CONTRAST-X: A MULTI-MODAL CONTRAST IMAGE SYNTHESIS BENCHMARK AND UNIVERSAL MODALITY FLOW MATCHING

Anonymous authors
Paper under double-blind review

ABSTRACT

Contrast-enhanced imaging is central to oncologic diagnosis, but contrast agents can be contraindicated for many of the patients who need them most. Synthesizing contrast scans from non-contrast inputs is the natural response. Two obstacles stand in the way: no benchmark provides paired contrast data with lesion-level evaluation, and no single model handles the arbitrary missing patterns seen in practice. We introduce Contrast-X, a benchmark of paired contrast-enhanced and non-contrast imaging spanning 10 organs in CT (1,526 patients) and multi-phase breast DCE-MRI (1,116 patients). Every case carries radiologist-verified phase labels and tumor masks. We further propose FlowMI, a single model that handles arbitrary subsets of available modalities through a unified multi-modal latent space and flow matching. We benchmark a range of missing-modality configurations, reporting standard image-quality metrics, radiologist reader studies, and downstream lesion analysis on the synthesized scans. We further evaluate cross-organ generalization to test whether the model has learned a transferable contrast-enhancement operation. Dataset, code, and leaderboard will be released.

1 INTRODUCTION

A radiology report is, in many cases, only as informative as the contrast agent that made it possible. Iodinated and gadolinium-based agents render tumors, vessels, and inflammation visible against tissue that would otherwise appear uniform. They are, by quiet consensus, the difference between a scan that diagnoses and a scan that merely shows.

Yet contrast is not a universally applicable resource. Patients with renal impairment, prior allergic reaction, hyperthyroidism, or pregnancy cannot safely receive it; many sites cannot reliably administer it; and even when both patient and site permit, scheduling, cost, and workflow constraints routinely produce studies in which the contrast portion is missing. The clinical consequence is a recurring asymmetry: some patients are imaged with the full diagnostic signal (with contrast), others without it.

Closing this asymmetry computationally has become an active research direction. If a non-contrast scan can be transformed into its contrast counterpart, the patient who cannot receive the agent is not penalized for that fact. Recent advances in image-to-image translation Atli et al. (2024); Chartsias et al. (2018), GANs Zhu et al. (2017); Isola et al. (2017), diffusion modeling Saharia et al. (2022); Xu et al. (2024); Rombach et al. (2022); Peebles & Xie (2023), and flow matching Lipman et al. (2023); Chen et al. (2025) produce synthetic medical images of plausible visual quality.

However, current contrast imaging tasks lack a benchmark, and contrast/non-contrast data (referred to as contrast pairs hereafter) is scarce and limited in scope. Many resources, such as AMOS Ji et al. (2022) and TCIA Clark et al. (2013), either lack pairing or have no contrast labels (i.e., it is unclear which images are contrast-enhanced and which are not). Furthermore, they rarely include associated lesion segmentation masks for further analysis and verification of the generated results.

A second limitation is methodological. Clinical scans frequently present with missing components in arbitrary configurations: sometimes only a single phase is available, sometimes two of three. The prevailing approach is to train a dedicated model for each configuration, requiring a separate network

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.

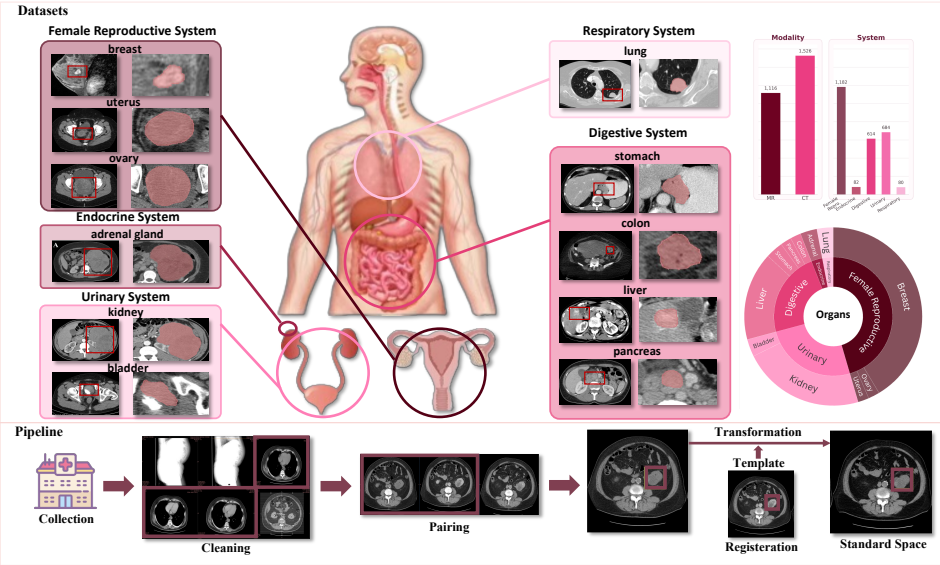

Figure 1: *Top:* Representative organ systems with paired contrast and non-contrast scans. The charts show organ-wise composition and modality balance. *Bottom:* Standardized curation pipeline including data collection, cleaning, pairing, and registration.

for every source–target combination. This scales poorly (on the order of 2^N models for N phases), making it impractical to maintain or deploy in real-world scenarios.

In this paper, we address these two limitations that hinder the development of multi-modal contrast image synthesis (CIS). We aim to build a unified benchmark for comprehensively evaluating translation-based generative models, along with a universal method (one model for all configurations) for these tasks. In summary, our contributions are as follows:

- We construct Contrast-X, covering pan-cancer contrast pairs across 11 organs from 2,642 patients. Every case includes tumor masks manually delineated by radiologists and subsequently verified by oncologists. It comprises two sets of challenges: the first is a multi-organ generation task using non-contrast and contrast-enhanced computed tomography (CT/CTC) across 10 organs; the second uses multi-phase dynamic contrast-enhanced magnetic resonance imaging (DCE-MRI) sequences to evaluate different missing-modality settings. Additionally, we validate the effectiveness of the generated images through downstream segmentation tasks.
- We introduce Flow Modal Imputation (FlowMI) to recast multi-modal contrast synthesis as a single transport problem in a shared latent space. Arbitrary subsets of available phases are encoded as distinct starting points, all flowing toward a common endpoint: the complete latent representation in which no modality is missing. Intermediate states along this trajectory are further constrained to semantically aligned with flow matching.

2 CONTRAST-X

2.1 WHAT CONTRAST-X CONTAINS

Contrast-X is assembled from public oncology imaging archives and restructured into a single benchmark with consistent format, registration, and labeling, all curated directed by senior radiologists. It contains 2,642 patients across 11 organs.

The benchmark is organized into two cohorts that probe different axes of contrast synthesis.

The CT cohort covers 1,526 patients across 10 organs frequently imaged in oncology workflows: adrenal gland, ovary, uterus, stomach, pancreas, liver, colon, bladder, kidney, and lung. Each study contains a non-contrast CT and its contrast-enhanced counterpart (CTC), representing the standard pre-/post-contrast acquisition used in abdominal and thoracic CT. The challenge here is anatomical:

Table 1: **Composition of Contrast-X.** The benchmark aggregates 25 public source collections into 2,642 paired cases spanning 11 organs and two domains (MR: 1,116 breast; CT: 1,526 across 10 organs). Sources are grouped by organ and listed in descending order of case count within each organ; organs themselves are ordered by total case count. All released under CC BY 3.0, except those marked † (CC BY 4.0). #Case denotes per-source case count.

Source Dataset	Organ	#Case	Source Dataset	Organ	#Case	Source Dataset	Organ	#Case
MR Track *— Breast, 1,116 cases*								
I-SPY 1	Breast	558	TCGA-BRCA	Breast	378	UCSF	Breast	180
CT Track *— 10 organs, 1,526 cases*								
TCGA-KIRC	Kidney	278	TCGA-STAD	Stomach	86	CPTAC-PDA	Pancreas	54
C4KC-KiTS	Kidney	216	Adrenal-ACC-Ki67-Seg†	Adrenal	82	CPTAC-UCEC	Uterus	40
CPTAC-CCRCC	Kidney	66	CMB-LCA†	Lung	28	TCGA-UCEC	Uterus	14
TCGA-KIRP	Kidney	26	CPTAC-LSCC	Lung	28	CMB-CRC	Colon	18
TCGA-KICH	Kidney	12	CPTAC-LUAD	Lung	10	TCGA-COAD	Colon	4
HCC-TACE-Seg	Liver	360	Lung-PET-CT-Dx†	Lung	6	TCGA-OV	Ovary	12
TCGA-LIHC	Liver	72	Anti-PD-1_Lung	Lung	6			
TCGA-BLCA	Bladder	86	TCGA-LUSC	Lung	2			

contrast uptake patterns differ markedly across organs, and a synthesis model is asked to generalize across them. The MR cohort covers 1,116 breast cases, each with a complete three-phase dynamic contrast-enhanced sequence (DCE1, DCE2, DCE3). Rather than a single pre-/post pair, DCE samples contrast uptake at successive time points, tracing the wash-in and wash-out kinetics that distinguish malignant from benign tissue. The challenge here is multi-phase: a synthesis model must learn the relationships across phases and recover any missing phase from the configurations that remain.

