# Semantic-Aware Organ-Level Esophageal Tumor Synthesis via Latent Rectified Flow

Qinji Yu[1,2,3⋆✉], Guangyu Guo[1,2⋆], Jiawen Yao[1,2⋆], Zhilin Zheng[1,2⋆], Tiancheng Lin[1,2⋆], Yingda Xia[1,2], Yi Guo[3], Qifeng Wang[4], Jian Zhou[5], and Ling Zhang[1,2✉]

[1]DAMO Academy, Alibaba Group, Hangzhou, China
[2]Hupan Lab, 310023 Hangzhou, China
[3]Fudan University, Shanghai, China
[4]Sichuan Cancer Hospital, Chengdu, China
[5]Sun Yat-sen University Cancer Center, Guangzhou, China
cscyqj@gmail.com, fabiozhang0722@gmail.com

**Abstract.** Generative models offer a promising solution to the data scarcity and annotation bottleneck of tumors, particularly early-stage cases. However, existing methods are predominantly designed for solid organs, relying on naively scaling local lesion textures to simulate tumor progression. This paradigm is unsuitable for hollow organs like the esophagus, where the tumor stage is dictated by the depth of invasion rather than absolute tumor size. Moreover, because esophageal tumors inherently induce complex organ-level morphological changes, realistic synthesis demands a paradigm shift from localized inpainting to comprehensive organ-level editing. To bridge this gap, we propose the Semantic-aware Organ-level Flow for Tumor synthesis (SOFT) framework. Built upon Latent Rectified Flow and conditioned on semantic embeddings of clinical attributes (e.g., T-stage) through a language model, SOFT simultaneously deforms the organ structure and synthesizes realistic tumors. Crucially, to resolve the label shift induced by organ-level editing, we introduce a synchronous segmentation head co-trained with the flow model, dynamically yielding perfectly aligned image-label pairs. Extensive experiments demonstrate that augmenting real training cohorts with our synthetic samples consistently enhances overall downstream performance, while significantly boosting the detection sensitivity for early-stage (T1) tumors on an external cohort from 63.3% to 71.1% ($p = 0.03$).

**Keywords:** Esophageal tumor synthesis · Rectified flow · Hollow organs

## 1 Introduction

Deep learning has driven remarkable progress in automated tumor recognition [24,1,18,3,8,16,25]. In practice, training robust models heavily relies on large-scale, densely annotated datasets, especially for early-stage cancers that are crucial for timely intervention. However, early-stage lesions typically exhibit

⋆ Equal contribution.

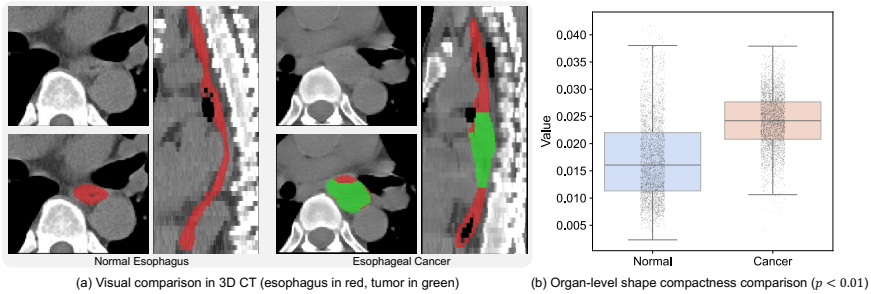


Fig. 1: **Morphological deformations induced by esophageal cancer (EC).** (a) 3D CT visualization illustrating irregular wall thickening and mass effects compared to normal anatomy. (b) Statistical confirmation of significant shifts in organ-level shape compactness ($p < 0.01$).

highly subtle and ambiguous imaging features, masking the manual annotation substantially challenging, thus limits the scale of available training data.

Generative models [11,6,7,22,17] offer a promising solution to this data scarcity by synthesizing artificial lesions to augment training datasets [12,9,4,23,26,15,20,14]. However, most existing 3D CT generative methods are primarily designed for solid organs such as the liver, kidney, or brain [9,23,4,26,14]. These approaches commonly formulate tumor synthesis as a lesion-only texture inpainting task within a predefined mask. Consequently, previous works control tumor progression and clinical staging by simply scaling the absolute size of the lesion mask [9,4,14], leaving the underlying organ morphology largely unaffected.

