	20.93	0.706	0.230	90.0
$4 \times 4 \times 8$	128	256	22.41	0.697	0.128	40.2	20.97	0.706	0.228	89.1

Table 15 further examines retraction strategy, supervision strategy, and gate inputs. Global amplification falls below even Full-prefix AR on both tasks, showing that amplifying the prefix correction is detrimental rather than beneficial in these paired translation settings. Endpoint zero-regret and minimum-loss supervision also underperform SPIRA across all eight dataset–metric comparisons, supporting continuous maximal zero-regret retraction. The direct target-NLL gate, which retains the same complete gate inputs, capacity, and retraction path but directly optimizes GT target loss, also underperforms SPIRA across all eight metrics. This suggests that directly optimizing local target loss does not recover the benefit of constructing and distilling the maximal zero-regret predictive action. This advantage is further supported by the GT-assisted policy comparisons in Figs. 6(a) and 10(b), which use the same predictor and avoid gate-training variation. Removing the source or correction features, or replacing the correction residual with the full-prefix feature, consistently degrades performance, while endpoint statistics provide smaller complementary gains. Together, these results support selective retraction, maximal zero-regret supervision, predictive-action distillation, and the proposed gate design.

Impact of retraction granularity. Table 16 compares block-wise, sub-block-wise, and token-wise retraction while keeping the AR block size fixed. Performance consistently improves as the retraction granularity becomes finer, with token-wise control achieving the best results across both datasets and all metrics. These findings suggest that sharing one action across multiple tokens can obscure local differences in prefix utility, while token-wise retraction better adapts the source–prefix balance to local prediction needs.

Effect of block size. Table 17 compares different AR block sizes, which trade off prediction granularity and the number of autoregressive steps. Smaller blocks achieve comparable performance but require twice as many AR blocks, while larger blocks reduce the number of steps at the cost of translation quality. Balancing translation quality and efficiency, we use $(4 \times 4 \times 4)$ as the default block size in all main experiments.

E Further Analysis

E.1 Additional Utility–Effect Analysis

Building on Sec. 4.4, we provide additional analyses of the utility–effect gap across effect levels, tasks, anatomical regions, and controlled matched-effect interventions.

Effect measures. We distinguish the empirical TV-based prediction change used in Sec. 2 from the theoretical KL effect $E_i(1)$ used in the regret decomposition. Both compare the same source-only and full-prefix prediction endpoints, with TV providing a bounded descriptive measure for the empirical analysis, whereas KL is used in Sec. 4.4 and here because it appears exactly in Eq. 5.

Do matched effects imply matched utilities? Fig. 9(a) extends the utility analysis across eight translation tasks and the full effect spectrum. Helpful and harmful corrections coexist within the same effect levels across all tasks, indicating that effect magnitude alone does not determine correction utility.

Does the pattern hold across tasks? Fig. 9(b) summarizes utility mixing and partial retraction across tasks. Opposite-utility mass remains 21.0–34.7%, while partial retraction accounts for 9.5–19.4% of predictions, indicating that both patterns persist across datasets and translation directions.

