ECLIPSE: EHR-Constrained Localized Inference for Pan-Tumor Segmentation

Runqi Meng[1], Zonglin Liu[2, 6], Qiong Ma[2, 6], Lin Teng[1], Yulin Wang[1], Sifan Song[7], Pengfei Jin[5], Yiqun Sun[1, 2, 6, (✉)], and Dinggang Shen[1, 3, 4, (✉)]

[1]School of Biomedical Engineering & State Key Laboratory of Advanced Medical Materials and Devices, ShanghaiTech University
[2]Department of Radiology, Fudan University Shanghai Cancer Center
[3]Shanghai United Imaging Intelligence Co., Ltd.
[4]Shanghai Clinical Research and Trial Center
[5]Center of Advanced Medical Computing and Analysis, Massachusetts General Hospital and Harvard Medical School
[6]Department of Oncology, Shanghai Medical College, Fudan University
[7]Xi'an Jiaotong-Liverpool University
12211230032@fudan.edu.cn , Dinggang.Shen@gmail.com

Abstract. Accurate tumor segmentation is essential for tumor-centric clinical modeling, yet most existing methods remain organ-specific or rely primarily on imaging and generic semantic prompts, failing to incorporate patient-specific clinically recorded tumor burden into spatial inference. In routine practice, electronic health record (EHR) systematically record lesion count, longest diameter, and coarse anatomical subregion, providing structured but indirect spatial cues. Mapping structured clinical descriptors to voxel-level tumor delineation is challenging in pan-tumor settings, where cross-organ variability expands the spatial hypothesis space and sparse EHR descriptors alone cannot uniquely determine tumor location. To address this, we propose **ECLIPSE** (EHR-Constrained Localized Inference for Pan-Tumor Segmentation). Specifically, through a dedicated Clinically Constrained Hypothesis Generation (CCHG) (CCHG) module, ECLIPSE reformulates pan-tumor segmentation as a feasibility-constrained localization–refinement process. First, a tumor-type prompt–conditioned spatial response map is estimated within the decoder space of a pretrained model to generate dense tumor hypotheses. Second, CCHG injects structured EHR descriptors as soft spatial constraints to enforce lesion count consistency, size plausibility, and regional coherence, producing clinically consistent bounding box prompts. Finally, the inferred bounding boxes are encoded as spatial prompts to refine voxel-wise delineation. Across 14 public and one private tumor datasets comprising 5,360 cases for training, evaluation and testing, ECLIPSE achieves a mean DSC of 61.41%, outperforming the strongest baseline by +7.44%. It further improves performance on mixed tumor data (+8.91%), private clinical data (+3.88%), and attains the best DSC (71.23%) on an unseen dataset, demonstrating robust generalization.

Keywords: Pan-Tumor Segmentation · EHR-Guided Diagnosis

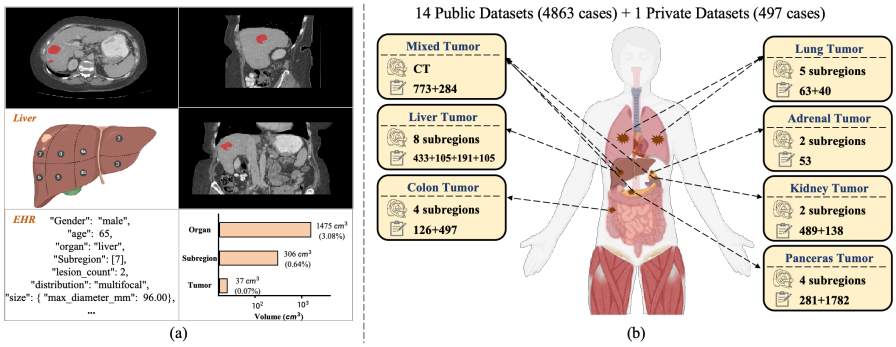

Fig. 1. (a) Representative liver case with tumors in Couinaud segment VII, as specified in the EHR. Due to the small tumor volume, subregion-level guidance provides tighter spatial constraints than organ-level guidance. (b) Overview of the multi-organ tumor datasets used in this study, showing annotated anatomical subregions and corresponding case numbers from public and private cohorts.

