KNOWLEDGE-GUIDED MULTI-TASK DIFFUSION WITH TYPE-AWARE BOUNDARY MODELING FOR LESION SYNTHESIS

Anonymous authors
Paper under double-blind review

ABSTRACT

Annotated pathological CT data remain scarce, limiting the generalization of lesion segmentation models. Diffusion-based synthesis offers scalable data augmentation, but mask- or prompt-conditioned generation alone does not enforce measurable radiological properties or realistic lesion–tissue transitions; joint generation and segmentation can also produce conflicting gradients. We propose a knowledge-guided lesion synthesis framework with two complementary components. Knowledge-guided multi-task lesion synthesis (KG-MTS) incorporates CT priors as differentiable constraints on lesion density, tissue composition, and internal heterogeneity, and introduces auxiliary segmentation through asymmetric conflict-aware weighting that attenuates harmful segmentation updates while preserving the generation direction. Type-aware boundary transition modeling (TA-BTM) represents the lesion boundary as a phenotype-dependent intensity transition zone and aligns its signed contrast, effective width, and smoothness with real-lesion statistics. Given a CT region of interest, a prescribed mask, and a structured radiological prompt, the model synthesizes controllable lesions while preserving the surrounding anatomy. We evaluate pulmonary nodule and renal tumor synthesis on LIDC-IDRI and KiTS23 using three downstream segmentation architectures. Compared with the reproduced LeFusion baseline, our method achieves Dice and NSD improvements of 2.80–3.20% and 2.04–2.38% on LIDC-IDRI, and 1.73–3.05% and 1.31–1.99% on KiTS23, respectively. Ablations further show that KG-MTS and TA-BTM provide complementary improvements in volumetric overlap and boundary agreement. Code: https://anonymous.4open.science/r/lesion-synthesis-review-1C7D

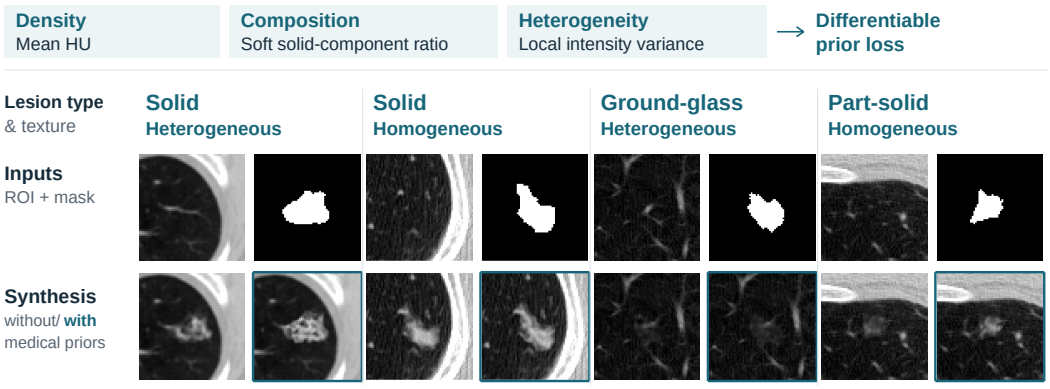

Figure 1: Medical knowledge guidance. Paired results without (left) and with (right) priors share the same ROI, mask, prompt, and initial noise.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Clinical motivation: data scarcity in lesion analysis. Accurate CT lesion segmentation requires diverse pathological images with voxel-wise annotations, which are costly to obtain. Even LIDC-IDRI (Armato et al., 2011) exhibits substantial variation in lesion size, density, texture, and morphology, with rare phenotypes remaining underrepresented. Lesion synthesis addresses this scarcity by inserting pathology into lesion-free scans to produce paired image–mask data. Diffusion models (Ho et al., 2020) support this approach through background-preserving inpainting (Lugmayr et al., 2022), while LeFusion restricts generation to pathological regions (Zhang et al., 2025a). Useful augmentation nevertheless requires control of lesion phenotype and realistic interaction with surrounding tissue, including in both pulmonary nodules and renal tumors.

Limitations of existing mask-guided lesion synthesis. Two limitations motivate our work. **(i) Insufficient knowledge guidance and task integration.** A mask specifies where a lesion appears, but not its radiological phenotype. Histogram and text conditions provide appearance control (Zhang et al., 2025a; Lei et al., 2025), yet conditional encoding alone does not explicitly enforce measurable CT density, tissue composition, or internal heterogeneity. Moreover, generators trained separately from segmentation receive no task-specific structural feedback. Joint training can provide such feedback, but conflicting gradients may cause segmentation supervision to interfere with synthesis (Yu et al., 2020). **(ii) Oversimplified boundary representation.** Binary masks specify geometry, whereas real lesion–tissue interfaces occupy transition regions whose contrast, width, and smoothness depend on lesion phenotype and anatomical context. Ignoring these transitions can introduce synthetic edge artifacts even when the prescribed shape is respected.

Our approach: KG-MTS with TA-BTM. We introduce two complementary components. **Knowledge-guided multi-task lesion synthesis (KG-MTS)** expresses lesion density, tissue composition, and local texture variation as differentiable constraints on the generated clean estimate. Prompt-specific intervals derived from real training lesions guide these descriptors while allowing variation within each valid range. An auxiliary ROI segmentation task adds structural supervision. Asymmetric conflict-aware weighting retains aligned segmentation updates and attenuates conflicting ones while leaving the primary generation gradient term unchanged. **Type-aware boundary transition modeling (TA-BTM)** measures signed contrast, effective transition width, and normalized second-order variation around the prescribed mask. Type-specific real-lesion statistics constrain these measurements, linking boundary appearance to lesion phenotype rather than geometry alone. Together, the components address internal lesion appearance and its transition to surrounding tissue. We evaluate their augmentation utility on LIDC-IDRI and KiTS23 using three downstream segmentation architectures.

Contributions. Our contributions are twofold:

- **KG-MTS.** To address weak radiological control and task conflict, we combine medical constraints with adaptive segmentation weighting, reducing medical NID from 0.669 to 0.049 on LIDC-IDRI and 0.782 to 0.094 on KiTS23.
- **TA-BTM.** To overcome binary boundary modeling, we constrain type-specific intensity transitions, improving MedNeXt NSD by 3.62% on LIDC-IDRI and 2.87% on KiTS23 over Prompt-only.

