See More, Change Less: Anatomy-Aware Diffusion for Contrast Enhancement

Junqi Liu[1,2] Zejun Wu[1,3] Pedro R. A. S. Bassi[1,4,5] Xinze Zhou[1] Wenxuan Li[1]
Ibrahim E. Hamamci[6,7] Sezgin Er[8] Tianyu Lin[1] Yi Luo[1] Szymon Płotka[9]
Bjoern Menze[6,7] Daguang Xu[10] Kai Ding[11] Kang Wang[12] Yang Yang[12]
Yucheng Tang[10] Alan L. Yuille[1] Zongwei Zhou[1,*]

[1]Johns Hopkins University [2]University of Copenhagen [3]University of Virginia
[4]University of Bologna [5]Italian Institute of Technology [6]University of Zurich
[7]ETH AI Center [8]Istanbul Medipol University [9]Jagiellonian University
[10]NVIDIA [11]Johns Hopkins Medicine [12]University of California, San Francisco

Code, Dataset, and Models: https://github.com/MrGiovanni/SMILE

Abstract

*Image enhancement improves visual quality and helps reveal details that are hard to see in the original image. In medical imaging, it can support clinical decision-making, but current models often over-edit. This can distort organs, create false findings, and miss small tumors because these models do not understand anatomy or contrast dynamics. We propose SMILE, an anatomy-aware diffusion model that learns how organs are shaped and how they take up contrast. It enhances only clinically relevant regions while leaving all other areas unchanged. SMILE introduces three key ideas: (1) structure-aware supervision that follows true organ boundaries and contrast patterns; (2) registration-free learning that works directly with unaligned multi-phase CT scans; (3) unified inference that provides fast and consistent enhancement across all contrast phases. Across six external datasets, SMILE outperforms existing methods in image quality (**14.2%** higher SSIM, **20.6%** higher PSNR, **50%** better FID) and in clinical usefulness by producing anatomically accurate and diagnostically meaningful images. SMILE also improves cancer detection from non-contrast CT, raising the F1 score by up to 10 percent.*

1. Introduction

Image enhancement is an important task in computer vision that aims to improve image quality or emphasize key details. Recent advances in image enhancement are fueled by generative AI, especially diffusion models [12–14, 22, 30, 50, 71], which can make enhancements through conditional generation. However, these models often become too creative, adding unwanted changes to the original image that reduce its usefulness. This happens because

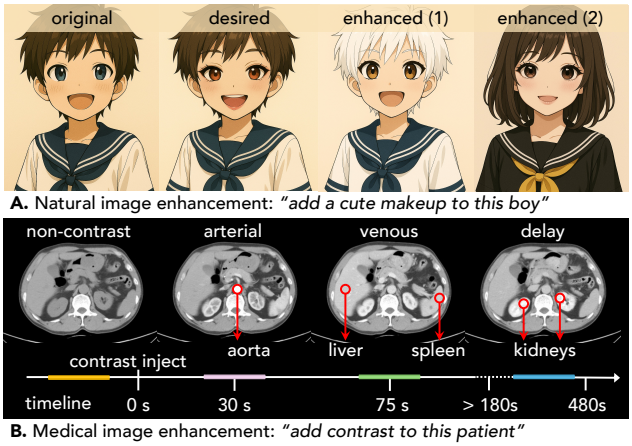

Figure 1. **A. Natural image enhancement.** When generative models add makeup to a photo, they often do too much—changing not just the face, but the hair and clothes as well. That's fine for social media filters, but in medical imaging, such 'over-creativity' can hide real tumors or create fake ones. **B. Medical image enhancement.** Doctors use contrast agents (a liquid injected into the body) to make internal organs easier to see. After injection, CT scans are taken at four times: (1) Non-contrast (N) is the baseline before enhancement; (2) Arterial (A) highlights arteries; (3) Venous (V) enhances organs such as the liver and spleen; and (4) Delay (D) shows mainly urinary system. The contrast makes certain tissues absorb more X-rays and appear brighter, revealing subtle tumors or vascular structures that would otherwise be invisible. These phases reflect how contrast flows through organs, helping radiologists detect and diagnose disease more accurately.

*Correspondence to Zongwei Zhou (ZZHOU82@JH.EDU)

they lack clear knowledge of what objects are in the image and which parts should be changed. For example, think of how an AI model adds makeup to a face. Ideally, it should touch only what's needed—the lips, skin tone, or eyes—while keeping everything else unchanged. Yet most generative models can't resist being 'creative': they also modify the hair, the clothes, or even the background [63], as shown in Figure 1A. These changes maybe okay in beauty filters, but dangerous in medicine. In CT scans, for example, a model that tries to enhance contrast[1] might unintentionally alter healthy organs or fabricate false tumors. Our goal is to build a model that knows exactly *what to change* and *what to keep*, ensuring every enhancement looks realistic—and remains clinically trustworthy.

We hypothesize that if generative models can recognize objects and their fine structures (e.g., through semantic segmentation), they can focus edits on the right regions while keeping everything else unchanged. This would enable image enhancement to make necessary, but minimal, changes. Clinically, this problem is highly relevant because approximately half of abdominal CT scans are acquired without contrast [6, 11, 31]. If a reliable AI model could create realistic contrast-enhanced versions of these scans, it would make tumor detection easier and more accessible. Radiologists could identify tumors earlier without additional scans or injections. This highlights the need for enhancement models that are anatomically and physiologically accurate.

To study this, we first build a large-scale medical image dataset that offers two key advantages over existing datasets. **First,** our dataset includes per-pixel annotations of many organs, vessels, bones, and disease regions, enabling structural evaluation. In contrast, other datasets lack such detailed annotations of fine structures (e.g., hair, face, clothing), making them less suitable for assessing generative models. **Second,** our dataset provides paired scans for more than 477 patients from over 112 hospitals, each with non-contrast, arterial, venous, and delay phases (see Figure 1B). Other datasets, by comparison, lack paired "before-and-after" images. For example, in natural image datasets, there are no exact pairs of faces with and without makeup. We expect the insights from this study to generalize to natural image enhancement once comparable datasets become available for training and evaluation.

We introduce SMILE (Super Modality Image Learning and Enhancement), an anatomy-aware diffusion model designed for clinically reliable CT enhancement. Here, *anatomy-aware* has two meanings. **First,** the enhanced image should still represent the same patient as the original. Many existing models change organ shapes or spatial structures so much that the result no longer looks like the same patient [9, 10, 55] (see Figure 4A). **Second,** anatomy-aware means understanding basic physiology, i.e., how different organs, vessels, and tissues take up contrast over time. Tumors should neither disappear nor be artificially created. Current generative models completely lack this knowledge, often producing unrealistic contrast intensity that ignore the actual dynamics of the human body [26, 32–34, 43, 44, 51, 69] (see Figure 4B).

To our knowledge, this is the *first* study to quantitatively evaluate AI-generated CT enhancements for both **structural consistency** and **intensity accuracy**. Structural consistency ensures the enhanced image still represents the same patient, while intensity accuracy verifies that tissue density values remain clinically correct. Such evaluation was impossible before, as it requires precise organ and vessel masks—now enabled by our dataset and method. In summary, we make two contributions that mark a significant step toward clinically useful contrast image enhancement:

1. **An open, multi-phase dataset.** We present CTVerse, a large CT dataset that includes four contrast phases (non-contrast, arterial, venous, and delay) for all 477 patients from 112 hospitals. Each scan is annotated with 88 anatomical structures and tumors in the pancreas, liver, and kidney, resulting in 159,632 three-dimensional masks (detailed statistics are in Figure 3). This dataset enables open-source training and quantitative evaluation of cross-phase enhancement methods.
2. **An anatomy-aware method.** We develop SMILE, an anatomy-aware diffusion framework trained with multiple types of supervision to ensure both visual quality and clinical reliability (illustrated in Figure 2). The model introduces new loss functions that guide learning in three ways—preserving structure, maintaining clinical meaning, and ensuring realistic intensity. We conduct rigorously assessment not only in image quality but also in clinical relevance. Across six external datasets, SMILE achieves significant improvement in both image quality and clinical relevance: $+14\%$ SSIM, $+20\%$ PSNR, $+50\%$ FID, and $r > 0.95$ in intensity correlation.

We further demonstrate SMILE for opportunistic cancer screening using non-contrast CT scans—the most common type of abdominal imaging. These scans are widely available but often difficult for radiologists and AI systems to interpret because they lack contrast. Since many of these scans are taken for unrelated reasons, they represent a large and underused resource for early cancer detection. SMILE transforms non-contrast scans into realistic contrast versions (arterial, venous, or delay), revealing tumors that would otherwise remain invisible. When combined with existing detection models, the enhanced scans improve F1 score by 10%, turning non-contrast CT scans into a practical and accessible tool for early cancer screening without additional cost or contrast exposure.

[1] In medical imaging, contrast refers to a radiopaque substance injected into the bloodstream that temporarily increases tissue brightness, allowing radiologists to see blood vessels, organs, and potential tumors more clearly.

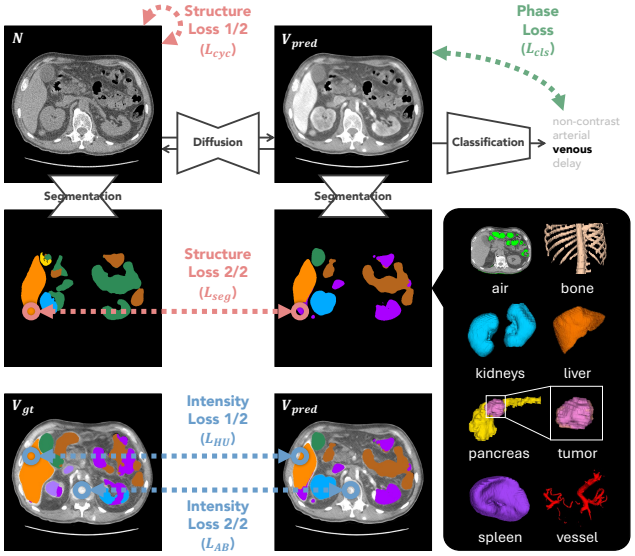

Figure 2. **SMILE employs anatomy-aware supervision.** Our framework integrates multiple anatomical constraints to guide contrast-phase enhancement. The structural segmentation loss ($\mathcal{L}_{\text{seg}}$) and cycle consistency loss ($\mathcal{L}_{\text{cyc}}$) preserve structural fidelity. The phase classification loss ($\mathcal{L}_{\text{cls}}$) ensures the enhanced CT shows the correct contrast-phase characteristics. Finally, the intensity HU loss $\mathcal{L}_{\text{HU}}$ and air/bone loss $\mathcal{L}_{\text{AB}}$ enforce realistic organ (mainly abdominal, e.g., liver) enhancement and maintain consistency in air and bone regions that should remain unchanged. As demonstrated in the figure, SMILE does not require registration for enhancement source and ground truth.

2. SMILE

SMILE is a registration-free diffusion framework for multi-phase CT enhancement built on Stable Diffusion [61]. It learns directly from unaligned data and uses three complementary supervision signals—structure, phase, and intensity—to guide the diffusion process. As shown in Figure 2, these anatomy-aware signals work together to preserve anatomy, enforce correct phase appearance, and maintain realistic tissue contrast. Overall, SMILE minimizes

$$\mathcal{L}_{\text{SMILE}} = \mathcal{L}_{\text{diff}} + \lambda_{\text{stu}}\mathcal{L}_{\text{stu}} + \lambda_{\text{cls}}\mathcal{L}_{\text{cls}} + \lambda_{\text{int}}\mathcal{L}_{\text{int}}, \tag{1}$$

where each term controls a specific form of structural or clinical supervision, enabling anatomically faithful and clinically reliable enhancement.

2.1. Structural Loss

Organ shapes vary noticeably across CT phases because the scans are not perfectly aligned. To handle this mismatch, SMILE adds structural supervision that does not require voxel-wise registration. The structural loss has two components: $\mathcal{L}_{\text{stu}} = \mathcal{L}_{\text{seg}} + \mathcal{L}_{\text{cyc}}$.

1. Segmentation Loss: Because the shapes of anatomical structures such as organs, vessels, bones, etc., remain consistent across CT phases, SMILE adds a segmentation loss to preserve anatomy. Both the enhanced CT $\hat{\mathbf{x}}_{\text{tgt}}$ and the input $\mathbf{x}_{\text{src}}$ are segmented using a pretrained nnU-Net [37] on large, publicly available datasets [5, 46, 47, 49, 59]. The loss $\mathcal{L}_{\text{seg}} = -\frac{1}{N}\sum G_i \log S_i(\hat{\mathbf{x}}_{\text{tgt}})$ penalizes structural differences between the enhanced and source masks.

2. Cycle Consistency Loss: Although CT phases differ in contrast appearance, their underlying anatomy remains the same. To reflect this property under unregistered training, SMILE applies a cycle reconstruction step in which the source phase $\mathbf{x}_{\text{src}}$ is recovered from the generated target phase $\hat{\mathbf{x}}_{\text{tgt}}$. Let $\boldsymbol{\epsilon}$ denote the sampled noise and $\boldsymbol{\epsilon}_\theta$ the corresponding prediction. The cycle loss is defined as

$$\mathcal{L}_{\text{cyc}} = \|\hat{\mathbf{x}}_{\text{cyc}} - \mathbf{x}_{\text{src}}\|_1 + \lambda_{\text{cdiff}}\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta\|_1, \tag{2}$$

where the second term stabilizes diffusion-based reconstruction by enforcing noise-prediction consistency, and is computed only within the cycle path to regularize cycle diffusion process, separate from the main diffusion loss.

2.2. Phase Loss

Different phases differ mostly in local contrast changes (e.g., arterial enhancement in vessels or liver), while most regions of the CT are unchanged across phases. As a result, patch-based discriminators—which judge realism on small image crops—often receive little or ambiguous supervision, because most small crops look identical in all phases. To address this, we simply add a pretrained phase classifier $C(\cdot)$ to guide the diffusion model, ensuring the generated image $\hat{\mathbf{x}}_{\text{tgt}}$ receives the correct phase label.

2.3. Intensity Loss

Contrast enhancement is mainly driven by intensity changes, so SMILE develops new losses that model how organ intensities should change or stay constant across phases, and is defined as $\mathcal{L}_{int} = \mathcal{L}_{HU} + \mathcal{L}_{AB}$.