Unlike solid organ tumors, esophageal cancer (EC) staging is clinically defined by the depth of wall invasion rather than absolute tumor size [13,5,21]. This necessitates modeling profound morphological deformations—such as asymmetric wall thickening and mass effects compressing (Fig. 1(a)). We quantitatively validate these structural transformations through a statistical analysis of organ-level shape compactness, revealing a significant shift between normal and cancerous esophagus ($p < 0.01$, Fig. 1(b)). Such evidence quantitatively mandates a paradigm shift from lesion-level synthesis to comprehensive organ-level editing to accurately reflect the geometric transformations of EC, particularly for early-stage case.

To bridge this gap, we propose **S**emantic-aware **O**rgan-level **F**low for **T**umor synthesis (**SOFT**), a unified generative framework tailored for esophageal tumor synthesis. Built upon Latent Rectified Flow that offers faster training convergence and substantially accelerated inference compared to latent diffusion models (LDM) [19,28], our method simultaneously generates tumors, deforms the esophageal structure, and preserves the surrounding anatomical context. To facilitate the synthesis of early-stage esophageal cancer cases, the generative process is explicitly controlled by structured clinical attributes (e.g., *T-stage* and *tumor location*), enabling fine-grained semantic manipulation of tumor extent and invasion locations. Crucially, organ-level editing induces a label shift dilemma,

where pre-defined anatomical masks become spatially invalid as the model deforms the underlying organ structure. Unlike static inpainting, this morphological reshaping makes dynamic label generation a logical necessity rather than an auxiliary feature. To resolve this, we introduce a synchronous segmentation head co-trained with the flow model. This unified design compels the network to dynamically delineate the boundaries of the anatomy it synthesizes, yielding perfectly aligned image-label pairs for robust training.

Our main contributions are three-fold: (1) We identify the unique morphological challenges of esophageal tumor synthesis and shift the prevailing paradigm from lesion-level texture inpainting to organ-level editing. (2) We propose a semantic-aware Latent Rectified Flow framework with a synchronous segmentation head, effectively resolving the label shift problem by unifying image synthesis and semantic prediction. (3) Extensive experiments on two large-scale esophageal CT datasets demonstrate that SOFT generates highly realistic image–label pairs and substantially improves downstream segmentation performance, particularly for early-stage tumor detection.

## 2 Methods

We propose a semantic-aware, organ-level generative framework for esophageal tumor synthesis that explicitly models morphological changes of the esophagus. Fig. 2 provides an overview of the proposed framework.

### 2.1 Problem Formulation

We formulate esophageal tumor synthesis as a conditional inpainting task with joint label generation. Given a 3D CT patch $\mathbf{x} \in \mathbb{R}^{H\times W\times D}$, an inpainting mask $\mathbf{M}$ covering the esophagus and part of the surrounding context, and structured clinical attributes $\mathbf{A}$, the goal is to synthesize the masked region $\mathbf{x} \odot \mathbf{M}$ such that it conforms to the morphological and semantic characteristics specified by the attributes $\mathbf{A}$, while the unmasked region remains unchanged. In parallel, the segmentation map $\mathbf{S}$ is learned to delineate the anatomical structures implied by $\mathbf{A}$, providing paired labels that accurately reflect the synthesized and potentially deformed anatomy.

### 2.2 Latent Rectified Flow for Anatomy Generation

To enable high-fidelity and efficient 3D synthesis, we adopt a Latent Rectified Flow framework [28]. Unlike standard LDMs based on stochastic differential equations, Rectified Flow [19] learns a deterministic Ordinary Differential Equation (ODE) that transports Gaussian noise to the target latent distribution along straight trajectories.

