Towards Text–Mask Consistency in Medical Image Segmentation

C2Seg: Consistency-enhanced Two-stage Segmentation Framework

Jie Gui, Hang Tu, Wen Sha, Xiuquan Du
Anhui University

ICLR 2026

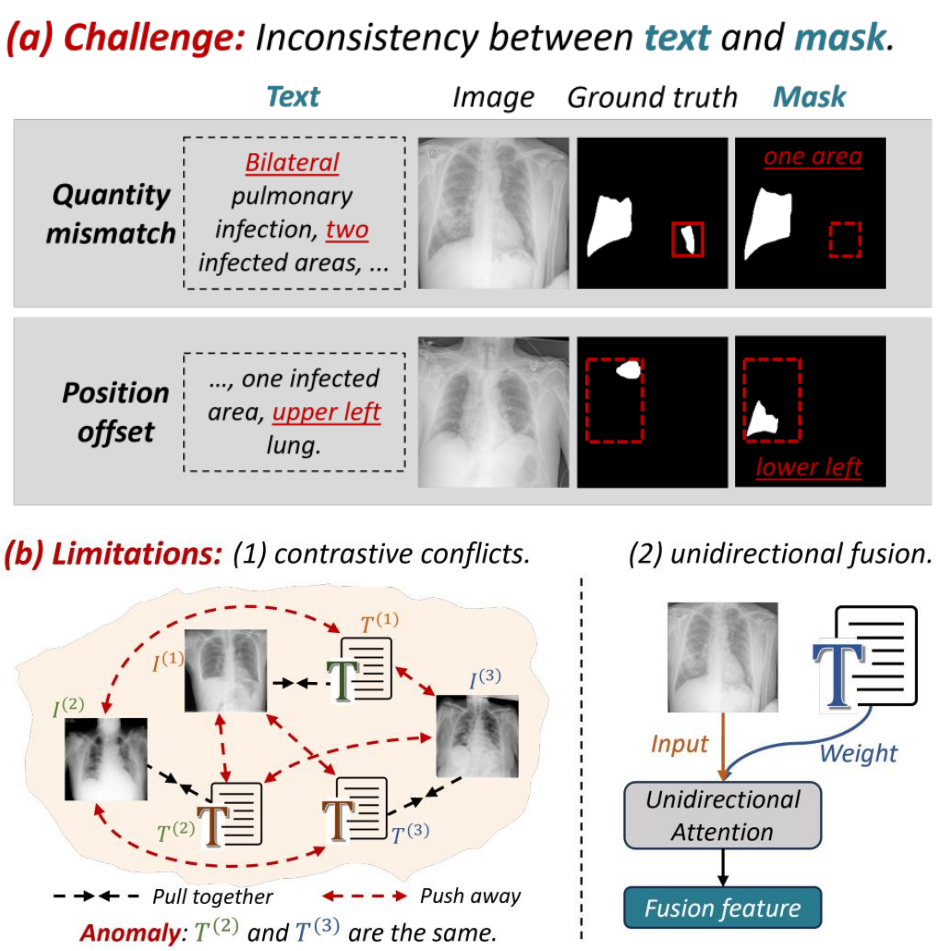
Modeling text–mask consistency for medical image segmentation

Why text-mask inconsistency persists

Masks may contradict the text even when prompts explicitly specify quantity or coarse location.

- Clinical descriptions are highly templated and repetitive, so hard one-to-one contrastive learning creates many false negatives.
- Most fusion modules are vision-centric: language modulates features, but does not form an explicit spatial representation on the pixel grid.
- The failure shows up most clearly for lesion count, laterality, and coarse position cues.

Key diagnosis: the problem is both in pretraining (alignment) and in fusion (spatial injection of language).



The illustration and investigation of text-mask inconsistency from Figure 1.

