NEGATIVE SEMANTICS DRIVE EXPERT SPECIALIZATION IN SEMI-SUPERVISED MEDICAL IMAGE SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent semi-supervised medical image segmentation methods have increasingly explored clinical text to alleviate the scarcity of pixel-level annotations. However, existing methods primarily use positive semantics to describe desired predictions, while overlooking information about which visual patterns should be rejected. This paper shows that structured negative semantics can provide explicit signals for learning complementary visual experts. Building on this insight, RCEA-MoE structures negative descriptions into disease, location, and extent factors to specialize heterogeneous visual experts. Positive–negative semantic contrasts induce factor-specific spatial corrections, which are adaptively aggregated by an image–positive-text router, while an exponential moving average (EMA) teacher provides confidence-aware supervision for unlabeled images. In a controlled ablation on QaTa-COV19 with 25% labeled training data, RCEA-MoE achieves 90.84% Dice and 83.22% mIoU, improving over the single-expert visual baseline by 3.87 and 6.27 percentage points, respectively. These findings highlight the potential of combining structured negative guidance, heterogeneous experts, and reliability-aware learning for label-efficient medical image segmentation. Code is available at: https://anonymous.4open.science/r/RCEA-MoE-91D1.

1 INTRODUCTION

Medical image segmentation supports quantitative disease assessment and treatment planning, but obtaining pixel-level annotations requires substantial clinical effort. Despite advances in segmentation networks(Ronneberger et al., 2015), limited annotations remain a practical constraint. Semi-supervised medical image segmentation addresses this problem by combining labeled and unlabeled images, with reliable supervision of the unlabeled images being a central challenge.

Existing approaches improve the reliability of supervision from unlabeled images through teacher–student consistency and uncertainty estimation(Tarvainen & Valpola, 2017; Yu et al., 2019), geometric constraints(Li et al., 2020; Luo et al., 2021), collaborative prediction (Wang et al., 2023; Shen et al., 2023), and image mixing or weak-to-strong perturbations(Bai et al., 2023; Yang et al., 2023). While these methods strengthen the consistency and robustness of visual predictions, they remain largely dependent on the information encoded in the images themselves. In particular, prediction agreement does not necessarily resolve ambiguities in disease appearance, anatomical location, or lesion extent, allowing multiple visual branches to consistently reinforce the same erroneous segmentation.

Clinical descriptions can provide complementary semantic cues that are not readily available from visual predictions alone. LViT and cross-position attention integrate medical descriptions with visual features(Li et al., 2024c; Lee et al., 2023), while TextDiff introduces text guidance into diffusion-based representations (Feng, 2024). TextMatch, VCLIPSeg, and DuSSS further explore textual guidance for semi-supervised medical segmentation (Li et al., 2024a;b; Pan et al., 2025). Recent work also investigates spatial and frequency-domain text conditioning in DD-CMD (Rahman & Hammond, 2026), and text-assisted expert decoding for joint CT–text and X-ray–text learning in TextMoE(Zeng et al., 2025a).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

Despite these advances, existing text-guided methods primarily use positive descriptions to specify the semantics that a segmentation should capture, while providing limited guidance on visual patterns that should be rejected. We therefore investigate structured negative descriptions as explicit signals for visual expert specialization. Disease, location, and extent contradictions correspond to different types of visual evidence, including local appearance, anatomical context, and lesion boundaries. Associating these negative factors with dedicated visual experts enables complementary semantic specialization beyond architectural diversity alone.

RCEA-MoE couples factor-guided heterogeneous experts with Reliability-Gated Contrastive Expert Attention (RCEA). As illustrated in Figure 1, each expert receives the positive report together with its assigned negative factor, and RCEA contrasts the two to generate spatially gated feature corrections. The corrected experts are then aggregated through an image–positive-text router, while expert-separation and cross-negative objectives encourage factor-specific specialization. An exponential moving average teacher further provides confidence-aware supervision for unlabeled images. Together, these components integrate structured semantic guidance with semi-supervised learning.

