TotalFM: An Organ-Separated 3D-CT Foundation Model Leveraging Large-Scale Routine Clinical Radiology Data

Kohei Yamamoto[1†⋆][0009−0008−6043−4069] and Tomohiro Kikuchi[1†][0000−0002−4222−4569]

Department of Radiology, Jichi Medical University, Tochigi, Japan
yamamoto.kohei@jichi.ac.jp, r1419kt@jichi.ac.jp

Abstract. Developing robust 3D-CT foundation models (FMs) using large-scale, real-world clinical data is hindered by three barriers: unstructured report curation, the resolution-batch size trade-off, and vision-language correspondence ambiguity. We propose TotalFM, an organ-separated framework that automatically converts raw clinical CT images and reports into FM-optimized, organ-level volume–finding pairs. By decomposing whole CT volumes into anatomically meaningful units, this approach enables high-resolution 3D encoding with large batch sizes for efficient contrastive learning. The proposed framework was trained on 286,667 CT series and 401,501 organ-level volume-finding pairs. In zero-shot tasks, it consistently outperformed state-of-the-art models in organ-wise anomaly detection and finding-wise AUROC across the majority of disease categories. These results demonstrate that organ-separated learning is a scalable, clinically interpretable principle for bridging the gap between raw clinical archives and 3D-CT FMs. The source code and pretrained models are publicly available at https://github.com/jichi-labo/TotalFM.

Keywords: Foundation Model · Computed Tomography · Contrastive Learning · Organ-separated Learning · Real-world Data.

1 Introduction

In diagnostic imaging AI, foundation models (FMs) complement annotation-limited supervised learning by enabling the acquisition of generalizable representations from large-scale clinical data [1, 3, 4, 7, 18, 25, 26, 29]. Achieving external validity in real-world settings requires that such representations reflect the diversity and complexity of routine practice. Nevertheless, in 3D-CT, efficient and reliable utilization of raw clinical data for foundation-model training remains hindered by three fundamental barriers [15, 21].

– **Curating clinical data for foundation-model training.** Routine clinical CT data are not directly suitable for foundation-model training: reports are

† These authors contributed equally to this work. ⋆ Corresponding author.

unstructured and mix findings across organs, and a single examination often includes multiple series with different scan ranges and contrast phases [9, 19]. Thus, simply collecting data does not guarantee reliable report–image correspondence.
- **The trade-off between resolution and batch size.** The high dimensionality of 3D-CT creates a conflict between the spatial resolution needed to capture subtle lesions and the batch size required for efficient contrastive learning [5, 8]. High-resolution 3D inputs drastically limit batch size due to GPU memory constraints, leading to stalled learning and unstable convergence.
- **Input granularity and representation reliability.** Contrastive representation learning based on CLIP has been reported to suffer from several limitations, including its inability to faithfully preserve geometric structures in images and its difficulty in capturing fine-grained contextual information in long-form text [11, 28]. In the context of 3D-CT foundation models, most prior studies adopt a paradigm in which the entire CT volume and/or the full radiology report are provided as input [1, 3, 4, 7, 25]. In medical applications, controlling the granularity at which embedding representations are constructed is not merely a design choice but a critical requirement, as it directly affects diagnostic performance, model reliability, and interpretability.

To address these three challenges simultaneously, we propose TotalFM, an organ-separated framework for building 3D-CT FMs. Our pipeline converts non-curated CT images and reports into organ-level volume–finding pairs (VFPs), enabling high-resolution 3D encoding with practical batch sizes and more reliable, fine-grained representations.

2 Methods

2.1 Dataset

Real-world axial CT series and Japanese radiology reports were obtained from the Japan Medical Image Database (J-MID) [2] for the year 2024. For self-supervised pretraining, we included 119,275 studies (8 institutions) acquired between January 1 and May 15. For model development (training/validation) and internal testing, we used 69,032 studies (8 institutions) acquired between October 1 and December 31. Among these institutions, one was held out for internal testing. The remaining institutions were split by acquisition date into Training (October 1–December 22) and Validation (December 23–December 31). For external validation, we used the test set of the Merlin Abdominal CT Dataset.

2.2 Dataset Construction Pipeline

We built a fully automated pipeline that combines an LLM and organ segmentation to generate foundation-model–optimized, organ-level VFPs from raw CT images and radiology reports (Figure 1).

