GuideGen: A Text-Guided Framework for Paired Full-torso Anatomy and CT Volume Generation

Presenter Linrui Dai[1,2*], Rongzhao Zhang[3*], Yongrui Yu[1], Xiaofan Zhang[1]✉

[1] Department of Computer Science & Engineering, School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University

[2] Department of Computer Science, School of Information Science and Technology, The University of Tokyo

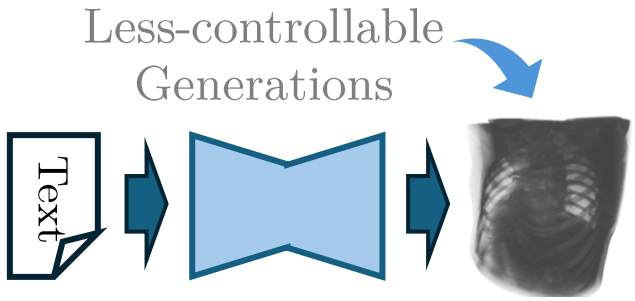
[3] Shanghai Artificial Intelligence Laboratory

Presentation Slides for AAAI-26

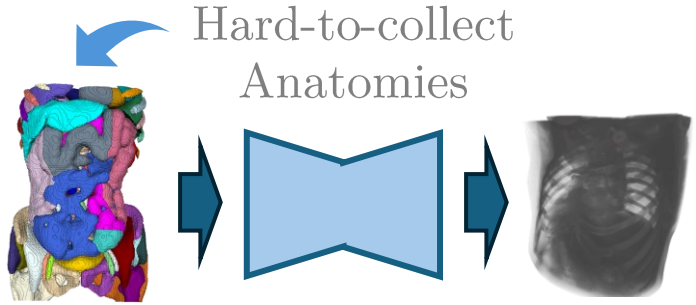
Accessible Full-torso Segmentation Dataset Synthesis

We propose to enhance the accessibility of medical CT synthesis models for segmentation by identifying two key drawbacks for concurrent models:

- *Text-based generation frameworks* often yields anatomies uncontrollable for segmentation purposes;
- *Mask-based generation frameworks* lack the flexibility for effortless sample synthesis at downstream;



Text-based generation frameworks

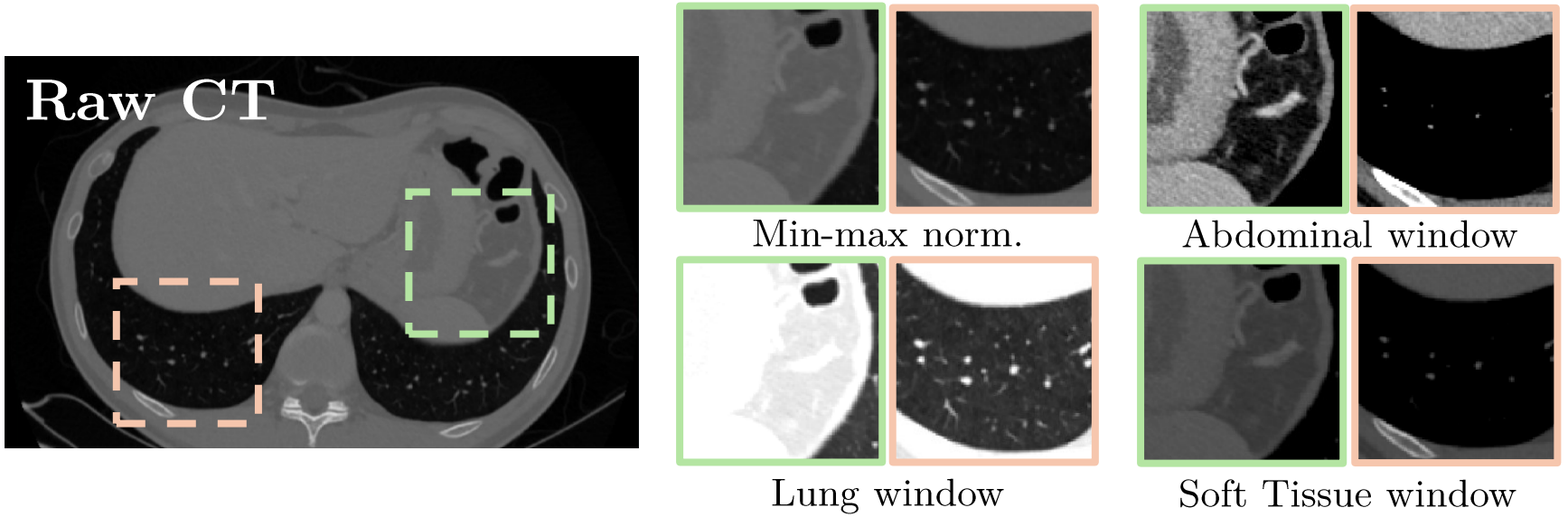


Mask-based generation frameworks

- Our *GuideGen* incorporates a separate mask synthesis module to extract valid anatomy from input textual prompts, thereby retrieving CT samples aligned with both the textual and semantic information and ready for a wide range of downstream tasks