1 Introduction

Accurate tumor segmentation is essential for cancer staging, treatment planning, risk stratification and prognosis. However, in clinical practice, voxel-level tumor annotations are rarely available due to the substantial time and expertise required [15]. Instead, radiologists record key findings, such as lesion count, maximal diameter, and anatomical subregion in structured electronic health records (EHR), which implicitly constrains the feasible spatial configuration of tumors (Fig. 1(a)). However, most current segmentation frameworks [17,1,2,6] optimize voxel-level likelihood purely from imaging data, rarely incorporating these clinically grounded constraints. This separation creates a structural disconnection between clinical reasoning and spatial inference.

The disconnection becomes more pronounced in pan-tumor settings. While tumors across anatomical sites share structural regularities [7,8], they exhibit substantial heterogeneity in morphology, intensity patterns, anatomical context, and spatial dispersion. Without explicit constraints, this heterogeneity significantly expands the space of plausible tumor configurations, rendering localization increasingly ambiguous. Meanwhile, structured EHR descriptors encode global yet indirect information. The mapping from lesion count, size, and region-level descriptors to voxel-level masks is inherently underdetermined. To enable consistency with clinical records, a principled pan-tumor segmentation framework must therefore address two coupled challenges: (1) reducing the combinatorial hypothesis space induced by cross-organ variability, and (2) translating sparse, structured clinical descriptors into spatially actionable constraints for precise tumor localization during inference.

Existing approaches fall into two categories, neither of which resolves the above challenges. On the one hand, prompt-driven foundation-model paradigms

[4,5] mitigate appearance variability by conditioning segmentation on spatial prompts such as points or bounding boxes. However, these prompts are manually specified and effectively externalize constraint definition, limiting scalability in routine workflows and multifocal disease scenarios. Multimodal fusion strategies, on the other hand, integrate imaging and EHR information primarily at the feature level to enhance outcome prediction, treating clinical descriptors as auxiliary covariates rather than spatial feasibility constraints [9,10,11]. Even report-derived signals used in training [12] rarely guide inference, decoupling delineation from clinical tumor burden.

To address this problem, we propose **ECLIPSE** (**E**HR-**C**onstrained **L**ocalized **I**nference for **P**an-Tumor **Se**gmentation), a feasibility-constrained framework that integrates structured tumor burden descriptors into spatial inference via a Clinically Constrained Hypothesis Generation (CCHG) module. Rather than performing late-stage multimodal fusion or relying on externally provided prompts, ECLIPSE first generates a tumor-type prompt–conditioned spatial response map in the decoder space of a pretrained model, producing dense but coarse localization hypotheses. The proposed CCHG module then reformulates structured EHR descriptors—including lesion count, maximal diameter, and anatomical subregion—as soft spatial feasibility constraints to enforce count consistency, size plausibility, and regional coherence during hypothesis selection, thereby generating clinically consistent bounding box proposals. These bounding boxes are subsequently encoded as automatically generated prompts and fed back into the pretrained decoder for refinement. Through this decoupled **visual evidence modeling → CCHG-based constraint injection → prompt-based refinement** paradigm, ECLIPSE reformulates pan-tumor segmentation as feasibility-constrained spatial inference while remaining compatible with foundation models. To our knowledge, this is the first framework to elevate structured EHR tumor descriptors to feasibility constraints within pan-tumor segmentation.

2 Method

In this paper, we propose **ECLIPSE**, a unified framework for EHR-constrained pan-tumor segmentation. As shown in Fig. 2, ECLIPSE augments a pretrained encoder–decoder backbone with a Clinically Constrained Hypothesis Generation (CCHG) (CCHG) module, enabling tumor delineation guided by structured EHR tumor burden descriptors. Instead of treating EHR as auxiliary covariates, CCHG integrates subregion, lesion count, and size into spatial inference to constrain the feasible hypothesis space. Specifically, ECLIPSE proceeds in three steps. **First**, the backbone extracts visual embeddings and estimates a dense voxel-wise tumor likelihood map, providing coarse localization hypotheses (Sec. 2.1). **Second**, CCHG converts structured EHR descriptors into spatial feasibility constraints, selects tumor centers under count and size consistency, and generates clinically coherent bounding box prompts (Sec. 2.2). **Finally**, these prompts are fed back into the decoder to refine voxel-level segmentation (Sec. 2.3). By coupling image evidence with CCHG-based feasibility constraints,