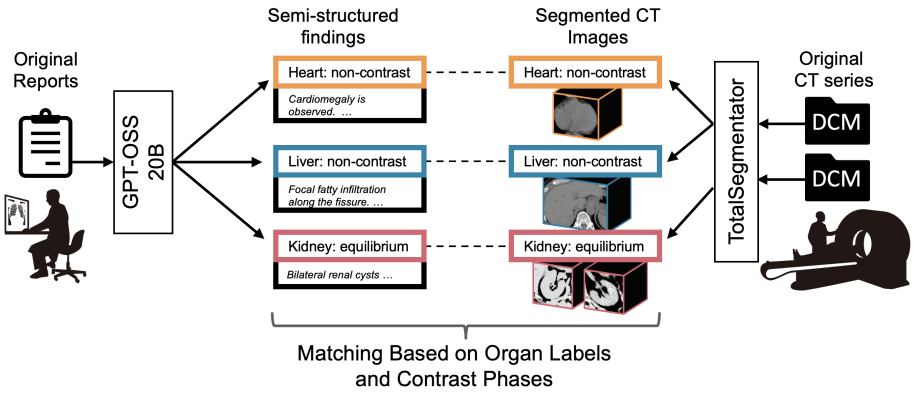

Fig. 1. Overview of the dataset construction pipeline.

- **Text parsing:** Raw Japanese radiology reports were first segmented into finding-level sentences using a BERT-based classifier [13]. The segmented findings were then input into GPT-OSS 20B to extract positive findings, identify the mentioned organ(s), and infer the optimal contrast phase for assessing each finding from four categories: non-contrast, arterial, portal, and equilibrium or later phases. We provide the prompts in our GitHub repository.
- **Image parsing:** We applied TotalSegmentator to CT series to obtain organ segmentation labels and contrast phase information, categorizing each series into the same four groups as defined in the text parsing step [24].
- **Pairing:** We integrated the above outputs to generate VFPs that link a specific organ volume to its corresponding finding sentence. Under radiologist supervision, we developed a matching algorithm to select the optimal image series when the preferred contrast phase for a specific finding was unavailable.
- **Negative augmentation:** To strengthen the learning of normal anatomy, we paired organs not explicitly mentioned in the report with rule-based normal sentences, which were treated as negative samples (defined as "Rule-based negative").

2.3 Model Architecture and Training

Our model architecture and training scheme were informed by InternVideo2 [23]. Following prior volumetric CT studies, we treat the third dimension as an ordered spatial axis rather than a temporal axis and adopt VideoMAE [20] solely as a framework for 3D token modeling and masked reconstruction. Training consisted of two stages: (i) self-supervised pretraining with VideoMAE for 10 epochs (approximately 6 days), and (ii) image–text contrastive learning using organ-level VFPs for 20 epochs (approximately 1 day).

For each organ, the input volume was cropped based on the corresponding segmentation mask and resized to $192 \times 192 \times 32$, which was empirically

determined to encompass the majority of target organs. To accommodate different windowing conditions, we applied three-channel windowing with window level/width settings for lung (-600, 1500), soft tissue (40, 400), and bone (300, 1500), each normalized to 0–1.

- **Image encoder:** a 3D ViT-based model (ViT-B scale; ≈131M parameters) that processes organ-specific volumes of $192 \times 192 \times 32$ with three-channel windowing [6].
- **Text encoder:** Japanese ModernBERT (`sbintuitions/modernbert-ja-130m`) projecting findings into a 512-dimensional shared embedding space [22].

This pipeline enabled a batch size of 32 samples per GPU. Table 1 summarizes key settings of representative existing FMs, including input resolution and batch size.

Table 1. Comparison of input resolution, computational cost, and training configurations among related works.

Model	Resolution	GFLOPs	GPU	GPUs	Batch	Per GPU
CT-CLIP [7]	$480 \times 480 \times 240$	484	A100	4	8	2
RadFM [25]	$256 \times 256 \times 64$	81	A100	32	32	1
Merlin [4]	$224 \times 224 \times 160$	667	A6000	1	18	18
Pillar-0 [1]	$384 \times 384 \times 384$	14,926	H100	8	64	8
TotalFM	$\mathbf{192 \times 192 \times 32}$	180	H100	2	**64**	**32**

2.4 Evaluation

Pipeline validity check. To assess the validity of the VFPs produced by our pipeline, we randomly sampled 500 pairs and asked a board-certified radiologist to evaluate the triplets consisting of the finding text, the predicted organ label, and the cropped organ volume.