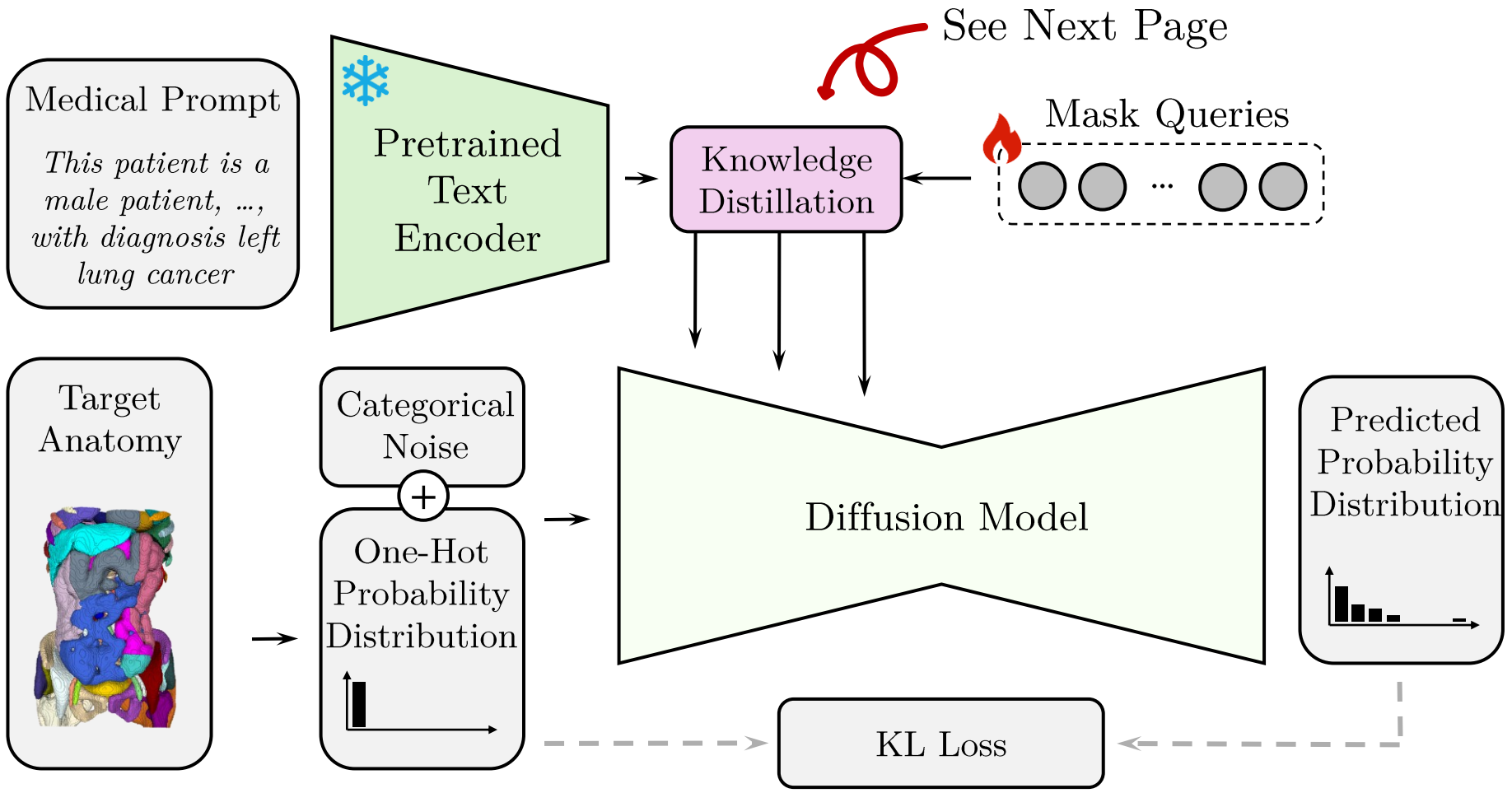
Accessible Full-torso Segmentation Dataset Synthesis

