This supplementary material provides additional details not included in the main paper due to space constraints. In Sec. A, we provide detaileds about the four datasets we employed. Sec. B includes information on training details and network hyperparameters. Sec. C details the features used as conditions in our model. Sec. D represents additional results for the rare sample generation. Sec. E describes future works. Finally, Sec. F details the training and inference algorithm.

A. Datasets

Kidney: We utilize the 2023 Kidney and Kidney Tumor Segmentation Challenge (KiTS23) dataset [19]. KiTS23 is a challenge dataset designed for segmenting kidneys and kidney tumors in CT scans. In our study, we used 406 tumor subjects out of 489 in the provided challenge training dataset, splitting them into 325 for training and 81 for the test set. During training, the 3D patch volume was cropped from the center of the tumor mask and patch volume size of $112 \times 112 \times 96$ was employed. We clipped the intensities to a range of [-175, 250] Hounsfield unit (HU) and further normalized them to [0,1].

Lung: We employ the non-small cell lung cancer (NSCLC) dataset [1]. This dataset consists of CT scans. In our study, we utilized 417 tumor subjects out of the provided training dataset of 422, dividing them into 334 for training and 83 for the test set. During training, the 3D patch volume was cropped from the center of the tumor mask and volume size of $112 \times 112 \times 80$ was used. The HU range of [-1000, 1000] was applied and normalized to [0,1].

Breast: We utilize a private dataset, which is a Dynamic Contrast Enhancement MRI (DCE-MRI) dataset. MRI scans were conducted using either a 1.5-T or a 3.0-T scanner from Philips. The scans included axial imaging with one pre-contrast and six post-contrast dynamic series. Contrast-enhanced images were acquired at 0.5, 1.5, 2.5, 3.5, 4.5, and 5.5 minutes after the contrast injection. We used the images from 0.5 minutes after contrast injection. The dataset comprises 110 breast cancer patients, of which 88 were used for the training set and 22 for the test set. During training, the 3D patch volume was cropped from the center of the tumor mask and patch volume size of $112 \times 112 \times 96$ with nonzero normalization (normalization excluding zero values, see Monai [3] framework) for MRI images was employed.

Brain: We utilize The Brain Tumor Segmentation Challenge 2021 (BraTS 2021) dataset [5, 6, 42], which includes multimodal MRI scans such as T1-weighted (T1), contrast-enhanced T1-weighted (T1ce), T2-weighted (T2), and Fluid Attenuated Inversion Recovery (FLAIR). Among these, T1ce, a modality enhanced with contrast agents to better visualize tumors, was employed in our study. From the BraTS 2021 training set, we utilized data from 1204 out of 1251 subjects, allocating 1000 for training and 204 for testing. During training, 3D patch volumes were centered around the enhancing-tumor mask, with a patch volume size of $112 \times 112 \times 96$ with nonzero normalization was applied to the MRI images.

[3] https://monai.io/

B. Implementation Details

B.1. Training Details

The network was trained using the Adam [34] optimizer with a learning rate set to 5×10^{-6}. The training was conducted on four A100 80GB GPUs with a batch size of one per GPU. The model was built using PyTorch version 1.13.1.

B.2. Comparison Methods

B.2.1 GAN-based Model.

The GAN-based model is adapted by modifying the 3D image translation model Ea-GAN [75] (based on pix2pix [25]), with the addition of cross-attention. The model architecture, based on the UNet [52] structure, is specified in Table A.

Stream	Cross Attn	Act.	Conv.	Norm.	Out ch.
In			C		64
DownBlock		LeakyReLU	C	IN	[128,256]
DownBlock	✓	LeakyReLU	C	IN	[256,512]
UPBlock	✓	ReLU	C^T	IN	[512, 256]
UPBlock		ReLU	C^T	IN	[256,128]
Out		ReLU	C^T	Tanh	1

Table A. Details of GAN-based Model. C is the convolution layer and C^T is the convolution transpose layer with $4 \times 4 \times 4$ kernel, $2 \times 2 \times 2$ stride, and $1 \times 1 \times 1$ padding. IN is the instance normalization layer. The Out layer uses tanh as the activation function to generate the final output.