2 RELATED WORK

Lesion synthesis for data augmentation. Medical image synthesis spans variational autoencoders, GANs (Kingma & Welling, 2014; Goodfellow et al., 2014), and diffusion models (Ho et al., 2020). Conditional GANs have been used for pulmonary nodule synthesis and free-form tumor insertion (Han et al., 2019; Jin et al., 2021), although jointly generating lesions and anatomy can alter the background. Copy-paste and rule-based insertion preserve surrounding anatomy; the latter constructs lesion shapes and textures using geometric primitives and intensity rules (Hu et al., 2023). These methods offer spatial control, but manually specified appearance models are difficult to extend across heterogeneous lesion types. Recent work studies synthetic tumors for scaling CT segmentation (Chen et al., 2025), lesion-conditioned autoregression for skin images (Sun et al., 2025), and clinically guided diffusion for laryngeal lesion detection (Baldini et al., 2025).

Controllable diffusion-based lesion generation. Diffusion inpainting (Lugmayr et al., 2022; Avrahami et al., 2022) preserves surrounding context by combining a denoised foreground with noised observations. DiffTumor conditions latent diffusion on masks and healthy anatomy for abdominal tumor synthesis (Chen et al., 2024). LeFusion uses lesion-focused denoising, histogram conditioning, multi-channel decomposition, and mask diffusion for appearance and shape control (Zhang et al., 2025a). Richer conditioning includes structured reports in LesionDiffusion (Lei et al., 2025), quantitative shape and texture descriptors in radiomics-conditioned synthesis (Kim et al., 2025), and radiomic-cluster-derived phenotype prompts in LesionLab (Lyu et al., 2026). Our method complements these conditioning strategies by measuring the generated clean estimate during training and directly penalizing deviations from prompt-specific real-data intervals.

Multi-task learning and gradient conflict. Shared representations support learning across related tasks (Caruana, 1997), but differences in loss scales and gradient directions complicate joint optimization. GradNorm balances gradient magnitudes (Chen et al., 2018), multi-objective optimization seeks Pareto solutions (Sener & Koltun, 2018), and PCGrad projects conflicting gradient components (Yu et al., 2020). Gradient-similarity weighting instead assesses whether an auxiliary update benefits the primary task (Du et al., 2018). KG-MTS adopts this asymmetric setting for knowledge-guided diffusion: segmentation supervision is weighted by timestep and prompt type, then attenuated when its gradient conflicts with generation.

Boundary-aware learning. Contour- and distance-based losses introduce boundary information into segmentation (Kervadec et al., 2019; Karimi & Salcudean, 2020). In synthesis, DiffBoost uses explicit edge maps to preserve structure (Zhang et al., 2025b), while LeFusion's mask diffusion controls lesion geometry (Zhang et al., 2025a). TA-BTM instead characterizes the intensity transition across the lesion–tissue interface using type-specific contrast, width, and smoothness statistics. The mask determines where these descriptors are measured, and real-lesion statistics determine their target ranges, without an additional edge-map input during synthesis.

3 METHOD

Given a three-dimensional CT image $\mathbf{x}_0$, a binary lesion mask $\mathbf{M}$, and a structured radiological prompt p, we synthesize a lesion inside $\mathbf{M}$ while preserving the surrounding anatomy. As illustrated in Fig. 2, the prompt-conditioned diffusion backbone is trained with two complementary components: knowledge-guided multi-task lesion synthesis (KG-MTS), which couples measurable radiological priors with structure-aware segmentation supervision, and type-aware boundary transition modeling (TA-BTM), which regularizes the lesion–tissue interface.

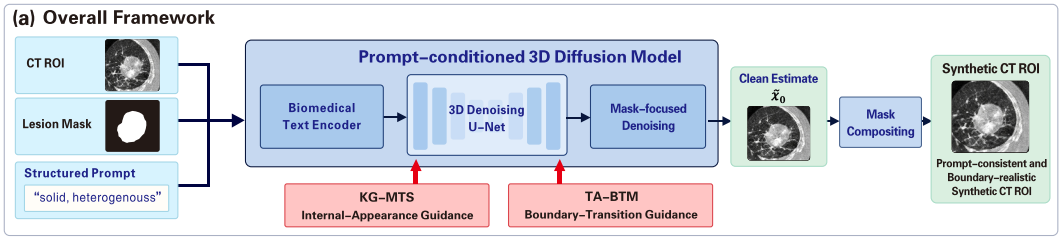

Figure 2: Overall workflow. A CT ROI, prescribed mask, and structured prompt condition lesion-focused diffusion. During training, KG-MTS guides internal appearance and auxiliary segmentation, while TA-BTM regularizes type-specific boundary transitions. The generated foreground is composited with the original exterior context to obtain the synthetic ROI.

3.1 PROMPT-CONDITIONED LESION DIFFUSION

The prompt describes measurable lesion attributes rather than a free-form appearance. For pulmonary nodules, it specifies the nodule type (ground-glass, part-solid, or solid) and internal pattern (homogeneous or heterogeneous). For renal tumors, the type is defined by attenuation relative to kidney parenchyma. A biomedical text encoder (Zhang et al., 2023) encodes the attributes in semantic

slots, which are injected into the bottleneck of a 3D denoising network through cross-attention. We retain the baseline lesion histogram as optional texture context and randomly drop it during training to prevent shortcut reliance.

Following DDPM (Ho et al., 2020), Gaussian noise is added to $\mathbf{x}_0$ at timestep t, producing $\mathbf{x}_t$. The network predicts the noise conditioned on t, p, and the optional histogram $\mathbf{h}$. Following lesion-focused diffusion (Zhang et al., 2025a), denoising is optimized only inside the lesion mask:

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{t,\boldsymbol{\epsilon}} \left[\left\| \mathbf{M} \odot \left(\boldsymbol{\epsilon} - \epsilon_\theta(\mathbf{x}_t, t, p, \mathbf{h}) \right) \right\|_1 \right]. \tag{1}$$

During reverse diffusion, the synthesized foreground is combined with the forward-diffused real background, so the exterior anatomy remains unchanged. This prompt-conditioned model provides the common initialization for the two proposed components.