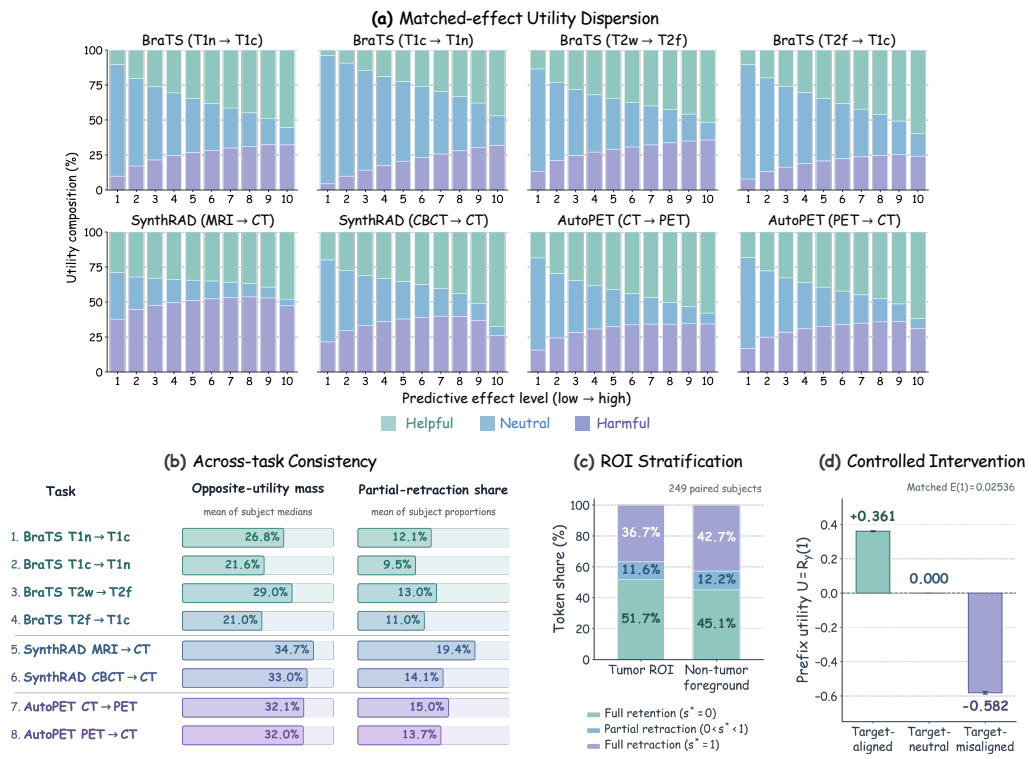

Figure 9: **Additional analysis of the utility–effect gap.** (a) Utility composition within task-specific predictive-effect deciles across eight translation tasks. (b) Utility mixing and partial retraction persist across tasks. (c) Oracle retraction regimes differ between tumor ROIs and non-tumor foreground on BraTS T1n→T1c. (d) At matched predictive effect, controlled changes in target alignment produce positive, neutral, and negative prefix utilities.

Does retraction vary across regions? Fig. 9(c) compares tumor ROIs with non-tumor foreground on BraTS T1n→T1c. Tumor regions have a higher full-retention share (51.7% vs. 45.1%) and a lower full-retraction share (36.7% vs. 42.7%), indicating spatial variation in the preferred source–prefix balance.

Does target alignment matter at fixed effect? For Fig. 9(d), we use real source-only predictions from BraTS T1n→T1c as anchors and fix the endpoint effect to $E(1) = 0.0254$. Target-aligned, neutral, and misaligned endpoints are constructed while numerically matching this effect. Their utilities are $+0.361$, 0.000, and -0.582, respectively, directly illustrating that matched effects can have different utilities when target alignment changes.

All three interventions are feasible for 81.6% of anchors, yielding 3.54M matched endpoints across 249 subjects. The maximum effect-matching error is 3.4×10^{-9}. Bars report subject-equal means with 95% confidence intervals from 1,000 subject-level bootstrap samples.

E.2 Additional Analysis of Selective Retraction

Actionability and gain definition. For subject n in task t, we measure baseline actionability on Full-prefix rollout states as

$$\mathcal{A}_{n,t} = \frac{\sum_{i \in \mathcal{V}_{n,t}} E_{n,t,i}(s^{*,F}_{n,t,i})}{\sum_{i \in \mathcal{V}_{n,t}} E_{n,t,i}(1) + \epsilon}, \tag{19}$$

where $s^{*,F}_{n,t,i}$ is the GT-assisted maximal zero-regret retraction computed on the Full-prefix rollout state, and $\mathcal{V}_{n,t}$ contains valid target tokens. Thus, $\mathcal{A}_{n,t}$ measures the fraction of endpoint predictive effect that can be safely retracted.

