

VasTSD: Learning 3D Vascular Tree-state Space Diffusion Model for Angiography Synthesis

Supplementary Material

We present additional experimental results and implementation details, supporting the main paper. We also provide a demo video and codes in supplementary materials.

Firstly, Sec. 1 provides more details of the pre-trained vision embedder. Sec. 2 provides further explanations of the proposed vascular tree scanning algorithm, highlighting the dynamic generation process of the vascular tree. In Sec. 3, We explain the pre-processing of ITKTubeTK dataset [4] and Topcow2024 dataset [10] and ISICDM 2020 dataset [5, 8]. Sec. 4 defines 3D blood vessel volume evaluation metrics. We discuss the effectiveness of different loss functions and provide corresponding ablation results in Sec. 5. Additionally, statistical analyses of 3D volumes are conducted in Section 6. In Sec. 7, we showcase additional examples of the vascular synthesis process.

1. Pre-training Vision Embedder

The non-angiography structures are processed using a variational autoencoder (VAE) to extract latent representations, resulting in a three-dimensional tensor, $z_{\text{Non-angio}}$, composed of z-axis slices of the non-contrast vessels. These latent representations are partitioned into multiple patches. Each patch is encoded via 2D convolution, producing an embedded feature of size $B \times L \times D$, where B is the batch size, L denotes the number of patches, and D represents the embedding dimension.

To capture global contextual information, global average pooling (GAP) is applied to the embedded features along each channel, reducing each channel to a scalar that reflects its international significance. These scalars are then processed through a two-layer multi-layer perceptron (MLP): the first layer performs dimensionality reduction to capture nonlinear inter-channel relationships, and the second restores the dimensionality to match the original channel count. Using a Sigmoid activation function, the resulting channel-wise weights are normalized to the range of (0, 1).

The normalized weights are attention masks applied channel-wise to adjust the embedded features. Specifically, each embedded feature channel is multiplied by its corresponding weight, yielding weighted feature maps that incorporate global contextual information while emphasizing task-relevant regions in the original non-angiography.

Subsequently, the adjusted embeddings are flattened into one-dimensional vectors (tokens). These tokens are L2-normalized and utilized to compute a similarity matrix across image sequences, optimizing the embedding function within a contrastive learning framework. These atten-

tion masks and tokens enable the model to capture subtle variations in non-contrast vascular images and provide an effective conditional representation for synthesizing contrast-enhanced vascular structures.

2. Vascular Tree Scanning Algorithm

2.1. Preliminaries

The state space models (SSMs) utilize methods similar to the Kalman filter. It maps the input sequence $\mathbf{u}(t)$ to the output sequence $\mathbf{y}(t)$, with the hidden state variable $\mathbf{x}(t)$ representing the internal dynamics of the system. The entire process can be described as Equation 1-2.

$$\dot{\mathbf{x}}(t) = \mathbf{A}\mathbf{x}(t) + \mathbf{B}\mathbf{u}(t), \quad (1)$$

$$\mathbf{y}(t) = \mathbf{C}\mathbf{x}(t) + \mathbf{D}\mathbf{u}(t), \quad (2)$$

where $\mathbf{x}(t) \in \mathbb{R}^n$ represents the system's internal state, $\dot{\mathbf{x}}(t)$ represents the updated state. The state transition matrix $\mathbf{A} \in \mathbb{R}^{n \times n}$ governs the state evolution, while the input matrix $\mathbf{B} \in \mathbb{R}^{n \times m}$ describes the influence of the input on the state. The output matrix $\mathbf{C} \in \mathbb{R}^{p \times n}$ maps the state to the output, and the feedthrough matrix $\mathbf{D} \in \mathbb{R}^{p \times m}$ represents the direct input-output relationship.

To improve the efficiency of state space models in handling long sequences, new models such as Structured State Space Sequence Models (S4) [3], Simplified State Space Layers for Sequence Modeling (S5) [7], and Selective Structured State Space Models (S6) [2] have emerged.