B.2.2 Latent Diffusion Model.

LDM [51] is adapted to a 3D format from its original method for comparison. LDM consists of two components: a pretrained VQGAN [15] and a DDPM [20]. The detailed structure of this model is outlined in Table B. During training, the number of time steps for the diffusion process was set to 1000. In the inference stage, the DDIM [60] method was employed, utilizing 200 sampling steps.

Table B. Model structural details of LDM and BBDM used for the tumor texture generation task. $|\mathcal{Z}|$ represents codebook size in the latent space.

Model	z-shape	$\lvert\mathcal{Z}\rvert$	Training Steps	Noise schedule	Channel Multiplier	Channels	Model Size
LDM-f2	$56 \times 56 \times 48 \times 2$	2048	1000	linear	[1,4,8]	128	658.32M
BBDM-f2	$56 \times 56 \times 48 \times 2$	2048	1000	linear	[1,4,8]	128	681.00M

Table C. Model structure details of the GigaGAN used for the tumor shape generation task.

Model	**z** dim	**w** dim	G Channels	D Channels	G Attention Resolution	D Attention Resolution	Attention Type
Shape 24	128	512	512	512	[6,12]	[6,12]	self + cross
Shape 24→96	128	512	512	512	[12,24]	[12,24]	self + cross

Table D. Hyperparameters of the Exponential Moving Average (EMA) and ReduceLROnPlateau learning rate scheduler used in the training process.

	EMA Parameters			LR Scheduler Parameters					
Model	Start Step	Decay	Update Interval	Max lr	Min lr	Factor	patience	Cool Down	Threshold
BBDM-f2	30000	0.995	8	5.0e-6	5.0e-7	0.5	3000	3000	1.0e-4

B.3. Network hyperparameters

B.3.1 Tumor Shape Generator.

The Tumor Shape Generator is based on GigaGAN [28], which has shown successful results in generating images from text in 2D natural images. This generator adapts GigaGAN to a 3D format and uses a shape feature-to-image approach to create tumor masks. Additionally, the Tumor Shape Generator employs cascaded generation processes where images are initially generated at a resolution of $24 \times 24 \times 24$ and then upsampled to $96 \times 96 \times 96$. Detailed information about its structure can be found in Table C.

B.3.2 Tumor Texture Generator.

The Tumor Texture Generator is based on BBDM [37] and has been modified for 3D application. This generator comprises two stages: a pretrained VQGAN and a BBDM. The detailed structure of the model is specified in Table B. During the training phase, the number of time steps for the Brownian Bridge was established at 1000, whereas, in the inference stage, 200 sampling steps were utilized. The training parameters for BBDM are specified in Table D.

C. Radiomics Features

In our study, we extracted radiomics [2] features using PyRadiomics [4] [67]. We obtained shape, histogram, and texture features for model training, and the details of each feature are described in the subsection. Additionally, most feature definitions used in this study are provided in a Zwanenburg et al [80]. The features used are specified in Table E. Additionally, Figure B depicts the correlation matrix of each feature to illustrate the relationships between them.

[4] https://pyradiomics.readthedocs.io/

C.1. Shape feature

For our Tumor Shape Generation task, we utilized a total of 16 shape features for each type of ROI. Shape features quantitatively represent the physical form and structure of the ROI. These morphological characteristics, such as the size, shape, and orientation of tumors, are analyzed for diagnosing diseases and evaluating prognoses.

The main shape features include:

- Volume: Measures the overall volume of the tumor.
- Surface Area: Measures the surface area of the tumor.
- Sphericity: Indicates the ratio of the tumor's length and width and thus reflects how close the tumor is to being spherical.

During the experiment to manipulate tumor shape, we focused on changing the volume and sphericity. Volume is just the sum of voxels in mm. Sphericity is made of two components, volume and surface area. Thus, we manipulated volume and surface area to adjust the two features. Other shaped features that depended on volume and surface area were adjusted accordingly.

