AMOS-MM: Abdominal Multimodal Analysis Challenge: Structured description of the challenge design

CHALLENGE ORGANIZATION

Title

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

AMOS-MM: Abdominal Multimodal Analysis Challenge

Challenge acronym

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

AMOS-MM

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.

Medical image analysis technology has played a significant role in enhancing the diagnostic and therapeutic capabilities of healthcare professionals, particularly in the area of abdominal CT image analysis. AI-driven localization and detection algorithms have significantly improved the efficiency of medical workflows and the accuracy of diagnoses. Clinicians can accurately analyse key indicators in images, such as the location and size of tumours, enabling them to develop appropriate treatment plans for patients and accurately assess disease prognosis.

Current models, while showing significant progress, are still limited to modes that only process image data. These methods do not effectively integrate other important clinical information, such as textual records and patient history, which limits a comprehensive understanding of cases. Large-scale Language Vision Models (LLVMs) have proven to be effective in handling tasks beyond those for which it was specifically designed.

However, the lack of flexibility and adaptability is no longer a limitation due to the rapid development of large-scale language vision models (LLVMs) [0,1] in recent years. These models creatively address a variety of real-world problems by demonstrating zero-shot generalization capabilities through the use of natural language queries. While recent LLVMs have made significant progress in processing complex, open-ended visual queries, they still fall short when it comes to understanding and interacting with biomedical images. This is partly due to a lack of high-quality datasets and a lack of adapting for medical fields.

The AMOS-MM: Abdominal Multimodal Analysis Challenge was launched to address the specific challenge of creating such a system, a unified knowledge framework that bridges vision and language in the domain of abdominal CT scans. Our high-quality image-text dataset integrates natural language processing and computer

vision technologies to enable deep understanding of medical images and natural language responses to diverse clinical queries.

The AMOS-MM Challenge is an extension of the AMOS22 Challenge (https://zenodo.org/records/6361922), which aimed to promote accurate segmentation of 15 different organs under multimodal conditions. Building on this, we have introduced two basic vision language tasks: medical report generation and medical visual question answering (VQA). The medical report generation task requires participants to automatically extract key medical information from CT images and generate comprehensive (focusing on the impression section), yet easy-to-understand medical reports. The medical VQA task requires models to integrate external knowledge and answer clinical questions based on CT images.

Building on the foundation of AMOS22, we have expanded our dataset to 2300 scans (training/validation/testing), with each case accompanied by comprehensive clinical narratives in both English and Chinese, and tailored VQA (Visual Question Answering) queries. To the best of our knowledge, this is the first publicly available 3D CT multimodal vision-text dataset and aims to push the boundaries of multimodal medical image analysis.

In summary, AMOS-MM offers the following remarkable features

- The first 3D CT multimodal image-text database: Containing 2300 cases, it supports voxel-based medical reporting and VQA tasks.
- Innovative dual tasks:
- Automated medical reporting: Provides 2300 image-text pairs for medical reports, bilingual English and Chinese versions; this task specifically targets the synthesis of the 'impression section' of medical reports, a crucial part that summarises key findings and diagnostic interpretations. The quality of the reports produced is assessed using BLEU [2], Rouge-L [3] and METEOR [4] scores, ensuring a comprehensive assessment of linguistic accuracy and coherence, which are essential for reliable medical documentation.
- Medical visual question answering: Based on 2300 cases, closed-ended, multiple-choice questions are provided, divided into eight key skill areas, categorised as 'Imaging Perception' and 'Clinical Reasoning'. The Imaging Perception category includes Anatomical Identification, Abnormality Detection, Sign Recognition and Quantitative Assessment, focusing on the fundamental aspects of medical image analysis. The Clinical Reasoning category includes Disease Diagnosis, Etiological Analysis, Therapeutic Suggestion, and Risk Assessment, addressing higher-level clinical insight and decision making based on imaging. This comprehensive scope is designed to cover essential aspects of medical image analysis and clinical application, with the accuracy of responses as the primary metric for evaluation.