C2Seg at a glance

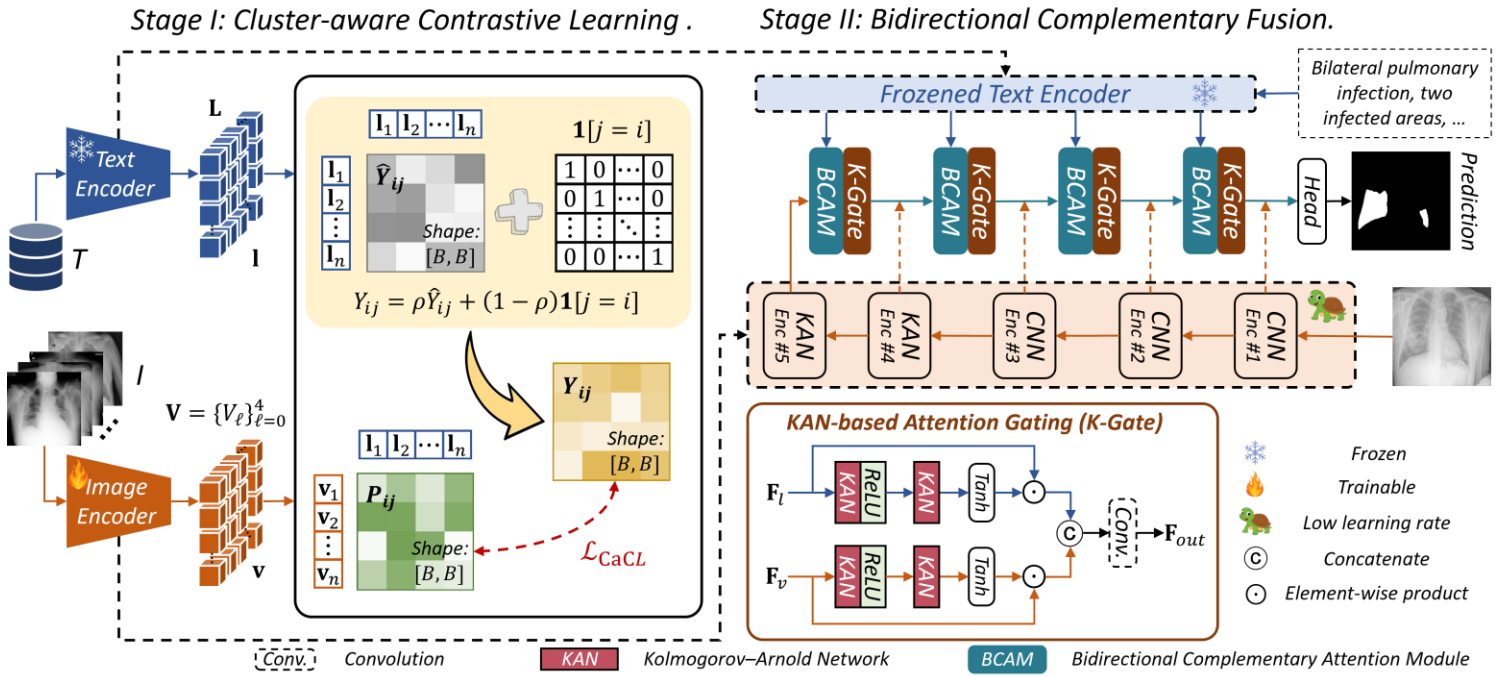
Stage I: CaCL

soft-label alignment in batch language neighborhoods

Stage II: BCAM + K-Gate

explicit bidirectional spatial interaction

Output: text-consistent mask



Overview of the proposed C2Seg from Figure 2.

Frozen text encoder + finetune visual encoder keeps semantics stable while learning better visual-spatial alignment.

Stage I: CaCL reduces false negatives

Mechanism

1. Compute batch text-text cosine similarity in frozen language space.
2. Apply row-mean debiasing and non-negativity clipping to suppress template bias.
3. Convert similarities into soft targets and mix them with the true matched pair.
4. Train bidirectionally so semantically similar non-matching samples are not over-repelled.

