VLM3D: Vision-Language Modeling in 3D Medical Imaging: Structured description of the challenge design

CHALLENGE ORGANIZATION

Title

Use the title to convey the essential information on the challenge mission.

VLM3D: Vision-Language Modeling in 3D Medical Imaging

Challenge acronym

Preferable, provide a short acronym of the challenge (if any).

VLM3D

Challenge abstract

Provide a summary of the challenge purpose. This should include a general introduction in the topic from both a biomedical as well as from a technical point of view and clearly state the envisioned technical and/or biomedical impact of the challenge.

VLM3D 2026 is a large-scale benchmark for vision language modeling in 3D medical imaging, evaluating two modalities within the same edition: chest CT and brain MRI. The challenge focuses on clinically grounded tasks that reflect real radiology workflows, including radiology report generation, multi-abnormality classification, localization and segmentation, and text-conditional 3D image synthesis. Our goal is to accelerate reproducible and generalizable 3D multimodal foundation models by providing standardized datasets, evaluation code, and container-based benchmarking.

This is the second edition of VLM3D. In the first edition (MICCAI 2025), the benchmark evaluated chest CT only; although Boston external validation was mentioned, the Boston external test set was not utilized for official evaluation/ranking, and no expert radiologist human evaluation was performed. In VLM3D 2026, we evaluate two modalities within the same benchmark (chest CT and brain MRI), introduce a new brain MRI track via MR-RATE, include mandatory external validation using a closed Boston University test set, and conduct expert radiologist human evaluation for the top-performing methods. Results will be reported with track-specific and task-specific leaderboards, alongside mandatory external generalization reporting to quantify robustness under dataset shift.

Challenge keywords

List the primary keywords that characterize the challenge.

3D Medical Imaging, Vision-Language Modeling, Computed Tomography, Magnetic Resonance Imaging, Radiology Report Generation, Multimodal AI, Foundation Models, Clinical Validation

Year

2026

Novelty of the challenge

Briefly describe the novelty of the challenge.

VLM3D 2026 is the second edition of the VLM3D challenge series and is designed to ensure continuity for the community while introducing concrete new evaluation components. VLM3D 2025 (first edition) evaluated chest CT only. Although an external Boston test set was mentioned, it was not used for official evaluation or ranking in the executed benchmark, and no expert radiologist's human evaluation of top methods was performed.

VLM3D 2026 expands the benchmark in three specific ways. First, it evaluates two modalities within the same edition, chest CT and brain MRI, with dedicated tracks and leaderboards. Second, it introduces a new modality and dataset for the benchmark, brain MRI via MR-RATE, extending 3D vision language research beyond chest CT. Third, it includes mandatory external validation on a closed Boston University test set for every submission and adds expert radiologist human evaluation of top-performing methods, reported alongside quantitative metrics and external generalization performance.

We keep “VLM3D” in the title to maintain continuity with the established benchmark series and community recognition from the first edition, and we explicitly specify the scope in the subtitle by naming the two modalities evaluated in this edition, chest CT and brain MRI.

Task description and application scenarios

Briefly describe the application scenarios for the tasks in the challenge.

The tasks in the VLM3D 2026 challenge reflect real clinical scenarios in which radiologists analyze 3D CT and MRI scans and produce diagnostic reports. In the CT task, models are evaluated on generating full radiology reports, detecting and classifying multiple abnormalities, localizing pathological regions, and synthesizing realistic 3D CT volumes from text descriptions. These capabilities support automated reporting, decision support, explainability, and data augmentation in routine radiology workflows.

In the MRI task, models are evaluated on generating brain MRI reports and synthesizing 3D brain MRI volumes from text, which are directly applicable to neuroimaging diagnostics, education, and simulation-based training. Together, these tasks aim to develop vision–language systems that can assist radiologists in daily practice, reduce workload, improve consistency of reporting, and enable large-scale clinical research.

FURTHER INFORMATION FOR CONFERENCE ORGANIZERS

Workshop

If the challenge is part of a workshop, please indicate the workshop.

The VLM3D 2026 challenge will be organized as a satellite event at MICCAI 2026 and is not associated with a specific workshop.

Duration

How long does the challenge take?

Half day

In case you selected half or full day, please explain why you need a long slot for your challenge.

If accepted, we will run an on-site 4-hour session optimized for engagement and clarity. All container-based evaluation and leaderboard computation will be completed before MICCAI; the on-site session focuses on presenting results, external generalization findings, clinical human evaluation outcomes, and discussion.

Proposed 4-hour agenda
00:00 to 00:10 Opening, overview of CT and MRI tracks, rules recap
00:10 to 00:35 Invited talk on 3D multimodal foundation models and generalization
00:35 to 01:35 Top methods spotlight, CT track (short talks)
01:35 to 02:10 Break and posters or demos
02:10 to 03:00 Top methods spotlight, MRI track (short talks)
03:00 to 03:40 External validation analysis (Boston) and expert radiologist evaluation summary
03:40 to 04:00 Panel discussion, awards, closing

Expected number of participants

Please explain the basis of your estimate (e.g. numbers from previous challenges) and/or provide a list of potential participants and indicate if they have already confirmed their willingness to contribute.

The expected number of participants for VLM3D 2026 is based on the strong engagement observed in the first edition of the challenge at MICCAI 2025, which attracted over 450 registered participants and 87 submitted methods across four tasks, making it the most popular challenge at the conference. Building on this momentum, and with the addition of the new MR-RATE dataset, external validation, and clinical evaluation, we conservatively expect participation to exceed last year, with more than 500 participants and at least 100 method submissions.

In addition, multiple research groups from academia and industry who participated in the 2025 edition have already expressed interest in joining the 2026 challenge, including teams from the University of Zurich, Boston University, Harvard University, Johns Hopkins University, Imperial College London, Shanghai Jiao Tong University, the National Institutes of Health, and NVIDIA.

Publication and future plans

Please indicate if you plan to coordinate a publication of the challenge results.

We plan to coordinate a joint publication of the challenge results. An overview paper describing the datasets, tasks, and evaluation protocol will be released on arXiv before the announcement of the final rankings. After the challenge, all results will be consolidated into a journal manuscript targeted to Nature Medicine. All teams with valid submissions will be invited as co-authors, ensuring broad recognition of participant contributions and fostering long-term collaboration.

MICCAI LNCS proceedings

Indicate if you want to offer MICCAI Springer LNCS proceedings to the participants. Publishing a proceedings volume is optional and at the discretion of each challenge's organizers. At a minimum, organizers must ensure that a description of each participant's submission is publicly available. Organizers who wish to publish MICCAI Springer LNCS proceedings must adhere to the MICCAI Satellite events publication process.

Yes

Collaboration with European Society of Radiology (ESR)

In collaboration with European Society of Radiology (ESR), we announce special clinical interest topics with associated clinicians who can help with the preparation of the proposals; the best 3 challenge proposals on these topics will get the opportunity to present their challenges at the European Congress of Radiology (ECR) 2027 in a special session. If you want to organize a challenge in collaboration with ESR on one of these topics, please reach out to the MICCAI Challenges Team (miccai-challenges-2026@dkfz-heidelberg.de) and we will put you in contact with the corresponding clinician.

Challenge in collaboration with ESR. Ticking 'Yes' implies that the challenge has been prepared in collaboration with the clinical contact point.

No

Space and hardware requirements

Organizers of on-site challenges must provide a fair computing environment for all participants. For instance, algorithms should run on the same computing platform provided to all.

For the on-site event, we will require standard technical equipment including a projector and screen for presentations, microphones and loudspeakers for speakers and audience interaction, and a computer for displaying results, rankings, and live demonstrations. We will coordinate closely with the MICCAI organizers to ensure that all necessary technical support is available.

TASK 1: Vision–Language Understanding in 3D Computed Tomography

SUMMARY

Abstract

Provide a summary of the challenge purpose. This should include a general introduction in the topic from both a biomedical as well as from a technical point of view and clearly state the envisioned technical and/or biomedical impact of the challenge.

This task focuses on understanding and generating clinical information from 3D chest CT scans. Participants will build models that can generate full radiology reports, classify multiple abnormalities, localize pathological regions, and generate realistic 3D CT volumes from text. The goal is to support radiologists by reducing reporting time, improving diagnostic consistency, and enabling explainable AI systems for thoracic imaging.

Keywords

List the primary keywords that characterize the task.

3D CT, Vision–Language Modeling, Radiology Report Generation, Multimodal AI, Abnormality Detection, Text-Conditional Image Generation

ORGANIZATION

Organizers

a) Provide information on the organizing team (names and affiliations).

University of Zurich, Switzerland
Ibrahim Ethem Hamamci; Sezgin Er; Suprosanna Shit; Ezequiel De la Rosa; Anjany Sekuboyina; Murong Xu; Chinmay Prabhakar; Bjoern Menze

University Hospital Zurich, Switzerland
Christian Bluethgen

Istanbul Medipol University, Turkey
Ayse Gulnihan Simsek; Omer Faruk Durugol; Neslihan Simsek; Gulhan Ertan Akan; Melih Akan; Mehmet Kemal Ozdemir

Boston University, USA
Chenyu Wang; Weicheng Dai; Kayhan Batmanghelich

Harvard University, USA
Xiaoman Zhang; Mohammed Baharoon; Luyang Luo; Pranav Rajpurkar

Johns Hopkins University, USA
Pedro R. A. S. Bassi; Jieneng Chen; Yixiong Chen; Wenxuan Li; Alan Yuille; Zongwei Zhou

Imperial College London, UK
Hadrien Reynaud; Bernhard Kainz

Shanghai Jiao Tong University, China
Chaoyi Wu; Weidi Xie

National Institutes of Health (NIH), USA
Benjamin Hou; Zhiyong Lu

NVIDIA, USA
Daguang Xu; Dong Yang; Pengfei Guo; Marc Edgar

b) Provide information on the primary contact person.

