OmniMedSeg: A Large-Scale Standardized Benchmark for Interactive Medical Image Segmentation

Zdravko Marinov[1] Alexander Jaus[1] Simon Reiß[1] Tobias Schlumberger[1]
Jiale Wei[1] Di Wen[1] Junwei Zheng[1] Jens Kleesiek[2]
Rainer Stiefelhagen[1]
[1]Karlsruhe Institute of Technology, Karlsruhe, Germany
{firstname.lastname}@kit.edu
[2]Institute for AI in Medicine, University Hospital Essen, Essen, Germany
{firstname.lastname}@uk-essen.de

Abstract

Medical image segmentation datasets form the foundation for developing deep learning models across diverse clinical tasks. As the field grows, the number of publicly annotated datasets has increased substantially. However, due to the sensitivity of medical data, many datasets are restricted by licenses, limited to specific challenges, or require complex access procedures. Moreover, medical imaging data is highly heterogeneous, existing in various formats that are often incompatible and difficult to use directly for model training or evaluation. In this paper, we address these issues by presenting *OmniMedSeg* - a large-scale, standardized benchmark for medical image segmentation that unifies 156 openly licensed datasets spanning nine imaging modalities. We introduce a modular three-layer framework (Download, Conversion, and Data layers) that automates data retrieval, standardizes formats to PNG for 2D and NIfTI for 3D data, and preserves all original metadata with full traceability. Beyond data aggregation, we provide standardized simulation and evaluation protocols for interactive segmentation, including click, scribble, bounding box, and polygon-based robot users. These protocols transform any existing metric into an interactive version via AUC and Final metric computation. *OmniMedSeg* is designed as a living, community-driven framework with an extensible architecture that supports continuous integration of new datasets, modalities, and protocols. This work establishes a robust foundation for fair, reproducible, and comparable benchmarking of both interactive and non-interactive segmentation methods. The converted dataset is available at: https://dx.doi.org/10.35097/qkc5m4kzran3g1yc

1 Introduction

Medical image segmentation remains one of the central areas in the medical image analysis community, with the availability of publicly annotated datasets enabling models to be trained on ever larger and more diverse data [1]. The rising popularity of segmentation challenges has further contributed to the number of datasets available [2]. However, due to the sensitivity of medical data, many datasets are restricted by licenses, limited to specific challenges, or require complex access procedures, such as signing data-use agreements or creating user accounts for tracking purposes. Moreover, medical imaging data is highly heterogeneous, existing in various formats that are often incompatible and difficult to use directly for model training or evaluation.

In this paper, we address these issues by unifying existing open-licensed datasets into *OmniMedSeg* - a large-scale, standardized benchmark for medical image segmentation. Section 2 discusses the motivation behind such a dataset and highlights the ongoing lack of *directly accessible* data resources for the community. Section 3 introduces *OmniMedSeg*, describing how we collected and standardized public datasets, and extended them with interactive segmentation simulation and evaluation protocols to enable offline benchmarking.

The contributions in this paper are based on Dr. Zdravko Marinov's PhD thesis, Chapter 8, which can be found here: https://publikationen.bibliothek.kit.edu/1000190952.

2 On the Accessibility of Medical Segmentation Datasets

Public medical segmentation datasets have become increasingly common as the field continues to grow. However, when researchers begin developing their own methods for a specific task and explore existing data resources, they often discover that *public* does not necessarily mean *directly accessible* and *open*.

In many cases, medical datasets are only available during the active period of a challenge and become inaccessible once the challenge ends, effectively closing the data permanently [3]. Even when data remains technically public, access is often restricted: some datasets require users to sign user agreements and wait several days for approval [4, 5, 6]. Other platforms are less restrictive and only require account registration [7, 8], but this still poses limitations, as users must manually complete these steps in a browser instead of accessing the data programmatically.

While such datasets can technically be classified as "public", they cannot be considered *directly accessible*. We consider directly accessible datasets as those hosted on platforms that provide a permanent and stable download link, allowing users to automatically retrieve the data through an API call or standard transfer protocols such as HTTP or FTP, without requiring additional authorization steps.

Moreover, even publicly available or directly accessible datasets often come with restrictive licenses that limit users' freedom to experiment or redistribute the data. For our purposes, we consider a dataset as *open-licensed* if its license permits redistribution, as this is essential for creating and sharing standardized versions of datasets. Conversely, datasets that prohibit redistribution, restrict use to specific purposes, or impose similar constraints are not considered open-licensed.