**Latent Space Compression** Following LesionDiffusion [15], we employ a pre-trained VQ-VAE to compress 3D CT patches into a low-dimensional latent space, reducing computational overhead. During training, the encoder $\mathcal{E}$ maps

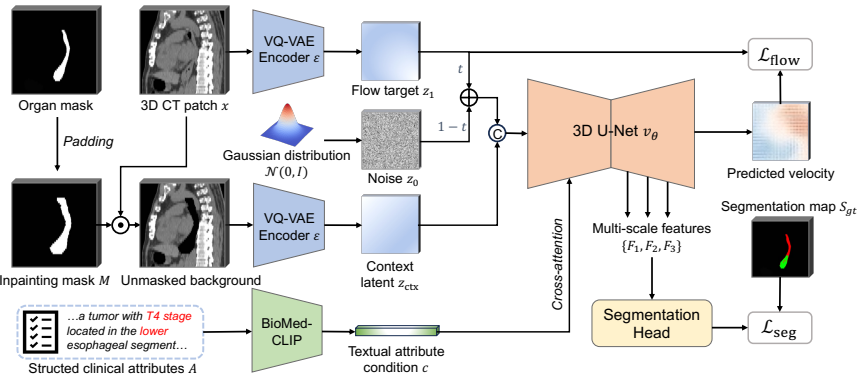


Fig. 2: Overview of the proposed SOFT framework. Conditioned on text-encoded clinical attributes and context-preserving latent features, the Latent Rectified Flow model synthesizes morphological deformations characteristic of EC. Concurrently, the multi-scale intermediate features from the 3D U-Net drive a Semantic-Aware Segmentation Head, dynamically generating updated masks to resolve the label shift induced by structural editing.

the complete input $\mathbf{x}$ to its latent flow target $\mathbf{z}_1 = \mathcal{E}(\mathbf{x})$, while the decoder $\mathcal{D}$ reconstructs the synthesized image as $\hat{\mathbf{x}} = \mathcal{D}(\hat{\mathbf{z}}_1)$. Simultaneously, to provide structural context for inpainting, the unmasked background is encoded into a context latent $\mathbf{z}_{\text{ctx}} = \mathcal{E}(\mathbf{x} \odot (1 - \mathbf{M}))$. All generative processes are performed exclusively within this efficient latent space.

**Flow Matching Objective** We define a linear interpolation between a noise sample $\mathbf{z}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and the target latent representation $\mathbf{z}_1$:

$$\mathbf{z}_t = t\mathbf{z}_1 + (1-t)\mathbf{z}_0, \quad t \in [0, 1]. \tag{1}$$

The generative model $v_\theta$, parameterized by a 3D U-Net, is trained to predict the velocity field of this transport. Specifically, $v_\theta$ takes the current noisy latent $\mathbf{z}_t$, the context latent $\mathbf{z}_{\text{ctx}}$, the timestep $t$, and the textual attribute condition $\mathbf{c}$ as inputs. Since the task is formulated as conditional inpainting, the flow matching loss is computed only within the masked region. Let $\mathbf{M}_z$ denote the inpainting mask downsampled to the latent spatial resolution. The objective minimizes the squared error between the predicted velocity and the GT direction $(\mathbf{z}_1 - \mathbf{z}_0)$ inside the mask:

$$\mathcal{L}_{\text{flow}} = \mathbb{E}_{t,\mathbf{z}_0,\mathbf{z}_1}\left[\left\|(v_\theta(\mathbf{z}_t, \mathbf{z}_{\text{ctx}}, t, \mathbf{c}) - (\mathbf{z}_1 - \mathbf{z}_0)) \odot \mathbf{M}_z\right\|_2^2\right]. \tag{2}$$

In practice, the context latent $\mathbf{z}_{\text{ctx}}$ is concatenated channel-wise with the noisy latent $\mathbf{z}_t$ to enforce anatomical continuity between the synthesized region and the preserved surrounding structures. Meanwhile, the structured clinical attributes $\mathbf{A}$ are encoded via BioMedCLIP [27] into text embeddings $\mathbf{c}$, which are injected into the U-Net through cross-attention layers to guide the generation toward attribute-consistent tumor characteristics.

### 2.3 Semantic-Aware Joint Segmentation Head

To yield accurate labels matching the synthesized anatomy, we introduce a semantic-aware segmentation head co-trained with the generative model. By dynamically delineating the generated tumor and deformed organ structures, this branch enforces semantic consistency during synthesis and provides valid supervision for downstream tasks.

