文本引导的半监督医学图像分割

AMVLM: Alignment-Multiplicity Aware Vision-Language Model for Semi-Supervised Medical Image Segmentation

Qingtao Pan , Zhengrong Li , Wenhao Qiao , Jingjiao Lou, Qing Yang ,Guang Yang, and Bing Ji*

https://github.com/QingtaoPan/AMVLM


IEEE TRANSACTIONS ON
MEDICAL IMAGING
A PUBLICATION OF
THE IEEE ENGINEERING IN MEDICINE AND BIOLOGY SOCIETY
THE IEEE NUCLEAR AND PLASMA SCIENCES SOCIETY
THE IEEE SIGNAL PROCESSING SOCIETY
THE IEEE ULTRASONICS, FERROELECTRICS, AND FREQUENCY CONTROL SOCIETY

NOVEMBER 2025 VOLUME 44 NUMBER 11 ITMID4 (ISSN 1558-254X)

PAPERS

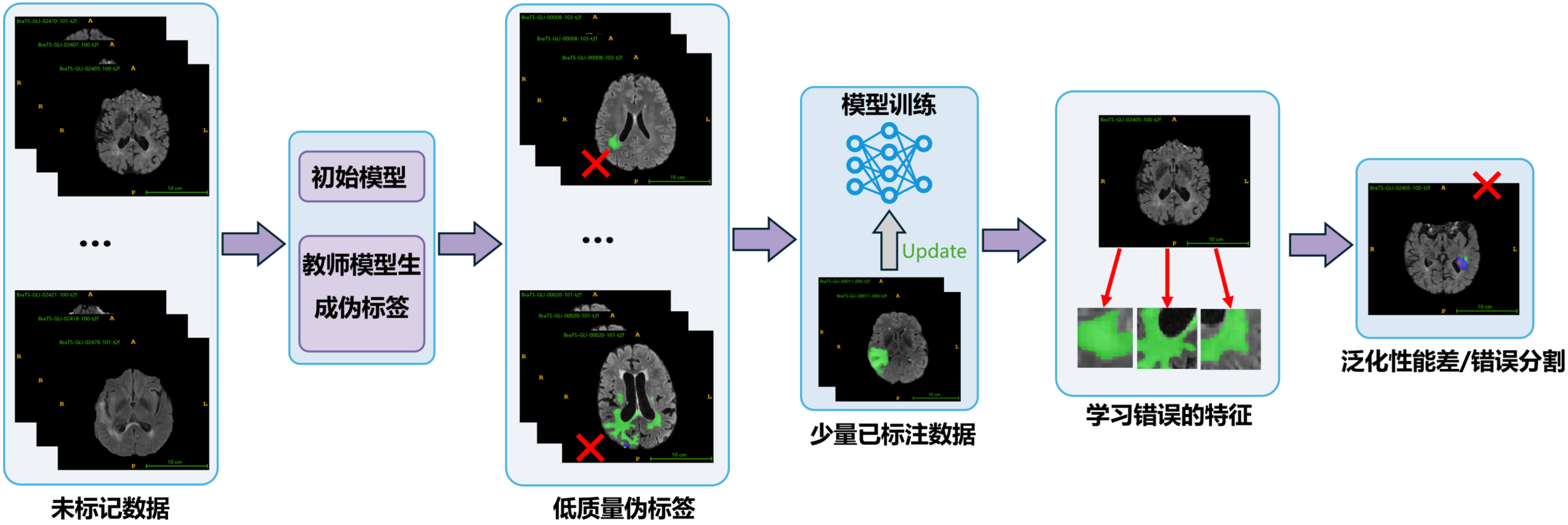
Domain-Generalized Discrete Diffusion Model for Cross-Domain Medical Image Segmentation ... *H. Yang, W. Hua, Z. Xu, and J. Sun* 4131
Improving CXR Bone Suppression by Exploiting Domain-Level and Instance-Level Information ... *L. Yang, Z. Chen, K. Wang, and L. Zhou* 4143
SA-Seg: Annotation-Efficient Segmentation for Airway Tree Using Saliency-Based Annotation ... *K. Zhou, N. Chen, Z. Yi, and X. Xu* 4156
HERON: High-Efficiency Real-Time Motion Quantification and Re-Acquisition for Fetal Diffusion MRI ... *J. A. Verdera, A. Bortolazzi, S. N. Silva, K. Payette, K. S. Clair, S. McElroy, S. Malik, J. Hajnal, R. Tomi-Tricot, M. Rutherford, and J. Hutter* 4171
Disentangled Pseudo-Bag Augmentation for Whole Slide Image Multiple Instance Learning ... *J. Dong, J. Jiang, K. Jiang, J. Li, L. Cai, and Y. Zhang* 4181
Dynamic Client Distillation for Semi-Supervised Federated Learning in a Realistic Scenario ... *N. Shen, T. Xu, S. Huang, Z. Chen, and J. Li* 4198
CLIF-Net: Intersection-Guided Cross-View Fusion Network for Infection Detection From Cranial Ultrasound ... *M. Yu, M. R. Peterson, K. Burgoine, T. Harbaugh, P. Olupot-Olupot, M. Gladstone, C. Hagmann, F. M. Cowan, A. Weeks, S. U. Morton, R. Mulondo, E. Mbabazi-Kabachelor, S. J. Schiff, and V. Monga* 4213
Advancing Volumetric Medical Image Segmentation via Global-Local Masked Autoencoders ... *J. Zhuang, L. Luo, Q. Wang, M. Wu, L. Luo, and H. Chen* 4226

(Contents Continued on Page 4129)

Indexed in PubMed® and MEDLINE®, products of the United States National Library of Medicine