Ibrahim Ethem Hamamci, MD
University of Zurich
ibrahim.hamamci@uzh.ch

c) Indicate whether clinicians are part of the organizing team. If yes, describe their role.

Yes.
Senior radiologists from University Hospital Zurich, Istanbul Medipol University Hospital, and Boston University Hospital are part of the organizing team and are responsible for data curation, report quality control, and clinical validation of the top-3 performing methods.

Life cycle type

Define the intended submission cycle of the challenge. Include information on whether/how the challenge will be continued after the challenge has taken place.Not every challenge closes after the submission deadline (one-time event). Sometimes it is possible to submit results after the deadline (open call) or the challenge is repeated with some modifications (repeated event).

Examples:

- One-time event with fixed conference submission deadline
- Open call (challenge opens for new submissions after conference deadline)
- Repeated event with annual fixed conference submission deadline

Repeated event with annual fixed conference submission deadline

Challenge venue and platform

a) Report the event (e.g. conference) that is associated with the challenge (if any).

MICCAI 2026 Satellite Event

b) Report the platform used to run the challenge.

https://vlm3dchallenge.com/

c) Do you agree that the your submission is shared with the platform (e.g., grand-challenge, synapse...) that you indicated?

Please note: 1) this purpose of such sharing is that the challenge chairs and the platform can communicate smoothly, your answer won't impact the review of your proposal; 2) regardless of your response to this question, it is your responsibility to perform all actions required by the platform (e.g. filling their submission request).

No

d) Provide the URL for the challenge website (if any).

https://vlm3dchallenge.com/

Participation policies

a) Define the allowed user interaction of the algorithms assessed. This includes the policy regarding any curation, (pre-)processing and (pre-)training steps.

No user interaction is allowed during testing. Users may curate training data and perform any preprocessing or pretraining prior to submission.

b) Define the policy on the usage of training data. The data used to train algorithms may, for example, be restricted to the data provided by the challenge or may also include publicly available data including (open) pre-trained nets. Clarify whether such additional data needs to be publicly available at the time of the challenge launch. Clarify whether adding (private) annotations of the public data is allowed.

Use of public data and pretrained models is allowed to reflect the current foundation-model setting. To improve interpretability despite heterogeneous training sources, each team will submit a brief data and model description summarizing public pretraining sources, fine-tuning datasets, and any pretrained checkpoints used. We will summarize performance trends post-challenge, including compute-performance tradeoffs and training regime comparisons.

c) Define the participation policy for members of the organizers' institutes. For example, members of the organizers' institutes may participate in the challenge but are not eligible for awards.

May participate but not eligible for awards and not listed in leaderboard

d) Define the award policy. In particular, provide details with respect to challenge prizes.

Top-3 teams will receive certificates and be invited to present at MICCAI 2026. Awards may be sponsored and announced later.

e) Define the policy for result announcement.

Examples:

- Top 3 performing methods will be announced publicly.
- Participating teams can choose whether the performance results will be made public.

Top-3 performing methods will be announced publicly.

f) Define the publication policy. In particular, provide details on ...

- ... who of the participating teams/the participating teams' members qualifies as author

- ... whether the participating teams may publish their own results separately, and (if so)
- ... whether an embargo time is defined (so that challenge organizers can publish a challenge paper first).

All teams with valid submissions will be invited as co-authors of the challenge paper. Teams may publish their own results independently after the official arXiv release.

Submission method

a) Describe the method used for result submission. Preferably, provide a link to the submission instructions.

Examples:

- Docker container on the Synapse platform. Link to submission instructions: <URL>
- Algorithm output was sent to organizers via e-mail. Submission instructions were sent by e-mail.

Algorithm container submission (Type 2) on our own website.

b) Provide information on the possibility for participating teams to evaluate their algorithms before submitting final results. For example, many challenges allow submission of multiple results, and only the last run is officially counted to compute challenge results.

Participants may submit multiple runs. Only the final valid submission will be used for ranking.

Challenge schedule

Provide a timetable for the challenge. Preferably, this should include

- the release date(s) of the training cases (if any)
- the registration date/period
- the release date(s) of the test cases and validation cases (if any)
- the submission date(s)
- associated workshop days (if any)
- the release date(s) of the results

Registration opens: March 2026

Training data release: March 2026

Test data release: June 2026

Submission deadline: August 2026

Results announced: September 2026 at MICCAI

Ethics approval

Indicate whether ethics approval is necessary for the data. If yes, provide details on the ethics approval, preferably institutional review board, location, date and number of the ethics approval (if applicable). Add the URL or a reference to the document of the ethics approval (if available).

Ethical approval obtained from Istanbul Medipol University Clinical Research Ethics Committee (E-10840098-772.02-6841, 27/10/2023). External validation data approved by Boston University IRB.

Data usage agreement

Clarify how the data can be used and distributed by the teams that participate in the challenge and by others during and after the challenge. This should include the explicit listing of the license applied.

Examples:

- CC BY (Attribution)
- CC BY-SA (Attribution-ShareAlike)
- CC BY-ND (Attribution-NoDerivs)
- CC BY-NC (Attribution-NonCommercial)
- CC BY-NC-SA (Attribution-NonCommercial-ShareAlike)
- CC BY-NC-ND (Attribution-NonCommercial-NoDerivs)

Please note that the data license should not differ among sources. In case a license has to be changed, it has to be reported to the MICCAI challenges team and changed in the proposal.

CC BY-NC-SA (Attribution-NonCommercial-ShareAlike)

Code availability

a) Provide information on the accessibility of the organizers' evaluation software (e.g. code to produce rankings). Preferably, provide a link to the code and add information on the supported platforms.

Evaluation code will be released publicly on GitHub prior to challenge start.

b) In an analogous manner, provide information on the accessibility of the participating teams' code.

Not required but encouraged.

Conflicts of interest

Provide information related to conflicts of interest. In particular provide information related to sponsoring/funding of the challenge. Also, state explicitly who had/will have access to the test case labels and when.

No direct commercial sponsorship. Test labels accessible only to designated evaluation organizers.

MISSION OF THE CHALLENGE

Field(s) of application

State the main field(s) of application that the participating algorithms target.

Examples:

- Diagnosis
- Education

- Intervention assistance
- Intervention follow-up
- Intervention planning
- Prognosis
- Research
- Screening
- Training
- Cross-phase
-

Diagnosis,Decision support,Education,Training

Task category(ies)

State the task category(ies)

Examples:

- Classification
- Detection
- Localization
- Modeling
- Prediction
- Reconstruction
- Registration
- Retrieval
- Segmentation
- Tracking

Classification,Segmentation,Localization,Reconstruction,Modeling

Cohorts

We distinguish between the target cohort and the challenge cohort. For example, a challenge could be designed around the task of medical instrument tracking in robotic kidney surgery. While the challenge could be based on ex vivo data obtained from a laparoscopic training environment with porcine organs (challenge cohort), the final biomedical application (i.e. robotic kidney surgery) would be targeted on real patients with certain characteristics defined by inclusion criteria such as restrictions regarding sex or age (target cohort).

a) Describe the target cohort, i.e. the subjects/objects from whom/which the data would be acquired in the final biomedical application.

Adult patients undergoing routine chest CT for thoracic disease assessment in clinical radiology. The intended application is automated reporting and decision support for common chest findings in daily workflows.

b) Describe the challenge cohort, i.e. the subject(s)/object(s) from whom/which the challenge data was acquired.

Chest CT scans and paired radiology reports collected as part of routine care. Training/validation are derived from the open CT-RATE cohort; evaluation uses a closed internal test set from the same source as CT-RATE and a closed external test set from Boston University.

Imaging modality(ies)

Specify the imaging technique(s) applied in the challenge.

Computed Tomography (CT)

Context information

Provide additional information given along with the images. The information may correspond ...

a) ... directly to the image data (e.g. tumor volume).

DICOM metadata (voxel spacing, reconstruction parameters, scanner information), and derived labels where applicable (e.g., abnormality labels extracted from reports for classification evaluation).

b) ... to the patient in general (e.g. sex, medical history).

Age and sex (where permitted by data governance).

Target entity(ies)

a) Describe the data origin, i.e. the region(s)/part(s) of subject(s)/object(s) from whom/which the image data would be acquired in the final biomedical application (e.g. brain shown in computed tomography (CT) data, abdomen shown in laparoscopic video data, operating room shown in video data, thorax shown in fluoroscopy video). If necessary, differentiate between target and challenge cohort.

Thorax/chest (lungs, pleura, mediastinum) in CT.

b) Describe the algorithm target, i.e. the structure(s)/subject(s)/object(s)/component(s) that the participating algorithms have been designed to focus on (e.g. tumor in the brain, tip of a medical instrument, nurse in an operating theater, catheter in a fluoroscopy scan). If necessary, differentiate between target and challenge cohort.

Algorithms must jointly reason over 3D CT volumes and text to produce: (i) a radiology report, (ii) multi-abnormality predictions, (iii) localization/segmentation outputs for target findings, and (iv) text-conditional 3D CT generation outputs for evaluation.

Assessment aim(s)

Identify the property(ies) of the algorithms to be optimized to perform well in the challenge. If multiple properties are assessed, prioritize them (if appropriate). The properties should then be reflected in the metrics applied (see below, parameter metric(s)), and the priorities should be reflected in the ranking when combining multiple metrics that assess different properties.

- Example 1: Find highly accurate liver segmentation algorithm for CT images.
- Example 2: Find lung tumor detection algorithm with high sensitivity and specificity for mammography images.

Corresponding metrics are listed below (parameter metric(s)).

