Reliable Text-Guided 3D Biomedical Segmentation with Scalable Volumetric Inference and Fragmentation-Aware Refinement

Zhou Jiang[1[0000−0001−8146−2028]], Boyan Lv[1[0009−0003−1075−4770]], Changkai Ji[1[0009−0007−7090−7360]], and Lisheng Wang[1[0000−0003−3234−7511]]*

School of Automation and Intelligent Sensing, Shanghai Jiao Tong University, Shanghai 200240, China
lswang@sjtu.edu.cn

Abstract. Text-guided 3D biomedical image segmentation requires models to handle heterogeneous imaging modalities, diverse anatomical and lesion targets, and large volumetric inputs under strict runtime and memory constraints. We present a reliable inference and refinement framework built on pretrained BiomedParseV2 for the CVPR 2025-2026 Foundation Models for Text-guided 3D Biomedical Image Segmentation challenge. Our method focuses on three practical aspects. First, we redesign the volumetric inference pipeline with adaptive slice batching, mixed-precision execution, chunked upsampling, text-embedding caching, and lossless output serialization to enable stable full-volume prediction across large 3D cases. Second, we introduce a fragmentation-aware instance mask refinement strategy that combines uncertainty-guided mask correction with conservative connected-component reassignment to suppress isolated fragmented false positives while preserving dominant anatomical or lesion instances. Third, we refine the pretrained checkpoint using a parameter-efficient decoder/head transfer stage while freezing the image backbone and text encoder. On the local full validation set, our final system produced all 2144 expected outputs without missing cases and achieved a mean DSC of 0.8272, mean NSD of 0.8925, mean $F1_{50}$ of 0.4904, and mean DSC-TP of 0.7152. Compared with the published BiomedParseV2 all-data baseline under our local validation protocol, this corresponds to improvements of 0.0272 DSC, 0.0096 NSD, 0.0295 $F1_{50}$, and 0.0270 DSC-TP. On the official final testing set, our team submission ranked first in the text-guided segmentation track, improving the BiomedParseV2 baseline by 0.013 average DSC, 0.012 average NSD, 0.027 average $F1_{50}$, and 0.040 average DSC-TP.

Keywords: Text-guided segmentation · 3D biomedical segmentation · Volumetric inference · Instance mask refinement · Transfer refinement

* Corresponding author.

1 Introduction

Text prompts provide a flexible interface for biomedical image segmentation because users can specify target organs, anatomical structures, lesions, or microscopic objects with natural language. Compared with box- or point-based interactive segmentation, text-guided segmentation can potentially support a much larger set of targets without manually drawing prompts for each object. However, extending text-guided segmentation to 3D biomedical images is challenging. The input volumes vary substantially in spatial size, slice thickness, intensity distribution, and anatomical field of view. The targets include large organs, thin vessels, small tumors, vertebrae, cardiac structures, ultrasound anatomy, PET lesions, and microscopy objects. A practical solution to this challenge must therefore balance segmentation accuracy, instance-level reliability, runtime, and GPU memory consumption.

Recent segmentation foundation models have made rapid progress in both natural and biomedical images. Segment Anything (SAM) [8] and SAM 2 [14] demonstrated strong promptable segmentation ability in natural images and videos.

Medical adaptations further explored promptable 2D and 3D medical segmentation [10,11,16], while SegVol, VISTA3D, and nnInteractive [2,5,3] focused on volumetric and interactive 3D settings. For text-guided biomedical segmentation, BiomedParseV2, CAT, and SAT [19,7,20] are representative approaches that connect textual target descriptions with biomedical visual representations. In particular, BiomedParseV2 provides a strong baseline for the challenge through a text-guided 3D inference pipeline.

The challenge also inherits the diversity and practical constraints of modern biomedical segmentation benchmarks. Public datasets and challenges such as the Medical Segmentation Decathlon [15], KiTS [6], FLARE [12], TotalSegmentator [17], LiTS [1], and autoPET [4] highlight the need for methods that generalize across organs, lesions, and imaging protocols. These benchmarks also make clear that practical systems must handle runtime and memory constraints, not only model architecture.

Our goal is not to replace the pretrained foundation model with a new architecture. Instead, we build a reliable challenge system around a strong pretrained BiomedParseV2 model and focus on the practical failure modes observed in text-guided 3D inference: large-volume runtime and memory pressure, fragmented instance masks, and the requirement that pretrained weights be incorporated through a training procedure rather than submitted unchanged. The resulting method combines scalable full-volume inference, fragmentation-aware instance refinement, and parameter-efficient transfer refinement.

The method makes three main contributions. First, we develop a scalable full-volume inference pipeline that combines adaptive slice batching, mixed-precision execution, chunked upsampling, text-embedding caching, and lossless output serialization, enabling reliable segmentation across large biomedical volumes with heterogeneous image sizes and modalities. Second, we introduce fragmentation-aware instance mask refinement, where uncertainty-guided mask correction and