3.2 Knowledge-Guided Multi-Task Lesion Synthesis

KG-MTS (Fig. 3) makes radiological knowledge part of the generation objective and uses the resulting knowledge-guided gradient as the protected direction in joint generation and segmentation learning.

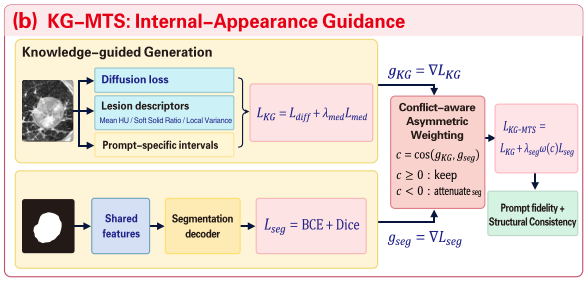

Figure 3: KG-MTS workflow. Mean HU, soft solid-component ratio, and local variance are compared with prompt-specific intervals to form the medical loss. Shared diffusion features feed an auxiliary segmentation decoder. Gradient cosine similarity determines its loss weight, attenuating conflicting updates while leaving the diffusion-plus-medical gradient term unchanged.

3.2.1 Medical Knowledge Guidance

We reconstruct the clean estimate $\hat{\mathbf{x}}_0$ from the predicted noise and convert it to CT intensity $\hat{\mathbf{I}}$ in HU. Within an eroded lesion core, we measure three differentiable descriptors: mean HU m_μ for lesion density, a soft solid-component ratio m_s obtained with a smooth HU threshold, and local intensity variance m_v for internal heterogeneity. Each prompt group selects descriptor intervals $[l_k, u_k]$ estimated from real training lesions. We use the following interval distance, which preserves natural intra-class diversity by penalizing only values outside the valid range:

$$d(m_k; l_k, u_k, a_k) = \text{ReLU}\left(\frac{l_k - m_k}{a_k}\right)^2 + \text{ReLU}\left(\frac{m_k - u_k}{a_k}\right)^2, \tag{2}$$

where a_k is a robust scale derived from the interquartile range. The medical knowledge loss is

$$\mathcal{L}_{\text{med}} = \sum_{k \in \{\mu, s, v\}} \gamma_k d(m_k; l_k, u_k, a_k). \tag{3}$$

Because $\hat{\mathbf{x}}_0$ is less reliable at highly corrupted timesteps, the loss is modulated by a low-noise weight $w_{\text{med}}(t)$ and a linearly warmed-up coefficient $\lambda_{\text{med}}(n)$. Consequently, the medical descriptors directly contribute to the denoising gradient rather than serving only as post-hoc evaluation measures.

3.2.2 Conflict-aware Segmentation Supervision

An auxiliary segmentation decoder receives shared diffusion encoder and bottleneck features from $\mathbf{x}_t$. The ground-truth mask is used only as supervision, not as decoder input. Its loss $\mathcal{L}_{\text{seg}}$ combines binary cross-entropy and soft Dice and is weighted by timestep and prompt type, yielding $\widetilde{\mathcal{L}}_{\text{seg}}$. This branch encourages the shared representation to preserve lesion structure, but its gradient may conflict with prompt-conditioned generation.

Let $\mathbf{g}_{\text{KG}}$ be the gradient of the diffusion and medical knowledge terms with respect to the shared parameters, and let $\mathbf{g}_{\text{seg}}$ be the corresponding segmentation gradient. We measure their compatibility

and convert it into an asymmetric attenuation factor:

$$c = \frac{\langle \mathbf{g}_{\mathrm{KG}}, \mathbf{g}_{\mathrm{seg}} \rangle}{\|\mathbf{g}_{\mathrm{KG}}\|_2 \|\mathbf{g}_{\mathrm{seg}}\|_2 + \varepsilon}, \qquad \omega(c) = \begin{cases} 1, & c \geq 0, \\ \mathrm{clip}(1 + c, 0, 1), & c < 0. \end{cases} \tag{4}$$

Inspired by gradient-conflict optimization (Yu et al., 2020), this rule fully retains helpful segmentation updates and continuously attenuates conflicting ones without modifying the primary generation direction. The complete KG-MTS objective is

$$\mathcal{L}_{\text{KG-MTS}} = \mathcal{L}_{\mathrm{diff}} + \lambda_{\mathrm{med}}(n)\mathbb{E}[w_{\mathrm{med}}(t)\mathcal{L}_{\mathrm{med}}] + \lambda_{\mathrm{seg}}(n)\,\mathrm{sg}[\omega(c)]\widetilde{\mathcal{L}}_{\mathrm{seg}}. \tag{5}$$

Here, $\mathrm{sg}[\cdot]$ denotes stop-gradient, such that $\omega(c)$ acts only as a scalar attenuation coefficient and introduces no higher-order gradient terms. Unlike direct gradient projection, the mechanism changes only a scalar auxiliary coefficient and therefore requires no additional optimization step.

3.3 Type-Aware Boundary Transition Modeling

TA-BTM treats the lesion boundary as an intensity transition region rather than merely a binary contour. Given the voxel spacing of each CT scan, we construct the inner and outer rings using physical distance thresholds around the lesion boundary. The predicted lesion inside $\mathbf{M}$ is combined with the real exterior context using stop-gradient, ensuring that boundary supervision updates only the generated lesion.

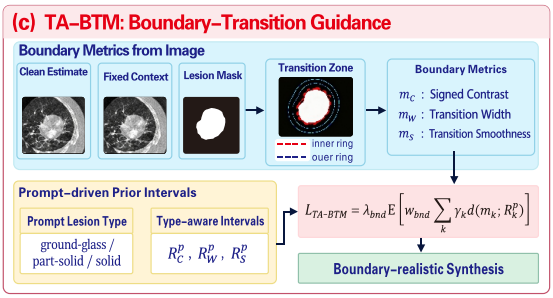

Figure 4: TA-BTM workflow. The generated lesion is combined with fixed exterior context, and the mask defines inner and outer rings in physical space. Signed contrast, transition width, and smoothness are compared with type-specific intervals selected by the prompt. Stop-gradient on the exterior restricts boundary supervision to the generated lesion.