Concurrent medical generation models seldomly demonstrate generalizability across multiple organs, as they often consider patches of local anatomy characterized by a fixed normalization window:

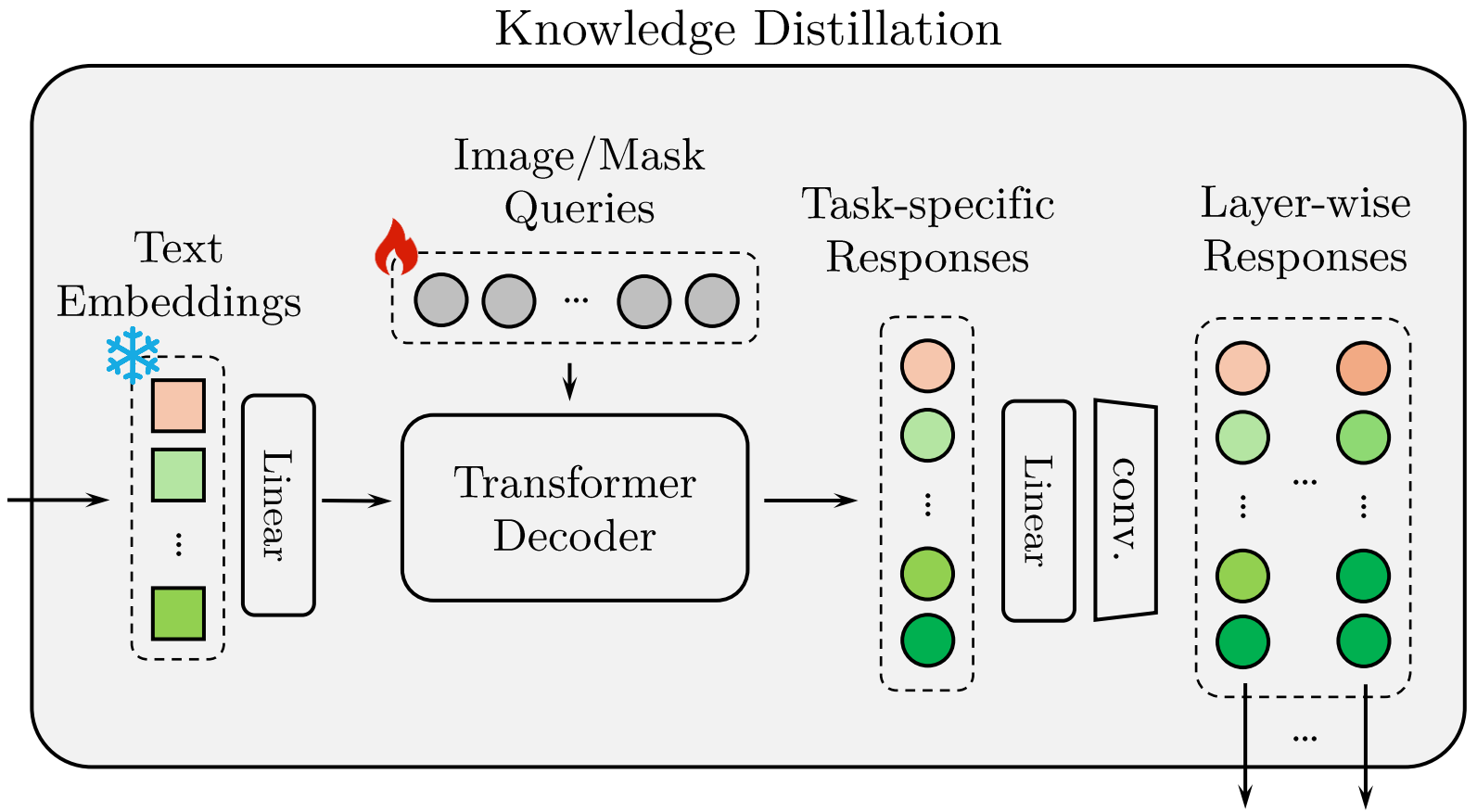


To incorporate comprehensive anatomical features across multiple contrast levels, our GuideGen utilizes an anatomy-aware HDR autoencoder to preserve details within the high dynamic range intrinsic to full-torso CTs.