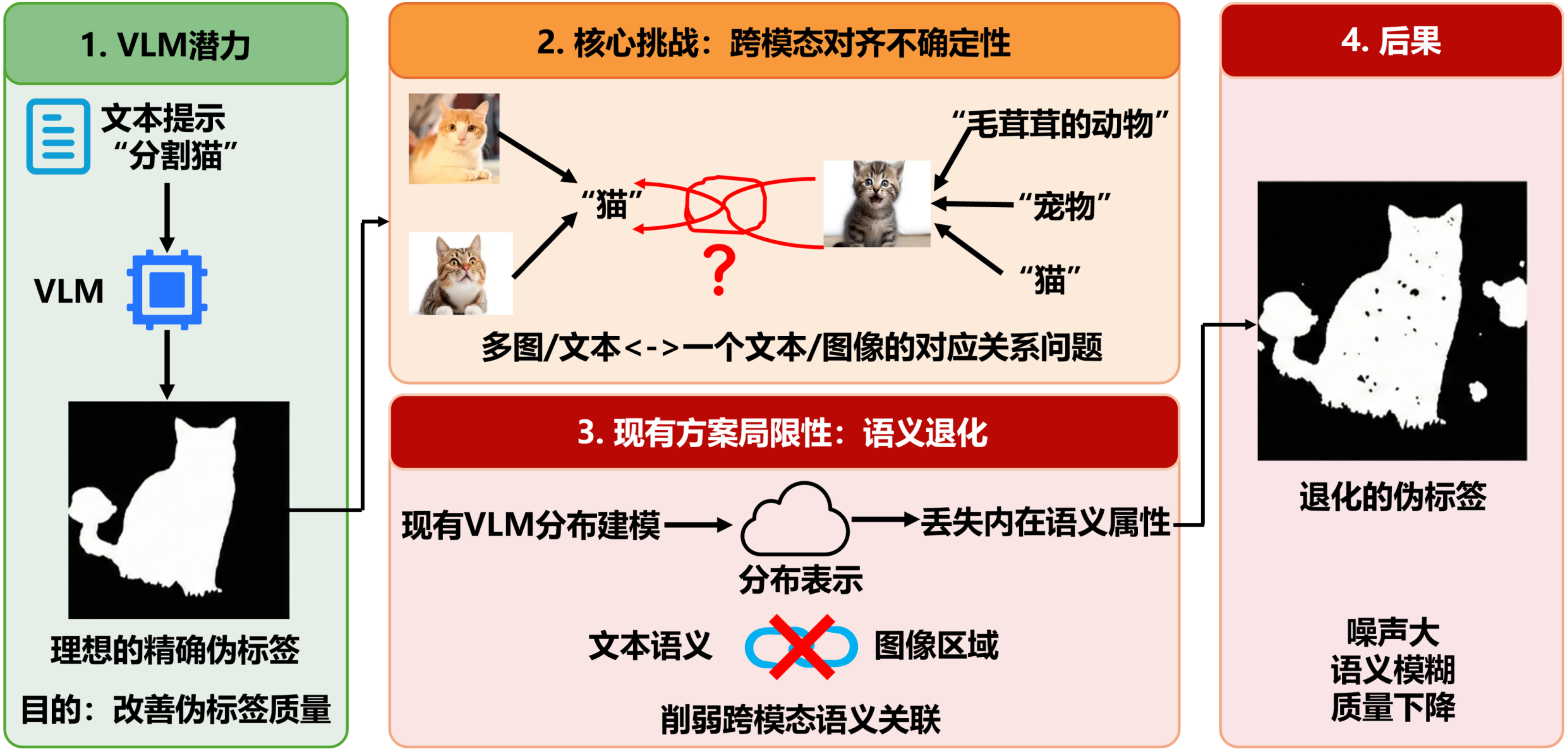
研究背景与意义

半监督医学图像分割中生成低质量的伪标签，导致模型学习错误特征



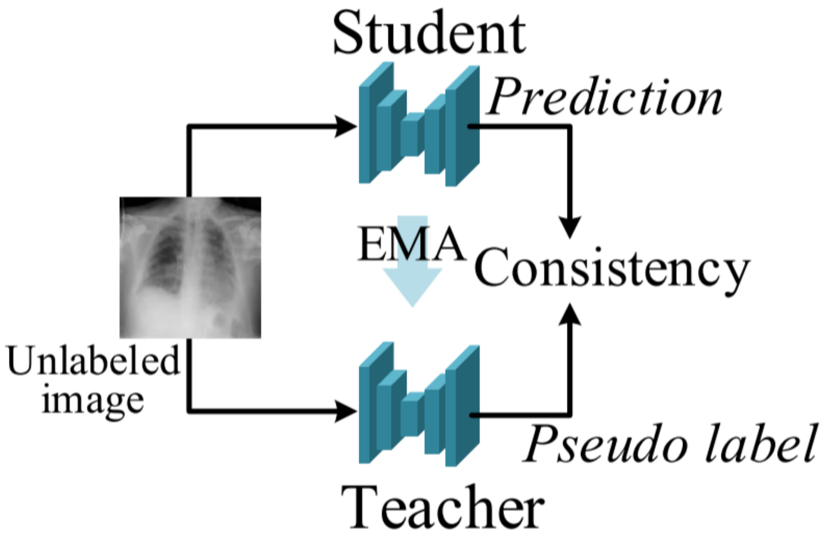
研究背景与意义

视觉语言模型面临跨模态对齐不确定性挑战



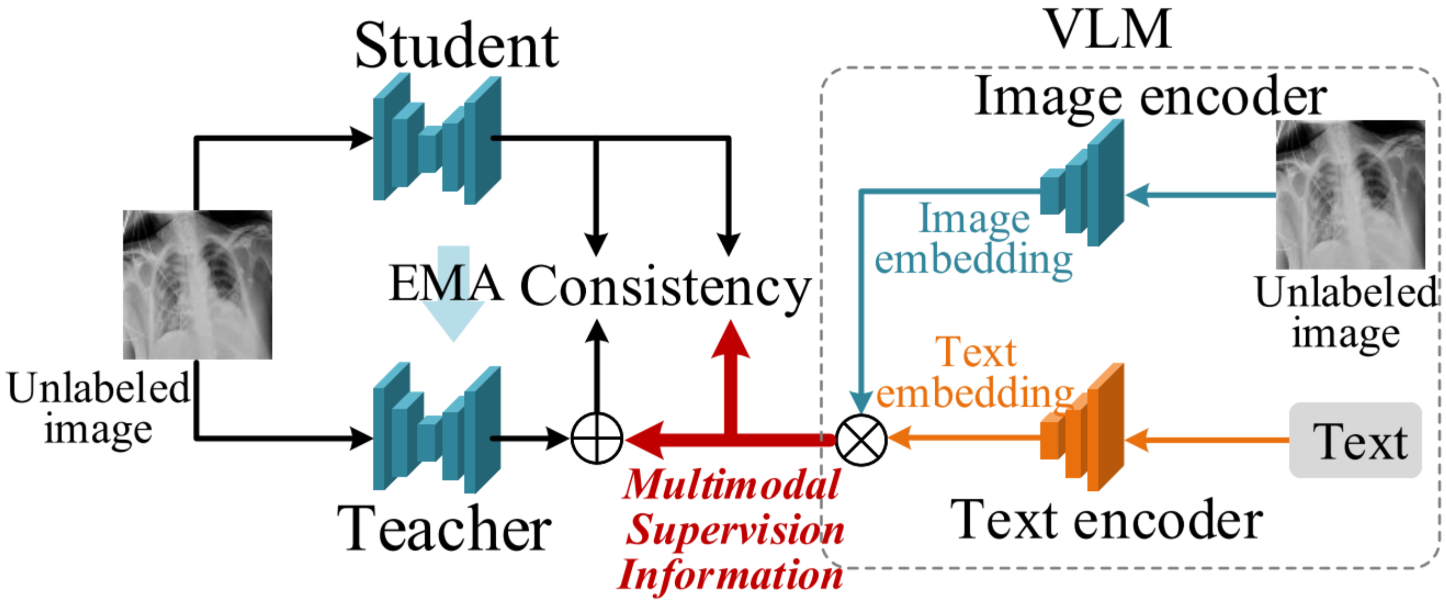
研究背景与意义

本文提出预训练范式AMVLM，并构建文本引导的SSMIS网络以增强伪标签质量



(a) 传统的伪标签半监督学习

缺点：产生低质量的伪标签

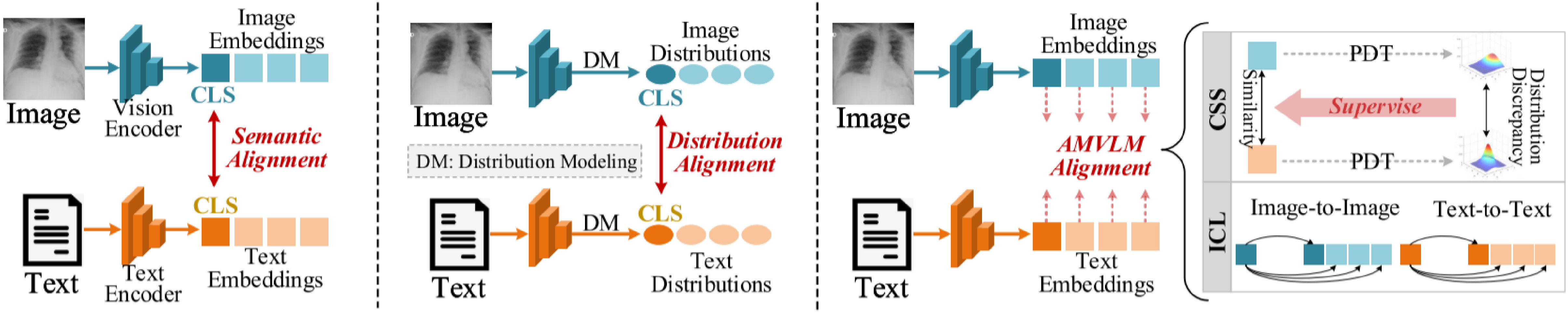


(b) 视觉语言模型引导的伪标签半监督学习

优点：利用多模监督信息产生高质量的伪标签

本文的贡献

- 为了解决跨模态对齐不确定性，提出了一种新的VLM预训练范式AMVLM
- 跨模态相似性监督提出了一种概率分布转换(PDT)来监督跨细粒度语义的相似性得分
- 模态内对比学习对粗-细粒度信息进行相似性度量

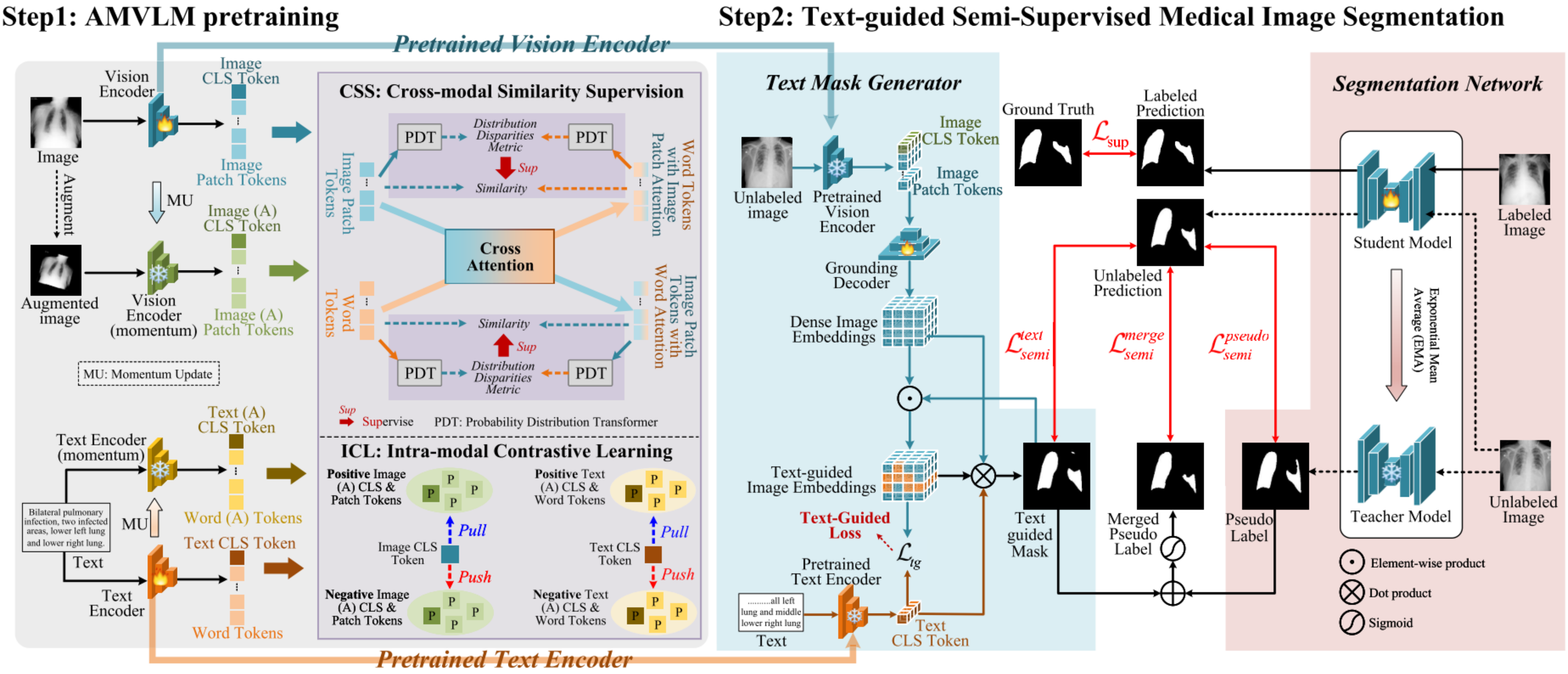


(a)基于语义的图像文本对齐　(b)基于分布的图像文本对齐　(c)基于AMVLM的图像文本对齐

研究方法

◻ AMVLM驱动的半监督医学图像分割

CLS Toke：汇总图像全局特征



研究方法

- 步骤一：AMVLM预训练

  a) 跨模态相似性监督(CSS)提出了概率分布转换(PDT)，通过学习跨模态分布表示差异来监督跨细粒度语义的相似性得分。
  b) 模态内对比学习(ICL)鼓励每个模态内粗-细粒度信息的相似性度量，以提高跨模态语义一致性。