The signed contrast m_C is the mean HU difference between the inner and outer rings. Image gradients and second-order differences are computed using axis-specific voxel spacing. The effective transition width m_W is the contrast magnitude divided by the mean physical gradient magnitude and is therefore measured in millimeters. The smoothness descriptor m_S is the spacing-normalized second-order variation divided by the physical gradient magnitude.

For each lesion type, the prompt selects real-data intervals for these descriptors. Reusing the interval distance in Eq. (2), the boundary term is

$$\mathcal{L}_{\text{TA-BTM}} = \lambda_{\mathrm{bnd}}(n)\mathbb{E}\left[w_{\mathrm{bnd}}(t)\sum_{k\in\{C,W,S\}} \gamma_k d(m_k; l_k^p, u_k^p, a_k^p)\right]. \tag{6}$$

Here, $w_{\mathrm{bnd}}(t)$ emphasizes reliable low-noise predictions, and the superscript p indicates type-specific statistics selected by the prompt. The mask determines where the transition is measured, whereas real CT statistics determine how it should appear.

3.4 Training and Inference

When both components are enabled, the overall objective is

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\text{KG-MTS}} + \mathcal{L}_{\text{TA-BTM}}. \tag{7}$$

Training proceeds in two stages. We first train the prompt-conditioned diffusion model with $\mathcal{L}_{\mathrm{diff}}$. We then initialize from this checkpoint and activate the proposed supervision with linear warm-up. For controlled ablations, KG-MTS and TA-BTM are also trained independently from the same prompt-only checkpoint. All prior intervals are estimated exclusively from the training set. The auxiliary segmentation and prior objectives are used to optimize the diffusion model, while lesion synthesis is conditioned only on a target CT, a prescribed lesion mask, and a structured prompt.

4 EXPERIMENTS

We evaluate synthesis-based augmentation for pulmonary nodule and renal tumor ROI segmentation on LIDC-IDRI and KiTS23. Experiments examine downstream performance, module ablations, and medical and boundary descriptor consistency. Detailed settings are provided in Appendix A.

4.1 COMPARISON WITH ALTERNATIVE SYNTHESIS METHODS

Our method achieves the highest Dice and NSD across all three backbones on both datasets (Tables 1 and 2). Compared with LeFusion, Dice/NSD improve by 2.80–3.20%/2.04–2.38% on LIDC-IDRI and 1.73–3.05%/1.31–1.99% on KiTS23, indicating consistent augmentation gains across organs and downstream architectures.

Table 1: Pulmonary nodule ROI segmentation on LIDC-IDRI. Dice/NSD (%) are averaged over three runs; best results are bold.

	nnU-Net		SwinUNETR		MedNeXt	
Method	Dice ↑	NSD ↑	Dice ↑	NSD ↑	Dice ↑	NSD ↑
Real Only	78.23	88.92	78.34	88.70	79.82	89.11
Hand-Crafted	77.46	87.50	77.45	86.95	77.84	87.69
Copy-Paste	77.15	87.63	77.60	88.54	78.11	88.90
Cond-Diffusion	78.43	88.68	78.55	87.05	78.90	88.20
DiffTumor	78.58	88.81	79.02	88.59	79.65	89.03
LeFusion	81.00	91.65	81.05	91.62	81.43	92.01
Ours	**84.20**	**94.03**	**83.85**	**93.71**	**84.25**	**94.05**

Table 2: Renal tumor ROI segmentation on the internal KiTS23 partition. Reporting follows Table 1.

	nnU-Net		SwinUNETR		MedNeXt	
Method	Dice ↑	NSD ↑	Dice ↑	NSD ↑	Dice ↑	NSD ↑
Real Only	78.88	86.30	78.92	86.71	79.00	86.54
Hand-Crafted	78.10	84.91	78.03	84.98	77.04	85.16
Copy-Paste	77.79	85.04	78.18	86.55	77.30	86.33
Cond-Diffusion	79.08	86.06	79.13	85.09	78.09	85.67
DiffTumor	79.23	86.19	79.61	86.60	78.83	86.45
LeFusion	81.67	88.93	81.65	89.56	80.60	89.34
Ours	**83.40**	**90.24**	**83.58**	**91.41**	**83.65**	**91.33**

4.2 ABLATION OF KG-MTS AND TA-BTM

Effects of the two modules. Both modules improve Dice and NSD over Prompt-only. KG-MTS yields larger Dice gains, while TA-BTM improves NSD more strongly, reflecting their focus on internal appearance and boundary transitions, respectively. The full model performs best on both datasets, supporting their complementarity.

Table 3 compares the four prompt-conditioned variants using MedNeXt on both datasets.

Table 3: MedNeXt ablation of KG-MTS and TA-BTM. Checkmarks indicate enabled modules; reporting follows Table 1.

			LIDC-IDRI		KiTS23	
Configuration	KG-MTS	TA-BTM	Dice ↑	NSD ↑	Dice ↑	NSD ↑
Prompt-only	–	–	80.96	89.09	80.34	86.60
+ KG-MTS	✓	–	83.30	91.51	82.55	87.61
+ TA-BTM	–	✓	81.47	92.71	81.29	89.47
Full model	✓	✓	**84.25**	**94.05**	**83.65**	**91.33**

4.3 Quantitative Evaluation of Lesion Realism

KG-MTS achieves the lowest average medical NID on both datasets, while TA-BTM achieves the lowest average boundary NID (Table 4), supporting their complementary focus on internal appearance and lesion–tissue transitions.

We assess medical and boundary descriptor consistency using normalized interval deviation (NID; lower is better). This dimensionless metric measures deviations from real-lesion reference ranges; evaluation details are provided in Appendix A.

Table 4: Descriptor realism on LIDC-IDRI and KiTS23 measured by NID (↓). Medical and boundary averages summarize their respective three descriptors. Best results within each dataset are bold.