S4 introduces a structured decomposition and discretization of the \mathbf{A} matrix, enhancing computational efficiency in long-sequence tasks:

$$\mathbf{x}_{k+1} = \mathbf{A}\mathbf{x}_k + \mathbf{B}\mathbf{u}_k. \quad (3)$$

S5 enhances adaptability by introducing a learnable time-scale parameter, Δ , for dynamic adjustment:

$$\mathbf{x}_{k+1} = e^{A\Delta}\mathbf{x}_k + \mathbf{B}\mathbf{u}_k. \quad (4)$$

S6 introduces a dynamic selection mechanism for the \mathbf{A} matrix, enabling content-based adjustments to prioritize critical information:

$$\mathbf{x}_{k+1} = \mathbf{A}(\mathbf{x}_k, \mathbf{u}_k)\mathbf{x}_k + \mathbf{B}\mathbf{u}_k. \quad (5)$$

These methods improve the efficiency and focus of state space models for long sequences but do not resolve long-range dependencies in vascular-type structured sequences.

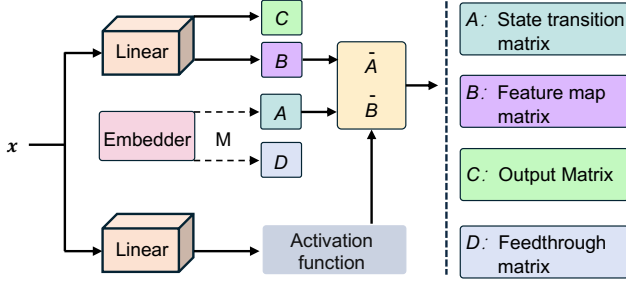


Figure 1. Schematic diagram illustrating the dynamic update process of vascular tree scanning parameters.

2.2. Dynamic Update of Tree Variables

To better model vascular structures in 3D volumes, an optimization mechanism can be introduced into the continuous state space framework to serialize 3D blood vessels effectively, as shown in Figure 1. These mechanisms must account for spatial coherence, topological consistency, and long-range dependencies within the vascular structures. Specifically, leveraging the cross-slice attention mechanism, as detailed in Section 3.3.2 in the main paper, allows for capturing the coarse spatial distribution of blood vessels between adjacent slices using soft masks. This facilitates the modeling of both the spatial coordinates and the extension directions of blood vessels through a state transition matrix A .

Given the spatial distribution of blood vascular characterized by their position $\mathbf{p}(t)$ and extension direction $\mathbf{d}(t)$ —the state transition matrix A is defined as follows:

$$\dot{\mathbf{x}}(t) = A(\mathbf{p}(t), \mathbf{d}(t))\mathbf{x}(t), \quad (6)$$

where $\mathbf{x}(t)$ represents the state of the system.

The input feature map matrix B is modeled as a dynamic matrix that governs how the input features influence state updates. It is defined as:

$$B = B(\mathbf{p}(t), \mathbf{d}(t))\mathbf{u}(t), \quad (7)$$

where $\mathbf{u}(t)$ denotes the input features, $\mathbf{p}(t)$ represents the spatial coordinates of the blood vessels in the 3D slices, and $\mathbf{d}(t)$ captures the direction of vascular extensions, inferred from the masks.

Thus, the core equation governing the continuous state-space scanning of 3D vascular structures is expressed as:

$$\dot{\mathbf{x}}(t) = A(\mathbf{p}(t), \mathbf{d}(t))\mathbf{x}(t) + B(\mathbf{p}(t), \mathbf{d}(t))\mathbf{u}(t). \quad (8)$$

This formulation enables the integration of spatial structural characteristics, ensuring a coherent and robust representation of 3D blood vessels.

