Dual-Domain Cross-Modal Decoding for Clinical Text-Guided Medical Image Segmentation

Md Maklachur Rahman and Tracy Hammond

Texas A&M University, College Station, TX 77843, USA
{maklachur,hammond}@tamu.edu

Abstract. Clinical text can narrow down what to segment, but recent text-guided designs emphasize spatial alignment while overlooking frequency content that governs texture and boundaries. We propose Dual-Domain Cross-Modal Decoding (DD-CMD) for clinical text-guided pulmonary infection segmentation, integrating two complementary forms of language guidance during decoding. In the spatial domain, Text-Guided Spatial Cross-Attention (TGSA) aligns multi-scale visual tokens with text semantics and updates features through gated residual fusion. In the frequency domain, Spectral-Text Adaptive Modulation (STAM) applies a 2D DCT to compute learnable band-energy statistics and predicts text-conditioned FiLM parameters to recalibrate decoder channels for frequency-aware decoding. DD-CMD embeds TGSA and STAM into a coarse-to-fine decoder ($7\times7 \rightarrow 56\times56$) and restores full-resolution masks using a lightweight two-stage refinement module. Experiments on QaTa-COV19 and MosMedData+ show that DD-CMD achieves 91.46% Dice / 84.26% mIoU and 81.95% Dice / 69.42% mIoU, respectively, with average gains of +1.96 Dice and +2.67 mIoU over the strongest prior baselines. Code: https://github.com/maklachur/DD-CMD.

Keywords: Medical Image Segmentation · Vision-Language Models · DCT · Text-Guided Segmentation · Frequency-Aware Decoding.

1 Introduction

Medical image segmentation delineates anatomical structures and pathological regions and is a key step in clinical image analysis. Accurate masks support diagnosis, treatment planning, and disease monitoring across diverse applications. Encoder–decoder networks, such as U-Net [19] and U-Net++ [25], are widely used baselines, while nnUNet [12] improves reliability through a self-configuring training pipeline. More recently, transformer-based designs such as Swin-UNet [5] strengthen long-range modeling. Despite this progress with image-only approaches, pulmonary infection segmentation remains challenging, where abnormalities can be diffuse and low-contrast, and boundaries often depend on subtle texture variations and acquisition-dependent noise.

To address these limitations, recent work has explored clinical text-guided medical image segmentation, where text provides complementary semantics about

what to delineate. LViT [13] demonstrates that pairing images with textual descriptions and enabling cross-modal interaction can improve over vision-only baselines. Subsequent approaches refine how language interacts with visual features: MAdapter [24] uses adaptor-style modules to facilitate bidirectional interaction, RecLMIS [11] enforces stronger cross-modal consistency via reconstruction-based conditioning, MG-UNet [7] introduces a learnable memory mechanism to retain multimodal information, and ViTexNet [3] adopts text-guided dynamic operations for lightweight conditioning [18].

While these methods highlight the benefits of language guidance, accurate lesion delineation remains challenging when abnormalities exhibit low contrast, diffuse appearance, or weak boundaries. Most text-guided fusion mechanisms primarily emphasize spatial alignment, whereas texture and boundary fidelity are also closely related to frequency content. A recent frequency-domain multimodal approach, FMISeg [22], explores this direction by decomposing inputs into low- and high-frequency components using DWT [17] and performing multimodal fusion through separate visual branches. This motivates exploring frequency-aware language guidance from a decoder-calibration perspective. In addition, full-resolution cross-modal interaction can be computationally demanding, motivating a coarse-to-fine design that reserves expensive interaction for lower resolutions.