- 步骤二：文本引导半监督医学图像分割

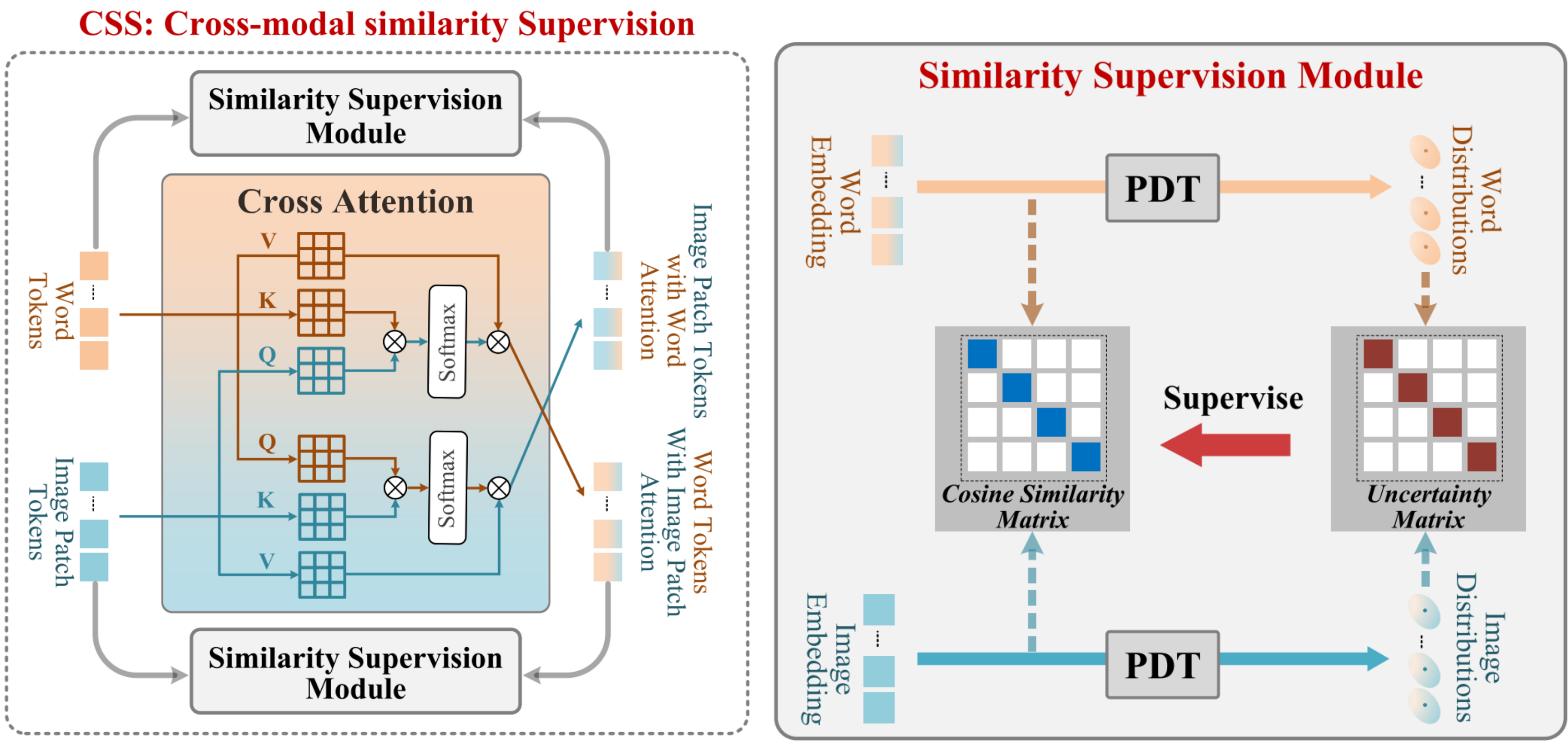
  a) 构建文本掩码生成器，利用预训练好的AMVLM生成多模态监督信息，并将其注入一致性正则化过程中，从而提高伪标签的质量和模型的一致性学习。

多重对齐感知视觉语言模型 (AMVLM)

◻ 跨模态相似性监督

a) 使用双向交叉注意力在生成的Image Patch Token 和Word Token之间进行软匹配。对齐局部图像和文本嵌入，确保文本中的每个单词与图像中的特定区域匹配。

b) 为了在软匹配过程中学习多个对齐，PDT将语义嵌入建模为分布表示，监督软匹配过程的余弦相似度。



多重对齐感知视觉语言模型 (AMVLM)

□ 跨模态相似性监督

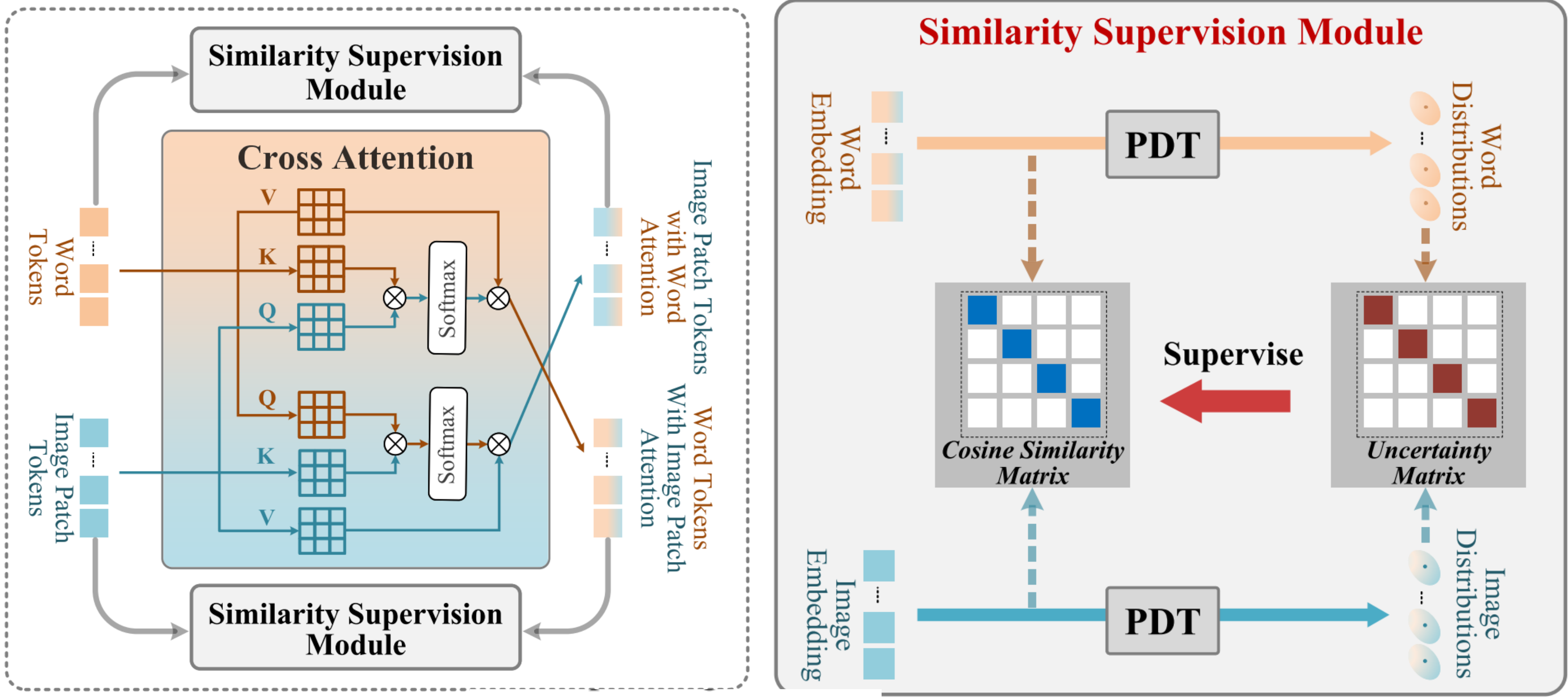
Image-Text Pair： $(x_{v,i}, x_{t,i})$

Image Patch Tokens: $V_i = \{v_i^1, v_i^2, \ldots, v_i^P\}$

Word Tokens: $T_i = \{t_i^1, t_i^2, \ldots, t_i^W\}$

v_i^p 关注 T_i 中的所有Word Token

CSS: Cross-modal similarity Supervision



基于注意力的第p个Word Token: $\alpha_i^{p2k} = \mathrm{softmax}\left(\frac{(Qv_i^p)^T (Kt_i^k)}{\sqrt{d}}\right), \quad \tilde{t}_i^p = \sum_{k=1}^{N} \alpha_i^{p2k} (Vt_i^k),$

图像到文本的细粒度对齐损失: $\mathcal{L}_{V2T} = -\frac{1}{2BP} \sum_{i=1}^{B} \sum_{p=1}^{P} \left(\log \frac{\exp(sim(v_i^p, \tilde{t}_i^p)/\tau)}{\sum_{k=1}^{P} \exp(sim(v_i^p, \tilde{t}_i^k)/\tau)} + \log \frac{\exp(sim(\tilde{t}_i^p, v_i^p)/\tau)}{\sum_{k=1}^{P} \exp(sim(\tilde{t}_i^p, v_i^k)/\tau)}\right),$