C.2. Texture feature

For our tumor texture generation task, we utilized a total of 74 features for each type of ROI, which included 18 histogram features, 24 GLCM features, 16 GLSZM features, and 16 GLRLM features. Detailed descriptions of each feature are provided in the subsections below.

Table E. **Radiomics features used in our study.** There are 16 shape, 18 histogram, 24 GLCM, 16 GLSZM, and 16 GLRLM features. The tumor shape generator utilizes 16 shape features for each region of interest (ROI) type, while the tumor texture generator employs 74 texture features for each ROI type.

Shape	**Histogram**	**GLCM**	**GLSZM**	**GLRLM**
Mesh Volume	Energy	Autocorrelation	Small Area Emphasis	Short Run Emphasis
Surface Area	Entropy	Joint Average	Large Area Emphasis	Long Run Emphasis
Surface Area to Volume ratio	Minimum	Cluster Prominence	Gray Level Non-Uniformity	Gray Level Non-Uniformity
Sphericity	Maximum	Cluster Shade	Gray Level Non-Uniformity Normalized	Gray Level Non-Uniformity Normalized
Compactness 1	Mean	Cluster Tendency	Size-Zone Non-Uniformity	Run Length Non-Uniformity
Compactness 2	Median	Contrast	Size-Zone Non-Uniformity Normalized	Run Length Non-Uniformity Normalized
Spherical Disproportion	10th percentile	Correlation	Zone Percentage	Run Percentage
Maximum 3D diameter	90th percentile	Difference Average	Gray Level Variance	Gray Level Variance
Maximum 2D diameter (Slice)	Interquartile Range	Difference Entropy	Zone Variance	Run Variance
Maximum 2D diameter (Column)	Range	Difference Variance	Zone Entropy	Run Entropy
Maximum 2D diameter (Row)	Mean Absolute Deviation	Joint Energy	Low Gray Level Zone Emphasis	Low Gray Level Run Emphasis
Major Axis Length	Robust Mean Absolute Deviation	Joint Entropy	High Gray Level Zone Emphasis	High Gray Level Run Emphasis
Minor Axis Length	Root Mean Squared	Informational Measure of Correlation 1	Small Area Low Gray Level Emphasis	Short Run Low Gray Level Emphasis
Least Axis Length	Standard Deviation	Informational Measure of Correlation 2	Small Area High Gray Level Emphasis	Short Run High Gray Level Emphasis
Elongation	Skewness	Inverse Difference Moment	Large Area Low Gray Level Emphasis	Long Run Low Gray Level Emphasis
Flatness	Kurtosis	Maximal Correlation Coefficient	Large Area High Gray Level Emphasis	Long Run High Gray Level Emphasis
	Variance	Inverse Difference		
	Uniformity	Inverse Difference Normalized		
		Inverse Difference Moment Normalized		
		Inverse Variance		
		Maximum Probability		
		Sum Average		
		Sum Entropy		
		Sum of Squares		

Histogram: Histogram features, also known as first-order features, are derived from the distribution of pixel or voxel intensities within the ROI. These features provide valuable insights into the texture of the tissue by analyzing the intensity histogram of the ROI.

The main histogram features include:

- Median: The middle-intensity value when all values are sorted.
- Skewness: The asymmetry in the intensity distribution.
- Energy: The sum of squared intensities, representing the magnitude of voxel values.
- Entropy: The randomness or complexity of the intensity distribution.

Gray Level Co-occurrence Matrix (GLCM): GLCM is a method used in image processing to examine the texture of an image by assessing how often pairs of pixels with specific values and in a specified spatial relationship occur in an image, creating a GLCM, and then extracting statistical measures from this matrix.

The main GLCM features include:

- Contrast: Measures the local variations in the GLCM.
- Correlation: Assesses how correlated a pixel is to its neighbors.
- Inverse Difference Moment (or Homogeneity): Measures the closeness of the distribution of elements in the GLCM to the GLCM diagonal.

