Rethinking CT Synthesis through Semantics-Structure Alignment

Riyu Qiu[1] **Qichao Zhou**[2] **Jiacheng Wang**[2*] **Liansheng Wang**[1,3,4*]

[1] Institute of Artificial Intelligence, Xiamen University, Xiamen, China
[2] Manteia Technology, Co., Ltd., Xiamen, China
[3] Department of Computer Science at the School of Informatics, Xiamen University, Xiamen, China
[4] National Institute for Data Science in Health and Medicine, Xiamen University, Xiamen, China

Abstract

Mask-guided CT synthesis bridges structural annotations with imaging appearance, facilitating data augmentation and clinical tasks such as radiotherapy planning. Despite plausible visual realism, current methods seek optimal solutions by treating all tissues uniformly, failing to capture intrinsic CT properties such as tissue-specific distributions and strict anatomy, leading to semantic misalignment and structural hallucinations. To address this, we decompose CT synthesis into semantics-structure alignment through the proposed Semantics and Structure Loss. Specifically, the Semantic Intensity Distribution (SID) Loss partitions the broad Hounsfield Unit (HU) range into semantic subspaces to achieve tissue-specific intensity alignment. Building on SID-based semantic alignment, the Structural Anatomy (SA) Loss further improves structural fidelity by regularizing gradient manifolds within clinically-informed spatial domains. Extensive experiments across GAN, Diffusion, flow, and foundation models demonstrate the broad compatibility and consistent gains of our approach. A user study further confirms the clinical realism and mask fidelity, while downstream segmentation demonstrates the practical utility. Code is available at https://github.com/16rq/Semantics-and-Structure-Loss-for-CT-synthesis.

1 Introduction

In recent years, artificial intelligence generated content (AIGC) has driven significant advances in image and video generation [1]. Mainstream generative paradigms, including generative adversarial networks (GANs), diffusion models, and flow-based models, typically optimize global distributions in pixel, latent, or frequency spaces [2, 3, 4]. Computed tomography (CT) synthesis is an important yet highly constrained generation task with applications in data augmentation and radiotherapy planning [5]. Unlike natural images, anatomical regions in CTs follow relatively stable Hounsfield Unit (HU) distributions and exhibit structural patterns determined by imaging physics and human anatomy. Additionally, CTs are quantified in HU, spanning a much wider dynamic range than RGB images, and exhibit window-dependent variations. Therefore, CT synthesis requires not only visual realism or global representation alignment, but also physically meaningful HU distributions and anatomically faithful structures.

However, existing generative models largely inherit training paradigms developed for natural images [6, 7]. Although they can generate visually realistic CTs, they often struggle to model clinically meaningful semantic units, including organ-specific HU distributions, window-dependent appearances, and global anatomical topology. Frequency-domain modeling improves high-frequency detail

*Corresponding authors.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

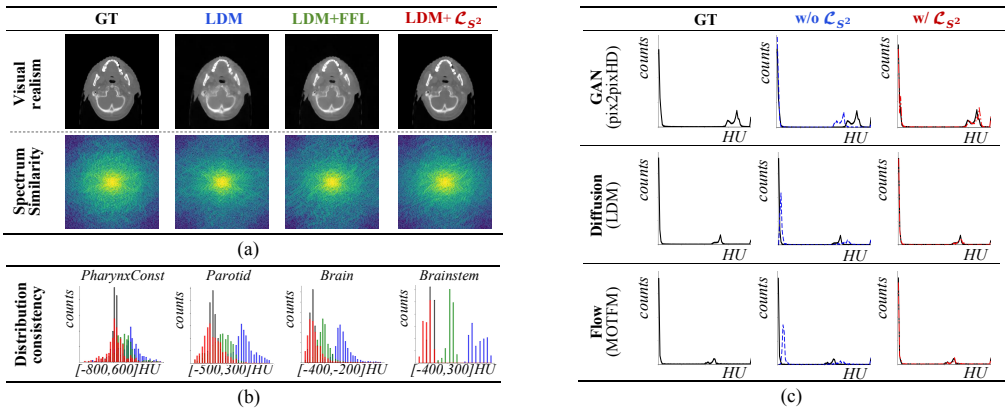

Figure 1: Visual, spectral, and semantic distribution comparison on SegRap. (a) LDM variants exhibit similar visual and spectral patterns. (b) Organ-wise HU histograms reveal distribution gaps between the pixel-modeling baseline (blue), frequency-domain method (green), and our distribution-based method (red), with our method better aligned to real CTs (black). (c) Our method consistently improves semantic distribution across GAN, diffusion, and flow paradigms. All CT images use window setting [-1000, 1000] HU.

compared with pixel- or latent-space modeling, but spectral alignment alone does not explicitly capture the physical principles of CTs [8]. Similarly, recent representation-alignment objectives, including metric learning and alignment regularization, mainly impose consistency on global image-level, cross-modal, or task-specific features, without addressing what should be aligned for clinically reliable CT synthesis [9]. As illustrated in Figure 1(a) and (b), this limitation manifests as semantic distribution mismatch and structural inconsistency, despite visually plausible synthesis results. To address this issue, we depart from conventional global consistency objectives and redefine the object of consistency modeling for CT synthesis.

We propose the semantics and structure loss ($\mathcal{L}_{S^2}$), which explicitly decomposes CT imaging space into clinically grounded semantic subspaces and aligns their region-specific physical and statistical properties. This design directly targets two major failure modes in synthetic CT, namely HU distribution shift that compromises intensity reliability and structural hallucination that weakens anatomical fidelity. $\mathcal{L}_{S^2}$ consists of two complementary components. The semantic intensity distribution loss ($\mathcal{L}_{Sid}$) employs semantic-driven constraints to decouple CT imaging into granular subspaces, facilitating multi-level alignment of heterogeneous tissue properties. To mitigate potential geometric hallucinations, we further introduce the structural anatomy loss ($\mathcal{L}_{Sa}$), which encourages synthesized images to preserve anatomical fidelity. The main contributions of this work are:

- We introduce a CT-specific consistency modeling perspective that shifts the alignment target from global representations to clinically grounded semantic subspaces.
- We propose a unified semantics-structure modeling framework with two complementary components. $\mathcal{L}_{Sid}$ enforces multi-level alignment of organ- and tissue-specific intensity distributions under high dynamic range. $\mathcal{L}_{Sa}$ promotes anatomically faithful structures consistent with clinical priors, while remaining robust to noise and adaptable to the inherent multi-solution nature of generative models.
- An empirical observation suggesting that improved frequency alignment may emerge as a byproduct of our unified semantics–structure modeling.
- Extensive experiments show that $\mathcal{L}_{S^2}$ can be seamlessly integrated into GAN, diffusion, flow, and pretrained foundation models, consistently improving mask-guided CT synthesis.

2 Related work

2.1 Generative modeling for CT imaging

Unconditional generation suffers from limited anatomical accuracy due to the lack of explicit constraints [10], while recent methods improve controllability by incorporating external priors such as text and anatomical maps [11, 12, 13]. Among them, segmentation masks offer precise anatomical guidance [14]. However, current CT synthesis methods, largely inherited from natural image generation, focus on visual realism or global distribution matching while neglecting CT-specific physical properties. This is critical because CT requires not only visual plausibility but also reliable HU distributions and anatomical structures [15]. Existing methods preserve coarse anatomy but suffer from issues like HU shifts, structural distortions, and missing structures not constrained by masks [16]. To address this, we propose shifting CT synthesis towards CT-specific semantic-structure modeling, aligning HU distributions and enforcing anatomical fidelity.

2.2 Segmentation-guided medical image synthesis

Mask-guided medical image synthesis shows great potential for rare disease augmentation and segmentation enhancement across modalities [14, 17, 18]. Recent advances are largely driven by diffusion models, leveraging improved sampling and conditional embeddings [17, 18, 19]. However, these approaches are model-specific, data-intensive, and computationally demanding. Therefore, a unified loss applicable across diverse generative paradigms offers a general and practical solution to address HU distribution mismatch and structural hallucination in existing CT synthesis models.

2.3 Frequency analysis in neural networks

Spectral bias [3] and F-Principle [4] indicate that neural networks prioritize learning low-frequency components, while the data manifold's structure nontrivially influences the learnability of high frequencies. This has motivated researchers to leverage the relationship between spatial and frequency domains, introducing techniques like adaptive frequency recalibration to enhance spatial analysis in tasks including camouflaged object detection, segmentation, and generative modeling with high-frequency constraints [20, 21, 22]. In CT synthesis, low-frequency textures and high-frequency anatomical details emphasize the importance of full-spectrum learning [21]. We clarify why existing frequency-domain fitting is insufficient for CT generation by contrasting CTs with natural images, decomposing CT space distributionally, and framing frequency fitting as a product of semantic-structural modeling.

3 Method

We redefine the generative modeling of CTs and propose semantic-structure learning (Section 3.1 and 3.2), which jointly improve CT synthesis. Furthermore, an empirical frequency-domain analysis is provided in Section 3.3 and its integration into different generative frameworks in Section 3.4.

3.1 Semantic intensity distribution alignment

Clinically, radiologists adjust window settings to discern subtle intensity variations within the vast HU range. This practice highlights that diagnostically critical information often lies in tissue-specific distributions that remain hidden under global observation. Consequently, global representative alignment fails to capture the fine-grained physical properties of heterogeneous tissues. To address this, we bridge clinical expertise with distribution modeling by proposing $\mathcal{L}_{Sid}$.

Formally, we define a set of masks $M = M^g \cup M^l$, where $M^g = \{M^g_w\}_{w=1}^{W}$ represents global clinical windows (e.g., bone, lung) and $M^l = \{M^l_k\}_{k=1}^{K}$ denotes local anatomical regions, as shown in Figure 2. These binary masks $M_i \in \{0,1\}^{H\times W}$ are derived from standardized clinical window parameters and expert-annotated labels, where w and k index clinical window and anatomical class.

Unlike conventional modeling in pixel or frequency domain, $\mathcal{L}_{Sid}$ anchors synthesized HU distributions to clinically relevant semantic manifolds. By minimizing the Wasserstein-1 distance [23]

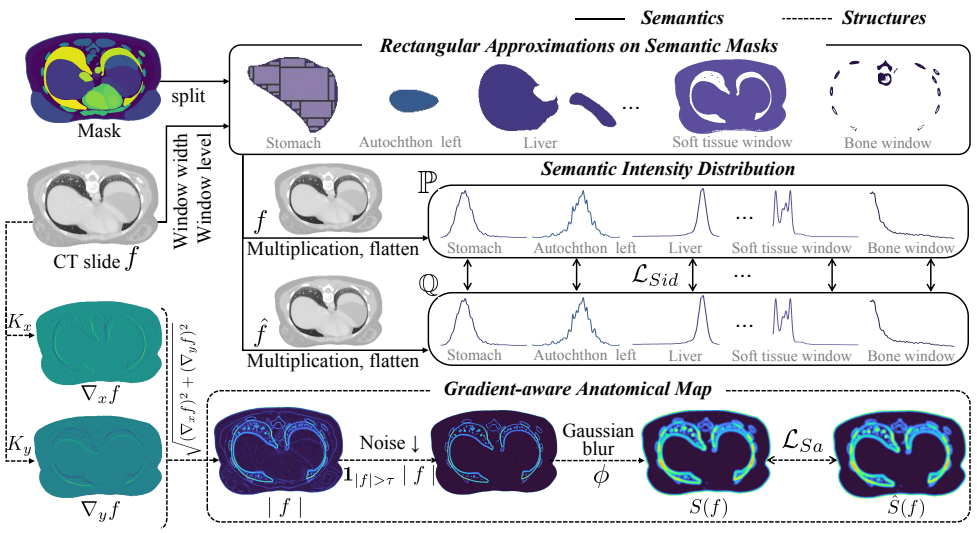

Figure 2: Illustration of the semantic and structure loss which integrates a semantic intensity distribution loss (solid lines) and a structural anatomy loss (dashed lines) to jointly enhance intensity consistency and structural fidelity.

between real and synthesized distributions, it corrects intensity drift, restores window-specific contrast, and suppresses over-smoothed intensity distributions.