For *OmniMedSeg*, we include only datasets that are both *directly accessible* and *open-licensed*. This enables us to construct a framework capable of automatically generating standardized datasets using only the original download links, while also allowing us to redistribute these standardized resources freely within the research community.

3 OmniMedSeg: A Large-Scale Standardized Medical Image Segmentation Dataset

To establish the *OmniMedSeg* dataset, we follow a structured three-step pipeline: (1) comprehensive data collection; (2) conversion of all data into a standardized format; and (3) implementation of interactive segmentation simulation and evaluation protocols.

3.1 Data Collection

We adopted a systematic search strategy similar to that employed in previous reviews. However, instead of identifying studies on deep interactive segmentation, our focus here was on locating publicly available datasets for medical image segmentation.

We queried several prominent open-data repositories, including Zenodo, Mendeley Data, Figshare, and others (summarized in Table 1), using targeted keyword combinations: [segmentation], [medical], and [image], in conjunction with modality-specific terms such as [fundus], [ultrasound], [MRI], [CT], [X-Ray], [dermoscopy], [endoscopy], [microscopy], and [OCT]. Our systematic search resulted in 314 segmentation datasets. Then, we manually reviewed the search results and included only datasets with: (1) directly accessible download links; (2) no prohibited redistribution; (3) no restricted use. This reduced the datasets to the final number of 156 datasets included in *OmniMedSeg*.

Dataset Platform	Link
Figshare	https://figshare.com/
Mendeley Data	https://data.mendeley.com/
Zenodo	https://zenodo.org/
Grand Challenge	https://grand-challenge.org/
The Cancer Imaging Archive	https://www.cancerimagingarchive.net/
Kaggle	https://www.kaggle.com/

Table 1: List of all dataset platforms used in our systematic search for public datasets.