Motivated by these observations, we propose Dual-Domain Cross-Modal Decoding (DD-CMD). Our key idea is to integrate two complementary forms of language guidance during decoding: spatial alignment indicates where the model should focus, while spectral calibration determines how decoder channels respond to frequency components that govern texture and boundary structure. Concretely, DD-CMD encodes images using ConvNeXt-Tiny [14] and encodes clinical text using a frozen PubMedBERT [8]. We then decode masks through three progressive stages ($7\times7 \rightarrow 14\times14 \rightarrow 28\times28 \rightarrow 56\times56$). At each stage, Text-Guided Spatial Cross-Attention (TGSA) aligns visual tokens with text semantics and updates features through a gated residual fusion. In parallel, Spectral-Text Adaptive Modulation (STAM) applies a 2D discrete cosine transform (DCT) [1], aggregates learnably gated band-energy statistics, and predicts text-conditioned feature-wise linear modulation (FiLM) parameters [16] to recalibrate decoder channels in a frequency-aware manner. Finally, we recover full 224×224 resolution through a two-stage refinement module that fuses shallow spatial features while preserving text consistency.

In summary, our contributions are threefold: (1) We propose a dual-domain cross-modal decoder that integrates two complementary forms of language guidance: TGSA for text-guided spatial cross-attention with gated residual fusion, and STAM for frequency-aware channel recalibration using DCT band-energy statistics and FiLM conditioning. (2) We introduce a coarse-to-fine cross-modal design that establishes image–text alignment through three decoder stages ($7\times7 \rightarrow 56\times56$) and recovers full-resolution structure with a two-stage refinement module. (3) We conduct extensive experiments on QaTa-COV19 and MosMedData+,

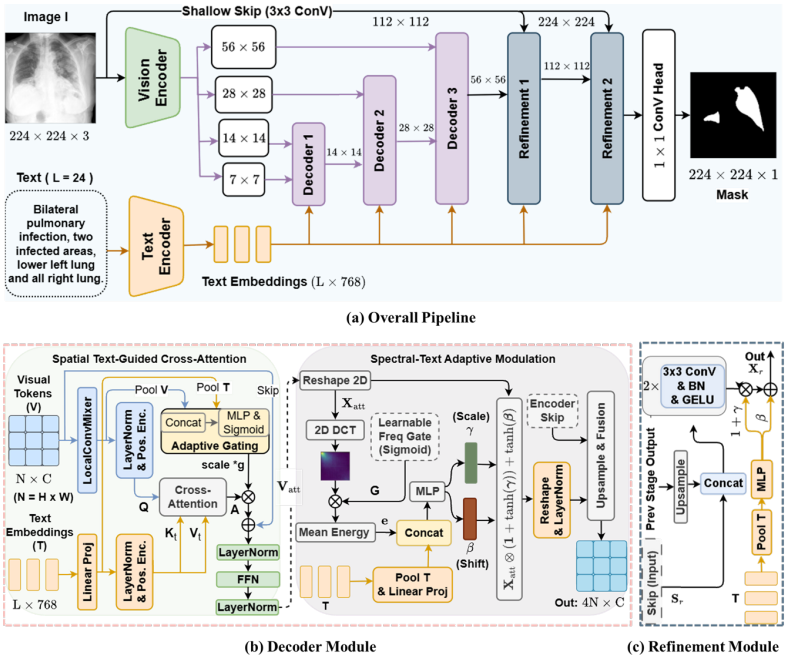

Fig. 1. (a) Overview of our proposed DD-CMD pipeline. (b) A single Decoder module. (c) A single Refinement module.

achieving 91.46% Dice / 84.26% mIoU on QaTa-COV19 and 81.95% Dice / 69.42% mIoU on MosMedData+, outperforming the recent baselines.

2 Methodology

An overview of DD-CMD is shown in Fig. 1. We propose DD-CMD, a text-guided medical image segmentation network that aggregates spatial cross-modal alignment with frequency-domain calibration. Given an input image $\mathbf{I} \in \mathbb{R}^{224\times224\times C}$ and a clinical text description $\mathbf{s}$, we predict a lesion mask in a coarse-to-fine decoder followed by lightweight high-resolution refinement. Our model has three key components. First, a vision encoder and a text encoder build multi-scale representations. Second, a dual domain cross-modal decoder reconstructs features from 7×7 to 56×56 for global semantic alignment. Third, a high-resolution refinement module restores full resolution from 56×56 to 224×224.