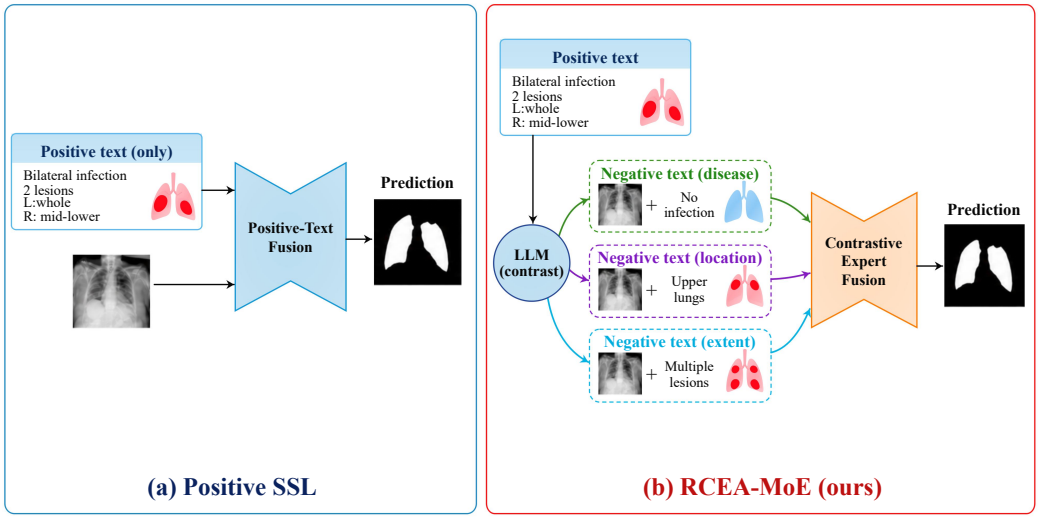


Figure 1: Comparison of semi-supervised medical image segmentation paradigms. Conventional SSL relies only on visual consistency, while positive-text SSL introduces affirmative semantic guidance. RCEA-MoE uses structured negative semantics to drive expert specialization: disease, location, and extent negatives guide the local, contextual, and boundary experts, respectively. Their complementary representations are adaptively combined by an image–positive-text router.

This paper evaluates RCEA-MoE on QaTa-COV19 and MosMedData+ under 25% and 50% labeled-data settings. Controlled ablations further disentangle the effects of the combined text-guidance design and the joint specialization objectives.

The main contributions are summarized as follows:

- **Factor-Guided Expert Specialization:** We align disease, location, and extent negatives with local, contextual, and boundary experts, enabling factor-specific and complementary visual processing.
- **Positive–Negative Spatial Correction:** RCEA converts positive–negative contextual contrasts into spatially gated corrections, selectively refining expert features according to their corresponding semantic factors.
- **Systematic Experimental Validation:** Experiments across two datasets and annotation budgets, together with controlled ablations, systematically assess the effectiveness of text guidance and the joint expert specialization objectives.

2 METHOD

2.1 PROBLEM FORMULATION AND FRAMEWORK OVERVIEW

Let $\mathcal{D}_l = \{(\mathbf{x}_i, \mathbf{t}_i^+, \mathbf{t}_{i,d}^-, \mathbf{t}_{i,s}^-, \mathbf{t}_{i,e}^-, \mathbf{y}_i)\}_{i=1}^{N_l}$ be the labeled set and $\mathcal{D}_u = \{(\mathbf{x}_i, \mathbf{t}_i^+, \mathbf{t}_{i,d}^-, \mathbf{t}_{i,s}^-, \mathbf{t}_{i,e}^-)\}_{i=1}^{N_u}$ the unlabeled set. Here, $\mathbf{x}$ is a medical image, $\mathbf{y} \in \{0,1\}^{H\times W}$ is its segmentation mask, and $\mathbf{t}^+$ is a positive clinical description. The three negative descriptions target contradictions in disease identity (d), anatomical location (s), and lesion extent (e), respectively. Our goal is to learn a segmentation model that uses these structured contradictions to reject plausible clinical alternatives while exploiting unlabeled images through reliable teacher supervision.

Figure 2 provides an overview of the proposed RCEA-MoE framework.

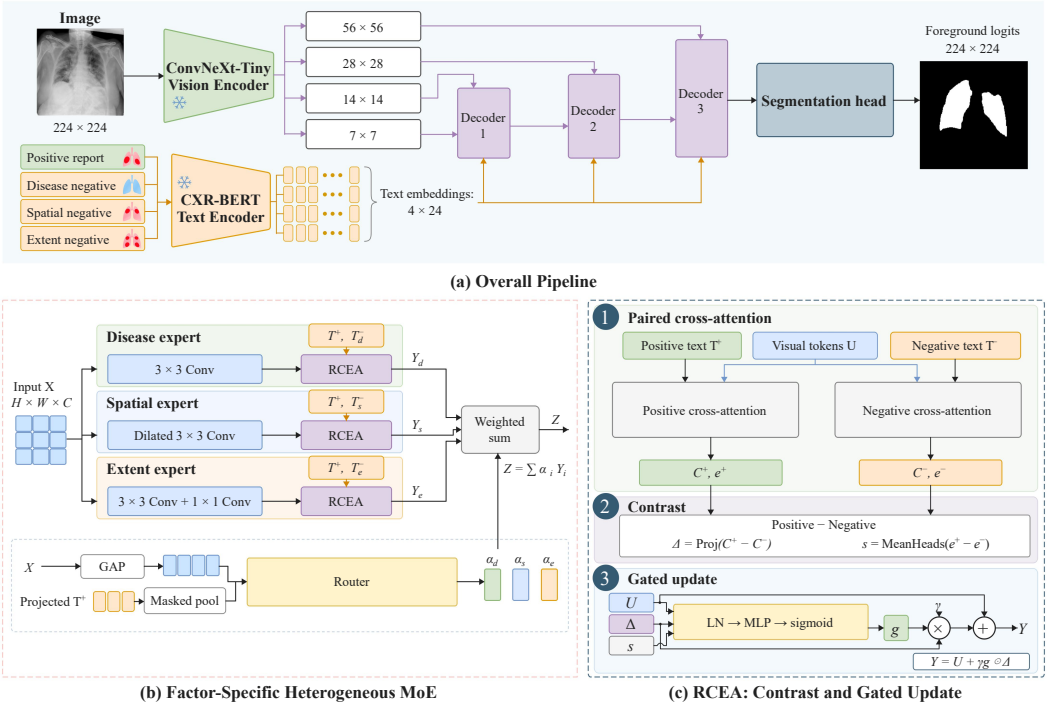


Figure 2: Overview of RCEA-MoE. (a) Image and text encoding with multi-scale segmentation decoding. (b) Factor-guided heterogeneous experts with image–positive-text routing. (c) RCEA generates spatially gated feature corrections through positive–negative contrast. PE and GAP denote positional encoding and global average pooling.