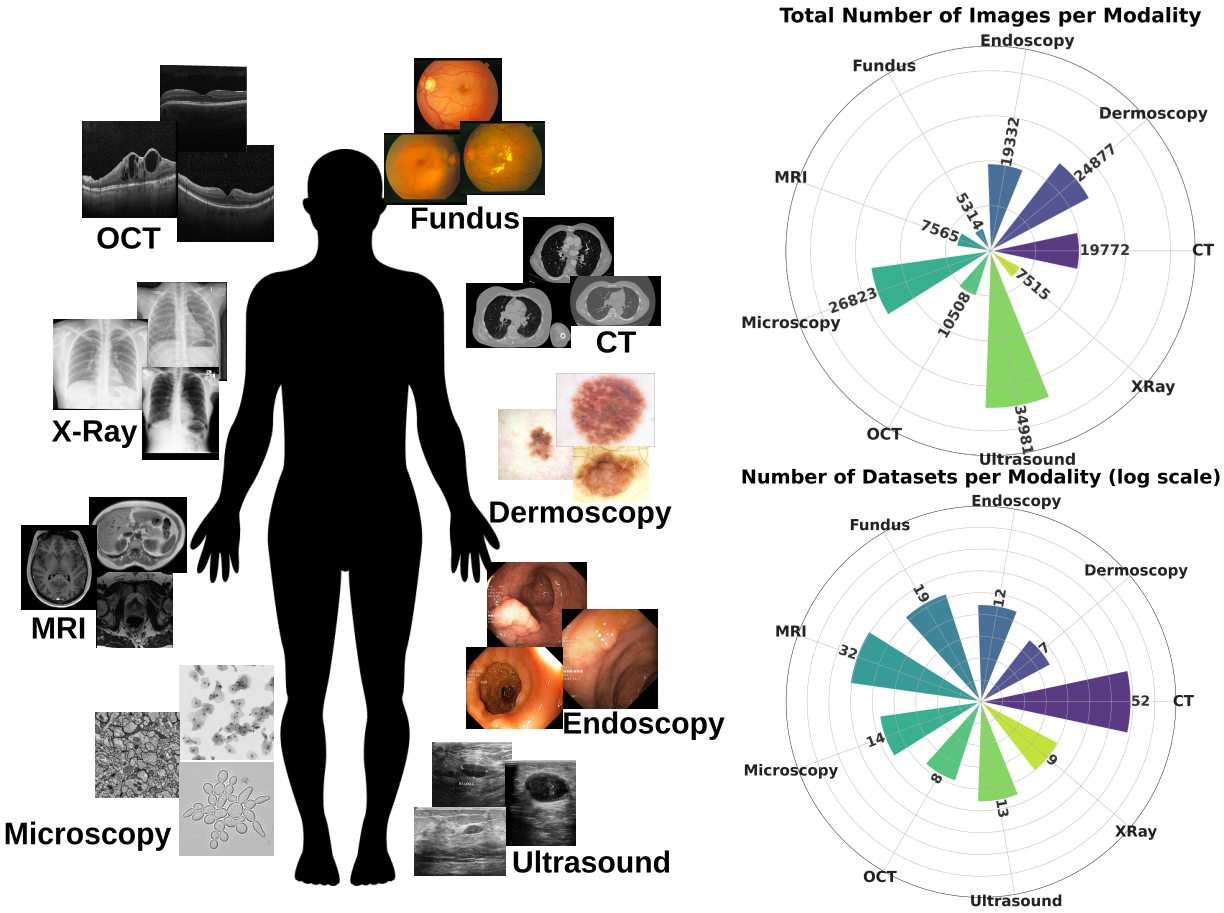

Figure 1: Overview of the *OmniMedSeg* dataset. We collected and standardized 156 open-licensed, directly accessible datasets from 9 imaging modalities.

The distribution of the datasets among the nine imaging modalities can be seen in Figure 1. The CT and MRI domains dominate in terms of dataset availability, with 52 and 32 datasets, respectively, reflecting their widespread use and application. In contrast, modalities such as Dermoscopy and OCT remain underrepresented, each contributing fewer than ten datasets. When considering image volume, ultrasound and microscopy collectively account for the largest number of images (34,981 and 26,823). However, it is important to note that the number of CT and MRI images refers to volumes, each consisting of hundreds of slices.

3.2 Conversion to a Standardized Format

3.2.1 On the Diversity of Medical Imaging Data

Medical imaging data exhibits substantial diversity across multiple dimensions, including image formats, directory structures, and label representations. This heterogeneity poses significant challenges for standardization and automation in data processing pipelines.

First, the variety of **image formats** is extensive. Common 2D formats include PNG and JPG, while legacy formats such as MAT also appear in certain datasets. For volumetric 3D data, formats such as NIfTI, DICOM, MHA, MHD+RAW, and H5 are prevalent. Second, **label formats** mirror this diversity. In 2D datasets, labels may be stored as binary masks or as text files describing contour coordinates. In 3D, labels are often encoded as binary NIfTI volumes, DICOM-SEG files, or RTSTRUCT objects. Third, the **dataset structure** itself can vary dramatically. Such inconsistencies necessitate custom preprocessing for each dataset, often requiring significant engineering effort before training can begin. Figure 2 illustrates representative examples of this diversity.

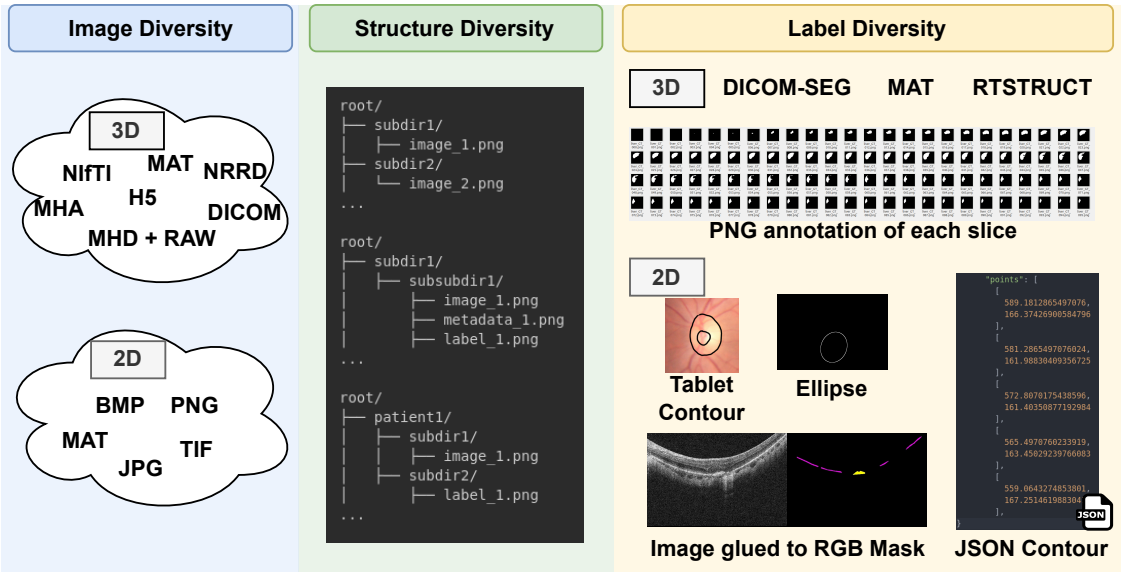

Figure 2: Examples illustrating the diversity of medical imaging datasets across image formats, directory structures, and label representations.