Given a real CT image f and its synthesized counterpart $\hat{f}$, the empirical intensity distributions for the i-th semantic subspace are:

$$\mathbb{P}_i = \{f(x,y) \mid M_i(x,y) = 1\}, \quad \mathbb{Q}_i = \left\{\hat{f}(x,y) \mid M_i(x,y) = 1\right\} \tag{1}$$

Here, $M_i \in M^g \cup M^l$ is the i-th semantic mask. $\mathcal{L}_{Sid}$ is defined as the average Wasserstein-1 distance between the real and generated distributions across all semantic subspaces:

$$\mathcal{L}_{Sid} := \frac{1}{W}\sum_{w=1}^{W} W_1(\mathbb{P}_w^g, \mathbb{Q}_w^g) + \frac{1}{K}\sum_{k=1}^{K} W_1(\mathbb{P}_k^l, \mathbb{Q}_k^l) \tag{2}$$

The Wasserstein-1 distance is computed as

$$W_1(\mathbb{P}, \mathbb{Q}) = \inf_{\gamma \in \Pi(\mathbb{P},\mathbb{Q})} \mathbb{E}_{(a,b)\sim\gamma}\left[|a-b|\right], \tag{3}$$

where a and b denote paired HU values sampled from $\mathbb{P}$ and $\mathbb{Q}$, respectively, and $\Pi(\mathbb{P}, \mathbb{Q})$ denotes the set of all joint distributions with marginals $\mathbb{P}$ and $\mathbb{Q}$.

3.2 Structural anatomy alignment

While $\mathcal{L}_{Sid}$ ensures physical accuracy in the intensity domain, it lacks explicit constraints on spatial topology and geometric continuity. Although semantic masks provide macro-structural guidance for major organs, this information is often incomplete; for instance, critical structures like bone cortices frequently lack explicit labels. Consequently, the synthesis remains an ill-posed inverse problem, where the one-to-many mapping leads to significant structural hallucinations, such as distorted organ boundaries or the omission of unmasked critical structures.

To bridge the gap between coarse, partial semantic guidance and high-fidelity anatomical realism, we propose $\mathcal{L}_{Sa}$, as shown in Figure 2. By leveraging intensity gradients, $\mathcal{L}_{Sa}$ provides essential supervision for both labeled organs and unlabeled but clinically vital structures. While standard gradient operators such as Sobel filters are intuitively suited for this task, raw gradients in CT imaging are notoriously sensitive to noise and high-frequency textural fluctuations. To provide a stable and anatomically meaningful structural anchor, we introduce the gradient-based anatomical attention mechanism to adapt classical gradient operators to the radiological domain.

Formally, we first compute the gradient magnitude $|\nabla f| = \sqrt{(\nabla_x f)^2 + (\nabla_y f)^2}$ of a CT image f using Sobel kernels. To suppress non-anatomical noise, we gate the gradient magnitude with a threshold τ, where $\tau = 1 \times 10^{-4}$ is adopted as a robust default. Instead of enforcing a rigid and pixel-aligned penalty which might conflict with the stochastic nature of generative models, we apply Gaussian blurring ϕ and normalization $\mathcal{N}$ to derive a spatially relaxed structural representation, $\mathcal{S}(f)$:

$$\mathcal{S}(f) = \mathcal{N}(\phi(\mathbf{1}_{|\nabla f|>\tau}|\nabla f|)) \tag{4}$$

By transforming sparse gradients into a continuous field, $\mathcal{S}(f)$ refines the search space within each mask while simultaneously acting as an implicit prior for unmasked regions. This ensures that while the model follows the available semantic guidance, it also recovers the fine-grained anatomical logic of structures that are not explicitly labeled. The structural modeling object is computed by Mean Squared Error (MSE) and Structural Similarity (SSIM):

$$\mathcal{L}_{Sa} := \lambda_m \|\mathcal{S} - \hat{\mathcal{S}}\|_2^2 + \lambda_s(1 - \mathrm{SSIM}(\mathcal{S}, \hat{\mathcal{S}})) \tag{5}$$

where $\hat{\mathcal{S}} = \mathcal{S}(\hat{f})$ and $\mathcal{S} = \mathcal{S}(f)$ for brevity. In our implementation, we set $\lambda_m = 10$, $\lambda_s = 0.1$.

The sensitivity analysis of parameters is provided in Appendix A. Overall, the semantic-structure modeling could be integrated into training paradigm via $\mathcal{L}_{S^2}$, which could be formulated as follows:

$$\mathcal{L}_{S^2} = \mathcal{L}_{Sid} + \mathcal{L}_{Sa}. \tag{6}$$

3.3 Empirical frequency-domain analysis

Empirical observations reveal that our method achieves superior performance in both perceptual quality and spectral fidelity, even surpassing generative models that incorporate explicit frequency-domain optimization. To better understand this phenomenon, we investigate the frequency-domain behavior of our semantic-structure modeling framework.

3.3.1 Low-frequency consistency via distribution anchoring

The low-frequency components correspond to homogeneous tissue regions. The proposed $\mathcal{L}_{Sid}$ aligns marginal HU distributions within anatomical masks but does not explicitly constrain spatial frequencies due to its invariance to spatial permutations. Nevertheless, anatomical regions often exhibit concentrated intensity distributions and smooth spatial variations. To investigate this association, we compare SID with global and organ-wise differentiable histogram losses [24] on AutoPET under identical settings (see Appendix B). Organ-wise histogram alignment reduces low-frequency error compared with global alignment, while SID achieves a further reduction with only modest high-frequency improvement. These findings suggest that semantic-region-aware intensity regularization can empirically improve low-frequency consistency without explicit frequency-domain constraints.

3.3.2 Spectral analysis of topological constraints

High-frequency correspond to sharp anatomy. We analyze the efficacy of $\mathcal{L}_{Sa}$ through the spectral response of Sobel operators K_x and K_y [25], which function as directional high-pass filters. The discrete Fourier transform (DFT) [26] could be formulated using Euler's formula $e^{i\theta} = \cos(\theta) + i\sin(\theta)$:

$$\begin{aligned} G_x[u,v] &= \sum_{x=-1}^{1}\sum_{y=-1}^{1} K_x[x,y]e^{-i2\pi\left(\frac{ux}{W}+\frac{vy}{H}\right)} = 4i \cdot \sin(2\pi\frac{u}{W}) \cdot (\cos(2\pi\frac{v}{H}) + 1) \\ G_y[u,v] &= \sum_{x=-1}^{1}\sum_{y=-1}^{1} K_y[x,y]e^{-i2\pi\left(\frac{ux}{W}+\frac{vy}{H}\right)} = 4i \cdot \sin(2\pi\frac{v}{W}) \cdot (\cos(2\pi\frac{u}{H}) + 1) \end{aligned} \tag{7}$$

The Sobel responses characterize sensitivity to spatial intensity transitions rather than directly enforcing topological invariants. By combining gradient thresholding and spatial relaxation, $\mathcal{L}_{Sa}$ encourages anatomically meaningful structural consistency while reducing sensitivity to local noise.

Our empirical results suggest that the joint regularization of intensity distributions and anatomical structures through $\mathcal{L}_{S^2}$ is associated with improved spectral alignment, even without explicit frequency-domain supervision (see Section 4.4).

3.4 Integration into different generative frameworks

The proposed $\mathcal{L}_{S^2}$ operates in image space and can be integrated into different generative frameworks through differentiable clean-image estimates. For GANs, the synthesized CT image $\hat{f}$ is directly obtained from the generator. For noise-prediction LDM, the clean-image estimate at timestep t is reconstructed as

$$\hat{f}_t = D\left(\frac{z_t - \sqrt{1-\bar{\alpha}_t}\,\epsilon_\theta(z_t, t, M)}{\sqrt{\bar{\alpha}_t}}\right), \tag{8}$$

where z_t is the noisy latent, ϵ_θ denotes the predicted noise, and D is the frozen VAE decoder. The complete training objective is

$$\mathcal{L}_{LDM+\mathcal{L}_{S^2}} = \mathbb{E}_{f,M,t,\epsilon}\left[\|\epsilon - \epsilon_\theta(z_t, t, M)\|_2^2 + \mathcal{L}_{S^2}^{(t)}\right]. \tag{9}$$

For MOTFM, using a linear velocity-prediction flow, the clean-image estimate is

$$\hat{f}_t = f_t - t\, v_\theta(f_t, t \mid M). \tag{10}$$

This formulation is validated for the specific linear flow parameterization used in MOTFM; other parameterizations require a corresponding differentiable clean-image estimator. Detailed derivations and implementation limitations are provided in Appendix C.

4 Experiments

4.1 Baselines, datasets and evaluation metrics

We evaluate $\mathcal{L}_{S^2}$ in both 2D and 3D CT synthesis settings. Our main evaluation uses 2D slice-based frameworks for consistent comparison, ablation, downstream segmentation, and user study. Specifically, we integrate $\mathcal{L}_{S^2}$ into representative GAN, diffusion, and flow-based models [27, 28, 29]. We compare it with pixel- and latent-space methods, including img2img-turbo [30], SegGuidedDiff [17], OASIS [31], and MOTFM, as well as frequency-domain losses, including FFL [22] and FDL [32]. To verify its scalability to volumetric generation, we further integrate $\mathcal{L}_{S^2}$ into the 3D CT foundation model MAISI-v2 [7]. The implementation details are presented in Appendix D.

Two datasets with different scales and anatomical coverage are used. SegRap [33] contains 120 patients and 15,312 head-and-neck CT slices with 45 organ-at-risk annotations for nasopharyngeal carcinoma, supplemented with body contours from AccuContour [34]. AutoPET [35] includes 533 patients and 197,304 whole-body CT slices with 144 anatomical structures, providing anatomical context for CT synthesis. Both datasets are split at the patient level into training, validation, and test sets, with 90/20/10 patients for SegRap and 370/108/55 for AutoPET, ensuring no patient overlap.

For evaluation, FID [36] measures realism, while PSNR and SSIM [37] assess intensity consistency and structural similarity. LFD [22] evaluates frequency discrepancy, and HU-W1 measures semantic HU distribution fidelity using Wasserstein-1 distance. We also conduct a blinded user study with one radiologist (13 years of experience) evaluating 100 randomly selected masks paired with real and synthesized CTs (200 image-mask pairs), presented in randomized order. Each pair is independently rated under two criteria: (1) Realism: categorized as unrealistic, uncertain, or realistic; (2) Anatomical faithfulness: scored from 0 to 2, with higher scores indicating better fidelity.

The Wilcoxon signed-rank test [38] is used to assess whether synthesized CTs achieve comparable anatomical fidelity to real CTs. Finally, we evaluate the synthetic-to-real transferability of generated CTs through downstream segmentation.