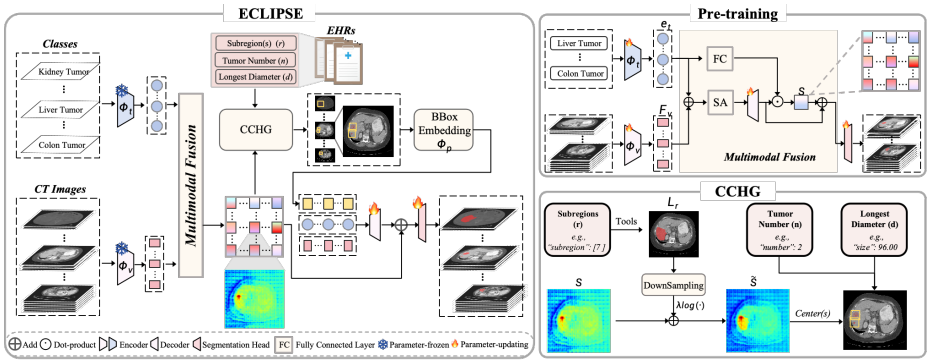

Fig. 2. Overview of ECLIPSE. ECLIPSE performs clinically constrained pan-tumor segmentation in three stages: (1) Text–image fusion produces a dense tumor response map; (2) Structured EHR descriptors (subregion, lesion count, diameter) impose feasibility constraints for hypothesis selection and bounding-box generation; (3) The inferred boxes are used as prompts for decoder-based refinement.

ECLIPSE produces segmentations consistent with both imaging and structured clinical descriptors.

2.1 Tumor-Type Prompt–Conditioned Spatial Hypothesis Generation

To initiate clinically constrained spatial inference, we construct a dense tumor-type prompt–conditioned spatial response map in the decoder feature space, serving as a proposal over plausible tumor locations.

Given an input CT volume $X \in \mathbb{R}^{1\times H\times W\times D}$, a 3D Vision Transformer (ViT) [28] encoder Φ_v extracts visual features

$$F_v = \Phi_v(X) \in \mathbb{R}^{C\times h\times w\times d}. \tag{1}$$

Let t denote the tumor-type prompt. We construct the text input using the template "*A computerized tomography of a* [t]", and encode it with the CLIP [29] text encoder Φ_t to obtain the semantic embedding

$$e_t = \Phi_t(t) \in \mathbb{R}^C. \tag{2}$$

The text embedding is broadcast to the spatial grid to match the shape of F_v and fused via element-wise addition, followed by two self-attention layers (SA):

$$F_f = \mathrm{SA}(F_v + \mathrm{Broadcast}(e_t)). \tag{3}$$

The fused representation F_f is mapped to the decoder to obtain an upscaled spatial feature map

$$U \in \mathbb{R}^{C_u\times H'\times W'\times D'}. \tag{4}$$

To compute voxel-wise tumor responses, the text embedding is linearly aligned to the decoder channel dimension, denoted as $\hat{e}_t \in \mathbb{R}^{C_u}$. The spatial response map is defined via dot-product similarity:

$$S(i,j,k) = \hat{e}_t^{\top} U_{i,j,k}, \tag{5}$$

where (i,j,k) indexes spatial locations on the decoder grid. This dense similarity field provides coarse spatial likelihood prior to clinical feasibility modeling. During pre-training, decoder features are fed to a segmentation head for voxel-wise supervision, aligning visual and textual representations.

2.2 Structured Clinical Feasibility Modeling

To enforce consistency with structured tumor burden descriptors, we introduce a CCHG module that incorporates EHR information as soft spatial feasibility constraints. Let the structured EHR descriptor be denoted as

$$\mathcal{E} = (r, n, d), \tag{6}$$

where r indicates a coarse anatomical subregion, n denotes lesion count, and d represents the longest tumor diameter.