2.2 MULTI-SCALE IMAGE AND TEXT ENCODING

Given a 224×224 image, ConvNeXt-Tiny extracts four feature maps $\{\mathbf{E}_1, \mathbf{E}_2, \mathbf{E}_3, \mathbf{E}_4\}$ with spatial sizes $[56, 28, 14, 7]$ along each axis. A three-stage decoder progressively upsamples the features and concatenates the corresponding encoder features through skip connections. Each stage contains two RCEA-MoE blocks. Finally, sub-pixel upsampling and a 1×1 prediction head produce the full-resolution segmentation output.

The positive report and three factor-specific negatives are tokenized to a maximum length of $L = 24$ and encoded by the same frozen CXR-BERT. The final hidden states provide four token sequences: $\mathbf{T}^+$, $\mathbf{T}_d^-$, $\mathbf{T}_s^-$, and $\mathbf{T}_e^-$. These sequences are supplied to every decoder stage. Within each RCEA-MoE block, a shared block-specific projection maps the positive and negative tokens to the visual feature dimension for subsequent contrastive attention.

2.3 FACTOR-SPECIFIC HETEROGENEOUS EXPERTS

At each decoder site, three parallel experts process the input feature $\mathbf{X}$ to capture local appearance, anatomical context, and lesion boundaries:

$$\mathbf{Z}_d = \mathcal{C}_d(\mathbf{X}), \quad \mathbf{Z}_s = \mathcal{D}_s(\mathbf{X}), \quad \mathbf{Z}_e = \mathcal{C}_e(\mathbf{X}) + \beta\mathcal{P}_e(\mathcal{S}(\mathbf{X})). \tag{1}$$

Disease expert: $\mathcal{C}_d$ is a standard 3×3 convolution that captures local opacity and texture patterns. This expert is paired with the disease negative. **Spatial expert:** $\mathcal{D}_s$ consists of a 3×3 convolution with dilation two, followed by GELU and a 1×1 projection. This expert captures broader anatomical context and is paired with the location negative. **Extent expert:** $\mathcal{C}_e$ is a separate 3×3 convolution, while $\mathcal{P}_e$ projects channel-wise Sobel edge magnitudes $\mathcal{S}(\mathbf{X})$ through a 1×1 convolution. The learnable scale β controls the boundary contribution, and this expert is paired with the extent negative.

Each output $\mathbf{Z}_k$, $k \in \{d, s, e\}$, is flattened into visual tokens and augmented with sinusoidal positional encoding to obtain $\mathbf{V}_k$. RCEA then refines each sequence using the positive report $\mathbf{T}^+$ and its assigned negative $\mathbf{T}_k^-$.

2.4 RELIABILITY-GATED CONTRASTIVE EXPERT ATTENTION

Clinical reports provide image-level semantic information, while segmentation requires decisions at individual spatial locations. RCEA constructs positive–negative feature corrections and controls their strength at individual spatial locations. For expert k, the positive report describes the target semantics, while the assigned negative introduces a contradiction in its corresponding factor. Figure 2c illustrates the following four steps.

1. Dual masked cross-attention. Each expert queries the positive report and its assigned negative using the same visual tokens $\mathbf{V}_k$:

$$\mathbf{C}_k^{\pm} = \mathrm{Attn}_k\left(\mathbf{V}_k, \mathbf{T}_k^{\pm}; \mathbf{M}_k^{\pm}\right). \tag{2}$$

Here, $\mathbf{T}_k^+ = \mathbf{T}^+$ is the shared positive report, and $\mathbf{T}_k^-$ is the factor-specific negative. The attention operator uses visual queries and textual keys and values. Within each expert, the two branches share visual queries, text normalization, and key/value projections, allowing the reports to be compared under the same visual context. The masks $\mathbf{M}_k^{\pm}$ exclude padded tokens. Each row of $\mathbf{C}_k^+$ and $\mathbf{C}_k^-$ therefore summarizes the text attended to at the same visual location under the two descriptions.

2. Feature contrast and attention evidence. We contrast the two contextual responses and map their difference back to the visual feature space:

$$\mathbf{\Delta}_k = \mathrm{Proj}_o\left(\mathbf{C}_k^+ - \mathbf{C}_k^-\right), \tag{3}$$

where Proj_o is a learnable affine projection. This operation makes the proposed correction sensitive to the semantic distinction between the positive report and the assigned negative. The resulting $\mathbf{\Delta}_k$ is a feature vector at each location, specifying the content of the visual update.

We additionally retain compatibility evidence from the pre-softmax attention logits. Unlike the contextual responses, which aggregate textual values, these logits measure query–text compatibility before attention normalization. For each location and attention head, log-mean-exp pooling over valid text tokens summarizes this evidence with sequence-length normalization. We subtract the negative evidence from the positive evidence and average across heads to obtain $s_{k,j}$. A positive score indicates stronger query–text compatibility with the positive report than with the assigned negative at location j. The evidence score complements the visual feature and proposed correction as an input to the spatial gate.