Every case in Contrast-X carries two annotations: a radiologist-verified contrast-phase label, and a voxel-level tumor segmentation mask. The phase label resolves an ambiguity that public collections frequently leave open or mislabel. The tumor mask enables evaluation criteria that operate at the level of the clinically meaningful lesion rather than the pixel grid. We return to this point in section 2.4.

2.2 How It Was Built

Contrast-X is constructed from publicly redistributable sources. The primary collections and cohorts are summarized in Table 1, with their licenses referenced.

Pairing. To identify pairs, all raw image series were parsed to inspect the three metadata fields most indicative of contrast state: `SeriesDescription`, `ContrastBolusAgent`, and `AcquisitionTime`. As an initial coarse filter, studies lacking a matchable contrast pair were excluded.

Labeling. Each remaining study then underwent fine-grained labeling and inspection by a team of trained annotators, with final confirmation by a senior radiologist for CT versus CTC and for DCE phase identity (DCE1, DCE2, DCE3). Cases with unresolved ambiguity, non-standard or distorted imaging, severe motion artifacts, or insufficient anatomical coverage were discarded. Tumor segmentation masks were inherited from the source dataset when available with sufficient quality and consistent with our protocol; otherwise, they were annotated by a radiologist.

Registration. Within each study, scan pairs were registered using the Elastix Klein et al. (2010) ITK backend. A four-level rigid and affine pyramid was applied to all studies, followed by a three-level B-spline deformable stage for organs prone to non-rigid motion (liver, lung), with mutual information as the similarity metric throughout. Each registration was visually inspected for anatomical consistency, and cases with residual misalignment were discarded.

Normalization. Finally, scan volumes were resampled to isotropic $1\,\mathrm{mm}^3$ spacing and cropped to organ-specific bounding boxes to remove irrelevant anatomy. CT intensities were windowed using organ-specific Hounsfield Unit ranges (e.g., $[-200, 300]$ HU for soft tissue) and then min-max normalized. MR intensities were standardized by per-scan z-score normalization to mitigate the inter-scanner variation typical of multi-site MR cohorts.

2.3 What Makes It Different

Medical image translation has progressed largely on benchmarks designed for adjacent problems, and none provides the per-patient contrast pairing across organs and modalities that clinical oncology

Table 2: Comparison of Contrast-X with existing medical imaging benchmarks. "CE Pair" denotes explicit per-patient pairing of contrast-enhanced and non-contrast scans; "Multi-Phase" indicates availability of multiple temporal contrast phases (e.g., DCE); "Mask" denotes availability of voxel-level segmentation annotations (anatomical structures or lesions); "#Organs" counts distinct organs covered; "Size" reports the number of patients. ● = full support, ○ = partial support, — = not provided. ∼ is the estimated number.

Benchmark	MR	CT	CE Pair	Multi-Phase	Mask	#Organs	Size	Primary Application
Brain-centric benchmarks								
BraTS-GLI 2024 de Verdier et al. (2024)	●	—	●	—	●	1	1,350	Post-treatment glioma seg.
BraSyn 2023 Li et al. (2024)	●	—	○	—	●	1	1,251	Missing-sequence synthesis
IXI IXI	●	—	—	—	—	1	∼600	Healthy brain reference
OASIS-3 LaMontagne et al. (2019)	●	—	—	—	○	1	1,378	Alzheimer's research
ADNI-3 Jack et al. (2008)	●	—	—	—	○	1	∼1,500	Alzheimer's research
AANLIB Summers (2003)	●	●	—	—	—	1	∼30	Normal & major brain diseases
Region-specific / cross-modality benchmarks								
CHAOS Valindria et al. (2018)	●	●	—	—	●	4	80	Healthy abdominal organ seg.
crossMoDA Dorent et al. (2023)	●	—	—	—	●	1	379	Cochlear domain adaptation
ACDC Bernard et al. (2018)	●	—	—	—	●	1	150	Cardiac diagnosis
MMWHS Zhuang (2018)	●	●	—	—	●	1	120	Cardiac structure seg.
SynthRAD2023 Thummerer et al. (2023)	●	●	●	—	○	2	1,080	MR-to-CT synthesis
Whole-body / PET benchmarks								
FDG-PET/CT Gatidis et al. (2022)	—	●	—	—	●	—	900	PET–CT tumor lesion analysis
Ultra-low Dose UDP (2024)	—	—	—	—	—	—	1,447	Low-dose PET synthesis
Multi-organ contrast-pair benchmarks								
Contrast-X (Ours)	●	●	●	●	●	**11**	**2,642**	**Pan-cancer contrast synthesis**

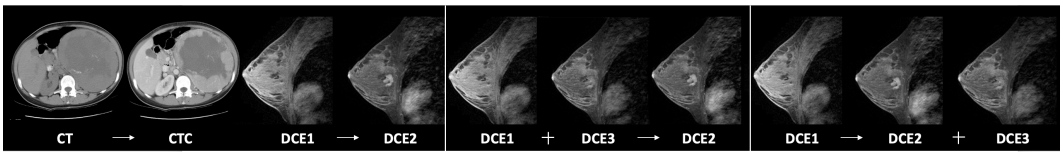

Figure 2: Representative tasks instantiated on Contrast-X. (a) CT → CTC: cross-organ contrast synthesis on the CT cohort. (b) $DCE_1 \rightarrow DCE_2$: neighbor-phase prediction on the MR cohort. (c) $DCE_{1,3} \rightarrow DCE_2$: recovery of an intermediate phase from its temporal neighbors. (d) $DCE_1 \rightarrow DCE_{2,3}$: synthesis of later phases from the earliest acquired phase. Settings (a) and (b) are *one-to-one*; (c) and (d) instantiate the more general case in which an arbitrary subset of modalities is observed and a disjoint subset is predicted (Table 2.4).

imaging requires. Brain MR datasets dominate the field: BraTS de Verdier et al. (2024) offers T1/T1ce pairs that have powered contrast-synthesis research, and IXI IXI, OASIS-3 LaMontagne et al. (2019), and ADNI Jack et al. (2008) have driven cross-sequence MR synthesis. Yet all are confined to the brain, and methods developed on them do not transfer to multi-organ settings. Beyond the brain, whole-body resources such as FDG-PET/CT Gatidis et al. (2022) and UDPET UDP (2024) achieve impressive scale, but they pair PET with CT rather than contrast pair, addressing a fundamentally different translation task. Cross-modality benchmarks like MMWHS Zhuang (2018) provide MR and CT data, but do not provide the supervised contrast pairing our setting requires.

Contrast-X fills this gap by combining the three properties no prior benchmark offers together: explicit per-patient pairing of contrast and non-contrast scans, broad anatomical coverage across both CT and MR, and tumor-level annotation for clinically meaningful evaluation. Table 2 summarizes the comparison along five axes.

Note. The term "multi-modal" carries different meanings across the literature, referring variously to distinct imaging modalities (e.g., MR and CT), multiple MR sequences within a study (as in BraTS), or multi-phase CT acquisitions. For clarity, throughout this paper we use *modality* to denote an individual MR sequence or CT phase, and *multi-domain* to denote the MR–CT setting.

2.4 What It Benchmarks

Contrast-X defines a family of contrast-synthesis tasks, which the two cohorts instantiate along complementary axes (Figure 2).

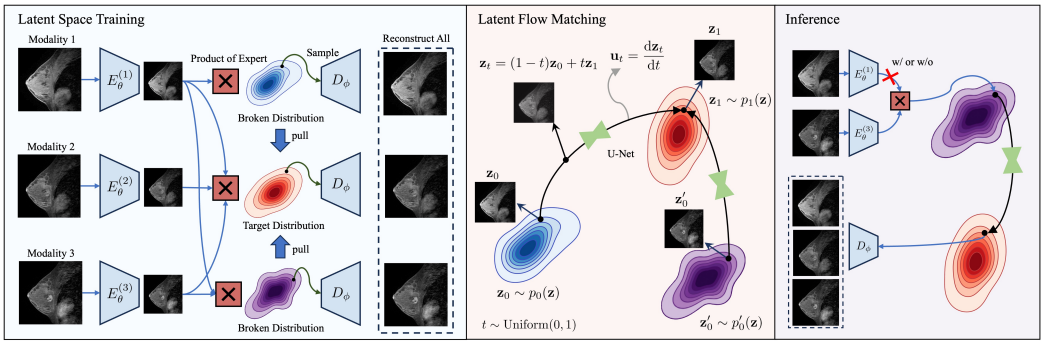

Figure 3: Overview of the proposed FlowMI framework. **Left:** Modality-specific encoders $E_\theta^{(i)}$ map inputs into a latent space, which are fused via a product-of-experts. The distribution with all modalities defines the *target*, while cases with missing modalities define the *broken* distribution. **Middle:** Latent flow matching learns a smooth mapping from $p_0(z)$ (broken) to $p_1(z)$ (target) using a U-Net parameterization of the velocity field, $\mathbf{u}_t = \frac{dz_t}{dt}$. **Right:** During inference, inputs with missing modalities are encoded and aligned through the learned flow, enabling consistent reconstruction or synthesis of complete modalities.