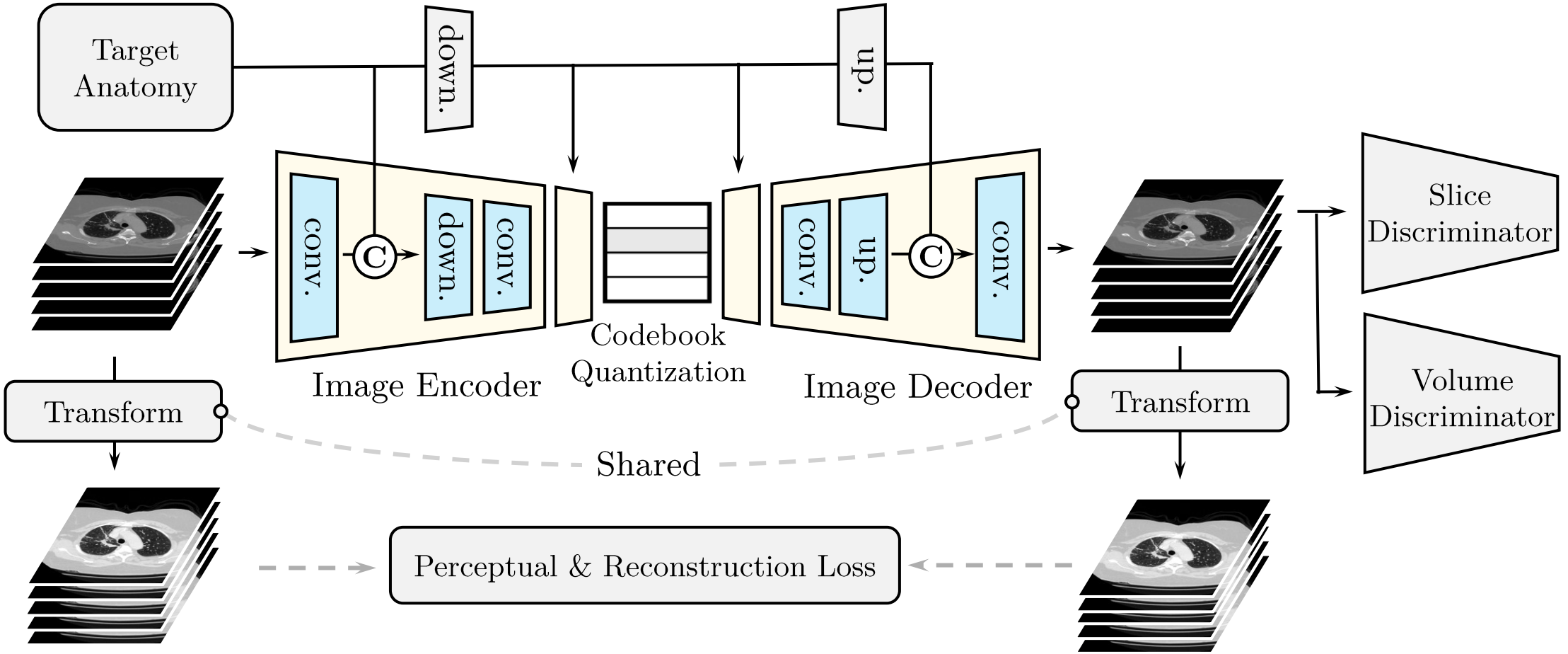
Stage-I: Text-conditional Semantic Synthesizer