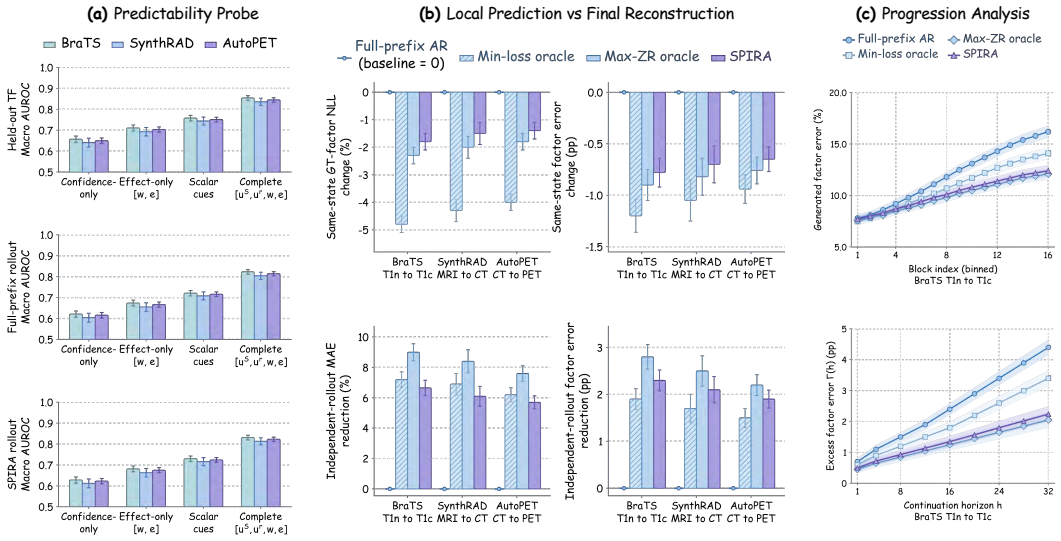

Figure 10: **Additional analysis of selective retraction.** (a) Predictability probe. Macro AUROC for predicting the three oracle retraction regimes from different GT-free feature sets on teacher-forced and generated-prefix states. (b) Local prediction versus final reconstruction. Same-state GT-factor NLL is compared with independently rolled-out reconstruction performance. (c) Progression and prefix sensitivity. Free-running factor error is tracked over block index, while controlled continuations from matched reference and generated prefixes measure the excess factor error $\Gamma^{\pi}(h)$.

For a policy π, we define its subject-level generation gain over Full-prefix AR as

$$g^{\pi}_{n,t} = 100\,\frac{\mathrm{MAE}^{F}_{n,t} - \mathrm{MAE}^{\pi}_{n,t}}{\mathrm{MAE}^{F}_{n,t}}. \tag{20}$$

We quantify how actionability relates to SPIRA's additional gain over effect-only using

$$g^{\mathrm{SPIRA}}_{n,t} - g^{\mathrm{effect}}_{n,t} = \alpha_t + \beta\,\widetilde{\mathcal{A}}_{n,t} + \varepsilon_{n,t}, \tag{21}$$

where $\widetilde{\mathcal{A}}_{n,t}$ is actionability standardized within each task and α_t is a task-specific intercept. Accordingly, β denotes the extra percentage-point gain of SPIRA over effect-only per one standard deviation increase in actionability.

Can oracle retraction be predicted without the target? We train lightweight probes on teacher-forced training states to predict the three oracle regimes, namely full retention, partial retraction, and full retraction. We utilize a one-hidden-layer 64-d MLP probe with subject-disjoint training and selection splits. We compare source confidence c^S, effect-only statistics $[w, e]$, scalar cues $[c^S, c^F, w, e, E(1)]$, and the complete GT-free gate features $[u^S, u^r, w, e]$. The same probe is evaluated without refitting on held-out teacher-forced, Full-prefix rollout, and SPIRA rollout states, using macro one-vs-rest AUROC. Complete features perform best across all settings (macro AUROC 0.804–0.854), exceeding confidence or effect alone and indicating that source–correction features contain additional oracle-relevant information.

How often does the learned gate violate the local zero-regret constraint? We further evaluate the learned policies on identical SPIRA-generated prefix states and compute the current-state regret $R_y(\hat{s}) = \ell_y(z(\hat{s})) - \ell_y(z^F)$. Within each token, previously generated SPIRA factors are kept fixed when evaluating regret, while GT factors are used only for scoring. We report two complementary metrics. The positive-regret rate (PRR) measures how often the learned action violates the local zero-regret condition, i.e., the fraction of actions with $R_y(\hat{s}) > 0$. Positive-regret magnitude measures how severe these violations are, computed as the mean factor-averaged NLL regret over violating actions. As shown in Table 18, SPIRA consistently reduces both quantities across all three tasks. Compared with the effect-only gate, PRR decreases from 26.4–31.0% to 15.1–17.4%, while positive-regret magnitude decreases from 0.020–0.025 to 0.011–0.014 nats. Accordingly, SPIRA better

Table 18: **Local regret audit on generated-prefix states.** We evaluate learned retraction policies on identical SPIRA-generated states, using the GT target only to measure current-state regret. Positive-regret rate reports the fraction of actions with $R_y(\hat{s}) > 0$, while positive-regret magnitude reports the mean regret over these violating actions.

	BraTS T1n→T1c		SynthRAD MRI→CT		AutoPET CT→PET	
Methods	**PRR (%)↓**	**Regret↓**	**PRR (%)↓**	**Regret↓**	**PRR (%)↓**	**Regret↓**
Effect-only gate	27.1	0.022	31.0	0.025	26.4	0.020
SPIRA	15.9	0.013	17.4	0.014	15.1	0.011

Table 19: **Training efficiency versus translation quality on BraTS 2024.** We compare the computational cost of representative internal variants under the same training setting, together with translation quality on the T1n→T1c task.