During our exploration of publicly available datasets, we encountered several unconventional labeling practices. For example, some CT scan labels are provided as per-slice PNG masks. Other datasets use tablet annotations overlaid on RGB images, requiring machine learning–based extraction just to obtain usable binary masks. Additional cases include ellipses that must be filled to form binary labels, JSON contour files that require interpolation and rasterization, and RGB masks concatenated next to the image, where each color corresponds to a different class.

To ensure consistent interaction simulation and evaluation across datasets, all data must first conform to a standardized format. We eliminate the heterogeneity of image and label formats and introduce a simple file structure to achieve this, as shown in Figure 3.

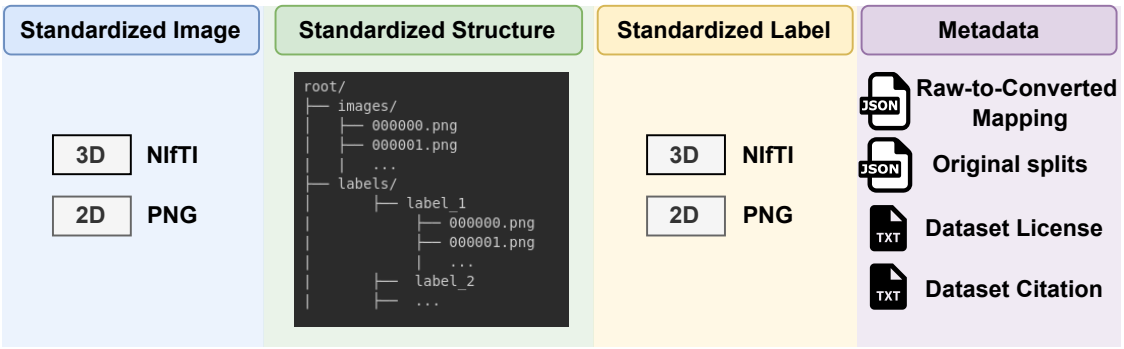

Figure 3: We use the NIfTI and PNG formats to store images and labels for all datasets in *OmniMedSeg*, and store a binary label for each class. We provide metadata: a JSON file mapping each converted image and label to the original file(s); a JSON file indicating original data splits; and the original license and citation.

Previous efforts [9, 10, 11] standardized data through heavy normalization and filtering, simplifying usage but discarding valuable details such as metadata and labels with small structures. In contrast, our work standardizes datasets while preserving the original data and labels, providing both a unified input format and full access to the original files to ensure flexibility and access to all available clinical metadata. We also provide a mapping of the converted images and labels to the original data to ensure traceability.

We designed *OmniMedSeg* with a modular architecture that allows the community to integrate additional datasets in the future without altering existing ones. The framework is organized into three core layers: (1) the **Download Layer**, (2) the **Conversion Layer**, and (3) the **Data Layer**, illustrated in Figure 4.