3. Spatial reliability gate. The usefulness of a text-derived correction can vary across the image. We therefore predict a separate gate for each visual token by jointly considering its original feature $\mathbf{v}_{k,j}$, proposed correction $\mathbf{\Delta}_{k,j}$, and evidence score $s_{k,j}$:

$$g_{k,j} = \sigma(\mathrm{MLP}([\mathbf{v}_{k,j}; \mathbf{\Delta}_{k,j}; s_{k,j}])), \qquad g_{k,j} \in [0, 1]. \tag{4}$$

Here, $[\cdot;\cdot]$ denotes concatenation, σ is sigmoid, and the MLP includes input layer normalization. Small gate values suppress the local correction, whereas larger values allow it to contribute more strongly. During semi-supervised training, the router-aggregated gate map at each decoder site is additionally supervised by EMA teacher certainty, as detailed in Section 2.6.

4. Gated residual update. The refined expert output is obtained by adding the spatially weighted correction to the original features:

$$\mathbf{Y}_k = \mathbf{V}_k + \gamma_k \mathbf{g}_k \odot \mathbf{\Delta}_k. \tag{5}$$

Here, $\mathbf{g}_k$ collects the scalar gates $g_{k,j}$ over all visual locations, and each gate is broadcast across feature channels. The learnable scalar γ_k controls the overall correction magnitude for the current expert at this decoder site, while $\mathbf{g}_k$ controls its spatial variation. We initialize γ_k to 0.1 to moderate the initial correction strength. The residual connection retains the original visual features as a direct path, with the semantic correction added according to the learned spatial gates.

2.5 Image–Positive-Text Router

At each decoder site, the router uses the globally pooled visual feature $\mathbf{p}$ and the masked mean of projected positive-text tokens $\mathbf{t}$. Their cosine similarity after a learned visual projection is denoted by ρ. The router computes sample-wise weights and combines the refined expert outputs $\mathbf{Y}_k$:

$$\boldsymbol{\alpha} = \mathrm{softmax}(\mathrm{MLP}([\mathbf{p};\mathbf{t};\rho])), \qquad \mathbf{Z} = \sum_{k\in\{d,s,e\}} \alpha_k \mathbf{Y}_k. \tag{6}$$

All three experts are densely aggregated. Negative reports guide expert refinement through RCEA but are not explicit routing inputs. The router assigns one weight to each expert for the entire image, whereas the RCEA gate controls the correction at each spatial location. An expert can therefore receive a large routing weight while suppressing its text-derived correction in specific regions. A mean-squared load-balancing loss $\mathcal{L}_{\mathrm{bal}}$, averaged over decoder sites, encourages the batch-average routing weights to remain close to a uniform distribution.

2.6 Reliable Semi-Supervised Optimization

Labeled supervision. The labeled objective $\mathcal{L}_{\mathrm{sup}}$ combines Dice and binary cross-entropy losses with boundary supervision, router balancing, and RCEA evidence supervision. Boundary supervision minimizes the mean-squared difference between the Sobel magnitudes of the channel-averaged prediction and ground truth. Router balancing encourages uniform batch-average expert usage. RCEA evidence supervision applies binary cross-entropy to the router-aggregated evidence logits, upsampled to the ground-truth mask resolution.

Expert specialization. Let $\mathbf{f}_k$ denote refined expert features pooled within the annotated foreground, and $\mathbf{u}_k$ the globally pooled features before RCEA. The text vectors $\mathbf{t}^+$ and $\mathbf{t}_k^-$ are masked averages of projected, position-encoded tokens. The separation loss favors the positive report over the assigned negative, while the cross-negative loss discourages similarity to other experts' negatives:

$$\begin{aligned} \mathcal{L}_{\mathrm{sep}} &= \mathbb{E}\Big[\big(m + \cos(\mathbf{f}_k, \mathbf{t}_k^-) - \cos(\mathbf{f}_k, \mathbf{t}^+)\big)_+\Big], \\ \mathcal{L}_{\mathrm{xneg}} &= \frac{1}{2}\,\mathbb{E}\Big[\sum_{\ell\neq k} \cos\big(\mathbf{u}_k, \mathrm{sg}(\mathbf{t}_\ell^-)\big)\Big]. \end{aligned} \tag{7}$$

Here, m is the separation margin, $(a)_+ = \max(a, 0)$, and sg denotes stop-gradient. The expectation averages over valid samples and experts within each site, then over applicable sites. Both losses use labeled images, and separation excludes samples without foreground. Their scaled combination defines $\mathcal{L}_{\mathrm{spec}}$.

Unlabeled segmentation consistency. An EMA teacher predicts two geometrically aligned views of each unlabeled image. Adaptive confidence thresholds identify reliable pixels $\mathcal{R}$ where both

views confidently agree. The remaining pixels with class agreement form $\mathcal{U}$. The hard and soft consistency losses are

$$\mathcal{L}^u_{\text{hard}} = \mathbb{E}_{j\in\mathcal{R}}\big[\ell_{\text{BCE}}(z^S_j, \hat{y}_j)\big],$$
$$\mathcal{L}^u_{\text{soft}} = \mathbb{E}_{j\in\mathcal{U}}\big[(p^S_j - \bar{p}^T_j)^2\big]. \tag{8}$$

Here, ℓ_{BCE} denotes binary cross-entropy with logits, z^S_j is the student logit, $p^S_j = \sigma(z^S_j)$, $\hat{y}_j$ is the hard pseudo-label, and $\bar{p}^T_j$ averages the two teacher predictions. The expectations denote averaging over the selected pixels. Cross-view disagreements are excluded from both terms.