2.3. Node Update Process

In order to better describe the changes in nodes and matrices during the vascular tree update process, we take the state update of a certain node as an example. The node representation obtained from the visual encoder (denoted as M) corresponds to the input feature vector for each node. The state transformation matrix $A \in \mathbb{R}^{N \times N}$ controls the state propagation of each node, determining how information is transmitted from parent nodes to child nodes in the tree structure. The feature vector B is generated by the input features:

$$B_i = W_B \cdot x_i + b_B, \quad (9)$$

where $W_B \in \mathbb{R}^{N \times d}$, $b_B \in \mathbb{R}^N$ are the weight and bias, respectively. The feature vector B is then transformed into node states for further processing.

The matrices C and D are used to update the transformations through loss functions. The matrix C is used to compute the final output transformations, and matrix D is responsible for the input transformations.

The update process is divided into two stages: the aggregation of the feature information and the transmission of this information across the tree structure:

$$S(E[i, j]) = \frac{\exp(-\psi \cdot \cos(X[i], X[j]))}{\sum_{k \in \text{Neigh}(i)} \exp(-\psi \cdot \cos(X[i], X[k]))}, \quad (10)$$

where $S(E[i, j])$ represents the edge weight between nodes, and $B[j]$ and $X[j]$ are the features and state of node j , respectively.

In the tree-like state space model, the state of each node is influenced by the parent node's state. The tree structure propagates states based on hierarchical relationships, and the aggregation of features follows a sequence of steps as shown below.

The feature vectors are aggregated through the parent-to-child transmission process, where the aggregation process of node i is defined as:

$$\eta_i = B_i \cdot \frac{\partial \text{Loss}}{\partial h_i} + \sum_{j \in \text{Children}(i)} \eta_j A_j, \quad (11)$$

where η_j and A_j are the transformations passed from the parent node to the child node j .

During the transmission from parent to child, the feature weights A and B are updated as follows:

$$A_i^- = A_i - \lambda \cdot \frac{\partial L}{\partial A_i}, \quad (12)$$

$$B_i^- = B_i - \lambda \cdot \frac{\partial L}{\partial B_i}, \quad (13)$$

$$\eta_i = A_i^- \eta, \quad \eta_j = B_i^- \eta_j. \quad (14)$$

The matrices C and D are updated similarly:

$$C = C - \lambda \cdot \frac{\partial L}{\partial C}, \quad (15)$$

$$D = D - \lambda \cdot \frac{\partial L}{\partial D}, \quad (16)$$

Finally, during the aggregation and transmission process, the loss function is computed as:

$$\mathcal{L}_{\text{scan}} = \sum_i \frac{1}{2} \|\eta_i\|^2. \quad (17)$$

3. Datasets

This study performs a series of experiments on brain and lung vascular data. The brain vascular data sets include the ITKTubeTK dataset [4] and the Topcow2024 dataset [10], while the lung vascular data comes from the ISICDM 2020 dataset [5, 8]. The specific parameters of the datasets are shown in Tab. 1.

ITKTubeTK Dataset. This dataset is a publicly available resource for analyzing healthy brain structures using MRI, from CASILab of the University of North Carolina. It includes high-quality data from 100 healthy subjects, evenly distributed across five age groups ($18 \sim 29$, $30 \sim 39$, $40 \sim 49$, $50 \sim 59$, ≥ 60), with balanced male and female distributions. Subjects with conditions that could impact brain structure, such as diabetes or head trauma, are excluded. Demographic information, including age, sex, and handedness, is also recorded. Some participants are excluded due to claustrophobia or motion artifacts.

The dataset includes T1, T2, Time-of-Flight MRA (TOF-MRA), and Diffusion Tensor Imaging (DTI) acquired under standardized protocols using a 3T MRI scanner. The voxel sizes are $1 \times 1 \times 1 \text{ mm}^3$ for T1 and T2, $0.5 \times 0.5 \times 0.8 \text{ mm}^3$ for TOF-MRA, and $2 \times 2 \times 2 \text{ mm}^3$ for DTI, providing high-resolution images suitable for detailed analysis. Data is stored in the metaImage format (.mha).