CT cohort: cross-organ contrast synthesis. The CT cohort defines a one-to-one task, CT $\rightarrow$ CTC, in which a non-contrast scan predicts its contrast-enhanced counterpart. It tests whether a single model can learn a contrast mapping that generalizes across anatomy rather than memorizing organ-specific behavior. We evaluate under two protocols: a full-organ setting that trains and tests on all organs jointly, and a leave-one-organ-out setting that probes generalization to unseen anatomy.

MR cohort: multi-phase DCE synthesis. The MR cohort captures the temporal dynamics of contrast enhancement through three tasks over the phases DCE_1, DCE_2, DCE_3 (collectively denoted $\text{DCE}_{1,2,3}$). $\text{DCE}_1 \rightarrow \text{DCE}_2$ tests direct one-to-one prediction, mirroring the CT $\rightarrow$ CTC setting. $\text{DCE}_{1,3} \rightarrow \text{DCE}_2$ and $\text{DCE}_1 \rightarrow \text{DCE}_{2,3}$ reflect the different clinical scenarios in which only a subset of phases is acquired and the remainder must be synthesized. We exploit this view in Section 3 to motivate a single model trained jointly across all configurations.

Splits. We partition Contrast-X at the organ-level into training (70%), validation (10%), and test (20%) sets, using stratified sampling to preserve the distribution over organ systems and cancer types. Main results in Section 4 and ablations are reported on the full test set.

3 Flow Modality Imputation (FlowMI)

We propose **FlowMI**, which unifies autoencoding and flow matching into a single framework for missing-modality imputation. The core idea is to treat any partial-modality configuration as a starting latent and the full-modality observation as the endpoint, with flow matching transporting one to the other. We further shape the latent space so that intermediate states along the flow remain meaningful, aligning the trajectory with the geometry of the latent manifold. We want to highlight that our contribution is not the use of flow matching or Product-of-Experts themselves, but rather the reformulation of these modules and the proposed training scheme that together make FlowMI a unified model capable of handling arbitrary observation configurations.

3.1 Preliminary: Flow Matching

Flow Matching (FM) Lipman et al. (2023); Liu et al. (2022) learns a continuous transport map between two distributions π_0 and π_1 over a space $\mathcal{Z} \subseteq \mathbb{R}^d$ from samples alone. A time-dependent vector field $v_\theta(\mathbf{z}, t)$ is trained to follow a prescribed interpolation path $\mathbf{z}_t$ between paired endpoints $(\mathbf{z}_0, \mathbf{z}_1) \sim \pi_0 \times \pi_1$ by regressing the analytic target velocity:

$$\mathcal{L}_{\text{FM}}(\theta) \;=\; \mathbb{E}_{t,\,(\mathbf{z}_0,\mathbf{z}_1)} \left[\left\| v_\theta(\mathbf{z}_t, t) - \dot{\mathbf{z}}_t \right\|^2 \right], \tag{1}$$

with $t \sim \mathcal{U}(0,1)$. We adopt straight-line interpolation $\mathbf{z}_t = (1-t)\mathbf{z}_0 + t\mathbf{z}_1$, for which the target velocity reduces to the constant $\mathbf{z}_1 - \mathbf{z}_0$. Once trained, v_θ transports samples by integrating $\mathrm{d}\mathbf{z}_t/\mathrm{d}t =$

$v_\theta(\mathbf{z}_t, t)$ from $t = 0$ to 1. Latent Flow Matching Dao et al. (2023); Chadebec et al. (2025) applies this construction inside the latent space of a pretrained autoencoder. FlowMI extends this construction to a clean missing-modality imputation regime: both endpoints are produced by *the same autoencoder under different observation masks*, casting any missing-modality imputation as a transport problem on a unify learned manifold.

3.2 Unified Latent Space via Product-of-Experts Autoencoder

A latent space that supports flow transport across observation regimes should satisfy two requirements: it can accept any subset of modalities without architectural changes, and it can provide a continuous manifold aligned with the flow trajectory. We meet both requirements with a unified latent space design.

For modality flexibility, we adopt a multi-modal VAE with modality-specific encoders and a shared decoder, and fuse the encoded distributions via a Product-of-Experts (PoE) Wu & Goodman (2018) to yield a latent space of fixed dimensionality regardless of which modality configuration is observed.

Let a multi-modal sample $\mathbf{x} = \{x^{(1)}, \ldots, x^{(M)}\}$ contains M modalities, indexed by $\mathcal{M} = \{1, \ldots, M\}$. A binary mask $\mathbf{m} \in \{0, 1\}^M$ encodes observability, with $\mathbf{m}^{(i)} = 1$ if modality i is observed and 0 otherwise. We write $\mathcal{O}(\mathbf{m}) = \{i : \mathbf{m}^{(i)} = 1\}$ for the observed index set, $\mathbf{x}^{\mathbf{m}}$ for the observed subset, and $\mathbf{x}^{\mathbb{1}}$ for the fully observed sample, where $\mathbb{1}$ is the all-ones mask. The objective is to predict $\mathbf{x}^{\neg\mathbf{m}}$ from $\mathbf{x}^{\mathbf{m}}$.

For each modality m, an encoder $E_\theta^{(m)}$ outputs a Gaussian expert capturing what modality m alone reveals about the shared latent $\mathbf{z} \in \mathbb{R}^d$:

$$q_\theta^{(m)}(\mathbf{z} \mid x^{(m)}) = \mathcal{N}\big(\mathbf{z};\, \boldsymbol{\mu}^{(m)},\, \mathrm{diag}(\boldsymbol{\sigma}^{(m)})^2\big), \qquad (\boldsymbol{\mu}^{(m)}, \boldsymbol{\sigma}^{(m)}) = E_\theta^{(m)}(x^{(m)}). \tag{2}$$

Given an observation mask $\mathbf{m}$, we fuse only the experts of observed modalities with the prior $p(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \mathbf{I})$:

$$q_\theta(\mathbf{z} \mid \mathbf{x}^{\mathbf{m}}) \propto p(\mathbf{z}) \prod_{m \in \mathcal{O}(\mathbf{m})} q_\theta^{(m)}(\mathbf{z} \mid x^{(m)}). \tag{3}$$

The prior plays two roles: it keeps the posterior well-defined for any mask (including when few or no modalities are observed) and provides the standard VAE regularization target. Since the product of Gaussians is Gaussian, the posterior has closed form,

$$\boldsymbol{\Lambda}^{\mathbf{m}} = \mathbf{I} + \sum_{m \in \mathcal{O}(\mathbf{m})} \mathrm{diag}(\boldsymbol{\sigma}^{(m)})^{-2}, \qquad \boldsymbol{\mu}^{\mathbf{m}} = (\boldsymbol{\Lambda}^{\mathbf{m}})^{-1} \sum_{m \in \mathcal{O}(\mathbf{m})} \mathrm{diag}(\boldsymbol{\sigma}^{(m)})^{-2}\, \boldsymbol{\mu}^{(m)}, \tag{4}$$

so each expert contributes proportionally and the prior anchors the posterior when evidence is sparse.

We then sample the fused code via reparameterization, $\mathbf{z}^{\mathbf{m}} = \boldsymbol{\mu}^{\mathbf{m}} + (\boldsymbol{\Lambda}^{\mathbf{m}})^{-1/2}\boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ at training, or take $\mathbf{z}^{\mathbf{m}} = \boldsymbol{\mu}^{\mathbf{m}}$ at inference. A single decoder D_ϕ then maps $\mathbf{z}^{\mathbf{m}}$ to image space for reconstruction.

The autoencoding process ensures that the endpoints $\mathbf{z}_0$ and $\mathbf{z}_1$ are meaningful, but the flow trajectory between them is not. In practice, latent flow matching never lands exactly on $\mathbf{z}_1$: the integrated trajectory drifts, and any drift in latent space is amplified by the decoder into unpredictable image artifacts. The only way to make this drift tolerable is to ensure that every point along the path, not just the endpoints, decodes to something meaningful. We therefore require the latent trajectory to mirror a known, well-behaved trajectory in image space.