To derive a spatial constraint corresponding to the subregion descriptor r, we first perform fine-grained subregion segmentation using established anatomical segmentation tools (*e.g.*, TotalSegmentator [15] for liver, lung, kidney, and adrenal). For colon and pancreas, we segment the whole organ and subsequently partition it into anatomical subregions using nnU-Net. The number of subregions defined for each organ is summarized in Fig. 1(b). The specified subregion(s) are then converted into a binary anatomical mask L_r defined on the original image grid and downsampled to the decoder resolution:

$$M_r = \text{Downsample}(L_r) \in [0,1]^{H' \times W' \times D'}. \tag{7}$$

Instead of imposing a hard spatial restriction, we introduce the anatomical constraint as a soft logit modulation of the response map:

$$\tilde{S}(i,j,k) = S(i,j,k) + \lambda \log\big(\epsilon + M_r(i,j,k)\big), \tag{8}$$

where λ controls the strength of the prior and ϵ ensures numerical stability.

Cardinality feasibility is enforced through diversity-aware peak selection on $\tilde{S}$ using radius-controlled 3D non-maximum suppression, from which the top-n spatially separated centers $\{c_i\}_{i=1}^{n}$ are retained. Scale feasibility is then imposed by expanding each center c_i into a loose bounding box based on EHR-reported tumor diameter d. Here, α denotes a safety-margin coefficient that controls the box expansion and accounts for uncertainty in the reported tumor size:

$$b_i = \text{Box}(c_i, d; \alpha). \tag{9}$$

2.3 Prompt-Based Segmentation Refinement

The clinically admissible bounding boxes $\{b_i\}_{i=1}^{n}$ are encoded as box features and used to refine voxel-wise delineation. Specifically, each bounding box b_i is converted into a box embedding

$$p_i = \Phi_p(b_i), \tag{10}$$

and aggregated with the tumor-type embedding e_t and visual features to form decoder inputs. Conditioned on $\{p_i\}_{i=1}^{n}$, the decoder produces a refinement feature map:

$$R = \mathcal{D}(F_v, e_t, \{p_i\}_{i=1}^{n}). \tag{11}$$

We then fuse this refinement feature with the response map via element-wise addition, followed by a segmentation head to obtain the final mask:

$$\hat{Y} = \mathcal{H}(R + S), \tag{12}$$

where $\mathcal{H}$ denotes the segmentation head, and S is the dot-product response map defined in Eq. (5).

3 Experiments

3.1 Dataset

To rigorously evaluate pan-tumor generalization, we constructed a multi-cohort benchmark comprising 14 public datasets and one private clinical dataset (Colon-FD), totaling 5,360 contrast-enhanced CT volumes spanning diverse tumor types and anatomical regions (Fig. 1(b)). AbdomenCT-1K [18], LiTS [19], KiTS [20], MSD-Liver [21], MSD-Lung [21], MSD-Pancreas [21], MSD-Colon [21], MSD-Hepatic Vessels [21], and PanTS-training [24] were used for training and validation with an 8:2 split. MSWAL [22], HCC [23], PanTS-testing [24], KAS [25], LUAD [26], and ACC [27] were reserved exclusively for testing under a unified protocol without cohort-specific fine-tuning. ACC (adrenal tumor) was treated as an unseen tumor type during testing to explicitly assess cross-tumor generalization. The private Colon-FD cohort contains 497 patients and was used solely for external clinical evaluation. In this cohort, tumor masks were independently delineated by experienced radiologists, while EHR descriptors were extracted from real clinical records, providing a real clinical setting for evaluating ECLIPSE.

3.2 Implementation Details

All experiments were implemented in PyTorch and conducted on an NVIDIA Tesla A100 GPU (80GB). The backbone was pretrained for 1000 epochs, followed by 300 epochs of end-to-end fine-tuning with the CCHG. Both stages were trained with the AdamW optimizer using a base learning rate of 5×10^{-5}, a batch size of 2, and an input size of $[32, 256, 256]$. We used a combined Dice and binary