	Medical descriptors				Boundary descriptors			
Method	Mean HU	Solid ratio	Local var.	Avg.	Contrast	Width	Smooth.	Avg.
				LIDC-IDRI				
Prompt-only	0.041	1.745	0.220	0.669	0.082	0.075	0.069	0.075
+ KG-MTS	**0.021**	**0.098**	**0.027**	**0.049**	**0.051**	0.060	0.069	0.060
+ TA-BTM	0.046	0.652	0.221	0.306	0.061	**0.052**	**0.061**	**0.058**
				KiTS23				
Prompt-only	0.993	1.115	0.238	0.782	0.128	0.117	0.145	0.130
+ KG-MTS	**0.037**	**0.170**	**0.074**	**0.094**	0.104	0.084	0.131	0.106
+ TA-BTM	0.610	0.781	0.221	0.537	**0.082**	**0.065**	**0.078**	**0.075**

4.4 Qualitative Analysis

Comparison of synthesized lesions. Figure 5 compares representative synthesis outputs under matched background ROIs and prescribed masks.

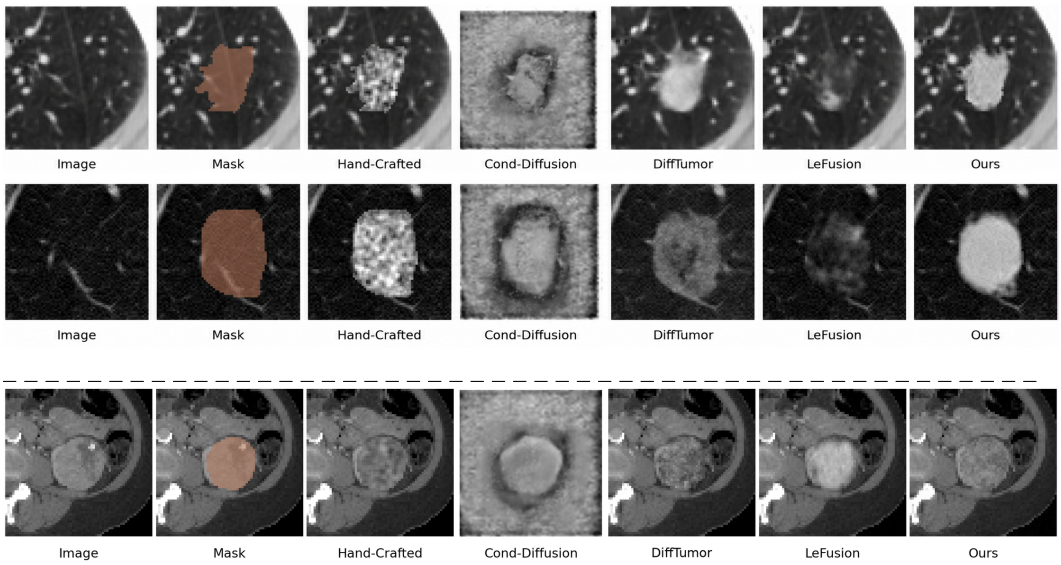

Figure 5: Lesion synthesis on LIDC-IDRI (top two rows) and KiTS23 (bottom row). The examples illustrate lesion–mask alignment and internal appearance across pulmonary and renal ROIs.

Downstream prediction consistency. Figure 6 visualizes how the choice of synthetic augmentation affects a downstream segmenter. Hand-Crafted augmentation produces scattered confidence variations inside the lesion, while the DiffTumor and LeFusion results contain local low-confidence regions or responses extending beyond the target contour. With our synthetic data, the predicted lesion probability is more spatially uniform and its outer extent follows the reference boundary more closely. This qualitative comparison is consistent with the simultaneous Dice and NSD improvements reported in Table 1. The maps complement overlap scores by showing where predictions

differ: low responses within the target contour indicate incomplete coverage, whereas responses outside it indicate spatial leakage. Inspecting both patterns helps distinguish consistent lesion coverage from a simple increase in foreground confidence. This distinction also helps interpret improvements in boundary agreement.