Accuracy,Applicability

DATA SETS

Data source(s)

a) Specify the device(s) used to acquire the challenge data. This includes details on the device(s) used to acquire the imaging data (e.g. manufacturer) as well as information on additional devices used for performance assessment (e.g. tracking system used in a surgical setting).

Clinical CT scanners (multi-vendor; e.g., Philips, Siemens, and others depending on site).

b) Describe relevant details on the imaging process/data acquisition for each acquisition device (e.g. image acquisition protocol(s)).

Routine non-contrast chest CT acquisition in clinical practice; reconstruction settings vary by scanner and protocol.

c) Specify the center(s)/institute(s) in which the data was acquired and/or the data providing platform/source (e.g. previous challenge). If this information is not provided (e.g. for anonymization reasons), specify why.

Primary source: Istanbul Medipol University Hospital (CT-RATE and updated internal test). External validation: Boston University.

d) Describe relevant characteristics (e.g. level of expertise) of the subjects (e.g. surgeon)/objects (e.g. robot) involved in the data acquisition process (if any).

Scans acquired by trained CT technologists; reports written by board-certified radiologists as part of routine care.

Training and test case characteristics

a) State what is meant by one case in this challenge. A case encompasses all data that is processed to produce one result that is compared to the corresponding reference result (i.e. the desired algorithm output).

Examples:

- Training and test cases both represent a CT image of a human brain. Training cases have a weak annotation (tumor present or not and tumor volume (if any)) while the test cases are annotated with the tumor contour (if any).
- A case refers to all information that is available for one particular patient in a specific study. This information always includes the image information as specified in data source(s) (see above) and may include context information (see above). Both training and test cases are annotated with survival (binary) 5 years after (first) image was taken.

One case consists of a 3D chest CT volume with associated metadata and (where applicable) the paired radiology report and derived labels for evaluation. Outputs are generated report text, abnormality predictions, localization/segmentation outputs, and text-conditional CT generation outputs.

b) State the total number of training, validation and test cases.

Training: CT-RATE training split (as released)

Validation: CT-RATE validation split (as released)

Internal test (closed): ~2,000 cases (updated test set from CT-RATE source)

External test (closed): 1,024 cases (Boston University)

c) How much of the data are already annotated (stratified by train test in percentage)?

Train/Val: 100% paired reports; classification labels are available/derivable from reports.

Test: 100% paired reports for report-generation evaluation; additional evaluation annotations for localization/segmentation are provided for the evaluation subset.

d) Explain why a total number of cases and the specific proportion of training, validation and test cases was chosen.

We follow the established CT-RATE training/validation split. Closed internal and external test sets are used to measure generalization and prevent overfitting to public benchmarks, with an external site for robustness assessment.

e) Mention further important characteristics of the training, validation and test cases (e.g. class distribution in classification tasks chosen according to real-world distribution vs. equal class distribution) and justify the choice.

Cases are sampled from routine clinical practice to preserve realistic prevalence of findings and report language variability, improving clinical relevance and robustness.

f) Challenge organizers are encouraged to (partly) use unseen, unpublished data for their challenges. Describe if new data will be used for the challenge and state the number of cases along with the proportion of new data.

Yes. The test sets are closed and unseen, including an updated internal test set from the CT-RATE source and an external test set from Boston University.

Annotation characteristics

a) Describe the method for determining the reference annotation, i.e. the desired algorithm output. Provide the information separately for the training, validation and test cases if necessary. Possible methods include manual image annotation, in silico ground truth generation and annotation by automatic methods.

If human annotation was involved, state the number of annotators.

Reference reports are clinical reports written by radiologists. Classification targets are derived from reports using a validated report-label extraction pipeline. For segmentation/localization evaluation, a labeled subset is curated for benchmark scoring and clinical validation of top methods.

b) Provide the instructions given to the annotators (if any) prior to the annotation. This may include description of a training phase with the software. Provide the information separately for the training, validation and test cases if necessary. Preferably, provide a link to the annotation protocol.

Reports follow standard clinical reporting practice at each institution. For any additional evaluation annotations (e.g., segmentation/localization subset), annotators follow a written labeling protocol and consensus guidelines provided by the organizing radiologists.

c) Provide details on the subject(s)/algorithm(s) that annotated the cases (e.g. information on level of expertise such as number of years of professional experience, medically-trained or not). Provide the information separately for the training, validation and test cases if necessary.

Clinical reports written by board-certified radiologists. Additional evaluation labels (where used) are created/verified by expert radiologists and trained annotators under clinician supervision.

d) Describe the method(s) used to merge multiple annotations for one case (if any). Provide the information separately for the training, validation and test cases if necessary.

Where multiple annotations exist, we use clinician-led consensus (adjudication) to create a single reference per case for evaluation.

Data pre-processing method(s)

Describe the method(s) used for pre-processing the raw training data before it is provided to the participating teams. Provide the information separately for the training, validation and test cases if necessary.

Images are released in a standardized format; minimal preprocessing is applied beyond de-identification and consistent resampling/formatting required for evaluation. All necessary metadata (e.g., spacing) is provided to support participant preprocessing choices.

Sources of error

a) Describe the most relevant possible error sources related to the image annotation. If possible, estimate the magnitude (range) of these errors, using inter-and intra-annotator variability, for example. Provide the information separately for the training, validation and test cases, if necessary.

Variability in clinical reporting style, inter-reader variability, and differences in acquisition protocols across scanners/sites. For derived labels from text, errors may arise from label extraction; this is mitigated via validation and clinician review.

b) In an analogous manner, describe and quantify other relevant sources of error.

Domain shift between institutions (scanner vendor, protocol, population), and distribution shift over time. This is explicitly measured through external validation.

ASSESSMENT METHODS

Metric(s)

a) Define the metric(s) to assess a property of an algorithm. These metrics should reflect the desired algorithm properties described in assessment aim(s) (see above). State which metric(s) were used to compute the ranking(s) (if any).

- Example 1: Dice Similarity Coefficient (DSC)
- Example 2: Area under curve (AUC)

Report generation: ROUGE-L, BLEU-4, METEOR, and clinical label F1 (labels extracted from generated vs reference reports)

Classification: AUROC (macro), F1 (macro), precision/recall

Segmentation/localization: Dice, IoU (on the labeled evaluation subset)

Text-conditional CT generation: feature-based similarity (e.g., FID on medical feature encoder) +

realism/consistency checks via a blinded classifier

b) Justify why the metric(s) was/were chosen, preferably with reference to the biomedical application.

We combine language-overlap metrics with clinically meaningful label-based scoring because high textual similarity alone does not guarantee correct clinical content. Classification/segmentation metrics are standard in medical imaging and directly reflect diagnostic utility. Generation metrics include feature-based similarity to capture realism beyond pixel-level overlap.

Ranking method(s)

a) Describe the method used to compute a performance rank for all submitted algorithms based on the generated metric results on the test cases. Typically the text will describe how results obtained per case and metric are aggregated to arrive at a final score/ranking. Ideally, provide the ranking scheme as a concrete pseudo code.

We use a robust point-based ranking similar to prior MICCAI challenges: for each metric, we compute case-level scores, perform pairwise comparisons between teams using a two-sided permutation test, and award points for statistically significant wins. Final rank is determined by total points aggregated across primary metrics (with predefined weights per subtask).

b) Describe the method(s) used to manage submissions with missing results on test cases.

Missing outputs for any subtask are treated as invalid for that subtask and receive the lowest possible score (or zero) for the corresponding metrics, ensuring fair penalization.

c) Justify why the described ranking scheme(s) was/were used.

Point-based ranking with significance testing is more stable than simple averaging because it accounts for variability across cases and reduces sensitivity to outliers, while providing a fair comparison when performance differences are small.

Statistical analyses

Provide an overview of the statistical approaches used in the scope of the challenge analysis. Details can be provided in the parameters below. For each parameter, justify why the described statistical method(s) was/were used and, if necessary, add a description of any method used to assess whether the data met the assumptions required for the particular statistical approach.

We report per-case metric distributions, compute confidence intervals via bootstrap, and assess statistical significance between methods using permutation testing. External validation performance is analyzed separately to quantify generalization. External test cases from Boston University are used for mandatory generalization evaluation and are reported as a separate external leaderboard for every team. The official ranking is computed on the closed internal test set only. We will additionally report internal-to-external generalization gaps and stratified analyses where feasible to characterize performance under dataset shift.

We will conduct expert human evaluation for the top-performing methods selected from the official internal leaderboard. Board-certified radiologists will assess outputs using a structured rubric focused on clinical usefulness, factual correctness, and safety-critical errors. Methods will be anonymized during review and cases will be randomized. We will report clinician preference rates and score distributions with confidence intervals, and we will include inter-rater agreement statistics appropriate to the rating format.

Provide a description of how the precision of the performance estimates of individual algorithms is assessed (e.g. confidence interval of the mean on the test set computed using percentile bootstrap, confidence interval of the accuracy on the test set computed using percentile bootstrap).

95% confidence intervals computed using percentile bootstrap over test cases (e.g., 1,000 bootstrap samples) for primary metrics.

Provide a description of how variability of the performance of individual algorithms across tests cases is assessed (e.g. SD across test cases, IQR, graphs, reporting outliers...).

We report SD/IQR across cases and provide boxplots/violin plots; we also highlight outliers and stratify by acquisition site (internal vs external).

Provide a description of how variability of rankings is assessed.

We assess ranking robustness via bootstrap resampling of test cases and report the distribution of ranks per team (rank stability).

Provide a description of statistical tests that are used to assess whether the differences in performance between algorithms are statistically significant.

Two-sided permutation tests for pairwise method comparisons on case-level metric values, with multiple-comparison correction where appropriate.

Provide a description of the missing data handling.