	Training Cost				BraTS 2024 T1n→T1c			
Variant	**Params (M)↓**	**FLOPs (G)↓**	**Time/Step (s)↓**	**GPU h ↓**	**PSNR↑**	**SSIM↑**	**LPIPS↓**	**FID↓**
Standard AR	77.73	921.1	0.489	13.57	22.16	0.696	0.137	42.5
Tuned global scalar	77.73	1802.5	0.920	25.57	22.28	0.699	0.133	42.8
Effect-only gate	77.80	1815.1	0.930	25.83	22.31	0.700	0.131	40.3
Direct s^* regression	77.86	1826.1	0.948	26.32	22.60	0.701	0.129	39.9
SPIRA	77.86	1827.3	0.948	26.32	**22.72**	**0.705**	**0.126**	**38.3**

approximates the oracle's local zero-regret behavior on generated-prefix states, while the remaining violations emphasize that this property does not extend as a guarantee to the learned gate.

Does lower current token loss imply better final reconstruction? We first compare Full-prefix, minimum-loss, maximal zero-regret, and SPIRA on identical Full-prefix rollout states, fixing the source, generated prefix, factor history, and endpoint predictions. Minimum-loss achieves the lowest same-state GT-factor NLL (a 4.0–4.8% reduction relative to Full-prefix versus 1.8–2.3% for maximal zero-regret), as expected from its objective. We then roll out each policy independently from the beginning. Despite its lower local NLL, minimum-loss yields a worse final decoded MAE than maximal zero-regret (6.2–7.2% versus 7.6–9.2% MAE reduction), showing that lower current token loss does not necessarily translate into better free-running reconstruction, while GT-free SPIRA achieves comparable gains of 5.7–6.8%.

How do policies respond to generated-prefix errors? We first report free-running factor error over 16 block-index bins within each 512-block latent window. To isolate prefix sensitivity, we additionally start controlled continuations at fixed 25%, 50%, and 75% positions from either the GT reference prefix P_{ref} or the matched Full-prefix generated prefix P_{gen}, and let each policy continue for $H = 32$ blocks. We define prefix sensitivity as $\Gamma^{\pi}(h) = \mathbb{E}[e^{\pi}(P_{\text{gen}}, h) - e^{\pi}(P_{\text{ref}}, h)]$, the excess factor error caused by replacing the reference prefix with the generated prefix at the same starting block. Maximal zero-regret and SPIRA show slower free-running error growth and lower Γ^{π}, consistent with reduced sensitivity to generated-prefix perturbations.

E.3 Computational Efficiency Analysis

To assess the practical efficiency of SPIRA, we analyze its computational cost during both AR training and inference. All efficiency measurements are conducted on a single NVIDIA RTX 5090 GPU under the same precision and implementation settings.

Wall-clock Cost of AR Training. Table 19 shows that SPIRA has a non-negligible training overhead relative to Standard AR, increasing total training cost from 13.57 to 26.32 GPU hours ($1.94\times$). This increase mainly results from dual-endpoint prediction, which raises FLOPs from 921.1G to 1827.3G and time per step from 0.489 to 0.948 seconds, while the retraction module itself adds only 0.13M parameters. Crucially, the matched dual-endpoint variants have nearly identical training costs but different translation performance. For example, Direct s^* regression has the same parameter count, time per step, and total GPU hours as SPIRA, yet performs worse across all four metrics. The

Table 20: **Inference efficiency comparison.** We report end-to-end inference latency and peak GPU memory for representative internal variants on BraTS 2024 T1n→T1c and SynthRAD 2025 MRI→CT. All variants are evaluated under the same hardware, precision, and batch-size settings.

Variant	BraTS 2024		SynthRAD 2025	
	Latency (s/vol)↓	**Peak VRAM** (GB)↓	**Latency** (s/vol)↓	**Peak VRAM** (GB)↓
Full-prefix AR	16.03	22.34	89.06	22.34
Tuned global scalar	16.49	22.34	92.94	22.34
Effect-only gate	16.67	22.34	93.59	22.34
Direct s^* regression	16.85	22.34	94.27	22.34
SPIRA	16.85	22.34	94.27	22.34

tuned global scalar and effect-only gate likewise operate at similar computational cost but achieve lower translation quality. Thus, while dual-endpoint training is substantially more expensive than Standard AR, the gains of SPIRA over compute-matched dual-endpoint variants cannot be explained by additional training compute alone.