Unlabeled auxiliary supervision. The auxiliary objective is

$$\mathcal{L}^u_{\text{aux}} = \mathcal{L}^u_{\text{bd}} + \mathcal{L}^u_{\text{bal}} + \mathcal{L}^u_{\text{rcea,hard}} + \mathcal{L}^u_{\text{rcea,soft}} + \mathcal{L}^u_{\text{gate}}, \tag{9}$$

where each term incorporates its training coefficient. Boundary supervision uses certainty-weighted Sobel matching against the confident-foreground pseudo-mask, while router balancing encourages uniform batch-average expert usage. The hard and soft evidence terms use BCE with pseudo-labels and MSE with teacher probabilities, weighted by reliability and its complement, respectively. The soft evidence term also includes cross-view disagreements.

Gate supervision matches the router-aggregated gate map to teacher certainty using MSE. Its target is $q_j = |2\tilde{p}^T_j - 1|^2$, where $\tilde{p}^T_j$ averages teacher probabilities over views and channels. Evidence and gate maps are upsampled to the target resolution before supervision.

Overall objective. The complete training objective is

$$\mathcal{L} = \mathcal{L}_{\text{sup}} + \mathcal{L}_{\text{spec}} + \mathcal{L}_u. \tag{10}$$

Here, $\mathcal{L}_u$ combines hard and soft segmentation consistency with auxiliary supervision. The training coefficients are absorbed into the grouped objectives, and $\mathcal{L}_u$ additionally includes Gaussian ramp-up during the initial semi-supervised epochs.

3 Experiments

3.1 Experimental Setup

Datasets and annotation budgets. We evaluate RCEA-MoE on two pulmonary lesion segmentation datasets: QaTa-COV19 for chest X-ray images (Degerli et al., 2022) and MosMedData+ for axial chest CT slices (Morozov et al., 2020; Li et al., 2024c). The construction of the image–text benchmarks is described by Li et al. (2024c). QaTa-COV19 contains 5,716 training images, 1,429 validation images, and 2,113 test images. MosMedData+ contains 2,183 training slices, 273 validation slices, and 273 test slices.

We consider annotation budgets of 25% and 50%, measured relative to the training pool, excluding validation and test images. For QaTa-COV19, these settings use 1,429 and 2,858 labeled images, respectively; for MosMedData+, they use 545 and 1,091 labeled slices. The remaining training images are used without segmentation masks. Within each dataset, the two settings share the same training, validation, and test partitions, and the 25% labeled subset is contained in the 50% subset. Our model uses a positive description and three factor-specific negative descriptions for each image, corresponding to disease, location, and extent. The positive descriptions are those supplied with the experimental data, and the negatives are prepared offline with language-model assistance. All four descriptions are supplied during training and inference. The annotation budgets count segmentation masks; text provides additional supervision.

Evaluation protocol. We evaluate segmentation performance using Dice and mIoU, reported as percentages. Literature-reported results follow the evaluation protocols of their respective sources. The average columns in Table 1 summarize performance across the two annotation budgets. For our model, predictions are binarized at a fixed threshold of 0.5. The final model uses the EMA weights from the semi-supervised checkpoint with the highest validation Dice. The selected checkpoint and prediction threshold are fixed before the reported test evaluation.

Implementation details. Images and masks are resized to 224×224. We use ConvNeXt-Tiny as the visual encoder and a frozen specialized CXR-BERT as the text encoder. The text encoder remains in evaluation mode during training, and the maximum text length is 24 tokens. The decoder contains three heterogeneous experts at each of six expert-mixture blocks.

Training consists of labeled pretraining followed by semi-supervised optimization. Each stage uses AdamW with an initial learning rate of 3×10^{-4}, weight decay of 0.01, and cosine annealing to 10^{-6}. Each stage runs for at least 40 and at most 200 epochs, with an early-stopping patience of 20 epochs. The batch size is 32. During semi-supervised training, a full batch contains 8 labeled and 24 unlabeled images at the 25% annotation budget, or 16 labeled and 16 unlabeled images at the 50% budget.

The EMA decay is capped at 0.999, and the unlabeled objective is gradually introduced over the first 10 semi-supervised epochs. Geometrically aligned teacher views are constructed using photometric perturbations. Student perturbations include contrast adjustment, cutout, and positive-text token masking with probability 0.15. RCEA-MoE uses the training objective described in Section 2.6 throughout the experiments.

3.2 Quantitative Comparison

Table 1 reports segmentation results using QaTa-COV19 and MosMedData+ with 25% and 50% labeled data. The image-only methods are UCMT (Shen et al., 2023), BCP (Bai et al., 2023), and LeFeD (Zeng et al., 2025b). The text-guided methods are LAVT (Yang et al., 2022), LViT (Li et al., 2024c), TeViD (Zeng et al., 2026), and DuSSS (Pan et al., 2025). The comparison includes local evaluations and literature-reported scores.