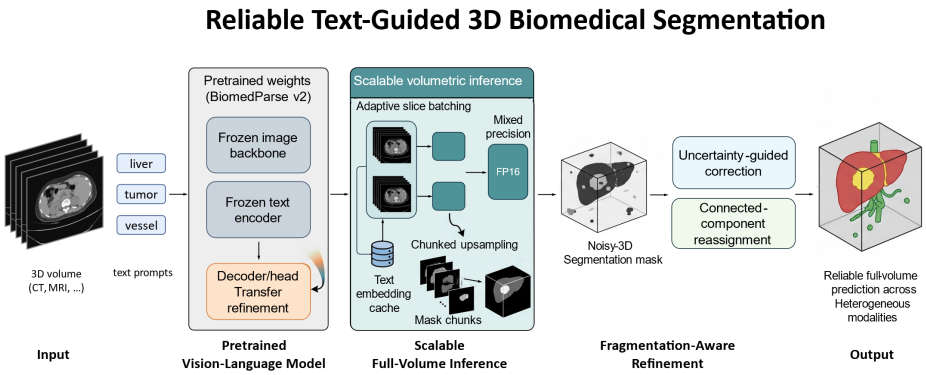

Fig. 1. Overview of the proposed reliable text-guided 3D biomedical segmentation system. The model is refined from pretrained BiomedParseV2, while the inference path is redesigned for scalable full-volume prediction and instance-mask stability.

conservative connected-component reassignment are used to suppress isolated false positives while preserving dominant anatomical or lesion instances. Third, we perform parameter-efficient transfer refinement of the pretrained BiomedParseV2 model by updating only the decoder/head components on the official challenge training data, while keeping the image backbone and text encoder frozen to retain pretrained biomedical vision-language knowledge.

2 Method

Figure 1 summarizes the proposed system. The method starts from a pretrained BiomedParseV2 model. The input volume and text prompts are converted into the format expected by the model. During inference, the volume is processed with memory-aware slice batching and mixed precision. The predicted prompt-specific masks are reconstructed with chunked upsampling and converted into a final segmentation map. For instance segmentation cases, we further apply uncertainty-guided refinement and connected-component reassignment to improve instance-mask stability.

2.1 Pretrained Text-Guided Segmentation Backbone

We use BiomedParseV2 as the base text-guided segmentation model. Given a 3D image volume and a set of text prompts, the model predicts prompt-specific mask logits together with object-existence scores. The original BiomedParseV2 inference strategy performs slice-based processing of the 3D volume and uses text prompts to condition the segmentation output. We preserve this foundation-model behavior because it already provides strong cross-modality biomedical recognition and segmentation ability.

For a case with target identifiers $\{c_i\}_{i=1}^K$ and text descriptions $\{t_i\}_{i=1}^K$, the text input is constructed by joining target prompts with a separator token. Denoting the input volume as V, the prompt-conditioned model outputs prompt-wise mask logits M_i and object-existence logits e_i:

$$\{M_i, e_i\}_{i=1}^K = f_\theta(V, \{t_i\}_{i=1}^K). \tag{1}$$

The object-existence score suppresses slices where the target is predicted to be absent. For voxel $\mathbf{x}$ on slice $z_\mathbf{x}$, we write the gated foreground probability as

$$p_i(\mathbf{x}) = \sigma(M_i(\mathbf{x}))\mathbf{1}\{\sigma(e_i(z_\mathbf{x})) > \gamma_i\}, \tag{2}$$

where $\sigma(\cdot)$ is the sigmoid function and γ_i is the object-existence threshold. Prompt-wise masks are then merged into a final multi-class segmentation map by comparing foreground prompt scores against a background prior:

$$i^\star(\mathbf{x}) = \arg\max_{1\le i\le K} p_i(\mathbf{x}), \quad \hat{y}(\mathbf{x}) = c_{i^\star(\mathbf{x})}\mathbf{1}\{\max_i p_i(\mathbf{x}) > \tau_{\text{bg}}\}. \tag{3}$$

2.2 Parameter-Efficient Transfer Refinement

The challenge rules allow pretrained weights only when they are incorporated through transfer learning, distillation, fine-tuning, or related training techniques. We therefore refine pretrained BiomedParseV2 weights on the official challenge training data. To reduce the risk of disrupting the pretrained biomedical vision-language representation and to make training feasible on limited GPU resources, we freeze the image backbone and text encoder and update only the decoder/head components.

This training stage uses 2D neighboring slices sampled from 3D training volumes. Each sampled slice stack is paired with the corresponding text prompt and segmentation target. The model is initialized from the pretrained BiomedParseV2 checkpoint, and the trainable decoder/head parameters are optimized with a low learning rate. Let θ_{img}, θ_{text}, and θ_{dec} denote the image backbone, text encoder, and decoder/head parameters, respectively. The refinement optimizes

$$\theta_{\text{dec}}^\star = \arg\min_{\theta_{\text{dec}}} \sum_{(V,T,Y)\in\mathcal{D}_{\text{train}}} \mathcal{L}\big(f_{\theta_{\text{img}},\theta_{\text{text}},\theta_{\text{dec}}}(V,T), Y\big), \quad \theta_{\text{img}}, \theta_{\text{text}} \text{ fixed.} \tag{4}$$

This yields a task-specific checkpoint while keeping the majority of pretrained parameters fixed.

2.3 Scalable Full-Volume Inference

Large 3D volumes are a central challenge in the competition. A fixed slice batch size can be inefficient for small cases and unsafe for large cases with many prompts. We therefore use an adaptive inference policy that selects slice batch

size, slice stride, and post-processing chunk size based on the number of prompts and volume shape:

$$(b, s, q) = \pi(D, H, W, K), \tag{5}$$

where π denotes a heuristic scheduling policy, (D, H, W) is the input volume shape, K is the number of prompts, b is the slice batch size, s is the slice stride, and q is the prompt chunk size for mask reconstruction. This policy is designed to preserve full-volume prediction while avoiding excessive intermediate tensors.