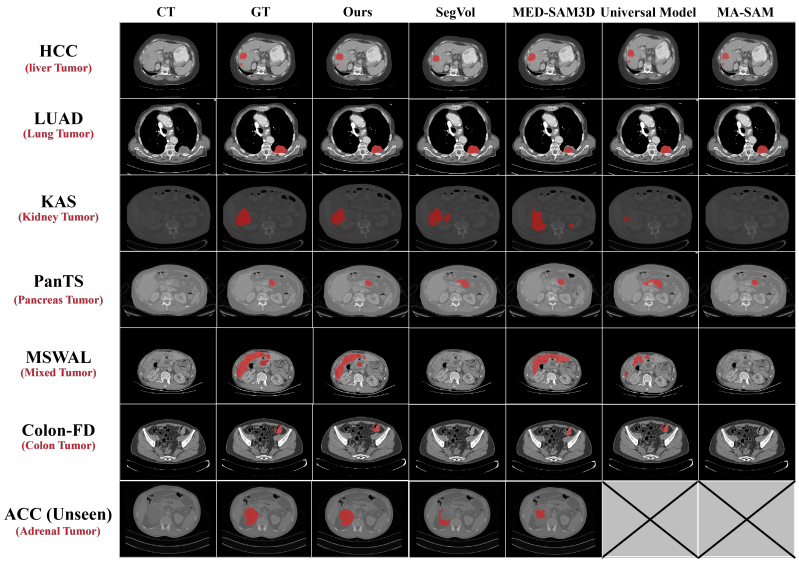

Fig. 3. Qualitative comparisons of the proposed ECLIPSE (Ours) with other methods for multi-tumor segmentation. Each row represents one type of tumor. Column 1 shows the original CT scans, column 2 presents the ground-truth segmentation, and columns 3-7 show the predictions from different SOTA methods.

cross-entropy (BCE) loss in both stages, with equal weighting between the two terms. All CT volumes were reoriented to RAS space. Intensities were clipped to $[-200, 300]$ Hounsfield units and linearly normalized. Data augmentation, including random 90° rotations and intensity perturbations, was applied during training to improve robustness. For public datasets without released EHR fields, pseudo-EHR descriptors were generated from tumor masks to simulate clinical tumor burden information. Lesion count was derived by connected-component analysis, maximal diameter by the largest physical lesion extent, and anatomical subregion by the overlaps with predefined organ subregions.

3.3 Comparison with State-of-the-Art Methods

To comprehensively evaluate the effectiveness of ECLIPSE, we conducted comparisons against representative state-of-the-art universal medical image segmentation approaches. Specifically, we included MA-SAM [13], Universal Model [14], Med-SAM3D (point) [4], and SegVol (point and text) [5]. All competing methods were evaluated following their original protocols to ensure a fair comparison.

Quantitative Comparison. Table 1 reports the DSC results across seven tumor datasets. Overall, ECLIPSE achieves the highest mean DSC of **61.41%**, outperforming the strongest baseline, SegVol (53.97%), by **+7.44%**. These results confirm the advantage of feasibility-constrained spatial inference for pan-tumor segmentation.

Table 1. Quantitative comparison with other SOTA methods across all datasets in terms of DSC (%). The best results are bolded.

Method	HCC	LUAD	KAS	PanTS	MSWAL	Colon-FD	ACC	Mean
MA-SAM [13]	44.62	22.81	21.83	17.79	15.76	24.61	-	24.57
Universal Model [14]	54.31	48.59	48.88	32.37	35.61	48.80	-	44.76
Med-SAM3D [4]	61.68	24.28	56.11	45.20	34.32	53.72	69.48	49.83
SegVol [5]	58.02	66.06	55.77	44.30	40.78	54.86	57.99	53.97
ECLIPSE (ours)	**66.57**	**68.06**	**60.78**	**54.83**	**49.69**	**58.74**	**71.23**	**61.41**

Performance gains are consistent across datasets with varying characteristics. On MSWAL, which contains mixed tumor data from multiple anatomical sites, ECLIPSE improves DSC by **+8.91%**, indicating robustness to heterogeneous data distributions. On Colon-FD, a private clinical dataset, our method achieves a **+3.88%** improvement, suggesting effective generalization under real clinical practice. On the unseen ACC dataset, ECLIPSE achieves the best DSC of **71.23%**, surpassing all baselines without any task-specific adaptation. This result demonstrates that integrating structured EHR constraints into spatial inference enables robust generalization beyond seen tumor distributions, rather than relying on dataset-specific appearance cues.