4.2 Evaluation across 2D generative framework

As shown in Figure 3, $\mathcal{L}_{S^2}$ consistently improves anatomical fidelity and realism across different generative frameworks. In **pix2pixHD**, $\mathcal{L}_{S^2}$ sharpens tissue boundaries and restores key anatomical structures, including clearer bone-air interfaces (row 1), sharper trachea-mediastinum separation (row 2), and continuous intestinal structures (row 3). In **LDM**, $\mathcal{L}_{S^2}$ improves intensity consistency and tissue contrast, correcting color and bone-intensity artifacts (row 1), smoothing mediastinal soft-tissue transitions (row 2), and producing complete intestinal walls (row 3). In **MOTFM**, $\mathcal{L}_{S^2}$ enhances

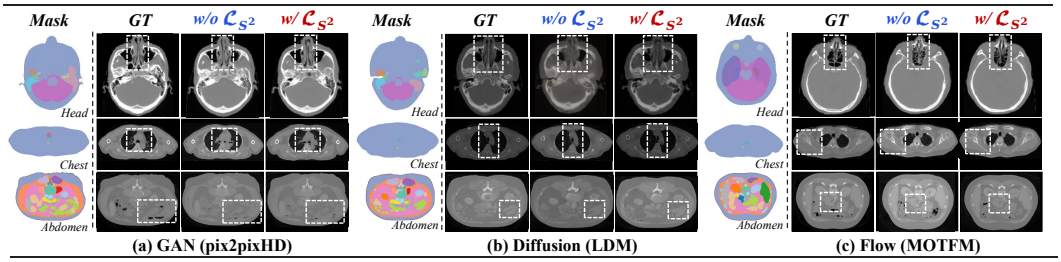

Figure 3: Visual comparison on SegRap (head, chest) and AutoPET (abdomen) showing the effectiveness of $\mathcal{L}_{S^2}$ on different paradigms. All images share the same window setting [-1000, 1000] HU.

Table 1: Quantitative evaluation of $\mathcal{L}_{S^2}$ across 2D and 3D CT synthesis frameworks on SegRap and AutoPET. The table compares representative generative models in pixel, latent, and frequency domains, evaluates the integration of $\mathcal{L}_{S^2}$ into GAN, diffusion, flow, and 3D foundation-model frameworks, and includes ablation studies based on pix2pixHD.

Method	Venue	Dim.	SegRap [33]					AutoPET [35]				
			PSNR↑	SSIM↑	FID↓	LFD↓	HU-W1↓	PSNR↑	SSIM↑	FID↓	LFD↓	HU-W1↓
pix2pixHD [27]	CVPR'18	2D	24.869	0.811	37.623	18.139	120.127	23.984	0.719	26.648	18.505	188.706
FFL [22]	ICCV'21	2D	24.188	0.791	40.767	18.060	114.115	24.134	0.740	24.260	18.522	93.232
OASIS [31]	IJCV'22	2D	21.072	0.753	53.629	17.683	181.954	19.702	0.705	37.841	18.983	31.497
LDM [28]	CVPR'22	2D	28.809	0.721	75.542	15.511	36.226	23.247	0.760	20.187	18.828	218.054
img2img-turbo [30]	arXiv'22	2D	20.260	0.728	80.789	17.781	198.448	19.467	0.714	34.665	18.947	32.759
SegGuidedDiff [17]	MICCAI'24	2D	21.685	0.284	60.322	18.623	293.397	11.217	0.300	73.991	19.001	200.001
FDL [32]	CVPR'24	2D	23.521	0.773	60.611	18.065	111.250	22.713	0.713	28.072	18.621	201.530
MOTFM [29]	MICCAI'25	2D	23.820	0.619	37.492	23.820	32.402	23.151	0.633	27.596	23.151	32.036
MAISI-v2 [7]	AAAI'26	3D	15.697	0.248	63.890	28.461	15.740	16.924	0.795	22.657	27.421	7.135
pix2pixHD+$\mathcal{L}_{Sid}$	This work	2D	25.017	0.815	35.553	16.413	110.231	24.094	0.736	23.986	16.625	162.303
pix2pixHD+$\mathcal{L}_{Sd}$	This work	2D	24.982	0.814	33.186	16.475	115.286	24.138	0.743	24.652	16.615	171.552
pix2pixHD+$\mathcal{L}_{S^2}$	This work	2D	25.068	0.816	33.932	16.411	107.981	**24.177**	0.752	23.889	**16.606**	159.242
MOTFM+$\mathcal{L}_{S^2}$	This work	2D	24.214	0.678	**32.535**	16.630	32.036	23.428	0.649	24.235	16.779	30.119
LDM+$\mathcal{L}_{S^2}$	This work	2D	**32.175**	**0.834**	51.496	**14.732**	16.204	23.515	0.771	**18.131**	18.816	216.000
MAISI-v2+$\mathcal{L}_{S^2}$	This work	3D	15.998	0.269	56.213	27.461	**13.737**	17.130	**0.810**	20.591	27.200	**6.803**

anatomical distinctiveness by recovering fine sinus and trabecular details (row 1), separating fused clavicle-rib regions (row 2), and delineating intestinal and muscular boundaries (row 3).

From the semantic distribution perspective in Figure 1(c), the baseline models (blue) show varying degrees of HU misalignment relative to real CT (black). **Pix2pixHD** exhibits shifted high-HU peaks, **LDM** shows discrepancies in both low- and high-HU density counts, and **MOTFM** retains subtle bias in low- and mid-range tissues. With $\mathcal{L}_{S^2}$, these deviations are consistently reduced, yielding distributions that better overlap with real CT across the HU range (red).

As shown in Table 1, in **pix2pixHD**, $\mathcal{L}_{S^2}$ improves PSNR and SSIM while reducing FID, LFD, and HU-W1, with AutoPET SSIM increasing to 0.752 (+3.3%), indicating better structural alignment. For **LDM**, $\mathcal{L}_{S^2}$ yields notable gains on the smaller SegRap dataset, improving PSNR to 32.175 (+3.366) and SSIM to 0.834 (+11.3%), while reducing FID to 51.496 (-24.046), LFD to 14.732 (-0.799), and HU-W1 from 36.226 to 16.204. Performance on AutoPET also improves accordingly. For **MOTFM**, $\mathcal{L}_{S^2}$ consistently improves PSNR/SSIM and reduces FID, LFD, and HU-W1 across both datasets, demonstrating its effectiveness in enhancing flow-based CT synthesis and distribution consistency.

4.3 Scalability to 3D CT foundation models

To examine whether $\mathcal{L}_{S^2}$ generalizes beyond 2D synthesis, we integrate it into the 3D CT foundation model MAISI-v2 and evaluate it on SegRap and AutoPET. As shown in Table 1, MAISI-v2 shows relatively limited absolute performance on head-and-neck CT synthesis, likely due to the domain and anatomical mismatch between its pretrained generative prior and the fine-grained structures in SegRap. Nevertheless, adding $\mathcal{L}_{S^2}$ consistently improves performance across both datasets, increasing PSNR and SSIM while reducing LFD and HU-W1, indicating enhanced intensity fidelity, structural consistency, spectral alignment, and semantic HU distribution. These results suggest that

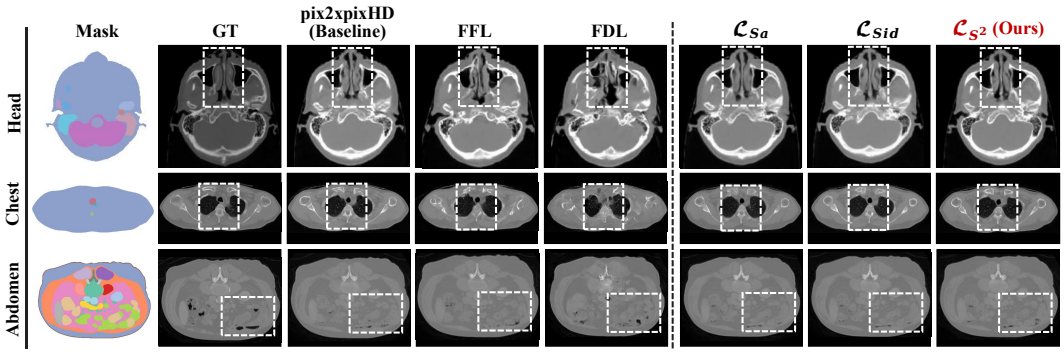

Figure 4: Comparison with frequency-modeling loss and ablation studies on SegRap (head and chest) and AutoPET (abdomen) through pix2pixHD. All images use the window setting [-1000, 1000] HU.

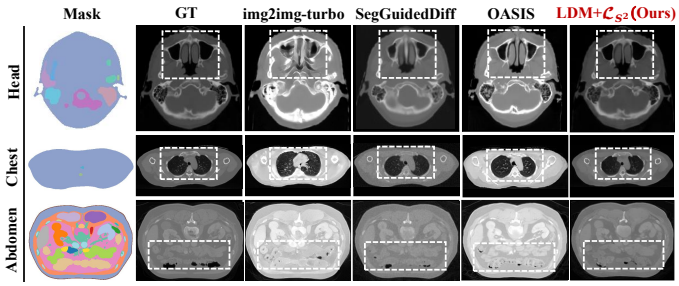

Figure 5: Comparison with pixel-modeling generators on SegRap (head, chest) and AutoPET (abdomen) through LDM. All images use the window setting $[-1000, 1000]$ HU.

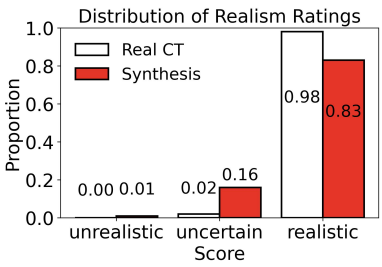

Figure 6: Results of user study.

$\mathcal{L}_{S^2}$ can serve as a plug-in fine-tuning objective for 3D CT foundation models, improving their adaptation to unseen CT datasets. Additional visual comparisons, slice-level and patient-level 3D evaluations, and cross-framework relative improvements are provided in Appendix E.

4.4 Comparison with relevant loss using frequency modeling

As shown in Figure 4, existing frequency-relevant losses (FFL and FDL) fail to address distribution shift and structural distortions, and may even produce anatomically implausible results. In contrast, $\mathcal{L}_{S^2}$ significantly improves structural fidelity and radiological realism across different anatomical regions. In the head region, the nasal and sinus structures are more clearly reconstructed (row 1). Additionally, the clavicle forms a complete, undistorted shape without explicit anatomical guidance, and the axial skeleton is more realistically restored. Moreover, the intestines exhibit natural segmentation and continuity, aligning well with anatomical boundaries indicated by mask condition.

After introducing $\mathcal{L}_{S^2}$, all evaluation metrics are improved on SegRap and AutoPET (see Table 1). Moreover, unlike FFL and FDL, which show instability across datasets, $\mathcal{L}_{S^2}$ provides robust and scalable improvements for CT synthesis at different dataset scales. Experiments on other generative paradigms are included in Appendix F.

4.5 Comparison with generators using pixel modeling

As shown in Figure 5, compared with pixel-modeling generators, LDM+$\mathcal{L}_{S^2}$ achieves the best anatomical accuracy and intensity realism. Specifically, LDM+$\mathcal{L}_{S^2}$ accurately reconstructs the nasal cavity and sinus structures, avoiding the over-smoothing and intensity collapse observed in other methods (row 1). In the chest region, ours generates anatomically correct ribs, clavicles, and the external body without explicit anatomical guidance, producing more realistic and consistent intensity distributions (row 2). Moreover, intestinal structures exhibit clear segmentation and well-defined boundaries, with more reasonable intensity contrasts among surrounding organs (row 3).

Table 2: Downstream segmentation performance using real data, LDM-generated data, and LDM generated data with our loss.