Mixed-precision execution is enabled during model inference. Most neural network operations are performed in FP16, while numerically sensitive post-processing steps are kept in appropriate precision. This reduces activation memory and improves throughput on GPUs with tensor-core support. For mask reconstruction, prompt-wise logits are upsampled in chunks rather than all at once. This chunked upsampling avoids large memory spikes when a case contains many target prompts or a long depth dimension.

Specifically, prompts are divided into chunks,

$$\mathcal{I}_m = \{(m-1)q + 1, \ldots, \min(mq, K)\}, \tag{6}$$

and upsampling is applied to one chunk at a time:

$$\widetilde{M}_{\mathcal{I}_m} = U(M_{\mathcal{I}_m}; D, H, W), \tag{7}$$

where $U(\cdot)$ restores logits to the original volume size. The final label map is accumulated across chunks, which keeps peak memory bounded by q rather than K.

Text prompts are constant within each case. We cache text embeddings keyed by the prompt tuple so that repeated forward passes for slice batches can reuse the same textual representation. The final segmentation is saved with lossless uncompressed serialization to reduce output I/O overhead while preserving exact label values.

2.4 Fragmentation-Aware Instance Mask Refinement

Instance segmentation cases, especially in microscopy and PET, can contain many small objects or lesions. A binary foreground mask with many isolated components can lead to fragmented false positives and unstable instance-level scores. We address this with a conservative instance-mask refinement stage.

First, we apply uncertainty-guided mask correction. Let $p_i(\mathbf{x})$ denote the predicted probability for prompt i at voxel $\mathbf{x}$ and let τ_i denote the foreground prior threshold. Voxels with probabilities close to the decision boundary have higher uncertainty than confident foreground or background voxels. We compute the Bernoulli entropy

$$H_i(\mathbf{x}) = -p_i(\mathbf{x}) \log p_i(\mathbf{x}) - (1 - p_i(\mathbf{x})) \log(1 - p_i(\mathbf{x})). \tag{8}$$

For candidate voxels whose probability is near the foreground prior and whose local neighborhood is supported by a 3D morphological closing operation, we

promote the probability slightly above the foreground prior. In practice, voxels satisfying both entropy and neighborhood support constraints are assigned foreground probabilities above τ_i. This correction is conservative: it only affects uncertain voxels spatially connected to existing foreground evidence.

Second, after the final instance segmentation map is obtained, we apply connected-component reassignment. Binary instance masks are converted to connected-component labels $\mathcal{C} = \{C_j\}_{j=1}^{J}$. We retain only components that pass a conservative size and rank criterion,

$$\mathcal{J} = \{j : |C_j| \geq a_{\min}\} \cap \mathrm{Top}_L(\{|C_j|\}_{j=1}^{J}), \tag{9}$$

and relabel components in $\mathcal{J}$ while suppressing the remaining small fragments. This step produces cleaner 3D instance masks and prevents unstable component proliferation while preserving dominant anatomical or lesion instances.

3 Experiments

3.1 Dataset and Evaluation Metrics

We evaluate the method on the text-guided 3D biomedical segmentation setting from the CVPR 2025-2026 challenge. The dataset covers multiple biomedical modalities, including computed tomography (CT), magnetic resonance imaging (MR), positron emission tomography (PET), ultrasound (US), and microscopy. The task contains both semantic segmentation and instance segmentation cases.

The validation cases include large multi-organ CT volumes, PET lesion cases, ultrasound anatomy, and microscopy instances, reflecting the heterogeneous data sources used in recent biomedical segmentation benchmarks [17,4].

For semantic segmentation cases, we report Dice Similarity Coefficient (DSC) and Normalized Surface Distance (NSD). DSC measures volumetric overlap between prediction and reference masks, while NSD measures boundary agreement. For instance segmentation cases, we report F1 score at an intersection-over-union threshold of 0.5 and DSC for true-positive instances (DSC-TP), following the need for task-aware metric selection in image analysis validation [13]. The official runtime constraint is 60 seconds per 3D case per class. Cases exceeding the limit can receive zero scores, making runtime reliability an important part of the system.

3.2 Implementation Details

Preprocessing. The challenge data are provided as 3D NPZ files with intensities normalized to the range $[0, 255]$. We follow the input format used by BiomedParseV2 and related MedSAM-style preprocessing pipelines [10,11]. During inference, each volume is padded and resized to the model input resolution, processed along a selected valid axis, and restored to the original spatial size before saving. No external data are used beyond the official challenge data and pretrained model weights.

Environment settings. Table 1 summarizes the local development environment. The final Docker submission is designed to match the challenge evaluation setting with a single visible GPU.

Table 1. Development environments and requirements. (mandatory table)

System	Ubuntu Linux server
Local CPU	Intel Core i9-10920X
Local RAM	128 GB
Local experiment GPU	NVIDIA GeForce RTX 3090 24 GB
Target evaluation GPU	NVIDIA Quadro RTX5000 16 GB
CUDA version	12.1 compatible runtime
Programming language	Python 3
Deep learning framework	PyTorch

Training protocols. The parameter-efficient transfer refinement stage freezes the image backbone and text encoder and updates only decoder/head components. The training script samples neighboring 2D slices from 3D volumes, constructs text-guided training pairs, and optimizes the trainable parameters with mixed precision and gradient accumulation. The objective combines classification and segmentation losses, following the common use of compound losses in medical image segmentation [9]. Because the refinement stage uses stochastic slice-stack sampling rather than complete passes over the full 3D training set, we report its training budget by optimization steps rather than dataset epochs. Table 2 lists the main training configuration.