Let $\mathbf{x}_0$ denote a missing-modality observation and $\mathbf{x}_1$ its complete-modality counterpart, with corresponding latents $\mathbf{z}_0$ and $\mathbf{z}_1$ obtained from the PoE encoder. We construct the image-space interpolant $\mathbf{x}_t = (1-t)\mathbf{x}_0 + t\mathbf{x}_1$ and the latent interpolant $\mathbf{z}_t = (1-t)\mathbf{z}_0 + t\mathbf{z}_1$, and require that $D_\phi(\mathbf{z}_t)$ recovers $\mathbf{x}_t$ for every $t \in [0, 1]$:

$$\mathcal{L}_{\text{Trajectory}} = \mathbb{E}_{\mathbf{x},\, \mathbf{m},\, t}\Big[\, \big\| D_\phi(\mathbf{z}_t) - \mathbf{x}_t \big\|_2^2 \,\Big]. \tag{5}$$

In practice, we apply a stop-gradient on $\mathbf{z}_0$ and $\mathbf{z}_1$ when forming $\mathbf{z}_t$ to stabilize training. This per-step reconstruction constraint anchors the entire latent path to image-space ground truth, so the flow has

Table 3: Quantitative comparison on the Contrast-X CT and MR paired pan-cancer contrast media dataset. Methods are grouped by generative mechanism (Direct regression, GAN, Diffusion, Flow Matching) and backbone architecture. PSNR (dB) and SSIM (%) are reported as Mean±Std across the test cohort. **Bold** marks the best result per column; underline marks the second-best. *Trans: Transformer; Arch.: Architecture.

Mechanism	Arch.	Method	CT → CTC		$DCE_1 \rightarrow DCE_2$		$DCE_1 \rightarrow DCE_{2,3}$		$DCE_{1,3} \rightarrow DCE_2$	
			PSNR↑	SSIM(%)↑	PSNR↑	SSIM(%)↑	PSNR↑	SSIM(%)↑	PSNR↑	SSIM(%)↑
Direct	UNet	HiNet Zhou et al. (2020)	22.23 ±3.70	77.7 ±10.3	23.41 ±3.48	68.1 ±11.9	23.88 ±3.37	72.9 ±8.32	26.47 ±3.32	70.2 ±12.2
	Trans	ResViT Dalmaz et al. (2022)	20.80 ±3.11	74.2 ±10.2	25.12 ±3.16	70.5 ±10.5	24.65 ±2.81	63.5 ±10.8	25.24 ±2.89	59.8 ±15.0
	Mamba	I2IMamba Atli et al. (2024)	20.97 ±3.20	74.6 ±10.3	23.25 ±2.37	55.9 ±14.5	23.57 ±2.50	58.0 ±13.0	26.82 ±3.32	70.1 ±11.8
GAN	UNet	CycleGAN Zhu et al. (2017)	21.90 ±3.98	75.8 ±13.2	24.18 ±3.33	65.7 ±13.3	24.46 ±3.52	71.4 ±12.0	25.74 ±3.28	72.4 ±12.2
	UNet	Pix2Pix Isola et al. (2017)	21.39 ±3.10	72.6 ±13.8	23.24 ±3.35	64.6 ±14.3	23.90 ±2.97	72.2 ±8.95	26.42 ±3.10	70.2 ±10.4
Diffusion	UNet	Palette Saharia et al. (2022)	15.88 ±5.30	33.2 ±21.8	9.54 ±2.17	9.70 ±6.38	10.19 ±1.99	14.5 ±6.16	4.79 ±1.25	5.47 ±2.55
	UNet	SelfRDB Arslan et al. (2025)	17.16 ±2.38	72.5 ±7.90	23.40 ±3.31	57.6 ±16.4	24.98 ±2.87	67.6 ±13.5	26.66 ±3.79	70.2 ±11.9
Flow Matching	UNet	ConcatFM Lipman et al. (2023)	23.10 ±4.12	77.1 ±9.28	26.31 ±2.67	72.9 ±6.61	26.25 ±2.49	71.8 ±6.18	29.06 ±2.73	76.1 ±6.63
	UNet	DirectFM Lipman et al. (2023)	22.84 ±3.88	76.7 ±11.4	25.74 ±3.14	68.5 ±8.09	25.98 ±3.05	71.1 ±8.22	27.90 ±3.18	74.2 ±7.69
	UNet	PMRF Ohayon et al. (2024)	21.91 ±3.95	76.6 ±11.6	25.06 ±3.60	67.5 ±8.28	26.11 ±3.80	70.1 ±8.32	27.53 ±4.25	74.9 ±5.31
	UNet	**FlowMI (Ours)**	**24.47** ±4.15	**78.5** ±8.62	**26.52** ±3.13	**74.2** ±9.33	**26.63** ±3.11	**73.7** ±8.06	**29.17** ±3.24	**76.2** ±7.65

no opportunity to drift into a region the decoder cannot interpret. Every intermediate $\mathbf{z}_t$ remains semantically meaningful and decodable, and the latent trajectory becomes a faithful counterpart to the image-space flow rather than an unconstrained curve through latent space.

3.3 Latent Flow Modality Imputation

The autoencoder defines, for every sample $\mathbf{x}$ and every mask $\mathbf{m}$, a paired endpoint structure $(\mathbf{z}^{\mathbf{m}}, \mathbf{z}^{\mathbb{1}})$ on a shared latent manifold. We model imputation as a flow between these endpoints, conditioned on the mask so that a single flow field handles arbitrary missing patterns. Concretely, we define a mask-conditioned vector field $v_\theta(\mathbf{z}, t, \mathbf{m})$ and use straight-line interpolation between endpoints,

$$\mathbf{z}_t = (1-t)\,\mathbf{z}^{\mathbf{m}} + t\,\mathbf{z}^{\mathbb{1}}, \tag{6}$$

for $t \in [0,1]$, training the flow by regressing the constant target velocity:

$$\mathcal{L}_{\text{LFM}} = \mathbb{E}_{\mathbf{x},\,\mathbf{m},\,t}\left[\,\left\|v_\theta(\mathbf{z}_t,\,t,\,\mathbf{m}) - (\mathbf{z}^{\mathbb{1}} - \mathbf{z}^{\mathbf{m}})\right\|^2\right], \tag{7}$$

where $\mathbf{m}$ is sampled uniformly over non-empty subsets of $\mathcal{M}$ at each step.

Inference. Given a test input $\mathbf{x}^{\mathbf{m}}$ with mask $\mathbf{m}$, we encode the partial-observation latent $\mathbf{z}_0 = E_\theta(\mathbf{x}^{\mathbf{m}}, \mathbf{m})$ and integrate the learned ODE $\mathrm{d}\mathbf{z}_t/\mathrm{d}t = v_\theta(\mathbf{z}_t, t, \mathbf{m})$ from $t = 0$ to $t = 1$ using a standard Euler scheme,

$$\mathbf{z}_{t+\Delta t} \;=\; \mathbf{z}_t + \Delta t \cdot v_\theta(\mathbf{z}_t, t, \mathbf{m}), \qquad t = 0, \Delta t, \ldots, 1 - \Delta t. \tag{8}$$

The terminal state $\hat{\mathbf{z}}^{\mathbb{1}} := \mathbf{z}_1$ approximates the full-observation latent, which the decoder maps to a complete reconstruction $\hat{\mathbf{x}}^{\mathbb{1}} = D_\phi(\hat{\mathbf{z}}^{\mathbb{1}})$. The imputed missing modalities $\hat{\mathbf{x}}^{\neg\mathbf{m}}$ are read off from the corresponding output channels.

4 Experiments

We benchmark four categories of generative models on Contrast-X, covering both multimodal translation and missing-modality synthesis: (i) Direct regression: HiNet Zhou et al. (2020), ResViT Dalmaz et al. (2022), and I2IMamba Atli et al. (2024); (ii) GAN-based: CycleGAN Zhu et al. (2017) and Pix2Pix Isola et al. (2017); (iii) Diffusion-based: Palette Saharia et al. (2022) and SelfRDB Arslan et al. (2025); (iv) Flow matching: ConcatFM Lipman et al. (2023), DirectFM Lipman et al. (2023), PMRF Ohayon et al. (2024), and our FlowMI. We use each method's public implementation throughout.

Table 4: LPIPS, FID, and KID comparison across methods on Contrast-X. Methods are grouped by generative mechanism and backbone architecture. **Bold** marks the best result per column; underline marks the second-best. *Trans: Transformer; Arch.: Architecture.