**Segmentation Head** Following [2], we extract multi-scale features from intermediate decoder layers of the 3D U-Net instead of the final output. Feature maps $\{F_1, F_2, F_3\}$ from three middle decoder blocks are upsampled to a common resolution, concatenated, and fed into a lightweight convolutional head to predict voxel-wise probability maps $\mathbf{S}$ for background, esophagus, and tumor classes.

**Semantic Consistency Loss** To enforce semantic alignment, we compute a segmentation loss $\mathcal{L}_{\text{seg}}$ between the predicted probability maps $\mathbf{S}$ and the ground-truth $\mathbf{S}_{\text{gt}}$. Since intermediate activations in generative models only form coherent semantic structures as they approach the target data distribution [2], we scale this loss directly by the timestep $t$. This prevents penalizing the network when predicting masks from noise-dominated latents:

$$\mathcal{L}_{\text{seg}} = t \cdot \left(\mathcal{L}_{\text{Dice}}(\mathbf{S}, \mathbf{S}_{\text{gt}}) + \mathcal{L}_{\text{CE}}(\mathbf{S}, \mathbf{S}_{\text{gt}})\right). \tag{3}$$

This auxiliary objective encourages the latent features to encode anatomically meaningful boundaries as the image structure emerges, ensuring the synthesized anatomy is both realistic and separable.

The overall training objective is defined as a combination of the flow matching loss and the semantic consistency loss:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{flow}} + \mathcal{L}_{\text{seg}}. \tag{4}$$

### 2.4 Inference and Controllable Co-Synthesis

For a given normal CT template, the inpainting mask $\mathbf{M}$ is constructed slice-wise: on each axial slice we first take the 2D bounding box of the esophagus mask, then apply outward padding to leave room for tumor-induced deformation, and finally stack these padded 2D masks axially to form the 3D mask $\mathbf{M}$. We avoid morphological dilation, which would leak the original esophageal contour into $\mathbf{M}$ and contradict organ-level editing, as well as only local masking around a presumed tumor site, which would fix the tumor location and weaken text-based control over T-stage and anatomy. We then initialize the generative process with Gaussian noise $\mathbf{z}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and solve the learned ODE up to $t = 1$, applying classifier-free guidance (CFG) to enforce strict adherence to the random sampled clinical attributes $\mathbf{A}$. The predicted latent $\hat{\mathbf{z}}_1$ is decoded to yield the synthesized volume $\hat{\mathbf{x}} = \mathcal{D}(\hat{\mathbf{z}}_1)$. To preserve the unmasked background anatomy, the final image is composited with the original context:

$$\hat{\mathbf{x}}_{\text{final}} = \hat{\mathbf{x}} \odot \mathbf{M} + \mathbf{x} \odot (1 - \mathbf{M}). \tag{5}$$

Concurrently, the synchronized segmentation masks $\mathbf{S}_{\text{final}}$ are directly extracted from the segmentation head at $t = 1$, completing the unified co-synthesis process.

# 3 Experiments

## 3.1 Experimental Setup

**Dataset:** We collected EC CT datasets from two medical centers, including 2,716 cases from Center 1 and 1,028 cases from Center 2. Each CT scan is accompanied by corresponding endoscopic reports and surgical pathology reports. Tumor masks were manually annotated by two radiologists with more than three years of experience. From Center 1, 80% of the EC cases were used for training the generative model and the remaining 20% for testing, resulting in 2,164 training samples and 552 testing samples. In addition, 20,622 normal chest CT scans were collected from a third center and used as anatomical templates for synthetic EC case generation. The EC dataset from Center 2 was treated as an external dataset and exclusively used to evaluate the generalization performance of downstream segmentation models.

**Implementation Details:** Our framework is trained in two sequential stages. First, the VQ-VAE is trained for 100 epochs (learning rate $1 \times 10^{-5}$) with $2 \times 2 \times 1$ spatial downsampling and a 4,096-entry codebook. The VQ-VAE weights are subsequently frozen. Second, the Latent Rectified Flow model and the synchronous segmentation head are jointly optimized for 1,000 epochs (learning rate $1 \times 10^{-4}$) using 1,000 discretized timesteps. This phase is distributed across 8 NVIDIA H20 GPUs with a batch size of 4 per GPU. At inference, we solve the ODE using 50 steps and a CFG scale of 2.0. For controllable synthesis, each template case is guided by a randomly sampled combination of clinical attributes, specifically the *T-stage* (`T1`–`T4`) and *tumor location* (`Upper`, `Middle`, `Lower`, `Esophago-Gastric Junction`).