4 Results and Discussion

4.1 Quantitative Results on Validation Set

All quantitative results reported in this section are validation-set results. Hidden-test labels were not available at the time of writing. Table 3 compares the published BiomedParseV2 all-data baseline from the challenge webpage with our local full validation results computed with the official metric definitions. The comparison should be interpreted as a validation reference rather than a controlled hidden-test leaderboard comparison. Best values in each metric column are shown in bold.

The final system completed all validation cases in the local full validation run. All 2144 validation cases produced outputs, with no missing cases and no extra files. The measured mean per-case inference time was 7.89 seconds, the 95th percentile was 31.37 seconds, and the maximum case time was 46.53 seconds. The peak GPU memory was approximately 12.37 GiB, below the 16 GB target hardware budget. Figure 2 visualizes both the validation metric gains and the runtime distribution over the local full validation set.

Table 2. Training protocols. (mandatory table)

Pre-trained Model	BiomedParseV2
Training data	Official challenge training data
Data augmentation	Random flips, 90-degree rotations, intensity jitter
Data sampling strategy	Neighboring slice stacks from 3D volumes
Trainable modules	Decoder/head only
Frozen modules	Image backbone and text encoder
Batch size	1 with gradient accumulation of 8
Patch size	512×512 slice-based input
Total epochs	N/A; selected 500-step checkpoint
Optimizer	AdamW
Initial learning rate (lr)	1×10^{-6}
Lr decay schedule	Fixed lr in selected run
Training time	Not recorded separately
Loss function	BiomedParseLossCLS + MedSamLoss
Precision	Mixed precision
Number of model parameters	371.80 M total
Number of trainable parameters	70.43 M
Number of flops	N/A
Optimal model selection criteria	Local validation sweep

Table 3. Validation performance. The official BiomedParseV2 baseline values are the published all-data baseline reported by the challenge webpage. Our results are computed on the local full validation set. Best values in each metric column are shown in bold.

Method	DSC	NSD	$F1_{50}$	DSC-TP
Published BiomedParseV2 baseline	0.8000	0.8829	0.4609	0.6882
Ours, local full validation	**0.8272**	**0.8925**	**0.4904**	**0.7152**

4.2 Qualitative Results on Validation Set

Figure 3 illustrates the qualitative behavior of the final validation model across heterogeneous modalities. The ROI-cropped CT, MR, and PET views make the local mask agreement easier to inspect than full-slice visualization. In the CT and MR examples, the predictions preserve the dominant anatomical extent and produce contiguous foreground regions, with most residual errors concentrated near organ boundaries. In the PET and microscopy examples, the targets are smaller and more instance-like; the predictions localize the main foreground objects, while remaining false positives and false negatives appear mainly around weak or ambiguous object boundaries. These examples support the quantitative results by showing stable full-volume segmentation behavior across both large anatomical structures and smaller lesion or cellular targets.

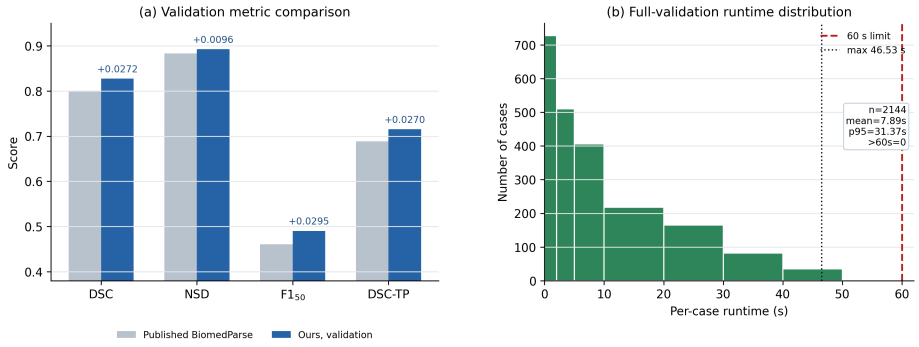

Fig. 2. Validation performance and runtime analysis. Panel (a) compares the published BiomedParseV2 all-data baseline with our local full validation scores, with absolute improvements annotated above the bars. Panel (b) shows the per-case runtime distribution over all 2144 validation cases, together with the 60-second runtime limit and the observed maximum runtime.

Table 4. Efficiency summary on the local full validation run.

Metric	Value
Number of validation cases	2144
Completed outputs	2144
Missing outputs	0
Extra outputs	0
Mean per-case time	7.89 s
95th percentile time	31.37 s
Maximum per-case time	46.53 s
Cases over 60 s	0
Peak GPU memory	12.37 GiB

4.3 Ablation and Design Analysis

We performed targeted ablation experiments to decide which components should be kept in the final system. Table 5 summarizes the main system-level comparison. The reproduced pretrained checkpoint already benefits substantially from the reliable full-volume inference path. Decoder-only fine-tuning provides a small additional gain in DSC and $F1_{50}$, but slightly decreases DSC-TP. The final refinement stack further improves $F1_{50}$ at a small DSC-TP cost. We therefore interpret the final gain as being driven mainly by reliable full-volume inference and instance-level cleanup rather than by fine-tuning alone. We hypothesize that this adaptation helps align the object-existence and mask prediction heads with the challenge prompt distribution, although the main validated gains come from the full-volume inference and refinement pipeline. NSD was computed for the published baseline and for our final local full validation run, but not for the intermediate reruns; therefore, the final NSD value should be read as the final system's complete validation result rather than as an incremental NSD ablation.