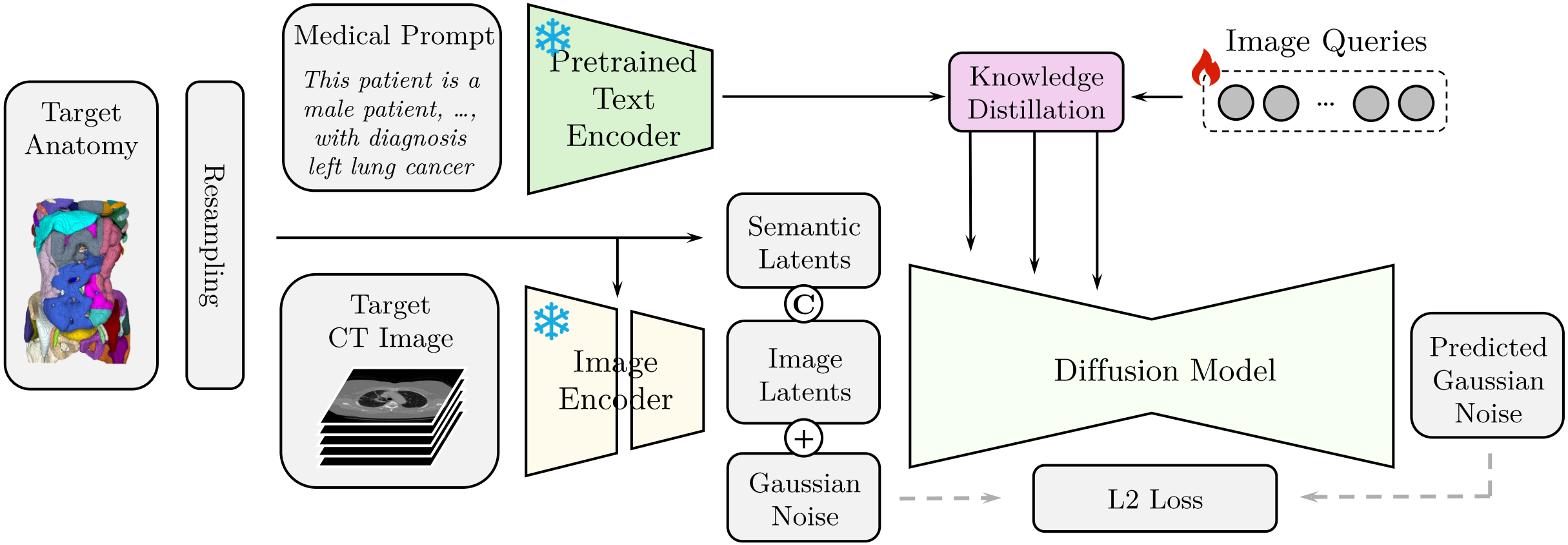
Stage-I: Text-conditional Semantic Synthesizer