Downstream evaluation. We evaluated the learned FM on two zero-shot downstream tasks. In both tasks, TotalFM received organ-cropped CT volumes and Japanese text inputs, whereas CT-CLIP and Merlin were evaluated using full CT volumes and English prompts translated from the original Japanese sentences with an LLM (GPT-OSS 20B [17]).

Organ-wise lesion classification using actual reports. A binary task was defined for each organ using balanced positive and negative samples (1:1). We compared the cosine similarity between the organ-cropped CT embedding and two prompts: the original finding and a rule-based negative template (e.g., "No significant abnormality is observed in the {organ_name}."). For negative samples,

a positive finding sentence from the same organ was randomly selected from the evaluation set. Cosine similarity was computed between the image embedding and each text embedding, and the label was assigned based on the higher similarity. Performance was measured using the F1-score for each organ.

Finding-wise lesion classification. Similarity between the image embedding and a rule-based positive prompt ("{abnormal_finding} is present.") served as the confidence score. When a single disease label corresponded to multiple organ labels (e.g., kidney), similarity was computed for each organ and averaged to obtain the final confidence score. AUROC was used to evaluate performance.

3 Results

3.1 Data Used and Preprocessing

Table 2 summarizes the scale of the data used in our pipeline construction and validation. We used 286,667 CT series for pretraining. For model training and internal testing, we generated 401,501 organ-level VFPs from 166,978 series contained in 69,032 CT examinations with associated radiology reports. In the pipeline validity check, invalid triplets were rare: 6/208 (2.9%) for positive findings, 2/154 (1.3%) for negative findings, and 2/138 (1.4%) for rule-based negative pairs.

In addition, we adopted the dataset of 5,125 cases used as the internal test set in the Merlin study as our external dataset.

Table 2. Dataset statistics for each split.

Dataset	Pretrain	Train	Valid	Test
Dates (2024)	Jan 1–May 15	Oct 1–Dec 22	Dec 23–Dec 31	Oct 1–Dec 31
Institutions	8	7	7	1
Studies	119,275	57,457	5,964	5,611
Series	286,667	141,554	14,347	11,077
Imaging regions (No. of series)				
Head	–	10,915	1,115	1,281
Neck	–	1,686	202	57
Chest	–	62,547	5,727	3,346
Abdomen/Pelvis	–	18,121	1,994	1,729
Torso	–	45,245	4,754	4,305
Whole body	–	3,040	329	359
Volume–finding pairs (Original sentence)	–	224,117	12,415	40,710
Volume–finding pairs (Rule–based negative)	–	107,035	5,503	11,721

3.2 Organ-wise lesion classification

Table 3 reports zero-shot lesion classification performance by organ. Averaged over commonly supported organs, TotalFM improved F1 over CT-CLIP (0.708 vs. 0.515; +37.5%) and Merlin (0.692 vs. 0.650; +6.5%). TotalFM outperformed CT-CLIP in all organs except the aorta and surpassed Merlin in 64% (9/14) of organs, with F1 > 0.6 for all evaluated labels.

Table 3. Zero-shot organ-wise paired image–text lesion classification (F1).

Organ	CT-CLIP	Merlin	TotalFM (ours)
brain	–	–	.618
thyroid gland	.354	–	.731
trachea	.478	–	.709
esophagus	.556	.662	.761
lung	.475	.379	.612
aorta	.650	.685	.611
heart	.574	.694	.824
liver	–	.788	.739
gallbladder	–	.705	.694
stomach	–	.590	.666
spleen	–	.624	.704
kidney	–	.447	.708
pancreas	–	.666	.678
small bowel	–	.692	.749
colon	–	.767	.652
urinary bladder	–	.590	.618
prostate	–	.817	.675

"–" indicates organs not included in a model's training data.

3.3 Finding-wise lesion classification

Figure 2 compares AUROC across finding categories on J-MID. Our model outperformed Merlin in 83% (25/30) of categories, while performance was comparable or slightly lower for elongated organs along the z-axis (e.g., aorta, lungs) or organs split into multiple TotalSegmentator labels. We also benchmarked on the Merlin Test dataset, evaluating 17 localizable categories. Merlin tended to perform better on aorta and lung-field findings, whereas performance was comparable for other findings.