Inference Cost. As shown in Table 20, SPIRA introduces modest inference overhead over Full-prefix AR, increasing latency from 16.03 to 16.85 seconds on BraTS 2024 (de Verdier et al., 2024) and from 89.06 to 94.27 seconds on SynthRAD 2025 (Thummerer et al., 2025). Using the tuned global scalar to isolate the endpoint cost, the additional source-only prediction accounts for 0.46 and 3.88 seconds on BraTS and SynthRAD, respectively, while adaptive gating adds only 0.36 and 1.33 seconds. Caching avoids repeated source-only evaluation but not its initial computation, indicating that most of the overhead comes from the second endpoint rather than the gate itself. The higher SynthRAD latency reflects sliding-window whole-volume generation, while peak VRAM remains unchanged at 22.34 GB because the memory peak occurs during VidTok decoding. Thus, despite the higher training cost of dual-endpoint learning, SPIRA adds only modest inference latency without increasing peak memory usage.

F Discussion

SPIRA studies source-conditioned autoregressive translation as a region-wise balance between fixed source evidence and evolving target context. By treating the prefix contribution as a controllable correction, it learns to retain target-useful information while retracting unnecessary or harmful corrections. Across eight 3D medical translation tasks, the results support both the effectiveness of this formulation and its connection to the utility–effect gap. Despite these encouraging results, several limitations remain and motivate future work.

Multiple-source translation. Our experiments focus on translation from a single source modality to a single target modality. Extending SPIRA to multiple-source settings requires jointly controlling the contributions of several source modalities and the generated prefix, which is an interesting direction for future work.

Teacher-forced gate supervision. The retraction oracle and gate supervision use GT prefixes, whereas inference operates on generated prefixes. Generated-context training methods reduce this mismatch by training on model-generated histories (Huang et al., 2025; Zhu et al., 2026; Chen et al., 2026), while SPIRA instead regulates prefix contributions token-wise without changing context training. Combining these complementary directions may further reduce the train–test gap.

Token-space and image-space objectives. Our retraction regret is defined by token-level target loss, while final translation quality is evaluated after decoding in image space. Although our analyses connect token-level decisions with reconstruction behavior, directly incorporating image-space or downstream utility into retraction learning remains an open direction.