Stage-II: Anatomy-aware HDR Autoencoder

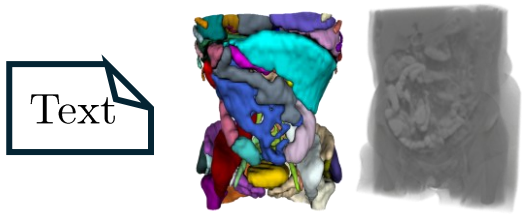


Stage-III: Latent-guided Feature Generator



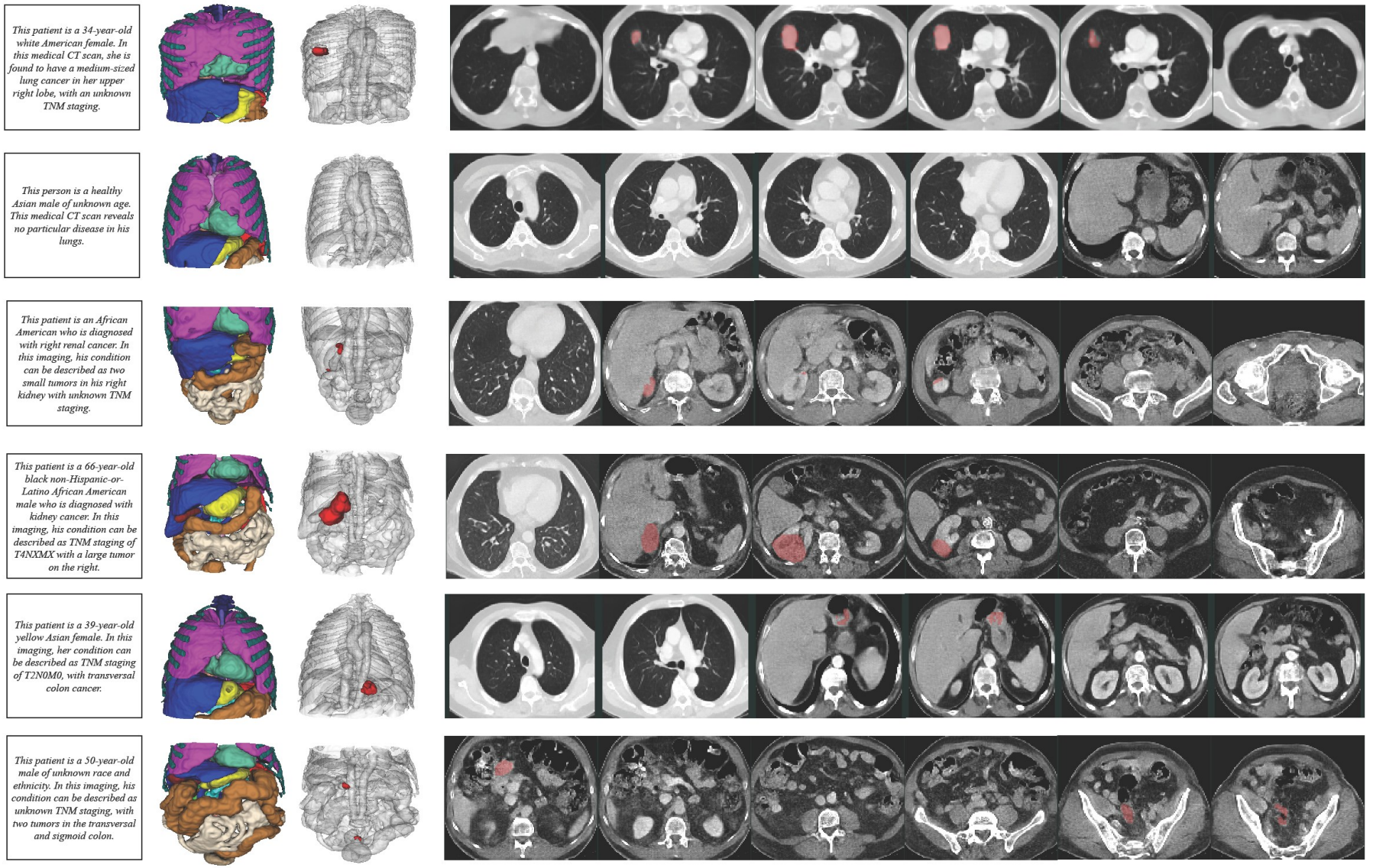
Synthesis Results

Synthesized
(input text)-mask-CT triplet $\mathbf{P}$



$$\mathbf{P}=\{\ \mathbf{t}\ ,\ \mathbf{m}\ ,\ \mathbf{v}\ \}$$

Stage-I: $\mathbf{t} \rightarrow \mathbf{m}$
Stage-II: $\mathbf{v} \overset{enc.}{\rightarrow} \mathbf{z}$
Stage-III: $\mathbf{t} + \mathbf{m} \rightarrow \mathbf{z} \overset{dec.}{\rightarrow} \mathbf{v}$



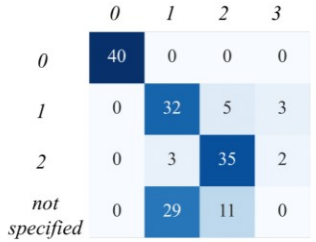
Modality Alignment Analysis

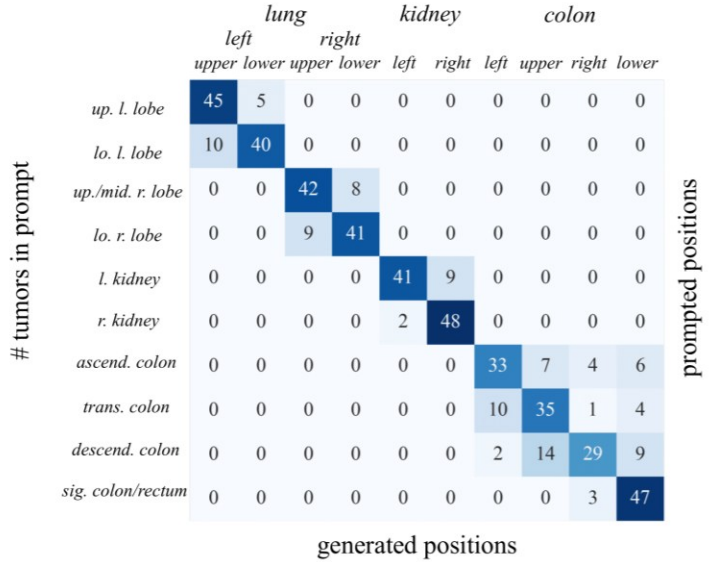
Methods	Accuracy↑ Age	Gender	Race	Tumor Loc.	Avg.
Pinaya's	0.06	0.35	0.10	0.17	0.17
GenerateCT	0.07	0.21	0.44	0.03	0.19
MedSyn	0.17	0.74	0.51	0.47	0.47
GuideGen	**0.39**	**0.90**	**0.60**	**0.89**	**0.69**

Image-Prompt alignment

Methods	DSC↑ Spl.	Kid.	Liver	Sto.	Pan.	Lung	S.B.	Duo.	Colon	Heart	Avg.
MAISI	0.73	0.72	0.80	0.60	0.43	0.84	0.49	0.35	0.55	0.40	0.59
Zhuang's	0.43	0.43	0.62	0.36	0.21	0.69	0.37	0.23	0.26	0.24	0.38
MedDDPM	0.35	0.36	0.39	0.54	0.29	0.34	0.47	**0.43**	0.67	0.22	0.41
MedSyn	0.52	0.51	0.51	0.40	0.07	0.59	0.12	0.01	0.54	0.22	0.35
GuideGen	**0.75**	**0.72**	**0.90**	**0.63**	**0.46**	**0.84**	**0.51**	0.41	**0.70**	**0.53**	**0.65**