Our key design principle is to separate the two complementary roles of language guidance. Spatial attention determines where the model should focus, while spectral modulation determines how strongly each channel should respond to different frequency content. This design improves semantic alignment at low resolution and preserves fine details at high resolution.

2.1 Vision and Text Encoders

Image encoder: We adopt ConvNeXt-Tiny [14] as the vision encoder to encode the input image $I \in \mathbb{R}^{224\times224\times3}$. The encoder produces a four-scale feature pyramid $\{F_{56}, F_{28}, F_{14}, F_7\}$ at spatial resolutions $\{56\times56, 28\times28, 14\times14, 7\times7\}$ with channel dimensions $\{96, 192, 384, 768\}$, respectively. Low-resolution features capture the global context for coarse lesion localization, while high-resolution features preserve edge and texture details for better boundary reconstruction.

Text encoder: We tokenize each clinical text, pad or truncate it to a maximum length of L=24, and encode it with a frozen PubMedBERT (base-uncased) encoder [8]. Freezing the language branch preserves biomedical semantics and reduces overfitting when text diversity is limited. The encoder outputs token embeddings $\mathbf{T} \in \mathbb{R}^{L\times d}$ with d=768. We use $\mathbf{T}$ for cross-modal interaction and mean pooling to obtain a global descriptor $\bar{\mathbf{t}} \in \mathbb{R}^d$ for stage-wise conditioning.

2.2 Dual-Domain Cross-Modal Decoder

The decoder operates across three progressive stages: $7\times7 \to 14\times14$, $14\times14 \to 28\times28$, and $28\times28 \to 56\times56$. At each stage s, we process visual tokens $\mathbf{V}^{(s)} \in \mathbb{R}^{N_s\times C_s}$ (with $N_s = H_sW_s$) by two complementary components: Spatial Text-Guided Cross-Attention (TGSA) and Spectral-Text Adaptive Modulation (STAM).

TGSA: To align visual features with text semantics while preserving local structure, we first add spatial inductive bias through a local convolutional mixer comprising depthwise 3×3 and pointwise 1×1 convolutions with BatchNorm and GELU. This step helps maintain local boundary continuity when attention flattens spatial dimensions. We then perform multi-head cross-attention where locally-mixed visual tokens serve as queries $\mathbf{Q}$ and text embeddings serve as keys $\mathbf{K_t}$ and values $\mathbf{V_t}$. Here, $\tilde{V}^{(s)}$ denotes the output of the LocalConvMix block before cross-attention. It provides the visual queries for TGSA and is also used with text features to estimate the image-text agreement gate.

Clinical text can occasionally be vague or uninformative. To reduce over-conditioning, we introduce an adaptive gate $g^{(s)} \in (0, 1)$ that scales text influence based on multi-modal agreement:

$$g^{(s)} = \sigma\Big(\mathrm{MLP}\left([\mathrm{Avg}(\tilde{\mathbf{V}}^{(s)});\ \mathrm{Avg}(\mathbf{T}^{(s)})]\right)\Big), \tag{1}$$

$$\mathbf{V}^{(s)}_{\mathrm{att}} = \mathrm{LN}\Big(\mathbf{V}^{(s)} + \alpha^{(s)} g^{(s)}\,\mathrm{MHA}(\mathbf{Q}, \mathbf{K_t}, \mathbf{V_t})\Big), \tag{2}$$

where $\alpha^{(s)}$ is a learnable scale parameter and LN denotes LayerNorm. We follow this with a feed-forward network (FFN) and another LN layer. The scale $\alpha^{(s)}$ is learned separately at each decoder stage, enabling the residual text update to be calibrated in a stage-specific manner.

STAM: While spatial attention determines where to focus, medical images often exhibit acquisition-dependent frequency characteristics that benefit from explicit control over which frequency components to preserve. Medical anomalies exhibit distinct frequency signatures. STAM calibrates features in the frequency domain conditioned on text semantics. DCT provides a real-valued, orthogonal, and parameter-free representation for compact channel-wise spectral energy estimation. STAM combines these spectral descriptors with the global text embedding to predict FiLM parameters for frequency-aware decoder calibration.