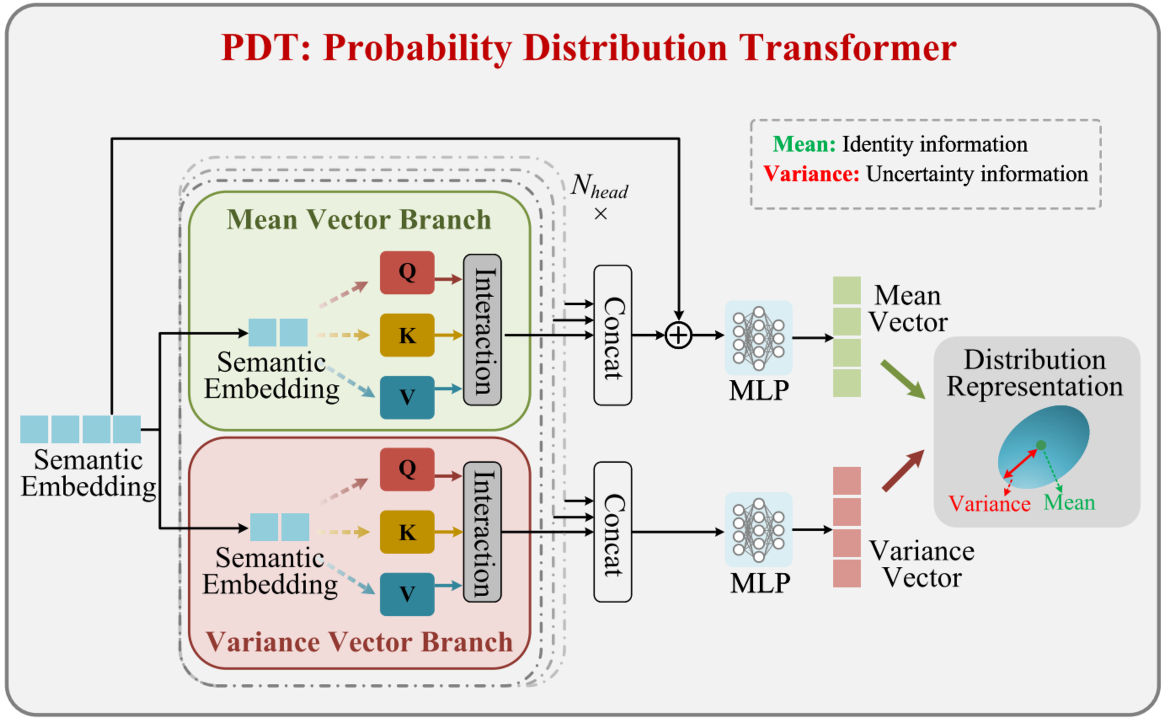
$$sim(v_i^p, \tilde{t}_i^p) = (v_i^p)^T \tilde{t}_i^p$$

多重对齐感知视觉语言模型 (AMVLM)

☐ 概率分布转换 (PDT)

a) 概率分布转换的目是衡量 Image Patchs和Words的对齐不确定性。

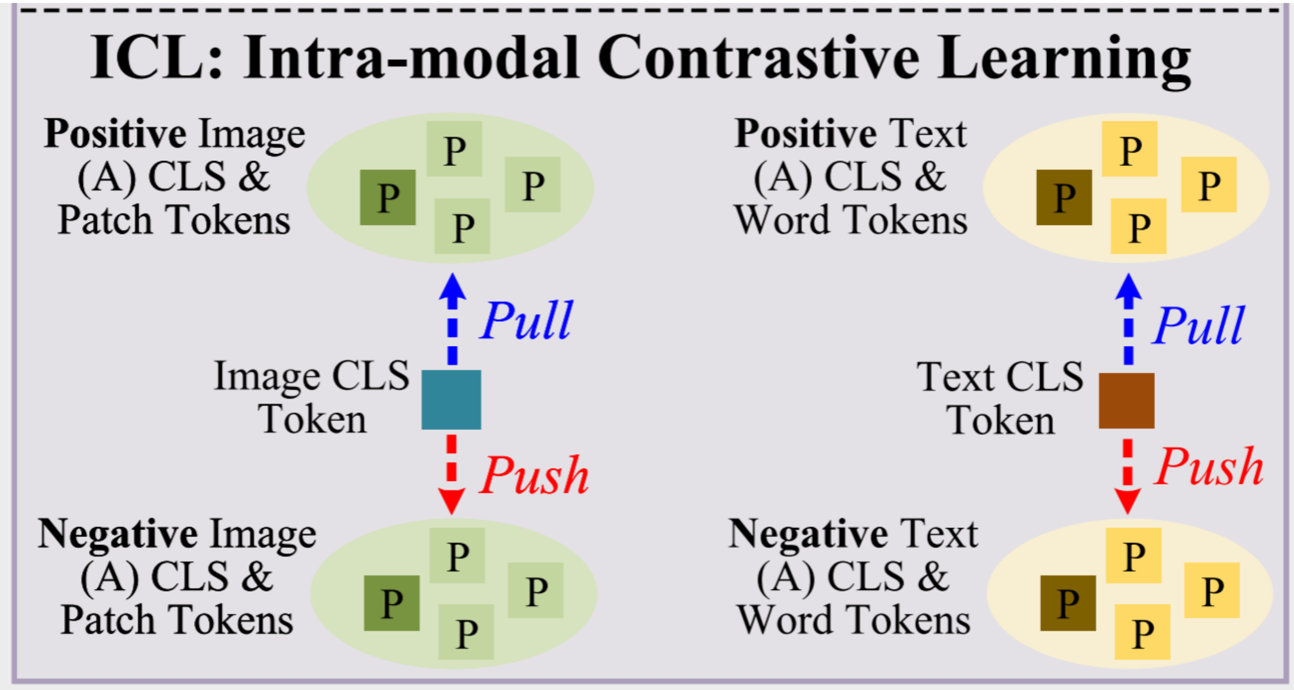
b) PDT预测每个输入语义嵌入的均值向量和方差向量，其中均值向量表示身份信息，方差向量表示分布范围。



多重对齐感知视觉语言模型 (AMVLM)

◻ 模态内对比学习 (ICL)

ICL在每个模态内进行粗-细粒度的对比学习，学习每个模态表示的相似性和差异性，从而提高跨模态语义一致性。



粗粒度图像到图像的对比学习损失：

$$\mathcal{L}_{V2V}^{coarse} = -\frac{1}{B}\sum_{i=1}^{B}(\log\frac{\exp(sim(v_i^{cls}, v_{i,a}^{cls})/\tau)}{\sum_{j=1}^{B}\exp(sim(v_i^{cls}, v_{j,a}^{cls})/\tau)},$$

细粒度图像到图像的对比学习损失：

$$\mathcal{L}_{V2T}^{fine} = -\frac{1}{BS}\sum_{i=1}^{B}\sum_{p=1}^{S}(\log\frac{\exp(sim(v_i^{cls}, v_{i,a}^{p})/\tau)}{\sum_{j=1}^{B}\exp(sim(v_i^{cls}, v_{j,a}^{p})/\tau)},$$

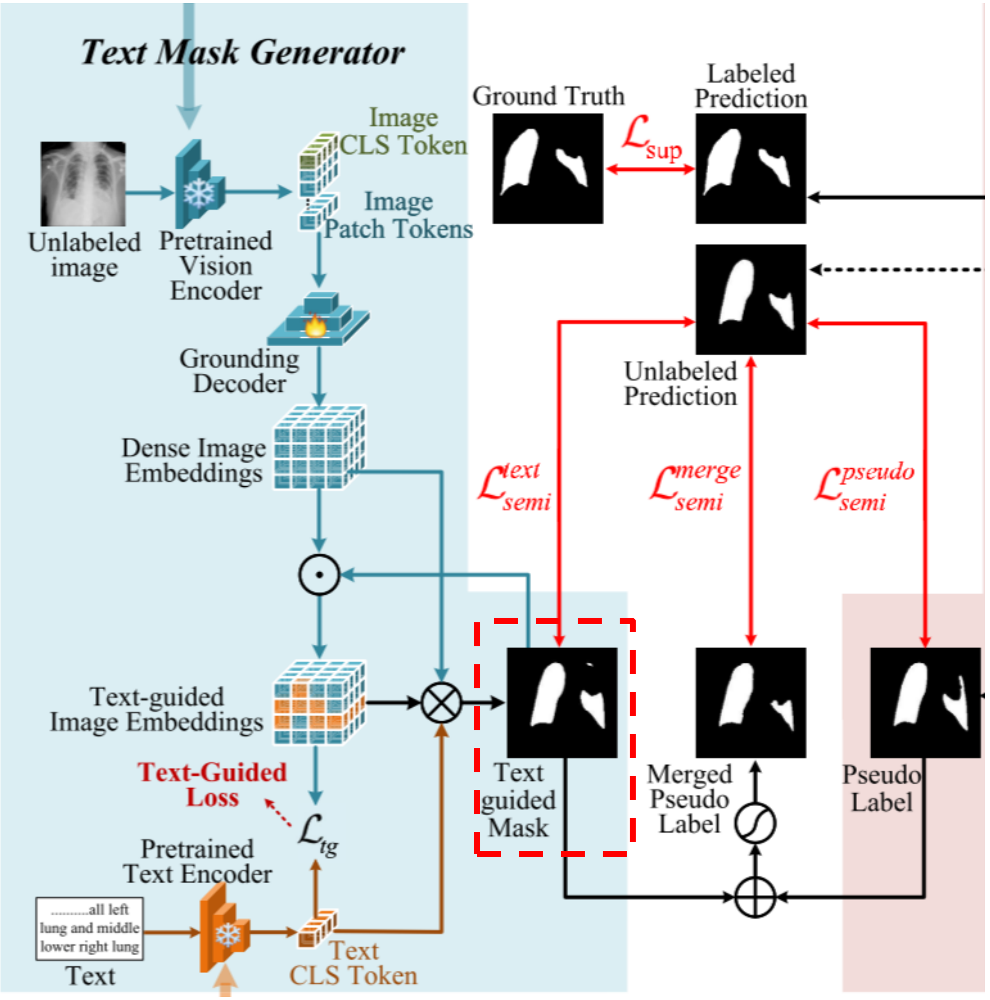
文本引导半监督医学图像分割(SSMIS)