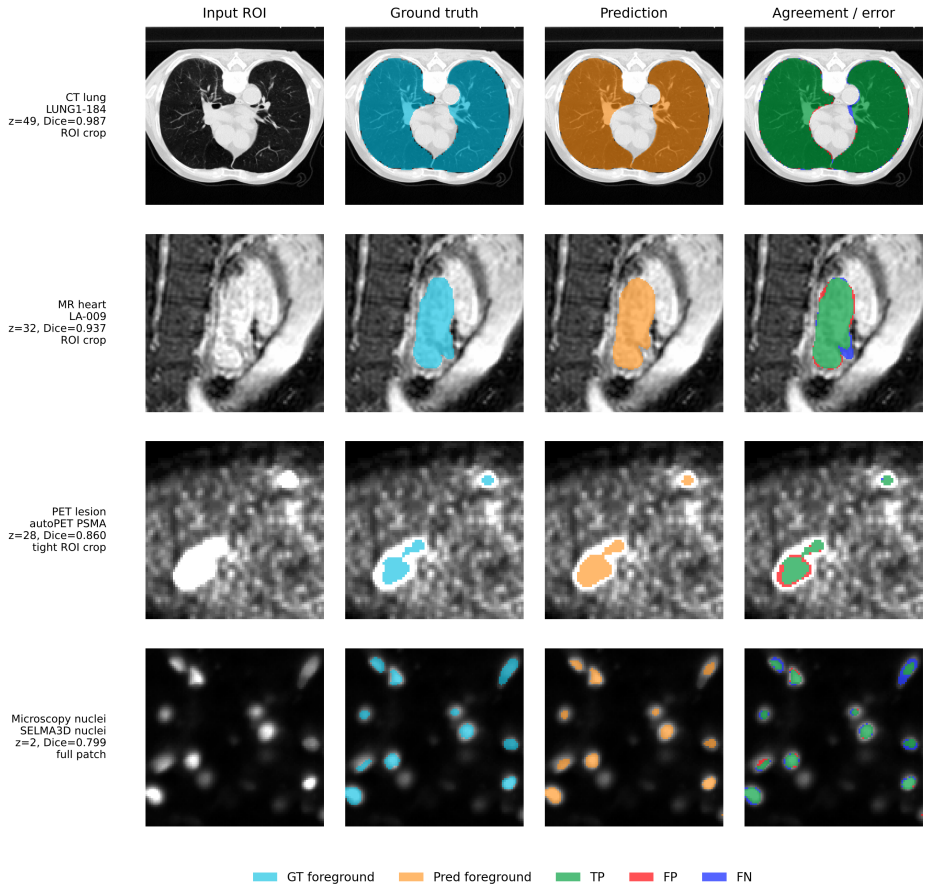

Fig. 3. Representative qualitative results on four validation cases. Cases were selected to cover representative modalities. The panel includes CT lung, MR heart, PET lesion, and microscopy nuclei examples. For CT, MR, and PET, the displayed view is cropped around the foreground ROI to make the segmentation details visible; the microscopy example is shown as the local patch. For each case, we show the input view, ground-truth foreground, predicted foreground, and an agreement/error map. Green denotes true-positive foreground overlap, red denotes false-positive voxels, and blue denotes false-negative voxels.

Table 5. Ablation of the proposed practical improvements on the validation set. The official BiomedParseV2 numbers are organizer-reported. Our reproduced and fine-tuned variants are evaluated on the same local validation split. NSD is reported for the organizer-reported baseline and for our final local full validation run; intermediate reruns did not include surface-distance evaluation. Best values in each metric column are shown in bold.

Method / Variant	Cases	DSC	NSD	$F1_{50}$	DSC-TP
Published BiomedParseV2 baseline	–	0.8000	0.8829	0.4609	0.6882
Reproduced checkpoint + full-volume inference	2144	0.8268	–	0.4758	**0.7180**
+ decoder-only fine-tuning	2144	**0.8272**	–	0.4768	0.7171
Final inference/refinement stack	2144	**0.8272**	**0.8925**	**0.4904**	0.7152

Table 6. Controlled evidence for each highlighted component. Fragmentation-aware refinement is evaluated on the 83 microscopy instance cases. Parameter-efficient fine-tuning is evaluated by comparing the selected decoder-only checkpoint against the reproduced original checkpoint on the local full validation set.

Highlight	Evidence/result
Scalable inference	2144/2144 outputs; 0 cases >60 s; 12.37 GiB peak GPU memory
Instance refinement	83 microscopy instance cases; $\Delta F1_{50} = +0.0226$ mean; 60 improved, 14 degraded, and 9 unchanged cases
Instance refinement	83 microscopy instance cases; ΔDSC-TP = +0.0001 mean; true-positive mask quality is preserved
PE fine-tuning	Local full validation, best decoder-only checkpoint; ΔDSC = +0.0004, $\Delta F1_{50} = +0.0011$; ΔDSC-TP = -0.0010

Table 6 provides controlled evidence for each highlighted component. Scalable full-volume inference is supported by complete output generation, runtime safety, and bounded GPU memory. Fragmentation-aware refinement is evaluated on microscopy instance cases, where fragmented false positives are the dominant failure mode. Parameter-efficient fine-tuning is evaluated by comparing the selected decoder-only checkpoint with the reproduced original checkpoint on the local full validation set.