Gray Level Size Zone Matrix (GLSZM): GLSZM is a method used in radiomics for texture analysis, particularly focusing on the size and distribution of continuous zones with the same gray level intensity in an image. GLSZM provides a way to quantify patterns and structures in an image that are not captured by first-order statistics or other texture matrices like the GLCM. A zone in GLSZM refers to a group of connected pixels that have the same gray level intensity.

The main GLSZM features include:

- Small Area Emphasis: Focuses on the distribution of small size zones.
- Large Area Emphasis: Highlights the presence of large size zones.
- Zone Percentage: The total number of zones relative to the size of the ROI, indicating textural uniformity.
- Low Gray Level Zone Emphasis: Reflects the proportion of zones with lower gray level values.
- High Gray Level Zone Emphasis: Indicates the presence of zones with higher gray level values.

Gray Level Run Length Matrix (GLRLM): GLRLM focuses on examining the length and directionality of continuous runs of pixels with the same gray level intensity in an image, thus providing important information about the texture and structural patterns. A run in GLRLM is defined as a set of consecutive, collinear pixels having the same gray level intensity. The length of a run is the number of pixels in this set.

The main GLRLM features include:

- Short Run Emphasis: Measures the distribution of short runs, indicating fine textural patterns.
- Long Run Emphasis: Highlights long runs, suggesting coarser textures.
- Run Length Nonuniformity: Quantifies the variability of run lengths, with higher values indicating more heterogeneous textures.
- Gray Level nuniformity: Measures the variability of gray levels in the runs.

D. Additional Results

To demonstrate the effectiveness of our proposed method in simulations, we attempted to generate large tumors, which are rare in clinical settings, This process is depicted in Figure A. Cases with such large tumors are unusual in medical imaging and difficult to find data. Our results prove that we can generate these rare samples. Furthermore, not only can we create hard-to-find samples, but we can also apply our method to simulate the growth of tumors. We anticipate this will contribute to prognostic research studies.

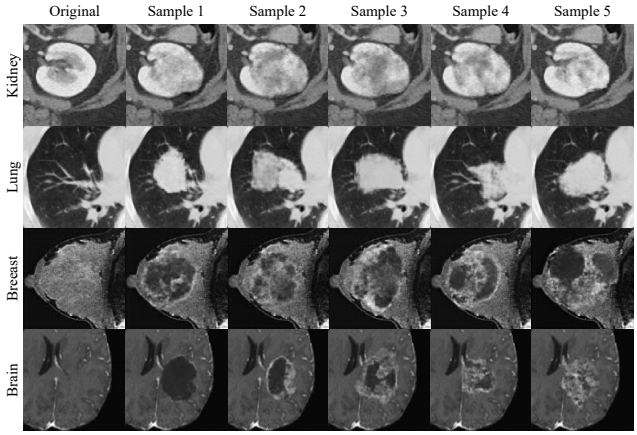

Figure A. Results of generating rare samples.

E. Future works

Medical imaging tends to be sensitive to parameter variations across different acquisition sites, which could lead to potential inconsistencies in extracting intensity-sensitive radiomics features, because minor alterations may result in significantly different feature values. This presents a challenge when attempting to apply the process across diverse domains and datasets simultaneously. For example, difficulties arise not only between CT and MRI but also across datasets from different anatomical organs like breast and brain. Advancements in research, in tandem with domain normalization for extracting and standardized radiomics features within a unified space, could pave the way for substantial progress.

F. Training and Inference Process

The training and inference processes of the texture generator involve scenarios where texture features are either provided (for generating tumors) or not provided (for generating normal tissue). During the training of non-tumor regions, an area that does not overlap with the tumor region is randomly chosen for masking. Subsequently, this masked area is reconstructed and utilized in training. These steps are summarized in Algorithm 1.

For sampling a normal image, the intended area for restoration to its normal state is masked, followed by a diffusion process to derive the normal image. This method is detailed in Algorithm 2. When sampling a tumor image, a mask is generated with specific shape features using a shape generator. This mask is applied to the intended area for tumor generation. A diffusion process is then carried out, conditioned on the texture features, to produce the tumor image. This method is detailed in Algorithm 3.