◻ 文本掩码生成器

a) 文本掩码生成器通过整合文本提示和密集图像嵌入来生成多模态监督信息(即文本引导掩码)。使用预训练的视觉和文本编码器提取image patch token和text CLS token。

b) 为了抑制无关的文本引导掩码生成，设计了一个文本引导损失 L_{tg}，它是通过文本引导图像嵌入和text CLS Token之间的InfoNCE损失来实现的:

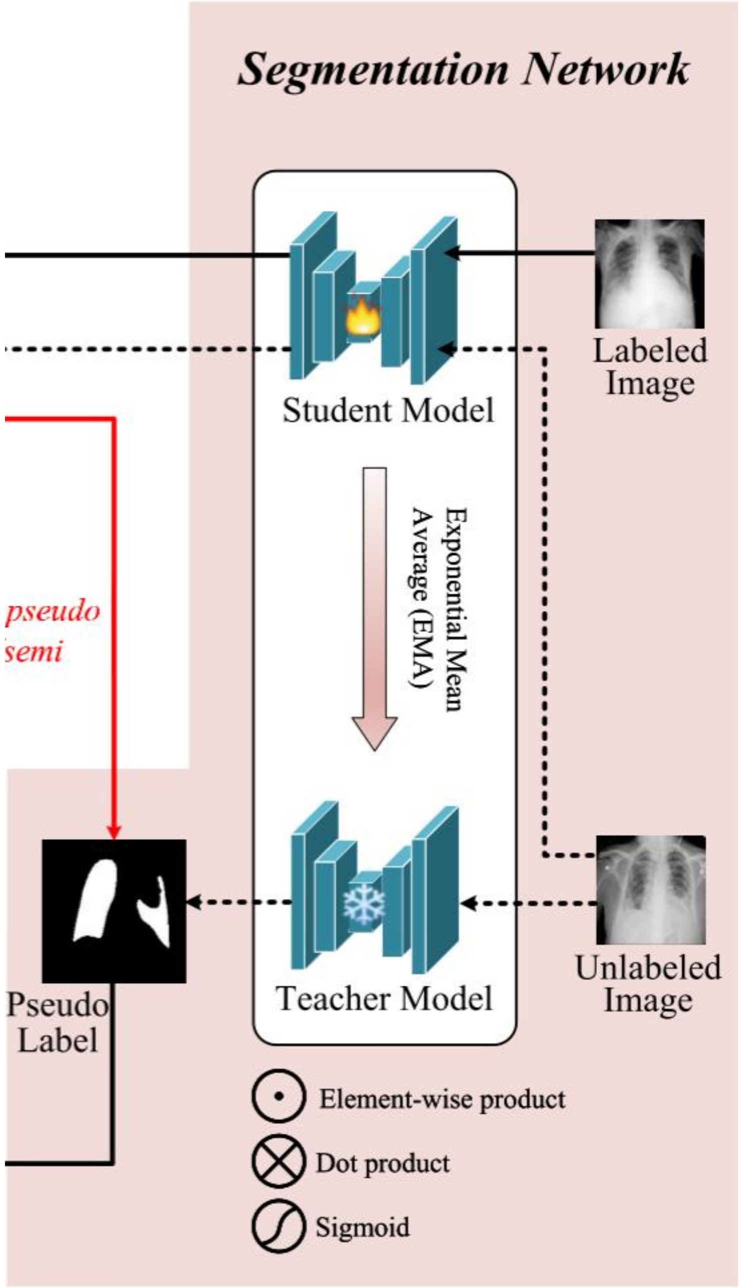
$$\mathcal{L}_{tg} = -\frac{1}{2B}\sum_{i=1}^{B}(\log\frac{\exp(sim(v_{text,i}, t_{cls,i})/\tau)}{\sum_{j=1}^{B}\exp(sim(v_{text,i}, t_{cls,j})/\tau)} + \log\frac{\exp(sim(t_{cls,i}, v_{text,i})/\tau)}{\sum_{j=1}^{B}\exp(sim(t_{cls,i}, v_{text,j})/\tau)}).$$



文本引导半监督医学图像分割(SSMIS)

◻ 分割网络

Mean-Teacher(MT)用于分割网络。它包括两个细分网络，即学生和教师。教师网络由学生的指数移动平均线(EMA)更新，学生网络预测标记的图像和未标记的图像，教师网络生成未标记图像的伪标签。

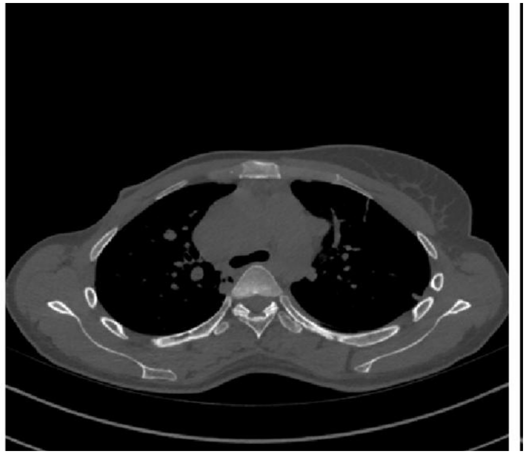


实验过程

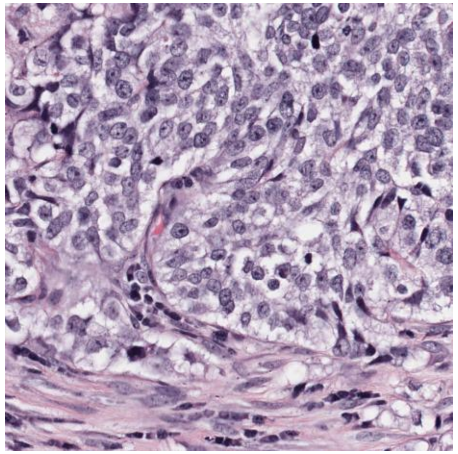
□ 数据集

QaTa-COV19、BM-Seg、MoNuSeg和MRSpineSeg

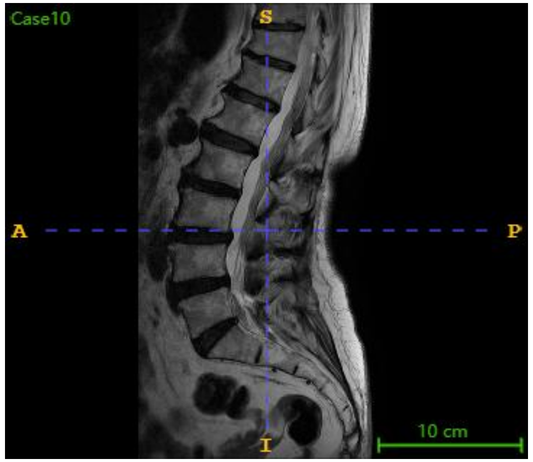
(a) COVID-19胸部X光片



(b) 骨转移分割



(c) 细胞核实例分割



(d) 多类别脊柱结构分割

□ 评估指标

Dice和mIoU C：类别数量 M：像素数量

$$Dice = \sum_{m=1}^{M}\sum_{c=1}^{C}\frac{1}{MC}\cdot\frac{2|p_{mc}\cap y_{mc}|}{(|p_{mc}|+|y_{mc}|)}$$

$$mIoU = \sum_{m=1}^{M}\sum_{c=1}^{C}\frac{1}{MC}\cdot\frac{|p_{mc}\cap y_{mc}|}{|p_{mc}\cup y_{mc}|}$$