Topcow 2024 Dataset. This dataset is a publicly available resource design for brain vascular analysis, including high-resolution 3D reconstructions from both MRI and CT scans. The dataset features a voxel size of $1.0 \sim 1.5 \text{ mm}$ isotropic resolution and matrix sizes typically around $256 \times 256 \times 256$. It contains detailed vascular imagery, ideal for segmentation and modeling tasks, and includes metadata with specific scanning parameters such as slice thickness and acquisition angles. The images were captured using clinical scanners from Siemens and Philips and are stored in metaImage (.mha) format for compatibility with medical imaging software.

ISICDM 2020 Dataset. This dataset is for pulmonary vascular analysis, comprising 10 pulmonary CT and 15 pulmonary CTA series. The dataset includes high-resolution

3D images with 204 to 536 slices per series, normalized voxel spacing of 1 mm , and a 512×512 pixels resolution. All images are stored in Nifti format for compatibility.

Data Preprocessing Procedures. To prepare the original 3D blood vessel volume dataset for analysis, redundant slices from the top and bottom were evenly removed to retain the central region of interest. The resulting volumes were subsequently cropped to a standardized size of $128 \times 448 \times 448$. Additionally, a subset of more reliable data was selected from the original dataset for experimental purposes. The specific data identifiers used in the experiments are detailed in Tab. 1.

$$S_{\text{mapped}} = 10 \times \frac{S_{\text{raw}} - S_{\text{min}}}{S_{\text{max}} - S_{\text{min}}} \quad (18)$$

The resulting map score indicates the degree of structural similarity, where higher scores reflect more consistent connectivity. This metric provides a robust, quantifiable means of comparing vascular modeling techniques.

4. Metrics

Dice Score. The Dice Score, also known as the Dice coefficient, is a widely used metric for evaluating the similarity between two sets or volumes. It is defined as the ratio of twice the volume of the intersection of the two sets to the sum of their individual volumes. Mathematically, it is expressed as:

$$8S_{\text{Dice}} = \frac{2 \times |A \cap B|}{|A| + |B|}, \quad (19)$$

where A and B represent the two volumes, and $|A \cap B|$ denotes the volume of their intersection. The Dice coefficient ranges from 0 to 1, with 1 indicating perfect overlap and 0 indicating no overlap. In the context of 3D vascular analysis, the Dice Score is used to assess the spatial overlap between the synthetic and real vascular structures. A high Dice Score implies a high degree of morphological similarity between the two vascular volumes, indicating that the synthetic blood vessel closely mimics the real one in terms of shape and size.

Jaccard Score. The Jaccard Index is another metric used to quantify the similarity between two sets, but it focuses on the ratio of the intersection to the union of the sets. It is defined as:

$$S_{\text{Jaccard}} = \frac{|A \cap B|}{|A \cup B|}, \quad (20)$$

where $|A \cup B|$ is the volume of the union of the two sets, and $|A \cap B|$ is the volume of their intersection. Similar to the Dice Score, the Jaccard Index ranges from 0 to 1, with higher values indicating better overlap. The Jaccard Index is particularly useful in 3D vascular analysis to evaluate the spatial consistency of vascular structures. A high Jaccard Score indicates that the synthetic and real vessels

Datasets	Voxel Size (mm^3)	Matrix Size	Scanning Parameters	Scanner	Dataset size
ITKTubeTK [4]	$0.5 \times 0.5 \times 0.8$	$448 \times 448 \times 128$	TR/TE (ms): 35/3.56, Flip Angle: 22°	3T	T2:87 T1-MPRAGE:51 T1-Flash:88
Topcow 2024 [10]	1.0 - 1.5	$256 \times 256 \times 256$	55	55	85
ISICDM 2020 [5, 8]	1	$512 \times 512 \times 512$	55	55	10

Table 1. **Overview of datasets used in the study.** Voxel size, matrix size, scanning parameters, and scanners for ITKTubeTK, Topcow2024, and ISICDM 2020 datasets. Missing information is marked with 55.