[0] OpenAI. (2023). ChatGPT (Mar 14 version) [Large language model]. https://chat.openai.com/chat
[1] Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2023). Visual instruction tuning. arXiv preprint arXiv:2304.08485.
[2] Lavie, Alon, and Michael J. Denkowski. "The METEOR metric for automatic evaluation of machine translation." Machine translation 23 (2009): 105-115.
[3] Papineni, Kishore, et al. "Bleu: a method for automatic evaluation of machine translation." Proceedings of the 40th annual meeting of the Association for Computational Linguistics. 2002.
[4] Lin, Chin-Yew. "Rouge: A package for automatic evaluation of summaries." Text summarization branches out. 2004.

Challenge keywords

List the primary keywords that characterize the challenge.challenge_

Multi-Modality, Languance Vision Model

Year

The challenge will take place in 2024

FURTHER INFORMATION FOR CONFERENCE ORGANIZERS

Workshop

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

NA

Duration

How long does the challenge take?

Half day.

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.

AMOS22 attracted over 1,000 team registrations, but limited by the subsequent two-stage validation, only 15 teams completed their final submissions. For AMOS-MM we have decided to move away from an offline validation design and expect 20-30 teams to complete their final submissions.

Publication and future plans

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

We plan to summarize the results and submit them to TMI or MIA.

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.

No specific requirements

TASK 1: Medical Report Generation

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.

In the rapidly evolving field of medical imaging, the generation of accurate and comprehensive medical reports from imaging data remains a paramount yet challenging task. The Medical Report Generation task of the AMOS-MM Challenge aims to catalyse progress in this critical area. This challenge focuses on the development of AI-driven tools that can automatically interpret abdominal CT scans and generate detailed, clinically relevant reports.

We present a unique dataset, unparalleled in its scope and detail, comprising over 2300 abdominal CT scans, each accompanied by a corresponding expert-validated medical report. These reports are meticulously produced in both English and Chinese, also providing a rich bilingual resource. The dataset spans two clinical sites and includes a wide range of patient demographics and pathologies, ensuring broad applicability and generalizability of the tools developed.

Participants in this challenge are invited to apply their algorithms to this dataset with the aim of generating medical reports that are not only accurate in their representation of imaging findings, but also coherent and understandable from a clinical perspective. The reports generated by the participating algorithms will be evaluated based on their fidelity to expert-validated reports (foucs on impression section), with particular emphasis on accuracy of medical information, clarity of language and overall coherence. The performance evaluation will employ BLEU [0], Rouge-L [1], and METEOR [2] score metrics, which are standard benchmarks in natural language processing, to quantitatively assess the correspondence between the generated reports and the gold-standard references.

[0] Lavie, Alon, and Michael J. Denkowski. "The METEOR metric for automatic evaluation of machine translation." Machine translation 23 (2009): 105-115.
[1] Papineni, Kishore, et al. "Bleu: a method for automatic evaluation of machine translation." Proceedings of the 40th annual meeting of the Association for Computational Linguistics. 2002.
[2] Lin, Chin-Yew. "Rouge: A package for automatic evaluation of summaries." Text summarization branches out. 2004.

The ultimate goal of this task is to bridge the gap between complex imaging data and actionable medical insights. Successful algorithms will demonstrate the ability to distill nuanced imaging findings into clear, concise reports that aid clinicians in diagnosis and treatment planning. This is not only a significant technical challenge, but also promises to have a profound impact on clinical practice by improving the efficiency and accuracy of medical reporting in radiology.

Keywords

List the primary keywords that characterize the task.

Multi-Modality, Medical Report Generation

ORGANIZATION

Organizers

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

Yuanfeng Ji ; The University of Hongkong
Chongjian Ge ; The University of Hongkong
Ruijiang Li; Stanford University
Ping Luo; The University of Hongkong

b) Provide information on the primary contact person.