Image-Mask alignment

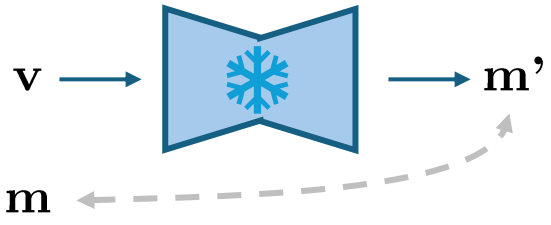




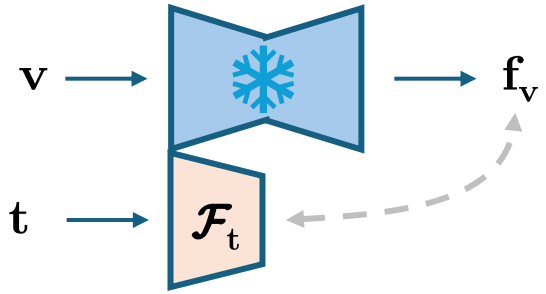
Mask-Prompt alignment

Alignment Score Computation (←--→):

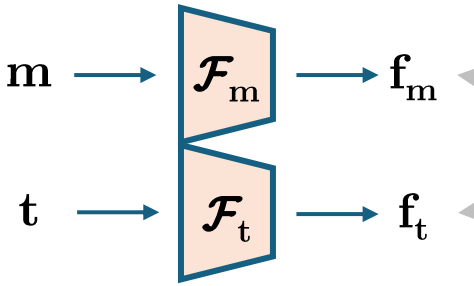
Pretrained multi-organ segmentation model



Pretrained feature classification model



Text feature extractor (age, gender, tumor loc., ...)



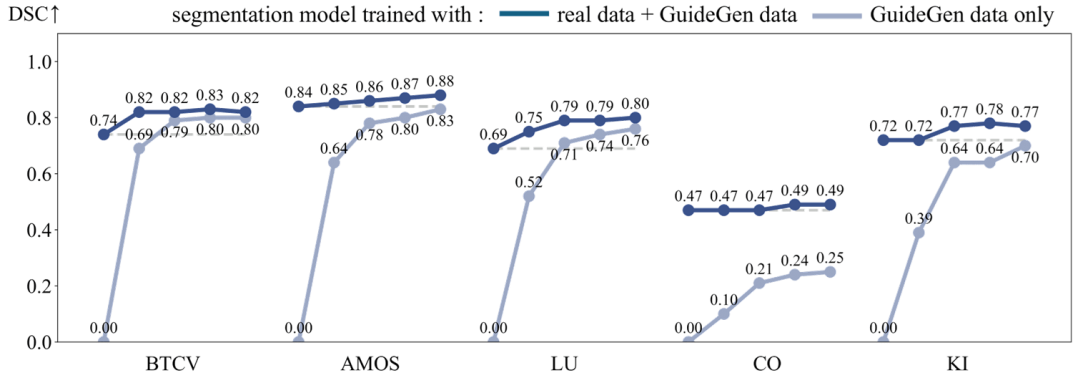
Mask feature extractor (#conn. components, tumor pos.)

Text feature extractor (#tumor, tumor pos.)

Downstream Usability of Synthesized Samples

Method	No. Train Cases	DSC↑ LU	CO	KI	Avg.	HD_{95} ↓ LU	CO	KI	Avg.
Real	50/100/313	0.69	0.47	0.72	0.63	11.9	212.7	70.9	98.5
MAISI	200	0.48	0.10	0.24	0.27	31.7	284.9	293.5	203.4
Zhuang's	200	0.12	0.07	0.13	0.11	716.0	274.1	511.5	500.5
MedDDPM	200	0.10	0.09	0.29	0.16	776.0	273.6	119.4	389.7
MedSyn	200	0.44	0.11	0.39	0.31	33.1	267.4	309.0	203.2
GuideGen	200	**0.71**	**0.21**	**0.64**	**0.52**	**8.4**	**227.0**	**84.5**	**106.6**

usability for tumor segmentation



Usability under different number of synthesized samples (0, 100, 200, 500, 1K)