Core intuition

Replace rigid one-positive/others-negative supervision with a continuous neighborhood-aware target distribution.

$$\hat{Y}_{ij} = \frac{\exp(M'_{ij}/\tau)}{\sum_k \exp(M'_{ik}/\tau)}.$$

$$Y_{ij} = \rho\,\hat{Y}_{ij} + (1-\rho)\,\mathbf{1}[j=i],$$

$$P_{ij}^{v\to l} = \mathrm{softmax}_j(s_{ij}/\tau)$$

$$P_{ij}^{l\to v} = \mathrm{softmax}_i(s_{ij}/\tau)$$

$$\mathcal{L}_{\mathrm{CaCL}} = -\frac{1}{B}\sum_{i=1}^{B}\sum_{j=1}^{B}\left[Y_{ij}\log P_{ij}^{v\to l} + Y_{ji}\log P_{ij}^{l\to v}\right].$$

$$\frac{\partial \mathcal{L}_{\mathrm{CaCL}}}{\partial s_{ij}} = \frac{1}{\tau}\left(P_{ij}^{v\to l} - Y_{ij} + P_{ij}^{l\to v} - Y_{ji}\right).$$

Stage II: BCAM + K-Gate inject language into space

Vision-dominant path

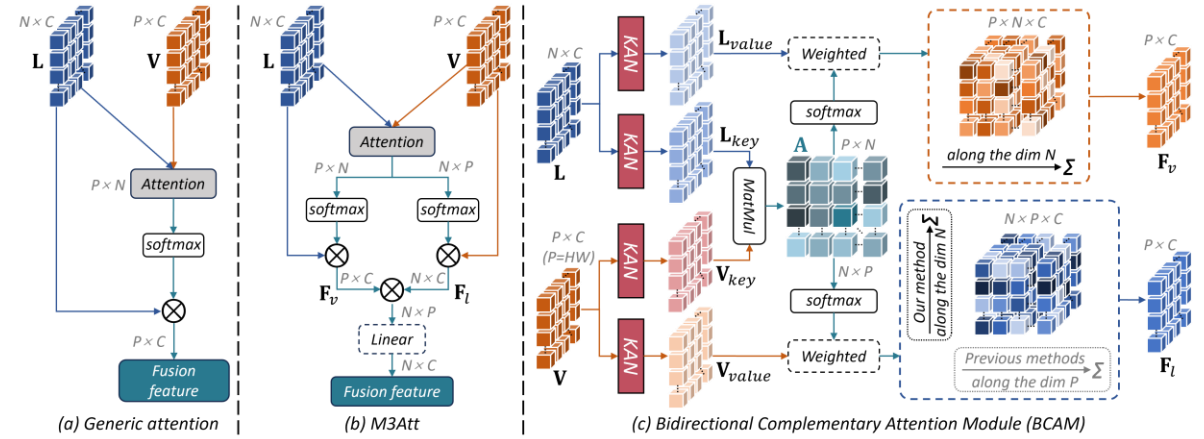
Each pixel attends to token semantics and stays on the original image grid.

Language-dominant path

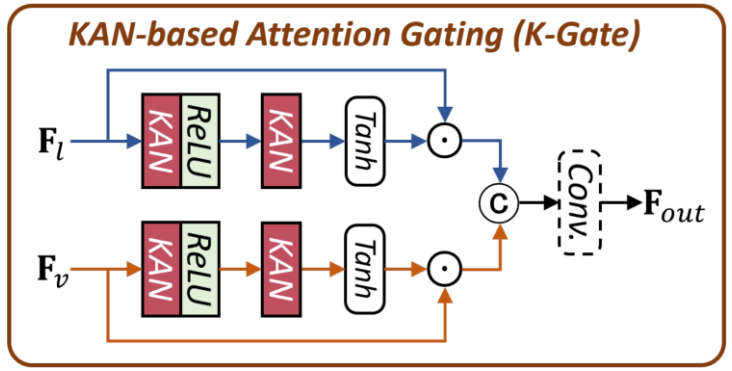
Each token projects influence back onto spatial locations, forming a language-guided feature map.

K-Gate

KAN-based nonlinear gating suppresses noisy modality-specific activations before fusion.



The illustration of different attention mechanisms from Figure 3.



Overview of the proposed K-Gate from Figure 2.