1. Organ HU Loss: Organ-level intensity patterns provide essential cues for distinguishing contrast phases. Organ Hounsfield Unit (HU) values [21, 72] follow predictable shifts across contrast phases. SMILE computes a normalized squared error between the enhanced and target mean HU for each organ:

$$\mathcal{L}_{\text{HU}} = \frac{1}{N_{\text{org}}}\sum_{i=1}^{N_{\text{org}}} \frac{(\bar{H}_i^{\text{pred}} - \bar{H}_i^{\text{gt}})^2}{(\bar{H}_i^{\text{gt}})^2}, \tag{3}$$

where $\bar{H}_i$ denotes the mean HU of the i^{th} organ. This loss guides predicted organ intensities toward phase-consistent attenuation patterns.

2. Air and Bone Loss: Air cavities and bones show minimal intensity variation across phases. A detector $D(\cdot)$ identifies these regions, and deviations between the enhanced and source images are penalized via:

$$\mathcal{L}_{\text{AB}} = \|D(\hat{\mathbf{x}}_{\text{tgt}}) - D(\mathbf{x}_{\text{src}})\|_2^2. \tag{4}$$

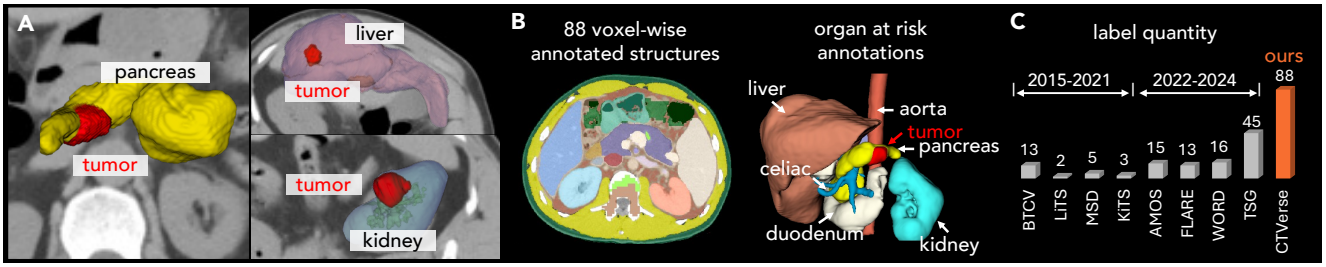

Figure 3. **CTVerse sets a new standard for multi-phase CT benchmarks by offering the most comprehensive, finely annotated organ and tumor labels.** We introduce CTVerse, a high-quality, multi-phase, and precisely annotated CT dataset designed for training and evaluating generative models. **A.** CTVerse provides detailed annotations for multiple tumor types (*pancreas*, *liver*, and *kidney*). **B.** CTVerse contains voxel-wise annotations for 88 anatomical structures, including organs at risk, vessels, bones, etc. **C.** Compared to existing publicly available datasets [5, 16, 17, 45–47, 49, 52, 53, 59], our CTVerse includes at least 1.5 times more labeled structures, making it a strong benchmark for generative models and medical research.

This term constrains intensity stability in these regions while leaving contrast-sensitive structures free to vary.

3. Experiments

3.1. Datasets and Evaluation

We test SMILE on a fully Out-of-Distribution (OOD) setup. The training and test data come from different sources, scanners, and patients. For evaluation, we use three common metrics: SSIM [65] to measure structural similarity, PSNR [41] to measure image quality, FID [29] to measure how realistic the generated images look. Moreover, we are the first to use HU correlation to measure how well the organ intensity values match the real ones. Details of datasets used can be found in Table1.

Training dataset: We trained SMILE using a subset of the PACN (private), RTCN (private) and FUS (private) dataset, which provides multi-phase abdominal CT volumes acquired from the same patients under different contrast conditions. The training set consists of 312 CT volumes from 85 patients, organized as intra-patient multi-phase pairs to ensure physiological consistency across modalities.

Testing dataset: Evaluation of SMILE is conducted on multiple independent external datasets, including WAW-TACE [3], CT-RATE [27], MSD-CT [1], PECN (private), JUS (private), and TUS (private). For SSIM/PSNR comparison, all volumes are voxel-wise registered using ANTs [2] solely to enable fair metric computation, since these metrics require phase-aligned images for valid comparison. The registration is not used during model training or inference, and does not affect SMILE's generation process.

A new CTVerse dataset (Figure 3): We will release a precisely annotated, phase-wise paired, and organ-wise registered high-quality dataset, CTVerse, containing 1814 CT volumes from 477 patients and voxel-wise annotations of 88 anatomical structures (details can be found in Table 4).

Table 1. **Dataset characteristics.** Our constructed CTVerse provides complete four-phase CT scans with full per-voxel annotations, whereas most public datasets contain only partial labels or limited phases. All public sources were checked for duplicates using 3D perceptual hashing.

dataset	patients	scans	phase	region
*our constructed CTVerse dataset, where **477** patients' scans will be released*				
BRA [40]	14	28	A V	BR
LiTS [7]	18	36	A V	DE
WAW-TACE [3]	163	652	N A V D	PL
MSD-CT [1]	15	60	N A V D	US
VinDr [20]	30	90	N A V	VN
PECN (private)	123	492	N A V D	CN
RTCN (private)	114	456	N A V D	CN
***85** patients used for AI training in this paper*				
PACN (private)	24	96	N A V D	CN
RTCN (private)	33	132	N A V D	CN
FUS (private)	28	84	N V D	US
***1,015** patients used for AI testing in this paper*				
WAW-TACE [3]	163	652	N A V D	PL
CT-RATE [27]	74	74	N	TR
MSD-CT [1]	15	60	N A V D	US
PECN (private)	123	492	N A V D	CN
JUS (private) [66]	55	220	N A V D	US
TUS (private)	585	585	N	US

BR: Brazil CN: China DE: Germany PL: Poland TR: Turkey
US: United States VN: Vietnam

The data were collected from over 112 hospitals, including BRA [40], LiTS [7], WAW-TACE [3], MSD-CT [1], VinDr [20], PECN (private), and RTCN (private) datasets. All volumes were registered at the voxel level using the ANTs [2] toolkit to ensure accurate spatial alignment across phases. The precise voxel-level registration enables accurate comparison across contrast phases, providing a reliable benchmark for evaluating enhancement fidelity and structural consistency in image enhancement.

3.2. Baseline and Implementation

We compare SMILE with six representative baselines with the three main stream generative model architectures: (1) Generative Adversarial Network (GAN): Pix2Pix [38], Cy-

Table 2. **SMILE improves enhancement quality across all 12 phase conversions, with +14.2% SSIM, +20.6% PSNR, and 50.5% better FID than the best baselines. In downstream tumor detection, SMILE further lifts F1-score by more than 10%, while all baseline enhancement methods fail to provide any meaningful improvement.** Quantitative comparison of multi-phase CT enhancement across 356 patients from WAW-TACE [3], MSD-CT [1], PECN (private), JUS (private) datasets. Each patient has four contrast phases: non-contrast (N), arterial (A), venous (V), and delay (D). For every source–target combination, representative scans were extracted from the CT volume after enhancement. We report SSIM, PSNR, and FID to assess structural similarity, perceptual fidelity, and distributional realism. The best and second-best results are marked with bold and underline. We also performed a one-sided Wilcoxon signed-rank test comparing our SMILE with all others, and statistically significant improvements at $P = 0.05$ are highlighted in pink.

	source	N			A			V			D			
method	target	A	V	D	N	V	D	N	A	D	N	A	V	average
Pix2Pix [38] *GAN-based*	SSIM (↑)	70.7	78.9	72.1	68.9	72.5	47.3	54.1	67.8	57.3	44.5	44.3	49.6	60.7
	PSNR (↑)	18.9	21.3	19.3	19.9	21.0	14.9	18.4	20.7	19.9	15.9	16.5	19.0	18.8
	FID (↓)	147.8	179.5	377.3	423.7	232.2	250.9	414.3	324.2	322.0	416.6	228.9	279.4	299.7
CycleGAN [18] *GAN-based*	SSIM (↑)	71.9	76.1	70.8	71.7	70.9	66.6	77.4	69.4	73.9	72.1	67.1	75.4	71.9
	PSNR (↑)	17.9	19.3	17.2	18.3	18.0	16.4	20.8	17.9	18.8	18.4	16.6	19.3	18.2
	FID (↓)	128.8	319.9	231.7	397.5	275.1	261.8	484.1	195.6	183.4	415.8	241.4	117.9	271.1
CyTran [60] *GAN-based*	SSIM (↑)	38.9	63.1	48.0	39.9	51.9	52.6	45.6	61.7	66.1	49.3	69.8	70.1	54.8
	PSNR (↑)	9.4	16.2	12.4	8.6	12.8	13.3	10.0	17.0	17.6	13.2	19.5	18.4	14.0
	FID (↓)	193.5	231.2	258.8	373.4	300.3	292.5	374.0	178.3	244.0	407.8	209.7	254.4	276.5
DALL-E [23] *VAE-based*	SSIM (↑)	51.9	47.6	54.6	51.6	50.9	50.7	47.2	49.0	60.7	45.7	47.5	59.8	51.4
	PSNR (↑)	17.4	16.0	16.1	16.3	16.3	16.8	14.9	15.1	18.8	14.4	15.0	17.9	16.3
	FID (↓)	471.1	482.4	454.5	240.8	511.1	468.6	233.6	505.0	466.5	243.6	511.7	495.8	423.7
MedDiffusion [39] *DIFF-based*	SSIM (↑)	63.2	72.2	72.0	69.1	62.4	62.7	55.3	60.3	67.2	64.3	70.4	55.7	64.6
	PSNR (↑)	16.7	18.0	15.6	17.0	16.2	16.0	14.1	15.3	17.1	13.7	15.4	16.9	16.0
	FID (↓)	490.0	298.9	299.4	468.5	516.1	503.1	310.9	506.5	205.8	298.6	482.4	495.3	406.3
CUT [57] *GAN-based*	SSIM (↑)	66.8	73.6	70.6	72.9	77.3	79.1	78.7	73.9	80.1	75.5	75.3	81.3	<u>75.4</u>
	PSNR (↑)	18.9	19.9	18.7	20.4	21.2	20.7	21.1	21.2	21.5	19.6	21.8	22.9	<u>21.4</u>
	FID (↓)	283.4	258.6	280.9	363.8	204.4	259.3	333.4	182.3	270.7	360.1	172.5	265.1	<u>269.5</u>
SMILE (ours)	SSIM (↑)	**85.1**	**89.5**	**82.9**	**84.8**	**86.9**	**85.7**	**86.0**	**83.5**	**86.1**	**89.1**	**85.6**	**88.3**	**86.1** (↑14.2%)
	PSNR (↑)	**25.5**	**27.5**	**22.0**	**24.4**	**26.2**	**25.0**	**26.4**	**25.5**	**26.0**	**28.1**	**25.8**	**27.3**	**25.8** (↑20.6%)
	FID (↓)	**117.3**	**152.8**	**156.7**	**148.6**	**127.9**	**129.2**	**138.1**	**134.3**	**129.6**	**112.6**	**129.4**	**124.8**	**133.4** (↓50.5%)

N: non-contrast **A**: arterial **V**: venous **D**: delay SSIM: structural similarity index measure PSNR: peak signal-to-noise ratio FID: Fréchet Inception Distance

cleGAN [18], CUT [57] and CyTran [60], (2) Variational Autoencoder (VAE): DALL-E pytorch [23], (3) Diffusion Models (DIFF): MedDiffusion [39]. These methods were carefully selected because they represent the most widely adopted paradigms in image enhancement.

Other potentially relevant methods were not included for three justified reasons: (1) *Limited reproducibility:* Many recent works report results on private medical datasets without releasing code or pretrained models, making fair comparison impossible. (2) *Dataset mismatch:* Some methods are trained on domain-specific data (e.g., brain MRI), which do not generalize to CT enhancement task. (3) *Dimensional constraint:* Most image-enhancement methods are designed for 2D images and cannot handle 3D CT data with volumetric consistency required.

SMILE is trained for 100k steps on a single NVIDIA H100 (80 GB). For implementation, SMILE uses five practical loss terms (`seg`, `cyc`, `cls`, `HU`, `AB`) although the method them into three fields (structure, phase, intensity). The distinction is explained in Appendix C. We progressively enable supervision: diffusion only (0–2k), add phase+cycle (10k), add seg+HU+AB (20k). Fixed weights are λ_{diff}=1, λ_{cls}=10^{-3}, λ_{cyc}=10, λ_{seg}=10^{-3}, λ_{HU}=10^{-2}, λ_{AB}=1, and become learnable after 80k via the Uncertainty Loss Module.

4. Results

4.1. SMILE Exceeds Other Enhancement Methods

We evaluate SMILE against six baselines on WAW-TACE [3], MSD-CT [1], PECN (private), and JUS (private) datasets. Each case contains four phases (non-contrast, arterial, venous, delay). To keep evaluation efficient while still reflecting whole-volume quality, we use uniformly sampled scans per source–target pair, which reliably represent overall CT enhancement without unnecessary computational cost. Table 2 reports SSIM, PSNR, and FID across all 12 enhancement directions. SMILE achieves the highest scores on all metrics and all phase conversions.

4.2. SMILE Maintains Anatomical Structures

We evaluate SMILE on 22 annotated organs[2] across all four contrast phases from WAW-TACE [3], MSD-CT [1], PECN (private), and JUS (private) datasets. Figure 5 shows the correlation between synthesized and real CT scans in organ-

[2]The full list is provided in Appendix Table 5.