Qualitative Comparison. Fig. 3 presents representative cases across diverse tumor types. Compared with competing methods, ECLIPSE produces more compact and anatomically confined segmentations, reducing fragmented responses and suppressing false positives outside plausible tumor regions. Across liver, kidney, pancreas, colon, and lung tumors, baseline models either miss small lesions or over-segment adjacent structures, whereas ECLIPSE preserves clearer lesion boundaries and more consistent spatial extent.

On the mixed tumor case (MSWAL), where lesions from multiple anatomical sites are jointly evaluated, competing methods often generate diffuse or disconnected predictions, while ECLIPSE maintains lesion-level specificity and avoids cross-organ spillover. On the unseen ACC dataset, despite no exposure during training, ECLIPSE delineates the tumor with stable localization and coherent shape, whereas baselines exhibit unstable boundaries or partial omission, indicating stronger cross-distribution robustness.

3.4 Ablation Study

Table 2 reports the effect of introducing CCHG into the baseline framework. Adding CCHG improves the mean DSC from 42.57% to 61.41% (+18.84%), demonstrating its critical role in constraining spatial inference.

Performance gains are consistent across all datasets. On MSWAL (mixed tumor data), DSC increases from 38.49% to 49.69%, indicating improved robustness under heterogeneous distributions. On Colon-FD (private clinical data), performance rises from 46.15% to 58.74%, suggesting better domain generalization. Most notably, on the unseen ACC dataset, DSC improves dramatically from

Table 2. Ablation study of CCHG across all datasets in terms of DSC (%).

Method	HCC	LUAD	KAS	PanTS	MSWAL	Colon-FD	ACC	Mean
Baseline	57.49	49.45	46.19	37.59	38.49	46.15	22.65	42.57
Baseline + CCHG	**66.57**	**68.06**	**60.78**	**54.83**	**49.69**	**58.74**	**71.23**	**61.41**

22.65% to 71.23%, highlighting that CCHG effectively reduces cross-distribution ambiguity and stabilizes localization under zero-shot conditions.

4 Conclusion

We have presented ECLIPSE for pan-tumor segmentation by integrating structured clinical feasibility modeling into prompt-based segmentation. Specifically, by introducing CCHG, our method reshapes unconstrained visual likelihoods into clinically admissible bounding box hypotheses using anatomical, cardinality, and scale descriptors. Extensive experiments across multiple tumor types demonstrate consistent improvement over state-of-the-art universal segmentation models. Ablation results confirm that structured feasibility modeling substantially enhances robustness and generalization. These findings highlight the importance of incorporating clinical priors into hypothesis generation, providing a principled direction for clinically aligned medical image segmentation. Nevertheless, this study has limitations. ECLIPSE currently relies on structured tumor burden descriptors and anatomical priors, and its robustness to incomplete or inaccurate EHR information remains to be further evaluated. In addition, because structured clinical records are unavailable in most public datasets, pseudo-EHR descriptors are used to simulate clinical tumor burden information. Future work will validate the model in larger real-world cohorts.

Acknowledgments. This work was supported in part by National Natural Science Foundation of China (82441023, U23A20295, 62131015, 82394432), the China Ministry of Science and Technology (STI2030-Major Projects-2022ZD0209 000, STI2030-Major Projects-2022ZD0213100), The Key R&D Program of Guangdong Province, China (2023B0303040001), and HPC Platform of ShanghaiTech University.

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

References

1. Wang, H., et al.: Dual-reference source-free active domain adaptation for nasopharyngeal carcinoma tumor segmentation across multiple hospitals. IEEE Transactions on Medical Imaging (2024)