4 Discussion

The pipeline developed in this study enables the transformation of "raw information" from real-world clinical reports into organ-level representations suitable

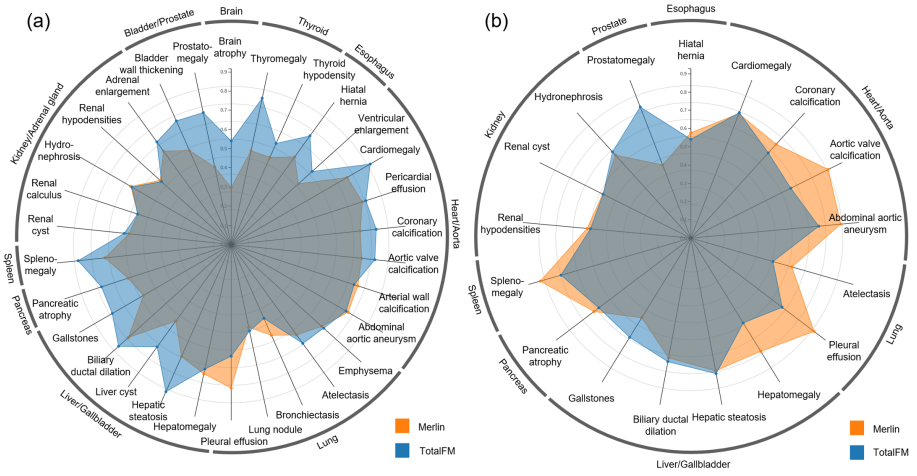

Fig. 2. Comparison of AUROC performance by finding category. (a) AUROC comparison for 30 finding categories defined in the J-MID dataset. (b) AUROC comparison for 17 finding categories selected from the Merlin Test dataset.

for FM training. This achievement is built upon the foundations of existing research in organ segmentation and open-source large language models; specifically, we utilized TotalSegmentator and GPT-OSS to achieve efficient dataset construction by maximizing existing resources [17, 24]. Furthermore, our TotalFM, trained on the constructed organ-level VFPs, demonstrated high performance across two zero-shot tasks, validating the feasibility of the training paradigm from clinical data to FMs. Specifically, in zero-shot organ-wise anomaly detection (F1-score), it outperformed CT-CLIP in 83% (5/6) and Merlin in 64% (9/14) of organs. In disease/finding-based AUROC evaluation, our model surpassed Merlin in 83% (25/30) of the finding categories. While its performance was lower than Merlin's in 10 out of 17 items on the Merlin test set, it still showed relatively competitive performance, considering that this test set is an internal dataset for Merlin. This performance gap may reflect domain shifts arising not only from language translation, but also from differences in reporting styles, patient populations, scanners, and acquisition protocols between the J-MID and Merlin datasets.

4.1 Organ-Separated Learning as a Design Principle for 3D-CT Foundation Models

A central challenge in developing 3D-CT FMs is the trade-off between spatial resolution and training efficiency. High-resolution volumetric encoding is computationally expensive and typically forces smaller batch sizes, which can weaken contrastive learning [5, 8].

By decomposing CT volumes into anatomically meaningful organ units, our framework preserves within-organ resolution while enabling large-batch contrastive learning. This reduces computation on background regions and improves GPU-memory utilization, allowing for larger batch sizes at higher resolutions by reallocating computational resources from the entire CT volume to specific organ-level volumes.

Although not directly evaluated here, whole-volume encoding may limit discrimination of sub-regional structures and laterality (e.g., left versus right): compressing an entire CT volume into a single representation can dilute fine-grained spatial cues [11]. In contrast, organ-separated learning constrains representations to localized anatomical units, potentially improving intra-organ discrimination and laterality awareness. Whether this advantage yields measurable gains in fine-grained anatomical reasoning warrants further study.

4.2 Clinical Implications and Translational Potential

The proposed framework has substantial clinical implications. By adopting organ-level encoders, TotalFM enables organ-specific abnormality detection and provides a flexible foundation for practical applications, including Computer-Aided Diagnosis and retrieval-based tasks. The learned image–text embedding space supports bidirectional retrieval—searching radiology reports from CT images or identifying relevant image regions from textual queries [27, 29]—potentially facilitating clinical decision support, case search, and educational use.

Furthermore, we propose a translational pipeline that automatically converts routine radiology department data into organ-level structured representations suitable for AI training. Unlike many existing studies that rely on carefully curated and manually annotated datasets [4, 7], our approach leverages real-world clinical reports and segmentation outputs to construct large-scale training data in a fully automated manner. This perspective represents an initial step toward overcoming the limitations of curated datasets and moving closer to continuous learning frameworks grounded in routine clinical practice. Importantly, the proposed pipeline may also serve as an infrastructure for institution-specific iterative training, allowing FMs to be continuously adapted to local imaging protocols, reporting styles, and patient populations.