Given $\mathbf{X}_{\text{att}}^{(s)} \in \mathbb{R}^{C_s \times H_s \times W_s}$ (reshaped from attended tokens), we apply channel-wise 2D DCT-II [1] to obtain frequency coefficients. Rather than fixed frequency band partitioning, we use a learnable gate $\mathbf{G}^{(s)} \in (0,1)^{H_s \times W_s}$ that softly weights informative frequency components when computing per-channel spectral energy:

$$e_c^{(s)} = \frac{1}{H_s W_s} \sum_{u,v} \sigma(\mathbf{G}_{u,v}^{(s)}) \left(\text{DCT}(\mathbf{X}_{\text{att}}^{(s)})_{c,u,v} \right)^2, \quad c = 1, \ldots, C_s. \tag{3}$$

This forms a compact spectral descriptor $\mathbf{e}^{(s)} \in \mathbb{R}^{C_s}$. We concatenate $\mathbf{e}^{(s)}$ with the global text vector $\bar{\mathbf{t}}$ and predict FiLM parameters [16] $(\boldsymbol{\gamma}^{(s)}, \boldsymbol{\beta}^{(s)}) \in \mathbb{R}^{C_s}$ using a two-layer MLP. We apply channel-wise affine modulation and bound both scale and shift with $\tanh(\cdot)$ for stability:

$$[\boldsymbol{\gamma}^{(s)}, \boldsymbol{\beta}^{(s)}] = \text{MLP}\Big([\mathbf{e}^{(s)};\ \bar{\mathbf{t}}] \Big), \tag{4}$$

$$\mathbf{X}_{\text{cal}}^{(s)} = \mathbf{X}_{\text{att}}^{(s)} \otimes \Big(1 + \tanh(\boldsymbol{\gamma}^{(s)}) \Big) + \tanh(\boldsymbol{\beta}^{(s)}), \tag{5}$$

where $\otimes$ denotes element-wise multiplication. At the end of each decoder stage, we upsample $\mathbf{X}_{\text{cal}}^{(s)}$ by $2\times$ via bilinear interpolation, concatenate the corresponding higher-resolution encoder skip feature $\mathbf{F}^{(s)}$, and refine the fused features using two 3×3 Conv-BN-GELU layers. Repeating this process across three stages yields text-conditioned features at 56×56.

2.3 High-Resolution Refinement Module

Computing full cross-attention at 112×112 and 224×224 is computationally expensive and largely redundant once global semantic alignment is established. We therefore stop cross-modal decoding at 56×56 and recover pixel-level details through two lightweight refinement blocks.

We extract shallow image features directly from the input via small convolutional layers, producing $\mathbf{S}_{112} \in \mathbb{R}^{48\times112\times112}$ (stride 2) and $\mathbf{S}_{224} \in \mathbb{R}^{24\times224\times224}$ (stride 1). These shallow pathways preserve edges and textures that are attenuated in deeper encoder stages. At each refinement resolution $r \in \{112, 224\}$, we bilinearly upsample the incoming features, concatenate the corresponding shallow skip $\mathbf{S}_r$, and apply two 3×3 Conv-BN-GELU layers. We then similarly apply lightweight FiLM conditioning from $\bar{\mathbf{t}}$ to preserve text consistency at high resolution. Finally, a 1×1 convolution produces segmentation logits at 224×224, then a sigmoid is applied to predict the final mask.

3 Experiments and Results

Datasets, Implementation Details, and Evaluation: We evaluate our method on two publicly available medical vision–language segmentation benchmarks, namely MosMedData+ [15, 13] and QaTa-COV19 [6, 13]. MosMedData+ provides 2,729 COVID-19 lung CT slices with ground-truth infection masks and corresponding clinical text. QaTa-COV19 contains 9,258 chest X-ray images with lesion annotations and accompanying textual descriptions. Following [13, 9, 3, 22], we adopt the exactly same train/val/test splits for fair comparison: MosMedData+ is divided into 2,183/273/273 samples, and QaTa-COV19 is split into 5,716/1,429/2,113 samples for training/validation/testing, respectively. Following prior text-guided benchmarks, each sample is treated as a slice-level image-text-mask triplet.