2. Liu, H., et al.: Multimodal brain tumor segmentation boosted by monomodal normal brain images. IEEE Transactions on Image Processing **33**, 1199–1210 (2024)
3. Kirillov, A., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023)
4. Wang, H., et al.: SAM-Med3D. arXiv preprint arXiv:2310.15161 (2023)
5. Du, Y., et al.: SegVol: Universal and interactive volumetric medical image segmentation. In: Advances in Neural Information Processing Systems **37**, 110746–110783 (2025)
6. Meng, R., et al.: A neighbor-sensitive multi-modal flexible learning framework for improved prostate tumor segmentation in anisotropic MR images. IEEE Transactions on Biomedical Engineering (2025)
7. Meng, R., et al.: MAST-Pro: Dynamic mixture-of-experts for adaptive segmentation of pan-tumors with knowledge-driven prompts. In: MICCAI 2025. Springer, Cham (2025)
8. Shah, A., Rojas, C.A.: Imaging modalities, indications, differential diagnosis and imaging characteristics of cystic mediastinal masses: A review. Mediastinum (2022)
9. Zhang, Y., et al.: Towards cardiac MRI foundation models: Comprehensive visual-tabular representations for whole-heart assessment and beyond. Medical Image Analysis, 103756 (2025)
10. Duenias, D., et al.: Hyperfusion: A hypernetwork approach to multimodal integration of tabular and medical imaging data for predictive modeling. Medical Image Analysis **102**, 103503 (2025)
11. Zeng, B., et al.: C2HFusion: Clinical context-driven hierarchical fusion of multimodal data for personalized prognostic assessment in pancreatic cancer. Medical Image Analysis, 103937 (2026)
12. Bassi, P.R.A.S., et al.: Learning segmentation from radiology reports. In: MICCAI 2025. Springer, Cham (2025)
13. Chen, C., et al.: MA-SAM: Modality-agnostic SAM adaptation for 3D medical image segmentation. Medical Image Analysis **98**, 103310 (2024)
14. Liu, J., et al.: CLIP-driven universal model for organ segmentation and tumor detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), pp. 21152–21164 (2023)
15. Wasserthal, J., et al.: TotalSegmentator: Robust segmentation of 104 anatomic structures in CT images. Radiology: Artificial Intelligence **5**(5), e230024 (2023)
16. Ma, J., et al.: Automatic organ and pan-cancer segmentation in abdomen CT: The FLARE 2023 challenge. arXiv preprint arXiv:2408.12534 (2024)
17. Meng, R., et al.: NaMa: Neighbor-aware multi-modal adaptive learning for prostate tumor segmentation on anisotropic MR images. In: Proceedings of the AAAI Conference on Artificial Intelligence **38**(5) (2024)
18. Ma, J., et al.: AbdomenCT-1K: Is abdominal organ segmentation a solved problem? IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) (2021)
19. Bilic, P., et al.: The liver tumor segmentation benchmark (LiTS). Medical Image Analysis **84**, 102680 (2023)
20. Heller, N., et al.: An international challenge to use artificial intelligence to define the state-of-the-art in kidney and kidney tumor segmentation in CT imaging. Medical Image Analysis (2020)
21. Antonelli, M., et al.: The medical segmentation decathlon. Nature Communications **13**(1), 4128 (2022)
22. Wu, Z., et al.: MSWAL: 3D multi-class segmentation of whole abdominal lesions dataset. In: MICCAI 2025
23. Moawad, A.W., et al.: Multimodality annotated HCC cases with and without advanced imaging segmentation. The Cancer Imaging Archive (2021)

24. Li, W., et al.: PanTS: The pancreatic tumor segmentation dataset. In: Advances in Neural Information Processing Systems (NeurIPS) (2025)
25. Humpire-Mamani, G.E., et al.: Dataset for kidney abnormality segmentation in thorax-abdomen CT scans. Zenodo (2023)
26. Goldgof, D., et al.: Long and short survival in adenocarcinoma lung CTs. The Cancer Imaging Archive (2017)
27. Moawad, A.W., et al.: Voxel-level segmentation of pathologically proven adrenocortical carcinoma with Ki-67 expression. The Cancer Imaging Archive (2023)
28. Hatamizadeh, A., Tang, Y., Nath, V., Yang, D., Myronenko, A., Landman, B., Roth, H., Xu, D.: UNETR: Transformers for 3D medical image segmentation. arXiv preprint arXiv:2103.10504 (2021)
29. Radford, A., et al.: Learning transferable visual models from natural language supervision. arXiv preprint arXiv:2103.00020 (2021)