Mechanism	Arch.	Method	CT → CTC			$DCE_1 \rightarrow DCE_2$			$DCE_1 \rightarrow DCE_{2,3}$			$DCE_{1,3} \rightarrow DCE_2$		
			LPIPS↓	FID↓	KID↓	LPIPS↓	FID↓	KID↓	LPIPS↓	FID↓	KID↓	LPIPS↓	FID↓	KID↓
Direct	UNet	HiNet Zhou et al. (2020)	0.084	0.01	14.62	0.181	0.02	37.68	0.163	0.01	30.63	0.113	0.01	18.83
	Trans	ResViT Dalmaz et al. (2022)	0.109	0.01	17.53	0.224	0.02	43.18	0.146	0.01	18.31	0.142	0.01	31.08
	Mamba	I2IMamba Atli et al. (2024)	0.099	**0.00**	12.21	0.217	0.02	40.17	0.181	0.01	28.88	0.099	**0.00**	**12.21**
GAN	UNet	CycleGAN Zhu et al. (2017)	0.079	**0.00**	**8.98**	0.150	0.01	20.88	0.159	0.01	22.25	0.114	0.01	21.40
	UNet	Pix2Pix Isola et al. (2017)	0.109	0.01	22.70	0.218	0.02	48.92	0.186	0.01	36.21	0.127	0.01	32.88
Diffusion	UNet	Palette Saharia et al. (2022)	0.216	0.03	41.19	0.755	0.32	270.9	0.747	0.29	255.7	0.929	0.52	412.3
	UNet	SelfRDB Arslan et al. (2025)	0.189	0.07	78.80	0.254	0.03	53.98	0.203	0.03	50.24	0.186	0.02	37.30
Flow Matching	UNet	ConcatFM Lipman et al. (2023)	0.080	0.01	18.47	0.137	**0.01**	19.18	0.128	0.02	18.44	0.109	0.02	20.67
	UNet	DirectFM Lipman et al. (2023)	0.074	0.01	15.08	0.214	0.03	26.74	0.201	0.03	28.83	0.169	0.02	26.41
	UNet	PMRF Ohayon et al. (2024)	0.090	0.02	24.74	0.212	0.02	28.37	0.149	0.02	22.64	0.214	0.02	22.18
	UNet	**FlowMI (Ours)**	**0.065**	0.01	11.55	**0.120**	**0.01**	**16.74**	**0.101**	**0.01**	**14.63**	**0.090**	0.01	12.85

4.1 FlowMI Implementation Details

Model Architecture. FlowMI consists of two stages: a 3D variational autoencoder (VAE) that maps volumes into a unified latent space, and a flow matching module that transports latents conditioned on the observation mask. Both stages use a 3D U-Net backbone with depth 4, residual blocks at each level, and group normalization. The VAE produces a latent representation with 256 channels; the flow module uses an identically structured U-Net to predict the continuous velocity field over latents.

Training Configuration. We train each stage with AdamW at a learning rate of 1×10^{-4}, using cosine annealing with linear warm-up over the first 5% of steps. Due to the memory cost of 3D volumes, we use a per-GPU batch size of 2 with gradient accumulation, yielding an effective batch size of 8. All models are trained for 200 epochs unless otherwise specified. Following standard continuous-time flow matching, we sample $t \sim \mathcal{U}(0, 1)$ uniformly at each training step. All experiments are run on NVIDIA A6000 GPUs. Notably, FlowMI uses a single shared model (one VAE and one flow module) across all three DCE tasks, whereas competing methods require a separate model for each setting.

Metrics. We evaluate synthesis quality at three levels. At the image level, we report 5 standard fidelity and perceptual metrics: Peak Signal-to-Noise Ratio (PSNR), Structural Similarity Index Measure (SSIM), Kernel Inception Distance (KID), Fréchet Inception Distance (FID), and Learned Perceptual Image Patch Similarity (LPIPS). At the lesion level, the tumor segmentation masks released with Contrast-X enable measuring whether synthesized lesions remain faithful to the ground truth, reported via downstream segmentation performance. Finally, we conduct a reader study in which users subjectively assess image quality and diagnostic plausibility.

4.2 Experiment Results

Quantitative comparison. Tables 3 and 4 show that flow matching methods consistently outperform the direct, GAN, and diffusion baselines, with FlowMI achieving the best PSNR, SSIM, and LPIPS, and competitive FID and KID across all tasks, with the largest margin on CT→CTC. On the three DCE tasks, the gain over the second-best baseline (ConcatFM) is more modest in absolute terms but more meaningful in practice: ConcatFM and the other baselines train a separate model for each DCE setting, whereas FlowMI handles all three settings with a single shared model.

Qualitative comparison. Figure 4 reinforces the quantitative findings. Most baselines produce visually plausible CT images but leave tumor regions under-enhanced, making lesions hard to distinguish. FlowMI instead reproduces the bright enhancement seen in the ground-truth CTC, and the residual maps (blue) confirm correspondingly lower reconstruction error in tumor regions.

Downstream lesion analysis. Beyond image-level metrics, we evaluate whether the synthesized contrast scans preserve the lesion information needed for downstream analysis. We train organ-specific tumor segmentation models on real ground-truth contrast scans, then apply them to the synthesized outputs of each method and measure patient-level Dice against the released tumor masks.

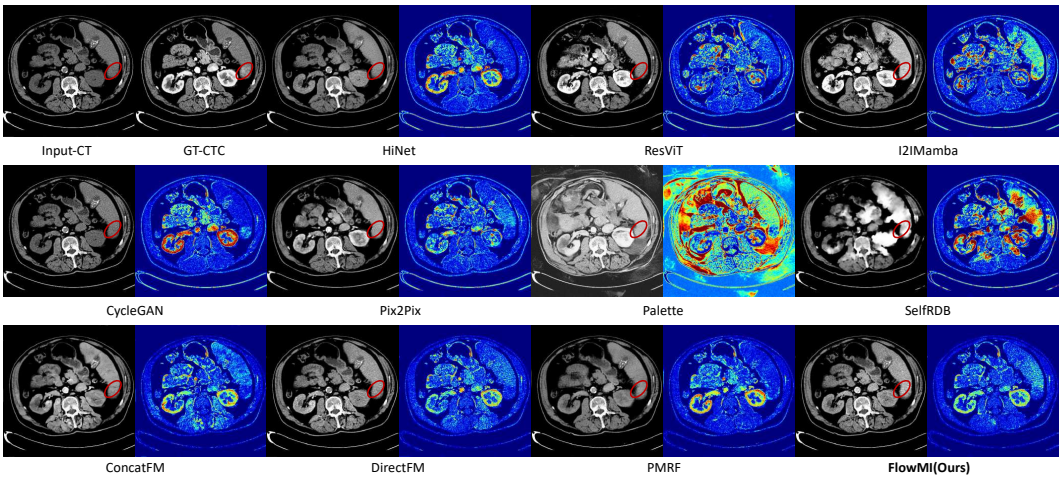

Figure 4: CT→CTC (liver). Red circles mark tumor regions.

Table 5: Downstream tumor segmentation performance. We report patient-level Dice mean ± Std averaged across all CT organs. The CT baseline is trained and tested on real CT; all synthesis methods and the ground-truth CTC upper bound use the same organ-specific nnU-Net trained on real CTC.

Method	CT	CycleGAN	Pix2Pix	HiNet	ResViT	I2I-Mamba	Palette	SelfRDB	ConcatFM	DirectFM	PMRF	FlowMI (Ours)	GT CTC
Dice ↑	0.5312 ± 0.1483	0.6065 ± 0.1094	0.6041 ± 0.1153	0.6118 ± 0.1036	0.6008 ± 0.1127	0.6157 ± 0.1018	0.5238 ± 0.1761	0.5827 ± 0.1284	0.6248 ± 0.0927	0.6216 ± 0.0941	0.6193 ± 0.0975	**0.6315 ± 0.0893**	**0.6349 ± 0.0851**

The score on real ground-truth scans serves as an upper bound; the closer a method's score approaches this bound, the more faithfully its synthesized images preserve clinically actionable lesion structure. Table 5 reports the mean and standard deviation across all CT organ categories. FlowMI achieves the best synthetic-image result and approaches the real-CECT upper bound.

Leave-organs-out generalization. To test whether FlowMI learns a transferable contrast-enhancement operation, we hold out organs from training and evaluate on the held-out organ cases, comparing against the two strong baselines. We use four organs spanning tissue types: Colon, Ovary, Uterus, and Pancreas. As shown in Table 6, FlowMI achieves the highest PSNR and SSIM on every held-out organ, with the largest margin on Colon (+2.16 PSNR, +6.5 SSIM). The roughly 5 dB PSNR drop and higher standard deviation relative to the in-distribution setting (Table 3) is expected for unseen anatomy, but FlowMI's consistent advantage across all four organs indicates that the latent transport generalizes rather than memorizes.

Table 6: Leave-organs-out generalization on the CT→CTC task. Each column holds out one organ from training and evaluates on the held-out cases. PSNR (dB) and SSIM (%) reported as Mean±Std.