Missing predictions are penalized as described (lowest score/invalid for that subtask). Partial submissions are ranked only on submitted subtasks if the rules allow; otherwise marked invalid.

Indicate any software product that is used for all data analysis methods.

Python (NumPy, SciPy, scikit-learn), with evaluation scripts released by organizers.

Further analyses

Present further analyses to be performed (if applicable), e.g. related to

- combining algorithms via ensembling,
- inter-algorithm variability,
- common problems/biases of the submitted methods, or
- ranking variability.

We will analyze (i) internal vs external generalization gaps, (ii) failure modes by pathology/protocol, (iii) clinical validation outcomes for top-3 methods, and (iv) potential ensembling of top methods.

TASK 2: Vision–Language Understanding in 3D Magnetic Resonance Imaging

SUMMARY

Abstract

Provide a summary of the challenge purpose. This should include a general introduction in the topic from both a biomedical as well as from a technical point of view and clearly state the envisioned technical and/or biomedical impact of the challenge.

This task introduces vision–language modeling for brain MRI using the MR-RATE dataset with over 50,000 paired scans and reports. Participants will develop models for brain MRI report generation and text-conditional MRI synthesis. The task aims to extend multimodal AI to neuroimaging, enabling scalable clinical reporting, simulation, and education.

Keywords

List the primary keywords that characterize the task.

Brain MRI, Vision–Language Modeling, Radiology Report Generation, Text-Conditional MRI Synthesis, Multimodal AI

ORGANIZATION

Organizers

a) Provide information on the organizing team (names and affiliations).

University of Zurich, Switzerland
Ibrahim Ethem Hamamci; Sezgin Er; Suprosanna Shit; Ezequiel De la Rosa; Anjany Sekuboyina; Murong Xu; Chinmay Prabhakar; Bjoern Menze

University Hospital Zurich, Switzerland
Christian Bluethgen

Istanbul Medipol University, Turkey
Ayse Gulnihan Simsek; Omer Faruk Durugol; Neslihan Simsek; Gulhan Ertan Akan; Melih Akan; Mehmet Kemal Ozdemir

Boston University, USA
Chenyu Wang; Weicheng Dai; Kayhan Batmanghelich

Harvard University, USA
Xiaoman Zhang; Mohammed Baharoon; Luyang Luo; Pranav Rajpurkar

Johns Hopkins University, USA
Pedro R. A. S. Bassi; Jieneng Chen; Yixiong Chen; Wenxuan Li; Alan Yuille; Zongwei Zhou

Imperial College London, UK
Hadrien Reynaud; Bernhard Kainz

Shanghai Jiao Tong University, China
Chaoyi Wu; Weidi Xie

National Institutes of Health (NIH), USA
Benjamin Hou; Zhiyong Lu

NVIDIA, USA
Daguang Xu; Dong Yang; Pengfei Guo; Marc Edgar

b) Provide information on the primary contact person.

Sezgin Er
University of Zurich
sezgin.er@uzh.ch

c) Indicate whether clinicians are part of the organizing team. If yes, describe their role.

Yes.
Senior radiologists from University Hospital Zurich, Istanbul Medipol University Hospital, and Boston University Hospital are part of the organizing team and are responsible for data curation, report quality control, and clinical validation of the top-3 performing methods.

Life cycle type

Define the intended submission cycle of the challenge. Include information on whether/how the challenge will be continued after the challenge has taken place.Not every challenge closes after the submission deadline (one-time event). Sometimes it is possible to submit results after the deadline (open call) or the challenge is repeated with some modifications (repeated event).

Examples:

- One-time event with fixed conference submission deadline
- Open call (challenge opens for new submissions after conference deadline)
- Repeated event with annual fixed conference submission deadline

Repeated event with annual fixed conference submission deadline

Challenge venue and platform

a) Report the event (e.g. conference) that is associated with the challenge (if any).

MICCAI 2026 Satellite Event

b) Report the platform used to run the challenge.

https://vlm3dchallenge.com/

c) Do you agree that the your submission is shared with the platform (e.g., grand-challenge, synapse...) that you indicated?

Please note: 1) this purpose of such sharing is that the challenge chairs and the platform can communicate smoothly, your answer won't impact the review of your proposal; 2) regardless of your response to this question, it is your responsibility to perform all actions required by the platform (e.g. filling their submission request).

No

d) Provide the URL for the challenge website (if any).

https://vlm3dchallenge.com/

Participation policies

a) Define the allowed user interaction of the algorithms assessed. This includes the policy regarding any curation, (pre-)processing and (pre-)training steps.

No user interaction is allowed during testing. Users may curate training data and perform any preprocessing or pretraining prior to submission.

b) Define the policy on the usage of training data. The data used to train algorithms may, for example, be restricted to the data provided by the challenge or may also include publicly available data including (open) pre-trained nets. Clarify whether such additional data needs to be publicly available at the time of the challenge launch. Clarify whether adding (private) annotations of the public data is allowed.

Use of public data and pretrained models is allowed to reflect the current foundation-model setting. To improve interpretability despite heterogeneous training sources, each team will submit a brief data and model description summarizing public pretraining sources, fine-tuning datasets, and any pretrained checkpoints used. We will summarize performance trends post-challenge, including compute-performance tradeoffs and training regime comparisons.

c) Define the participation policy for members of the organizers' institutes. For example, members of the organizers' institutes may participate in the challenge but are not eligible for awards.

May participate but not eligible for awards and not listed in leaderboard

d) Define the award policy. In particular, provide details with respect to challenge prizes.

Top-3 teams will receive certificates and be invited to present at MICCAI 2026. Awards may be sponsored and announced later.

e) Define the policy for result announcement.

Examples:

- Top 3 performing methods will be announced publicly.
- Participating teams can choose whether the performance results will be made public.

Top-3 performing methods will be announced publicly.

f) Define the publication policy. In particular, provide details on ...

- ... who of the participating teams/the participating teams' members qualifies as author

- ... whether the participating teams may publish their own results separately, and (if so)
- ... whether an embargo time is defined (so that challenge organizers can publish a challenge paper first).

All teams with valid submissions will be invited as co-authors of the challenge paper. Teams may publish their own results independently after the official arXiv release.

Submission method

a) Describe the method used for result submission. Preferably, provide a link to the submission instructions.

Examples:

- Docker container on the Synapse platform. Link to submission instructions: <URL>
- Algorithm output was sent to organizers via e-mail. Submission instructions were sent by e-mail.

Algorithm container submission (Type 2) on our own website.

b) Provide information on the possibility for participating teams to evaluate their algorithms before submitting final results. For example, many challenges allow submission of multiple results, and only the last run is officially counted to compute challenge results.

Participants may submit multiple runs. Only the final valid submission will be used for ranking.

Challenge schedule

Provide a timetable for the challenge. Preferably, this should include

- the release date(s) of the training cases (if any)
- the registration date/period
- the release date(s) of the test cases and validation cases (if any)
- the submission date(s)
- associated workshop days (if any)
- the release date(s) of the results

Registration opens: March 2026

Training data release: March 2026

Test data release: June 2026

Submission deadline: August 2026

Results announced: September 2026 at MICCAI

Ethics approval

Indicate whether ethics approval is necessary for the data. If yes, provide details on the ethics approval, preferably institutional review board, location, date and number of the ethics approval (if applicable). Add the URL or a reference to the document of the ethics approval (if available).

Ethical approval obtained from Istanbul Medipol University Clinical Research Ethics Committee (E-10840098-772.02-6841, 27/10/2023). External validation data approved by Boston University IRB.

Data usage agreement

Clarify how the data can be used and distributed by the teams that participate in the challenge and by others during and after the challenge. This should include the explicit listing of the license applied.

Examples:

- CC BY (Attribution)
- CC BY-SA (Attribution-ShareAlike)
- CC BY-ND (Attribution-NoDerivs)
- CC BY-NC (Attribution-NonCommercial)
- CC BY-NC-SA (Attribution-NonCommercial-ShareAlike)
- CC BY-NC-ND (Attribution-NonCommercial-NoDerivs)

Please note that the data license should not differ among sources. In case a license has to be changed, it has to be reported to the MICCAI challenges team and changed in the proposal.

CC BY-NC-SA (Attribution-NonCommercial-ShareAlike)

Code availability

a) Provide information on the accessibility of the organizers' evaluation software (e.g. code to produce rankings). Preferably, provide a link to the code and add information on the supported platforms.

Evaluation code will be released publicly on GitHub prior to challenge start.

b) In an analogous manner, provide information on the accessibility of the participating teams' code.

Not required but encouraged.

Conflicts of interest

Provide information related to conflicts of interest. In particular provide information related to sponsoring/funding of the challenge. Also, state explicitly who had/will have access to the test case labels and when.

No direct commercial sponsorship. Test labels accessible only to designated evaluation organizers.

MISSION OF THE CHALLENGE

Field(s) of application

State the main field(s) of application that the participating algorithms target.

Examples:

- Diagnosis
- Education

- Intervention assistance
- Intervention follow-up
- Intervention planning
- Prognosis
- Research
- Screening
- Training
- Cross-phase
-

Diagnosis,Decision support,Education,Training

Task category(ies)

State the task category(ies)

Examples:

- Classification
- Detection
- Localization
- Modeling
- Prediction
- Reconstruction
- Registration
- Retrieval
- Segmentation
- Tracking

Modeling,Reconstruction

Cohorts

We distinguish between the target cohort and the challenge cohort. For example, a challenge could be designed around the task of medical instrument tracking in robotic kidney surgery. While the challenge could be based on ex vivo data obtained from a laparoscopic training environment with porcine organs (challenge cohort), the final biomedical application (i.e. robotic kidney surgery) would be targeted on real patients with certain characteristics defined by inclusion criteria such as restrictions regarding sex or age (target cohort).

a) Describe the target cohort, i.e. the subjects/objects from whom/which the data would be acquired in the final biomedical application.