4.3 Limitations

Several limitations exist in this study. First, findings that do not correspond to the anatomical labels provided by TotalSegmentator were excluded. Specifically, findings related to the vascular system (excluding the aorta) and the lymphatic system were largely omitted from the training data. To address this, future dataset construction requires methods to localize regions within the CT volume that correspond to finding sentences, potentially utilizing techniques such as Open-Vocabulary Segmentation or Visual Grounding to learn richer local representations [10, 30]. Second, while most previous studies perform contrastive learning between whole reports and whole volumes, our approach adopts a finer

granularity by pairing specific finding sentences with organ volumes. A drawback of this approach is that the InfoNCE loss treats all non-paired elements in a mini-batch as negatives [16]. This may introduce noise, as certain non-paired combinations within a batch might actually represent positive findings, potentially leading to over-separation of representations or unstable optimization [12, 14].

5 Conclusion

We presented TotalFM, a 3D-CT FM trained using a fully automated pipeline that converts multi-institutional, real-world CT images and radiology reports into learning-ready organ-level data. By adopting an organ-separated training strategy, TotalFM enables larger effective batch sizes and more efficient learning, and achieves competitive or superior performance to state-of-the-art models on two zero-shot downstream tasks.

Acknowledgments. This research was partially supported by JSPS KAKENHI [Grant Number JP23K17234]. We would like to thank the departments of radiology that provided the J-MID database, including Juntendo University, Kyushu University, Keio University, The University of Tokyo, Okayama University, Kyoto University, Osaka University, Tokyo Medical and Dental University, Hokkaido University, Ehime University, and Tokushima University.

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

References

1. Agrawal, K.K., Liu, L., Lian, L., Nercessian, M., Harguindeguy, N., Wu, Y., Mikhael, P., Lin, G., Sequist, L.V., Fintelmann, F., et al.: Pillar-0: A new frontier for radiology foundation models. arXiv [cs.CV] (Nov 2025)
2. Akashi, T., Kumamaru, K.K., Wada, A., Hashimoto, M., Hirata, K., Hayakawa, Y., Sano, K., Kamagata, K., Hagiwara, A., Ikenouchi, Y., Aoki, S.: Japan-medical image database (J-MID): Medical big data supporting data science. Juntendo Med. J. **71**(3), 166–172 (Jun 2025)
3. Bai, F., Du, Y., Huang, T., Meng, M.Q.H., Zhao, B.: M3D: Advancing 3D medical image analysis with multi-modal large language models. arXiv [cs.CV] (Mar 2024)
4. Blankemeier, L., Cohen, J.P., Kumar, A., Van Veen, D., Gardezi, S.J.S., Paschali, M., Chen, Z., Delbrouck, J.B., Reis, E., Truyts, C., et al.: Merlin: A vision language foundation model for 3D computed tomography. arXiv [cs.CV] (Jun 2024)
5. Chen, T., Kornblith, S., Norouzi, M., Hinton, G.: A simple framework for contrastive learning of visual representations. In: International Conference on Machine Learning. pp. 1597–1607. PMLR (Nov 2020)
6. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (ICLR) (2021)
7. Hamamci, I.E., Er, S., Wang, C., Almas, F., Simsek, A.G., Esirgun, S.N., Doga, I., Durugol, O.F., Dai, W., Xu, M., et al.: Developing generalist foundation models from a multimodal dataset for 3D computed tomography. arXiv [cs.CV] (Mar 2024)
8. He, K., Fan, H., Wu, Y., Xie, S., Girshick, R.: Momentum contrast for unsupervised visual representation learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (June 2020)
9. Huang, W., Li, C., Zhou, H.Y., Yang, H., Liu, J., Liang, Y., Zheng, H., Zhang, S., Wang, S.: Enhancing representation in radiography-reports foundation model: a granular alignment algorithm using masked contrastive learning. Nat. Commun. **15**(1), 7620 (Sep 2024)
10. Ichinose, A., Hatsutani, T., Nakamura, K., Kitamura, Y., Iizuka, S., Simo-Serra, E., Kido, S., Tomiyama, N.: Visual grounding of whole radiology reports for 3D CT images. In: Lecture Notes in Computer Science, pp. 611–621. Lecture Notes in Computer Science, Springer Nature Switzerland, Cham (2023)
11. Kang, R., Song, Y., Gkioxari, G., Perona, P.: Is clip ideal? no. can we fix it? yes! (2025), https://arxiv.org/abs/2503.08723
12. Khosla, P., Teterwak, P., Wang, C., Sarna, A., Tian, Y., Isola, P., Maschinot, A., Liu, C., Krishnan, D.: Supervised contrastive learning. arXiv [cs.LG] (Apr 2020)
13. Kikuchi, T., Yamagishi, Y., Yamamoto, K., Akashi, T., Mori, H., Makimoto, H., Kohro, T.: Context-aware sentence classification of radiology reports using synthetic data: Development and validation study. J Med Internet Res **28**, e86365 (Apr 2026). https://doi.org/10.2196/86365, https://www.jmir.org/2026/1/e86365
14. Kim, M., Shim, S.W., Lee, B.J.: FALCON: False-negative aware learning of COntrastive negatives in vision-language alignment. arXiv [cs.CV] (Nov 2025)
15. Li, J., Li, D., Savarese, S., Hoi, S.C.H.: BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. ICML **2023**, 19730–19742 (Jan 2023)
16. van den Oord, A., Li, Y., Vinyals, O.: Representation learning with contrastive predictive coding. arXiv [cs.LG] (Jul 2018)