Data source	Overall				Organ-level Dice↑		
	mean Dice↑	mean IoU↑	mean ASSD↓	mean HD95↓	Submandibular_L	Brain	SpinalCord
True	0.9101 ± 0.0832	0.9057 ± 0.0279	0.2952 ± 0.2576	0.4959 ± 0.2821	0.9271 ± 0.0636	0.7141 ± 0.0438	0.5487 ± 0.0456
LDM	0.9565 ± 0.0254	0.9452 ± 0.0192	0.2173 ± 0.1178	0.3615 ± 0.1579	0.9363 ± 0.0404	0.9512 ± 0.0163	0.8086 ± 0.0238
LDM+$\mathcal{L}_{S^2}$	$\mathbf{0.9613 \pm 0.0214}$	$\mathbf{0.9497 \pm 0.0144}$	$\mathbf{0.1994 \pm 0.0827}$	$\mathbf{0.3287 \pm 0.0936}$	$\mathbf{0.9513 \pm 0.0271}$	$\mathbf{0.9681 \pm 0.0115}$	$\mathbf{0.8454 \pm 0.0198}$

Quantitative results in Table 1 also validate the superior performance of LDM+$\mathcal{L}_{S^2}$. On SegRap, it yields the highest PSNR (32.175, +10.490) and SSIM (0.834, +8.1%) with substantially lower FID and HU-W1 (16.204, -16.198), showing enhanced spatial and distribution fidelity. While SegGuidedDiff suffers from serious noise (SSIM 0.300 on AutoPET), LDM+$\mathcal{L}_{S^2}$ preserves both anatomical accuracy and perceptual realism, achieving the best overall quantitative results.

4.6 Ablation studies

As shown in Figure 4, with the addition of $\mathcal{L}_{Sa}$, anatomically realistic nasal septum structures compared to the baseline are generated. Specifically, more complete nasal cavities with clearer air–bone boundaries are generated (row 1). The introduction of $\mathcal{L}_{Sid}$ stabilizes the overall intensity distribution, preventing intensity blur or contrast inconsistencies in regions such as abdominal soft tissues (row 3). Under their constraints, models achieve intensity consistency and boundary alignment, producing CT with superior anatomical fidelity and radiological realism.

In addition, Table 1 confirms that collectively integrating $\mathcal{L}_{Sid}$ and $\mathcal{L}_{Sa}$ yields the best performance. $\mathcal{L}_{Sid}$ enhances semantic intensity fidelity through distribution constraints, increasing PSNR and decreasing FID on both datasets. Notably, after introducing $\mathcal{L}_{Sid}$, HU-W1 decreases from 120.127 to 110.231 on SegRap, and from 188.706 to 162.303 on AutoPET, which validates the effectiveness of $\mathcal{L}_{Sid}$ in semantic distribution alignment. $\mathcal{L}_{Sa}$ improves structural alignment, raising SSIM from 0.719 to 0.736 and surpassing the structural limit of pix2pixHD on AutoPET. Additional ablations are provided in Appendix G, while conditional diversity and mask robustness are further evaluated in Appendix H.

4.7 User study

As shown in Figure 6, the radiologist classified most real images as realistic, with only 2% rated as uncertain, while 16% of synthesized images were rated as suspicious. Notably, 83% of synthesized images were classified as realistic, suggesting encouraging perceptual realism under this single-reader evaluation. Anatomical fidelity scores were comparable between real and synthesized images (1.930 ± 0.256 vs. 1.980 ± 0.141). The Wilcoxon signed-rank test detected no statistically significant difference ($p = 0.0588$). As only one radiologist participated, inter-rater agreement could not be assessed. Further details are provided in Appendix I, while mask fidelity is additionally evaluated using a pre-trained segmentation model (Appendix J).

4.8 Improvement in downstream segmentation

We conduct a synthetic-to-real transfer and report 2D results on SegRap using the same split as the synthesis experiments. Segmentation models are trained and validated separately on real CTs, LDM-generated CTs, or LDM+$\mathcal{L}_{S^2}$-generated CTs, paired with the corresponding anatomical masks. Dice and IoU are computed independently for each organ on each 2D slice and macro-averaged within and across organs.

As shown in Table 2, LDM+$\mathcal{L}_{S^2}$ achieves the best downstream segmentation performance. Compared with LDM, LDM+$\mathcal{L}_{S^2}$ increases mean Dice from 0.9565 ± 0.0254 to 0.9613 ± 0.0214 and mean IoU from 0.9452 ± 0.0192 to 0.9497 ± 0.0144, while reducing mean HD95 and ASSD by 9.07% and 8.24%, respectively. These gains suggest that $\mathcal{L}_{S^2}$ improves not only region overlap but also anatomical boundary accuracy. At the organ level, LDM+$\mathcal{L}_{S^2}$ outperforms LDM in Dice across 36 out of 45 anatomical categories, with the most pronounced gains observed in structures exhibiting weak baseline performance, complex boundaries, or elongated geometries. Representative improvements include SpinalCord (0.8086 ± 0.0238 to 0.8454 ± 0.0198), Brain (0.9512 ± 0.0163 to 0.9681 ± 0.0115), and Submandibular_L (0.9363 ± 0.0404 to 0.9513 ± 0.0271), with additional gains in Larynx,

Mandible_R, and other categories. Small anatomical structures with already high Dice scores show marginal changes, with a maximum drop of 0.32%. Detailed evaluation protocols, organ-wise results, and patient-level 3D evaluations are provided in Appendix K.

5 Conclusion

Different from current generation modeling from a global perspective, we proposed $\mathcal{L}_{S^2}$, an objective that unifies semantics and structure learning in CT synthesis through distribution alignment in multiple subspaces and plausible gradient manifold constraints, which adapts to the physical and anatomical characteristics of CTs. Besides, this objective could be integrated seamlessly into diverse generative frameworks. Extensive experiments across representative generative paradigms (GANs, diffusion, and flow) and datasets of varying scales have shown consistent improvements on semantics and structure. However, mask-guided synthesis relies heavily on annotations, limiting the modeling of unannotated structures. Although $\mathcal{L}_{S^2}$ may be extended to lesion-aware synthesis, our experiments mainly focus on normal anatomical CT generation rather than pathological lesion generation. Future work will explore generation under incomplete anatomical priors and extend the framework to morphologically complex tissues and lesion-aware CT synthesis.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 62371409) and National High Level Hospital Clinical Research Funding (2025-PUMCH-D-002 and BJ-2025-138).

References

[1] Jiacheng Wang, Jing Yang, Qichao Zhou, and Liansheng Wang. Medical boundary diffusion model for skin lesion segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 427–436. Springer, 2023.

[2] Sanuwani Dayarathna, Kh Tohidul Islam, Sergio Uribe, Guang Yang, Munawar Hayat, and Zhaolin Chen. Deep learning based synthesis of mri, ct and pet: Review and analysis. *Medical image analysis*, 92:103046, 2024.

[3] Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. In *International conference on machine learning*, pages 5301–5310. PMLR, 2019.

[4] Zhi-Qin John Xu. Frequency principle: Fourier analysis sheds light on deep neural networks. *Communications in Computational Physics*, 28(5):1746–1767, 2020.

[5] Gleb Kumichev, Pavel Blinov, Yulia Kuzkina, Vasily Goncharov, Galina Zubkova, Nikolai Zenovkin, Aleksei Goncharov, and Andrey Savchenko. Medsyn: Llm-based synthetic medical text generation framework. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 215–230. Springer, 2024.

[6] Evi MC Huijben, Maarten L Terpstra, Suraj Pai, Adrian Thummerer, Peter Koopmans, Manya Afonso, Maureen Van Eijnatten, Oliver Gurney-Champion, Zeli Chen, Yiwen Zhang, et al. Generating synthetic computed tomography for radiotherapy: Synthrad2023 challenge report. *Medical image analysis*, 97:103276, 2024.

[7] Can Zhao, Pengfei Guo, Dong Yang, Yucheng Tang, Yufan He, Benjamin Simon, Mason Belue, Stephanie Harmon, Baris Turkbey, and Daguang Xu. Maisi-v2: Accelerated 3d high-resolution medical image synthesis with rectified flow and region-specific contrastive loss. *Proceedings of the 40th AAAI Conference on Artificial Intelligence (AAAI 2026)*, 2026.

[8] Zhe Wang, Ziqiu Chi, Yanbing Zhang, et al. Fregan: Exploiting frequency components for training gans under limited data. *Advances in Neural Information Processing Systems*, 35:33387–33399, 2022.

[9] Sihyun Yu, Sangkyung Kwak, Huiwon Jang, Jongheon Jeong, Jonathan Huang, Jinwoo Shin, and Saining Xie. Representation alignment for generation: Training diffusion transformers is easier than you think. In *International Conference on Learning Representations*, 2025.

[10] Ahmad Pesaranghader, Yiping Wang, and Mohammad Havaei. Ct-sgan: computed tomography synthesis gan. In *MICCAI Workshop on Deep Generative Models*, pages 67–79. Springer, 2021.

[11] Ibrahim Ethem Hamamci, Sezgin Er, Anjany Sekuboyina, Enis Simsar, Alperen Tezcan, Ayse Gulnihan Simsek, Sevval Nil Esirgun, Furkan Almas, Irem Doğan, Muhammed Furkan Dasdelen, et al. Generatect: Text-conditional generation of 3d chest ct volumes. In *European Conference on Computer Vision*, pages 126–143. Springer, 2024.

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

[13] Yanwu Xu, Li Sun, Wei Peng, Shuyue Jia, Katelyn Morrison, Adam Perer, Afrooz Zandifar, Shyam Visweswaran, Motahhare Eslami, and Kayhan Batmanghelich. Medsyn: text-guided anatomy-aware synthesis of high-fidelity 3-d ct images. *IEEE Transactions on Medical Imaging*, 43(10):3648–3660, 2024.

[14] Pengfei Guo, Can Zhao, Dong Yang, Ziyue Xu, Vishwesh Nath, Yucheng Tang, Benjamin Simon, Mason Belue, Stephanie Harmon, Baris Turkbey, et al. Maisi: Medical ai for synthetic imaging. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4430–4441. IEEE, 2025.

[15] Haijo Jung. Basic physical principles and clinical applications of computed tomography. *Progress in Medical Physics*, 32(1):1–17, 2021.

[16] 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.

[17] Nicholas Konz, Yuwen Chen, Haoyu Dong, and Maciej A Mazurowski. Anatomically-controllable medical image generation with segmentation-guided diffusion models. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 88–98. Springer, 2024.

[18] Riyu Qiu, Kun Xia, Feng Gao, Shuting Yang, Du Cai, Jiacheng Wang, Yinran Chen, and Liansheng Wang. Accurate boundary alignment and realism enhancement for colonoscopic polyp image-mask pair generation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 34–44. Springer, 2025.

[19] Meilong Xu, Saumya Gupta, Xiaoling Hu, Chen Li, Shahira Abousamra, Dimitris Samaras, Prateek Prasanna, and Chao Chen. Topocellgen: Generating histopathology cell topology with a diffusion model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 20979–20989, 2025.

[20] Yanguang Sun, Chunyan Xu, Jian Yang, Hanyu Xuan, and Lei Luo. Frequency-spatial entanglement learning for camouflaged object detection. In *European Conference on Computer Vision*, pages 343–360. Springer, 2024.

[21] Reza Azad, Afshin Bozorgpour, Maryam Asadi-Aghbolaghi, Dorit Merhof, and Sergio Escalera. Deep frequency re-calibration u-net for medical image segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3274–3283, 2021.

[22] Liming Jiang, Bo Dai, Wayne Wu, and Chen Change Loy. Focal frequency loss for image reconstruction and synthesis. In *ICCV*, 2021.

[23] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017.

[24] Mor Avi-Aharon, Assaf Arbelle, and Tammy Riklin Raviv. Differentiable histogram loss functions for intensity-based image-to-image translation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45:11642–11653, 2023.

[25] Nick Kanopoulos, Nagesh Vasanthavada, and Robert L Baker. Design of an image edge detection filter using the sobel operator. *IEEE Journal of solid-state circuits*, 23(2):358–367, 1988.

[26] Brad G Osgood. *Lectures on Fourier transform and its applications*. SIAM, 2019.

[27] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-resolution image synthesis and semantic manipulation with conditional gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018.

[28] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021.

[29] Milad Yazdani, Yasamin Medghalchi, Pooria Ashrafian, Ilker Hacihaliloglu, and Dena Shahriari. Flow matching for medical image synthesis: Bridging the gap between speed and quality. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 216–226. Springer, 2025.