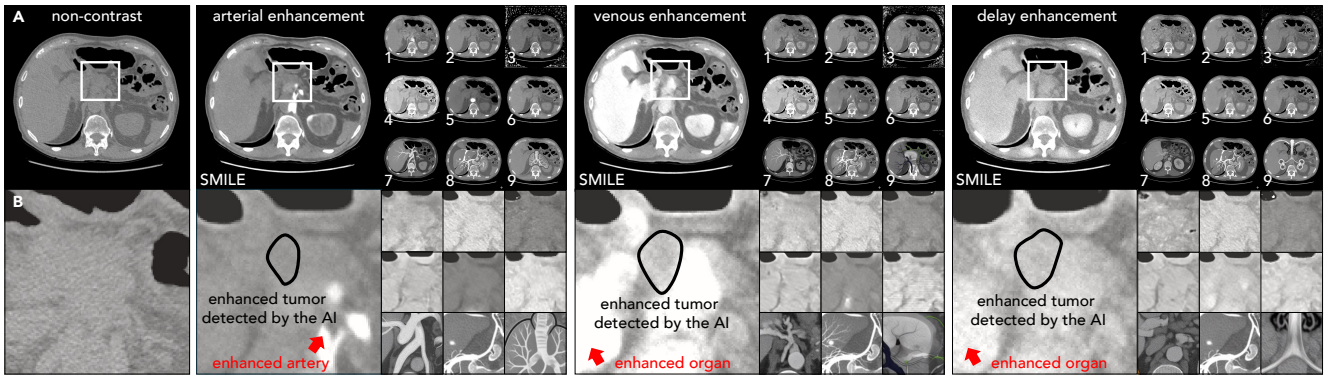

Figure 4. **SMILE ensures both structural consistency and intensity accuracy on enhanced scans, outperforming all competing models in both anatomical integrity and tumor visibility.** We evaluate whether the enhanced CT scans remain anatomically correct (no extra structures, no distortions, high image quality), and whether the added contrast is diagnostically correct by allowing the tumor to be clearly detected. Baselines used (labeled 1–9): (1) Pix2Pix [38], (2) CycleGAN [18], (3) CyTran [60], (4) DALL-E [23], (5) MedDiffusion [39], (6) CUT [57], (7) ChatGPT-5.1 [8], (8) Google Nano Banana [19], (9) Qwen-3 Max [68]. The last three large vision models are included only for visual reference to show that general image-editing systems cannot handle medical CT enhancement reliably. **A. Structural Consistency.** We evaluate SMILE's structural fidelity by comparing non-contrast CT enhancement results against 9 compelling baseline models. Compared to other methods, SMILE preserves organ boundaries and global anatomy without introducing extra structures or phase-inconsistent artifacts. This shows that SMILE maintains high structural reliability even under large intensity shifts. **B. Intensity Accuracy.** To further validate enhancement correctness, we apply a state-of-the-art tumor detector [48] on the enhanced arterial, venous, and delay phases. In all three enhanced phases, the tumor is successfully detected (marked by the black circle), confirming that SMILE restores clinically meaningful contrast cues needed for downstream diagnostic tasks.

level mean HU and volume. Mean HU is computed per organ, providing a stable and clinically meaningful metric. All scans are segmented using the state-of-the-art VISTA-3D [28] to obtain accurate organ measurements. SMILE closely matches real anatomical statistics, achieving average correlations >0.95 for both HU and size. This demonstrates that SMILE enhances image quality while preserving organ integrity and clinical realism across all phases.

4.3. SMILE Enhances Precise Contrasts

We assess phase accuracy using a radiologist reader study and a pretrained phase classifier (Figure 6). Both tests show that SMILE produces clear and correct phase-specific enhancements, with overall precision and recall above 95%.

Reader study. Three radiologists from two hospitals labeled a mixed set of SMILE-enhanced scans containing all four phases. Their confusion matrices show that each phase is easily and consistently identified.

Phase classification model. A pretrained DenseNet [35] classifier further evaluates phase correctness. For each source phase, we classify all SMILE-enhanced outputs, and the resulting confusion matrices show high precision and recall across all source–target phase pairs.

4.4. SMILE Detects Tumors in Non-Contrast Scans

We evaluate SMILE as a preprocessing module for two pancreatic tumor detectors: MedFormer [24] (winner of the Touchstone benchmark [4]) and ScaleMAI [48]. Both models represent the current state of the art in pancreatic tumor analysis, achieving strong detection and segmentation performance in recent benchmark studies. MedFormer reports 82–88% F1 on multi-institutional benchmarks, while ScaleMAI achieves over 90% sensitivity for early pancreatic lesions [24, 48]. We evaluate both detectors on external non-contrast CT scans from CT-RATE [27] and TUS (private), and report sensitivity, specificity, F1, as well as AUC

Table 3. **SMILE improves early cancer detection with 10–20% performance gains using non-contrast CT scans.** We evaluate SMILE as a preprocessing module for two state-of-the-art pancreatic tumor detectors, MedFormer [24] and ScaleMAI [48], on external non-contrast CT scans. Adding SMILE yields strong gains in sensitivity, f1-score, and AUC with only modest drops in specificity, demonstrating that SMILE consistently improves the performance of these SOTA detectors.

	CT-RATE ($N = 74$)				TUS ($N = 585$)			
method	Sen.	Spec.	F1.	AUC	Sen.	Spec.	F1.	AUC
Gao *et al.* [24]	53.9	95.1	60.9	0.75	78.0	98.7	81.2	0.88
+SMILE	69.2	91.8	71.4	0.84	90.0	98.7	88.2	0.94
Δ	+15.3	-3.3	+10.5	+0.09	+12.0	0.0	+7.0	+.06
Li *et al.* [48]	61.5	86.9	55.2	0.74	86.0	91.9	63.3	0.88
+SMILE	84.6	80.3	61.1	0.83	94.0	91.9	66.1	0.92
Δ	+23.1	-6.6	+5.9	+0.09	+8.0	0.0	+2.8	+.04

Sen.: sensitivity Spec.: specificity F1.: f1-score
AUC: area under the receiver operating characteristic curve

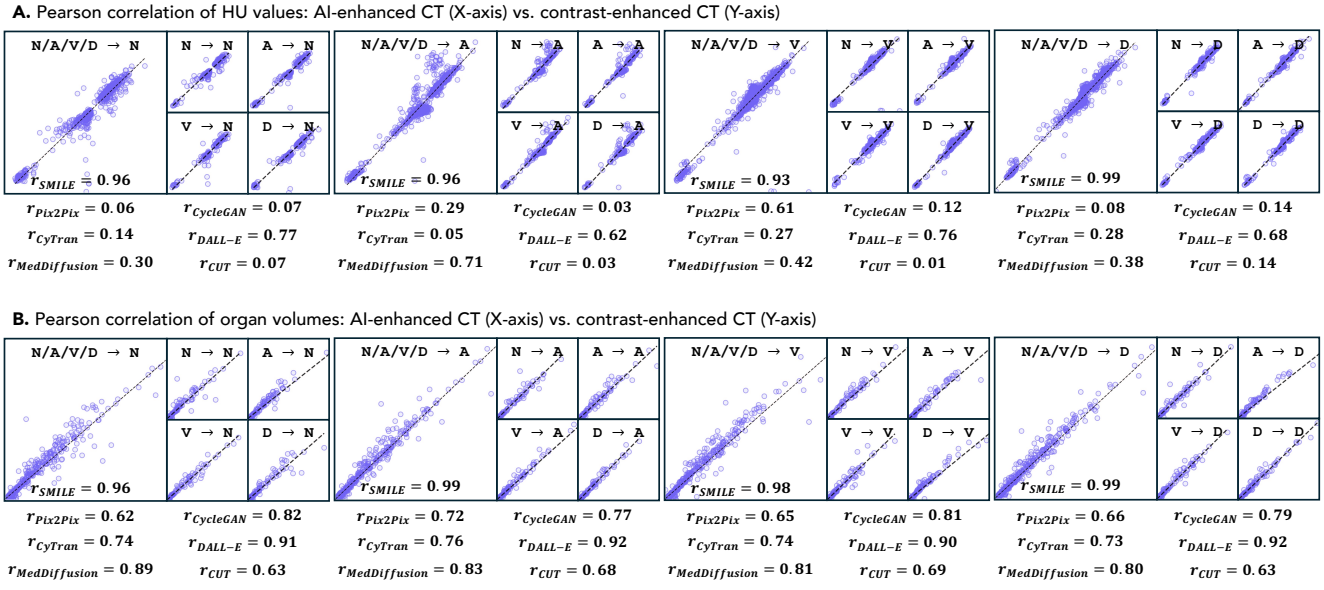

Figure 5. **Across all phases and organs, SMILE achieves an average HU and size correlation above 0.95, demonstrating strong anatomical and intensity consistency with real CT scans.** We evaluate how well SMILE preserves organ intensity (HU) and volume size compared with the real CT scans across 22 organs. Here the HU is averaged over the whole organ rather than per-pixel. Each large plot corresponds to a target phase, N: non-contrast, A: arterial, V: venous, D: delay, and the four small plots on the right show results as enhanced from other source phases. The smaller plots illustrate all source→target phase pairs, showing that SMILE delivers consistently strong organ HU and size alignment regardless of the enhancement direction. Overall, SMILE maintains high consistency in both HU and organ size, demonstrating stable and reliable enhancement performance.

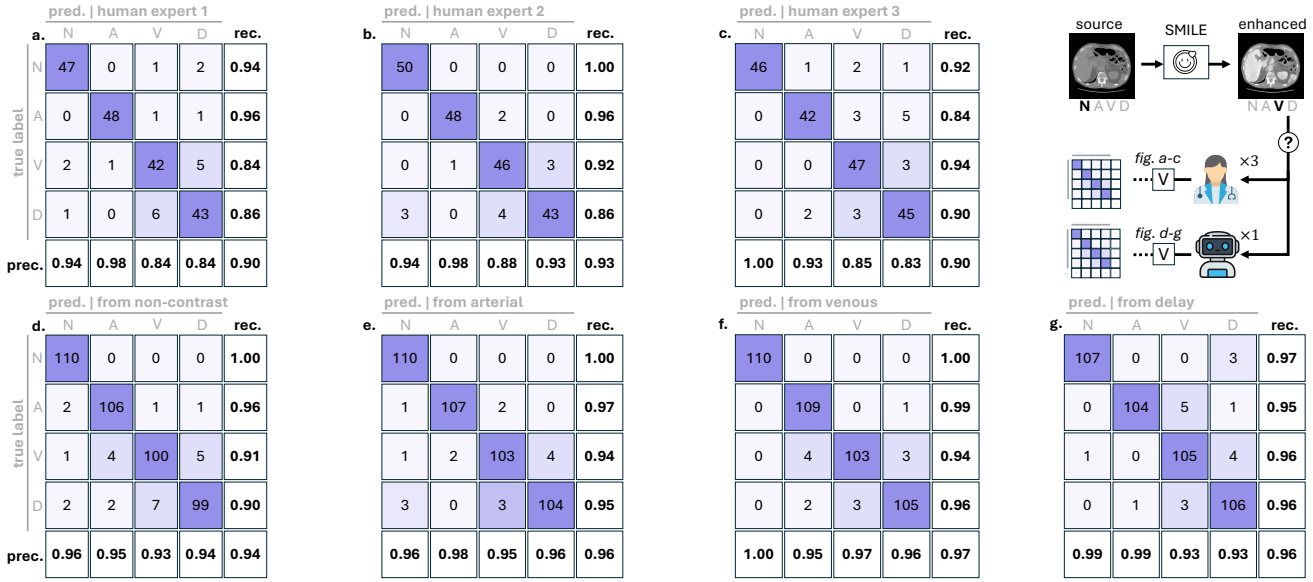

Figure 6. **SMILE consistently produces accurate and distinguishable phase-specific enhancements, achieving average over 95.0 precision and recall score. (a–c)** Reader study: three radiologists from two hospitals were shown a mixed set of SMILE-enhanced CT scans containing all four phases (N: non-contrast, A: arterial, V: venous, D: delayed) shuffled together. Each radiologist independently labeled the phase of every slice, and we plot confusion matrices from their annotations. **(d–g)** Classification model: We also evaluated SMILE phase enhancement correctness via a pretrained phase classifier. By feeding SMILE-enhanced scans enhanced from each source phase (N from N/A/V/D, A from N/A/V/D, V from N/A/V/D, D from N/A/V/D), and plot the confusion metrics respectively.

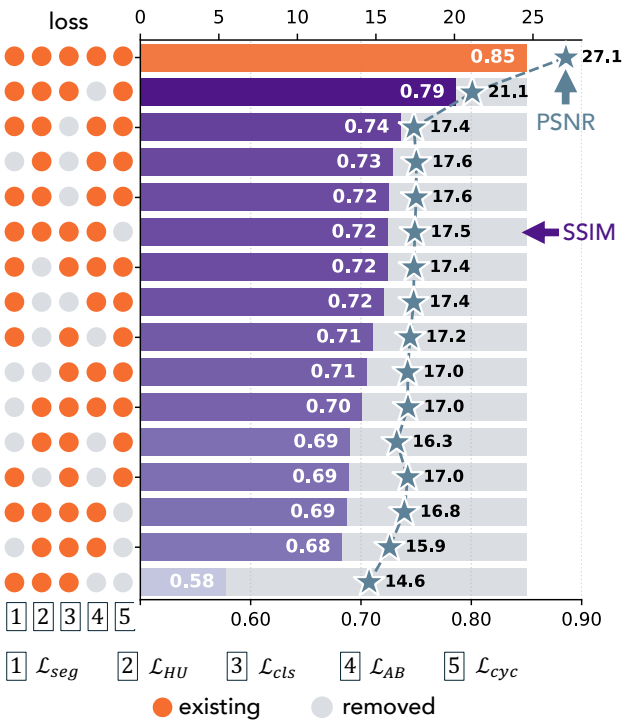

Figure 7. **Structural and intensity losses drive the largest gains.** Removing any loss module reduces performance, confirming that all five components contribute to SMILE's stability. The biggest drops occur when structural losses ($\mathcal{L}_{\text{seg}}$, $\mathcal{L}_{\text{cyc}}$) are removed, showing they are essential for preserving anatomy. Intensity losses ($\mathcal{L}_{\text{HU}}$, $\mathcal{L}_{\text{AB}}$) and the phase loss ($\mathcal{L}_{\text{cls}}$) also provide strong improvements by enforcing realistic contrast behavior, while phase supervision adds a smaller but consistent benefit.

in Table 3. Adding SMILE increases sensitivity, f1-score and AUC across all settings, with modest changes in specificity, indicating improved early tumor detectability without introducing false-positive drift. These results suggest that SMILE can serve as a general enhancement module to improve a wide range of tumor detection systems operating on low-contrast CT scans.

4.5. Ablation Study of SMILE Components

We conduct an ablation study on 40 patients from our test datasets to assess the contribution of each loss module in SMILE. A small paired subset is sufficient to show the effect of each component. The model uses five losses: $\mathcal{L}_{\text{seg}}$ and $\mathcal{L}_{\text{cyc}}$ for structure, $\mathcal{L}_{\text{cls}}$ for phase, and $\mathcal{L}_{\text{HU}}$ and $\mathcal{L}_{\text{AB}}$ for intensity. In each experiment, we remove one or two losses and train the model under the same settings as SMILE. As shown in Figure 7, removing any module reduces SSIM and PSNR, demonstrating that all five losses are essential for stable, high-quality enhancement.

5. Related Work

For CT phase translation, existing image enhancement and translation methods (summarized in Table 2) do not meet three key needs at the same time: (1) preserving anatomy, (2) handling unregistered multi-phase CT scans, and (3) producing clinically meaningful enhancement.