17. OpenAI: gpt-oss-120b & gpt-oss-20b model card (2025), https://arxiv.org/abs/2508.10925
18. Pai, S., Hadzic, I., Bontempi, D., Bressem, K., Kann, B.H., Fedorov, A., Mak, R.H., Aerts, H.J.W.L.: Vision foundation models for computed tomography. arXiv [eess.IV] (Jan 2025)
19. Paschali, M., Chen, Z., Blankemeier, L., Varma, M., Youssef, A., Bluethgen, C., Langlotz, C., Gatidis, S., Chaudhari, A.: Foundation models in radiology: What, how, why, and why not. Radiology **314**(2), e240597 (Feb 2025)
20. Tong, Z., Song, Y., Wang, J., Wang, L.: VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. arXiv [cs.CV] pp. 10078–10093 (Mar 2022)
21. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: LLaMA: Open and efficient foundation language models. arXiv [cs.CL] (Feb 2023)
22. Tsukagoshi, H., Li, S., Fukuchi, A., Shibata, T.: ModernBERT-Ja (2025), https://huggingface.co/collections/sbintuitions/modernbert-ja-67b68fe891132877cf67aa0a
23. Wang, Y., Li, K., Li, X., Yu, J., He, Y., Wang, C., Chen, G., Pei, B., Yan, Z., Zheng, R., et al.: InternVideo2: Scaling foundation models for multimodal video understanding. arXiv [cs.CV] (Mar 2024)
24. 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. Radiol. Artif. Intell. **5**(5), e230024 (Sep 2023)
25. Wu, C., Zhang, X., Zhang, Y., Wang, Y., Xie, W.: Towards generalist foundation model for radiology by leveraging web-scale 2D&3D medical data. arXiv [cs.CV] (Aug 2023)
26. Xu, H., Usuyama, N., Bagga, J., Zhang, S., Rao, R., Naumann, T., Wong, C., Gero, Z., González, J., Gu, Y., et al.: A whole-slide foundation model for digital pathology from real-world data. Nature **630**(8015), 181–188 (Jun 2024)
27. Yamamoto, K., Kikuchi, T.: Feasibility study of CLIP-based key slice selection in CT images and performance enhancement via lesion- and organ-aware fine-tuning. Bioengineering (Basel) **12**(10), 1093 (Oct 2025)
28. Zhang, B., Zhang, P., Dong, X., Zang, Y., Wang, J.: Long-clip: Unlocking the long-text capability of clip (2024), https://arxiv.org/abs/2403.15378
29. Zhang, S., Xu, Y., Usuyama, N., Xu, H., Bagga, J., Tinn, R., Preston, S., Rao, R., Wei, M., Valluri, N., et al.: BiomedCLIP: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs. arXiv [cs.CV] (Mar 2023)
30. Zhao, Z., Zhang, Y., Wu, C., Zhang, X., Zhou, X., Zhang, Y., Wang, Y., Xie, W.: Large-vocabulary segmentation for medical images with text prompts. NPJ Digit. Med. **8**(1), 566 (Sep 2025)