Method	**Colon** (22)		**Ovary** (12)		**Uterus** (54)		**Pancreas** (54)		**Average**	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
ConcatFM	17.12 ±8.18	65.3 ±15.4	16.08 ±8.27	59.1 ±17.9	18.94 ±7.83	61.7 ±13.6	17.85 ±7.49	62.2 ±15.1	17.50	62.1
DirectFM	16.71 ±8.42	58.9 ±16.2	17.63 ±8.91	60.4 ±16.8	17.62 ±7.07	59.8 ±14.3	17.46 ±7.13	63.5 ±14.7	17.36	60.7
FlowMI (Ours)	**19.28** ±7.71	**71.8** ±10.8	**18.94** ±7.63	**63.7** ±15.4	**20.41** ±6.52	**67.9** ±12.9	**19.87** ±6.89	**66.2** ±13.5	**19.63**	**67.4**

5 Conclusion

In this work, we introduced Contrast-X, a comprehensive pan-cancer benchmark for multimodal image translation and missing-modality synthesis in clinically realistic settings. It provides high-quality, per-patient paired CT and MR scans across 11 organs with both contrast-enhanced and non-contrast modalities, curated through a standardized and reproducible pipeline. We defined three benchmark tasks and reported reference results using representative generative models. While limitations remain, such as biases in public data sources and variability in clinical imaging, Contrast-X offers a foundation for developing robust multimodal translation methods and advancing clinically reliable decision support. Future work could extend the dataset with additional modalities, larger cohorts, and more advanced synthesis techniques to further enhance generalization and applicability.

REFERENCES

Ixi dataset (rrid:scr_005839). http://brain-development.org/ixi-dataset/. 4

Ultra-low dose pet imaging challenge 2024 (udpet). Dataset and challenge information available online, 2024. Accessed via https://udpet-challenge.github.io/. 4

Oguz Akin, Pierre Elnajjar, Matthew Heller, Rose Jarosz, Bradley J Erickson, Shanah Kirk, Yueh Lee, Marston W Linehan, Rabindra Gautam, Raghu Vikram, et al. The cancer genome atlas kidney renal clear cell carcinoma collection (tcga-kirc)(version 3)[data set]. *Cancer Imaging Arch*, 2016. 16

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. 7, 8

Omer F Atli, Bilal Kabas, Fuat Arslan, Arda C Demirtas, Mahmut Yurt, Onat Dalmaz, and Tolga Cukur. I2i-mamba: Multi-modal medical image synthesis via selective state space modeling. *arXiv preprint arXiv:2405.14022*, 2024. 1, 7, 8

Olivier Bernard, Alain Lalande, Clement Zotti, Frederick Cervenansky, Xin Yang, Pheng-Ann Heng, Irem Cetin, Karim Lekadir, Oscar Camara, Miguel Angel Gonzalez Ballester, et al. Deep learning techniques for automatic mri cardiac multi-structures segmentation and diagnosis: is the problem solved? *IEEE transactions on medical imaging*, 37(11):2514–2525, 2018. https://ieeexplore.ieee.org/document/8360453. 4

C Biobank. Cancer moonshot biobank-lung cancer collection (cmb-lca)(version 3)[dataset]. *The Cancer Imaging Archive*, 2022a. 16

Cancer Moonshot Biobank. Cancer moonshot biobank – colorectal cancer collection (cmb-crc) (version 8) [data set], 2022b. The Cancer Imaging Archive. https://doi.org/10.7937/djg7-gz87. 16

Clément Chadebec, Onur Tasar, Sanjeev Sreetharan, and Benjamin Aubin. Lbm: Latent bridge matching for fast image-to-image translation. *arXiv preprint arXiv:2503.07535*, 2025. 6

Agisilaos Chartsias, Thomas Joyce, Mario Valerio Giuffrida, and Sotirios A. Tsaftaris. Multimodal mr synthesis via modality-invariant latent representation. *IEEE Transactions on Medical Imaging*, 37:803–814, 3 2018. ISSN 1558254X. doi: 10.1109/TMI.2017.2764326. 1

Hao Chen, Rui Yin, Yifan Chen, Qi Chen, and Chao Li. Learning patient-specific disease dynamics with latent flow matching for longitudinal imaging generation. *arXiv preprint arXiv:2512.09185*, 2025. 1

Kenneth Clark, Bruce Vendt, Kirk Smith, John Freymann, Justin Kirby, Paul Koppel, Stephen Moore, Stanley Phillips, David Maffitt, Michael Pringle, Lawrence Tarbox, and Fred Prior. The cancer imaging archive (tcia): maintaining and operating a public information repository, 2013. TCIA website: https://www.cancerimagingarchive.net/. 1

National Cancer Institute Clinical Proteomic Tumor Analysis Consortium et al. The clinical proteomic tumor analysis consortium lung squamous cell carcinoma collection (cptac-lscc). *The Cancer Imaging Archive*, 2018a. 16

National Cancer Institute Clinical Proteomic Tumor Analysis Consortium et al. The clinical proteomic tumor analysis consortium lung adenocarcinoma collection (cptac-luad). *The Cancer Imaging Archive.*, 2018b. 16

National Cancer Institute Clinical Proteomic Tumor Analysis Consortium (CPTAC). The clinical proteomic tumor analysis consortium clear cell renal cell carcinoma collection (cptac-ccrcc), 2018. The Cancer Imaging Archive. https://doi.org/10.7937/k9/tcia.2018.oblamn27. 16

Onat Dalmaz, Mahmut Yurt, and Tolga Çukur. Resvit: Residual vision transformers for multimodal medical image synthesis. *IEEE Transactions on Medical Imaging*, 41(10):2598–2614, 2022. 7, 8

Quan Dao, Hao Phung, Binh Nguyen, and Anh Tran. Flow matching in latent space. *arXiv preprint arXiv:2307.08698*, 2023. 6

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. https://arxiv.org/abs/2405.18368. 4

Reuben Dorent, Aaron Kujawa, Marina Ivory, Spyridon Bakas, Nicola Rieke, Samuel Joutard, Ben Glocker, Jorge Cardoso, Marc Modat, Kayhan Batmanghelich, et al. Crossmoda 2021 challenge: Benchmark of cross-modality domain adaptation techniques for vestibular schwannoma and cochlea segmentation. *Medical Image Analysis*, 83:102628, 2023. https://www.sciencedirect.com/science/article/pii/S1361841522002560. 4

BJ Erickson, D Mutch, L Lippmann, and R Jarosz. The cancer genome atlas uterine corpus endometrial carcinoma collection (tcga-ucec). *The Cancer Imaging Archive*, 2016a. 15

Bradley J Erickson, Shanah Kirk, Y Lee, Oliver Bathe, Melissa Kearns, C Gerdes, Kimberly Rieger-Christ, and John Lemmerman. The cancer genome atlas liver hepatocellular carcinoma collection (tcga-lihc)(version 5)[data set]. *The Cancer Imaging Archive*, 2016b. 16

Sergios Gatidis, Tobias Hepp, Marcel Früh, Christian La Fougère, Konstantin Nikolaou, Christina Pfannenberg, Bernhard Schölkopf, Thomas Küstner, Clemens Cyran, and Daniel Rubin. A whole-body fdg-pet/ct dataset with manually annotated tumor lesions. *Scientific Data*, 9(1):601, 2022. https://www.scopus.com/pages/publications/85139504476. 4

Nicholas Heller, Nithesh Sathianathen, Arveen Kalapara, Ethan Walczak, Kenneth Moore, Holly Kaluzniak, Jacob Rosenberg, Paul Blake, Zachary Rengel, Michael Oestreich, Joel Dean, Matthew Tradewell, Adeel Shah, Rishi Tejpaul, Zachary Edgerton, Matthew Peterson, Sohaib Raza, Samip Regmi, Nikolaos Papanikolopoulos, and Christopher Weight. Data from c4kc-kits [data set], 2019. The Cancer Imaging Archive. https://doi.org/10.7937/TCIA.2019.IX49E8NX. 16

Chandra Holback, Rose Jarosz, Fred Prior, David G Mutch, Priya Bhosale, Kimberly Garcia, Yueh Lee, Shanah Kirk, Cheryl A Sadow, Seth Levine, et al. The cancer genome atlas ovarian cancer collection (tcga-ov)(version 4)[data set]. *The Cancer Imaging Archive*, 10:K9, 2016. 15

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1125–1134, 2017. 1, 7, 8

Clifford R. Jr. Jack, Matt A. Bernstein, Nick C. Fox, Paul Thompson, Gene Alexander, Danielle Harvey, Bret Borowski, Paula J. Britson, Jennifer L. Whitwell, Chad Ward, et al. The alzheimer's disease neuroimaging initiative (adni): Mri methods. *Journal of Magnetic Resonance Imaging*, 27 (4):685–691, 2008. https://doi.org/10.1002/jmri.21049. 4

Nazia F Jafri, David C Newitt, John Kornak, Laura J Esserman, Bonnie N Joe, and Nola M Hylton. Optimized breast mri functional tumor volume as a biomarker of recurrence-free survival following neoadjuvant chemotherapy. *Journal of Magnetic Resonance Imaging*, 40(2):476–482, 2014. 16