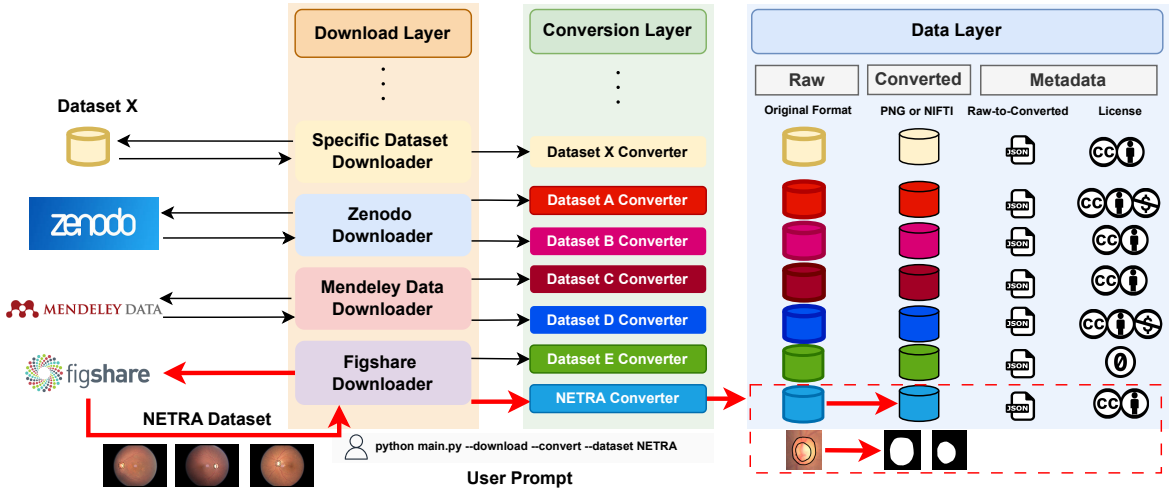

Figure 4: Overview of the three layers behind our *OmniMedSeg* dataset. Here, the user requests the download and conversion of the NETRA dataset [12] (red arrows).

3.2.2 Download Layer

The download layer automates the retrieval of each of the 156 collected datasets through API calls or standard HTTP/FTP requests. We assign a dedicated *downloader* to each dataset platform (e.g., Figshare or Zenodo), which can be reused for all datasets hosted on that platform. For datasets stored on institutional or lab-specific servers, we implement custom downloaders. Each downloader operates independently, ensuring modularity and enabling seamless integration of new datasets.

3.2.3 Conversion Layer

The conversion layer assigns a unique *converter* to each dataset to handle its specific format and structure. While converters are dataset-specific, common operations, such as DICOM-to-NIfTI conversion, are shared across implementations to avoid redundant code. Each converter transforms the dataset from its *raw format* into a *standardized format*: PNG for 2D images and NIfTI for 3D volumes. To extend the framework to new datasets, users simply need to implement one additional *converter* function.

3.2.4 Data Layer

The data layer manages both the converted and optional raw data. For each dataset, we provide a comprehensive JSON file mapping every converted image and label to its corresponding raw source files. For example, a single NIfTI file may be derived from multiple DICOM files, all of which are linked to the NIfTI filename in the JSON structure. This traceability allows users to reference original metadata, verify data integrity, or extract additional information. Furthermore, we include the original license and citation information for every dataset.

3.3 Interactive Segmentation Evaluation and Simulation Protocols

Given the standardized structure of *OmniMedSeg*, we can define interaction simulation and evaluation protocols in a dataset-agnostic manner. This abstraction allows the same protocol to be applied consistently across all 156 included datasets, as well as to future datasets added to *OmniMedSeg*.

3.3.1 Standardized Evaluation Protocols

For interactive evaluation, any non-interactive metric can be transformed into an interactive one by analyzing its evolution across interaction steps, specifically by computing the *Area Under the Curve (AUC)* and the *Final* metric value, which together characterize the interaction curve. This approach has been validated in multiple independent challenges to rank participants' approaches.

In *OmniMedSeg*, we formalize this approach by implementing the *AUC* and *Final* metrics as lightweight wrappers around an abstract `Metric` class. Given a sequence of metric values collected over successive interaction steps, the AUC is computed via trapezoidal numerical integration, while the Final metric simply corresponds to the last recorded value. Algorithm 1 demonstrates this modular design.

Algorithm 1 Interactive Metric Computation in *OmniMedSeg*

```
procedure InteractiveMetric
    abstract class Metric():
        def compute(preds, gts):
            # User-defined metric implementation here
            return metric_values

    class InteractiveMetric():
        def compute_interactive(metric_values):
            AUC ← trapezoidal_integration(metric_values)
            Final ← metric_values[-1]
            return {AUC, Final}

    Usage Example:
    user_metric = Metric()
    metric_values = user_metric.compute(predictions, ground_truths)
    results = InteractiveMetric().compute_interactive(metric_values)
    print(results)
end procedure
```

3.3.2 Standardized Simulation Protocols

At a fundamental level, a robot user can be understood as a function that selects interaction points based on an **eligibility mask** $\mathcal{M}$, a binary mask indicating valid sampling regions. The eligibility mask can represent ground-truth labels $\mathcal{L}$, model errors $\mathcal{L} \neq Y$ (where Y is the model prediction), or other logical combinations derived from the image I, label $\mathcal{L}$, and prediction Y. Formally, $\mathcal{M}$ can be expressed as the output of a function $f(I, \mathcal{L}, Y)$, which encodes task-specific eligibility criteria.

Within this abstraction, we define standardized protocols for four interaction types: clicks, scribbles, bounding boxes, and polygon vertices, each with multiple variations inspired by robot users from prior work.