Natural image enhancement. Paired methods, like Pix2Pix [38], rely on pixel-level losses and require perfectly aligned data, which is unrealistic for multi-phase CT because scans from different contrast phases are rarely voxel-aligned. Unpaired models, like CycleGAN [73], MUNIT [36], and CUT [57], remove the need for alignment, but often cause geometric shifts or hallucinated structures. Diffusion-based methods (DiffuseIT [42], UNIT-DDPM [62]) improve visual quality, yet their evaluation focuses on appearance rather than anatomical or clinical accuracy.

Medical image enhancement. In medical imaging, unpaired translation models have been extensively studied for cross-modality and cross-contrast synthesis (e.g., MRI–CT or T1–T2) [15, 54, 56, 58, 64]. ContourDiff [15] introduces edge-aware constraints to better preserve structure, while SynDiff [56] incorporates diffusion-based priors to improve robustness under unpaired settings. However, these methods are not specialized for CT contrast-phase enhancement, where HU values impose physical constraints and phase-dependent changes are confined to a limited set of anatomically meaningful regions. Recent GAN and diffusion models for CT contrast synthesis [25, 67, 70, 74–76] depend on voxel-level supervision or pre-registration, which is rarely available in routine clinical settings. CyTran [60] can operate without pairing, but still requires post-hoc registration.

6. Conclusion and Discussion

We introduced SMILE, an anatomy-aware diffusion framework that enhances CT images by modifying only clinically relevant regions and preserving all unchanged anatomy. With structural, phase, and intensity supervision, SMILE generates realistic contrast-enhanced images that follow true organ geometry and physiological patterns without creating false structures. The results are both visually convincing and clinically reliable.

SMILE also addresses an important clinical need. Many patients cannot receive contrast due to kidney issues, allergies, or unstable conditions, leaving clinicians to interpret limited non-contrast scans. SMILE provides a safe alternative by producing diagnostic-quality contrast images directly from non-contrast CT, without extra injections or radiation. This can improve care for vulnerable patients, aid rapid decisions in emergency settings, and simplify radiation oncology workflows by removing the need for cross-phase registration.

Acknowledgments. This work was supported by the Lustgarten Foundation for Pancreatic Cancer Research and the National Institutes of Health (NIH) under Award Number R01EB037669. We would like to thank the Johns Hopkins Research IT team in IT@JH for their support and infrastructure resources where some of these analyses were conducted; especially DISCOVERY HPC. We thank Jaimie Patterson for writing a news article about this project. Paper content is covered by patents pending.

References

[1] Michela Antonelli, Annika Reinke, Spyridon Bakas, Keyvan Farahani, Annette Kopp-Schneider, Bennett A Landman, Geert Litjens, Bjoern Menze, Olaf Ronneberger, Ronald M Summers, et al. The medical segmentation decathlon. *Nature communications*, 13(1):4128, 2022. 4, 5

[2] Brian B Avants, Nicholas J Tustison, Gang Song, Philip A Cook, Arno Klein, and James C Gee. A reproducible evaluation of ants similarity metric performance in brain image registration. *Neuroimage*, 54(3):2033–2044, 2011. 4, 2

[3] Krzysztof Bartnik, Tomasz Bartczak, Mateusz Krzyziński, Krzysztof Korzeniowski, Krzysztof Lamparski, Piotr Węgrzyn, Eric Lam, Mateusz Bartkowiak, Tadeusz Wróblewski, Katarzyna Mech, et al. Waw-tace: a hepatocellular carcinoma multiphase ct dataset with segmentations, radiomics features, and clinical data. *Radiology: Artificial Intelligence*, 6(6):e240296, 2024. 4, 5

[4] Pedro RAS Bassi, Wenxuan Li, Yucheng Tang, Fabian Isensee, Zifu Wang, Jieneng Chen, Yu-Cheng Chou, Yannick Kirchhoff, Maximilian Rokuss, Ziyan Huang, Jin Ye, Junjun He, Tassilo Wald, Constantin Ulrich, Michael Baumgartner, Saikat Roy, Klaus H. Maier-Hein, Paul Jaeger, Yiwen Ye, Yutong Xie, Jianpeng Zhang, Ziyang Chen, Yong Xia, Zhaohu Xing, Lei Zhu, Yousef Sadegheih, Afshin Bozorgpour, Pratibha Kumari, Reza Azad, Dorit Merhof, Pengcheng Shi, Ting Ma, Yuxin Du, Fan Bai, Tiejun Huang, Bo Zhao, Haonan Wang, Xiaomeng Li, Hanxue Gu, Haoyu Dong, Jichen Yang, Maciej A. Mazurowski, Saumya Gupta, Linshan Wu, Jiaxin Zhuang, Hao Chen, Holger Roth, Daguang Xu, Matthew B. Blaschko, Sergio Decherchi, Andrea Cavalli, Alan L. Yuille, and Zongwei Zhou. Touchstone benchmark: Are we on the right way for evaluating ai algorithms for medical segmentation? *Conference on Neural Information Processing Systems*, 2024. 6

[5] Pedro RAS Bassi, Mehmet Can Yavuz, Ibrahim Ethem Hamamci, Sezgin Er, Xiaoxi Chen, Wenxuan Li, Bjoern Menze, Sergio Decherchi, Andrea Cavalli, Kang Wang, Yang Yang, Alan Yuille, and Zongwei Zhou. Radgpt: Constructing 3d image-text tumor datasets. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23720–23730, 2025. 3, 4

[6] Pedro RAS Bassi, Xinze Zhou, Wenxuan Li, Szymon Płotka, Jieneng Chen, Qi Chen, Zheren Zhu, Jakub Prządo, Ibrahim E Hamamci, Sezgin Er, Xiaoxi Chen, Mehmet Can Yavuz, Yu-Cheng Chou, Tianyu Lin, Kang Wang, Yucheng Tang, Jaroslaw B Cwikla, Sergio Decherchi, Andrea Cavalli, Yang Yang, Alan L Yuille, and Zongwei Zhou. Scaling artificial intelligence for multi-tumor early detection with more reports, fewer masks. *arXiv preprint arXiv:2510.14803*, 2025. 2

[7] Patrick Bilic, Patrick Christ, Hongwei Bran Li, Eugene Vorontsov, Avi Ben-Cohen, Georgios Kaissis, Adi Szeskin, Colin Jacobs, Gabriel Efrain Humpire Mamani, Gabriel Chartrand, et al. The liver tumor segmentation benchmark (lits). *Medical image analysis*, 84:102680, 2023. 4

[8] Sébastien Bubeck, Christian Coester, Ronen Eldan, Timothy Gowers, Yin Tat Lee, Alexandru Lupsasca, Mehtaab Sawhney, Robert Scherrer, Mark Sellke, Brian K Spears, et al. Early science acceleration experiments with gpt-5. *arXiv preprint arXiv:2511.16072*, 2025. 6

[9] Yuanhao Cai, Yixun Liang, Jiahao Wang, Angtian Wang, Yulun Zhang, Xiaokang Yang, Zongwei Zhou, and Alan Yuille. Radiative gaussian splatting for efficient x-ray novel view synthesis. *arXiv preprint arXiv:2403.04116*, 2024. 2

[10] Yuanhao Cai, Jiahao Wang, Alan Yuille, Zongwei Zhou, and Angtian Wang. Structure-aware sparse-view x-ray 3d reconstruction. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11174–11183, 2024. 2

[11] Kai Cao, Yingda Xia, Jiawen Yao, Xu Han, Lukas Lambert, Tingting Zhang, Wei Tang, Gang Jin, Hui Jiang, Xu Fang, et al. Large-scale pancreatic cancer detection via non-contrast ct and deep learning. *Nature medicine*, 29(12): 3033–3043, 2023. 2

[12] Qi Chen, Xiaoxi Chen, Haorui Song, Zhiwei Xiong, Alan Yuille, Chen Wei, and Zongwei Zhou. Towards generalizable tumor synthesis. In *IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 11147–11158, 2024. 1

[13] Qi Chen, Yuxiang Lai, Xiaoxi Chen, Qixin Hu, Alan Yuille, and Zongwei Zhou. Analyzing tumors by synthesis. *Generative Machine Learning Models in Medical Image Computing*, pages 85–110, 2024.

[14] Qi Chen, Xinze Zhou, Chen Liu, Hao Chen, Wenxuan Li, Zekun Jiang, Ziyan Huang, Yuxuan Zhao, Dexin Yu, Junjun He, Yefeng Zheng, Ling Shao, Alan Yuille, and Zongwei Zhou. Scaling tumor segmentation: Best lessons from real and synthetic data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 24001–24013, 2025. 1

[15] Yuwen Chen, Nicholas Konz, Hanxue Gu, Haoyu Dong, Yaqian Chen, Lin Li, Jisoo Lee, and Maciej A Mazurowski. Contourdiff: Unpaired image-to-image translation with structural consistency for medical imaging. *arXiv preprint arXiv:2403.10786*, 2024. 8

[16] Yixiong Chen, Wenjie Xiao, Pedro RAS Bassi, Xinze Zhou, Sezgin Er, Ibrahim Ethem Hamamci, Zongwei Zhou, and Alan Yuille. Are vision language models ready for clinical diagnosis? a 3d medical benchmark for tumor-centric visual question answering. *arXiv preprint arXiv:2505.18915*, 2025. 4

[17] Yu-Cheng Chou, Zongwei Zhou, and Alan Yuille. Embracing massive medical data. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 24–35. Springer, 2024. 4

[18] Casey Chu, Andrey Zhmoginov, and Mark Sandler. Cyclegan, a master of steganography. *arXiv preprint arXiv:1712.02950*, 2017. 5, 6, 7, 8, 9

[19] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 6

[20] Binh T Dao, Thang V Nguyen, Hieu H Pham, and Ha Q Nguyen. Phase recognition in contrast-enhanced ct scans based on deep learning and random sampling. *Medical Physics*, 49(7):4518–4528, 2022. 4

[21] Tami D DenOtter and Johanna Schubert. Hounsfield unit. 2019. 3

[22] Shiyi Du, Xiaosong Wang, Yongyi Lu, Yuyin Zhou, Shaoting Zhang, Alan Yuille, Kang Li, and Zongwei Zhou. Boosting dermatoscopic lesion segmentation via diffusion models with visual and textual prompts. In *2024 IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 1–5. IEEE, 2024. 1

[23] Patrick Esser, Robin Rombach, and Bjorn Ommer. Taming transformers for high-resolution image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12873–12883, 2021. 5, 6, 7, 8, 9

[24] Yunhe Gao, Mu Zhou, Di Liu, Zhennan Yan, Shaoting Zhang, and Dimitris N Metaxas. A data-scalable transformer for medical image segmentation: architecture, model efficiency, and benchmark. *arXiv preprint arXiv:2203.00131*, 2022. 6

[25] Syed Jamal Safdar Gardezi, Lucas Aronson, Peter Wawrzyn, Hongkun Yu, E Jason Abel, Daniel D Shapiro, Meghan G Lubner, Joshua Warner, Giuseppe Toia, Lu Mao, et al. Resnct: A deep learning model for the synthesis of nephrographic phase images in ct urography. *arXiv preprint arXiv:2405.04629*, 2024. 8

[26] Pengfei Guo, Can Zhao, Dong Yang, Yufan He, Vishwesh Nath, Ziyue Xu, Pedro RAS Bassi, Zongwei Zhou, Benjamin D Simon, Stephanie Anne Harmon, Ali B Syed, Holger Roth, and Daguang Xu. Text2ct: Towards 3d ct volume generation from free-text descriptions using diffusion model. *arXiv preprint arXiv:2505.04522*, 2025. 2

[27] Ibrahim Ethem Hamamci, Sezgin Er, Chenyu Wang, Furkan Almas, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Irem Dogan, Omer Faruk Durugol, Benjamin Hou, Suprosanna Shit, et al. Developing generalist foundation models from a multimodal dataset for 3d computed tomography. *arXiv preprint arXiv:2403.17834*, 2024. 4, 6

[28] Yufan He, Pengfei Guo, Yucheng Tang, Andriy Myronenko, Vishwesh Nath, Ziyue Xu, Dong Yang, Can Zhao, Benjamin Simon, Mason Belue, et al. Vista3d: A unified segmentation foundation model for 3d medical imaging. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 20863–20873, 2025. 6

[29] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 4

[30] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 1

[31] Can Hu, Yingda Xia, Zhilin Zheng, Mengxuan Cao, Guoliang Zheng, Shangqi Chen, Jiancheng Sun, Wujie Chen, Qi Zheng, Siwei Pan, et al. Ai-based large-scale screening of gastric cancer from noncontrast ct imaging. *Nature Medicine*, pages 1–9, 2025. 2

[32] Qixin Hu, Junfei Xiao, Yixiong Chen, Shuwen Sun, Jie-Neng Chen, Alan Yuille, and Zongwei Zhou. Synthetic tumors make ai segment tumors better. *NeurIPS Workshop on Medical Imaging meets NeurIPS*, 2022. 2

[33] Qixin Hu, Yixiong Chen, Junfei Xiao, Shuwen Sun, Jieneng Chen, Alan L Yuille, and Zongwei Zhou. Label-free liver tumor segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7422–7432, 2023.