Experimental setting

QaTa-COV19

9,258 chest X-ray images

COVID-19 lesion masks + paired text

train/val/test
5716 /1429 / 2113

MosMedData+

2,729 chest CT slices

infection masks + laterality/location text

train/val/test
2183 /273 / 273

CVC-ClinicDB

612 colonoscopy images

polyp masks + clinical-style descriptions

train/test
550 / 62

Kvasir

1,000 colonoscopy images

polyp masks + clinical-style descriptions

train/test
900 / 100

Evaluation metrics

- Overlap: Dice and mIoU
- Boundary quality: HD95 and ASSD

Training defaults

- Pretrain batch 256; segmentation batch 32
- $\tau = 0.07$, $\rho = 0.8$; Adam + cosine annealing
- Language encoder frozen; BCE + Dice in Stage II

Main quantitative results

QaTa-COV19
85.25 Dice
+0.98 vs. best prior (84.27, MedLangViT)

MosMedData+
77.81 Dice
+1.86 vs. best prior (75.95, MedLangViT)

CVC-ClinicDB
91.82 Dice
+1.86 vs. best prior (89.96, MMIUNet)

Kvasir
91.92 Dice
+1.09 vs. best prior (90.83, LAVT)

Method	Params(M)	Text	QaTa-COV19				MosMedData+			
			Dice(%)↑	mIoU(%)↑	HD95↓	ASSD↓	Dice(%)↑	mIoU(%)↑	HD95↓	ASSD↓
U-Net (Ronneberger et al., 2015)	31.4	×	79.02	69.46	33.98	9.03	64.60	50.73	23.52	6.35
U-Net++ (Zhou et al., 2018)	74.5	×	79.62	70.25	36.14	9.91	71.75	58.39	24.06	5.45
nnUNet (Isensee et al., 2021)	19.1	×	80.42	70.81	28.14	9.86	72.59	60.36	22.75	5.56
Swin-Unet (Cao et al., 2023)	82.3	×	78.07	68.34	31.51	9.20	63.29	50.19	25.31	7.69
TransUNet (Chen et al., 2024)	105.0	×	78.63	69.13	29.88	8.42	71.24	58.44	23.41	6.38
UKAN (Li et al., 2025a)	**9.4**	×	79.30	69.85	31.89	8.79	72.56	59.05	29.38	7.25
MM-UKAN++ (Zhang et al., 2025a)	9.9	×	79.20	69.70	35.26	9.76	71.82	58.37	32.63	8.96
CLIP (Radford et al., 2021)	87.0	✓	79.81	70.66	23.25	5.54	71.97	59.64	26.24	6.58
GLoRIA (Huang et al., 2021)	45.6	✓	79.94	70.68	26.47	5.24	72.42	60.18	28.61	6.79
ViLT (Kim et al., 2021)	87.4	✓	79.63	70.12	25.32	5.96	72.36	60.15	24.85	5.69
TGANet (Tomar et al., 2022)	19.8	✓	77.17	64.39	29.54	7.83	69.48	55.81	26.39	6.12
ConVIRT (Zhang et al., 2022)	35.2	✓	79.72	70.58	22.36	6.03	72.06	59.73	22.38	6.36
LAVT (Yang et al., 2022)	118.6	✓	80.48	67.01	15.70	4.87	68.51	55.32	17.28	4.18
SLViT (Ouyang et al., 2023)	131.5	✓	79.25	68.87	15.18	4.35	72.57	60.78	21.23	6.10
LViT (Li et al., 2024)	29.7	✓	81.52	68.63	18.62	5.32	72.10	57.35	18.94	4.82
UniLSeg (Liu et al., 2024)	28.7	✓	72.88	59.58	15.15	4.11	65.89	52.01	19.98	4.96
RefSegformer (Wu et al., 2024)	195.0	✓	81.63	69.71	20.22	5.29	70.25	57.31	19.70	4.78
MedLangViT (Wang et al., 2025b)	27.7	✓	84.27	75.93	14.51	3.97	75.95	63.17	18.29	4.12
ARSeg (Wang et al., 2026)	30.1	✓	84.09	72.64	19.90	5.24	73.24	59.82	31.88	7.65
C2Seg (Ours)	18.92	✓	**85.25**	**76.97**	**12.71**	**3.38**	**77.81**	**65.17**	**15.02**	**3.76**