For downstream segmentation tasks, we train a separate segmentation network using the official nnU-Netv2 [10] implementation. Segmentation performance is evaluated under three training settings: (1) training on real EC data from the Center 1 training split; (2) training on synthetic EC cases generated from normal CT templates; and (3) training on a combination of synthetic and real EC training data. The same experimental settings are applied to all synthesis baselines for fair comparison. Evaluation is conducted on the Center 1 test set and the external Center 2 dataset using segmentation metrics, including Dice Similarity Coefficient (DSC) and Normalized Surface Dice (NSD), as well as detection performance measured by patient-level sensitivity (defined as detected if DSC $> 0.1$).

**Comparing Methods:** We compare our method against two state-of-the-art diffusion-based inpainting baselines, **LeFusion** [26] and **LesionDiffusion** [15], both of which follow a two-stage pipeline that first generates a lesion mask and then synthesizes lesion texture conditioned on the generated mask.

In addition to the full model (**SOFT**), we evaluate three variants to isolate the contribution of key components: (i) **SOFT (LDM)** replaces Rectified Flow with a latent DDPM while keeping the remaining architecture and conditioning unchanged; (ii)**SOFT (w/o attr.)** removes the text-based clinical attributes

Table 1: Comparison of downstream segmentation performance. Sen. T1 specifically denotes detection sensitivity for early-stage (T1) tumors. **Bold** numbers indicate the best performance in each setting.

| Methods | Internal Test (Center 1) | | | | External Test (Center 2) | | | |
|---|---|---|---|---|---|---|---|---|
| | Sen. All (%)↑ | Sen. T1 (%)↑ | DSC (%)↑ | NSD (%)↑ | Sen. All (%)↑ | Sen. T1 (%)↑ | DSC (%)↑ | NSD (%)↑ |
| ***Baseline: Real EC*** | 92.57 (511/552) | 79.82 (91/114) | 72.81±26.07 | 77.04±26.25 | 85.12 (875/1028) | 63.33 (57/90) | 64.54±30.41 | 67.54±31.17 |
| ***Data: Synthetic EC*** | | | | | | | | |
| LeFusion [26] | 28.62 (158/552) | 15.79 (18/114) | 8.24±12.62 | 14.94±17.22 | 36.48 (375/1028) | 17.78 (16/90) | 11.85±16.02 | 18.15±18.99 |
| LesionDiffusion [15] | 27.72 (153/552) | 20.18 (23/114) | 9.95±16.69 | 14.54±20.38 | 52.43 (539/1028) | 35.56 (32/90) | 20.45±21.91 | 27.28±24.02 |
| SOFT (LDM) | 84.96 (469/552) | 66.67 (76/114) | 62.64±31.31 | 66.76±32.18 | 81.32 (836/1028) | 61.11 (55/90) | 57.81±30.73 | 60.38±31.65 |
| SOFT (w/o attr.) | 84.78 (468/552) | 68.42 (78/114) | 63.92±31.09 | 67.47±32.09 | 84.14 (865/1028) | **63.33** (57/90) | 60.22±29.82 | 62.96±30.74 |
| SOFT (w/o pad.) | 78.44 (433/552) | 50.88 (58/114) | 57.02±33.36 | 60.41±34.47 | 80.64 (829/1028) | 57.78 (52/90) | 54.86±30.69 | 57.31±31.62 |
| **SOFT (Ours)** | **90.22** (498/552) | **78.07** (89/114) | **66.39±28.00** | **71.15±28.43** | **86.09** (885/1028) | 62.22 (56/90) | **61.25±29.05** | **64.47±29.69** |
| ***Data: Real+Synthetic EC*** | | | | | | | | |
| LeFusion [26] | 93.30 (515/552) | 84.21 (96/114) | 71.89±25.26 | 76.73±25.33 | 84.44 (868/1028) | 61.11 (55/90) | 62.55±30.79 | 65.69±30.85 |
| LesionDiffusion [15] | 92.03 (508/552) | 77.19 (88/114) | 72.53±26.24 | 76.81±26.59 | 85.80 (882/1028) | 65.56 (59/90) | 64.31±30.15 | 67.47±30.56 |
| SOFT (LDM) | 93.66 (517/552) | 84.21 (96/114) | 73.14±24.96 | 77.67±24.91 | 86.87 (893/1028) | 68.89 (62/90) | 64.96±29.39 | 68.15±29.81 |
| SOFT (w/o attr.) | 92.93 (513/552) | 83.33 (95/114) | 72.84±25.45 | 77.44±25.27 | 86.58 (890/1028) | 64.44 (58/90) | 64.57±29.69 | 67.59±30.07 |
| SOFT (w/o pad.) | 92.75 (512/552) | 81.58 (93/114) | 72.66±25.59 | 76.96±25.64 | 85.99 (884/1028) | 61.11 (55/90) | 64.26±29.99 | 67.30±30.37 |
| **SOFT (Ours)** | **93.84** (518/552) | **85.09** (97/114) | **73.57±24.51** | **78.30±24.45** | **87.84** (903/1028) | **71.11** (64/90) | **65.31±28.81** | **68.49±29.21** |