Patients undergoing routine brain MRI for clinical neuroimaging assessment. The intended application is automated report drafting and clinically plausible image synthesis for education, simulation, and model

development.

b) Describe the challenge cohort, i.e. the subject(s)/object(s) from whom/which the challenge data was acquired.

Brain MRI scans and paired radiology reports from the MR-RATE cohort (≥50k exams with reports). Evaluation uses a closed internal test set and a closed external test set from Boston University to assess generalization.

Imaging modality(ies)

Specify the imaging technique(s) applied in the challenge.

Magnetic Resonance Imaging (MRI)

Context information

Provide additional information given along with the images. The information may correspond ...

a) ... directly to the image data (e.g. tumor volume).

DICOM metadata (voxel spacing, sequence/protocol information where available, scanner information) and paired report text.

b) ... to the patient in general (e.g. sex, medical history).

Age and sex (where permitted by data governance).

Target entity(ies)

a) Describe the data origin, i.e. the region(s)/part(s) of subject(s)/object(s) from whom/which the image data would be acquired in the final biomedical application (e.g. brain shown in computed tomography (CT) data, abdomen shown in laparoscopic video data, operating room shown in video data, thorax shown in fluoroscopy video). If necessary, differentiate between target and challenge cohort.

Brain in MRI.

b) Describe the algorithm target, i.e. the structure(s)/subject(s)/object(s)/component(s) that the participating algorithms have been designed to focus on (e.g. tumor in the brain, tip of a medical instrument, nurse in an operating theater, catheter in a fluoroscopy scan). If necessary, differentiate between target and challenge cohort.

Algorithms must jointly reason over 3D brain MRI volumes and text to produce: (i) a radiology report, and (ii) text-conditional 3D brain MRI generation outputs for evaluation.

Assessment aim(s)

Identify the property(ies) of the algorithms to be optimized to perform well in the challenge. If multiple properties are assessed, prioritize them (if appropriate). The properties should then be reflected in the metrics applied (see below, parameter metric(s)), and the priorities should be reflected in the ranking when combining multiple metrics that assess different properties.

- Example 1: Find highly accurate liver segmentation algorithm for CT images.
- Example 2: Find lung tumor detection algorithm with high sensitivity and specificity for mammography images.

Corresponding metrics are listed below (parameter metric(s)).

Accuracy

DATA SETS

Data source(s)

a) Specify the device(s) used to acquire the challenge data. This includes details on the device(s) used to acquire the imaging data (e.g. manufacturer) as well as information on additional devices used for performance assessment (e.g. tracking system used in a surgical setting).

Clinical MRI scanners (multi-vendor; site-dependent; typical 1.5T/3T clinical systems).

b) Describe relevant details on the imaging process/data acquisition for each acquisition device (e.g. image acquisition protocol(s)).

Routine clinical brain MRI protocols (multi-sequence exams; protocol composition varies by site and indication).

c) Specify the center(s)/institute(s) in which the data was acquired and/or the data providing platform/source (e.g. previous challenge). If this information is not provided (e.g. for anonymization reasons), specify why.

Primary source: MR-RATE contributing center(s) (Istanbul Medipol University Hospital).

d) Describe relevant characteristics (e.g. level of expertise) of the subjects (e.g. surgeon)/objects (e.g. robot) involved in the data acquisition process (if any).

Scans acquired by trained MRI technologists; reports written by board-certified radiologists as part of routine care.

Training and test case characteristics

a) State what is meant by one case in this challenge. A case encompasses all data that is processed to produce one result that is compared to the corresponding reference result (i.e. the desired algorithm output).

Examples:

- Training and test cases both represent a CT image of a human brain. Training cases have a weak annotation (tumor present or not and tumor volume (if any)) while the test cases are annotated with the tumor contour (if any).
- A case refers to all information that is available for one particular patient in a specific study. This information always includes the image information as specified in data source(s) (see above) and may include context information (see above). Both training and test cases are annotated with survival (binary) 5 years after (first) image was taken.

One case consists of a 3D brain MRI exam (volume(s) per case according to the released format) with associated metadata and the paired radiology report. Outputs are generated report text and text-conditional MRI generation outputs.

b) State the total number of training, validation and test cases.

Training: ~45,000 cases (from MR-RATE)

Validation: ~2,000 cases

Internal test (closed): ~1,500 cases

c) How much of the data are already annotated (stratified by train test in percentage)?

Train/Val/Test: 100% paired reports.

d) Explain why a total number of cases and the specific proportion of training, validation and test cases was chosen.

We allocate the majority of MR-RATE to training to support foundation-model scale learning and maintain separate closed internal/external test sets to enable unbiased evaluation and generalization analysis.

e) Mention further important characteristics of the training, validation and test cases (e.g. class distribution in classification tasks chosen according to real-world distribution vs. equal class distribution) and justify the choice.

Cases are sampled from routine clinical practice to preserve realistic report style and pathology prevalence, supporting generalizable neuroimaging models.

f) Challenge organizers are encouraged to (partly) use unseen, unpublished data for their challenges. Describe if new data will be used for the challenge and state the number of cases along with the proportion of new data.

Yes. The internal test set is closed and unseen at submission time.

Annotation characteristics

a) Describe the method for determining the reference annotation, i.e. the desired algorithm output. Provide the information separately for the training, validation and test cases if necessary. Possible methods include manual image annotation, in silico ground truth generation and annotation by automatic methods.

If human annotation was involved, state the number of annotators.

Reference reports are clinical reports written by radiologists. Generation evaluation uses the reference imaging data and report text as ground truth targets.

b) Provide the instructions given to the annotators (if any) prior to the annotation. This may include description of a training phase with the software. Provide the information separately for the training, validation and test cases if necessary. Preferably, provide a link to the annotation protocol.

Reports follow standard clinical reporting practice at each institution. Any additional evaluation curation follows institutional clinical guidelines.

c) Provide details on the subject(s)/algorithm(s) that annotated the cases (e.g. information on level of expertise such as number of years of professional experience, medically-trained or not). Provide the information separately for the training, validation and test cases if necessary.

Clinical reports written by board-certified radiologists.

d) Describe the method(s) used to merge multiple annotations for one case (if any). Provide the information separately for the training, validation and test cases if necessary.

Not applicable (single clinical report per case); if multiple readings exist, clinician adjudication is applied.

Data pre-processing method(s)

Describe the method(s) used for pre-processing the raw training data before it is provided to the participating teams. Provide the information separately for the training, validation and test cases if necessary.

Images are released in a standardized format; minimal preprocessing is applied beyond de-identification and consistent resampling/formatting required for evaluation. All necessary metadata (e.g., spacing) is provided to

support participant preprocessing choices.

Sources of error

a) Describe the most relevant possible error sources related to the image annotation. If possible, estimate the magnitude (range) of these errors, using inter-and intra-annotator variability, for example. Provide the information separately for the training, validation and test cases, if necessary.

Variability in clinical reporting style and imaging protocol differences across sites/scanners.

b) In an analogous manner, describe and quantify other relevant sources of error.

Domain shift between institutions (scanner vendor, protocol, population), and distribution shift over time. This is explicitly measured through external validation.

ASSESSMENT METHODS

Metric(s)

a) Define the metric(s) to assess a property of an algorithm. These metrics should reflect the desired algorithm properties described in assessment aim(s) (see above). State which metric(s) were used to compute the ranking(s) (if any).

- Example 1: Dice Similarity Coefficient (DSC)
- Example 2: Area under curve (AUC)

Report generation: ROUGE-L, BLEU-4, METEOR, and clinical label F1 (derived from reports)

Text-conditional MRI generation: feature-based similarity (FID-like) + consistency checks via a blinded classifier

b) Justify why the metric(s) was/were chosen, preferably with reference to the biomedical application.

Same rationale: combine language quality with clinical correctness for reports, and use feature-based similarity and consistency checks for generation to evaluate clinical plausibility and generalization.

Ranking method(s)

a) Describe the method used to compute a performance rank for all submitted algorithms based on the generated metric results on the test cases. Typically the text will describe how results obtained per case and metric are aggregated to arrive at a final score/ranking. Ideally, provide the ranking scheme as a concrete pseudo code.

We use a robust point-based ranking similar to prior MICCAI challenges: for each metric, we compute case-level scores, perform pairwise comparisons between teams using a two-sided permutation test, and award points for statistically significant wins. Final rank is determined by total points aggregated across primary metrics (with predefined weights per subtask).

b) Describe the method(s) used to manage submissions with missing results on test cases.

Missing outputs for any subtask are treated as invalid for that subtask and receive the lowest possible score (or zero) for the corresponding metrics, ensuring fair penalization.

c) Justify why the described ranking scheme(s) was/were used.

Point-based ranking with significance testing is more stable than simple averaging because it accounts for variability across cases and reduces sensitivity to outliers, while providing a fair comparison when performance differences are small.

Statistical analyses

Provide an overview of the statistical approaches used in the scope of the challenge analysis. Details can be provided in the parameters below. For each parameter, justify why the described statistical method(s) was/were used and, if necessary, add a description of any method used to assess whether the data met the assumptions required for the particular statistical approach.

We report per-case metric distributions, compute confidence intervals via bootstrap, and assess statistical significance between methods using permutation testing. External validation performance is analyzed separately to quantify generalization. External test cases from Boston University are used for mandatory generalization evaluation and are reported as a separate external leaderboard for every team. The official ranking is computed on the closed internal test set only. We will additionally report internal-to-external generalization gaps and stratified analyses where feasible to characterize performance under dataset shift.