Table 1: Quantitative comparison on QaTa-COV19 and MosMedData+ with 25% and 50% labeled data. Scores are reported in percentages. I and T denote image and text inputs, respectively. Avg. averages the two annotation budgets. The best and second-best listed scores are highlighted in bold and underlined, respectively.

		25% labeled		50% labeled		Avg.	
Method	Modality	Dice	mIoU	Dice	mIoU	Dice	mIoU
QaTa-COV19							
UCMT	I	76.09	64.13	77.81	68.65	76.95	66.39
BCP	I	74.79	65.26	75.57	65.31	75.18	65.29
LeFeD	I	78.15	67.12	78.19	69.21	78.17	68.16
LAVT	I+T	77.08	67.21	79.10	70.88	78.09	69.05
LViT	I+T	78.12	66.75	80.32	72.16	79.22	69.46
DuSSS	I+T	79.00	68.21	82.52	75.87	80.76	72.04
TeViD	I+T	<u>88.23</u>	<u>78.94</u>	<u>90.49</u>	<u>82.63</u>	<u>89.36</u>	<u>80.79</u>
Ours	I+T	**90.84**	**83.22**	**91.35**	**84.08**	**91.10**	**83.65**
MosMedData+							
UCMT	I	68.73	52.86	70.80	55.09	69.76	53.97
BCP	I	69.56	53.42	71.38	55.76	70.47	54.59
LeFeD	I	70.72	54.50	72.06	56.21	71.39	55.36
LAVT	I+T	71.02	54.73	72.11	56.29	71.57	55.51
LViT	I+T	71.38	54.82	72.25	56.33	71.82	55.58
DuSSS	I+T	72.39	55.60	73.18	57.32	72.79	56.46
TeViD	I+T	<u>74.66</u>	<u>59.56</u>	<u>75.91</u>	<u>61.17</u>	<u>75.29</u>	<u>60.37</u>
Ours	I+T	**79.76**	**66.33**	**80.55**	**67.44**	**80.16**	**66.89**

RCEA-MoE achieves 90.84% Dice and 83.22% mIoU for QaTa-COV19 with 25% labeled data, exceeding TeViD by 2.61 and 4.28 percentage points, respectively. With 50% labeled data, the scores increase to 91.35% Dice and 84.08% mIoU, exceeding TeViD by 0.86 and 1.45 percentage points. Averaged across the two annotation budgets, RCEA-MoE achieves 91.10% Dice and 83.65% mIoU.

Evaluation using MosMedData+ yields 79.76% Dice and 66.33% mIoU with 25% labeled data, improving over TeViD by 5.10 and 6.77 percentage points, respectively. With 50% labeled data, RCEA-MoE achieves 80.55% Dice and 67.44% mIoU, exceeding TeViD by 4.64 and 6.27 percentage points. The corresponding averages across annotation budgets are 80.16% Dice and 66.89% mIoU.

3.3 Ablation Study

We conduct a progressive ablation study on QaTa-COV19 with annotations available for 25% of the training pool. Starting from a single-expert visual baseline, we introduce heterogeneous experts (HE), factorized negative text with RCEA (FN and RCEA), and the separation and cross-negative objectives (Sep. and XNeg.). Configurations with FN and RCEA also enable image–positive-text routing and RCEA auxiliary supervision. All configurations are trained under the same protocol. Table 2 summarizes the results.

Table 2: Progressive ablation on QaTa-COV19 with 25% labeled data. HE: heterogeneous experts; FN: factorized negative text; Sep.: expert separation; XNeg.: cross-negative constraint. The visual baseline uses a single expert. Scores are reported in percentages, with the best results highlighted in bold.

Configuration	HE	FN	RCEA	Sep.	XNeg.	Dice	mIoU
Visual baseline	×	×	×	×	×	86.97	76.95
HE	✓	×	×	×	×	87.09	77.14
HE & FN & RCEA	✓	✓	✓	×	×	90.36	82.42
Full model	✓	✓	✓	✓	✓	**90.84**	**83.22**

Heterogeneous expert mixture. Replacing the single visual expert with HE increases Dice from 86.97% to 87.09% and mIoU from 76.95% to 77.14%. The local, contextual, and boundary branches provide different visual inductive biases, but their introduction alone yields only modest improvements in this setting.

Factorized negative text and RCEA. The HE & FN & RCEA configuration raises Dice to 90.36% and mIoU to 82.42%, improving over HE by 3.27 and 5.28 percentage points, respectively. This is the largest improvement in the ablation series. The result is consistent with our design rationale: positive–negative contrasts provide targeted feature corrections, while image–positive-text routing adapts the contributions of the refined experts to each input. Because factor-specific text, RCEA, adaptive routing, and auxiliary supervision are enabled together, this comparison evaluates their combined contribution.

Expert specialization constraints. Adding Sep. and XNeg. jointly further improves Dice by 0.48 points and mIoU by 0.80 points. The separation objective increases the relative cosine similarity of lesion features to the positive report over the assigned negative, while the cross-negative objective reduces similarity to negative factors assigned to other experts. These results support their combined contribution beyond the feature correction and routing used in the preceding configuration. The full model achieves 90.84% Dice and 83.22% mIoU, exceeding the visual baseline by 3.87 and 6.27 percentage points, respectively.