We implement our method in PyTorch with PyTorch-Lightning and run experiments on a single NVIDIA RTX 3090Ti GPU (24 GB). We resize all input images to 224×224 and apply random zoom, horizontal/vertical flips, and rotation, followed by intensity normalization. We train the network for 160 epochs with AdamW and cosine annealing (initial learning rate 5×10^{-5}, $\eta_{\min} = 10^{-6}$) using a batch size of 8. We optimize an equally weighted Dice and cross-entropy objective [2]: $\mathcal{L} = \mathcal{L}_{Dice} + \mathcal{L}_{CE}$. Consistent with prior work [13, 9, 3, 22], we report Dice and mIoU (%) as primary metrics. We additionally assess boundary quality using HD95 (pixels) and include it in our ablations.

Comparison with SOTA Methods: Table 1 compares text-free backbones and recent text-guided methods on QaTa-COV19 and MosMedData+. Text guidance consistently improves performance, highlighting the importance of clinical descriptions for infection delineation. DD-CMD achieves the best results on both datasets, with 91.46/84.26 Dice/mIoU on QaTa-COV19 and 81.95/69.42 on MosMedData+. Compared to MMI-UNet [4] on QaTa-COV19, DD-CMD improves by +0.58 Dice and +0.98 mIoU. On MosMedData+, it surpasses the best prior Dice by +3.33 (MAdapter [24]) and the best prior mIoU by +4.35 (ReCLMIS [11]), while using 47.77M parameters and 20.31 GFLOPs, comparable to several baselines. We also provide the qualitative comparison with the baselines in Fig. 2, showing better overlap on both datasets.

Ablation Study on the Effect of Different Core Modules: Table 2 analyzes the contribution of each component in DD-CMD. The image-only baseline performs the worst on both datasets, highlighting the benefit of clinical text for target specification. When we keep text but disable both TGSA and STAM, performance improves only slightly, indicating that naive conditioning is insufficient. Enabling TGSA brings a clear jump by improving text–image alignment, and adding STAM further boosts overlap while reducing boundary error, showing that spectral calibration complements spatial alignment. The high-resolution refinement primarily sharpens boundaries, while FiLM-style parameters and local mixing in TGSA bring additional consistent gains. Overall, the full model is

Table 1. Quantitative comparison on QaTa-COV19 and MosMedData+. Best and second-best are shown in **bold** and underlined. Arch.: CNN, Transformer(Trans.), SAM, Hybrid (CNN–Transformer). Text indicates whether the method uses clinical texts; N/R indicates Not Reported. P and F indicate Parameter and FLOPs.

Method	Arch.	Text	Venue	P ↓ (M)	F↓ (G)	QaTa-COV19		MosMedData+	
						Dice ↑	mIoU ↑	Dice ↑	mIoU ↑
U-Net [19]	CNN	×	MICCAI'15	14.8	50.3	79.02	69.46	64.60	50.73
U-Net++ [25]	CNN	×	MICCAI'18	74.5	94.6	79.62	70.25	71.75	58.39
nnUNet [12]	CNN	×	Nature'21	19.1	412.7	80.42	70.81	72.59	60.36
Swin-UNet [5]	Hybrid	×	ECCV'22	82.3	67.3	78.07	68.34	63.29	50.19
LAVT [21]	Trans.	✓	CVPR'22	118.6	83.8	79.28	69.89	73.29	60.41
LViT [13]	Hybrid	✓	IEEE TMI'23	29.7	54.1	83.66	75.11	74.57	61.33
LGA [10]	Trans.	✓	MICCAI'24	**8.24**	381.1	84.65	76.23	75.63	62.52
MAdapter [24]	CNN	✓	MICCAI'24	N/R	N/R	90.22	82.16	78.62	64.78
TGCAM [9]	CNN	✓	MICCAI'24	N/R	N/R	90.60	82.81	77.82	63.69
MMI-UNet [4]	Hybrid	✓	MICCAI'24	56.2	22.1	90.88	83.28	78.42	64.50
RecLMIS [11]	Hybrid	✓	IEEE TMI'24	23.7	24.1	85.22	77.00	77.48	65.07
ARSeg [20]	CNN	✓	MICCAI'25	N/R	N/R	84.09	72.64	73.24	59.82
MG-UNet [7]	Hybrid	✓	MICCAI'25	30.5	**11.0**	88.10	77.80	76.39	61.79
TextMoE [23]	Hybrid	✓	MICCAI'25	N/R	N/R	89.08	80.32	74.66	59.57
ViTexNet [3]	Hybrid	✓	MICCAI'25	37.7	11.5	90.76	83.25	78.19	64.04
Ours	Hybrid	✓	–	47.77	20.31	**91.46**	**84.26**	**81.95**	**69.42**