[34] Qixin Hu, Alan Yuille, and Zongwei Zhou. Synthetic data as validation. *arXiv preprint arXiv:2310.16052*, 2023. 2

[35] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017. 6

[36] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 172–189, 2018. 8

[37] Fabian Isensee, Paul F Jaeger, Simon AA Kohl, Jens Petersen, and Klaus H Maier-Hein. nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. *Nature methods*, 18(2):203–211, 2021. 3

[38] 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*, pages 1125–1134, 2017. 4, 5, 6, 8, 7, 9

[39] Firas Khader, Gustav Müller-Franzes, Soroosh Tayebi Arasteh, Tianyu Han, Christoph Haarburger, Maximilian Schulze-Hagen, Philipp Schad, Sandy Engelhardt, Bettina Baeßler, Sebastian Foersch, et al. Denoising diffusion probabilistic models for 3d medical image generation. *Scientific Reports*, 13(1):7303, 2023. 5, 6, 7, 8, 9

[40] 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*. 2016. 4

[41] Jari Korhonen and Junyong You. Peak signal-to-noise ratio revisited: Is simple beautiful? In *2012 Fourth international workshop on quality of multimedia experience*, pages 37–38. IEEE, 2012. 4

[42] Gihyun Kwon and Jong Chul Ye. Diffusion-based image translation using disentangled style and content representation. *arXiv preprint arXiv:2209.15264*, 2022. 8

[43] Yuxiang Lai, Xiaoxi Chen, Angtian Wang, Alan Yuille, and Zongwei Zhou. From pixel to cancer: Cellular automata in computed tomography. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 36–46. Springer, 2024. 2

[44] Bowen Li, Yu-Cheng Chou, Shuwen Sun, Hualin Qiao, Alan Yuille, and Zongwei Zhou. Early detection and localization of pancreatic cancer by label-free tumor synthesis. *MICCAI Workshop on Big Task Small Data, 1001-AI*, 2023. 2

[45] Jianning Li, Zongwei Zhou, Jiancheng Yang, Antonio Pepe, Christina Gsaxner, Gijs Luijten, Chongyu Qu, Tiezheng Zhang, Xiaoxi Chen, Wenxuan Li, Yuan Jin, and Jan Egger. Medshapenet–a large-scale dataset of 3d medical shapes for computer vision. *Biomedical Engineering/Biomedizinische Technik*, (0), 2024. 4

[46] Wenxuan Li, Chongyu Qu, Xiaoxi Chen, Pedro RAS Bassi, Yijia Shi, Yuxiang Lai, Qian Yu, Huimin Xue, Yixiong Chen, Xiaorui Lin, Yutong Tang, Yining Cao, Haoqi Han, Zheyuan Zhang, Jiawei Liu, Tiezheng Zhang, Yujiu Ma, Jincheng Wang, Guang Zhang, Alan Yuille, and Zongwei Zhou. Abdomenatlas: A large-scale, detailed-annotated, & multi-center dataset for efficient transfer learning and open algorithmic benchmarking. *Medical Image Analysis*, page 103285, 2024. 3

[47] Wenxuan Li, Alan Yuille, and Zongwei Zhou. How well do supervised models transfer to 3d image segmentation? In *International Conference on Learning Representations*, 2024. 3, 4

[48] Wenxuan Li, Pedro RAS Bassi, Tianyu Lin, Yu-Cheng Chou, Xinze Zhou, Yucheng Tang, Fabian Isensee, Kang Wang, Qi Chen, Xiaowei Xu, Jin Ye, Zheren Zhu, Sergio Decherchi, Andrea Cavalli, Alan L Yuille, and Zongwei Zhou. Scalemai: Accelerating the development of trusted datasets and ai models. *arXiv preprint arXiv:2501.03410*, 2025. 6

[49] Wenxuan Li, Xinze Zhou, Qi Chen, Tianyu Lin, Pedro RAS Bassi, Szymon Plotka, Jaroslaw B Cwikla, Xiaoxi Chen, Chen Ye, Zheren Zhu, Yu-Cheng Chou, Kang Wang, Yucheng Tang, Alan L Yuille, and Zongwei Zhou. Pants: The pancreatic tumor segmentation dataset. *arXiv preprint arXiv:2507.01291*, 2025. 3, 4

[50] Xinran Li, Yi Shuai, Chen Liu, Qi Chen, Qilong Wu, Pengfei Guo, Dong Yang, Can Zhao, Pedro RAS Bassi, Daguang Xu, and Zongwei Zhou. Text-driven tumor synthesis. *arXiv preprint arXiv:2412.18589*, 2024. 1

[51] Tianyu Lin, Xinran Li, Chuntung Zhuang, Qi Chen, Yuanhao Cai, Kai Ding, Alan L Yuille, and Zongwei Zhou. Are pixel-wise metrics reliable for sparse-view computed tomography reconstruction? *arXiv preprint arXiv:2506.02093*, 2025. 2

[52] Jie Liu, Yixiao Zhang, Jie-Neng Chen, Junfei Xiao, Yongyi Lu, Bennett A Landman, Yixuan Yuan, Alan Yuille, Yucheng Tang, and Zongwei Zhou. Clip-driven universal model for organ segmentation and tumor detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21152–21164, 2023. 4

[53] Jie Liu, Yixiao Zhang, Kang Wang, Mehmet Can Yavuz, Xiaoxi Chen, Yixuan Yuan, Haoliang Li, Yang Yang, Alan Yuille, Yucheng Tang, and Zongwei Zhou. Universal and extensible language-vision models for organ segmentation and tumor detection from abdominal computed tomography. *Medical Image Analysis*, page 103226, 2024. 4

[54] Yimin Luo, Qinyu Yang, Ziyi Liu, Zenglin Shi, Weimin Huang, Guoyan Zheng, and Jun Cheng. Target-guided diffusion models for unpaired cross-modality medical image translation. *IEEE Journal of Biomedical and Health Informatics*, 28(7):4062–4071, 2024. 8

[55] Jiawei Mao, Yuhan Wang, Yucheng Tang, Daguang Xu, Kang Wang, Yang Yang, Zongwei Zhou, and Yuyin Zhou. Medsegfactory: Text-guided generation of medical image-mask pairs. *arXiv preprint arXiv:2504.06897*, 2025. 2

[56] Muzaffer Özbey, Onat Dalmaz, Salman UH Dar, Hasan A Bedel, Şaban Özturk, Alper Güngör, and Tolga Cukur. Unsupervised medical image translation with adversarial diffusion models. *IEEE Transactions on Medical Imaging*, 42 (12):3524–3539, 2023. 8

[57] Taesung Park, Alexei A Efros, Richard Zhang, and Jun-Yan Zhu. Contrastive learning for unpaired image-to-image translation. In *European conference on computer vision*, pages 319–345. Springer, 2020. 5, 6, 8, 7, 9

[58] Vu Minh Hieu Phan, Yutong Xie, Bowen Zhang, Yuankai Qi, Zhibin Liao, Antonios Perperidis, Son Lam Phung, Johan W Verjans, and Minh-Son To. Structural attention: Rethinking transformer for unpaired medical image synthesis. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 690–700. Springer, 2024. 8

[59] Chongyu Qu, Tiezheng Zhang, Hualin Qiao, Jie Liu, Yucheng Tang, Alan Yuille, and Zongwei Zhou. Abdomenatlas-8k: Annotating 8,000 abdominal ct volumes for multi-organ segmentation in three weeks. In *Conference on Neural Information Processing Systems*, 2023. 3, 4

[60] Nicolae-Cătălin Ristea, Andreea-Iuliana Miron, Olivian Savencu, Mariana-Iuliana Georgescu, Nicolae Verga, Fahad Shahbaz Khan, and Radu Tudor Ionescu. Cytran: A cycle-consistent transformer with multi-level consistency for non-contrast to contrast ct translation. *Neurocomputing*, 538: 126211, 2023. 5, 6, 8, 7, 9

[61] 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*, pages 10684–10695, 2022. 3, 4

[62] Hiroshi Sasaki, Chris G Willcocks, and Toby P Breckon. Unit-ddpm: Unpaired image translation with denoising diffusion probabilistic models. *arXiv preprint arXiv:2104.05358*, 2021. 8

[63] Md Mahfuzur Rahman Siddiquee, Zongwei Zhou, Nima Tajbakhsh, Ruibin Feng, Michael B Gotway, Yoshua Bengio, and Jianming Liang. Learning fixed points in generative adversarial networks: From image-to-image translation to disease detection and localization. In *IEEE International Conference on Computer Vision (ICCV)*, pages 191–200, 2019. 2

[64] Haoshen Wang, Xiaodong Wang, and Zhiming Cui. Structure-preserving diffusion model for unpaired medical

image translation. In *International Workshop on Machine Learning in Medical Imaging*, pages 218–227. Springer, 2024. 8

[65] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 4

[66] Yingda Xia, Qihang Yu, Linda Chu, Satomi Kawamoto, Seyoun Park, Fengze Liu, Jieneng Chen, Zhuotun Zhu, Bowen Li, Zongwei Zhou, Alan L Yuille, Elliot K Fishman, and Ralph H Hruban. The felix project: Deep networks to detect pancreatic neoplasms. *medRxiv*, 2022. 4

[67] Ning Xiao, Zhenyu Li, Shaobo Chen, Liangtian Zhao, Yuer Yang, Hao Xie, Yang Liu, Yujuan Quan, and Junwei Duan. Contrast-enhanced ct image synthesis of thyroid based on transfomer and texture branching. In *2022 5th International Conference on Artificial Intelligence and Big Data (ICAIBD)*, pages 94–100. IEEE, 2022. 8

[68] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 6

[69] Yijun Yang, Zhao-Yang Wang, Qiuping Liu, Shuwen Sun, Kang Wang, Rama Chellappa, Zongwei Zhou, Alan Yuille, Lei Zhu, Yu-Dong Zhang, and Jieneng Chen. Medical world model: Generative simulation of tumor evolution for treatment planning. *arXiv preprint arXiv:2506.02327*, 2025. 2

[70] Juanjuan Yin, Jinye Peng, Xiaohui Li, and Jun Wang. Enhanced aortic ct synthesis based on multiscale information fusion. *IEEE MultiMedia*, 2025. 8

[71] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023. 1

[72] Zongwei Zhou, Michael B Gotway, and Jianming Liang. Interpreting medical images. In *Intelligent Systems in Medicine and Health*, pages 343–371. Springer, 2022. 3

[73] 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*, pages 2223–2232, 2017. 8

[74] Qikui Zhu, Hao Wu, Yanyan Zhang, and Shuo Li. Causality-driven spatio-temporal generator for multi-phase contrast-enhanced ct synthesis. 8

[75] Qikui Zhu, Andrew L Wentland, and Shuo Li. Contrast-aware network with aggregated-interacted transformer and multi-granularity aligned contrastive learning for synthesizing contrast-enhanced abdomen ct imaging. *IEEE Transactions on Computational Imaging*, 2025.

[76] Qikui Zhu, Shaoming Zhu, Bo Du, and Yanqing Wang. Cross-domain distribution adversarial diffusion model for synthesizing contrast-enhanced abdomen ct imaging. *Pattern Recognition*, page 111695, 2025. 8

Appendix

Table of Contents

A. Dataset **2**
- A.1. Anatomical Structures in CTVerse Dataset . . . 2
- A.2. Anatomical Structures in SMILE Test Datasets . . . 2

B. CT Contrast Phases **3**
- B.1. Contrast Phases Change Anatomical Appearance . . . 3
- B.2. Contrast Phases Improve Tumor Diagnosis . . . 3

C. SMILE Implementations Details **4**
- C.1. Why SMILE describes three loss fields but uses five loss weights? . . . 4
- C.2. Why SMILE does not need registration? . . . 4
- C.3. Why SMILE introduces supervision progressively? . . . 4
- C.4. Why SMILE can adjust loss weights automatically? . . . 5

D. SMILE Comparison in Early Tumor Detection with Baselines **6**

E. SMILE Comparison in Enhancement Quality with Baselines (Ground-Truth Provided) **8**

F. SMILE Compared with Commercialized Generative Vision Models **10**

G SMILE Performance on Specific Organs **12**

A. Dataset

A.1. Anatomical Structures in CTVerse Dataset

We will present a precisely annotated, phase-wise paired, and organ-wise registered high-quality dataset, CTVerse, containing 1814 CT volumes from 477 patients, collecting from over 112 hospitals. For each CT scan, CTVerse provides 88 anatomical structures. All volumes were registered at the voxel level using the ANTs [2] toolkit to ensure accurate spatial alignment across phases.

Table 4. Class mapping in the CTVerse dataset.

name	type	label	**name**	type	label	**name**	type	label	**name**	type	label
adrenal gland left	O	1	kidney left	O	23	rib left 3	B	45	vertebrae C3	B	67
adrenal gland right	O	2	kidney lesion	L	24	rib left 4	B	46	vertebrae C4	B	68
aorta	O	3	kidney right	O	25	rib left 5	B	47	vertebrae C5	B	69
bile duct (common)	O	4	liver	O	26	rib left 6	B	48	vertebrae C6	B	70
bladder	O	5	liver segment 1	P	27	rib left 7	B	49	vertebrae C7	B	71
celiac trunk	O	6	liver segment 2	P	28	rib left 8	B	50	vertebrae L1	B	72
colon	O	7	liver segment 3	P	29	rib left 9	B	51	vertebrae L2	B	73
duodenum	O	8	liver segment 4	P	30	rib left 10	B	52	vertebrae L3	B	74
esophagus	O	9	liver segment 5	P	31	rib left 11	B	53	vertebrae L4	B	75
gall bladder	O	10	liver segment 6	P	32	rib left 12	B	54	vertebrae L5	B	76
hepatic vessel	O	11	liver segment 7	P	33	rib right 1	B	55	vertebrae T1	B	77
intestine	O	12	liver segment 8	P	34	rib right 2	B	56	vertebrae T2	B	78
kidney left	O	13	liver lesion	L	35	rib right 3	B	57	vertebrae T3	B	79
kidney right	O	14	lung left	O	36	rib right 4	B	58	vertebrae T4	B	80
kidney lesion	L	15	lung right	O	37	rib right 5	B	59	vertebrae T5	B	81
liver	O	16	pancreas	O	38	rib right 6	B	60	vertebrae T6	B	82
liver segment 1	P	17	pancreas body	P	39	rib right 7	B	61	vertebrae T7	B	83
liver segment 2	P	18	pancreas head	P	40	rib right 8	B	62	vertebrae T8	B	84
liver segment 3	P	19	pancreas tail	P	41	rib right 9	B	63	vertebrae T9	B	85
liver segment 4	P	20	pancreatic lesion	L	42	rib right 10	B	64	vertebrae T10	B	86
liver segment 5	P	21	portal vein and splenic vein	O	43	rib right 11	B	65	vertebrae T11	B	87
liver segment 6	P	22	postcava	O	44	rib right 12	B	66	vertebrae T12	B	88

type: **B**: bone structure **L**: lesion **O**: whole organ **P**: organ sub-region

A.2. Anatomical Structures in SMILE Test Datasets

We use these 22 organs to evaluate how well SMILE preserves anatomical structure and tissue characteristics during CT phase conversion. For each organ, we measure two quantities:

- HU (Hounsfield Unit) correlation – whether the tissue density changes follow the correct pattern across phases.
- Size correlation – whether the enhanced organ keeps a similar volume as in the real CT.

Table 5. Anatomical structure-level HU analysis summary.

name	label	volume	name	label	volume
aorta	1	medium	intestine	12	large
bladder	2	medium	kidney left	13	large
celiac trunk	3	small	kidney right	14	large
colon	4	large	liver	15	large
duodenum	5	medium	pancreas body	16	medium
esophagus	6	small	pancreas head	17	medium
gall bladder	7	small	pancreas tail	18	small
hepatic vessel	8	small	portal vein and splenic vein	19	small
prostate	9	small	postcava	20	medium
rectum	10	small	spleen	21	medium
stomach	11	large	superior mesenteric artery	22	small

B. CT Contrast Phases

B.1. Contrast Phases Change Anatomical Appearance

A multi-phase CT scan takes several images of the same patient at different times after a contrast agent (iodine) is injected (see Figure 8). Think of it like taking four photos of the same scene under different lighting conditions. The anatomy does not change, but the brightness of specific tissues changes over time, because the contrast agent flows through arteries → organs → veins → then slowly washes out. Each phase highlights different organs and blood vessels, which is important for detecting pancreatic tumors, because they often look very similar to normal tissue in one phase but become visible in another.