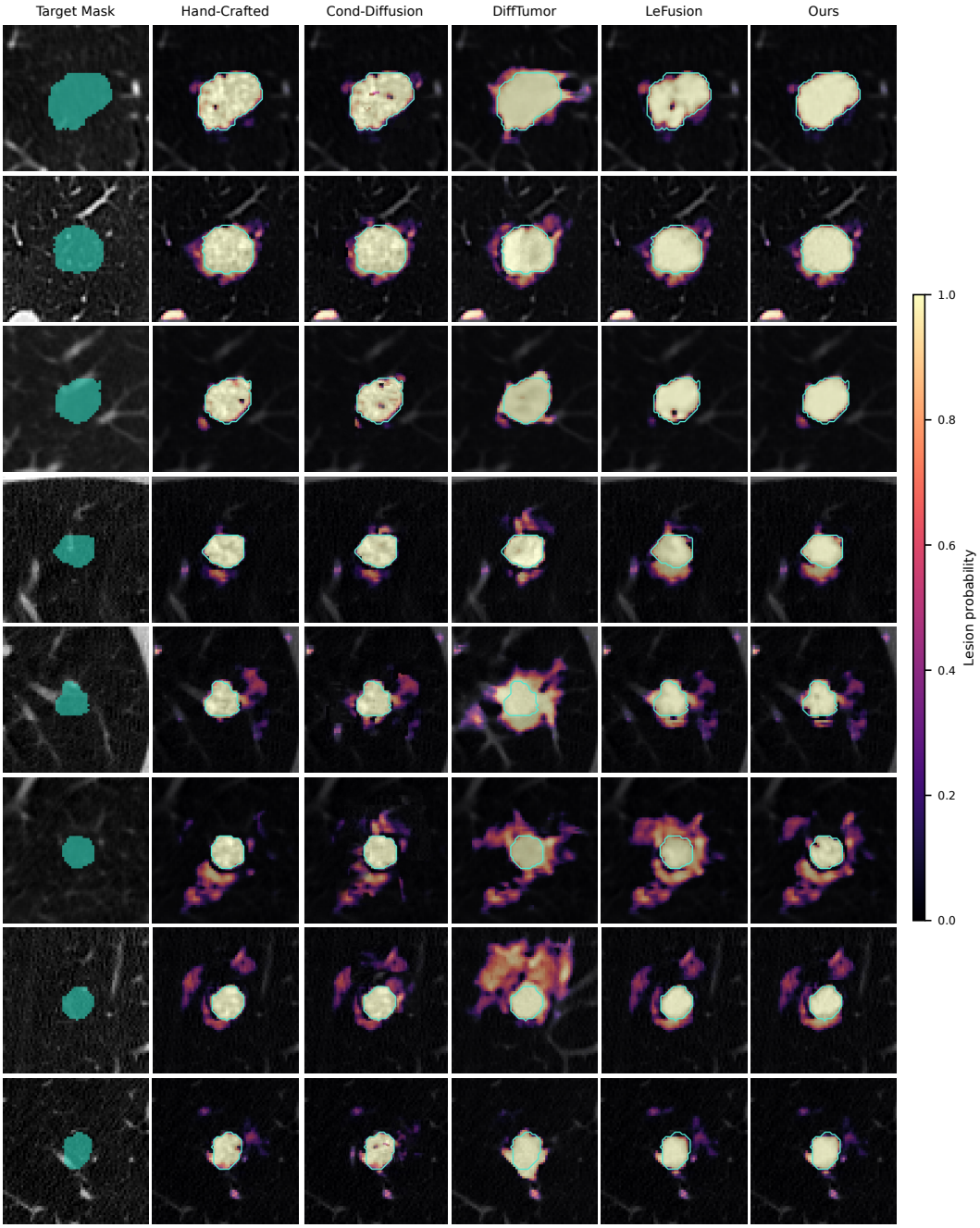

Figure 6: Downstream segmentation probabilities for eight lesion ROIs. Cyan contours mark the target boundary; warmer colors indicate higher probability. The probability maps illustrate more uniform lesion responses and closer agreement with the target contour.

Effect of boundary-transition modeling. Figure 7 compares synthesis before and after TA-BTM across eight pulmonary-nodule examples with different prescribed geometries. TA-BTM produces more gradual intensity transitions while retaining the overall lesion shape specified by the target

mask. The examples illustrate its effect on both compact and irregular boundaries, complementing the downstream NSD improvements in Table 3. The comparison emphasizes the transition at the lesion–tissue interface: preserving a target shape alone does not determine how intensities change across its boundary. Type-aware supervision addresses this appearance component, complementing the geometric information already supplied by the mask.

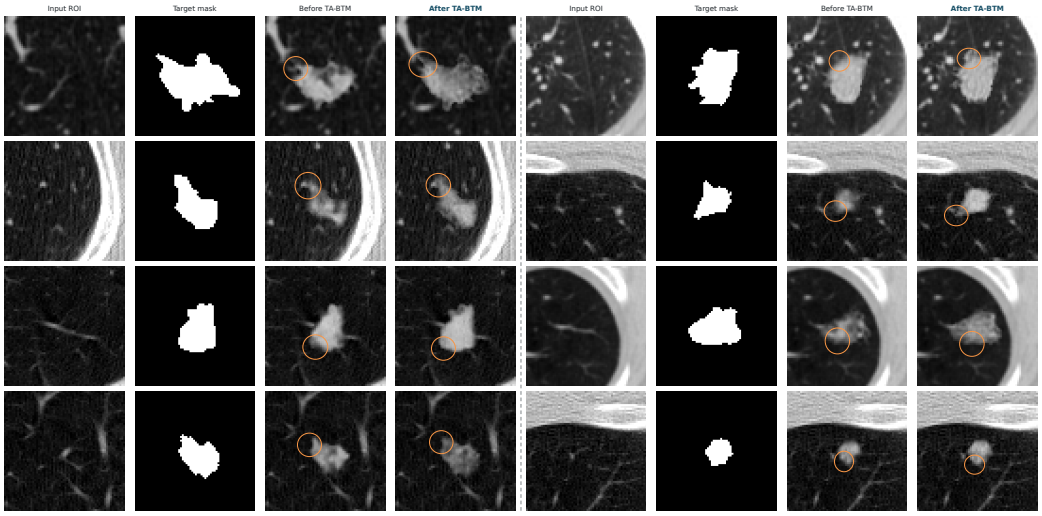

Figure 7: Boundary details before and after TA-BTM under matched conditions. Orange circles highlight local lesion–tissue interfaces. The paired examples illustrate changes in local boundary transitions while retaining the prescribed lesion shape.

5 Conclusion

We presented a knowledge-guided diffusion framework for controllable CT lesion synthesis with two complementary components. KG-MTS translates measurable radiological properties into differentiable constraints and incorporates auxiliary segmentation supervision through asymmetric conflict-aware weighting. TA-BTM regularizes the lesion–tissue interface using type-specific statistics of signed contrast, effective transition width, and smoothness. Together, these components guide internal lesion appearance and boundary transitions while preserving the surrounding anatomy. Experiments on LIDC-IDRI and KiTS23 show consistent improvements over the reproduced LeFusion baseline in downstream Dice and NSD across three segmentation architectures. Ablations further support the complementary contributions of knowledge-guided multi-task learning and boundary-transition modeling. These findings demonstrate the value of combining measurable radiological priors with structural supervision for synthetic data augmentation.

The current evaluation is limited to lesion ROI segmentation with prescribed masks and does not establish performance in full-volume lesion localization or zero-shot cross-organ transfer. Future work will investigate broader anatomical settings and full-volume evaluation to assess the framework's utility beyond the current ROI-based setting.

AI Use Statement

In this work, we used generative AI tools to aid or polish writing. We have not used generative AI tools to support retrieval and discovery (e.g., finding related work), to support research ideation and execution, to draft sections of the paper, to generate synthetic datasets, or to prove mathematical claims.

REFERENCES

Samuel G. Armato et al. The lung image database consortium (LIDC) and image database resource initiative (IDRI): A completed reference database of lung nodules on CT scans. *Medical Physics*, 38(2):915–931, 2011. doi: 10.1118/1.3528204.

Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18208–18218, 2022.