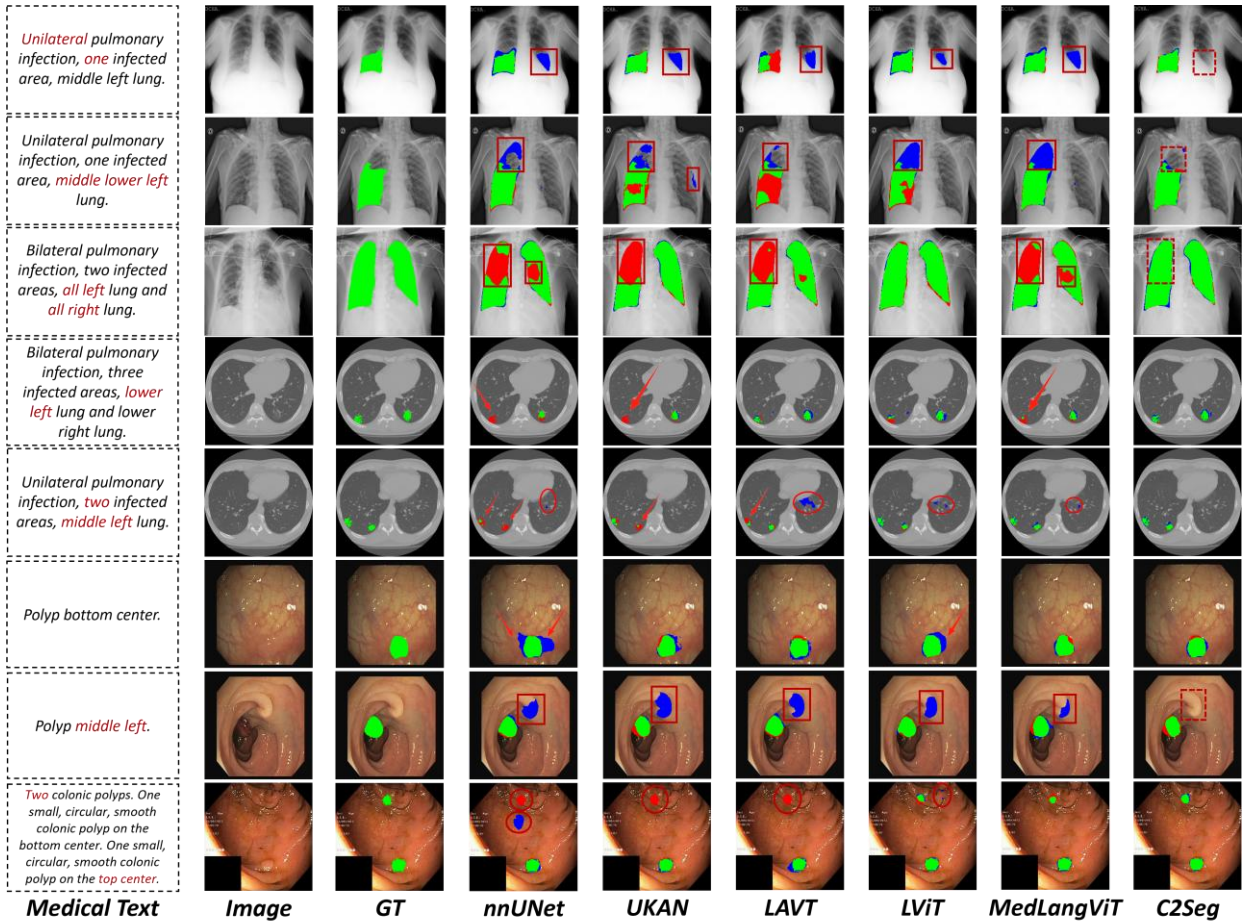
Method	Params(M)	Text	CVC-ClinicDB				Kvasir			
			Dice(%)↑	mIoU(%)↑	HD95↓	ASSD↓	Dice(%)↑	mIoU(%)↑	HD95↓	ASSD↓
U-Net (Ronneberger et al., 2015)	31.4	×	57.57	44.93	49.40	19.87	75.77	65.20	40.29	12.11
U-Net++ (Zhou et al., 2018)	74.5	×	88.94	82.91	12.16	3.99	87.00	79.71	20.87	6.63
nnUNet (Isensee et al., 2021)	105.0	×	85.69	77.72	13.48	5.84	86.95	79.19	20.39	5.81
Swin-Unet (Cao et al., 2023)	82.3	×	81.19	71.64	26.38	8.67	77.24	66.90	21.25	8.80
UKAN (Li et al., 2025a)	**9.4**	×	89.74	84.47	13.19	3.72	87.77	81.13	21.37	5.82
MM-UKAN++ (Zhang et al., 2025a)	9.9	×	89.52	82.15	13.26	3.36	85.63	78.03	24.81	7.12
LAVT (Yang et al., 2022)	118.6	✓	88.13	82.76	9.33	3.85	90.83	84.90	15.90	4.15
TGANet (Tomar et al., 2022)	19.8	✓	89.93	84.56	8.41	**2.11**	90.44	84.19	17.18	4.33
SLViT (Ouyang et al., 2023)	131.5	✓	80.55	72.86	26.91	9.87	85.69	77.97	19.57	5.56
LViT (Li et al., 2024)	29.7	✓	88.27	80.81	15.18	4.15	87.36	79.85	24.18	6.40
RefSegformer (Wu et al., 2024)	195.0	✓	80.73	71.94	23.10	7.28	86.87	78.77	25.30	6.46
MMIUNet (Bui et al., 2024)	56.2	✓	89.96	84.14	11.55	4.02	90.27	84.29	15.13	4.35
RecLMIS (Huang et al., 2025)	23.7	✓	81.31	73.04	38.44	12.13	90.63	84.35	17.93	4.43
MedLangViT (Wang et al., 2025b)	27.7	✓	88.35	81.92	10.66	4.50	90.57	84.21	14.12	4.09
ARSeg (Wang et al., 2026)	30.1	✓	89.74	82.64	13.71	4.29	88.45	81.34	22.79	5.66
C2Seg (Ours)	18.92	✓	**91.82**	**86.81**	**6.53**	2.23	**91.92**	**85.27**	**13.62**	**3.98**