from conditioning; (iii) **SOFT (w/o pad.)** removes organ-level padding, i.e., the inpainting region is restricted to the esophagus only.

### 3.2 Results

**Downstream Segmentation Performance:** Table 1 reports the downstream segmentation performance under different training settings on both the internal and external test sets. When trained solely on synthetic data, lesion-level synthesis methods (LeFusion and LesionDiffusion) yield limited detection sensitivity and poor segmentation accuracy, indicating that lesion-only texture inpainting is insufficient for generating clinically meaningful esophageal tumors. In contrast, SOFT trained on synthetic data alone achieves substantially higher sensitivity, DSC, and NSD on both test sets, reaching 86.09% overall sensitivity on the external dataset, which even surpasses the real-data-only baseline (85.09%). Moreover, SOFT consistently outperforms its ablated variants, highlighting the importance of semantic attribute conditioning and organ-level padding. When synthetic data are combined with real training data, SOFT further improves performance over the real-only baseline and all competing methods, increasing external overall sensitivity from 85.09% to 87.84% and, more notably, boosting external T1 sensitivity from 63.33% to 71.11%. Furthermore, SOFT achieves the best overall DSC and NSD across both test sets, demonstrating that our organ-level synthesis effectively enriches the training distribution with rare early-stage phenotypes and significantly enhances cross-center generalization.

**Data Efficiency:** Fig. 3 evaluates data efficiency by varying the proportion of real training data. Across all data regimes, models augmented with SOFT synthetic data consistently outperform those trained on real data alone. The performance gap is most pronounced in low-data settings, highlighting the ability of SOFT to effectively alleviate data scarcity. Notably, using only 50% of real training data augmented with SOFT synthetic samples achieves performance on par with 100% real data on the internal test set, and even surpasses it on the external dataset. As the amount of real data increases, the benefit of synthetic

augmentation gradually diminishes but remains consistently positive, indicating that SOFT does not introduce bias in data-rich scenarios.

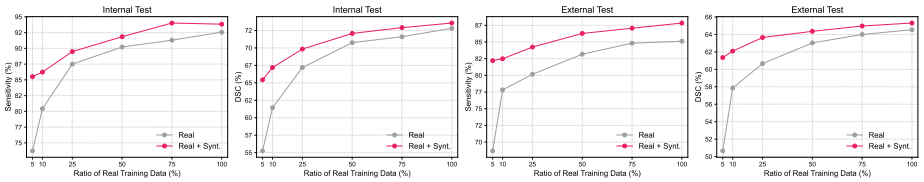


Fig. 3: Data efficiency analysis showing the downstream performance on the internal and external test sets. The models are trained with varying proportions of real data (10% to 100%), comparing the baseline (Real data only) against our augmented strategy (Real + Synthetic data from SOFT).