Chiara Baldini, Kaisar Kushibar, Richard Osuala, Simone Balocco, Oliver Diaz, Karim Lekadir, and Leonardo S. Mattos. Clinically-guided data synthesis for laryngeal lesion detection. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, pp. 54–63, 2025. doi: 10.1007/978-3-032-05141-7_6.

Rich Caruana. Multitask learning. *Machine Learning*, 28:41–75, 1997. doi: 10.1023/A:1007379606734.

Qi Chen, Xiaoxi Chen, Haorui Song, Zhiwei Xiong, Alan L. Yuille, Wei Chen, and Zongwei Zhou. Towards generalizable tumor synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/papers/Chen_Towards_Generalizable_Tumor_Synthesis_CVPR_2024_paper.pdf.

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

Zhao Chen, Vijay Badrinarayanan, Chen-Yu Lee, and Andrew Rabinovich. GradNorm: Gradient normalization for adaptive loss balancing in deep multitask networks. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 794–803, 2018.

Yunshu Du, Wojciech M. Czarnecki, Siddhant M. Jayakumar, Mehrdad Farajtabar, Razvan Pascanu, and Balaji Lakshminarayanan. Adapting auxiliary losses using gradient similarity. *arXiv preprint arXiv:1812.02224*, 2018.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, volume 27, pp. 2672–2680, 2014.

Changhee Han, Yoshiro Kitamura, Akira Kudo, Akimichi Ichinose, Leonardo Rundo, Yujiro Furukawa, Kazuki Umemoto, Yuanzhong Li, and Hideki Nakayama. Synthesizing diverse lung nodules wherever massively: 3d multi-conditional GAN-based CT image augmentation for object detection. In *International Conference on 3D Vision*, pp. 729–737, 2019.

Ali Hatamizadeh, Vishwesh Nath, Yucheng Tang, Dong Yang, Holger Roth, and Daguang Xu. Swin UNETR: Swin transformers for semantic segmentation of brain tumors in MRI images. *arXiv preprint arXiv:2201.01266*, 2022.

Nicholas Heller, Fabian Isensee, Dasha Trofimova, Resha Tejpaul, Zhongchen Zhao, Huai Chen, Lisheng Wang, Alex Golts, Daniel Khapun, Daniel Shats, et al. The KiTS21 challenge: Automatic segmentation of kidneys, renal tumors, and renal cysts in corticomedullary-phase CT. *arXiv preprint arXiv:2307.01984*, 2023.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851, 2020.

Qixin Hu, Yixiong Chen, Junfei Xiao, Shuwen Sun, Jieneng Chen, Alan L. Yuille, and Zongwei Zhou. Label-free liver tumor segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7422–7432, 2023.

Fabian Isensee, Paul F. Jaeger, Simon A. A. Kohl, Jens Petersen, and Klaus H. Maier-Hein. nnU-Net: A self-configuring method for deep learning-based biomedical image segmentation. *Nature Methods*, 18:203–211, 2021. doi: 10.1038/s41592-020-01008-z.

Qiangguo Jin, Hui Cui, Changming Sun, Zhaopeng Meng, and Ran Su. Free-form tumor synthesis in computed tomography images via richer generative adversarial network. *Knowledge-Based Systems*, 218:106753, 2021. doi: 10.1016/j.knosys.2021.106753.

Davood Karimi and Septimiu E. Salcudean. Reducing the Hausdorff distance in medical image segmentation with convolutional neural networks. *IEEE Transactions on Medical Imaging*, 39 (2):499–513, 2020. doi: 10.1109/TMI.2019.2930068.

Hoel Kervadec, Jihene Bouchtiba, Christian Desrosiers, Eric Granger, Jose Dolz, and Ismail Ben Ayed. Boundary loss for highly unbalanced segmentation. In *Proceedings of the 2nd International Conference on Medical Imaging with Deep Learning*, volume 102 of *Proceedings of Machine Learning Research*, pp. 285–296, 2019.

Jonghun Kim, Inye Na, Eun Sook Ko, and Hyunjin Park. Tumor synthesis conditioned on radiomics. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3635–3646, 2025. doi: 10.1109/WACV61041.2025.00358.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, 2014.

Wenhui Lei, Hengrui Tian, Linrui Dai, Hanyu Chen, and Xiaofan Zhang. LesionDiffusion: Towards text-controlled general lesion synthesis. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume 15964 of *Lecture Notes in Computer Science*, pp. 327–336, 2025. doi: 10.1007/978-3-032-04971-1_31.

Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. RePaint: Inpainting using denoising diffusion probabilistic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11461–11471, 2022.

Fei Lyu, Jingwen Xu, Ye Zhu, Grace Lai-Hung Wong, and Pong C. Yuen. Enhancing lesion segmentation via medical image-mask pair synthesis using phenotype-conditioned diffusion models. *IEEE Journal of Biomedical and Health Informatics*, 2026. doi: 10.1109/JBHI.2026.3695075.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10684–10695, 2022.

Saikat Roy, Gregor Koehler, Constantin Ulrich, Michael Baumgartner, Jens Petersen, Fabian Isensee, Paul F. Jaeger, and Klaus H. Maier-Hein. MedNeXt: Transformer-driven scaling of ConvNets for medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, 2023. doi: 10.1007/978-3-031-43901-8_39.

Ozan Sener and Vladlen Koltun. Multi-task learning as multi-objective optimization. In *Advances in Neural Information Processing Systems*, volume 31, pp. 525–536, 2018.

Jiajun Sun, Zhen Yu, Siyuan Yan, Jason J. Ong, Zongyuan Ge, and Lei Zhang. Controllable skin synthesis via lesion-focused vector autoregression model. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, pp. 129–139, 2025. doi: 10.1007/978-3-032-05325-1_13.

Tianhe Yu, Saurabh Kumar, Abhishek Gupta, Sergey Levine, Karol Hausman, and Chelsea Finn. Gradient surgery for multi-task learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 5824–5836, 2020.

Hantao Zhang, Yuhe Liu, Jiancheng Yang, Shouhong Wan, Xinyuan Wang, Wei Peng, and Pascal Fua. LeFusion: Controllable pathology synthesis via lesion-focused diffusion models. In *International Conference on Learning Representations*, pp. 13232–13253, 2025a.