Clicks. Given an eligibility mask $\mathcal{M}$, we simulate clicks following six annotation styles. The simulation proceeds in two stages: (1) selecting either the largest or a random connected component in $\mathcal{M}$, and (2) placing the click either at the center, boundary, or a random location within the component. Center and boundary clicks are generated by computing the Euclidean Distance Transform (EDT) of $\mathcal{M}$ and selecting from either its maximum or minimum points. Representative examples are shown in Figure 5.

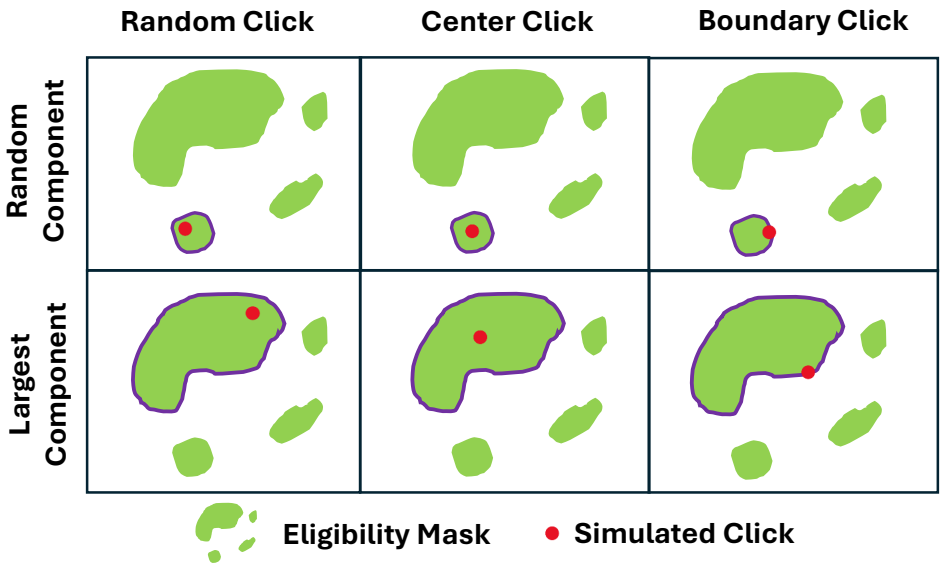

Figure 5: Standardized click-based robot users in *OmniMedSeg*.

Scribbles. We implement six types of analogous robot users for scribble-based interactions, following the established ScribblePrompt methodology [13]. For centerline scribbles, we extract the skeleton of $\mathcal{M}$, apply a random mask to fragment the structure, and introduce a deformation field. Boundary scribbles are generated by Gaussian-smoothing $\mathcal{M}$, thresholding to isolate edge regions, and applying the same random masking and deformation process. Random scribbles are produced by sampling random points within $\mathcal{M}$, connecting them with lines, and applying the same fragmentation and deformation steps. Illustrative examples are shown in Figure 6.

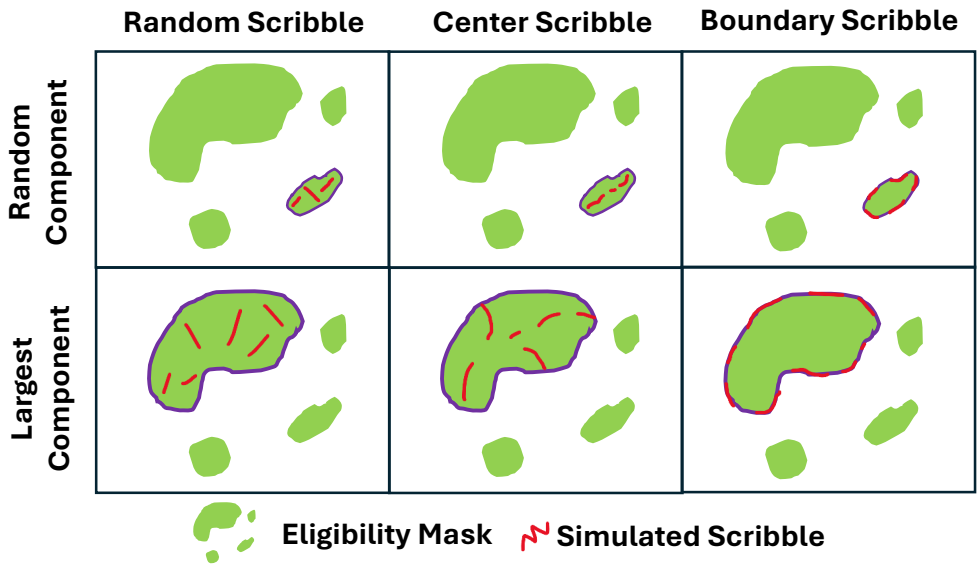

Figure 6: Standardized scribble-based robot users in *OmniMedSeg*.