- **Non-contrast: This is the scan before contrast injection.** Organs appear in their natural density. Pancreatic tumors are often hard to see, because both healthy pancreas and tumors look similar.
- **Arterial: Taken shortly after contrast agent injection.** Arteries and hyper-vascular structures become bright.
- **Venous: Contrast moves into veins and most abdominal organs.** Pancreatic tumors typically appear hypo-enhanced (darker) compared to the enhanced pancreas, making them easier to localize.
- **Delay: Contrast gradually washes out.** Tumors may retain different contrast patterns compared to surrounding tissue, providing another chance for detection.

B.2. Contrast Phases Improve Tumor Diagnosis

Contrast phases play a crucial role in detecting pancreatic tumors because they highlight how different tissues respond to injected contrast over time. In non-contrast scans, tumors often look similar to the surrounding pancreas, making them difficult to spot. Once contrast is injected, however, each phase reveals a different pattern of brightness across organs and vessels.

Blood-rich organs become bright in the arterial phase, most abdominal organs enhance in the venous phase, and contrast slowly fades in the delayed phase. Pancreatic tumors, however, usually stay darker than the surrounding pancreas in all these phases (see Figure 8), creating a clear visual difference that does not appear in non-contrast scans. These predictable brightness changes make contrast phases crucial: they reveal tumors that would otherwise stay hidden and give radiologists multiple chances to confirm where the tumor is and how big it is.

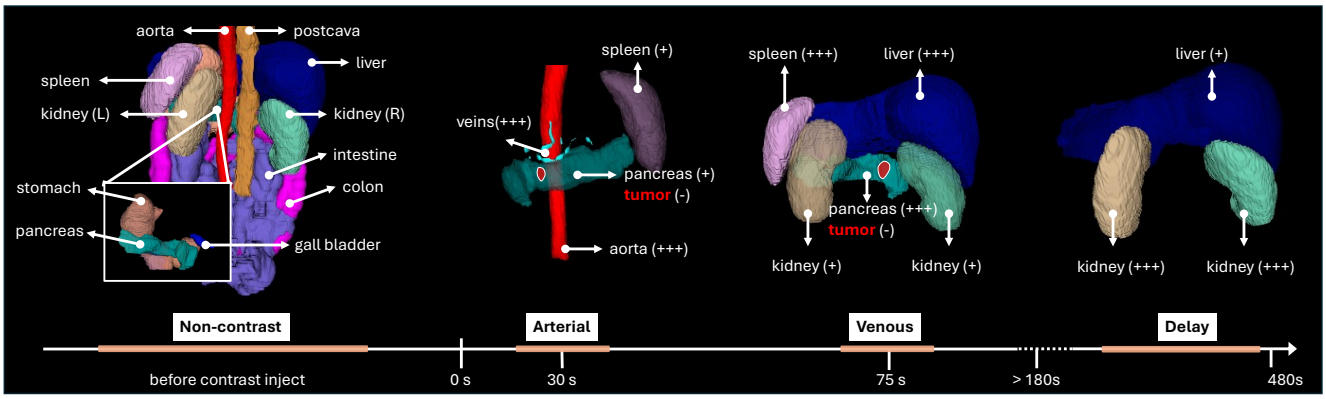

Figure 8. **Illustration of CT contrast phases.** From left to right: non-contrast phase (with anatomical structure labels), arterial phase, venous phase, and delay phase. In the enhancement phases, brighter organs are marked with "+" symbols (more "+" indicates stronger enhancement), while darker regions—such as tumors—are marked with "–". *Non-contrast Phase:* before contrast injection, most organs have similar brightness. *Arterial Phase:* after contrast injection, arteries, spleen, and pancreas brighten first. *Venous Phase:* after the arterial phase, the liver, pancreas and spleen brightness appear stronger. Pancreatic tumors often remain darker in the arterial and venous phase, making them easier to detect. *Delay Phase:* after the venous phase, kidneys reach peak enhancement, and liver enhancement gradually washes out.

C. SMILE Implementations Details

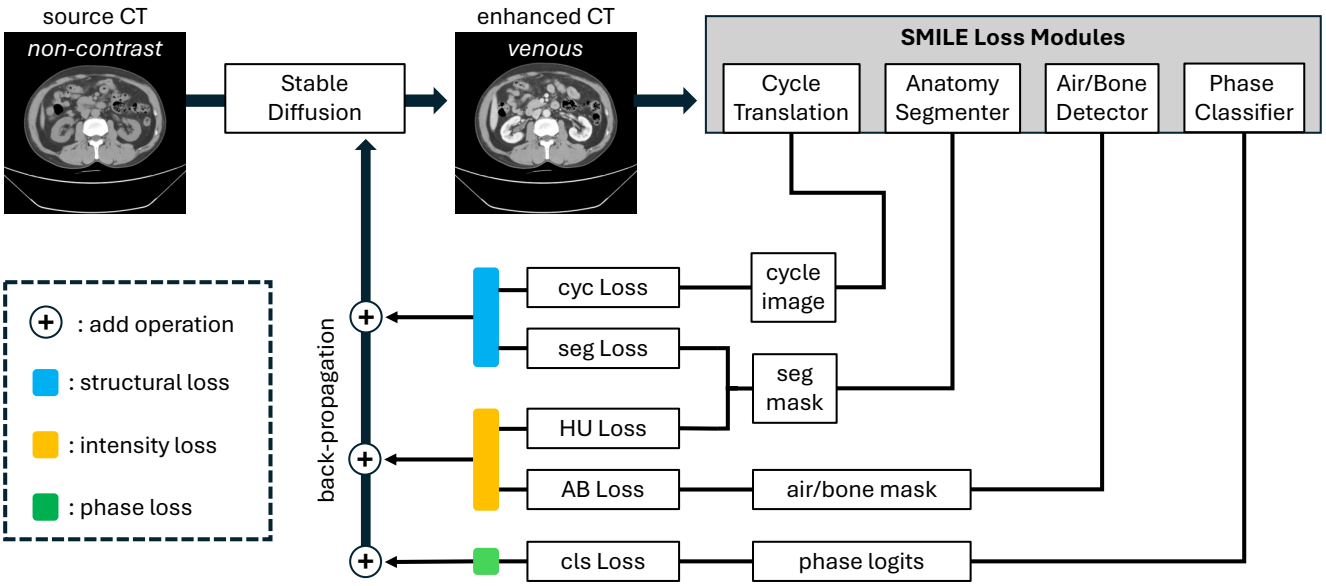

Figure 9. **Anatomical-aware design.** SMILE uses StableDiffusion [61] as the diffusion backbone, and use unique anatomy-aware loss design to ensure enhancement quality. *Structural loss:* Cycle translation and anatomy segmentation ensure anatomical structure preservation; *Intensity loss:* HU and air/bone losses enforce realistic intensities; *Phase loss:* A pretrained phase classifier ensures correct contrast phase. All losses back-propagate into the diffusion model to produce anatomically correct and phase-accurate enhanced CT scans.

C.1. Why SMILE describes three loss fields but uses five loss weights?

SMILE groups its supervision into three conceptual fields—structure, phase, and intensity—because these correspond to the three clinical requirements of contrast enhancement: preserving anatomy, achieving correct phase, and producing realistic tissue intensities. However, As shown in Figure 9, each field is implemented using practical loss components, resulting in five trainable losses: segmentation and cycle (structural), classifier (phase), and HU and air/bone (intensity). Thus, to reflect these components precisely, the full objective is:

$$\mathcal{L}_{\text{SMILE}} = \mathcal{L}_{\text{diff}} + \lambda_{\text{seg}}\mathcal{L}_{\text{seg}} + \lambda_{\text{cyc}}\mathcal{L}_{\text{cyc}} + \lambda_{\text{cls}}\mathcal{L}_{\text{cls}} + \lambda_{\text{HU}}\mathcal{L}_{\text{HU}} + \lambda_{\text{AB}}\mathcal{L}_{\text{AB}}. \tag{5}$$

C.2. Why SMILE does not need registration?

SMILE is trained on unregistered multi-phase CT by cleanly separating structural and intensity supervision. All geometry-related signals (segmentation + cycle consistency) come exclusively from the **source** CT, ensuring that no spatial information from the target phase is ever used. In contrast, the intensity and phase losses use only **alignment-free** cues from the target CT—mean organ HU values and slice-wise phase labels—which depend on global contrast behavior rather than voxel-wise correspondence. SMILE bypasses any need for voxel-level registration and remains robust to the natural misalignment between phases.

C.3. Why SMILE introduces supervision progressively?

We progressively activate loss modules during training to ensure stable optimization. Early in training (0–2k steps), the model learns only the diffusion prior, preventing supervision signals from dominating before the generator produces meaningful structures. Phase and cycle losses are added at 10k steps to guide global enhancement behavior once the model forms basic anatomy. At 20k steps, segmentation, HU, and air/bone losses are enabled to refine structural boundaries and organ-level intensity patterns.

This staged strategy avoids unstable gradients and ensures that high-level supervision is added only when the model is ready to benefit from it, leading to more stable and reliable enhancement quality.Let t denote the training step. The active

loss at step t is:

$$\mathcal{L}^{(t)} = \begin{cases} \mathcal{L}_{\text{diff}}, & 0 \le t < 2\text{k}, \\ \mathcal{L}_{\text{diff}} + \lambda_{\text{cyc}}\mathcal{L}_{\text{cyc}} + \lambda_{\text{cls}}\mathcal{L}_{\text{cls}}, & 2\text{k} \le t < 20\text{k}, \\ \mathcal{L}_{\text{diff}} + \lambda_{\text{cyc}}\mathcal{L}_{\text{cyc}} + \lambda_{\text{cls}}\mathcal{L}_{\text{cls}} + \lambda_{\text{seg}}\mathcal{L}_{\text{seg}} + \lambda_{\text{HU}}\mathcal{L}_{\text{HU}} + \lambda_{\text{AB}}\mathcal{L}_{\text{AB}}, & 20\text{k} \le t < 80\text{k}, \\ \mathcal{L}_{\text{diff}} + \lambda_{\text{cyc}}^{(t)}\mathcal{L}_{\text{cyc}} + \lambda_{\text{cls}}^{(t)}\mathcal{L}_{\text{cls}} + \lambda_{\text{seg}}^{(t)}\mathcal{L}_{\text{seg}} + \lambda_{\text{HU}}^{(t)}\mathcal{L}_{\text{HU}} + \lambda_{\text{AB}}^{(t)}\mathcal{L}_{\text{AB}}, & t \ge 80\text{k}, \end{cases} \tag{6}$$

where $\lambda_i^{(t)}$ are learnable weights estimated by the Uncertainty Loss Module.

C.4. Why SMILE can adjust loss weights automatically?

The idea of learnable loss weight is simple: if a supervision term is noisy or conflicting with others, its weight should be reduced; if a term is stable and provides consistent gradients, its weight should be increased.

Concretely, following standard uncertainty weighting for multi-task learning, we introduce one scalar parameter s_i per loss term $\mathcal{L}_i \in \{\mathcal{L}_{\text{cyc}}, \mathcal{L}_{\text{cls}}, \mathcal{L}_{\text{seg}}, \mathcal{L}_{\text{HU}}, \mathcal{L}_{\text{AB}}\}$. The combined objective is written as

$$\mathcal{L}_{\text{SMILE}} = \mathcal{L}_{\text{diff}} + \sum_i \exp(-s_i)\,\mathcal{L}_i + s_i, \tag{7}$$

and the effective weight for loss $\mathcal{L}_i$ at step t is $\lambda_i^{(t)} = \exp(-s_i)$. During training, s_i is optimized together with the network parameters via backpropagation. Intuitively, if a loss $\mathcal{L}_i$ is large or highly variable, the gradient w.r.t. s_i will increase s_i and thus decrease $\lambda_i^{(t)}$, down-weighting that term. Conversely, when a loss is small and consistent, s_i is reduced, which increases its weight.

This design is reasonable for two reasons. **First,** it avoids manual tuning of five various loss scales (segmentation, cycle, classification, HU, and air/bone), which naturally live on different numeric ranges. **Second,** the weighting is data-driven: the model learns which supervisions are currently reliable instead of relying on a fixed heuristic schedule.

D. SMILE Comparison in Early Tumor Detection with Baselines

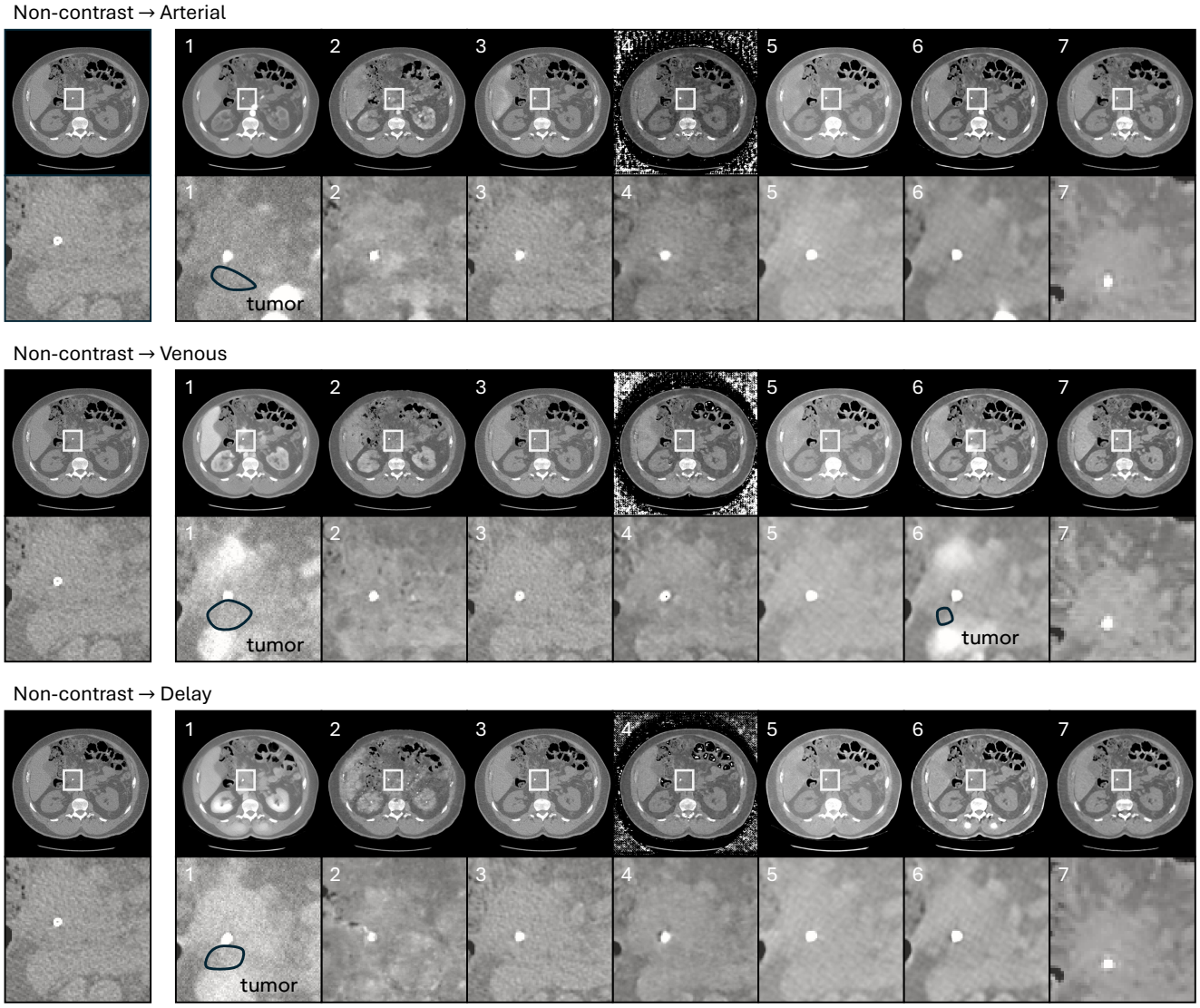

Figure 10. **SMILE outperforms all competing models in both anatomical integrity and tumor visibility (patient example #1).** For each source–target pair (non-contrast to arterial, venous, delay), the top row shows full slices and the bottom row zooms into the pancreas region. Methods are shown in the figure following this order: (1) SMILE, (2) Pix2Pix [38], (3) CycleGAN [18], (4) CyTran [60], (5) DALL-E [23], (6) MedDiffusion [39], (7) CUT [57].