Methods	ITKTubeTK						Topcow 2024	
	T1-Flash		T2		T1-MPRAGE		CTA	
	PSNR \uparrow	SSIM \uparrow	PSNR \uparrow	SSIM \uparrow	PSNR \uparrow	SSIM \uparrow	PSNR \uparrow	SSIM \uparrow
VasTSD	29.86	0.9478	29.72	0.9384	29.92	0.9454	27.36	0.9372
w/o $\mathcal{L}_{\text{InfoNCE}}$	28.12	0.8976	28.23	0.8847	28.42	0.8875	25.89	0.8735
w/o $\mathcal{L}_{\text{diff}}$	28.43	0.9214	28.56	0.9062	28.53	0.9125	26.21	0.9015
w/o $\mathcal{L}_{\text{scan}}$	<u>29.16</u>	<u>0.9314</u>	<u>29.25</u>	<u>0.9163</u>	<u>29.25</u>	<u>0.9286</u>	<u>26.57</u>	<u>0.9148</u>

Table 2. **Ablations of loss functions.** The best and second-best results are highlighted in **bold** and underlined.

Datasets	T		ANOVA		Pearson corr.
	T-stat	P-value	F-stat	P-value	
T1-Flash	-7.59	0.0016	98.4	2.59E-05	0.628
T1-MPRAGE	-6.83	0.0012	85.2	2.41E-05	0.592

Table 3. Results of statistical analysis of 3D volumes.

have a large common region relative to their combined size, reflecting a close structural match between them.

Connectivity Score. In this study, we propose the Connectivity Score to quantify the structural connectivity differences between two 3D vascular volumes. The process begins with thresholding the 3D images to generate binary masks, where v is the voxel intensity and T is the threshold:

$$\text{mask}(v) = \begin{cases} 1 & \text{if } v > T \\ 0 & \text{otherwise.} \end{cases} \quad (21)$$

Connected component analysis is then applied to the binary masks to identify distinct vascular structures. The Connectivity Score is calculated based on the mismatched connected components between the real and synthetic volumes:

$$S_{\text{connectivity}} = \sum_{i \in C_1} \sum_{j \in C_2} \mathbb{I}(C_1(i) \neq C_2(j)), \quad (22)$$

where C_1 and C_2 represent the connected components of the real and synthetic volumes, respectively, and $\mathbb{I}(\cdot)$ is the indicator function.

To standardize the score, we map the raw Connectivity Score to a $[0, 10]$ scale using the minimum and maximum values in the dataset:

5. Ablation Study of Loss Functions

We perform an ablation study on different components of the loss function to understand their contributions to the performance of our proposed model. The loss function is crucial in guiding the optimization process, and its design significantly affects the model’s ability to generalize and perform well on unseen data.

We design the ablation experiments by systematically removing or modifying key components of the loss function. The following variations are evaluated: removal of the InfoNCE loss, removal of the Denoising loss, and removal of the Tree-scanning loss.

Experimental results demonstrate that the complete VasTSD model outperforms all alternatives in angiography synthesis tasks, with significantly higher PSNR and SSIM, highlighting its superior signal fidelity and structural consistency. Among its components, the vision embedder $\mathcal{L}_{\text{InfoNCE}}$ is critical, as its removal causes the largest performance drop, underscoring the importance of pre-trained embeddings for capturing local and global consistency and generating coherent vascular structures.

Removing the diffusion loss $\mathcal{L}_{\text{diff}}$ reduces detail quality and increases blurriness, while eliminating the scan loss $\mathcal{L}_{\text{scan}}$ has minimal impact, with performance remaining close to the full model. These results emphasize the synergistic design of VasTSD’s loss functions, with $\mathcal{L}_{\text{InfoNCE}}$ as the core, complemented by $\mathcal{L}_{\text{diff}}$ and $\mathcal{L}_{\text{scan}}$, which refine detail and local consistency to enhance angiography synthesis.