**Clinical Stratification and Semantic Control:** To evaluate the clinical utility and generative fidelity of SOFT, we conduct subgroup analyses stratified by tumor T-stage with qualitative evaluations of semantic control (Fig. 4). Quantitatively, models trained solely on real data already achieve near-saturated sensitivities for advanced-stage tumors (>91.0% for T2, >95.0% for T3-4), suggesting that the primary clinical challenge—and the greatest potential for synthetic augmentation—resides in early-stage (T1) detection. SOFT achieves significant performance gains in T1 tumors on both internal and external testsets ($p < 0.05$), validating the framework's ability to bridge the gap in early-stage detection. Qualitatively, visual examples demonstrate the framework's capacity to translate textual clinical attributes into anatomically distinct phenotypes, ensuring precise alignment between the synthesized EC and generated labels.

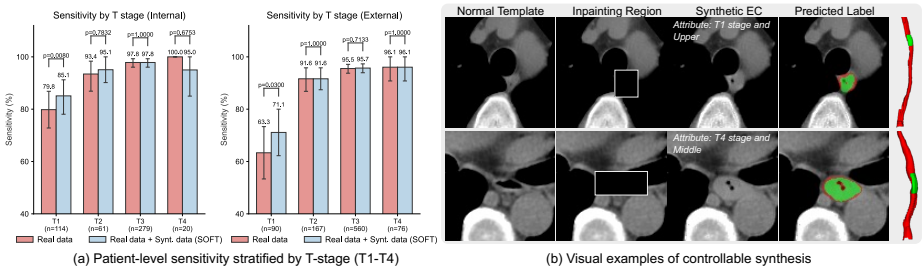


Fig. 4: **Evaluation of clinical utility and semantic control.** (a) Quantitative results of supplementing real data with SOFT synthetics on T1-4. (b) SOFT accurately translates clinical attributes into distinct tumor phenotypes (e.g., subtle T1 lesions vs. advanced T4 mass effects).

## 4 Conclusion

We presented SOFT, a semantic-aware generative framework that shifts the paradigm of esophageal tumor synthesis from localized texture inpainting to comprehensive organ-level editing. By unifying a Latent Rectified Flow model with a synchronous segmentation head, SOFT simultaneously synthesizes realistic morphological deformations and dynamically yields perfectly aligned image-label pairs. Extensive evaluations demonstrate that augmenting training cohorts with these synthetic samples significantly improves downstream performance, offering a highly effective solution to the data scarcity of early-stage (T1) tumors.

**Disclosure of Interests.** The authors have no competing interests to declare that are relevant to the content of this article.

## References

1. Antonelli, M., Reinke, A., Bakas, S., Farahani, K., Kopp-Schneider, A., Landman, B.A., Litjens, G., Menze, B., Ronneberger, O., Summers, R.M., et al.: The medical segmentation decathlon. Nature communications **13**(1), 4128 (2022)
2. Baranchuk, D., Voynov, A., Rubachev, I., Khrulkov, V., Babenko, A.: Label-efficient semantic segmentation with diffusion models. In: ICLR (2022)
3. Cao, K., Xia, Y., Yao, J., Han, X., Lambert, L., Zhang, T., Tang, W., Jin, G., Jiang, H., Fang, X., et al.: Large-scale pancreatic cancer detection via non-contrast ct and deep learning. Nature medicine **29**(12), 3033–3043 (2023)
4. Chen, Q., Chen, X., Song, H., Xiong, Z., Yuille, A., Wei, C., Zhou, Z.: Towards generalizable tumor synthesis. In: CVPR. pp. 11147–11158 (2024)
5. Gehrung, M., Crispin-Ortuzar, M., Berman, A.G., O'Donovan, M., Fitzgerald, R.C., Markowetz, F.: Triage-driven diagnosis of barrett's esophagus for early detection of esophageal adenocarcinoma using deep learning. Nature medicine **27**(5), 833–841 (2021)
6. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial networks. Communications of the ACM **63**(11), 139–144 (2020)
7. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. NeurIPS **33**, 6840–6851 (2020)
8. Hu, C., Xia, Y., Zheng, Z., Cao, M., Zheng, G., Chen, S., Sun, J., Chen, W., Zheng, Q., Pan, S., et al.: Ai-based large-scale screening of gastric cancer from noncontrast ct imaging. Nature Medicine pp. 1–9 (2025)
9. Hu, Q., Chen, Y., Xiao, J., Sun, S., Chen, J., Yuille, A.L., Zhou, Z.: Label-free liver tumor segmentation. In: CVPR. pp. 7422–7432 (2023)
10. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods **18**(2), 203–211 (2021)
11. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)
12. Koetzier, L.R., Wu, J., Mastrodicasa, D., Lutz, A., Chung, M., Koszek, W.A., Pratap, J., Chaudhari, A.S., Rajpurkar, P., Lungren, M.P., et al.: Generating synthetic data for medical imaging. Radiology **312**(3), e232471 (2024)