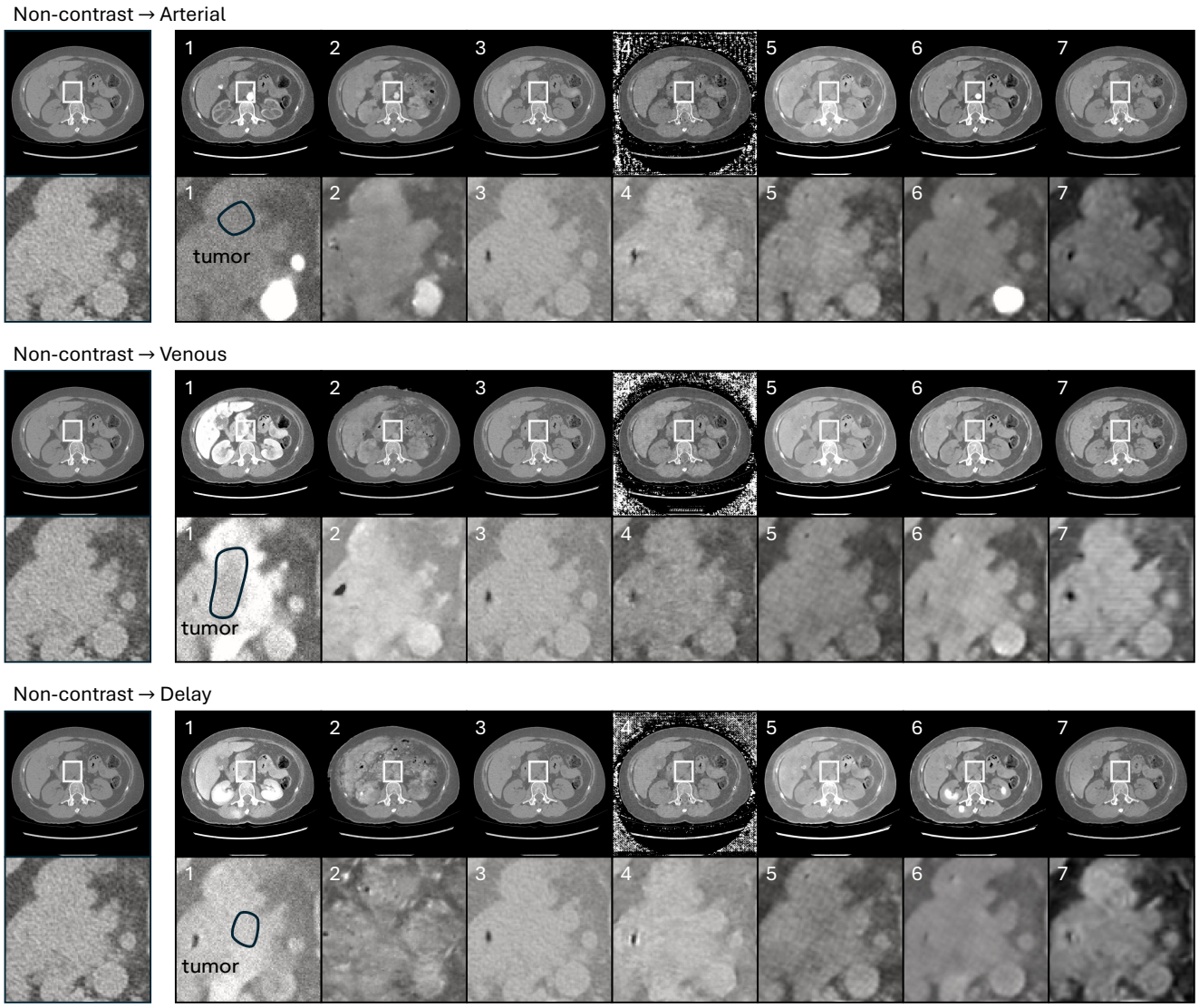

Figure 11. **SMILE outperforms all competing models in both anatomical integrity and tumor visibility (patient example #2).** For each source–target pair (non-contrast to arterial, venous, delay), the top row shows full slices and the bottom row zooms into the pancreas region. Methods are shown in the figure following this order: (1) SMILE, (2) Pix2Pix [38], (3) CycleGAN [18], (4) CyTran [60], (5) DALL-E [23], (6) MedDiffusion [39], (7) CUT [57]. SMILE improves tumor visibility, allowing the AI detector to identify the tumor reliably, while other baselines failed.

E. SMILE Comparison in Enhancement Quality with Baselines (Ground-Truth Provided)

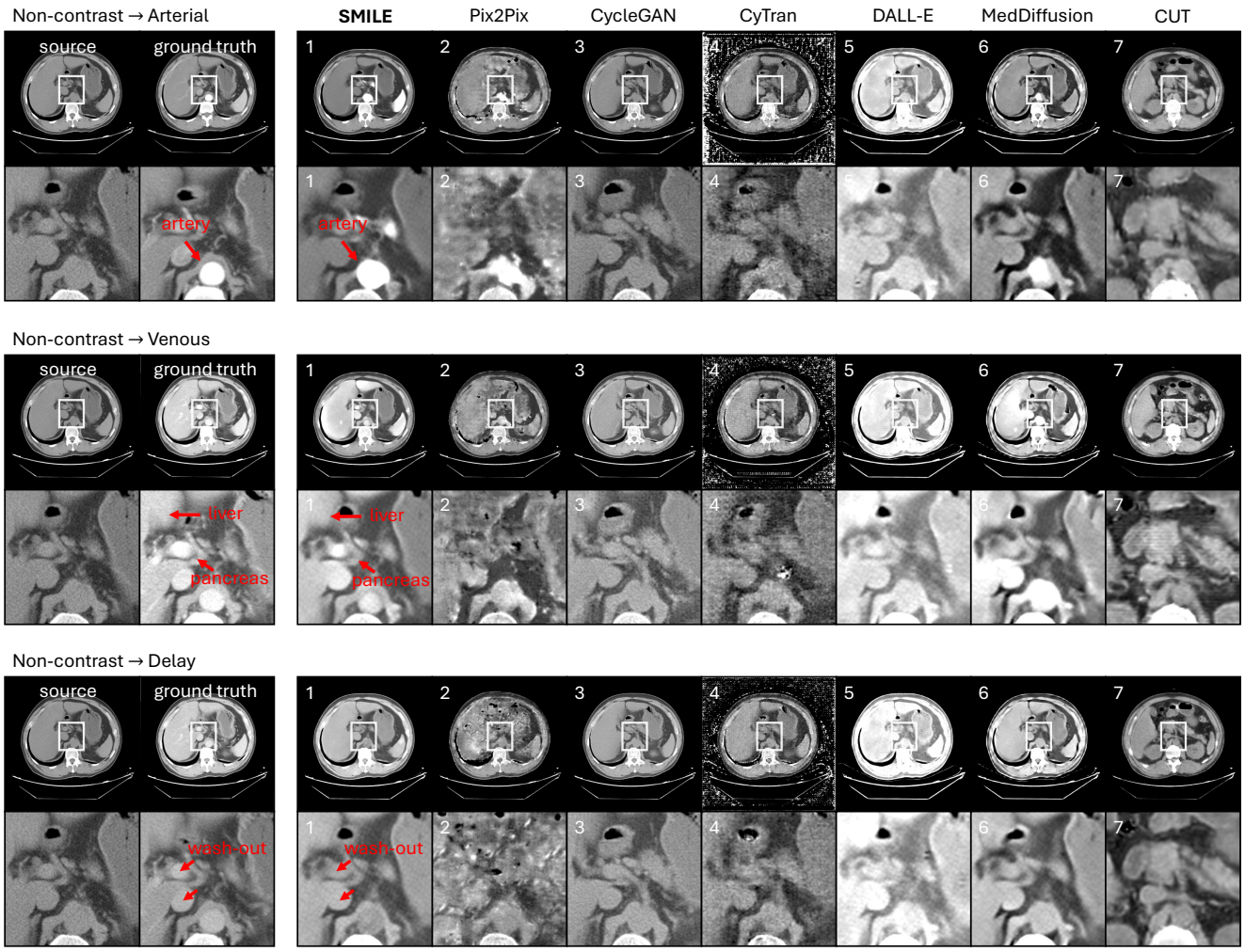

Figure 12. **SMILE produces the clearest and most phase-consistent enhancement (patient example #3).** Above shows the qualitative comparison of non-contrast to arterial, venous, and delay enhancement, with *registered ground truth provided*. The left-most column shows the source non-contrast CT and the ground-truth enhanced CT (arterial, venous, delay). All columns on the right show the enhanced results from different methods. For each source–target pair (non-contrast to arterial, venous, delay), the top row shows full slices and the bottom row zooms into the pancreas region. We compare SMILE (label 1) with six representative baselines: (2) Pix2Pix [38], (3) CycleGAN [18], (4) CyTran [60], (5) DALL-E [23], (6) MedDiffusion [39], and (7) CUT [57]. Baseline models often blur organ boundaries, introduce artifacts, or fail to reproduce the expected enhancement, especially around the pancreas and tumor region. On the contrary, across the three phase conversions—arterial, venous, and delay—SMILE shows the expected clinical patterns: In the arterial phase, SMILE correctly highlights the arteries, matching real contrast flow. In the venous phase, SMILE brightens the liver, pancreas, and veins, as seen in real venous enhancement. In the delay phase, SMILE reproduces the wash-out effect, where organ brightness slowly fades. These phase-specific changes match real CT behavior and help reveal tumors that are difficult to see in non-contrast scans.

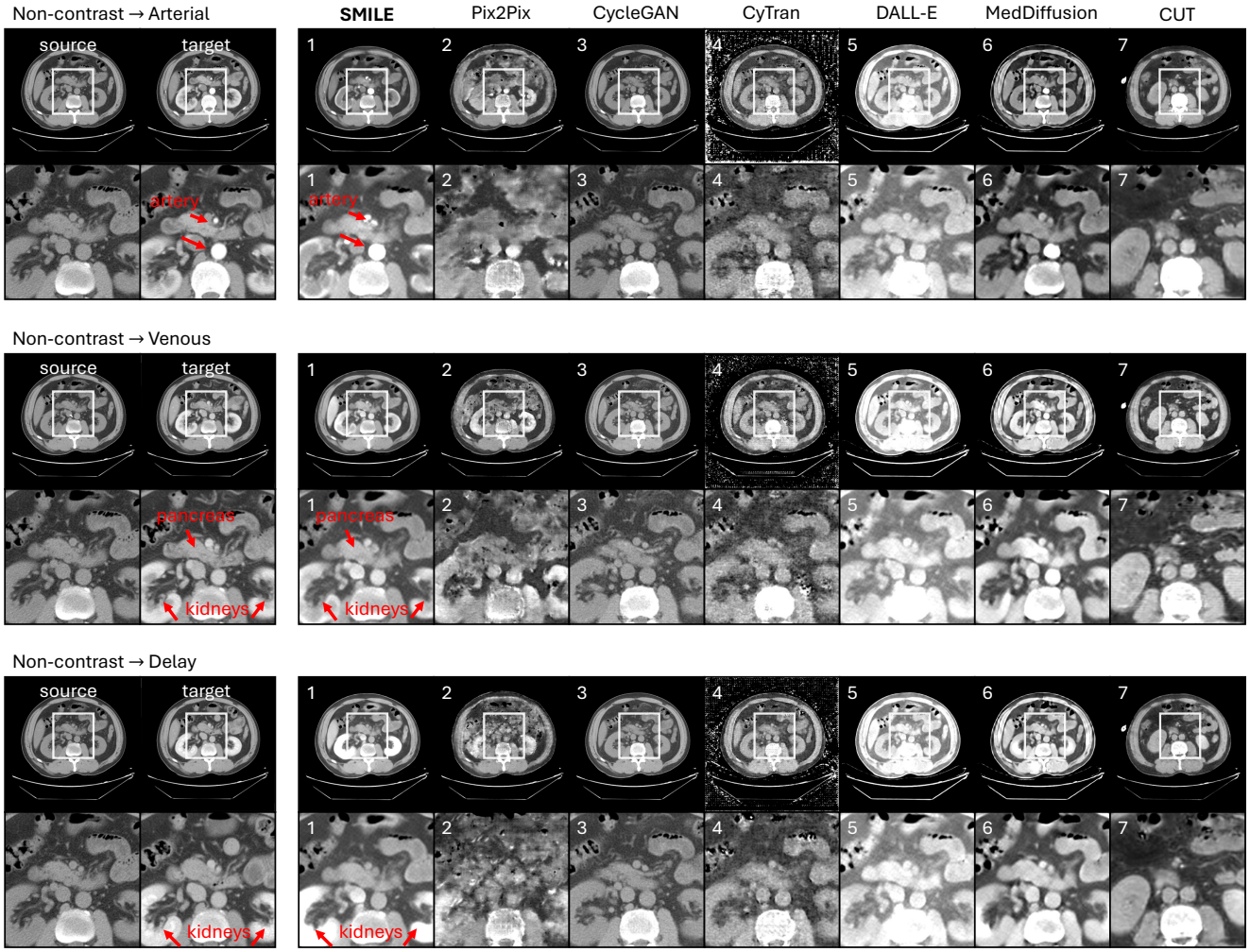

Figure 13. **SMILE produces the clearest and most phase-consistent enhancement (patient example #4).** Above shows the qualitative comparison of non-contrast to arterial, venous, and delay enhancement, with *registered ground truth provided*. The left-most column shows the source non-contrast CT and the ground-truth enhanced CT (arterial, venous, delay). All columns on the right show the enhanced results from different methods. For each source–target pair (non-contrast to arterial, venous, delay), the top row shows full slices and the bottom row zooms into the pancreas region. We compare SMILE (label 1) with six representative baselines: (2) Pix2Pix [38], (3) CycleGAN [18], (4) CyTran [60], (5) DALL-E [23], (6) MedDiffusion [39], and (7) CUT [57]. Baseline models often blur organ boundaries, introduce artifacts, or fail to reproduce the expected enhancement, especially around the pancreas and tumor region. On the contrary, across the three phase conversions—arterial, venous, and delay—SMILE shows the expected clinical patterns: In the arterial phase, SMILE correctly highlights the arteries, matching real contrast flow. In the venous phase, SMILE brightens the, pancreas, and veins, and part of kidney as seen in real venous enhancement. In the delay phase, SMILE correctly shows the wash-out pattern: only excretory organs such as the kidneys remain bright, while most other organs lose contrast. These phase-specific changes match real CT behavior and help reveal tumors that are difficult to see in non-contrast scans.