Table 7 summarizes representative runtime and resource-oriented experiments. A representative subset of 12 validation cases was used for these efficiency ablations to cover both fast and slow cases while keeping the experiment manageable. Text-embedding caching produced a small runtime reduction without changing the segmentation semantics. SDPA and pinned transfer were tested but not adopted because they did not consistently reduce maximum runtime. Weight-only INT8 quantization reduced GPU memory but increased runtime, so it was not used in the final accuracy submission.

For instance segmentation, the fragmentation-aware refinement was mainly designed to improve output stability. On the 83 microscopy instance cases, it

Table 7. Representative ablation experiments for inference efficiency on validation cases. Best values in each metric column are shown in bold; lower maximum time is better.

Variant	Cases	DSC	$F1_{50}$	Max time
Baseline speed stack	12	0.8275	0.3543	44.62 s
With text cache	12	0.8290	0.3543	**44.04 s**
SDPA trial	12	**0.8292**	0.3472	45.28 s
Pinned transfer trial	12	0.8285	**0.3734**	44.84 s

Table 8. Effect of instance-oriented refinement on local full validation instance metrics. Best values in each metric column are shown in bold.

Variant	$F1_{50}$	DSC-TP
Base full-volume inference	0.4886	**0.7153**
With fragmentation-aware refinement	**0.4904**	0.7152
Microscopy-only mean delta	+0.0226	+0.0001

increased mean $F1_{50}$ by 0.0226, with 60 improved, 14 degraded, and 9 unchanged cases. On the local full validation set, $F1_{50}$ increased from 0.4886 to 0.4904, while DSC-TP changed from 0.7153 to 0.7152. This small DSC-TP reduction is an explicit trade-off: the refinement improves instance detection stability by suppressing fragmented false positives, while avoiding aggressive reshaping of true-positive masks.

4.4 Discussion

The method works well for large 3D volumes where stable full-volume processing is required. The adaptive inference design reduces the risk of memory spikes caused by many prompts or long volumes. It also avoids unnecessary recomputation of text embeddings and reduces I/O overhead during output saving. These changes are especially important in a challenge setting where failed or timed-out cases can severely affect the final score.

The fragmentation-aware refinement is most relevant for instance-rich modalities such as microscopy and PET, where isolated small components can create false-positive instances. By limiting refinement to uncertain regions and applying conservative component reassignment, the method aims to clean the output without removing dominant anatomical or lesion structures.

Potential failure cases remain. First, the method still depends on the recognition ability of the pretrained BiomedParseV2 model. If the textual target is poorly represented by the pretrained model or if the target has very low image contrast, the refinement stage cannot recover completely missing predictions. Second, the transfer refinement stage is intentionally lightweight; it is not a replacement for full task-specific training. Third, connected-component reassignment is conservative and may not fully resolve difficult instance separation when adjacent objects touch each other.

4.5 Results on Final Testing Set

The official final testing results for the text-guided segmentation track are summarized in Table 9. Our submission ranked first among the reported submissions. Compared with the BiomedParseV2 baseline, it improved the average DSC from 0.750 to 0.763, the average NSD from 0.775 to 0.787, the average $F1_{50}$ from 0.238 to 0.265, and the average DSC-TP from 0.448 to 0.488.

Table 9. Official final testing results for the text-guided segmentation track. The values are organizer-reported leaderboard results. Best values in each metric column are shown in bold.

Rank	Team	Avg. DSC	Avg. NSD	Avg. $F1_{50}$	Avg. DSC-TP
1	Ours	**0.763**	**0.787**	**0.265**	**0.488**
2	drsam3	0.698	0.726	0.194	0.368
Baseline	BiomedParseV2	0.750	0.775	0.238	0.448

4.6 Limitation and Future Work

The current system prioritizes reliability and practical deployment under constrained runtime and memory budgets. More aggressive methods, such as image test-time augmentation, multi-axis ensembling, full-model fine-tuning, or TensorRT-based compilation, may further improve accuracy or efficiency but were not used because they introduce additional runtime, memory, or engineering risks. Future work will explore more principled uncertainty modeling, stronger instance separation for densely packed microscopy objects, and efficient 3D-native text-guided architectures.

5 Conclusion

We presented a reliable text-guided 3D biomedical segmentation system with scalable volumetric inference and fragmentation-aware refinement. Starting from pretrained BiomedParseV2, we performed parameter-efficient transfer refinement and redesigned the inference pipeline for memory-aware full-volume prediction. We further introduced uncertainty-guided correction and conservative connected-component reassignment for more stable 3D instance masks. Local full validation results show competitive segmentation performance while maintaining complete output generation and runtime safety across large heterogeneous 3D biomedical cases. On the official final testing leaderboard, our submission achieved the best overall performance among all participating teams, consistently outperforming the BiomedParseV2 baseline across DSC, NSD, $F1_{50}$, and DSC-TP.