Bounding Boxes. We provide three robot user variants for bounding boxes. The *perfect user* always draws an exact ground-truth bounding box. The *careful user* expands each dimension uniformly by 5–10%, simulating conservative over-annotation. Finally, the *sloppy user* introduces asymmetric perturbations of 5–10% in random directions along each dimension.

Polygon Vertices. Polygon-based interactions are implemented only for 2D data. Given a user-specified sampling budget N, N vertices are sampled uniformly over the boundary of the eligibility mask $\mathcal{M}$.

All robot users are integrated within the *OmniMedSeg* benchmark and can operate in both iterative and non-iterative settings, covering all areas in the interactive segmentation taxonomy [14]. We provide reference implementations and example scripts with a dummy model to demonstrate evaluation using any robot user.

4 Data Quality Assurance

To ensure the reliability and usability of *OmniMedSeg*, we conducted a rigorous manual quality assurance (QA) process to verify the integrity of our standardized conversion pipeline. Given the scale and heterogeneity of the 156 datasets, automated validation alone is insufficient to detect subtle errors such as misalignment between image and label, incorrect orientation, missing labels, or systematic artifacts introduced during format conversion.

For each dataset in *OmniMedSeg*, we randomly sampled at least 20 images. Additionally, to ensure adequate coverage across semantic classes, we sampled at least 5 images per segmentation class within each dataset. This stratified sampling strategy guarantees that both rare and frequent classes are represented in the QA process.

All sampled image-label pairs were then manually inspected by a panel of three independent reviewers. Each reviewer has substantial experience in computer vision (Ph.D. level or postdoctoral researcher) and has contributed to prior work in segmentation or medical image analysis. Furthermore, for every sampled image, at least two of the three reviewers possess a strong background specifically in medical image analysis, ensuring domain-specific scrutiny of clinical plausibility and anatomical correctness.

During inspection, reviewers evaluated the following criteria:

- **Alignment:** The segmentation mask must exactly overlay the corresponding anatomical structures in

the image without translational, rotational, or scaling offsets.

- **Orientation:** Volumetric data (NIfTI) must preserve the original radiological orientation (e.g., RAS) without unintended axis flips or transpositions.
- **Label Completeness:** No missing labels or empty masks for annotated structures present in the original dataset.
- **Systematic Errors:** Absence of artifacts introduced by the conversion process, such as interpolation artifacts, color space shifts, or loss of bit depth in 2D PNGs.
- **Structural Integrity:** For 3D volumes, contiguous slices must form a coherent volume without missing slices or corrupted headers.

Any dataset failing any of these criteria for more than 5% of sampled cases triggered a full review of that dataset's conversion pipeline, followed by re-conversion and re-inspection. In practice, the conversion process achieved a success rate of 98.2% on the first pass, with the remaining 1.8% requiring targeted fixes (primarily related to handling of non-standard DICOM metadata or multi-label RGB masks).

This multi-stage manual QA protocol ensures that *OmniMedSeg* maintains high data integrity across all 156 datasets, providing users with confidence that the standardized data faithfully represents the original annotations without corruption or systematic bias.

5 Experiments and Results

Due to space constraints, we present here representative results demonstrating the standardization and benchmarking capabilities of *OmniMedSeg*.

5.1 Standardization Validation

We validated our conversion pipeline on all 156 datasets. For each dataset, we verified that:

- All images are converted to PNG (2D) or NIfTI (3D) format
- All labels are stored as binary masks with consistent naming
- JSON metadata files correctly map converted files to original sources
- Original licenses and citations are preserved

The conversion preserved all original information, including small structures and metadata that previous standardization efforts discarded [9, 10].

6 Conclusion

We present *OmniMedSeg* - a large-scale, multimodal dataset that unifies 156 openly licensed datasets spanning nine imaging modalities. *OmniMedSeg* establishes a standardized foundation for training and evaluating both non-interactive and interactive segmentation methods, promoting accessibility, reproducibility, and fair comparison within the medical image segmentation community. The framework is designed to be transparent and extensible, built upon generic, dataset-agnostic concepts that can be uniformly applied across all included datasets and easily expanded with new contributions. Its open-source nature encourages community-driven development, enabling researchers to extend, refine, and maintain the resource collaboratively. Beyond data aggregation, *OmniMedSeg* provides standardized simulation and evaluation protocols for interactive segmentation, ensuring consistent benchmarking conditions.