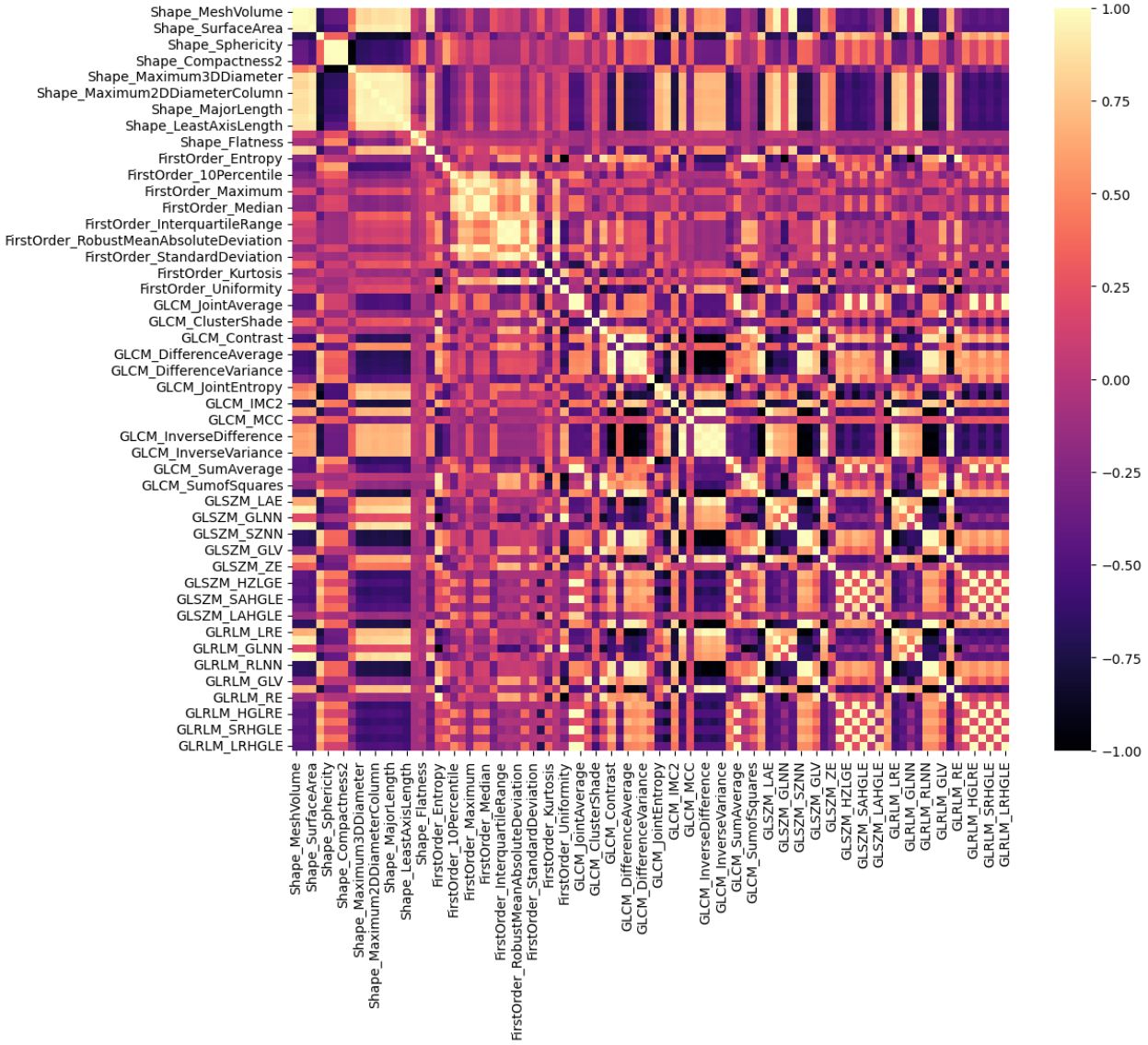

Figure B. **Radiomcis feature correlation matrix** extracted from the tumor region of a breast MRI. A higher positive correlation between features approaches 1, a higher negative correlation approaches -1, and a lower correlation approaches 0.