Quantitative Comparison from Tables 1 and 2.

What stands out

- Best HD95 on all four datasets, indicating sharper localization and cleaner boundaries.
- 18.92M parameters - much lighter than many multimodal competitors while still improving every benchmark.
- Gains are consistent across both chest imaging and colonoscopy settings.

Qualitative evidence: better adherence to text



Visualization of different methods from Figure 4.

1. Correct lesion count

Matches prompts such as “two infected areas” rather than merging or hallucinating lesions.

2. Better location cues

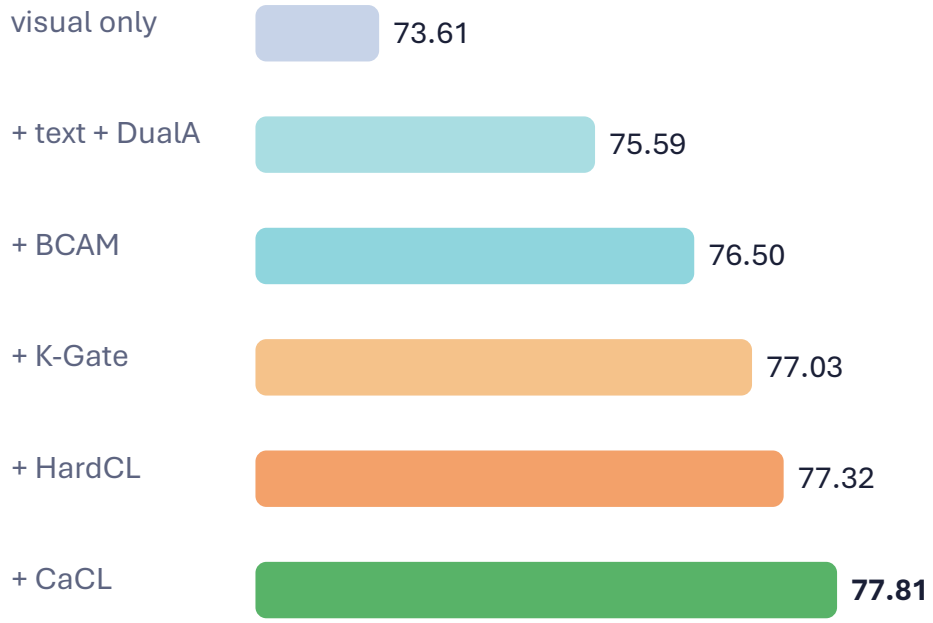
Handles laterality and coarse position cues such as “middle lower left” or “bilateral”.