F. SMILE Compared with Commercialized Generative Vision Models

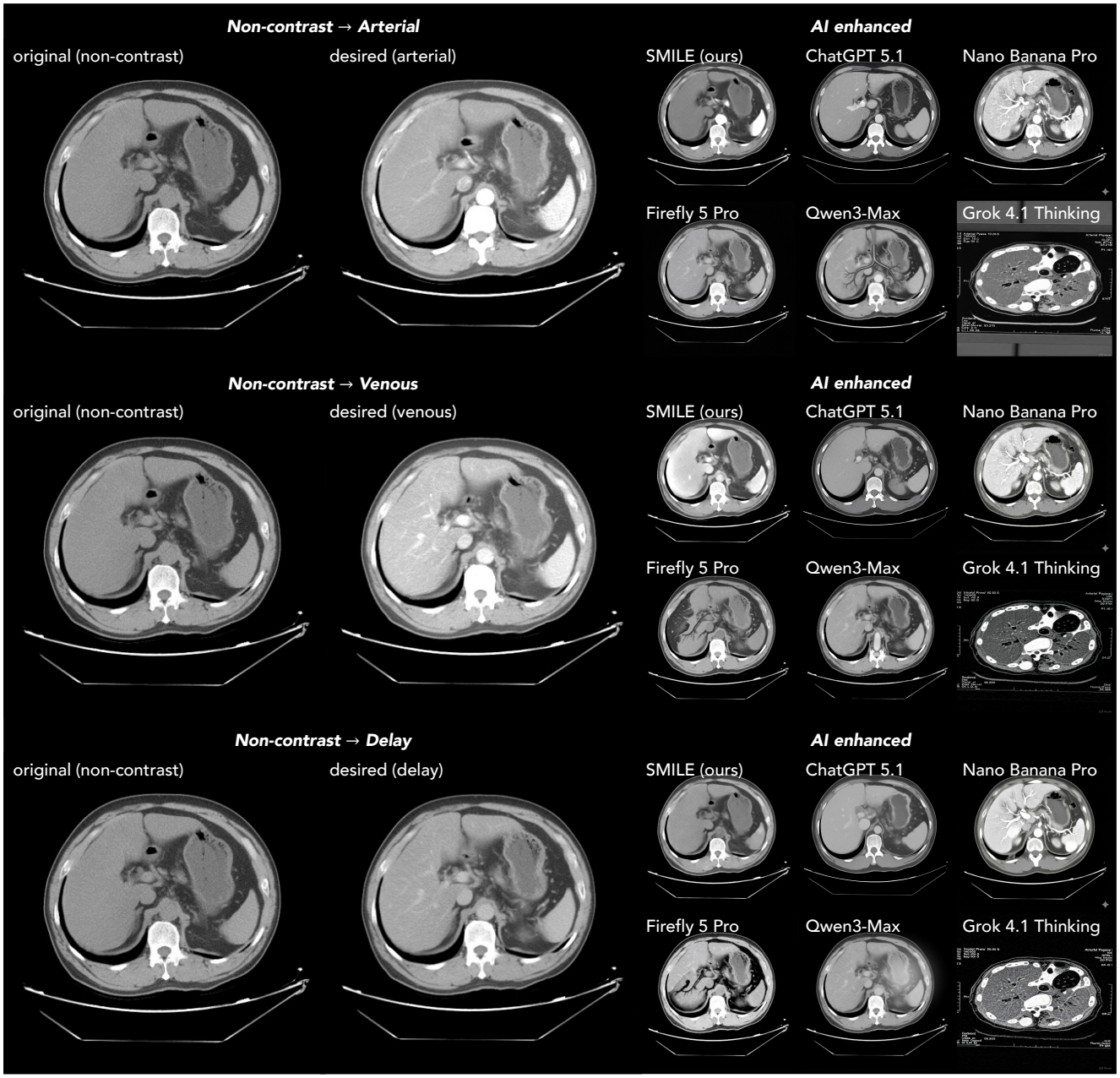

Figure 14. SMILE consistently produces the most anatomically accurate and clinically meaningful enhancement across all contrast phases, compares to current commercialized generative vision models.

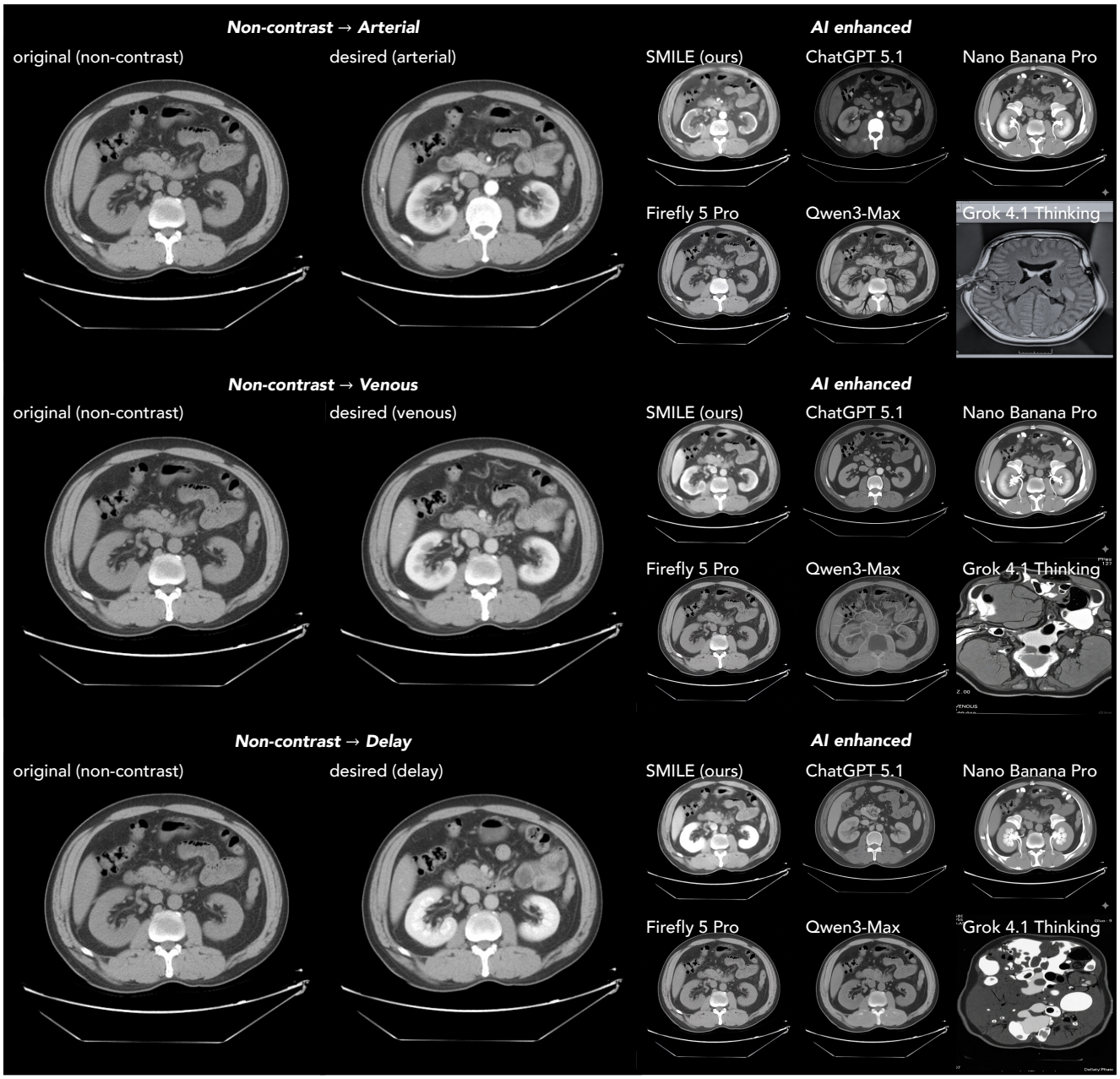

Figure 15. SMILE consistently produces the most anatomically accurate and clinically meaningful enhancement across all contrast phases, compares to current commercialized generative vision models.

G. SMILE Performance on Specific Organs

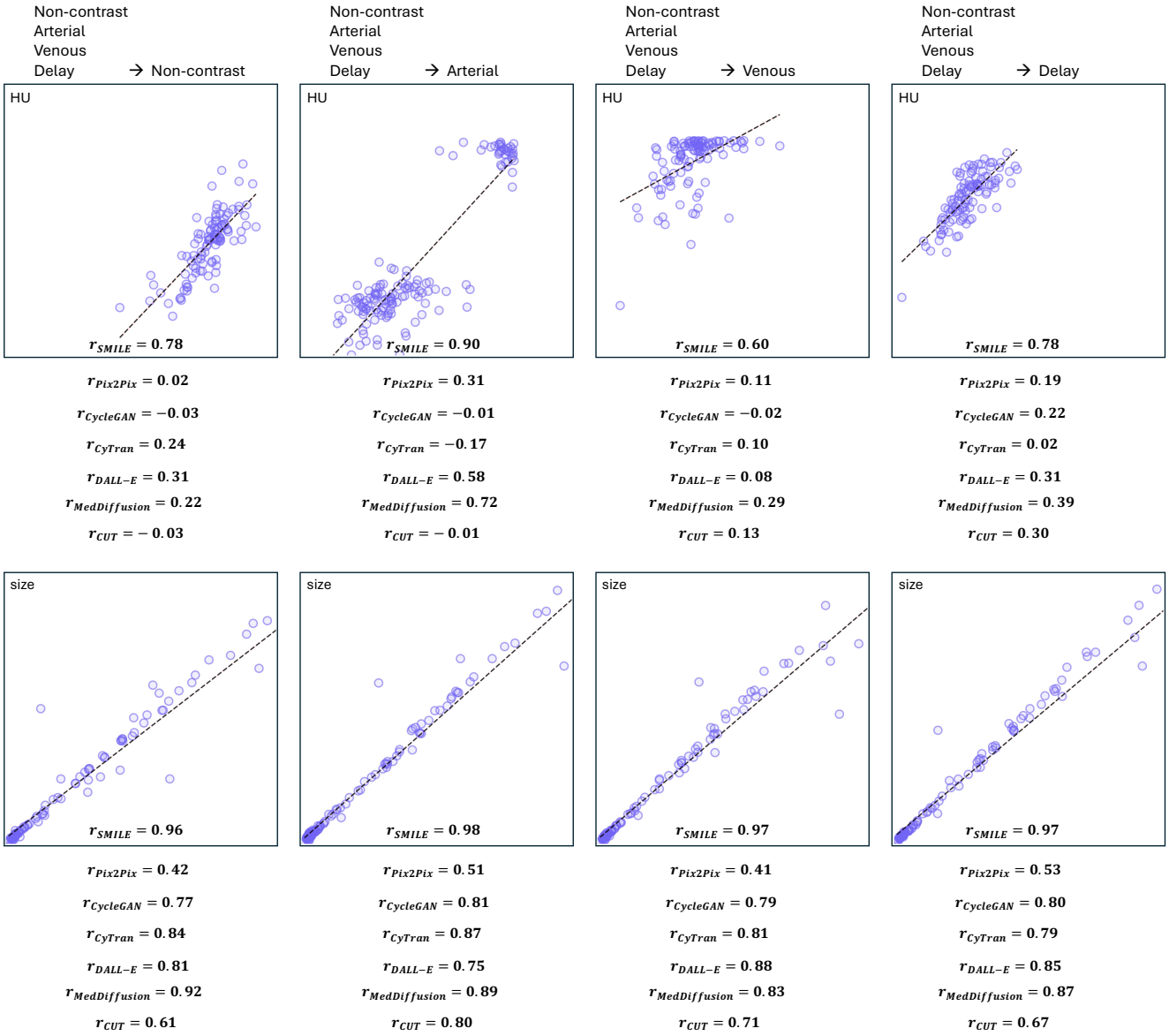

Figure 16. **Contrast enhancement of five key organs, i.e., aorta, liver, pancreas, kidneys, and spleen.** We evaluate SMILE's enhancement accuracy by measuring HU and size correlation between enhanced CT and ground truth on selected organs. They are selected because these structures show the strongest contrast changes across phases and are clinically important for tumor diagnosis. Each column shows one enhancement target. SMILE achieves high HU and size correlations in all settings compared to other baselines. Performance is lower in the venous phase, which is expected because venous enhancement depends on subtle global perfusion patterns that are harder to learn. Even so, SMILE remains consistently better than all baseline methods. Overall, these results show that SMILE produces anatomically and intensity-consistent enhancement for the most clinically relevant organs.