实验过程

□ 对比实验

Methods	Text	QaTa-COV19				BM-Seg				MoNuSeg				MRSpineSeg			
		25% labeled		50% labeled		25% labeled		50% labeled		25% labeled		50% labeled		25% labeled		50% labeled	
		Dice	mIoU	Dice	mIoU	Dice	mIoU	Dice	mIoU	Dice	mIoU	Dice	mIoU	Dice	mIoU	Dice	mIoU
U-Net [44]	✗	–	–	79.02	69.46	–	–	74.15	60.32	–	–	78.66	68.46	–	–	82.36	74.07
MT [19]	✗	77.88	67.53	79.64	71.88	66.31	53.35	72.31	56.99	72.80	56.70	75.15	59.64	77.11	67.65	79.31	69.05
CCT [5]	✗	78.02	67.03	80.25	71.69	70.66	55.27	72.65	60.54	74.25	56.55	76.23	55.87	78.51	68.06	80.26	70.11
BCP [2]	✗	74.79	65.26	75.57	65.31	71.26	55.28	74.21	61.25	72.06	54.69	72.17	50.88	70.35	58.97	70.68	59.68
MC-Net [3]	✗	74.58	64.26	74.98	61.97	69.67	53.36	73.37	60.74	71.53	48.12	71.59	48.75	73.34	62.27	75.09	64.18
SS-Net [45]	✗	67.93	58.13	68.37	56.73	70.21	54.68	73.59	60.89	71.80	47.80	73.05	49.88	73.71	63.22	74.60	64.30
UCMT [1]	✗	76.09	64.13	77.81	68.65	71.22	55.82	74.32	60.41	75.53	54.82	76.36	59.29	76.42	65.78	77.09	66.81
SemiSAM [51]	✗	78.05	66.84	80.19	71.55	69.28	54.12	73.69	57.44	72.17	56.94	75.44	61.37	77.68	68.11	80.08	68.76
KnowSAM [52]	✗	77.42	66.74	80.41	71.68	70.39	55.27	73.98	58.01	74.66	57.06	77.68	64.52	78.36	68.47	79.72	69.54
CPC-SAM [53]	✗	76.84	65.26	76.59	66.41	67.25	52.07	72.85	58.46	71.32	52.07	75.44	58.71	75.88	65.40	78.11	67.61
LViT [41]	✓	78.12	66.75	80.32	72.16	69.45	54.26	73.15	60.14	75.69	56.14	76.57	65.44	79.64	68.11	80.32	70.48
CLIP [8]	✓	78.05	65.41	80.45	69.87	71.22	56.87	75.16	59.43	75.72	56.14	77.42	65.21	77.35	67.24	77.92	67.56
MAP [11]	✓	78.65	65.19	79.69	70.31	71.68	57.69	75.13	59.65	75.42	54.86	76.55	64.33	79.45	70.18	80.04	71.93
CMITM [47]	✓	78.04	65.84	81.33	72.84	70.63	54.66	74.96	60.18	75.14	54.28	77.63	66.31	78.22	68.43	79.64	70.21
ASG [48]	✓	77.92	65.09	80.47	71.84	70.26	54.33	74.22	59.47	75.69	55.92	76.79	65.91	77.61	67.92	79.11	69.45
MGCA [10]	✓	78.17	67.03	81.24	73.56	70.19	53.67	75.17	**61.49**	76.14	56.22	77.06	65.23	80.06	70.45	81.42	**72.11**
LA-VLM [54]	✓	77.43	65.21	79.64	70.28	71.33	55.96	74.20	61.22	75.06	54.89	77.34	66.35	78.29	67.15	79.71	68.11
MedCLIP [9]	✓	78.62	65.55	80.93	70.04	70.12	56.39	74.46	59.11	75.61	55.74	77.91	64.51	77.01	66.59	78.30	67.99
DuSSS [49]	✓	79.00	68.21	82.52	**75.87**	71.48	55.97	74.61	61.08	76.51	58.08	**78.03**	66.93	79.88	69.41	80.55	71.03
Ours	✓	**80.62**	**69.87**	**82.57**	72.37	**73.52**	**59.71**	**77.63**	60.65	**76.63**	**58.90**	77.37	**67.34**	**80.63**	**71.39**	**81.60**	72.05

➡ 总体上，本文的文本引导半监督医学图像分割效果最好

实验过程

□ 消融实验

Num	Methods				QaTa-COV19		BM-Seg		MoNuSeg		MRSpineSeg	
	MT	CSS	ICL	$\mathcal{L}_{tg}$	Dice	mIoU	Dice	mIoU	Dice	mIoU	Dice	mIoU
No. 1	✓				77.88	67.53	66.31	53.35	72.80	56.70	77.11	67.65
No. 2	✓	✓			80.07	68.14	70.09	**58.40**	75.93	58.41	79.67	69.05
No. 3	✓		✓		78.75	66.32	69.57	57.14	75.71	57.21	78.64	67.84
No. 4	✓	✓	✓		80.45	68.92	72.49	59.52	76.44	58.62	80.51	71.15
No. 5	✓	✓	✓	✓	**80.62**	**69.87**	**73.52**	59.71	**76.63**	**58.90**	**80.63**	**71.39**

总体上，CSS+ICL+L_{tg}的效果最好，证明跨模态相似性度量和模态间对比学习的有效性

实验过程

□ 少样本任务分析

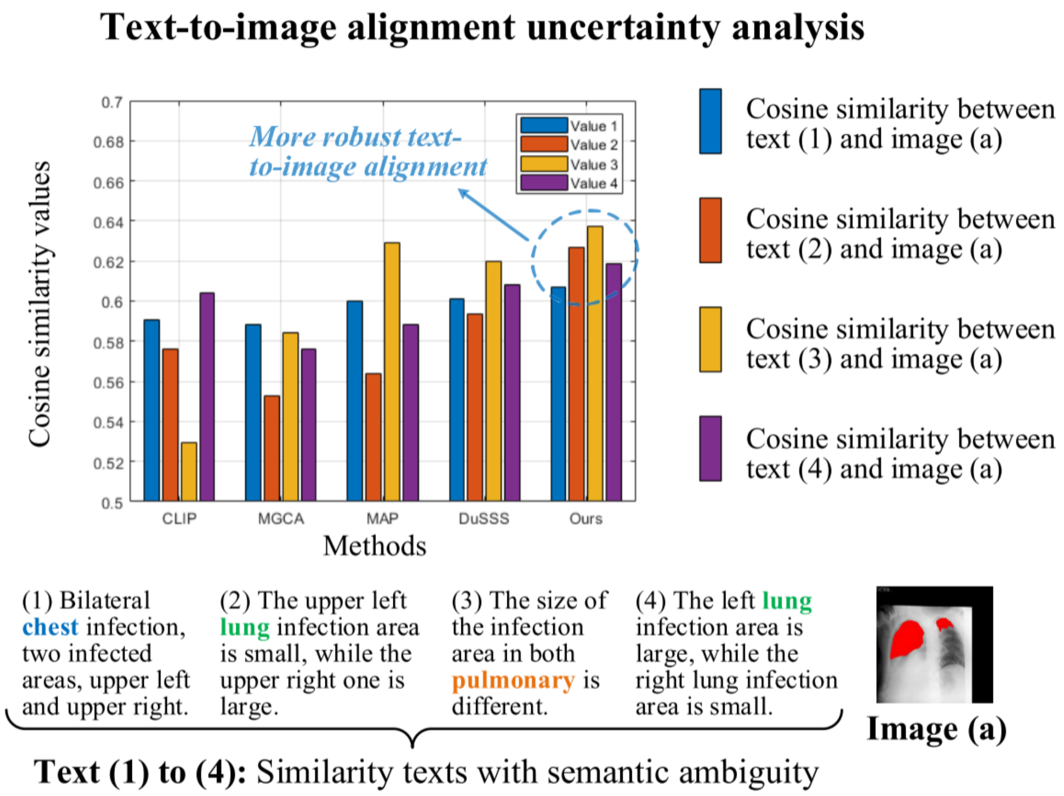
Methods	QaTa-COV19				BM-Seg				MoNuSeg				MRSpineSeg			
	1% Training data Train/Test: 71/2113		10% Training data Train/Test: 710/2113		1% Training data Train/Test: 2/50		10% Training data Train/Test: 22/50		1% Training data Train/Test: 0/14		10% Training data Train/Test: 2/14		1% Training data Train/Test: 22/542		10% Training data Train/Test: 217/542	
	Dice	mIoU	Dice	mIoU	Dice	mIoU	Dice	mIoU	Dice	mIoU	Dice	mIoU	Dice	mIoU	Dice	mIoU
CLIP [8]	56.25	41.06	59.52	45.83	54.04	42.12	67.93	**57.59**	–	–	35.96	24.47	21.86	13.55	48.24	35.14
CMITM [47]	58.03	42.17	61.42	46.88	52.09	41.16	65.32	51.44	–	–	34.65	24.13	21.45	13.22	46.41	34.77
ASG [48]	55.61	39.02	61.43	46.22	50.24	47.55	67.08	55.43	–	–	36.11	25.42	23.52	14.83	50.36	37.19
MGCA [10]	61.47	51.89	63.67	46.72	65.04	**52.63**	68.52	55.20	–	–	35.91	**26.23**	26.52	16.90	53.98	39.97
MAP [11]	60.55	50.05	65.82	47.99	51.78	41.33	61.39	45.72	–	–	35.89	24.57	24.55	15.51	52.91	38.90
BiomedParse [55]	60.97	50.26	66.43	48.59	61.44	48.08	68.13	54.92	–	–	37.04	24.88	25.16	15.69	51.97	39.84
LA-VLM [54]	61.26	49.33	64.87	47.21	63.22	49.60	66.74	55.18	–	–	36.29	23.94	24.85	16.02	52.66	39.05
MedCLIP [9]	57.35	44.30	62.45	45.83	60.23	47.44	67.25	56.26	–	–	36.05	24.71	23.70	15.09	50.41	37.26
DuSSS [49]	60.11	50.31	66.72	48.03	62.10	47.78	67.74	53.18	–	–	36.24	23.80	25.30	15.22	52.77	35.38
Ours	**62.21**	**52.22**	**68.65**	**50.07**	**65.61**	51.86	**70.42**	56.98	–	–	**37.28**	23.75	**27.36**	**17.20**	**55.61**	**41.62**