[30] Gaurav Parmar, Taesung Park, Srinivasa Narasimhan, and Jun-Yan Zhu. One-step image translation with text-to-image models. *arXiv e-prints*, pages arXiv–2403, 2024.

[31] Vadim Sushko, Edgar Schönfeld, Dan Zhang, Juergen Gall, Bernt Schiele, and Anna Khoreva. Oasis: only adversarial supervision for semantic image synthesis. *International Journal of Computer Vision*, 130:2903–2923, 2022.

[32] Zhangkai Ni, Juncheng Wu, Zian Wang, Wenhan Yang, Hanli Wang, and Lin Ma. Misalignment-robust frequency distribution loss for image transformation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2910–2919, 2024.

[33] Xiangde Luo, Jia Fu, Yunxin Zhong, Shuolin Liu, Bing Han, Mehdi Astaraki, Simone Bendazzoli, Iuliana Toma-Dasu, Yiwen Ye, Ziyang Chen, et al. Segrap2023: A benchmark of organs-at-risk and gross tumor volume segmentation for radiotherapy planning of nasopharyngeal carcinoma. *Medical image analysis*, 101:103447, 2025.

[34] W Zhang, Z Chen, Z Liang, Y Hu, and Q Zhou. Acculearning: a user-friendly deep learning auto-segmentation platform for radiotherapy. *International Journal of Radiation Oncology, Biology, Physics*, 111(3):e122, 2021.

[35] 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.

[36] 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.

[37] 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.

[38] SM Taheri and Gholamreza Hesamian. A generalization of the wilcoxon signed-rank test and its applications. *Statistical Papers*, 54(2):457–470, 2013.

[39] 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.

Table 3: Sensitivity analysis of the gradient threshold τ on MOTFM over the SegRap dataset. $\uparrow$ indicates that higher values are better, while $\downarrow$ indicates that lower values are better.

Value	PSNR↑	SSIM↑	LFD↓	FID↓	HU-W1↓
0	21.817	0.632	36.288	21.897	39.431
10^{-2}	22.524	0.659	34.145	18.693	35.394
10^{-3}	23.592	0.677	31.935	16.022	32.285
10^{-4}	24.214	0.678	32.535	16.630	32.036
10^{-5}	24.234	0.679	32.926	16.276	32.531

Table 4: Robustness evaluation of the relative weight ratio λ_m/λ_s on MOTFM over the SegRap dataset.

Scale	SSIM
100	0.678
10	0.682
1	0.680
0.1	0.683
0.01	0.679

A Sensitivity analysis of parameters

The sensitivity analysis of each parameter is implemented on MOTFM over SegRap.

Gradient threshold τ. We use τ of 0, 1e-2, 1e-3, 1e-4, and 1e-5. As shown in Table 3, quality is stable from 1e-3 to 1e-5, but degrades at τ=0 and τ=1e-2, especially at tau=0 where loss overreacts to noise. Based on this, we chose τ1e-4 as a robust default to suppress noise.

Relative weight ratio λ_m/λ_s in $\mathcal{L}_{Sa}$. We test λ_m/λ_s ratio in $\mathcal{L}_{Sa}$ from 0.01 to 100. As shown in Table 4, SSIM remains stable (0.678–0.683). This confirms the method is robust to weight choices and requires no careful tuning. Finally, we used a ratio of 100 to balance optimization scales.

Relative weight ratio $\mathcal{L}_{Sa}/\mathcal{L}_{Sid}$ in $\mathcal{L}_{S^2}$. We explore relative weight scales from 0.1 to 10 of $\mathcal{L}_{Sa}$ with respect to $\mathcal{L}_{Sid}$. As shown in Table 5, relevant metrics remain stable across this range, indicating that the method is robust to the weight choice and does not require careful tuning. We thus choose ratio with 1 for simplicity. Methodologically, the two losses address complementary rather than redundant failure modes, whose theoretical evidence is discussed in Section 3.3.

B Frequency-domain evaluation and analysis

Frequency-domain evaluation metrics. To assess spectral consistency between real and synthesized CT images, we compute low-frequency (LF) and high-frequency (HF) errors using 2D Fourier transform. For each image I_k, where $k \in \{\text{real}, \text{syn}\}$, the centered log-magnitude spectrum is

$$S_k = \log\left(1 + |\text{fftshift}(\text{FFT2}(I_k))|\right). \tag{11}$$

The absolute spectral discrepancy is $E(u,v) = |S_{\text{real}}(u,v) - S_{\text{syn}}(u,v)|$. Using the normalized radial frequency $\rho(u,v)$, defined as the distance from the spectral center divided by the maximum radius, we define

$$\Omega_{\text{LF}} = \{\rho < 0.25\}, \quad \Omega_{\text{HF}} = \{\rho > 0.50\}. \tag{12}$$

The corresponding errors are

$$e_b = \frac{1}{|\Omega_b|} \sum_{(u,v)\in\Omega_b} E(u,v), \quad b \in \{\text{LF}, \text{HF}\}. \tag{13}$$

Image intensities are normalized to $[0,1]$ without additional FFT magnitude normalization. Frequency thresholds are relative to the discrete Fourier grid.

Table 5: Effect of different relative weight scales of the structural anatomy loss with respect to the distribution loss on MOTFM over the SegRap dataset.

Scale	SSIM↑	HU-W1↓
10	0.671	31.896
1	0.678	32.036
0.1	0.672	31.981

Table 6: Comparison of $\mathcal{L}_{Sid}$ with differentiable histogram losses on AutoPET. Lower HU-W1 and frequency errors indicate better consistency.

Method	SSIM ↑	HU-W1 ↓	LF error ↓	HF error ↓
pix2pixHD	0.719 ± 0.057	188.706 ± 4.960	0.491 ± 0.061	0.418 ± 0.043
Histogram (global)	0.714 ± 0.055	170.190 ± 14.672	0.480 ± 0.062	0.418 ± 0.033
Histogram (organ)	0.710 ± 0.053	167.553 ± 15.672	0.451 ± 0.058	0.416 ± 0.042
$\mathcal{L}_{Sid}$	0.736 ± 0.051	162.303 ± 21.320	0.391 ± 0.051	0.408 ± 0.037
$\mathcal{L}_{Sa}$	0.743 ± 0.052	171.552 ± 4.596	0.482 ± 0.060	0.295 ± 0.013
$\mathcal{L}_{S^2}$	$\mathbf{0.752 \pm 0.048}$	$\mathbf{159.242 \pm 4.668}$	$\mathbf{0.389 \pm 0.047}$	$\mathbf{0.287 \pm 0.022}$

Comparison with differentiable histogram matching. To investigate the relationship between organ-aware intensity distribution alignment and spectral consistency, we compare the proposed semantic intensity distribution loss ($\mathcal{L}_{Sid}$) with the differentiable histogram loss [24]. The histogram loss is applied either globally or independently within anatomical masks, using the same pix2pixHD backbone and training settings on AutoPET. We evaluate SSIM, HU-W1, and LF/HF spectral errors. As shown in Table 6, global histogram matching slightly reduces the LF error from 0.491 to 0.480, whereas organ-wise histogram matching further reduces it to 0.451. The proposed $\mathcal{L}_{Sid}$ achieves a lower LF error of 0.391, corresponding to an approximately 20.4% reduction relative to the pix2pixHD baseline. It also improves HU-W1 from 188.706 to 162.303, while the HF error remains relatively close to the baseline (0.408 vs. 0.418). In contrast, $\mathcal{L}_{Sa}$ substantially reduces the HF error from 0.418 to 0.295, with only a minor change in LF error (0.491 to 0.482). Combining both objectives, $\mathcal{L}_{S^2}$ achieves the lowest LF and HF errors of 0.389 and 0.287, respectively, together with the highest SSIM and lowest HU-W1.

C Integration details into different generative frameworks

GAN-based models. For GANs, the synthesized CT image $\hat{f}$ is directly obtained from the generator conditioned on the anatomical mask. The proposed $\mathcal{L}_{S^2}$ is applied between $\hat{f}$ and the corresponding real CT image f as an additional image-space regularization term.

Latent diffusion models. Given a real CT image f and its conditioning mask M, the frozen VAE encoder E maps the image to the latent representation $z_0 = E(f)$. The forward diffusion process is

$$z_t = \sqrt{\bar{\alpha}_t} z_0 + \sqrt{1 - \bar{\alpha}_t}\epsilon, \quad \epsilon \sim \mathcal{N}(0, \mathbf{I}), \tag{14}$$

where $\bar{\alpha}_t$ denotes the cumulative noise schedule. Given the noise prediction $\epsilon_\theta(z_t, t, M)$, the clean-image estimate at timestep t is reconstructed as

$$\hat{f}_t = D\left(\frac{z_t - \sqrt{1 - \bar{\alpha}_t}\epsilon_\theta(z_t, t, M)}{\sqrt{\bar{\alpha}_t}}\right), \tag{15}$$

where D is the frozen VAE decoder. Following the definitions in Section 3.1 and Section 3.2, the image-space regularization is

$$\mathcal{L}_{S^2}^{(t)} = \mathcal{L}_{Sid}(\hat{f}_t, f, M) + \mathcal{L}_{Sa}(\hat{f}_t, f). \tag{16}$$

The complete LDM training objective is

$$\mathcal{L}_{LDM+\mathcal{L}_{S^2}} = \mathbb{E}_{f,M,t,\epsilon}\left[\|\epsilon - \epsilon_\theta(z_t, t, M)\|_2^2 + \lambda \mathcal{L}_{S^2}^{(t)}\right], \quad \lambda = 1. \tag{17}$$

Both VAE encoder and decoder remain frozen, while gradients from $\mathcal{L}_{S^2}$ propagate through the decoder to update the denoising network.

VAE reconstruction fidelity. Since the VAE is frozen, its reconstruction accuracy may limit the fidelity of image-space regularization. The frozen VAE achieves MAE/SSIM/PSNR of $0.0039/0.9824/36.97$ on AutoPET and $0.0028/0.9893/38.52$ on SegRap. The same VAE is used for the baseline LDM and LDM+$\mathcal{L}_{S^2}$, ensuring consistent reconstruction conditions.

Flow matching. For the linear velocity-prediction flow used in MOTFM, the intermediate state is defined as

$$f_t = (1-t)f + t\epsilon, \tag{18}$$

where ϵ denotes the source sample and the target velocity is $v = \epsilon - f$. Given the predicted velocity $v_\theta(f_t, t \mid M)$, the clean-image estimate is

$$\hat{f}_t = f_t - t\, v_\theta(f_t, t \mid M). \tag{19}$$

The proposed $\mathcal{L}_{S^2}$ is then applied between $\hat{f}_t$ and f in image space.

More generally, for an affine probability path $f_t = \alpha_t f + \sigma_t \epsilon$, the clean-image estimate can be expressed as

$$\hat{f}_t = \frac{\dot{\sigma}_t f_t - \sigma_t v_\theta(f_t, t \mid M)}{\alpha_t \dot{\sigma}_t - \dot{\alpha}_t \sigma_t}, \tag{20}$$

provided that $\alpha_t\dot{\sigma}_t - \dot{\alpha}_t\sigma_t \neq 0$. The linear MOTFM formulation corresponds to $\alpha_t = 1 - t$ and $\sigma_t = t$. Our implementation and empirical validation are limited to this linear velocity-prediction formulation; extensions to other flow parameterizations require corresponding differentiable clean-image estimators and further validation.