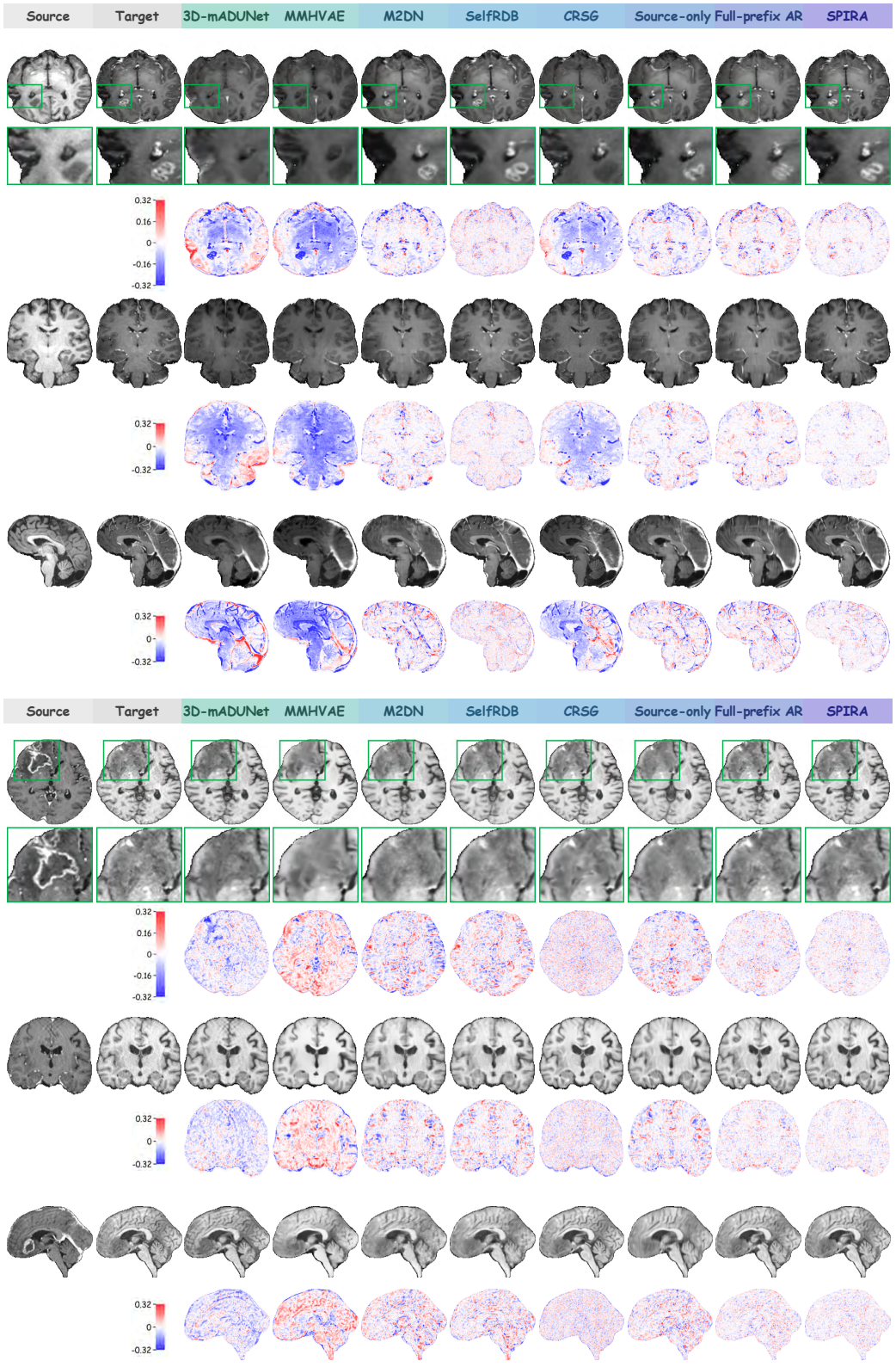

Figure 11: **Additional qualitative comparisons** on BraTS 2024 T1n→T1c (upper) and T1c→T1n (lower), showing axial, coronal, and sagittal views from top to bottom with corresponding error maps.