6. Statistical analyses of 3D volumes

To validate the contribution of individual modules in the ablation study, comprehensive statistical analyses were per-

formed, including T-tests (assessing inter-group mean differences), ANOVA (comparing multi-group means), and Pearson correlation coefficient (quantifying linear associations between continuous variables). Experimental results on both T1-Flash and T1-MPRAGE datasets (Tab. 3) demonstrated statistically significant inter-group discrepancies with robust ANOVA validation. Additionally, Pearson correlation coefficient suggested consistent linear relationships across experimental conditions.

7. Example of Angiography Synthesis

In this section, we present more examples of 2D slices and 3D effects of angiography synthesis. In Fig. 2- 4, we provide a visual comparison between our method and traditional modality conversion-based methods in synthesizing 2D slices. The classic modality conversion methods including cGAN [1], SynDiff [6] and DiffMa [9]. To facilitate the visualization of 3D effects, we include a demo video which can be found in supplementary materials.

References

- [1] Salman UH Dar, Mahmut Yurt, Levent Karacan, Aykut Erdem, Erkut Erdem, and Tolga Cukur. Image synthesis in multi-contrast mri with conditional generative adversarial networks. *IEEE transactions on medical imaging*, 38(10): 2375–2388, 2019. 5
- [2] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023. 1
- [3] Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*, 2021. 1
- [4] Alper Güngör, Baris Askin, Damla Alptekin Soydan, Can Barış Top, Emine Ulku Saritas, and Tolga Çukur. Deq-mpi: A deep equilibrium reconstruction with learned consistency for magnetic particle imaging. *IEEE Transactions on Medical Imaging*, 43(1):321–334, 2023. 1, 3, 4
- [5] Jiachen Han, Naixin He, Qiang Zheng, Lin Li, and Chaoqing Ma. 3d pulmonary vessel segmentation based on improved residual attention u-net. *Medicine in Novel Technology and Devices*, 20:100268, 2023. 1, 3, 4
- [6] Muzaffer Özbey, Onat Dalmaz, Salman UH Dar, Hasan A Bedel, Şaban Öztürk, Alper Güngör, and Tolga Çukur. Un-supervised medical image translation with adversarial diffusion models. *IEEE Transactions on Medical Imaging*, 2023. 5
- [7] Jimmy TH Smith, Andrew Warrington, and Scott W Linderman. Simplified state space layers for sequence modeling. *arXiv preprint arXiv:2208.04933*, 2022. 1
- [8] Wenjun Tan, Luyu Zhou, Xiaoshuo Li, Xiaoyu Yang, Yufei Chen, and Jinzhu Yang. Automated vessel segmentation in lung ct and cta images via deep neural networks. *Journal of X-ray science and technology*, 29(6):1123–1137, 2021. 1, 3, 4
- [9] Zhenbin Wang, Lei Zhang, Lituan Wang, and Zhenwei Zhang. Soft masked mamba diffusion model for ct to mri conversion. *arXiv preprint arXiv:2406.15910*, 2024. 5
- [10] Kaiyuan Yang, Fabio Musio, Yihui Ma, Norman Juchler, Johannes C Paetzold, Rami Al-Maskari, Luciano Höher, Hongwei Bran Li, Ibrahim Ethem Hamamci, Anjany Sekuboyina, et al. Benchmarking the cow with the topcow challenge: Topology-aware anatomical segmentation of the circle of willis for cta and mra. *ArXiv*, 2023. 1, 3, 4