3. Cleaner masks

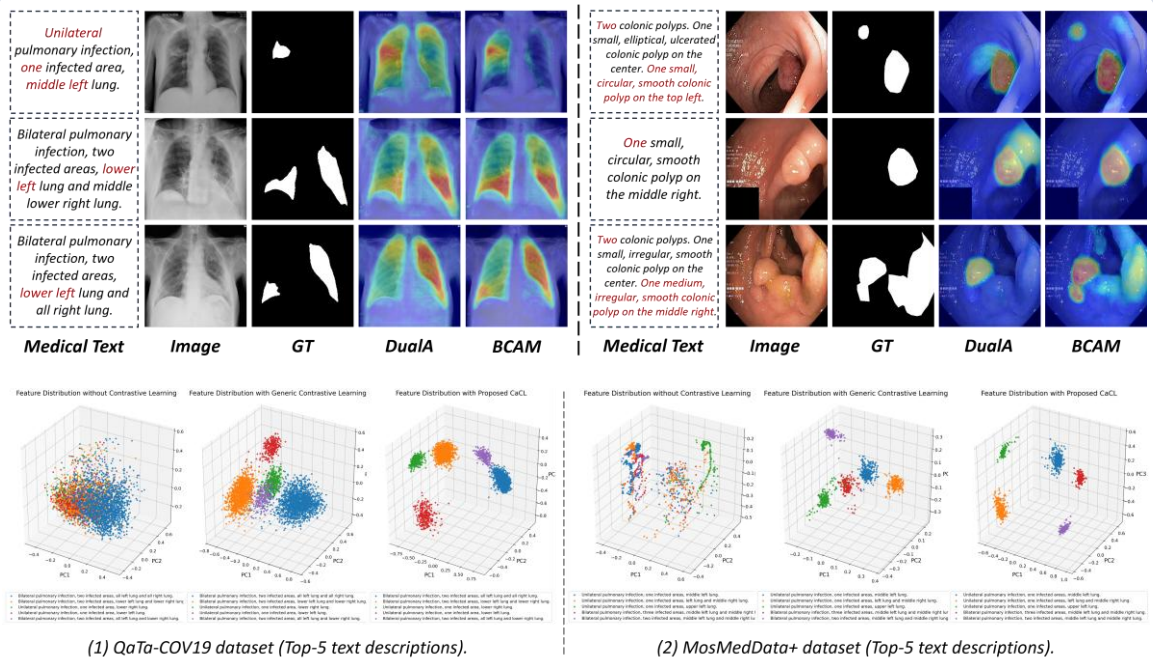
Lower HD95 / ASSD and visibly fewer false positives indicate sharper boundary quality.

Why the gains happen

Stepwise improvement (Table 3)



CLIP remains the best tested text encoder because it is already aligned to visual semantics.



(1) QaTa-COV19 dataset (Top-5 text descriptions). *(2) MosMedData+ dataset (Top-5 text descriptions).*

Visualizations of ablation experiments from Figures 5 and 6.

Sensitivity (Table 4)

- The model is only weakly sensitive to ρ and τ; the paper uses ρ = 0.8 and τ = 0.07.
- Bigger contrastive batches help, but 256 is almost as good as 512 and is much cheaper.

Limitations and future directions

Current limitations

- No dedicated metric for text-mask consistency yet.
- Robustness to noisy text (misspellings, abbreviations, missing findings) is not tested.

Promising next steps

- Build a metric for count, position, and laterality constraints.
- Stress-test C2Seg with noisy prompts and richer clinical wording.

Bottom line

C2Seg addresses both halves of the consistency problem:

1. Better alignment in pretraining via soft neighborhood-aware contrastive targets

2. Better fusion in segmentation via explicit bidirectional spatial interaction

3. Better feature selection via KAN-based nonlinear gating

Together these choices improve both segmentation accuracy and semantic consistency across four public datasets.

Thank you

C2Seg in one sentence:

use soft neighborhood-aware alignment and bidirectional spatial fusion so the predicted mask actually follows the clinical text.

- Stronger text-mask consistency
- Higher Dice / mIoU and lower HD95 / ASSD
- Lightweight 18.92M-parameter model