Timestep stability and limitations. To assess timestep-dependent stability, we divide diffusion timesteps into low-noise $[1, 334)$, medium-noise $[334, 667)$, and high-noise $[667, 1000]$ intervals. Using training checkpoints on SegRap, we evaluate denoising loss, $\mathcal{L}_{S^2}$, total loss, and their gradient norms over non-overlapping 10k-step training intervals. As noise increases, clean-image reconstruction PSNR decreases from 34.321 to 32.982 and 28.125, while $\mathcal{L}_{S^2}$ and its gradients become larger, indicating greater sensitivity at high-noise timesteps. Nevertheless, within the high-noise interval, denoising loss decreases from 0.026030 to 0.009135, and $\mathcal{L}_{S^2}$ decreases from 0.062312 to 0.021928, with finite gradient norms and no observed divergence. Similarly, the overall denoising loss decreases from 0.103973 to 0.025865 for LDM+$\mathcal{L}_{S^2}$, indicating stable optimization under the evaluated training settings. Although no timestep-dependent weighting is adopted to maintain a consistent loss formulation across frameworks, future timestep-adaptive or noise-aware regularization may further improve performance. Clean-image reconstruction also introduces additional training-time computational and memory overhead without affecting inference cost.

D Implementation details

In data preprocessing, CT slices and their corresponding masks were resized to 256 × 256 for 2D experiments and 256 × 256 × 128 for 3D experiments, and a consistent intensity window of [–1000, 1000] HU was applied to all CTs. For fair comparison, all methods were trained using four NVIDIA RTX 3090 GPUs in PyTorch and following their official default configurations. Importantly, our method does not introduce any additional inference overhead, so the inference time is identical to that of all baselines.

E Visual and quantitative evaluation of CT synthesis

Visual demonstration of relative improvements across generative frameworks. To provide a clearer comparison of the benefits of $\mathcal{L}_{S^2}$, we calculate the relative improvements over each corresponding baseline on SegRap and AutoPET. As shown in Figure 7, $\mathcal{L}_{S^2}$ improves all five evaluation metrics across GAN-, diffusion-, flow-, and 3D foundation-model-based frameworks, although the magnitude varies across models and datasets. Notably, LDM exhibits pronounced improvements on SegRap, particularly in HU-W1 (55.3%), FID (31.8%), and SSIM (15.7%). MOTFM achieves substantial LFD reductions on SegRap (30.2%) and AutoPET (27.5%). These findings further illustrate the cross-framework applicability of the proposed regularization.

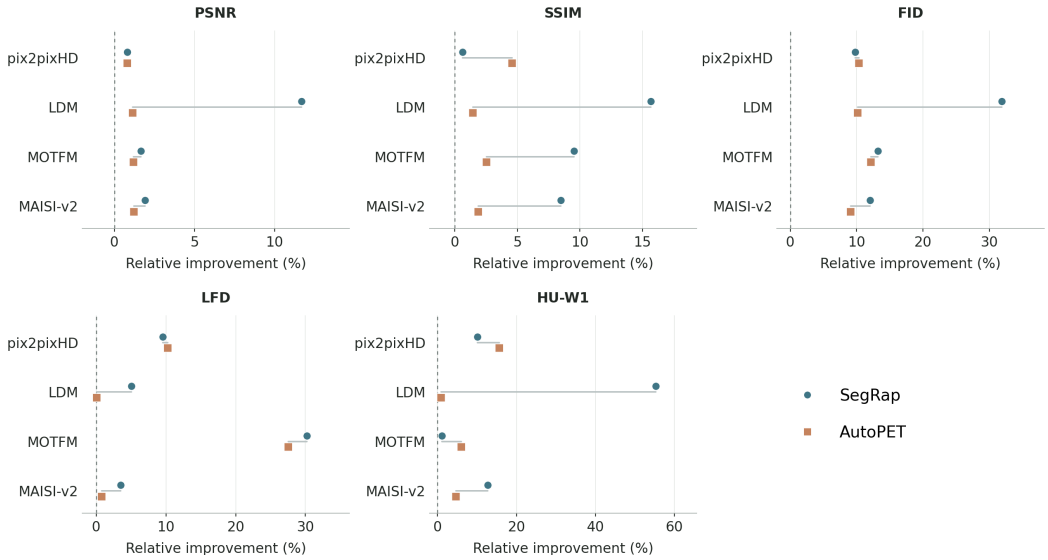

Figure 7: Relative improvements from incorporating $\mathcal{L}_{S^2}$ across different generative frameworks on SegRap and AutoPET. Each point represents the percentage improvement over the corresponding baseline. Improvements are defined as relative increases for PSNR and SSIM, and relative decreases for FID, LFD, and HU-W1. Positive values indicate better performance.

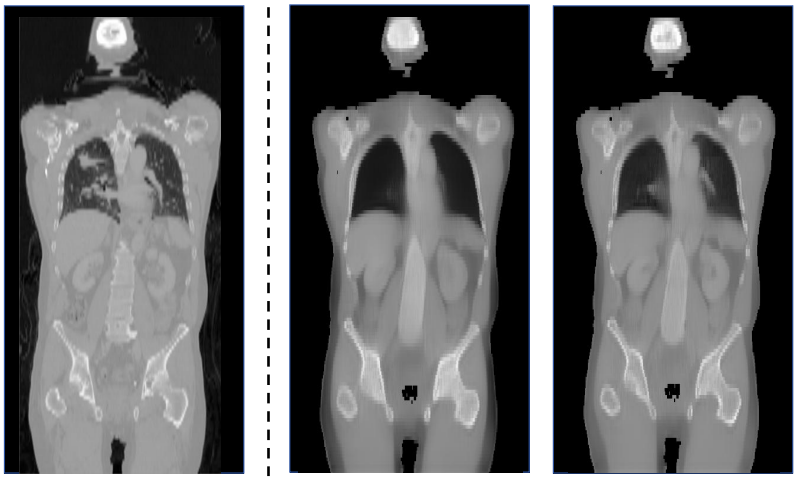

Figure 8: Qualitative comparison of 3D CT synthesis using MAISI-v2 with and without $\mathcal{L}_{S^2}$, illustrated by a representative example from AutoPET.

Visual enhancement on MAISI-v2. As shown in Figure 8, the baseline MAISI-v2 produces anatomically coarse structures with blurred boundaries and suboptimal intensity consistency. With $\mathcal{L}_{S^2}$, the synthesized results exhibit clearer anatomical structures and more faithful intensity distributions, approaching real CT appearance. These observations suggest that $\mathcal{L}_{S^2}$ can enhance the adaptation of 3D foundation models to CT synthesis tasks.

Slice-level and patient-level evaluation. To assess quantitative variability and volumetric fidelity, we additionally report slice-level metrics and patient-level 3D PSNR/SSIM on SegRap. Background-only slices are excluded from evaluation. For slice-level evaluation, PSNR and SSIM are computed on individual 2D slices, while FID, LFD, and HU-W1 are evaluated using 2D images or image distributions. For volumetric evaluation, consecutive synthesized and real CT slices are reconstructed into patient-level 3D volumes, on which 3D PSNR and SSIM are computed. Slice-level and patient-

Table 7: Slice-level and patient-level 3D CT synthesis performance on SegRap (mean ± standard deviation). Slice-level metrics are evaluated on 2D images, while 3D PSNR and SSIM are computed on reconstructed patient volumes.

Method	PSNR ↑	SSIM ↑	FID ↓	LFD ↓	HU-W1 ↓	3D PSNR ↑	3D SSIM ↑
pix2pixHD	24.869 ± 2.899	0.811 ± 0.078	37.623 ± 0.744	18.139 ± 0.668	120.127 ± 11.007	24.078 ± 0.932	0.829 ± 0.030
FFL	24.188 ± 2.915	0.791 ± 0.137	40.767 ± 0.824	18.060 ± 0.671	114.115 ± 20.443	23.329 ± 0.997	0.813 ± 0.031
OASIS	21.072 ± 4.136	0.753 ± 0.116	53.629 ± 0.876	17.683 ± 0.952	181.954 ± 61.509	19.624 ± 0.599	0.779 ± 0.037
LDM	28.809 ± 4.320	0.721 ± 0.180	75.542 ± 1.348	15.511 ± 0.995	36.226 ± 58.802	26.047 ± 2.068	0.738 ± 0.055
img2img-turbo	20.260 ± 4.367	0.728 ± 0.120	80.789 ± 1.249	17.781 ± 0.996	198.448 ± 66.885	18.593 ± 0.444	0.741 ± 0.037
SegGuideDiff	21.685 ± 2.150	0.284 ± 0.089	60.322 ± 1.576	18.623 ± 0.495	293.397 ± 73.184	19.503 ± 0.198	0.411 ± 0.029
FDL	23.521 ± 3.065	0.773 ± 0.140	60.611 ± 1.045	18.065 ± 0.706	111.250 ± 19.877	22.575 ± 0.993	0.797 ± 0.034
MOTFM	23.820 ± 2.912	0.619 ± 0.159	37.492 ± 0.827	23.820 ± 0.670	32.402 ± 26.073	22.692 ± 0.922	0.654 ± 0.027
MAISI-v2	15.697 ± 0.803	0.248 ± 0.076	63.890 ± 1.767	28.461 ± 0.185	15.740 ± 2.920	16.623 ± 0.803	0.301 ± 0.076
pix2pixHD+SID	25.017 ± 3.716	0.815 ± 0.151	35.553 ± 0.840	16.413 ± 0.856	110.231 ± 87.320	24.568 ± 1.052	0.834 ± 0.024
pix2pixHD+SA	24.982 ± 2.885	0.814 ± 0.136	33.186 ± 0.793	16.475 ± 0.664	115.286 ± 23.744	24.187 ± 0.925	0.836 ± 0.035
pix2pixHD+$\mathcal{L}_{S^2}$	25.068 ± 2.880	0.816 ± 0.135	33.932 ± 0.899	16.411 ± 0.663	107.981 ± 20.964	24.678 ± 0.932	0.840 ± 0.030
MOTFM+$\mathcal{L}_{S^2}$	24.214 ± 2.626	0.678 ± 0.145	32.535 ± 0.757	16.630 ± 0.605	32.036 ± 25.310	23.692 ± 0.922	0.714 ± 0.027
LDM+$\mathcal{L}_{S^2}$	32.175 ± 4.863	0.834 ± 0.159	51.496 ± 0.931	14.732 ± 1.120	16.204 ± 17.022	30.659 ± 3.515	0.875 ± 0.062
MAISI-v2+$\mathcal{L}_{S^2}$	15.998 ± 0.803	0.269 ± 0.076	56.213 ± 1.059	27.461 ± 0.185	13.737 ± 2.922	16.999 ± 0.811	0.322 ± 0.072

Table 8: Comparison with relevant loss in frequency domain through LDM and MOTFM on SegRap and AutoPET.

Method	SegRap [33]			AutoPET [35]		
	PSNR↑	SSIM↑	FID↓	PSNR↑	SSIM↑	FID↓
LDM [28] (baseline)	28.8092	0.7211	75.5422	23.2471	0.7597	20.1871
FFL [22]	30.6523	0.7983	54.2547	23.3995	0.7658	19.4318
FDL [32]	28.9654	0.8209	81.4184	23.0661	0.7424	27.8807
$\mathcal{L}_{S^2}$ **(Ours)**	**32.1748**	**0.8339**	**51.4962**	**23.5149**	**0.7705**	**18.1307**
MOTFM [29] (baseline)	23.8204	0.6185	37.4921	23.1514	0.6329	27.5958
FFL [22]	23.7959	0.6284	36.0017	23.4265	0.6423	24.7278
FDL [32]	23.9419	0.6613	33.8281	23.3668	0.6205	27.5005
$\mathcal{L}_{S^2}$ **(Ours)**	**24.2141**	**0.6776**	**32.5347**	**23.4275**	**0.6487**	**24.2353**

level results are reported as mean ± standard deviation over their respective evaluation units. As shown in Table 7, incorporating $\mathcal{L}_{S^2}$ generally improves image fidelity across GAN-, diffusion-, flow-, and foundation-model-based synthesis.