We summarize the scientific impact of this work in three key contributions:

Contribution 1: *OmniMedSeg* - a comprehensive multimodal dataset. It introduces a standardized, large-scale resource that unifies 156 publicly available datasets across nine imaging modalities, enabling consistent data access and integration across the medical imaging domain.

Contribution 2: A standardized interactive evaluation and simulation framework built on top of *OmniMedSeg*, ensuring fair, reproducible, and comparable benchmarking conditions for both existing and future interactive segmentation methods.

Contribution 3: An extensible and community-driven design. *OmniMedSeg* is designed as a living framework rather than a fixed solution. Its modular and extensible architecture supports continuous integration of new datasets, modalities, and protocols.

OmniMedSeg represents the first unified, large-scale effort to standardize medical image segmentation data, simulation, and evaluation under a single open framework. By standardizing heterogeneous datasets and formalizing reproducible interactive evaluation and simulation protocols, it provides a robust foundation for developing and benchmarking both interactive and non-interactive segmentation methods. More importantly, its open and modular design ensures that *OmniMedSeg* will continue to evolve alongside the field, fostering collaboration, transparency, and sustained progress in medical image analysis research.

Acknowledgments

We thank all dataset authors who made their data openly available under permissive licenses. We are also grateful to RADAR4KIT for providing long-term storage infrastructure for the converted datasets, enabling the community to benefit from this resource.

References

[1] Chongyu Qu et al. AbdomenAtlas-8K: Annotating 8,000 CT volumes for multi-organ segmentation in three weeks. *Advances in Neural Information Processing Systems*, pages 36620–36636, 2023.

[2] Michela Antonelli et al. The medical segmentation decathlon. *Nature Communications*, 13(1), 2022, Art. no. 4128.

[3] Gongning Luo et al. Tumor detection, segmentation and classification challenge on automated 3d breast ultrasound: The tdsc-abus challenge. *arXiv:2501.15588*, 2025.

[4] Junde Wu et al. Gamma challenge: glaucoma grading from multi-modality images. *Medical Image Analysis*, 90:102938, 2023.

[5] Joes Staal, Michael D Abràmoff, Meindert Niemeijer, Max A Viergever, and Bram Van Ginneken. Ridge-based vessel segmentation in color images of the retina. *IEEE Transactions on Medical Imaging*, 23(4):501–509, 2004.

[6] Martina Melinščak, M Radmilovič, Zoran Vatavuk, and Sven Lončarić. Aroi: Annotated retinal oct images database. *International Convention on Information, Communication and Electronic Technology*, pages 371–376, 2021.

[7] Prasanna Porwal et al. Indian diabetic retinopathy image dataset (idrid): a database for diabetic retinopathy screening research. *Data*, 3(3):25, 2018.

[8] Mingchao Li et al. Octa-500: a retinal dataset for optical coherence tomography angiography study. *Medical Image Analysis*, 93:103092, 2024.

[9] Jun Ma, Yuting He, Feifei Li, Lin Han, Chenyu You, and Bo Wang. Segment anything in medical images. *Nature Communications*, 15(1):654, 2024.

[10] Junlong Cheng et al. Interactive medical image segmentation: A benchmark dataset and baseline. *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20841–20851, 2025.

[11] Jin Ye et al. Sa-med2d-20m dataset: Segment anything in 2d medical imaging with 20 million masks. *arXiv:2311.11969*, 2023.

[12] Rimsa Goperma, Rojan Basnet, Pragati Gautam Adhikari, Sagun Narayan Joshi, and Liang Zhao. Netra: Enhancing glaucoma diagnosis through deep learning-a comparative clinical validation study. *IEEE Region 10 Humanitarian Technology Conference*, pages 691–698, 2023.

[13] Hallee E Wong, Marianne Rakic, John Guttag, and Adrian V Dalca. Scribbleprompt: fast and flexible interactive segmentation for any biomedical image. *European Conference on Computer Vision*, pages 207–229, 2024.

[14] Zdravko Marinov, Paul F. Jäger, Jan Egger, Jens Kleesiek, and Rainer Stiefelhagen. Deep interactive segmentation of medical images: A systematic review and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10998–11018, 2024.