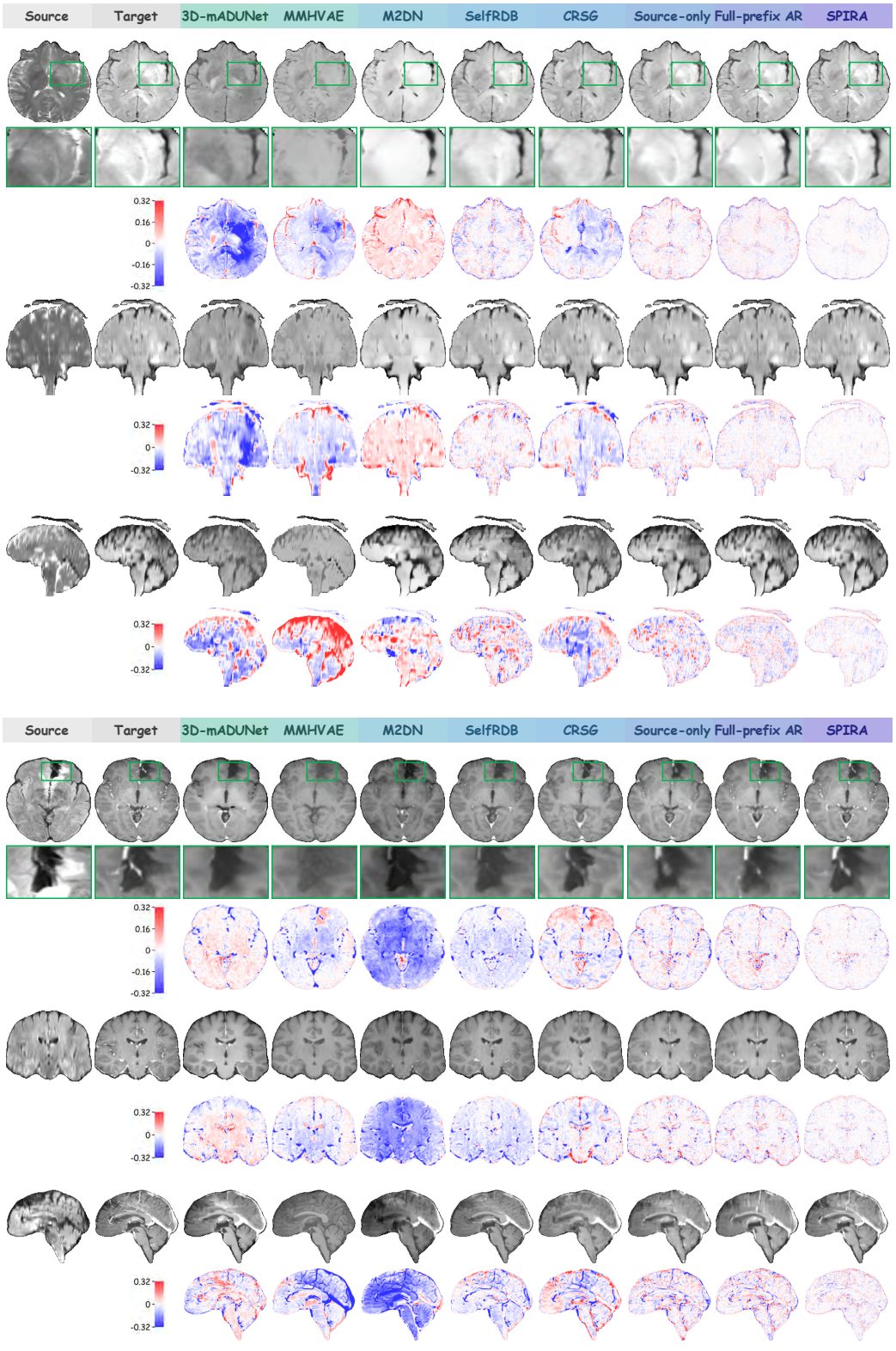

Figure 12: **Additional qualitative comparisons** on BraTS 2024 T2w→T2f (upper) and T2f→T1c (lower), showing axial, coronal, and sagittal views from top to bottom with corresponding error maps.

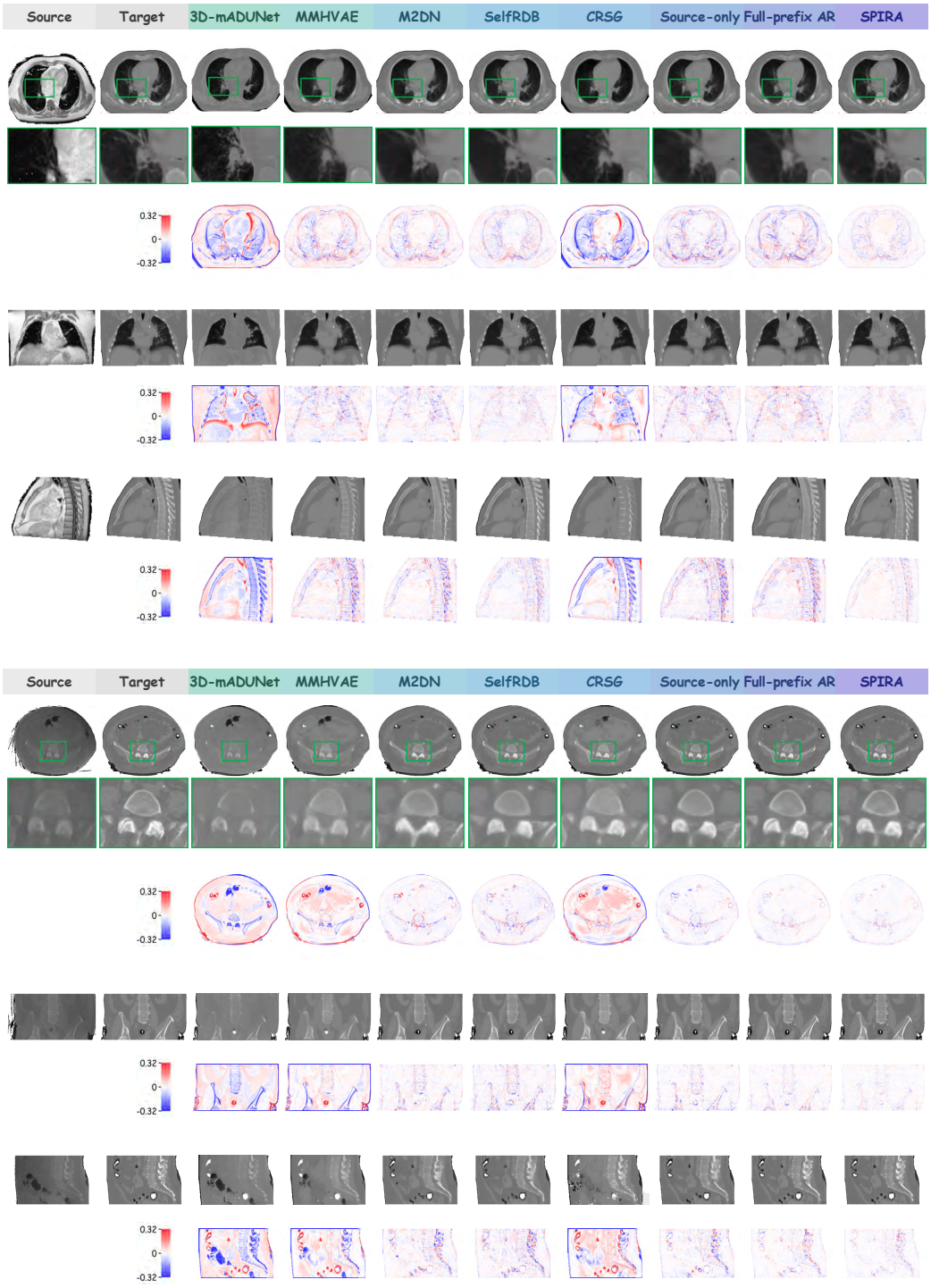

Figure 13: **Additional qualitative comparisons** on SynthRAD 2025 MRI→CT (upper) and CBCT→CT (lower), showing axial, coronal, and sagittal views from top to bottom with corresponding error maps.