13. Lagergren, J., Smyth, E., Cunningham, D., Lagergren, P.: Oesophageal cancer. The Lancet **390**(10110), 2383–2396 (2017)
14. Lai, Y., Chen, X., Wang, A., Yuille, A., Zhou, Z.: From pixel to cancer: Cellular automata in computed tomography. In: MICCAI. pp. 36–46. Springer (2024)
15. Lei, W., Tian, H., Dai, L., Chen, H., Zhang, X.: Lesiondiffusion: Towards text-controlled general lesion synthesis. In: MICCAI. pp. 327–336. Springer (2025)
16. Liang, Z., Niu, Q., Wang, J., Han, C., Li, Q., Wu, Y., Zhu, B., Han, X., Wang, Z., Wang, X., et al.: A foundation model for breast and lung cancer screening using non-contrast computed tomography. Nature Health pp. 1–13 (2026)
17. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. In: ICLR (2023)
18. Liu, J., Zhang, Y., Chen, J.N., Xiao, J., Lu, Y., A Landman, B., Yuan, Y., Yuille, A., Tang, Y., Zhou, Z.: Clip-driven universal model for organ segmentation and tumor detection. In: CVPR. pp. 21152–21164 (2023)
19. Liu, X., Gong, C., Liu, Q.: Flow straight and fast: Learning to generate and transfer data with rectified flow. arXiv preprint arXiv:2209.03003 (2022)
20. Mao, J., Wang, Y., Tang, Y., Xu, D., Wang, K., Yang, Y., Zhou, Z., Zhou, Y.: Medsegfactory: Text-guided generation of medical image-mask pairs. In: ICCV. pp. 21525–21535 (2025)
21. Sheikh, M., Roshandel, G., McCormack, V., Malekzadeh, R.: Current status and future prospects for esophageal cancer. Cancers **15**(3), 765 (2023)
22. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: ICLR (2021)
23. Wu, L., Zhuang, J., Zhou, Y., He, S., Ma, J., Luo, L., Wang, X., Ni, X., Zhong, X., Wu, M., et al.: Large-scale generative tumor synthesis in computed tomography images for improving tumor recognition. Nature Communications **16**(1), 11053 (2025)
24. Yan, K., Wang, X., Lu, L., Summers, R.M.: Deeplesion: automated mining of large-scale lesion annotations and universal lesion detection with deep learning. Journal of medical imaging **5**(3), 036501–036501 (2018)
25. Yao, J., Ye, X., Xia, Y., Zhou, J., Shi, Y., Yan, K., Wang, F., Lin, L., Yu, H., Hua, X.S., et al.: Effective opportunistic esophageal cancer screening using noncontrast ct imaging. In: MICCAI. pp. 344–354. Springer (2022)
26. Zhang, H., Liu, Y., Yang, J., Wan, S., Wang, X., Peng, W., Fua, P.: Lefusion: Controllable pathology synthesis via lesion-focused diffusion models. In: ICLR (2025)
27. Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., et al.: Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. arXiv preprint arXiv:2303.00915 (2023)
28. Zhao, C., Guo, P., Yang, D., Tang, Y., He, Y., Simon, B., Belue, M., Harmon, S., Turkbey, B., Xu, D.: Maisi-v2: Accelerated 3d high-resolution medical image synthesis with rectified flow and region-specific contrastive loss. arXiv preprint arXiv:2508.05772 (2025)