Dataset	Method	No. Train Cases	DSC↑ Spleen	Kidneys	Liver	Sto.	Pan.	Adr.	Eso.	Aorta	IVC	Gall.	Duo.	Blad.	PV&SV	Avg.
BTCV	Real	24	0.92	0.79	0.94	0.86	0.7	0.6	0.71	0.89	0.81	0.52	-	-	0.52	0.74
	MAISI	200	0.91	0.89	0.94	0.80	0.61	0.44	0.60	0.84	0.78	0.29	-	-	0.48	0.69
	Zhuang's	200	0.90	0.88	0.95	0.83	0.65	0.54	0.69	0.88	0.82	0.49	-	-	0.56	0.74
	MedDDPM	200	0.92	0.90	0.95	0.87	0.66	0.54	0.68	0.88	0.84	0.39	-	-	0.49	0.74
	MedSyn	200	0.89	0.90	0.96	0.81	0.65	0.56	0.70	0.86	0.86	0.39	-	-	0.32	0.72
	GuideGen	200	**0.96**	**0.91**	**0.98**	**0.90**	**0.76**	**0.62**	**0.74**	**0.92**	**0.90**	**0.49**	-	-	**0.57**	**0.79**
AMOS	Real	240	0.95	0.94	0.96	0.89	0.81	0.67	0.78	0.92	0.87	0.72	0.76	0.82	-	0.84
	MAISI	200	0.83	0.84	0.91	0.74	0.60	0.50	0.60	0.81	0.71	0.34	0.53	0.43	-	0.65
	Zhuang's	200	0.85	0.87	0.90	0.66	0.63	0.46	0.61	0.82	0.73	0.26	0.55	0.62	-	0.66
	MedDDPM	200	0.86	0.89	0.90	0.74	0.61	0.46	0.61	0.86	0.76	0.43	0.55	0.60	-	0.69
	MedSyn	200	0.85	0.90	0.91	0.70	0.64	0.52	0.62	0.80	0.69	0.21	0.53	0.64	-	0.67
	GuideGen	200	**0.95**	**0.92**	**0.95**	**0.90**	**0.70**	**0.52**	**0.73**	**0.88**	**0.82**	**0.60**	**0.67**	**0.72**	-	**0.78**

usability for multi-organ segmentation

Contributions

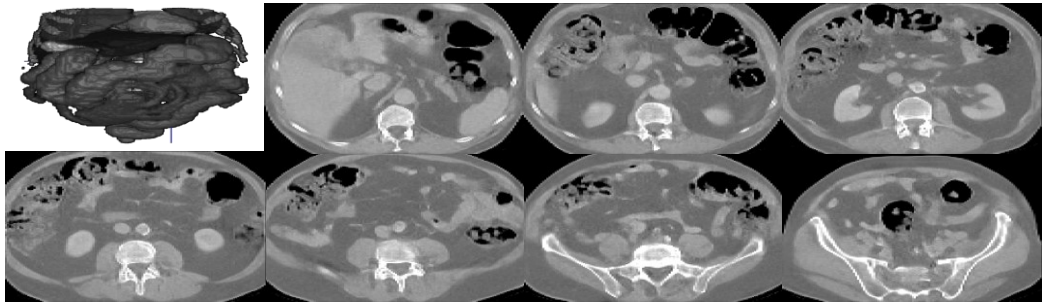
- We propose *GuideGen*, a viable and novel solution capable of performing text-based full-torso anatomy and CT generation.
- We introduce a text-conditional semantic synthesizer that avoids ambiguity in current anatomy generations, an anatomy-aware HDR autoencoder to accommodate different types of anatomical structures as well as high intensity variations in full-torso CT volumes and a latent-guided feature generator to generate faithful CT representations.
- We conduct thorough experiments comparing the samples generated by *GuideGen* with 6 state-of-the-art generative frameworks from their quality, conditional consistency and downstream usability, on all of which *GuideGen* achieves better performances.

Limitations

While our *GuideGen* achieves strong performance to synthesize full-torso semantic and CTs from structured prompts given by a medical Large Language Model (by giving it a template and let it fill the blanks inside the template with information extracted from radiology reports/structured data), its cannot directly generate coherent images from free-text inputs (with failure case shown below). We consider this as a potential future direction for enabling more versatile control.

Failure Case (no tumor mask generated):

Input: *A person diagnosed with colon cancer*



Thanks for your interest to our work !

Arxiv

https://arxiv.org/abs/2403.07247

Github

https://github.com/OvO1111/GuideGen/tree/main

Presentation Slides for AAAI-26