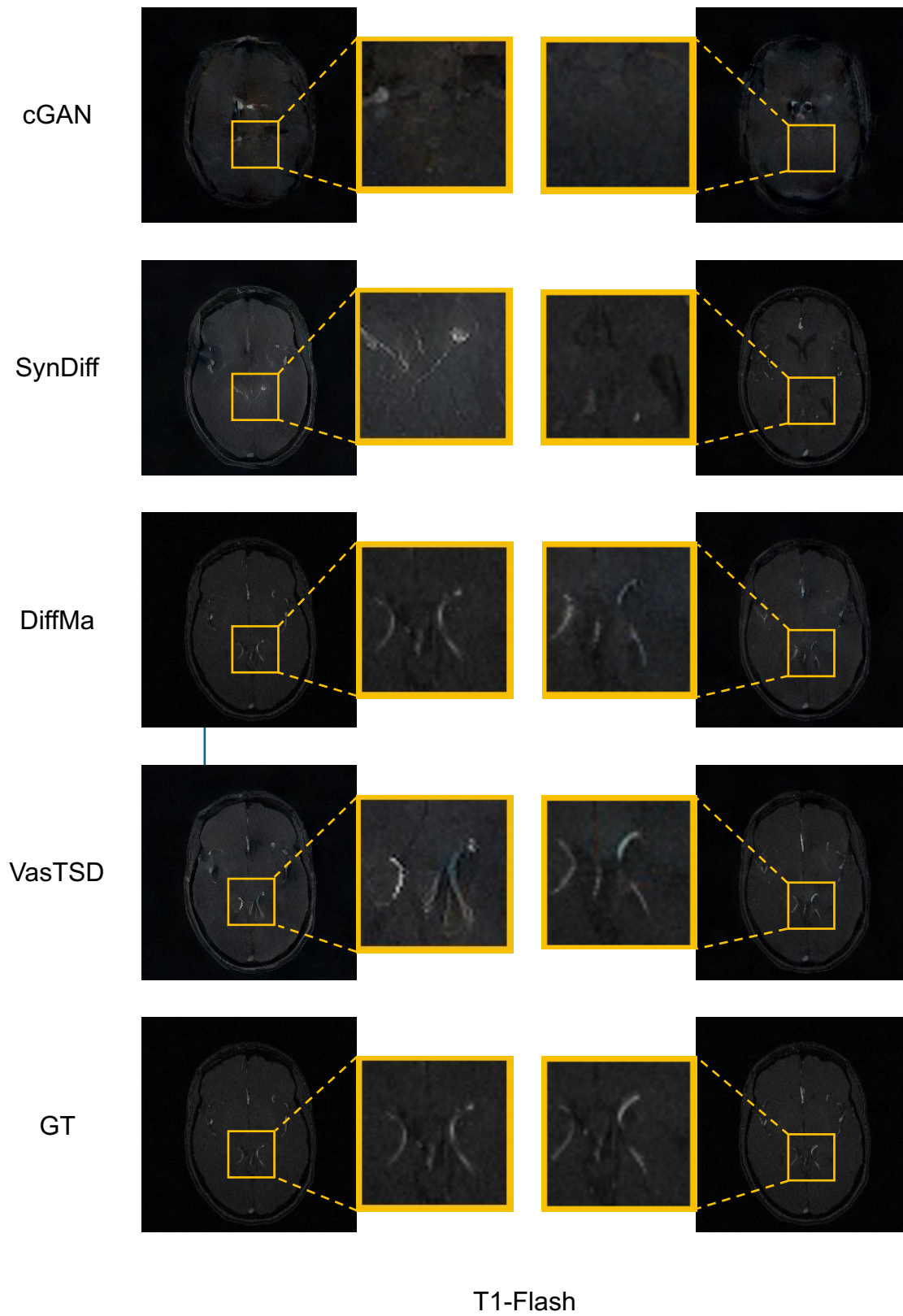


Figure 2. Comparison of 2D slices of angiography generated based on T1-Flash.



Figure 3. Comparison of 2D slices of angiography generated based on T1-MPRage.