Multiplanar evaluation. To further assess volumetric fidelity, we evaluate sagittal and coronal SSIM on reconstructed CT volumes. On SegRap, MAISI-v2+$\mathcal{L}_{S^2}$ improves sagittal SSIM from 0.2542 ± 0.1080 to 0.2722 ± 0.1009 and coronal SSIM from 0.2515 ± 0.0063 to 0.2694 ± 0.0078. This suggests improved multiplanar fidelity, although they do not directly establish inter-slice consistency.

F Comparison with relevant loss using frequency modeling through LDM and MOTFM

As shown in Table 8, across both datasets (SegRap and AutoPET) and two generative frameworks (LDM and MOTFM), introducing $\mathcal{L}_{S^2}$ consistently improves performance on all pixel-space metrics (PSNR, SSIM, FID). In contrast, existing frequency-domain losses exhibit unstable behavior. Specifically, while they occasionally provide moderate gains, they can also degrade performance. For instance, FDL substantially worsens results on AutoPET for both LDM and MOTFM.

A particularly noteworthy observation arises on the small-scale SegRap dataset. LDM is generally not regarded as a strong generator under limited data (as also reflected in its lower baseline compared with pix2pixHD in the main text). However, injecting frequency constraints markedly improves its performance, and $\mathcal{L}_{S^2}$ yields the largest and most stable enhancement, lifting LDM to a significantly stronger regime. This demonstrates that $\mathcal{L}_{S^2}$ can effectively compensate for data scarcity and suboptimal baseline performance by semantics-structure learning, and then naturally enforcing low-frequency intensity coherence and high-frequency anatomical fidelity. Overall, the results indicate

Table 9: Ablation studies through LDM and MOTFM on SegRap and AutoPET.

Method	SegRap [33]			AutoPET [35]		
	PSNR↑	SSIM↑	FID↓	PSNR↑	SSIM↑	FID↓
LDM [28] (baseline)	28.8092	0.7211	75.5422	23.2471	0.7597	20.1871
$\mathcal{L}_{Sa}$	29.0628	0.7922	60.1438	23.4421	0.7683	18.6429
$\mathcal{L}_{Sid}$	30.1754	0.7743	61.5420	23.4810	0.7696	18.3949
$\mathcal{L}_{S^2}$ **(Ours)**	**32.1748**	**0.8339**	**51.4962**	**23.5149**	**0.7705**	**18.1307**
MOTFM [29] (baseline)	23.8204	0.6185	37.4921	23.1514	0.6329	27.5958
$\mathcal{L}_{Sa}$	23.9129	0.6667	33.5398	22.9898	0.6467	27.0439
$\mathcal{L}_{Sid}$	24.1011	0.6198	36.1982	23.4106	0.6484	25.5575
$\mathcal{L}_{S^2}$ **(Ours)**	**24.2141**	**0.6776**	**32.5347**	**23.4275**	**0.6487**	**24.2353**

Table 10: Conditional diversity on AutoPET (mean $\pm$ standard deviation). Higher values indicate greater sample-to-sample diversity.

Method	LPIPS	$1 - \mathrm{SSIM}$	Sample-to-sample HU-W1
LDM	0.0506 ± 0.0195	0.1314 ± 0.0436	1.8949 ± 0.5328
LDM+$\mathcal{L}_{S^2}$	0.0353 ± 0.0148	0.0948 ± 0.0284	1.7671 ± 0.3670

that $\mathcal{L}_{S^2}$ generalizes well across architectures, datasets, and evaluation dimensions, offering a reliable and substantial performance boost.

G Ablation studies through LDM and MOTFM

As shown in Table 9, the results of ablation studies through LDM and MOTFM are similar to pix2pixHD in the main text. Specifically, $\mathcal{L}_{Sa}$ and $\mathcal{L}_{Sid}$ could produce meaningful but limited gains, whereas $\mathcal{L}_{S^2}$ (the combination of $\mathcal{L}_{Sid}$ and $\mathcal{L}_{Sa}$) yields the largest and most stable improvements, outperforming every single-loss variant on all metrics.

H Conditional diversity and mask robustness

Overall conditional diversity. To assess the fidelity-diversity trade-off, we generate ten CT samples per conditioning mask on AutoPET test set using LDM and LDM+$\mathcal{L}_{S^2}$, covering head, chest, and abdominal regions. We measure pairwise LPIPS, $1 - \mathrm{SSIM}$, and sample-to-sample HU-W1 to quantify perceptual, structural, and intensity-distribution diversity, respectively. As shown in Table 10, introducing $\mathcal{L}_{S^2}$ reduces LPIPS from 0.0506 to 0.0353 and $1 - \mathrm{SSIM}$ from 0.1314 to 0.0948, indicating reduced perceptual and structural diversity. And sample-to-sample HU-W1 decreases only moderately from 1.8949 to 1.7671, suggesting that intensity-distribution variability remains.

Qualitative analysis of conditional diversity. To complement the quantitative diversity evaluation, we visualize ten independently generated CT samples conditioned on the same anatomical mask. As shown in Figure 9, the generated images maintain broadly consistent anatomical structures while exhibiting variations in local tissue textures, intensity patterns, and soft-tissue appearance.

Fidelity-diversity stratification. To investigate whether greater diversity is associated with reduced anatomical fidelity, we compute the mean pairwise LPIPS of each generated image relative to the other nine samples from the same mask. Within each mask and method, the ten samples are ranked by LPIPS and divided into Low (three samples), Middle (four samples), and High (three samples) diversity groups. For each sample, we compute HU-W1 against the corresponding real CT and anatomical Dice using an independent segmentation model trained on real AutoPET data.

As shown in Table 11, higher LPIPS is associated with larger HU distribution discrepancies and lower anatomical Dice. For LDM, HU-W1 increases from 197.56 to 231.14, while Dice decreases from 0.9404 to 0.9140 across the Low and High groups. A similar trend is observed for LDM+$\mathcal{L}_{S^2}$.

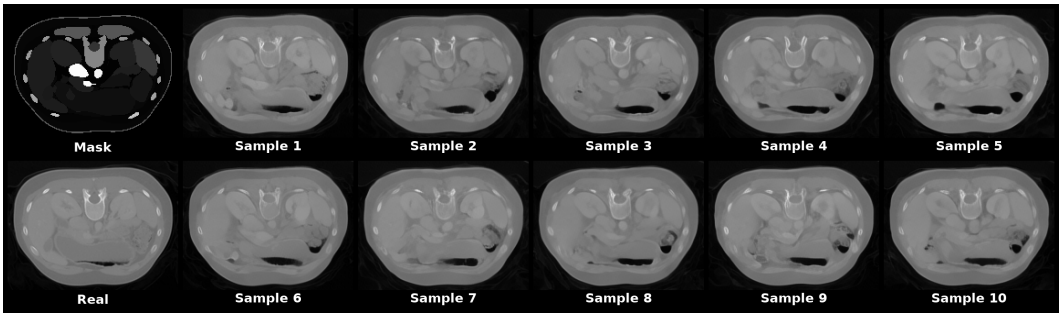

Figure 9: Qualitative illustration of conditional diversity in CT synthesis. Ten CT samples are independently generated from the same anatomical mask using different random seeds, alongside the corresponding real CT. The generated samples preserve the overall anatomical configuration while exhibiting variations in local tissue appearance and intensity distributions.

Table 11: Fidelity-diversity stratification on AutoPET (mean $\pm$ standard deviation). HU-W1 measures discrepancy against real CT, rather than sample-to-sample diversity.

Method	Group	LPIPS	$1 - \text{SSIM}$	HU-W1 $\downarrow$	Dice $\uparrow$
LDM	Low	0.0292 ± 0.0107	0.0906 ± 0.0309	197.56 ± 29.69	0.9404 ± 0.0248
LDM	Middle	0.0508 ± 0.0085	0.1330 ± 0.0301	218.47 ± 34.15	0.9267 ± 0.0250
LDM	High	0.0719 ± 0.0117	0.1702 ± 0.0315	231.14 ± 33.19	0.9140 ± 0.0219
LDM+$\mathcal{L}_{S^2}$	Low	0.0185 ± 0.0074	0.0679 ± 0.0201	199.42 ± 22.41	0.9482 ± 0.0177
LDM+$\mathcal{L}_{S^2}$	Middle	0.0353 ± 0.0067	0.0958 ± 0.0205	207.81 ± 22.36	0.9423 ± 0.0206
LDM+$\mathcal{L}_{S^2}$	High	0.0520 ± 0.0073	0.1204 ± 0.0188	223.50 ± 21.58	0.9352 ± 0.0190

Within the High-diversity group, our method reduces HU-W1 from 231.14 to 223.50 and improves Dice from 0.9140 to 0.9352 compared with LDM. These observations suggest an association between greater perceptual variation and reduced anatomical or intensity fidelity.

Quality-filtered diversity analysis. To further examine whether reduced diversity primarily reflects the suppression of low-quality samples, we perform a progressive sample-removal analysis using $1 - \text{Dice}$ and HU-W1 as quality indicators. Within each mask and method, samples are ranked by each criterion, and the worst 10%, 20%, and 30% are progressively removed. Pairwise LPIPS is recomputed among the remaining samples, with matched sample counts across methods.

As shown in Table 12, removing low-quality samples reduces LPIPS gap between LDM and LDM+$\mathcal{L}_{S^2}$. Under $1 - \text{Dice}$ filtering, removing the worst 30% reduces the gap from 0.0153 to 0.0068, a 55.6% reduction. Similarly, HU-W1 filtering reduces the gap to 0.0074. This suggest that part of the diversity reduction is associated with fewer low-quality samples. However, the remaining LPIPS differences indicate that some plausible stochastic variation may also be suppressed by $\mathcal{L}_{S^2}$.

Robustness to incomplete and coarse masks. We further assess robustness on AutoPET using two test-time perturbations: (1) randomly dropping 50% of organ labels and (2) downsampling masks by a factor of two followed by upsampling to the original resolution. As shown in Table 13, both methods deteriorate under substantial label removal. Under 50% label dropout, LDM+$\mathcal{L}_{S^2}$ achieves slightly higher SSIM and lower FID than LDM, but lower PSNR. Under coarse masks, our method improves SSIM, PSNR, and FID over the matched LDM baseline. These findings suggest improved tolerance to moderate mask coarsening, while substantial annotation incompleteness remains challenging.

I Wilcoxon signed-rank test for anatomical fidelity

As shown in Figure 10, synthesized CTs achieve score distributions comparable to real CT. The Wilcoxon signed-rank test indicates no statistically significant difference (p = 0.0588). Notably, this single-reader evaluation does not establish clinical equivalence between real and synthesized CTs.

Table 12: Quality-filtered conditional diversity on AutoPET. Samples with the highest $1 - \mathrm{Dice}$ or HU-W1 are progressively removed, and pairwise LPIPS is recomputed among retained samples (mean $\pm$ standard deviation).

Filtering criterion	Removed	LDM LPIPS	LDM+$\mathcal{L}_{S^2}$ LPIPS	Difference
None	0%	0.0506 ± 0.0068	0.0353 ± 0.0045	0.0153
$1 - \mathrm{Dice}$	10%	0.0474 ± 0.0061	0.0349 ± 0.0043	0.0125
	20%	0.0437 ± 0.0053	0.0343 ± 0.0041	0.0094
	30%	0.0406 ± 0.0048	0.0338 ± 0.0039	0.0068
HU-W1	10%	0.0478 ± 0.0062	0.0350 ± 0.0044	0.0128
	20%	0.0442 ± 0.0055	0.0345 ± 0.0042	0.0097
	30%	0.0414 ± 0.0049	0.0340 ± 0.0040	0.0074

Table 13: Robustness to test-time mask perturbations on AutoPET (mean $\pm$ standard deviation).