3.4 Qualitative Comparison

Figure 3 compares predicted masks with the reference annotations. The examples include QaTa-COV19 images at both annotation budgets and two MosMedData+ slices at the 50% budget. Green denotes correctly segmented foreground, red denotes false positives, and blue denotes false negatives.

In the first QaTa-COV19 example, several competing methods miss a substantial portion of the annotated upper lesion region. RCEA-MoE recovers more of this region, although localized false

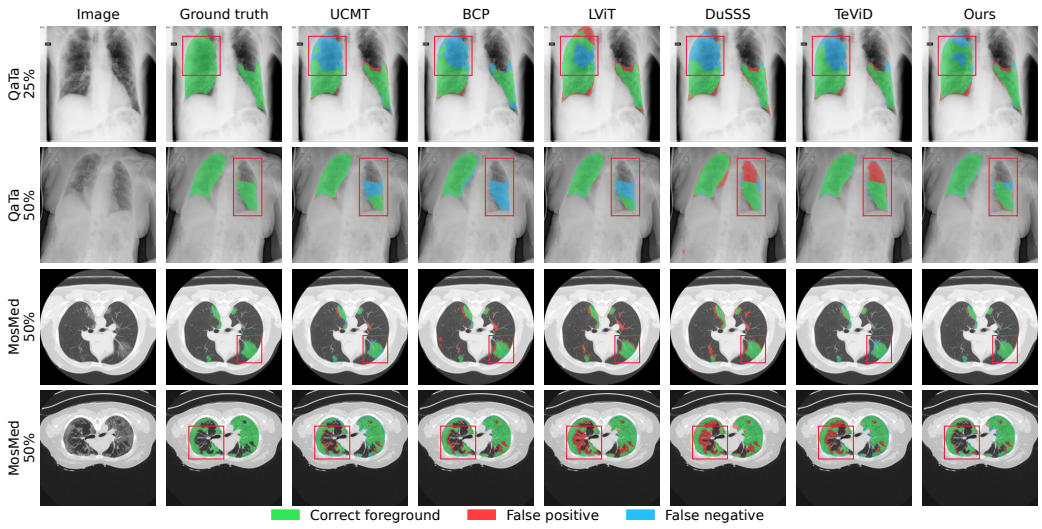


Figure 3: Qualitative comparison on QaTa-COV19 and MosMedData+. Columns show the input image, ground truth, and predictions from UCMT, BCP, LViT, DuSSS, TeViD, and RCEA-MoE. The first two rows correspond to QaTa-COV19 with 25% and 50% labeled data; the last two rows show MosMedData+ with 50% labeled data. Green indicates correctly segmented foreground, red indicates false positives, and blue indicates false negatives.

negatives remain. In the second example, it captures most of the two annotated regions while retaining errors near the boundary of the smaller region. The MosMedData+ examples exhibit different segmentation errors. In the slice with several separated lesions, RCEA-MoE delineates multiple annotated components with limited visible false-positive extensions. In the slice with extensive involvement, the prediction captures much of the annotated lesion extent, although false positives and false negatives remain along irregular internal boundaries. These selected examples illustrate segmentation behavior but do not isolate the contributions of individual model components.

4 Conclusion

In this paper, we address the challenge of exploiting rich clinical semantics for semi-supervised medical image segmentation, where limited pixel-level annotations make effective guidance from complementary visual cues particularly important. To this end, we present RCEA-MoE, a text-guided framework that combines factor-guided heterogeneous experts, positive–negative spatial correction, and reliability-aware learning. Experiments on QaTa-COV19 and MosMedData+ demonstrate its effectiveness under limited annotation budgets, while controlled ablations validate the complementary text-guidance design and joint expert specialization objectives. Future work will explore broader semantic factors and larger-scale multimodal benchmarks to further examine the potential of negative semantics for semi-supervised medical image segmentation.

AI use statement

Generative AI tools were used for language editing, figure preparation assistance, code development assistance, technical feedback, and generating initial negative clinical descriptions from provided positive descriptions. The authors verified the manuscript, implementations, experimental procedures, and reported results, and take responsibility for the final content and interpretation.

REPRODUCIBILITY STATEMENT

Section 2 describes the architecture and training objectives. Section 3.1 specifies the data partitions, annotation budgets, training settings, and model-selection rule. Section 3.3 defines the controlled ablation variants. Local experiment records retain fixed data manifests, configurations, source snapshots, checkpoints, and evaluation outputs. The evaluation metrics used in our experiments are described in Section 3.1. Results taken from prior studies follow the metric conventions reported in those studies.

REFERENCES

Yunhao Bai, Duowen Chen, Qingli Li, Wei Shen, and Yan Wang. Bidirectional copy-paste for semi-supervised medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11514–11524, 2023.

Aysen Degerli, Serkan Kiranyaz, Muhammad E. H. Chowdhury, and Moncef Gabbouj. OSegNet: Operational segmentation network for COVID-19 detection using chest x-ray images. In *IEEE International Conference on Image Processing*, pp. 2306–2310, 2022. doi: 10.1109/ICIP46576.2022.9897412. URL https://arxiv.org/abs/2202.10185.

Chun-Mei Feng. Enhancing label-efficient medical image segmentation with text-guided diffusion models. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume 15008, pp. 253–262. Springer Nature Switzerland, 2024. doi: 10.1007/978-3-031-72111-3_24.