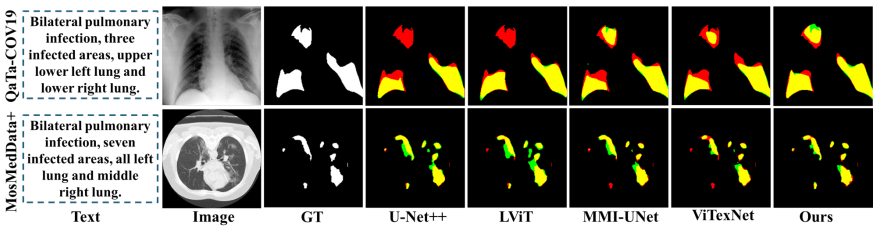

Fig. 2. Qualitative comparison on QaTa-COV19 and MosMedData+. Overlays: yellow = true positives, red = false negatives, green = false positives. Best viewed zoomed in.

best and most stable across datasets, showing that our decoder modules interact effectively with visual features and align lesion regions with textual semantics.

Ablation Study on the Effect of Text Length (Text Tokens): We vary text length ($L : 8 \rightarrow 40$) to assess how much text is needed for strong text–image alignment. Table 3 shows that very short prompts reduce overlap and boundary quality because truncation can remove lesion descriptors needed for localization and shape. Performance improves as L increases and then plateaus: L=24 provides a strong, stable default across datasets, while longer texts (L=32, 40) provide no consistent gains and can slightly degrade due to diminishing returns and added noise from extra tokens. We therefore use L=24 in all main experiments for our model.

Table 2. Effect of different core modules on QaTa-COV19 and MosMedData+. ↑/↓ indicate higher/lower is better; *w/o* means without. Best values are bolded.

Model Variant	QaTa-COV19			MosMedData+		
	Dice ↑	mIoU ↑	HD95 ↓	Dice ↑	mIoU ↑	HD95 ↓
No Text (Image-only)	88.17	78.84	26.90	79.76	66.33	18.60
w/o TGSA & STAM	89.35	80.75	22.70	80.52	67.39	16.29
w/o TGSA	90.47	82.60	16.95	81.01	68.08	15.03
w/o STAM	90.92	83.35	15.71	81.19	68.34	14.48
w/o Refinement	91.07	83.60	14.44	81.35	68.56	14.12
w/o FiLM-Style Params (γ, β)	91.27	83.94	13.06	81.54	68.83	14.56
w/o Adapting Gating in TGSA	91.10	83.65	14.40	81.41	68.65	14.09
w/o LocalConvMix in TGSA	91.33	84.04	13.44	81.68	69.03	14.20
Full model (Ours)	**91.46**	**84.26**	**12.76**	**81.95**	**69.42**	**13.94**

Table 3. Results on varied text token length (L) on QaTa-COV19 and MosMedData+.