Condition	Method	SSIM ↑	PSNR ↑	FID ↓
Original	LDM	0.760 ± 0.054	23.247 ± 2.492	20.187 ± 0.121
	LDM+$\mathcal{L}_{S^2}$	0.771 ± 0.053	23.515 ± 2.548	18.131 ± 0.099
50% label drop	LDM	0.745 ± 0.084	22.845 ± 2.683	24.586 ± 0.210
	LDM+$\mathcal{L}_{S^2}$	0.749 ± 0.077	22.723 ± 2.648	21.973 ± 0.189
Coarse mask ($\times 2$)	LDM	0.759 ± 0.057	23.240 ± 2.513	18.256 ± 0.101
	LDM+$\mathcal{L}_{S^2}$	0.768 ± 0.047	23.517 ± 2.573	15.169 ± 0.103

J Mask fidelity evaluation via a pre-trained segmentation model

To evaluate the anatomical consistency of the generated images, we utilize a pre-trained nnU-Net [39] segmentation model. The model was trained on the SegRap training and validation sets to segment the body contour and 45 organs-at-risk (OARs). The assessments are implemented using two metrics. Pixel accuracy (Pixel Acc.) computes the ratio of correctly classified pixels across the generated images. mIOU is the area-weighted Mean Intersection over Union, providing a robust measure of spatial overlap across multi-scale anatomical structures.

As shown in Table 14, the LDM backbone combined with our proposed $\mathcal{L}_{S^2}$ loss achieves the highest performance (0.9746 Pixel Acc. and 0.8326 mIoU). This demonstrates that our method best preserves complex anatomical geometries among all tested baselines. Besides, effectiveness of the $\mathcal{L}_{S^2}$ Loss is validated. Across all backbones (LDM, pix2pixHD, and MOTFM), the inclusion of $\mathcal{L}_{S^2}$ consistently outperforms the baseline (No Loss) and alternative constraints such as FFL and FDL. This underscores the generalizability of our loss function. While diffusion-based models (LDM, MOTFM) generally exhibit higher fidelity than GAN-based approaches (pix2pixHD), LDM remains the most responsive to our structural optimization, yielding the most precise results for the 45 OARs and body contour.

K Downstream Segmentation Evaluation

Evaluation protocol. Segmentation experiments follow the same patient-wise split as synthesis on SegRap (90/20/10 patients for training/validation/testing), with no patient overlap. Segmentation models are trained and validated separately using real CTs, LDM-generated CTs, or LDM+$\mathcal{L}_{S^2}$-generated CTs, paired with the corresponding real anatomical masks. The same patients, masks, sample sizes, and training settings are used across all experiments. Evaluation is performed exclusively on real CTs and masks from unseen test patients. Thus, the synthetic-data experiments assess synthetic-to-real transfer rather than real-data augmentation.

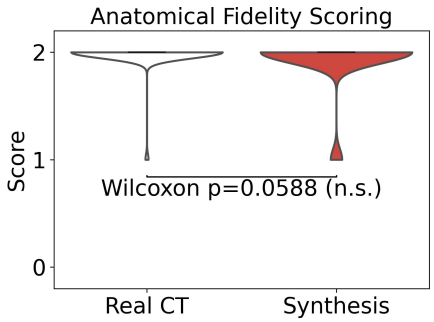

Figure 10: Anatomical fidelity scores from the single-reader user study.

Table 14: Comparison of mask fidelity on generation based on SegRap.

Method		Pixel Acc. (↑)	mIoU (↑)
Backbone	Loss		
img2img-turbo [30]	-	0.8820	0.5010
SegGuidedDiff [17]	-	0.9631	0.7562
OASIS [31]	-	0.9300	0.6948
LDM [28]	-	0.9617	0.7598
	FFL [22]	0.9728	0.8112
	FDL [32]	0.9699	0.7947
	$\mathcal{L}_{S^2}$(Ours)	**0.9746**	**0.8326**
pix2pixHD [27]	-	0.9238	0.7148
	FFL [22]	0.9321	0.6994
	FDL [32]	0.9316	0.6937
	$\mathcal{L}_{S^2}$(Ours)	0.9364	0.7180
MOTFM [29]	-	0.9374	0.8096
	FFL [22]	0.9598	0.8100
	FDL [32]	0.9586	0.8135
	$\mathcal{L}_{S^2}$(Ours)	0.9606	0.8149

Organ-wise segmentation analysis. Table 15 reports 2D-level organ-wise Dice scores for segmentation models trained on real CTs, LDM-generated CTs, and LDM+$\mathcal{L}_{S^2}$-generated CTs. Compared with LDM, our method improves Dice scores for most anatomical structures, with notable gains in challenging regions such as the SpinalCord, Brain, Submandibular glands, Mandible, Mastoid, and TemporalLobe. These results suggest improved synthetic-to-real segmentation transfer across diverse anatomical structures.

Patient-level 3D evaluation. To further assess volumetric segmentation performance, we stack the 2D predictions and ground-truth masks into 3D volumes for each test patient. Dice, IoU, HD95, and ASSD are computed for each organ within each patient volume. For overall evaluation, metrics are first averaged across organs within each patient, followed by calculating the mean and standard deviation across the 10 test patients. As shown in Table 16, LDM+$\mathcal{L}_{S^2}$ improves mean Dice from 0.4765 ± 0.0193 to 0.6022 ± 0.0322 and mean IoU from 0.3751 ± 0.0182 to 0.4714 ± 0.0300 compared with LDM. It also reduces HD95 from 4.7790 ± 0.5811 to 3.7903 ± 0.4388 mm and ASSD from 1.9145 ± 0.3267 to 1.5178 ± 0.2733 mm. Consistent numerical improvements are observed for representative organs (Table 17).

Limitations. Although synthetic-only training outperforms real-only training in this setting, the advantage may partly reflect greater image-label consistency from mask-conditioned synthesis. This remains a hypothesis rather than an established causal explanation. Given the single-center dataset and limited test cohort, these findings do not establish that synthetic data are generally superior to real clinical data.

Table 15: Organ-wise Dice performance comparison.

Organ	Dice↑		
	True	LDM	LDM+$\mathcal{L}_{S^2}$
SpinalCord	0.5487	0.8086	0.8454
Brain	0.7141	0.9512	0.9681
TMjoint_R	0.9729	0.9817	0.9880
TympanicCavity_R	0.9685	0.9728	0.9794
BrainStem	0.8326	0.9471	0.9550
Mastoid_R	0.9054	0.9573	0.9666
TemporalLobe_L	0.8403	0.9422	0.9537
Mastoid_L	0.8997	0.9459	0.9576
Larynx	0.9057	0.9321	0.9441
Mandible_R	0.8307	0.9247	0.9367
Submandibular_L	0.9271	0.9363	0.9513
TMjoint_L	0.9701	0.9822	0.9823
VestibulSemi_L	0.9818	0.9818	0.9821
OpticNerve_L	0.9798	0.9819	0.9825
MiddleEar_L	0.9805	0.9841	0.9848
Parotid_R	0.8590	0.9369	0.9377
IAC_L	0.9868	0.9868	0.9878
Chiasm	0.9818	0.9857	0.9871
Parotid_L	0.8561	0.9381	0.9401
Larynx_Glottic	0.9438	0.9561	0.9583
ETbone_R	0.9870	0.9858	0.9881
Larynx_Supraglot	0.9424	0.9575	0.9604
ETbone_L	0.9840	0.9838	0.9871
Cochlea_R	0.9840	0.9838	0.9871
Trachea	0.7887	0.9348	0.9386
Esophagus	0.6824	0.8773	0.8813
Submandibular_R	0.9411	0.9530	0.9571
Eye_L	0.8863	0.9739	0.9783
IAC_R	0.9859	0.9858	0.9903
OpticNerve_R	0.9777	0.9786	0.9832
Hippocampus_R	0.9413	0.9527	0.9581
PharynxConst	0.8130	0.8679	0.8735
TemporalLobe_R	0.8503	0.9508	0.9566
TympanicCavity_L	0.9685	0.9715	0.9774
MiddleEar_R	0.9824	0.9821	0.9881
Mandible_L	0.8319	0.9309	0.9369
Thyroid	0.8757	0.9249	0.9217
Hippocampus_L	0.9354	0.9614	0.9595
Cochlea_L	0.9838	0.9838	0.9821
Lens_L	0.9777	0.9777	0.9766
Eye_R	0.9198	0.9780	0.9769
Pituitary	0.9858	0.9870	0.9864
VestibulSemi_R	0.9818	0.9818	0.9814
Lens_R	0.9747	0.9747	0.9745
OralCavity	0.8884	0.9678	0.9677

L Broader impacts

This work proposes a CT-specific semantics-structure modeling framework for improving the fidelity of synthetic CT generation. The proposed method may have several positive impacts on medical imaging research and clinical AI development.

First, high-fidelity synthetic CT images can be used for data augmentation to improve the robustness and generalization of downstream AI systems, particularly in scenarios with limited annotated data or rare anatomical distributions. Second, our method directly targets two clinically important failure modes of synthetic CT generation: HU distribution shift and structural hallucination. Reducing

Table 16: Patient-level 3D segmentation performance on SegRap (mean ± standard deviation across 10 test patients).

Training data	Dice ↑	IoU ↑	HD95 (voxel) ↓	ASSD (voxel) ↓	HD95 (mm) ↓	ASSD (mm) ↓
Real CT	0.3591 ± 0.0392	0.3036 ± 0.0192	6.0969 ± 1.6045	2.0958 ± 1.1711	8.8589 ± 2.3314	3.1293 ± 1.7486
LDM	0.4765 ± 0.0193	0.3751 ± 0.0182	3.2734 ± 0.4782	1.2673 ± 0.2578	4.7790 ± 0.5811	1.9145 ± 0.3267
LDM+$\mathcal{L}_{S^2}$	$\mathbf{0.6022 \pm 0.0322}$	$\mathbf{0.4714 \pm 0.0300}$	$\mathbf{2.6211 \pm 0.4881}$	$\mathbf{1.0286 \pm 0.2101}$	$\mathbf{3.7903 \pm 0.4388}$	$\mathbf{1.5178 \pm 0.2733}$

Table 17: Patient-level 3D Dice scores for representative organs (mean ± standard deviation across patients).

Training data	Submandibular-L	Brain	SpinalCord
Real CT	0.4494 ± 0.0994	0.6527 ± 0.1693	0.3794 ± 0.0823
LDM	0.5523 ± 0.1210	0.9512 ± 0.0084	0.6976 ± 0.0610
LDM+$\mathcal{L}_{S^2}$	$\mathbf{0.6664 \pm 0.0894}$	$\mathbf{0.9620 \pm 0.0086}$	$\mathbf{0.7448 \pm 0.0465}$

HU distribution mismatch may improve intensity reliability for radiotherapy dose evaluation, while improving anatomical fidelity may benefit contouring and anatomy-aware clinical decision-making in radiotherapy planning. In addition, the proposed loss can be integrated into diverse generative frameworks and foundation models, potentially facilitating broader adoption of generative AI in medical imaging.

However, this work may also introduce potential negative impacts if misused. Synthetic medical images could be exploited to fabricate clinical records, manipulate insurance claims, generate misleading medical evidence, or construct falsified medical datasets. To mitigate these risks, synthetic CT images generated by our method should be used under appropriate clinical supervision and ethical regulations.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: See the abstract and Section 1.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See the Section 5 (especially the last three sentences).

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: The paper does not claim formal theoretical guarantees.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Experiment details for reproducibility are provided in Section 4, and the code will be released upon acceptance.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: The code needs to go through an internal approval process for release. We will submit it post paper acceptance.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: See Section 4.1, Appendix B and C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Wilcoxon signed-rank test is implemented in user study.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: See Appendix C.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: All experiments are conducted on publicly available datasets.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: See Appendix J.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper proposes a training loss rather than releasing a standalone generative system, public model, or scraped dataset. Potential misuse of synthetic medical images is discussed in the broader impacts section.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: In Section 4, we provide references for the datasets and baselines.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: The paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.