Code Availability. Our code is available at https://github.com/zhoujiang18/text-guided-3d-biomedseg-fragrefine.

Acknowledgements We thank the challenge organizers and data contributors for providing the benchmark and evaluation platform. We also thank CodaBench [18] for hosting the challenge.

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

References

1. Bilic, P., Christ, P., Li, H.B., Vorontsov, E., Ben-Cohen, A., Kaissis, G., Szeskin, A., Jacobs, C., Mamani, G.E.H., Chartrand, G., LohÃűfer, F., Holch, J.W., Sommer, W., Hofmann, F., Hostettler, A., Lev-Cohain, N., Drozdzal, M., Amitai, M.M., Vivanti, R., Sosna, J., Ezhov, I., Sekuboyina, A., Navarro, F., Kofler, F., Paetzold, J.C., Shit, S., Hu, X., LipkovÃą, J., Rempfler, M., Piraud, M., Kirschke, J., Wiestler, B., Zhang, Z., HÃijlsemeyer, C., Beetz, M., Ettlinger, F., Antonelli, M., Bae, W., Bellver, M., Bi, L., Chen, H., Chlebus, G., Dam, E.B., Dou, Q., Fu, C.W., Georgescu, B., i Nieto, X.G., Gruen, F., Han, X., Heng, P.A., Hesser, J., Moltz, J.H., Igel, C., Isensee, F., JÃďger, P., Jia, F., Kaluva, K.C., Khened, M., Kim, I., Kim, J.H., Kim, S., Kohl, S., Konopczynski, T., Kori, A., Krishnamurthi, G., Li, F., Li, H., Li, J., Li, X., Lowengrub, J., Ma, J., Maier-Hein, K., Maninis, K.K., Meine, H., Merhof, D., Pai, A., Perslev, M., Petersen, J., Pont-Tuset, J., Qi, J., Qi, X., Rippel, O., Roth, K., Sarasua, I., Schenk, A., Shen, Z., Torres, J., Wachinger, C., Wang, C., Weninger, L., Wu, J., Xu, D., Yang, X., Yu, S.C.H., Yuan, Y., Yue, M., Zhang, L., Cardoso, J., Bakas, S., Braren, R., Heinemann, V., Pal, C., Tang, A., Kadoury, S., Soler, L., van Ginneken, B., Greenspan, H., Joskowicz, L., Menze, B.: The liver tumor segmentation benchmark (lits). Medical Image Analysis **84**, 102680 (2023) 2
2. Du, Y., Bai, F., Huang, T., Zhao, B.: Segvol: Universal and interactive volumetric medical image segmentation. In: Advances in Neural Information Processing Systems. vol. 37, pp. 110746–110783 (2024) 2
3. Fabian, I., Maximilian, R., Lars, K., Stefan, D., Ashis, R., Florian, S., Benjamin, H., Tassilo, W., Moritz, L., Constantin, U., Jonathan, D., Ralf, F., Klaus, M.H.: nninteractive: Redefining 3D promptable segmentation. arXiv preprint arXiv:2503.08373 (2025) 2
4. Gatidis, S., Hepp, T., Früh, M., La Fougère, C., Nikolaou, K., Pfannenberg, C., Schölkopf, B., Küstner, T., Cyran, C., Rubin, D.: A whole-body fdg-pet/ct dataset with manually annotated tumor lesions. Scientific Data **9**(1), 601 (2022) 2, 6
5. He, Y., Guo, P., Tang, Y., Myronenko, A., Nath, V., Xu, Z., Yang, D., Zhao, C., Simon, B., Belue, M., Harmon, S., Turkbey, B., Xu, D., Li, W.: VISTA3D: A unified segmentation foundation model for 3D medical imaging. In: Proceedings of the IEEE/CVF International Conference on Computer Vision and Pattern Recognition (2024) 2
6. Heller, N., Isensee, F., Maier-Hein, K.H., Hou, X., Xie, C., Li, F., Nan, Y., Mu, G., Lin, Z., Han, M., Yao, G., Gao, Y., Zhang, Y., Wang, Y., Hou, F., Yang, J.,

Xiong, G., Tian, J., Zhong, C., Ma, J., Rickman, J., Dean, J., Stai, B., Tejpaul, R., Oestreich, M., Blake, P., Kaluzniak, H., Raza, S., Rosenberg, J., Moore, K., Walczak, E., Rengel, Z., Edgerton, Z., Vasdev, R., Peterson, M., McSweeney, S., Peterson, S., Kalapara, A., Sathianathen, N., Papanikolopoulos, N., Weight, C.: The state of the art in kidney and kidney tumor segmentation in contrast-enhanced ct imaging: Results of the kits19 challenge. Medical Image Analysis **67**, 101821 (2021) 2