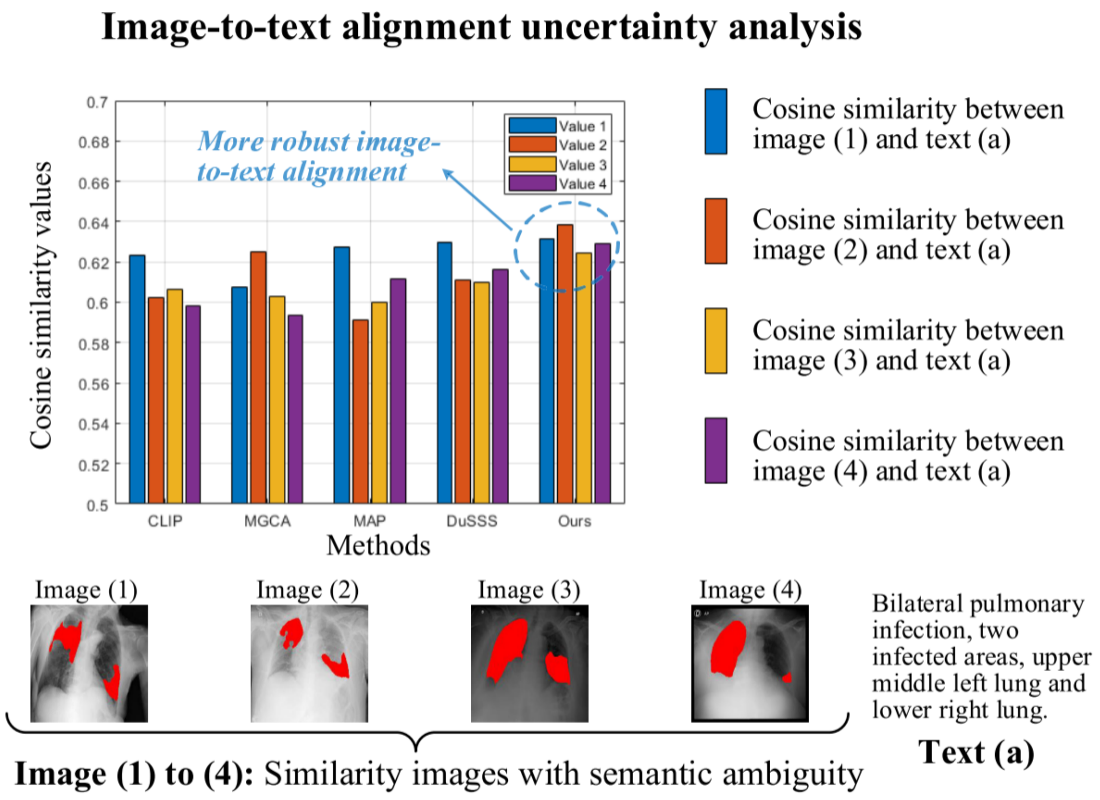
□ 与会议论文的比较

Methods	QaTa-COV19				BM-Seg				MoNuSeg				MRSpineSeg			
	SSMIS		Few Shot		SSMIS		Few Shot		SSMIS		Few Shot		SSMIS		Few Shot	
	Dice	mIoU	Dice	mIoU	Dice	mIoU	Dice	mIoU	Dice	mIoU	Dice	mIoU	Dice	mIoU	Dice	mIoU
Baseline	77.88	67.53	59.52	45.83	66.31	53.35	67.93	57.59	72.80	56.70	35.96	24.47	77.11	67.65	48.24	35.14
CMC (AAAI)	78.44	65.29	64.06	45.32	67.41	53.54	68.02	55.43	73.66	55.84	36.11	24.06	78.54	65.21	50.48	38.05
CSS (Journal)	**80.07**	**68.14**	**66.22**	**48.08**	**70.09**	**58.40**	**69.46**	**56.07**	**75.93**	**58.41**	**37.05**	**25.13**	**79.67**	**69.05**	**53.67**	**40.18**
DCL (AAAI)	78.06	65.11	61.24	43.56	68.25	55.43	67.93	55.14	73.19	54.93	35.99	**23.80**	78.09	64.98	50.14	38.26
ICL (Journal)	**78.75**	**66.32**	**64.87**	**47.33**	**69.57**	**57.14**	**68.21**	**55.74**	**75.71**	**57.21**	**36.52**	23.44	**78.64**	**67.84**	**51.64**	**38.93**
MLP (AAAI)	79.24	67.22	65.10	**52.44**	70.39	56.07	67.85	54.64	**77.14**	57.60	36.43	22.76	78.91	68.25	53.70	40.21
PDT (Journal)	**80.62**	**69.87**	**68.65**	50.07	**73.52**	**59.71**	**70.42**	**56.98**	76.63	**58.90**	**37.28**	**23.75**	**80.63**	**71.39**	**55.61**	**41.62**

实验过程

◻ 不确定性分析





AMVLM对语义歧义的相似文本显示出更稳定的余弦相似度

实验过程

◻ 文本贡献分析

Methods	QaTa-COV19		BM-Seg		MoNuSeg		MRSpineSeg	
	Dice	mIoU	Dice	mIoU	Dice	mIoU	Dice	mIoU
Vision	78.71	67.60	69.21	54.43	73.22	54.80	79.26	69.27
Vision+Text	**80.62**	**69.87**	**73.52**	**59.71**	**76.63**	**58.90**	**80.63**	**71.39**

具有文本引导的视觉方法在四个数据集上都比单独的视觉方法好

实验过程

□ 统计学分析

为了评价所提出的方法是否具有统计学意义，我们的方法与其他方法进行配对t检验，其中a = 0.05

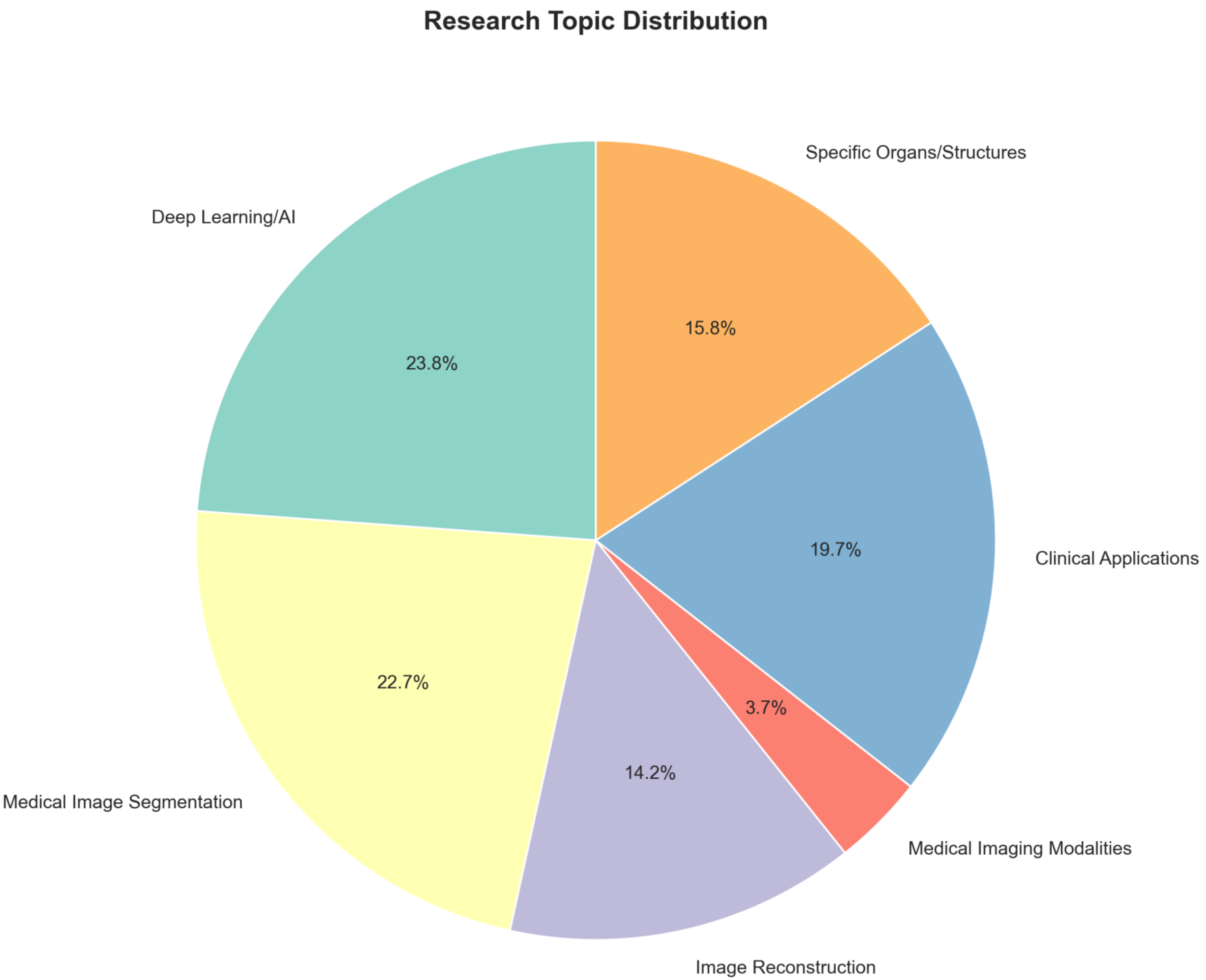
Methods	p-values			
	QaTa-COV19	BM-Seg	MoNuSeg	MRSpineSeg
vs. MT	0.0058	0.0001	0.0037	0.0003
vs. CCT	0.0108	0.0031	0.0717	0.0143
vs. BCP	0.0002	0.0011	0.0003	0.0001
vs. MC-Net	0.0003	0.0133	0.0008	0.0004
vs. SS-Net	0.0000	0.0034	0.0007	0.0002
vs. UCMT	0.0008	0.0077	0.0115	0.0008
vs. LViT	0.0112	0.0284	0.0302	0.0862
vs. CLIP	0.0104	0.0062	0.0073	0.0031
vs. MAP	0.0007	0.0098	0.0276	0.0102
vs. CMITM	0.0114	0.0005	0.0976	0.0102
vs. ASG	0.0005	0.0041	0.0118	0.0092
vs. MGCA	0.0091	0.0244	0.1881	0.0391
vs. BiomedParse	0.0241	0.0084	0.0063	0.0369
vs. SemiSAM	0.0354	0.0080	0.0150	0.0241
vs. KnowSAM	0.0075	0.0213	0.0374	0.0192
vs. CPC-SAM	0.0045	0.0017	0.0136	0.0011
vs. LA-VLM	0.0219	0.0643	0.0421	0.0516
vs. MedCLIP	0.0541	0.0073	0.0371	0.0118
vs. DuSSS	0.0321	0.0164	0.0792	0.0284