text Length (L)	QaTa-COV19			MosMedData+		
	Dice ↑	mIoU ↑	HD95 ↓	Dice ↑	mIoU ↑	HD95 ↓
$L = 8$	88.85	79.93	19.70	81.27	68.45	17.18
$L = 16$	91.22	83.85	13.28	81.46	68.72	14.97
$L = 24$ (Ours)	**91.46**	**84.26**	**12.76**	**81.95**	**69.42**	**13.94**
$L = 32$	91.26	83.92	13.17	81.22	68.38	14.79
$L = 40$	91.17	83.77	13.80	80.90	67.93	16.47

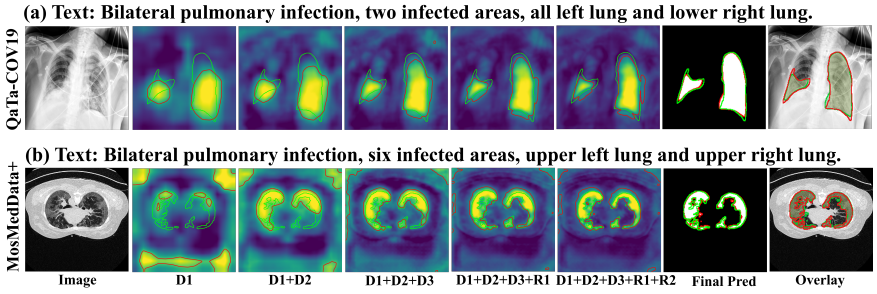

Fig. 3. Progressive feature map and boundary visualization. Decoder stages (D1–D3) and refinement modules (R1–R2) are progressively added to visualize stage-wise changes. Boundaries: *green = ground truth*, *red = prediction*. The final columns show the predicted mask and boundary overlay. Best viewed when zoomed in.

Progressive Feature Visualization and Boundary Updates: Fig. 3 shows how the prediction improves as we progressively add decoder and refinement blocks. Decoder (D1) gives a rough, low-resolution localization with broad activations and loose boundaries. Adding D2 and D3 strengthens lesion-specific responses and suppresses background, so the mask becomes semantically reliable and better covers the infection extent. After this point, refinement mainly

improves boundaries: R1 sharpens edges and reduces leakage into nearby tissue, and R2 closes small gaps and stabilizes thin structures.

4 Conclusion

We proposed DD-CMD for clinical text-guided pulmonary infection segmentation, integrating spatial and spectral language guidance within a unified decoding framework. TGSA uses text-guided spatial cross-attention to align visual tokens with clinical semantics, whereas STAM calibrates decoder features through DCT band-energy statistics and text-conditioned FiLM parameters for frequency-aware decoding. This dual-domain design is implemented in a coarse-to-fine decoder ($7\times7 \rightarrow 56\times56$), with a refinement pathway to recover full-resolution masks. Experiments on QaTa-COV19 and MosMedData+ demonstrate the effectiveness of DD-CMD with strong overlap and boundary accuracy.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Ahmed, N., Natarajan, T., Rao, K.R.: Discrete cosine transform. IEEE transactions on Computers **100**(1), 90–93 (1974)
2. Azad, R., Heidary, M., Yilmaz, K., Hüttemann, M., Karimijafarbigloo, S., Wu, Y., Schmeink, A., Merhof, D.: Loss functions in the era of semantic segmentation: A survey and outlook. arXiv preprint arXiv:2312.05391 (2023)
3. Bhardwaj, R., Tambe, U.Y., Neog, D.R.: Vitexnet: Vision-text guided dynamic convolution network for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 690–699. Springer (2025)
4. Bui, P.N., Le, D.T., Choo, H.: Visual-textual matching attention for lesion segmentation in chest images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 702–711. Springer (2024)
5. Cao, H., Wang, Y., Chen, J., Jiang, D., Zhang, X., Tian, Q., Wang, M.: Swin-unet: Unet-like pure transformer for medical image segmentation. In: European conference on computer vision. pp. 205–218. Springer (2022)
6. Degerli, A., Kiranyaz, S., Chowdhury, M.E., Gabbouj, M.: Osegnet: Operational segmentation network for covid-19 detection using chest x-ray images. In: 2022 IEEE International Conference on Image Processing (ICIP). pp. 2306–2310. IEEE (2022)
7. Ding, S., Li, M., Wang, C.: Mg-unet: A memory-guided unet for lesion segmentation in chest images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 355–365. Springer (2025)
8. Gu, Y., Tinn, R., Cheng, H., Lucas, M., Usuyama, N., Liu, X., Naumann, T., Gao, J., Poon, H.: Domain-specific language model pretraining for biomedical natural language processing. ACM Transactions on Computing for Healthcare **3**(1), 2:1–2:23 (Jan 2022). https://doi.org/10.1145/3458754