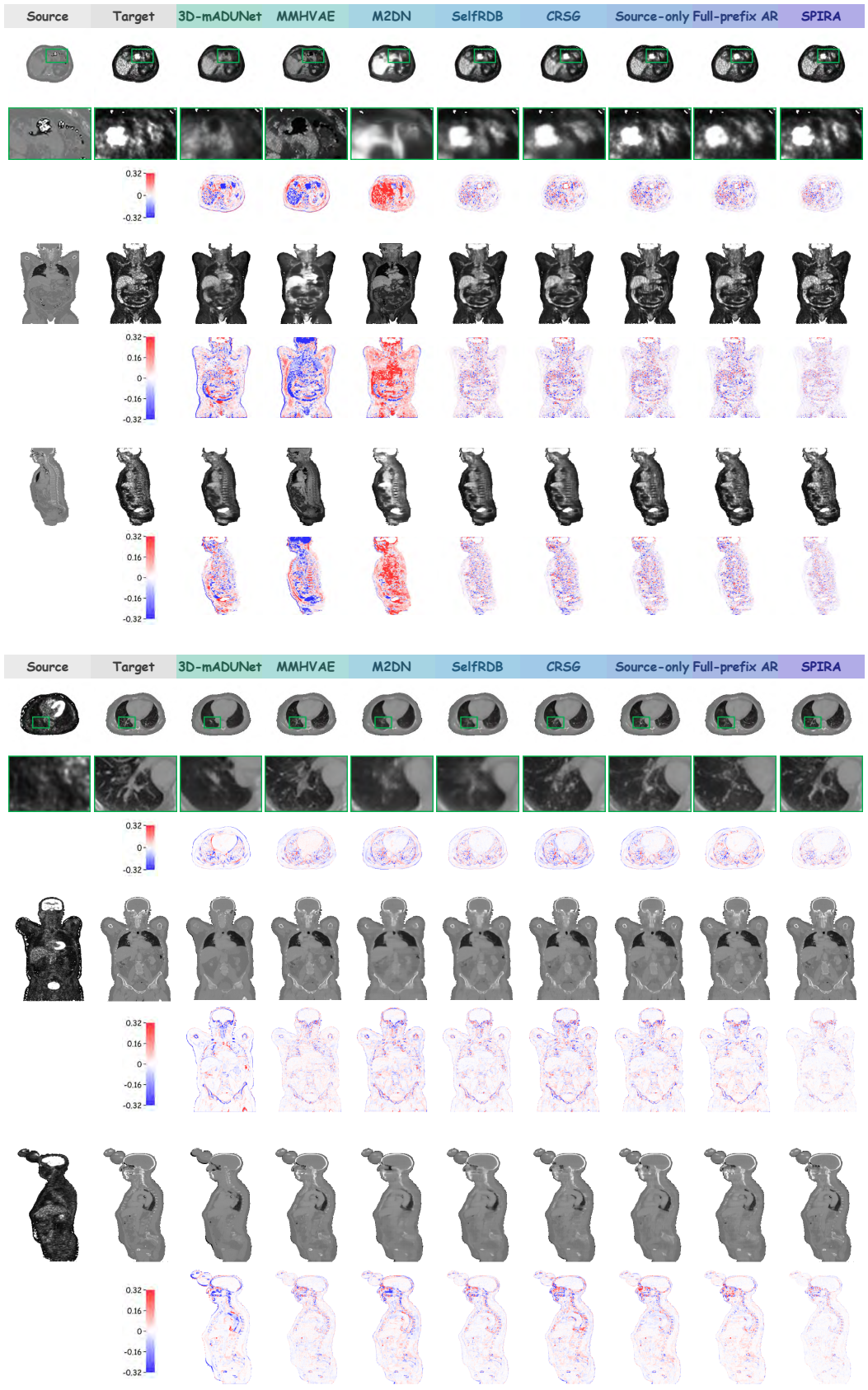

Figure 14: **Additional qualitative comparisons** on AutoPET CT→PET (upper) and PET→CT (lower), showing axial, coronal, and sagittal views from top to bottom with corresponding error maps.