We will conduct expert human evaluation for the top-performing methods selected from the official internal leaderboard. Board-certified radiologists will assess outputs using a structured rubric focused on clinical usefulness, factual correctness, and safety-critical errors. Methods will be anonymized during review and cases will be randomized. We will report clinician preference rates and score distributions with confidence intervals, and we will include inter-rater agreement statistics appropriate to the rating format.

Provide a description of how the precision of the performance estimates of individual algorithms is assessed (e.g. confidence interval of the mean on the test set computed using percentile bootstrap, confidence interval of the accuracy on the test set computed using percentile bootstrap).

95% confidence intervals computed using percentile bootstrap over test cases (e.g., 1,000 bootstrap samples) for primary metrics.

Provide a description of how variability of the performance of individual algorithms across tests cases is assessed (e.g. SD across test cases, IQR, graphs, reporting outliers...).

We report SD/IQR across cases and provide boxplots/violin plots; we also highlight outliers and stratify by acquisition site (internal vs external).

Provide a description of how variability of rankings is assessed.

We assess ranking robustness via bootstrap resampling of test cases and report the distribution of ranks per team (rank stability).

Provide a description of statistical tests that are used to assess whether the differences in performance between algorithms are statistically significant.

Two-sided permutation tests for pairwise method comparisons on case-level metric values, with multiple-comparison correction where appropriate.

Provide a description of the missing data handling.

Missing predictions are penalized as described (lowest score/invalid for that subtask). Partial submissions are ranked only on submitted subtasks if the rules allow; otherwise marked invalid.

Indicate any software product that is used for all data analysis methods.

Python (NumPy, SciPy, scikit-learn), with evaluation scripts released by organizers.

Further analyses

Present further analyses to be performed (if applicable), e.g. related to

- combining algorithms via ensembling,
- inter-algorithm variability,
- common problems/biases of the submitted methods, or
- ranking variability.

We will analyze (i) internal vs external generalization gaps, (ii) failure modes by pathology/protocol, (iii) clinical validation outcomes for top-3 methods, and (iv) potential ensembling of top methods.

TASK 3: Report-Supervision for Multi-Tumor Segmentation in 3D Computed Tomography

SUMMARY

Abstract

Provide a summary of the challenge purpose. This should include a general introduction in the topic from both a biomedical as well as from a technical point of view and clearly state the envisioned technical and/or biomedical impact of the challenge.

This task focuses on using radiology reports to train AI models to detect, segment, and diagnose (malignant vs. benign) multiple tumor types. Multi-tumor segmentation enables interpretable AI models with strong potential to support earlier cancer detection. However, segmentation models are traditionally trained using tumor masks, which are scarce in the public domain, especially for multiple tumor types. As a result, current public AI models cannot accurately detect or segment many clinically relevant tumors.

In contrast to masks, radiology reports are widely available at large scale and contain rich tumor information such as tumor count, size, location, and attenuation. Therefore, reports can potentially replace or complement masks for training tumor segmentation models (Bassi et al., MICCAI Best Paper Award Runner-up, 2025).

Participants will be provided with a large-scale dataset containing over 25,000 CT scans with radiology reports covering 13 tumor types: esophagus, stomach, duodenum, bladder, gallbladder, spleen, adrenal glands, kidney, uterus, prostate, colon, liver, and pancreas. For 7 of these tumor types, 100 tumor segmentation masks will also be provided. This task aims to foster new AI architectures and training strategies that learn tumor detection, segmentation, and diagnosis from language supervision in radiology reports, with or without a small number of segmentation masks.

Report supervision enables scalable learning and supports segmentation of tumor types that currently lack public annotations (8 of the 13 tumor types in this task). Ultimately, this task aims to improve interpretable AI systems that assist radiologists in opportunistic early detection of multiple cancers.

Keywords

List the primary keywords that characterize the task.

3D CT, Report-Supervision, Radiology Reports, Tumor Detection, Tumor Segmentation, Cancer Diagnosis, Vision-Language Models

ORGANIZATION

Organizers

a) Provide information on the organizing team (names and affiliations).

University of Zurich, Switzerland
Ibrahim Ethem Hamamci; Sezgin Er; Suprosanna Shit; Ezequiel De la Rosa; Anjany Sekuboyina; Murong Xu; Chinmay Prabhakar; Bjoern Menze

University Hospital Zurich, Switzerland
Christian Bluethgen

Istanbul Medipol University, Turkey
Ayse Gulnihan Simsek; Omer Faruk Durugol; Neslihan Simsek; Gulhan Ertan Akan; Melih Akan; Mehmet Kemal Ozdemir

Boston University, USA
Chenyu Wang; Weicheng Dai; Kayhan Batmanghelich

Harvard University, USA
Xiaoman Zhang; Mohammed Baharoon; Luyang Luo; Pranav Rajpurkar

Johns Hopkins University, USA
Pedro R. A. S. Bassi; Jieneng Chen; Yixiong Chen; Wenxuan Li; Alan Yuille; Zongwei Zhou

Imperial College London, UK
Hadrien Reynaud; Bernhard Kainz

Shanghai Jiao Tong University, China
Chaoyi Wu; Weidi Xie

National Institutes of Health (NIH), USA
Benjamin Hou; Zhiyong Lu

NVIDIA, USA
Daguang Xu; Dong Yang; Pengfei Guo; Marc Edgar

b) Provide information on the primary contact person.

Pedro Salvador Bassi
Johns Hopkins University
psalvad2@jh.edu

c) Indicate whether clinicians are part of the organizing team. If yes, describe their role.

Yes.
Senior radiologists are part of the organizing team and are responsible for data curation, report quality control, and clinical validation of the top-3 performing methods.

Life cycle type

Define the intended submission cycle of the challenge. Include information on whether/how the challenge will be continued after the challenge has taken place.Not every challenge closes after the submission deadline (one-time event). Sometimes it is possible to submit results after the deadline (open call) or the challenge is repeated with some

modifications (repeated event).

Examples:

- One-time event with fixed conference submission deadline
- Open call (challenge opens for new submissions after conference deadline)
- Repeated event with annual fixed conference submission deadline

Repeated event with annual fixed conference submission deadline

Challenge venue and platform

a) Report the event (e.g. conference) that is associated with the challenge (if any).

MICCAI 2026 Satellite Event

b) Report the platform used to run the challenge.

https://vlm3dchallenge.com/

c) Do you agree that the your submission is shared with the platform (e.g., grand-challenge, synapse...) that you indicated?

Please note: 1) this purpose of such sharing is that the challenge chairs and the platform can communicate smoothly, your answer won't impact the review of your proposal; 2) regardless of your response to this question, it is your responsibility to perform all actions required by the platform (e.g. filling their submission request).

No

d) Provide the URL for the challenge website (if any).

https://vlm3dchallenge.com/

Participation policies

a) Define the allowed user interaction of the algorithms assessed. This includes the policy regarding any curation, (pre-)processing and (pre-)training steps.

No user interaction is allowed during testing. Users may curate training data and perform any preprocessing or pretraining prior to submission.

b) Define the policy on the usage of training data. The data used to train algorithms may, for example, be restricted to the data provided by the challenge or may also include publicly available data including (open) pre-trained nets. Clarify whether such additional data needs to be publicly available at the time of the challenge launch. Clarify whether adding (private) annotations of the public data is allowed.

Use of public data and pretrained models is allowed to reflect the current foundation-model setting. To improve interpretability despite heterogeneous training sources, each team will submit a brief data and model description summarizing public pretraining sources, fine-tuning datasets, and any pretrained checkpoints used. We will summarize performance trends post-challenge, including compute-performance tradeoffs and training regime comparisons.

c) Define the participation policy for members of the organizers' institutes. For example, members of the organizers' institutes may participate in the challenge but are not eligible for awards.

May participate but not eligible for awards and not listed in leaderboard

d) Define the award policy. In particular, provide details with respect to challenge prizes.

Top-3 teams will receive certificates and be invited to present at MICCAI 2026. Awards may be sponsored and announced later.

e) Define the policy for result announcement.

Examples:

- Top 3 performing methods will be announced publicly.
- Participating teams can choose whether the performance results will be made public.

Top-3 performing methods will be announced publicly.

f) Define the publication policy. In particular, provide details on ...

- ... who of the participating teams/the participating teams' members qualifies as author
- ... whether the participating teams may publish their own results separately, and (if so)
- ... whether an embargo time is defined (so that challenge organizers can publish a challenge paper first).

All teams with valid submissions will be invited as co-authors of the challenge paper. Teams may publish their own results independently after the official arXiv release.

Submission method

a) Describe the method used for result submission. Preferably, provide a link to the submission instructions.

Examples:

- Docker container on the Synapse platform. Link to submission instructions: <URL>
- Algorithm output was sent to organizers via e-mail. Submission instructions were sent by e-mail.

CT scans from an internal validation dataset will be released publicly.

For each CT scan in the public test set, participants must provide:

Probability of any tumor (benign or malignant) in each of the 13 organs

Tumor segmentation mask

Probability of malignant tumor in each of the 13 organs

Malignant tumor segmentation mask

Participants must also submit their trained model and evaluation code.

The top-5 models will be evaluated on a private large-scale external test dataset from the University of California San Francisco (UCSF).

b) Provide information on the possibility for participating teams to evaluate their algorithms before submitting final results. For example, many challenges allow submission of multiple results, and only the last run is officially counted to compute challenge results.

Participants may submit multiple runs. Only the final valid submission will be used for ranking.

Challenge schedule

Provide a timetable for the challenge. Preferably, this should include

- the release date(s) of the training cases (if any)
- the registration date/period
- the release date(s) of the test cases and validation cases (if any)
- the submission date(s)
- associated workshop days (if any)
- the release date(s) of the results

Registration opens: March 2026

Ethics approval

Indicate whether ethics approval is necessary for the data. If yes, provide details on the ethics approval, preferably institutional review board, location, date and number of the ethics approval (if applicable). Add the URL or a reference to the document of the ethics approval (if available).