9. Guo, Y., Zeng, X., Zeng, P., Fei, Y., Wen, L., Zhou, J., Wang, Y.: Common vision-language attention for text-guided medical image segmentation of pneumonia. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 192–201. Springer (2024)
10. Hu, J., Li, Y., Sun, H., Song, Y., Zhang, C., Lin, L., Chen, Y.W.: Lga: A language guide adapter for advancing the sam model's capabilities in medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 610–620. Springer (2024)
11. Huang, X., Li, H., Cao, M., Chen, L., You, C., An, D.: Cross-modal conditioned reconstruction for language-guided medical image segmentation. IEEE Transactions on Medical Imaging **44**(4), 1821–1835 (2024)
12. Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H.: nnu-net: a self-configuring method for deep learning-based biomedical image segmentation. Nature methods **18**(2), 203–211 (2021)
13. Li, Z., Li, Y., Li, Q., Wang, P., Guo, D., Lu, L., Jin, D., Zhang, Y., Hong, Q.: Lvit: language meets vision transformer in medical image segmentation. IEEE transactions on medical imaging **43**(1), 96–107 (2023)
14. Liu, Z., Mao, H., Wu, C.Y., Feichtenhofer, C., Darrell, T., Xie, S.: A convnet for the 2020s. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 11976–11986 (2022)
15. Morozov, S.P., Andreychenko, A.E., Pavlov, N.A., Vladzymyrskyy, A., Ledikhova, N.V., Gombolevskiy, V.A., Blokhin, I.A., Gelezhe, P.B., Gonchar, A., Chernina, V.Y.: Mosmeddata: Chest ct scans with covid-19 related findings dataset. arXiv preprint arXiv:2005.06465 (2020)
16. Perez, E., Strub, F., De Vries, H., Dumoulin, V., Courville, A.: Film: Visual reasoning with a general conditioning layer. In: Proceedings of the AAAI conference on artificial intelligence. vol. 32 (2018)
17. Rahman, M.M.: A dwt, dct and svd based watermarking technique to protect the image piracy. arXiv preprint arXiv:1307.3294 (2013)
18. Rahman, M.M., Jung, S.K., Hammond, T.: Mambaliteunet: Cross-gated adaptive feature fusion for robust skin lesion segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8556–8565 (2026)
19. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical image computing and computer-assisted intervention. pp. 234–241. Springer (2015)
20. Wang, Q., Lin, X., Yan, Z.: Towards robust medical image referring segmentation with incomplete textual prompts. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 636–646. Springer (2025)
21. Yang, Z., Wang, J., Tang, Y., Chen, K., Zhao, H., Torr, P.H.: 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)
22. Yu, B., Yang, J., Du, Z., Huang, Y., Li, C., Wang, L.: Frequency-domain multimodal fusion for language-guided medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 278–288. Springer (2025)
23. Zeng, Q., Luo, H., Ma, X., Lu, Z., Hu, Y., Xia, Y.: Exploring text-enhanced mixture-of-experts for semi-supervised medical image segmentation with composite data. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 226–236. Springer (2025)

24. Zhang, X., Ni, B., Yang, Y., Zhang, L.: Madapter: A better interaction between image and language for medical image segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 425–434. Springer (2024)
25. Zhou, Z., Rahman Siddiquee, M.M., Tajbakhsh, N., Liang, J.: Unet++: A nested u-net architecture for medical image segmentation. In: Deep learning in medical image analysis and multimodal learning for clinical decision support, pp. 3–11. Springer (2018)