Yuanfeng Ji, Haotian Bai, Chongjian Ge, Jie Yang, Ye Zhu, Ruimao Zhang, Zhen Li, Lingyan Zhanng, Wanling Ma, Xiang Wan, et al. Amos: A large-scale abdominal multi-organ benchmark for versatile medical image segmentation. *Advances in neural information processing systems*, 35: 36722–36732, 2022. 1

S Kirk, Y Lee, P Kumar, J Filippini, B Albertina, M Watson, K Rieger-Christ, and J Lemmerman. The cancer genome atlas lung squamous cell carcinoma collection (tcga-lusc)(version 4)[data set]. *The Cancer Imaging Archive*, 2016a. 16

S Kirk, Y Lee, CA Sadow, S Levine, C Roche, E Bonaccio, and J Filiippini. The cancer genome atlas colon adenocarcinoma collection (tcga-coad)(version 3)[data set]. *The Cancer Imaging Archive. https://doi. org/10.7937 K*, 9, 2016b. 16

Shanah Kirk, Yueh Lee, Fabiano R Lucchesi, Natalia D Aredes, Nicholas Gruszauskas, James Catto, Kimberly Garcia, Rose Jarosz, Vinay Duddalwar, Bino Varghese, et al. The cancer genome atlas urothelial bladder carcinoma collection (tcga-blca). In *The Cancer Imaging Archive*. 2016c. 16

Stefan Klein, Margreet Staring, Koen Murphy, and et al. elastix: a toolbox for intensity-based medical image registration. *IEEE Transactions on Medical Imaging*, 29(1):196–205, 2010. doi: 10.1109/TMI.2009.2035616. 3

Pamela J LaMontagne, Tammie L. S. Benzinger, John C Morris, Sarah Keefe, Russ Hornbeck, Chengjie Xiong, Elizabeth Grant, Jason Hassenstab, Krista Moulder, Andrei G. Vlassenko, et al. Oasis-3: Longitudinal neuroimaging, clinical, and cognitive dataset for normal aging and alzheimer disease. *medRxiv*, 2019. https://doi.org/10.1101/2019.12.13.19014902. 4

Hongwei Bran Li, Gian Marco Conte, Qingqiao Hu, Syed Muhammad Anwar, Florian Kofler, Ivan Ezhov, Koen van Leemput, Marie Piraud, Maria Diaz, Byrone Cole, et al. The brain tumor segmentation (brats) challenge 2023: Brain mr image synthesis for tumor segmentation (brasyn). *ArXiv*, pp. arXiv–2305, 2024. https://pmc.ncbi.nlm.nih.gov/articles/PMC10441440/. 4

Ka-Loh Li, Savannah C Partridge, Bonnie N Joe, Jessica E Gibbs, Ying Lu, Laura J Esserman, and Nola M Hylton. Invasive breast cancer: predicting disease recurrence by using high-spatial-resolution signal enhancement ratio imaging. *Radiology*, 248(1):79–87, 2008. 16

Peng Li, Shaoke Wang, Tianyu Li, Jie Lu, Yufan HuangFu, and Dong Wei Wang. A large-scale ct and pet/ct dataset for lung cancer diagnosis (lung-pet-ct-dx) [data set], 2020. The Cancer Imaging Archive. https://doi.org/10.7937/TCIA.2020.NNC2-0461. 16

M Linehan, R Gautam, C Sadow, and SJ Levine. The cancer genome atlas kidney chromophobe collection (tcga-kich)(version 3). *The Cancer Imaging Archive*, 2016a. 16

Marston Linehan, R Gautam, S Kirk, Y Lee, C Roche, E Bonaccio, J Filippini, K Rieger-Christ, J Lemmerman, and R Jarosz. The cancer genome atlas cervical kidney renal papillary cell carcinoma collection (tcga-kirp), version 4. *The Cancer Imaging Archive*, 2016b. 16

Wilma Lingle, Bradley J Erickson, Margarita L Zuley, Rose Jarosz, Ermelinda Bonaccio, Joe Filippini, Jose M Net, Len Levi, Elizabeth A Morris, Gloria G Figler, et al. The cancer genome atlas breast invasive carcinoma collection (tcga-brca). *(No Title)*, 2016. 16

Yaron Lipman, Ricky TQ Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations*, 2023. 1, 5, 7, 8

Xingchao Liu, Chengyue Gong, et al. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *The Eleventh International Conference on Learning Representations*, 2022. 5

FR Lucchesi and ND Aredes. The cancer genome atlas stomach adenocarcinoma collection (tcga-stad). *The Cancer Imaging Archive*, 2016. 15

Pranathi Madhavi, Shweta Patel, and Anne S. Tsao. Data from anti-pd-1 immunotherapy lung [data set], 2019. The Cancer Imaging Archive. https://doi.org/10.7937/tcia.2019.zjjwb9ip. 16

Ahmed W Moawad, Ayahallah A Ahmed, Mohab ElMohr, Mohamed Eltaher, Mouhammed Amir Habra, Sarah Fisher, Nancy Perrier, Miao Zhang, David Fuentes, and Khaled Elsayes. Voxel-level segmentation of pathologically-proven adrenocortical carcinoma with ki-67 expression (adrenal-acc-ki67-seg)[data set]. *The Cancer Imaging Archive*, 8, 2023. 15

AW Moawad, D Fuentes, A Morshid, AM Khalaf, MM Elmohr, A Abusaif, JD Hazle, AO Kaseb, M Hassan, A Mahvash, et al. Multimodality annotated hcc cases with and without advanced imaging segmentation. *The Cancer Imaging Archive (TCIA)*, 2021. 16

National Cancer Institute Clinical Proteomic Tumor Analysis Consortium (CPTAC). The clinical proteomic tumor analysis consortium pancreatic ductal adenocarcinoma collection (cptac-pda) (version 15) [data set], 2018. The Cancer Imaging Archive. https://doi.org/10.7937/K9/TCIA.2018.SC20FO18. 16

National Cancer Institute Clinical Proteomic Tumor Analysis Consortium (CPTAC). The clinical proteomic tumor analysis consortium uterine corpus endometrial carcinoma collection (cptac-ucec) (version 10) [data set], 2019. The Cancer Imaging Archive. 15

David Newitt, Nola Hylton, et al. Multi-center breast dce-mri data and segmentations from patients in the i-spy 1/acrin 6657 trials. *Cancer Imaging Arch*, 10(7):2016, 2016. 16

Guy Ohayon, Tomer Michaeli, and Michael Elad. Posterior-mean rectified flow: Towards minimum mse photo-realistic image restoration. *arXiv preprint arXiv:2410.00418*, 2024. 7, 8

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, pp. 4195–4205, 2023. 1

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022. 1

Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 conference proceedings*, pp. 1–10, 2022. 1, 7, 8

D Summers. Harvard whole brain atlas: www. med. harvard. edu/aanlib/home. html. *Journal of Neurology, Neurosurgery & Psychiatry*, 74(3):288–288, 2003. https://jnnp.bmj.com/content/74/3/288. 4

Adrian Thummerer, Erik Van der Bijl, Arthur Galapon Jr, Joost JC Verhoeff, Johannes A Langendijk, Stefan Both, Cornelis (Nico) AT van den Berg, and Matteo Maspero. Synthrad2023 grand challenge dataset: Generating synthetic ct for radiotherapy. *Medical physics*, 50(7):4664–4674, 2023. https://ieeexplore.ieee.org/document/8354170. 4

Vanya V Valindria, Nick Pawlowski, Martin Rajchl, Ioannis Lavdas, Eric O Aboagye, Andrea G Rockall, Daniel Rueckert, and Ben Glocker. Multi-modal learning from unpaired images: Application to multi-organ segmentation in ct and mri. In *2018 IEEE winter conference on applications of computer vision (WACV)*, pp. 547–556. IEEE, 2018. . 4

Jia Wu, Chao Li, Michael Gensheimer, Sukhmani Padda, Fumi Kato, Hiroki Shirato, Yiran Wei, Carola-Bibiane Schönlieb, Stephen John Price, David Jaffray, et al. Radiological tumour classification across imaging modality and histology. *Nature machine intelligence*, 3(9):787–798, 2021. 16

Mike Wu and Noah Goodman. Multimodal generative models for scalable weakly-supervised learning. In *Advances in Neural Information Processing Systems*, volume 31, 2018. https://papers.neurips.cc/paper/7801-multimodal-generative-models-for-scalable-weakly-supervised-learning. 6

Guangkai Xu, Yongtao Ge, Mingyu Liu, Chengxiang Fan, Kangyang Xie, Zhiyue Zhao, Hao Chen, and Chunhua Shen. Diffusion models trained with large data are transferable visual models. *arXiv e-prints*, pp. arXiv–2403, 2024. 1

Tao Zhou, Huazhu Fu, Geng Chen, Jianbing Shen, and Ling Shao. Hi-net: hybrid-fusion network for multi-modal mr image synthesis. *IEEE transactions on medical imaging*, 39(9):2772–2781, 2020. 7, 8

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pp. 2223–2232, 2017. 1, 7, 8

Xiahai Zhuang. Multivariate mixture model for myocardial segmentation combining multi-source images. *IEEE transactions on pattern analysis and machine intelligence*, 41(12):2933–2946, 2018. https://ieeexplore.ieee.org/abstract/document/8458220. 4

Appendix

A Additional Experiments

Reader Study.