可以得出结论，该方法具有统计显著性！

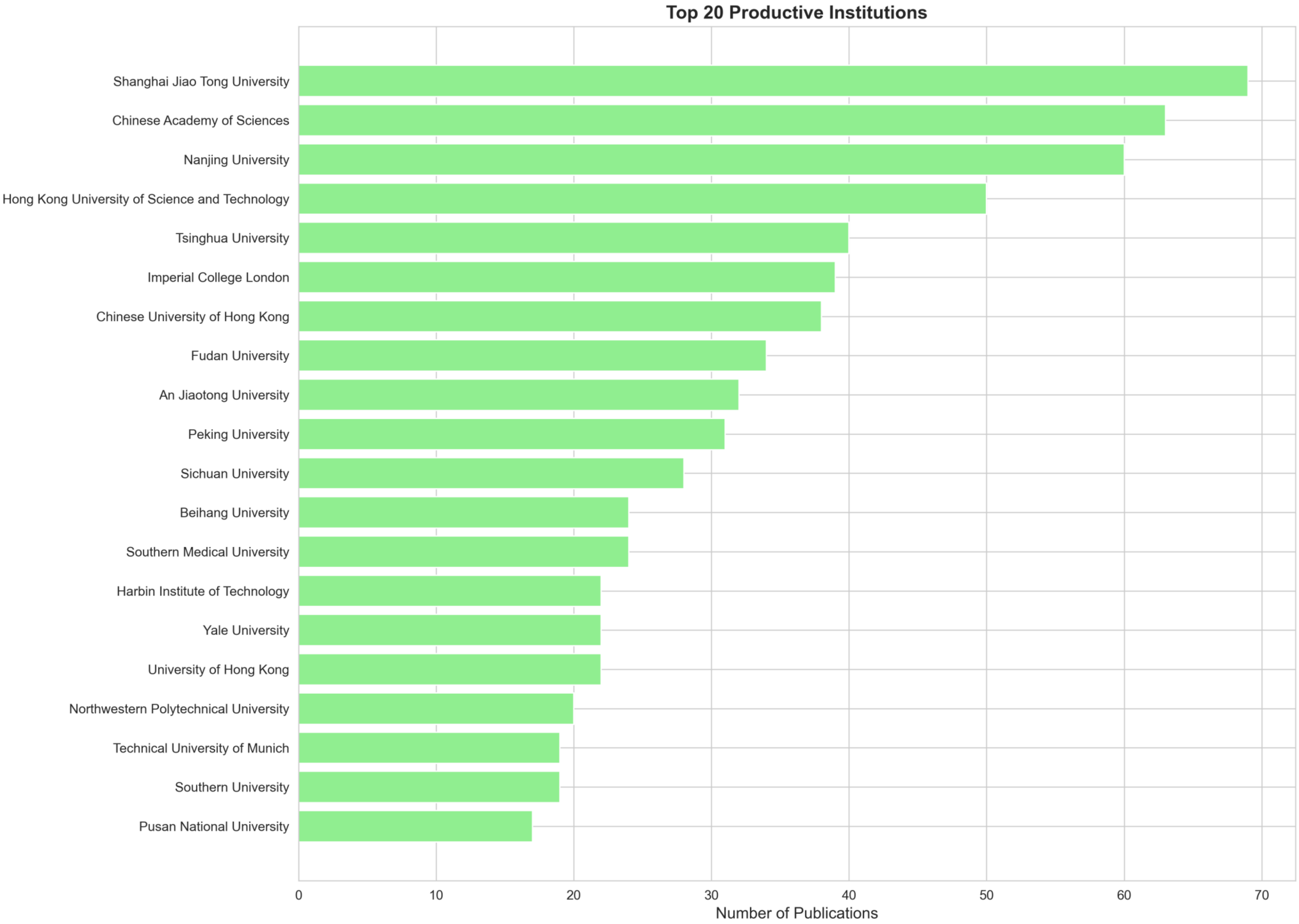
结论

- 本研究提出了一种创新的视觉语言模型预训练范式：AMVLM，该模型通过跨模态相似度监督和模态内对比学习有效解决了跨模态对齐不确定性问题，增强了跨模态表示的多重对应关系学习。
- 在此基础上，构建了文本引导的半监督医学图像分割（SSMIS）框架，在四个公开医学影像数据集（QaTa-COV19、BM-Seg、MoNuSeg、MRSpineSeg）上的实验表明，该方法在半监督分割任务中性能优于现有半监督学习方法和视觉语言模型驱动的分割方法

TMI期刊分析

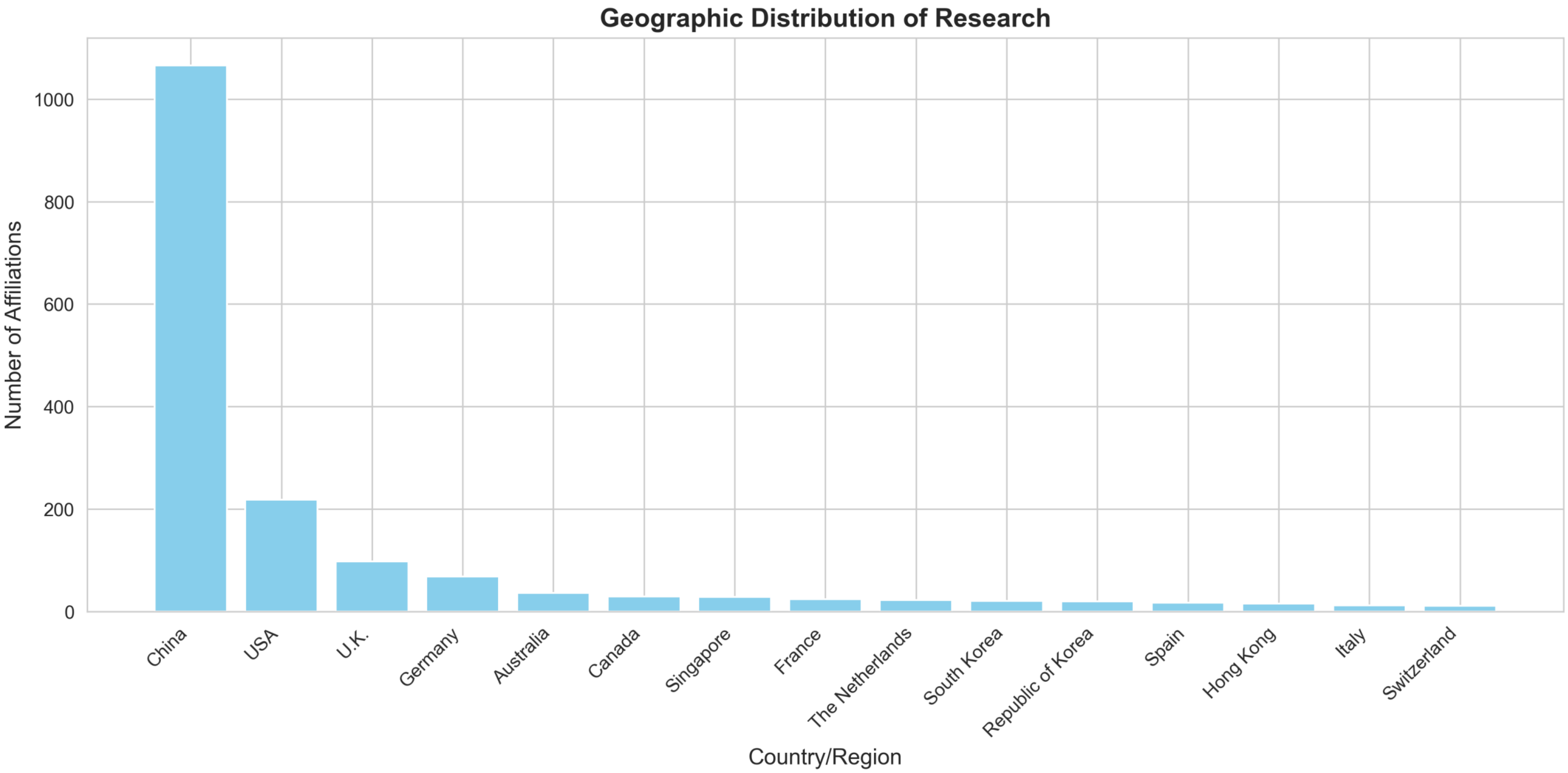

Research Topic Distribution
Specific Organs/Structures
Deep Learning/AI
15.8%
23.8%
19.7%
Clinical Applications
22.7%
3.7%
14.2%
Medical Image Segmentation
Medical Imaging Modalities
Image Reconstruction

TMI期刊分析



Top 20 Productive Institutions
Shanghai Jiao Tong University
Chinese Academy of Sciences
Nanjing University
Hong Kong University of Science and Technology
Tsinghua University
Imperial College London
Chinese University of Hong Kong
Fudan University
An Jiaotong University
Peking University
Sichuan University
Beihang University
Southern Medical University
Harbin Institute of Technology
Yale University
University of Hong Kong
Northwestern Polytechnical University
Technical University of Munich
Southern University
Pusan National University
0
10
20
30
40
50
60
70
Number of Publications

TMI期刊分析



Geographic Distribution of Research
Number of Affiliations
0
200
400
600
800
1000
China
USA
U.K.
Germany
Australia
Canada
Singapore
France
The Netherlands
South Korea
Republic of Korea
Spain
Hong Kong
Italy
Switzerland
Country/Region

Thank you