7. Huang, Z., Jiang, Y., Zhang, R., Zhang, S., Zhang, X.: Cat: Coordinating anatomical-textual prompts for multi-organ and tumor segmentation. In: Globerson, A., Mackey, L., Belgrave, D., Fan, A., Paquet, U., Tomczak, J., Zhang, C. (eds.) Advances in Neural Information Processing Systems. vol. 37, pp. 3588–3610 (2024) 2
8. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollar, P., Girshick, R.: Segment anything. In: Proceedings of the International Conference on Computer Vision. pp. 4015–4026 (2023) 2
9. Ma, J., Chen, J., Ng, M., Huang, R., Li, Y., Li, C., Yang, X., Martel, A.L.: Loss odyssey in medical image segmentation. Medical Image Analysis **71**, 102035 (2021) 7
10. Ma, J., He, Y., Li, F., Han, L., You, C., Wang, B.: Segment anything in medical images. Nature Communications **15**, 654 (2024) 2, 6
11. Ma, J., Yang, Z., Kim, S., Chen, B., Baharoon, M., Fallahpour, A., Asakereh, R., Lyu, H., Wang, B.: Medsam2: Segment anything in 3d medical images and videos. arXiv preprint arXiv:2504.03600 (2025) 2, 6
12. Ma, J., Zhang, Y., Gu, S., An, X., Wang, Z., Ge, C., Wang, C., Zhang, F., Wang, Y., Xu, Y., Gou, S., Thaler, F., Payer, C., Ǎǎtern, D., Henderson, E.G., McSweeney, D.M., Green, A., Jackson, P., McIntosh, L., Nguyen, Q.C., Qayyum, A., Conze, P.H., Huang, Z., Zhou, Z., Fan, D.P., Xiong, H., Dong, G., Zhu, Q., He, J., Yang, X.: Fast and low-gpu-memory abdomen ct organ segmentation: The flare challenge. Medical Image Analysis **82**, 102616 (2022) 2
13. Maier-Hein, L., Reinke, A., Godau, P., Tizabi, M.D., Buettner, F., Christodoulou, E., Glocker, B., Isensee, F., Kleesiek, J., Kozubek, M., Reyes, M., Riegler, M.A., Wiesenfarth, M., et al.: Metrics reloaded: recommendations for image analysis validation. Nature Methods **21**(2), 195–212 (2024). https://doi.org/10.1038/s41592-023-02151-z 6
14. Ravi, N., Gabeur, V., Hu, Y.T., Hu, R., Ryali, C., Ma, T., Khedr, H., Rädle, R., Rolland, C., Gustafson, L., Mintun, E., Pan, J., Alwala, K.V., Carion, N., Wu, C.Y., Girshick, R., Dollár, P., Feichtenhofer, C.: Sam 2: Segment anything in images and videos. In: International Conference on Learning Representations (2025) 2
15. Simpson, A.L., Antonelli, M., Bakas, S., Bilello, M., Farahani, K., van Ginneken, B., Kopp-Schneider, A., Landman, B.A., Litjens, G., Menze, B., Ronneberger, O., Summers, R.M., Bilic, P., Christ, P.F., Do, R.K.G., Gollub, M., Golia-Pernicka, J., Heckers, S.H., Jarnagin, W.R., McHugo, M.K., Napel, S., Vorontsov, E., Maier-Hein, L., Cardoso, M.J.: A large annotated medical image dataset for the development and evaluation of segmentation algorithms. Nature Communications **10**, 4122 (2019) 2
16. Wang, H., Guo, S., Ye, J., Deng, Z., Cheng, J., Li, T., Chen, J., Su, Y., Huang, Z., Shen, Y., Fu, B., Zhang, S., He, J., Qiao, Y.: Sam-med3d: Towards general-

purpose segmentation models for volumetric medical images. arXiv preprint arXiv:2310.15161 (2024) 2

17. Wasserthal, J., Breit, H.C., Meyer, M.T., Pradella, M., Hinck, D., Sauter, A.W., Heye, T., Boll, D.T., Cyriac, J., Yang, S., Bach, M., Segeroth, M.: Totalsegmentator: Robust segmentation of 104 anatomic structures in ct images. Radiology: Artificial Intelligence **5**(5), e230024 (2023) 2, 6
18. Xu, Z., Escalera, S., PavÃčo, A., Richard, M., Tu, W.W., Yao, Q., Zhao, H., Guyon, I.: Codabench: Flexible, easy-to-use, and reproducible meta-benchmark platform. Patterns **3**(7), 100543 (2022) 14
19. Zhao, T., Gu, Y., Yang, J., Usuyama, N., Lee, H.H., Kiblawi, S., Naumann, T., Gao, J., Crabtree, A., Abel, J., et al.: A foundation model for joint segmentation, detection and recognition of biomedical objects across nine modalities. Nature Methods **22**, 166–176 (2025) 2
20. Zhao, Z., Zhang, Y., Wu, C., Zhang, X., Zhang, Y., Wang, Y., Xie, W.: One model to rule them all: Towards universal segmentation for medical images with text prompt. arXiv preprint arXiv:2312.17183 (2023) 2

Table 10. Checklist table for manuscript preparation. Delete this table in the camera-ready submission.

Requirements	Answer
A meaningful title	Yes
The number of authors ($\leq$6)	4
Author affiliations and ORCID	Yes
Corresponding author email is presented	Yes
Validation scores are presented in the abstract	Yes
Introduction includes at least three parts: background, related work, and motivation	Yes
A pipeline/network figure is provided	Figure 1
Pre-processing	Section 3.2
Strategies to data augmentation	Table 2, data augmentation row
Strategies to improve model inference	Sections 2.3 and 4.3
Post-processing	Section 2.4
Environment setting table is provided	Table 1
Training protocol table is provided	Table 2
Ablation study	Section 4.3
Efficiency evaluation results are provided	Table 4
Visualized segmentation example is provided	Figure 3
Limitation and future work are presented	Yes
Reference format is consistent	Yes
Main text $\geq$ 8 pages, excluding references and appendix	Yes