Image-quality metrics correlate imperfectly with diagnostic utility, so we complement them with a reader study evaluating clinical plausibility. We recruited 8 experienced readers with backgrounds in radiology and medical imaging research to evaluate the synthesized scans across three complementary tasks. In the *authenticity classification* task, readers are shown a single scan (real or synthesized, presented in randomized order) and asked to judge whether it is real; an ideal synthesis method drives accuracy toward chance (50%). In the *quality scoring* task, readers rate each synthesized scan on a 5-point Likert scale for overall clinical realism. In the *paired preference* task, readers are shown two synthesized scans from different methods side by side and select which is more realistic; we report the win rate of each method against all others. Each reader evaluates 34 cases sampled uniformly across organs, with case order randomized and method labels blinded. Table 7 summarizes the results. FlowMI achieves the lowest authenticity-classification accuracy (56.3%, closest to the 50% ideal), the highest mean Likert score (3.99), and the highest paired-preference win rate (72.5%), indicating that radiologists find its outputs the hardest to distinguish from real scans and the most clinically plausible.

Table 7: Reader study results across three tasks. **Authenticity**: classification accuracy on real-vs-synthesized judgments (closer to 50% is better; red indicates deviation from the 50% ideal). **Likert**: mean rating on a 1–5 scale for clinical realism (higher is better). **Preference**: pairwise win rate against all other methods (higher is better). Values are mean ± std across readers. Best results in **bold**.

	Authenticity	Likert	Preference
Method	**Acc. (%) → 50**	**Score ↑**	**Win Rate (%) ↑**
CycleGAN	$68.4_{\pm 14.7}$ (+18.4)	$3.58_{\pm 0.63}$	$69.1_{\pm 33.2}$
I2IMamba	$61.2_{\pm 15.3}$ (+11.2)	$3.05_{\pm 0.69}$	$58.7_{\pm 34.7}$
ResViT	$64.6_{\pm 20.1}$ (+14.6)	$3.34_{\pm 0.74}$	$65.3_{\pm 23.8}$
FlowMI (Ours)	$\mathbf{56.3}_{\pm 15.6}$ (+6.3)	$\mathbf{3.99}_{\pm 0.79}$	$\mathbf{72.5}_{\pm 20.5}$
Real (GT)	—	$4.60_{\pm 0.34}$	—

Ablation Study. We ablate three components of FlowMI: (i) the trajectory reconstruction loss $\mathcal{L}_{\text{Trajectory}}$ (Eq. 5), which anchors intermediate latents to the image-space interpolant; (ii) mask conditioning, which signals the observation pattern to the velocity field; and (iii) Product-of-Experts (PoE) fusion, replaced with mean-pooling over modality-specific latents. As shown in Table 8, removing $\mathcal{L}_{\text{Trajectory}}$ causes the largest drop (-1.49 avg. PSNR), confirming that trajectory anchoring is critical across both CT and DCE settings. Mask conditioning and PoE have negligible effect on single-modality CT→CTC (within ± 0.05 dB), as expected since both reduce to near-identity operations with one input. Their value emerges on the multi-modality DCE tasks, where mask conditioning lets the velocity field distinguish $1 \to n$ from $n \to 1$ configurations and PoE's precision-weighted fusion outperforms uniform averaging, costing -0.88 and -0.33 avg. PSNR respectively when removed.

B Additional Dataset Details

Here we provide extended statistics for Contrast-X as well as the original data sources of each collection.

All scans were obtained by downloading raw DICOM or NIfTI files from publicly available repositories. The included collections are: Adrenal-ACC-Ki67-Seg Moawad et al. (2023) TCGA-OV Holback et al. (2016), TCGA-UCEC Erickson et al. (2016a), CPTAC-UCEC National Cancer Institute Clinical Proteomic Tumor Analysis Consortium (CPTAC) (2019), TCGA-STAD Lucchesi

Table 8: Ablation of FlowMI components. PSNR (dB) and SSIM (%) reported as Mean±Std on the four tasks. Average columns report the unweighted mean across the four tasks.

Variant	CT→CTC		DCE_1 →DCE_2		DCE_1 →$DCE_{2,3}$		$DCE_{1,3}$ →DCE_2		Average	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
w/o $\mathcal{L}_{\text{Trajectory}}$	22.81 ±4.32	75.2 ±9.74	24.97 ±3.41	70.4 ±10.2	25.18 ±3.38	70.1 ±8.96	27.89 ±3.52	73.4 ±8.41	25.21	72.3
w/o mask conditioning	24.49 ±4.18	78.4 ±8.83	25.73 ±3.27	71.8 ±9.65	25.42 ±3.24	70.6 ±8.72	27.64 ±3.41	72.1 ±8.27	25.82	73.2
PoE → mean-pooling	24.46 ±4.21	78.1 ±9.01	26.08 ±3.19	73.1 ±9.48	26.21 ±3.16	72.5 ±8.34	28.74 ±3.31	75.1 ±7.92	26.37	74.7
FlowMI (full)	**24.47** ±4.15	**78.5** ±8.62	**26.52** ±3.13	**74.2** ±9.33	**26.63** ±3.11	**73.7** ±8.06	**29.17** ±3.24	**76.2** ±7.65	**26.70**	**75.7**

& Aredes (2016), CPTAC-PDA National Cancer Institute Clinical Proteomic Tumor Analysis Consortium (CPTAC) (2018), HCC-TACE-Seg Moawad et al. (2021), TCGA-LIHC Erickson et al. (2016b), CMB-CRC Biobank (2022b), TCGA-COAD Kirk et al. (2016b), TCGA-BLCA Kirk et al. (2016c), TCGA-KIRC Akin et al. (2016), C4KC-KiTS Heller et al. (2019), CPTAC-CCRCC (CPTAC), TCGA-KIRP Linehan et al. (2016b), TCGA-KICH Linehan et al. (2016a), CMB-LCA Biobank (2022a), CPTAC-LSCC Consortium et al. (2018a), CPTAC-LUAD Consortium et al. (2018b), Lung-PET-CT-Dx Li et al. (2020), TCGA-LUSC Kirk et al. (2016a), Anti-PD-1_Lung Madhavi et al. (2019), TCGA-BRCA Lingle et al. (2016)Wu et al. (2021), UCSF Li et al. (2008)Jafri et al. (2014)Wu et al. (2021), I-SPY 1 Newitt et al. (2016)Wu et al. (2021).

C Registration Details

Registration was implemented via the 3D Slicer Elastix extension using ITK-Elastix backend. The rigid and affine stages used Advanced Mattes Mutual Information (32 bins), a 4-level resolution pyramid (shrink factors [8,4,2,1]), and adaptive stochastic gradient descent optimization. The deformable stage employed a third-order B-spline transform (FinalGridSpacingInVoxels = 10) with 3 resolution levels and up to 2000 iterations per level. All registrations were visually inspected for anatomical consistency before inclusion.

We performed rigid and affine alignment between CE and NCE scans using the **3D Slicer** Elastix module, followed by deformable registration for motion-prone organs (e.g., liver, lung). All stages used the ITK-elastix backend with the following representative settings: **rigid** (Euler3DTransform) and **affine** (AffineTransform) stages employed the Advanced Mattes Mutual Information (32 bins) as the similarity metric, a multi-resolution pyramid with 4 levels (shrink factors $[8, 4, 2, 1]$; smoothing sigmas $[3, 2, 1, 0]$ vox), the Adaptive Stochastic Gradient Descent optimizer (max 1024 iterations per level), and linear interpolation for resampling; the **deformable** stage used a third-order B-spline transform (BSpline order $= 3$) with 3 resolution levels (shrink factors $[4, 2, 1]$; smoothing sigmas $[2, 1, 0]$ vox), max 2000 iterations per level, and a final control-point spacing of 10 voxels (FinalGridSpacingInVoxels$= 10$), while final resampling used B-spline interpolation. This configuration achieved accurate anatomical correspondence while maintaining computational efficiency, reducing potential supervision bias due to CE/NCE misalignment.

D Ethics and Data Availability

We verified the licenses and acquisition procedures of all resources to ensure redistribution compliance. Dataset access is provided upon submission of an agreement form and approval from the original data sources. It will be released on our project website under a CC BY-NC 4.0 license, prohibiting private redistribution. All users must obtain permission through the website before downloading.

AI use statement

Generative AI tools were used to aid and polish the writing of this paper.