Ethical approval obtained from Istanbul Medipol University Clinical Research Ethics Committee (E-10840098-772.02-6841, 27/10/2023). External validation data approved by Boston University IRB.

Data usage agreement

Clarify how the data can be used and distributed by the teams that participate in the challenge and by others during and after the challenge. This should include the explicit listing of the license applied.

Examples:

- CC BY (Attribution)
- CC BY-SA (Attribution-ShareAlike)
- CC BY-ND (Attribution-NoDerivs)
- CC BY-NC (Attribution-NonCommercial)
- CC BY-NC-SA (Attribution-NonCommercial-ShareAlike)
- CC BY-NC-ND (Attribution-NonCommercial-NoDerivs)

Please note that the data license should not differ among sources. In case a license has to be changed, it has to be reported to the MICCAI challenges team and changed in the proposal.

CC BY-NC-SA (Attribution-NonCommercial-ShareAlike)

Code availability

a) Provide information on the accessibility of the organizers' evaluation software (e.g. code to produce rankings). Preferably, provide a link to the code and add information on the supported platforms.

Evaluation code will be released publicly on GitHub prior to challenge start.

b) In an analogous manner, provide information on the accessibility of the participating teams' code.

Not required but encouraged.

Conflicts of interest

Provide information related to conflicts of interest. In particular provide information related to sponsoring/funding of the challenge. Also, state explicitly who had/will have access to the test case labels and when.

No direct commercial sponsorship. Test labels accessible only to designated evaluation organizers.

MISSION OF THE CHALLENGE

Field(s) of application

State the main field(s) of application that the participating algorithms target.

Examples:

- Diagnosis
- Education
- Intervention assistance
- Intervention follow-up
- Intervention planning
- Prognosis
- Research
- Screening
- Training
- Cross-phase

Diagnosis,Treatment planning,Decision support

Task category(ies)

State the task category(ies)

Examples:

- Classification
- Detection

- Localization
- Modeling
- Prediction
- Reconstruction
- Registration
- Retrieval
- Segmentation
- Tracking

Localization,Segmentation,Classification,Detection

Cohorts

We distinguish between the target cohort and the challenge cohort. For example, a challenge could be designed around the task of medical instrument tracking in robotic kidney surgery. While the challenge could be based on ex vivo data obtained from a laparoscopic training environment with porcine organs (challenge cohort), the final biomedical application (i.e. robotic kidney surgery) would be targeted on real patients with certain characteristics defined by inclusion criteria such as restrictions regarding sex or age (target cohort).

a) Describe the target cohort, i.e. the subjects/objects from whom/which the data would be acquired in the final biomedical application.

Adult patients undergoing CT disease assessment in clinical radiology. The intended application is automated tumor detection, localization, and diagnosis as benign or malignant for decision support and opportunistic early cancer detection.

b) Describe the challenge cohort, i.e. the subject(s)/object(s) from whom/which the challenge data was acquired.

Abdominal, pelvic, and chest CT scans from the BodyMaps dataset collected at Istanbul Medipol University Hospital (over 16,000 CT scans) and University Hospital Basel (9,000 CT scans with reports). A total of 700 tumor segmentation masks will be created by radiologists. Evaluation uses an open internal test set and a closed external test set from UCSF.

Imaging modality(ies)

Specify the imaging technique(s) applied in the challenge.

Computed Tomography (CT)

Context information

Provide additional information given along with the images. The information may correspond ...

a) ... directly to the image data (e.g. tumor volume).

DICOM metadata (voxel spacing, protocol information, scanner details), paired radiology reports, and 700 tumor segmentation masks.

b) ... to the patient in general (e.g. sex, medical history).

Age and sex (where permitted by data governance).

Target entity(ies)

a) Describe the data origin, i.e. the region(s)/part(s) of subject(s)/object(s) from whom/which the image data would be acquired in the final biomedical application (e.g. brain shown in computed tomography (CT) data, abdomen shown in laparoscopic video data, operating room shown in video data, thorax shown in fluoroscopy video). If necessary, differentiate between target and challenge cohort.

Abdomen, pelvis, and chest in CT.

b) Describe the algorithm target, i.e. the structure(s)/subject(s)/object(s)/component(s) that the participating algorithms have been designed to focus on (e.g. tumor in the brain, tip of a medical instrument, nurse in an operating theater, catheter in a fluoroscopy scan). If necessary, differentiate between target and challenge cohort.

Algorithms must process CT scans (reports are not used during inference) and output:
(i) probability of tumor presence in each of the 13 organs,
(ii) probability of malignancy in each organ,
(iii) tumor segmentation masks,
(iv) malignant tumor segmentation masks.

Assessment aim(s)

Identify the property(ies) of the algorithms to be optimized to perform well in the challenge. If multiple properties are assessed, prioritize them (if appropriate). The properties should then be reflected in the metrics applied (see below, parameter metric(s)), and the priorities should be reflected in the ranking when combining multiple metrics that assess different properties.

- Example 1: Find highly accurate liver segmentation algorithm for CT images.
- Example 2: Find lung tumor detection algorithm with high sensitivity and specificity for mammography images.

Corresponding metrics are listed below (parameter metric(s)).

Accuracy

DATA SETS

Data source(s)

a) Specify the device(s) used to acquire the challenge data. This includes details on the device(s) used to acquire the imaging data (e.g. manufacturer) as well as information on additional devices used for performance assessment (e.g. tracking system used in a surgical setting).

Clinical CT scanners (multi-vendor; e.g., Philips, Siemens, and others depending on site).

b) Describe relevant details on the imaging process/data acquisition for each acquisition device (e.g. image acquisition protocol(s)).

Routine CT acquisition in clinical practice; reconstruction settings vary by scanner and protocol.

c) Specify the center(s)/institute(s) in which the data was acquired and/or the data providing platform/source (e.g. previous challenge). If this information is not provided (e.g. for anonymization reasons), specify why.

Contrast-enhanced CT scans acquired for multiple clinical indications with varying reconstruction settings.
Primary sources: Istanbul Medipol University Hospital and University Hospital Basel.

External validation: University of California San Francisco.
Scans acquired by trained CT technologists; reports written by board-certified radiologists; tumor masks created by board-certified radiologists.

d) Describe relevant characteristics (e.g. level of expertise) of the subjects (e.g. surgeon)/objects (e.g. robot) involved in the data acquisition process (if any).

Each case consists of one 3D CT volume. Training cases are paired with radiology reports, and 700 training cases also include tumor segmentation masks. Test cases include reports, extracted labels, and tumor masks, but during inference only the CT image is available.

Training: ~20,000 cases

Internal test: ~5,000 cases

External test (closed): ~1,000 cases (UCSF)

Train/Val: 100% paired reports; 700 training cases include tumor masks.
Test: 100% paired reports and tumor masks.

We allocate 80% of the data to training and 20% to internal testing. The external test set is fully closed.

Training and test case characteristics

a) State what is meant by one case in this challenge. A case encompasses all data that is processed to produce one result that is compared to the corresponding reference result (i.e. the desired algorithm output).

Examples:

- Training and test cases both represent a CT image of a human brain. Training cases have a weak annotation (tumor present or not and tumor volume (if any)) while the test cases are annotated with the tumor contour (if any).
- A case refers to all information that is available for one particular patient in a specific study. This information always includes the image information as specified in data source(s) (see above) and may include context information (see above). Both training and test cases are annotated with survival (binary) 5 years after (first) image was taken.

One case corresponds to a single 3D CT examination of a patient covering the chest, abdomen, and/or pelvis. Each case includes the CT volume and associated metadata. For training, each case is paired with a radiology report, and for a subset of 700 cases, tumor segmentation masks are additionally provided. For testing, each case includes radiology reports and ground-truth labels and segmentation masks for evaluation; however, during inference participants only have access to the CT volume. Each case produces one set of outputs consisting of tumor presence probabilities, malignancy probabilities, and tumor segmentation masks for all 13 target organs, which are compared against the corresponding reference annotations.

b) State the total number of training, validation and test cases.

Training: approximately 20,000 CT cases with paired radiology reports, including 700 cases with tumor segmentation masks

Internal test (open): approximately 5,000 CT cases with paired reports and segmentation masks

External test (closed, UCSF): approximately 1,000 CT cases

c) How much of the data are already annotated (stratified by train test in percentage)?

Reference segmentations and annotation consistency
Segmentation annotation and quality control were already part of VLM3D 2025, and we follow the same protocol in VLM3D 2026. Reference masks are created under a written annotation protocol that defines lesion boundaries, inclusion and exclusion criteria, and quality checks in a standardized labeling tool. An annotation calibration step is performed on a shared subset to align annotators before large-scale labeling.

Annotations are produced by board-certified radiologists under clinician-led oversight. A subset of cases is double-annotated to quantify inter-observer variability using standard overlap and boundary metrics, and disagreements are handled through clinician-led consolidation or adjudication to produce stable reference labels. We will report the measured inter-observer variability so participants can interpret segmentation performance in the context of label noise and ambiguity.

d) Explain why a total number of cases and the specific proportion of training, validation and test cases was chosen.

Similar as CT-RATE.

e) Mention further important characteristics of the training, validation and test cases (e.g. class distribution in classification tasks chosen according to real-world distribution vs. equal class distribution) and justify the choice.

Tumor detection labels are extracted from reports. Malignancy labels are extracted from pathology and clinical history. Segmentation masks were created by 31 board-certified radiologists using 3D Slicer.

f) Challenge organizers are encouraged to (partly) use unseen, unpublished data for their challenges. Describe if new data will be used for the challenge and state the number of cases along with the proportion of new data.