Go-Eun Lee, Seon Ho Kim, Jungchan Cho, Sang Tae Choi, and Sang-Il Choi. Text-guided cross-position attention for segmentation: Case of medical image. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, volume 14224, pp. 537–546. Springer Nature Switzerland, 2023. doi: 10.1007/978-3-031-43904-9_52.

Aibing Li, Xinyi Zeng, Pinxian Zeng, Sixian Ding, Peng Wang, Chengdi Wang, and Yan Wang. Textmatch: Using text prompts to improve semi-supervised medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume 15008, pp. 699–709. Springer Nature Switzerland, 2024a. doi: 10.1007/978-3-031-72111-3_66.

Lei Li, Sheng Lian, Zhiming Luo, Beizhan Wang, and Shaozi Li. VCLIPSeg: Voxel-wise CLIP-enhanced model for semi-supervised medical image segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume 15009, pp. 692–701. Springer Nature Switzerland, 2024b. doi: 10.1007/978-3-031-72114-4_66.

Shuailin Li, Chuyu Zhang, and Xuming He. Shape-aware semi-supervised 3d semantic segmentation for medical images. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pp. 552–561. Springer, 2020.

Zihan Li, Yunxiang Li, Qingde Li, Puyang Wang, Dazhou Guo, Le Lu, Dakai Jin, You Zhang, and Qingqi Hong. Lvit: Language meets vision transformer in medical image segmentation. *IEEE Transactions on Medical Imaging*, 43(1):96–107, 2024c. doi: 10.1109/TMI.2023.3291719.

Xiangde Luo, Jieneng Chen, Tao Song, and Guotai Wang. Semi-supervised medical image segmentation through dual-task consistency. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(10):8801–8809, 2021. doi: 10.1609/aaai.v35i10.17066.

S. P. Morozov, A. E. Andreychenko, N. A. Pavlov, A. V. Vladzymyrskyy, N. V. Ledikhova, V. A. Gombolevskiy, I. A. Blokhin, P. B. Gelezhe, A. V. Gonchar, and V. Yu. Chernina. MosMedData: Chest CT scans with COVID-19 related findings dataset. *arXiv preprint arXiv:2005.06465*, 2020. URL https://arxiv.org/abs/2005.06465.

Qingtao Pan, Wenhao Qiao, Jingjiao Lou, Bing Ji, and Shuo Li. DuSSS: Dual semantic similarity-supervised vision-language model for semi-supervised medical image segmentation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(6):6299–6307, 2025. doi: 10.1609/aaai.v39i6.32674.

Md Maklachur Rahman and Tracy Hammond. Dual-domain cross-modal decoding for clinical text-guided medical image segmentation. *arXiv preprint arXiv:2608.11335*, 2026. URL https://arxiv.org/abs/2608.11335. Accepted at MICCAI 2026.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, pp. 234–241. Springer, 2015.

Zhiqiang Shen, Peng Cao, Hua Yang, Xiaoli Liu, Jinzhu Yang, and Osmar R. Zaiane. Co-training with high-confidence pseudo labels for semi-supervised medical image segmentation. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pp. 4199–4207, 2023. doi: 10.24963/ijcai.2023/467.

Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Yongchao Wang, Bin Xiao, Xiuli Bi, Weisheng Li, and Xinbo Gao. MCF: Mutual correction framework for semi-supervised medical image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15651–15660, 2023.

Lihe Yang, Lei Qi, Litong Feng, Wayne Zhang, and Yinghuan Shi. Revisiting weak-to-strong consistency in semi-supervised semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7236–7246, 2023.

Zhao Yang, Jiaqi Wang, Yansong Tang, Kai Chen, Hengshuang Zhao, and Philip H. S. Torr. LAVT: Language-aware vision transformer for referring image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18155–18165, 2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Yang_LAVT_Language-Aware_Vision_Transformer_for_Referring_Image_Segmentation_CVPR_2022_paper.html.

Lequan Yu, Shujun Wang, Xiaomeng Li, Chi-Wing Fu, and Pheng-Ann Heng. Uncertainty-aware self-ensembling model for semi-supervised 3d left atrium segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, pp. 605–613. Springer, 2019. doi: 10.1007/978-3-030-32245-8_67.

Qingjie Zeng, Huan Luo, Xinke Ma, Zilin Lu, Yang Hu, and Yong Xia. Exploring text-enhanced mixture-of-experts for semi-supervised medical image segmentation with composite data. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, volume 15965, pp. 226–236. Springer Nature Switzerland, 2025a. doi: 10.1007/978-3-032-04978-0_22. URL https://papers.miccai.org/miccai-2025/0315-Paper1870.html.

Qingjie Zeng, Yutong Xie, Zilin Lu, Mengkang Lu, Jingfeng Zhang, and Yong Xia. Consistency-guided differential decoding for enhancing semi-supervised medical image segmentation. *IEEE Transactions on Medical Imaging*, 44(1):44–56, 2025b. doi: 10.1109/TMI.2024.3429340.

Qingjie Zeng, Huan Luo, Xinke Ma, Zilin Lu, Yang Hu, Mengkang Lu, Yanning Zhang, and Yong Xia. Decoupling target semantics via text-anchored visual contrast for semi-supervised medical image segmentation. *IEEE Transactions on Image Processing*, 35:3228–3239, 2026. doi: 10.1109/TIP.2026.3674365.