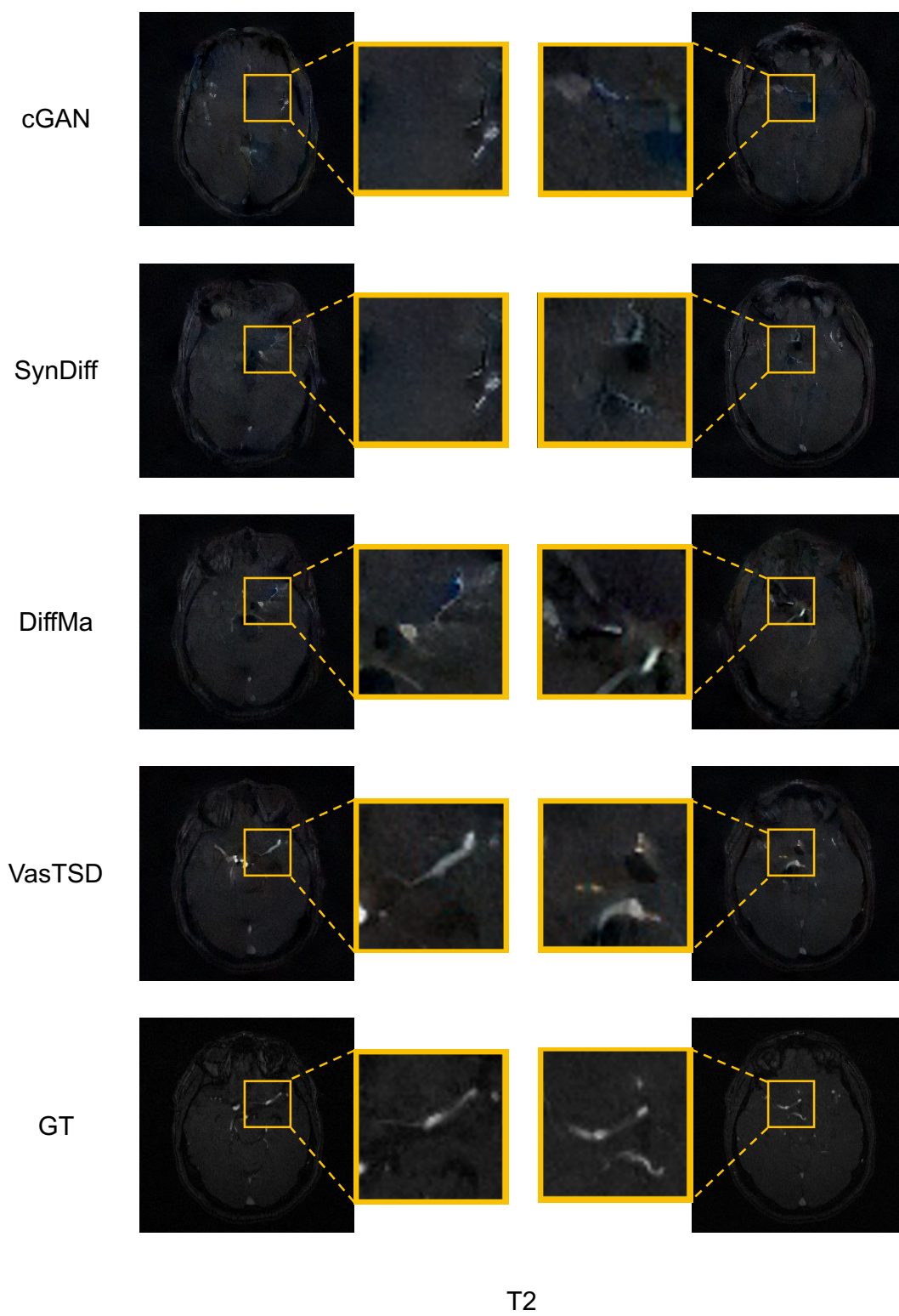


Figure 4. Comparison of 2D slices of angiography generated based on T2.