Reference segmentations and annotation consistency
Segmentation annotation and quality control were already part of VLM3D 2025, and we follow the same protocol in VLM3D 2026. Reference masks are created under a written annotation protocol that defines lesion boundaries, inclusion and exclusion criteria, and quality checks in a standardized labeling tool. An annotation calibration step is performed on a shared subset to align annotators before large-scale labeling.

Annotations are produced by board-certified radiologists under clinician-led oversight. A subset of cases is double-annotated to quantify inter-observer variability using standard overlap and boundary metrics, and disagreements are handled through clinician-led consolidation or adjudication to produce stable reference labels. We will report the measured inter-observer variability so participants can interpret segmentation performance in the context of label noise and ambiguity.

Annotation characteristics

a) Describe the method for determining the reference annotation, i.e. the desired algorithm output. Provide the information separately for the training, validation and test cases if necessary. Possible methods include manual image

annotation, in silico ground truth generation and annotation by automatic methods.

If human annotation was involved, state the number of annotators.

Tumor detection labels are extracted from reports. Malignancy labels are extracted from pathology and clinical history. Segmentation masks were created by 31 board-certified radiologists using 3D Slicer.

b) Provide the instructions given to the annotators (if any) prior to the annotation. This may include description of a training phase with the software. Provide the information separately for the training, validation and test cases if necessary. Preferably, provide a link to the annotation protocol.

N/A

c) Provide details on the subject(s)/algorithm(s) that annotated the cases (e.g. information on level of expertise such as number of years of professional experience, medically-trained or not). Provide the information separately for the training, validation and test cases if necessary.

Clinical reports written by board-certified radiologists.

d) Describe the method(s) used to merge multiple annotations for one case (if any). Provide the information separately for the training, validation and test cases if necessary.

Not applicable (single clinical report per case); if multiple readings exist, clinician adjudication is applied.

Data pre-processing method(s)

Describe the method(s) used for pre-processing the raw training data before it is provided to the participating teams. Provide the information separately for the training, validation and test cases if necessary.

Images are released in a standardized format; minimal preprocessing is applied beyond de-identification and consistent resampling/formatting required for evaluation. All necessary metadata (e.g., spacing) is provided to support participant preprocessing choices.

Sources of error

a) Describe the most relevant possible error sources related to the image annotation. If possible, estimate the magnitude (range) of these errors, using inter-and intra-annotator variability, for example. Provide the information separately for the training, validation and test cases, if necessary.

Variability in clinical reporting style, inter-reader variability, and differences in acquisition protocols across scanners/sites. For derived labels from text, errors may arise from label extraction; this is mitigated via validation and clinician review.

b) In an analogous manner, describe and quantify other relevant sources of error.

Domain shift between institutions (scanner vendor, protocol, population), and distribution shift over time. This is explicitly measured through external validation.

ASSESSMENT METHODS

Metric(s)

a) Define the metric(s) to assess a property of an algorithm. These metrics should reflect the desired algorithm properties described in assessment aim(s) (see above). State which metric(s) were used to compute the ranking(s) (if any).

- Example 1: Dice Similarity Coefficient (DSC)
- Example 2: Area under curve (AUC)

The challenge evaluates four complementary aspects of multi-tumor analysis.

Tumor detection (image-level, per organ):
Sensitivity, Specificity, F1-Score, Area Under the ROC Curve (AUC)

Malignancy diagnosis (image-level, per organ):
Sensitivity, Specificity, F1-Score, Area Under the ROC Curve (AUC)

Tumor localization (lesion-level):
Localization-adjusted sensitivity as defined in Lee et al., npj Digital Medicine 2022.

Tumor segmentation (voxel-level, per organ):
Dice Similarity Coefficient (DSC), Normalized Surface Dice (NSD), 95th percentile Hausdorff Distance (HD95)

The final ranking will be computed using a weighted aggregation of the primary metrics across these four evaluation categories.

b) Justify why the metric(s) was/were chosen, preferably with reference to the biomedical application.

Early cancer detection requires algorithms that are not only accurate but also clinically reliable across different tumor types and organs. Therefore, tumor detection and malignancy diagnosis are evaluated using sensitivity, specificity, F1-Score, and AUC, which together quantify the trade-off between missing true cancers and generating false alarms, a critical consideration in clinical screening and decision support.

Tumor localization is assessed using localization-adjusted sensitivity, which explicitly measures whether the algorithm detects true tumor lesions rather than exploiting dataset biases, and is therefore well aligned with the goal of robust lesion detection in real-world settings.

Tumor segmentation quality is evaluated using Dice Similarity Coefficient, Normalized Surface Dice, and HD95, which are standard metrics in medical image analysis that capture both volumetric overlap and boundary accuracy. These metrics are directly linked to downstream clinical usability, such as accurate tumor burden estimation and therapy planning.

Ranking method(s)

a) Describe the method used to compute a performance rank for all submitted algorithms based on the generated metric results on the test cases. Typically the text will describe how results obtained per case and metric are aggregated to arrive at a final score/ranking. Ideally, provide the ranking scheme as a concrete pseudo code.

We use a robust point-based ranking similar to prior MICCAI challenges: for each metric, we compute case-level scores, perform pairwise comparisons between teams using a two-sided permutation test, and award points for statistically significant wins. Final rank is determined by total points aggregated across primary metrics (with predefined weights per subtask).

b) Describe the method(s) used to manage submissions with missing results on test cases.

We use a robust point-based ranking similar to prior MICCAI challenges: for each metric, we compute case-level scores, perform pairwise comparisons between teams using a two-sided permutation test, and award points for statistically significant wins. Final rank is determined by total points aggregated across primary metrics (with predefined weights per subtask).

c) Justify why the described ranking scheme(s) was/were used.

Point-based ranking with significance testing is more stable than simple averaging because it accounts for variability across cases and reduces sensitivity to outliers, while providing a fair comparison when performance differences are small.

Statistical analyses

Provide an overview of the statistical approaches used in the scope of the challenge analysis. Details can be provided in the parameters below. For each parameter, justify why the described statistical method(s) was/were used and, if necessary, add a description of any method used to assess whether the data met the assumptions required for the particular statistical approach.

We report per-case metric distributions, compute confidence intervals via bootstrap, and assess statistical significance between methods using permutation testing. External validation performance is analyzed separately to quantify generalization.

Provide a description of how the precision of the performance estimates of individual algorithms is assessed (e.g. confidence interval of the mean on the test set computed using percentile bootstrap, confidence interval of the accuracy on the test set computed using percentile bootstrap).

95% confidence intervals computed using percentile bootstrap over test cases (e.g., 1,000 bootstrap samples) for primary metrics.

Provide a description of how variability of the performance of individual algorithms across tests cases is assessed (e.g. SD across test cases, IQR, graphs, reporting outliers...).

We report SD/IQR across cases and provide boxplots/violin plots; we also highlight outliers and stratify by acquisition site (internal vs external).

Provide a description of how variability of rankings is assessed.

We assess ranking robustness via bootstrap resampling of test cases and report the distribution of ranks per team (rank stability).

Provide a description of statistical tests that are used to assess whether the differences in performance between algorithms are statistically significant.

Two-sided permutation tests for pairwise method comparisons on case-level metric values, with multiple-comparison correction where appropriate.

Provide a description of the missing data handling.

Missing predictions are penalized as described (lowest score/invalid for that subtask). Partial submissions are ranked only on submitted subtasks if the rules allow; otherwise marked invalid.

Indicate any software product that is used for all data analysis methods.

Python (NumPy, SciPy, scikit-learn), with evaluation scripts released by organizers.

Further analyses

Present further analyses to be performed (if applicable), e.g. related to

- combining algorithms via ensembling,
- inter-algorithm variability,
- common problems/biases of the submitted methods, or
- ranking variability.

We will analyze (i) internal vs external generalization gaps, (ii) failure modes by pathology/protocol, (iii) clinical validation outcomes for top-3 methods, and (iv) potential ensembling of top methods.

ADDITIONAL POINTS

References

Please include any reference important for the challenge design, for example publications on the data, the annotation process or the chosen metrics as well as DOIs referring to data or code.

N/A

Further comments

Further comments from the organizers.

We have contacted the organizers of the related proposal “HEADLINE: A Benchmark for Vision Language Models in Head CT Reporting” and have already had an initial discussion about running a shared challenge event structure. We are aligned on exploring a coordinated program that leverages expertise across teams while keeping the rules and leaderboards clear to participants. We will also ensure diversity, equity, and inclusion considerations in the merged organizing structure and speaker selection.

To improve transparency and interpretability, each submission must include a brief resource report: GPU type(s), approximate training GPU-hours, model parameter count, inference runtime per case on the evaluation server, peak GPU memory, and peak CPU memory. Submissions will be executed under practical inference constraints to ensure feasibility and consistent evaluation across teams. Methods exceeding the enforced runtime or memory limits will be considered invalid for the affected subtask.

Evaluation sets and leaderboards
Internal and external test sets are not pooled.

Official internal leaderboard
The official ranking is computed on the closed internal test set only. We will publish separate official rankings for each track, including the chest CT track, the brain MRI track, and the tumor segmentation track (where applicable).

Mandatory external generalization leaderboard
All submissions will also be evaluated on a closed Boston University external test set (for the relevant tracks). We will publish a separate mandatory external leaderboard for every team and report internal-to-external

generalization gaps to quantify robustness under dataset shift. External performance is reported side-by-side with internal performance and is not merged into the official internal score.

Task-level leaderboards

For transparency, each subtask will also have its own leaderboard, enabling participants to compare strengths across report generation, classification, localization and segmentation, and text-conditional synthesis.