Sheng Zhang, Yanbo Xu, Naoto Usuyama, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Matthew P. Lungren, Tristan Naumann, and Hoifung Poon. Large-scale domain-specific pretraining for biomedical vision-language processing. *arXiv preprint arXiv:2303.00915*, 2023.

Zheyuan Zhang, Lanhong Yao, Bin Wang, Debesh Jha, Gorkem Durak, Elif Keles, Alpay Medetalibeyoglu, and Ulas Bagci. DiffBoost: Enhancing medical image segmentation via text-guided diffusion model. *IEEE Transactions on Medical Imaging*, 44(9):3670–3682, 2025b. doi: 10.1109/TMI.2024.3519307.

A Detailed Experimental Setup

Datasets. For LIDC-IDRI (Armato et al., 2011), our ROI collection contains 2,624 nodules from 875 patients. The fixed partition comprises 2,104 ROIs from 700 patients for training and 520 ROIs from 175 patients for testing, without patient overlap. The synthesis loader excludes samples in the fixed test list. Following the LIDC data preparation protocol of LeFusion (Zhang et al., 2025a), we additionally use 3,076 normal ROIs from 135 healthy patients, sampled from anatomical regions in which pulmonary nodules commonly occur, as backgrounds for lesion synthesis. For KiTS23 (Heller et al., 2023), we use renal tumors as target lesions; renal cysts are not merged into their masks. We extract 26-connected tumor components and center each ROI on the component's bounding box. The pipeline produces 567 ROIs from 489 cases, of which 13 are excluded because fewer than eight tumor voxels remain after resampling. The case-level partition, constructed with random seed 2026, leaves 443 training ROIs from 391 cases, 53 validation ROIs from 48 cases, and 58 test ROIs from 49 retained cases. This is an internal ROI partition, not the official challenge test set. We additionally extract 507 normal ROIs from non-tumor regions to provide backgrounds for renal tumor synthesis.

Preprocessing. Synthesis inputs have $32 \times 64 \times 64$ voxels for both datasets. Intensities are clipped to $[-1000, 400]$ HU for LIDC and $[-200, 400]$ HU for KiTS23, then linearly mapped to $[-1, 1]$. LIDC ROIs retain their stored, case-dependent voxel spacing without additional resampling in the synthesis loader. KiTS23 ROIs are resampled to $5.0 \times 2.5 \times 2.5$ mm in array-axis order, using linear interpolation for CT and nearest-neighbor interpolation for labels. Equal input dimensions therefore do not imply equal physical resolution. The voxel spacing associated with each ROI is passed to TA-BTM during training. Boundary rings are constructed in physical space, and axis-specific finite differences are normalized by the corresponding spacing. Consequently, transition width is measured in millimeters rather than voxels.

Image–mask pairs undergo synchronized random flips during synthesis training. Structured prompts describe lesion type and internal appearance (Section 3.1); renal tumor type denotes attenuation relative to kidney parenchyma, rather than histological diagnosis.

Evaluation protocol. We use nnU-Net (Isensee et al., 2021), SwinUNETR (Hatamizadeh et al., 2022), and MedNeXt (Roy et al., 2023) as downstream segmentation models, distinct from the auxiliary decoder used during KG-MTS training. We report Dice similarity coefficient (Dice) and normalized surface Dice (NSD) as percentages. Dice measures volumetric overlap, while NSD measures surface agreement within a tolerance. Higher values are better; improvements are absolute differences between the reported percentage scores. Within each run, Dice and NSD are computed separately for each test ROI and then averaged over ROIs. Each experimental configuration is run three times, and the tables report the mean of the three run-level scores.

Implementation details. The generation models are trained on four NVIDIA H20 GPUs with a batch size of 32. We use the Adam optimizer, a diffusion process with 300 timesteps and a cosine noise schedule, and an exponential moving average with decay 0.995. The biomedical text encoder remains trainable. Training follows the two-stage procedure in Section 3.4: prompt-conditioned pretraining followed by training with the proposed auxiliary objectives. For downstream augmentation, we combine all available real training samples with all synthetic training samples, while Real Only uses only the real training samples. For all synthesis-based methods, we generate exactly N synthetic training ROIs using the same target backgrounds and prescribed masks, and maintain the same synthetic-to-real ratio for downstream segmentation training.

Descriptor evaluation. In addition to downstream segmentation, we directly evaluate whether the synthesized lesions exhibit measurable characteristics consistent with real lesions. To isolate the effect of each proposed component, we compare the Prompt-only checkpoint with independently trained KG-MTS and TA-BTM variants. Each checkpoint synthesizes 300 LIDC-IDRI samples using matched backgrounds, masks, prompts, and initial noise, with 50 samples from each of the six type–pattern prompt groups. The reference set contains all 520 held-out real test lesions.

We report normalized interval deviation (NID) for the three medical descriptors used by KG-MTS (mean HU, soft solid-component ratio, and local variance) and the three boundary descriptors used

by TA-BTM (signed contrast, transition width, and transition smoothness). Before normalization, transition width is measured in millimeters, while the boundary derivatives are computed with respect to physical distance. NID is dimensionless because each deviation is normalized by the IQR of the corresponding real-lesion distribution.

For each prompt group, the valid real-lesion interval is defined by the 10th and 90th percentiles of the held-out real distribution. A synthetic value outside this interval is penalized by its squared distance to the nearest bound, normalized by the interquartile range of the corresponding real group. Values within the interval receive zero penalty. NID is averaged over all synthetic samples; lower values indicate better agreement with clinically plausible real-lesion ranges.

Comparison methods. Our comparators include Real Only (no synthetic augmentation), Hand-Crafted, Copy-Paste, Cond-Diffusion, DiffTumor, and LeFusion. Hand-Crafted uses rule-based lesion generation (Hu et al., 2023), while Copy-Paste transfers real lesions to target backgrounds. Cond-Diffusion uses mask-conditioned diffusion (Ho et al., 2020; Rombach et al., 2022) and is evaluated separately from DiffTumor, a generalizable conditional tumor generator (Chen et al., 2024). LeFusion focuses diffusion on the prescribed lesion region (Zhang et al., 2025a).