Algorithm 1 Texture generator training loop with image I, mask M, radiomics texture feature $\mathbf{r}_{\text{tx}}$, and VQGAN encoder $\mathcal{E}$

repeat
  if Training for Tumor **then**
    $I^M \leftarrow I \odot M$ ▷ Masking image with tumor mask
  else
    $M' \leftarrow \text{Shift}(M)$ ▷ Shift mask position without overlapping tumor
    $I^M \leftarrow I \odot M'$ ▷ Masking image with tumor mask
    $\mathbf{r}_{\text{tx}} \leftarrow$ None ▷ Initialize radiomics texture feature to none
  end if
  $\boldsymbol{x}_0 = \mathcal{E}(I), \boldsymbol{y} = \mathcal{E}(I^M)$ ▷ Image compression with VQGAN encoder
  $\boldsymbol{x}_0 \sim q(\boldsymbol{x}_0), \boldsymbol{y} \sim q(\boldsymbol{y})$
  timestep $t \sim Uniform(1, ..., T)$
  Gaussian noise $\epsilon \sim \mathcal{N}(0, \mathbf{I})$
  Forward diffusion $\boldsymbol{x}_t = (1 - m_t)\boldsymbol{x}_0 + m_t\boldsymbol{y} + \sqrt{\delta_t}\epsilon$
  Take gradient descent step on
    $\bigtriangledown_\theta ||m_t(\boldsymbol{y} - \boldsymbol{x}_0) + \sqrt{\delta_t}\epsilon - \epsilon_\theta(\boldsymbol{x}_t, t, \mathbf{r}_{\text{tx}})||^2$
until converged

Algorithm 2 Normal image sampling with VQGAN decoder $\mathcal{D}$

$I^M \leftarrow I \odot M$ ▷ Masking image with tumor mask
$\boldsymbol{y} = \mathcal{E}(I^M)$ ▷ Image compression with VQGAN encoder
$\boldsymbol{x}_T = \boldsymbol{y} \sim q(\boldsymbol{y})$ ▷ Sample conditional input
for $t = T, ..., 1$ **do**
  $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I})$ if $t > 1$, else $\mathbf{z} = 0$
  $\boldsymbol{x}_{t-1} = \boldsymbol{c}_{xt}\boldsymbol{x}_t + \boldsymbol{c}_{yt}\boldsymbol{y} - \boldsymbol{c}_{\epsilon t}\epsilon_\theta(\boldsymbol{x}_t, t) + \sqrt{\tilde{\delta}_t}\boldsymbol{z}$
end for
return $\mathcal{D}(\boldsymbol{x}_0)$ ▷ Reconstruction with VQGAN decoder

Algorithm 3 Tumor image sampling with shape generator G and radiomics shape feature $\mathbf{r}_{\text{sh}}$

$z \sim \mathcal{N}(0, \mathbf{I})$
$\mathcal{M} = G(z, \mathbf{r}_{\text{sh}})$ ▷ Generate tumor mask with shape feature
$I^{\mathcal{M}} \leftarrow I \odot \mathcal{M}$ ▷ Masking image with tumor mask
$\boldsymbol{y} = \mathcal{E}(I^{\mathcal{M}})$ ▷ Image compression with VQGAN encoder
$\boldsymbol{x}_T = \boldsymbol{y} \sim q(\boldsymbol{y})$ ▷ Sample conditional input
for $t = T, ..., 1$ **do**
  $\mathbf{z} \sim \mathcal{N}(0, \mathbf{I})$ if $t > 1$, else $\mathbf{z} = 0$
  $\boldsymbol{x}_{t-1} = \boldsymbol{c}_{xt}\boldsymbol{x}_t + \boldsymbol{c}_{yt}\boldsymbol{y} - \boldsymbol{c}_{\epsilon t}\epsilon_\theta(\boldsymbol{x}_t, t, \mathbf{r}_{\text{tx}}) + \sqrt{\tilde{\delta}_t}\boldsymbol{z}$
end for
return $\mathcal{D}(\boldsymbol{x}_0)$ ▷ Reconstruction with VQGAN decoder