Yuanfeng Ji; u30008013@connect.hku.hk

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

Open call (used data will always be available to the community after the competition is over)

Challenge venue and platform

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

This challenge will be hosted as part of a half-day thematic event on Abdominal analysis, together with two additional challenges:
24: CURVAS: Calibration and Uncertainty for multiRater Volume Assessment in multiorgan Segmentation (Meritxell Riera-Marin)
101: Abdominal Circumference Operator-agnostic UltraSound measurement in Low-Income Countries using Artificial Intelligence (María-Sofía Sappia)

During this event, each challenge will be assigned a timeslot for each organizer to present an overview of the task(s) and dataset involved, and for top-performing teams to present their solutions. Additionally, an opening and closing session hosted by the organizers of all three challenges will be held. The proposed program for this event is as follows:

00:00 - 00:10: Opening session (all challenges)

00:10 - 01:10: ACOUSLIC-AI
01:10 - 01:30: Break
01:30 - 02:30: CURVAS
02:30 - 02:50: Break
02:50 - 03:50: AMOS-MM
03:50 - 04:00: Closing session (all challenges)

b) Report the platform (e.g. grand-challenge.org) used to run the challenge.

CodaLab https://competitions.codalab.org/

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

NA, will be publiced later

Participation policies

a) Define the allowed user interaction of the algorithms assessed (e.g. only (semi-) automatic methods allowed).

only automatic method allowed

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 to publicly available data including (open) pre-trained nets.

Publicly available data is allowed. However, teams using additional datasets need to publish/share out the additional data during competition

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.

Can participate and be listed in the leaderboard. The final data used in the competition is only accessible to Yuanfeng Ji and Chongjian Ge.

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

Prizes will be awarded to the top three teams, the top five teams will be invited to co-author a paper, and cash prizes are being sought.

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 10 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).

Participants from each of the top five groups, with a maximum of two individuals per group, may receive an invitation to co-author a paper. Additionally, all participants are encouraged to publish their papers at their convenience.

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.

Due to the challenges posed by the storage and computational demands of the large-scale CT dataset, we will deviate from the AMOS22 docker submission process. AMOS-MM will be divided into two distinct phases: online validation and online testing. In both phases, participants will perform metric computations by submitting their prediction results to our platform (hosted in CodaLab platform). The online validation phase will become publicly accessible following the release of the training data. As for the online testing phase, it will open to participants 48 hours after final phase started. During this phase, teams are required to complete their analysis and submit their results within the specified deadline.

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.

Results based on last submission

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

Training Data Release & Registration Opening:
Date: May 1, 2024
Release of training cases with URL on the official challenge website.
Registration for the challenge opens on the same day.

Online Validation Phase:
Start Date: June 1, 2024
Following the release of the training data, this phase opens for participants to begin submitting their results for validation.

Online Test Phase & Submission of Final Results:
Start Date: September 1, 2024
Duration: 48 hours
This critical phase starts 48 hours after the release of the test cases. Teams must complete their analysis and submit the results within this timeframe.

Release of Challenge Results:
Date: September 10, 2024
Final challenge results will be announced, a month prior to MICCAI 2024.

MICCAI 2024 Conference:
Date: October 6-10, 2024
The conference will possibly include associated challenge days, details of which will be confirmed closer to the event.

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).

The data utilized for this challenge has already undergone and received ethics approval during the AMOS22 project.

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)

We follow the CC BY-SA (Attribution-ShareAlike) adopted by AMOS22.

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.

We plan to simultaneously release the evaluation code alongside the training data using the GitHub platform. This will provide clear guidance on the input and output formats required for the challenge.

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

We intend to systematically gather the codes submitted by participants and, with their consent, make them available for communal sharing and access.

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.

Funding for the data labeling was secured through academic grants from the University of Hong Kong. Exclusive access to the labels, which were critical for the final ranking calculations, was limited to Yuanfengji and Chongjiange.

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, Education

Task category(ies)

State the task category(ies)

Examples:

- Classification
- Detection
- Localization

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

Prediction

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.

The target cohort for our task comprises patients who have undergone CT scans at two distinct hospitals. This diverse group has been carefully selected to ensure a wide representation in the final biomedical application. The data from these patients were captured using five different imaging devices, offering a variety of imaging conditions and technical specifications. This selection aims to reflect the real-world variability and complexity encountered in clinical settings, thereby enhancing the applicability and robustness of the participating algorithms. The inclusion criteria for this cohort do not specify restrictions based on sex or age, allowing for a comprehensive representation of the general patient population typically undergoing CT scans for various medical reasons.

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

The challenge cohort consists of a meticulously curated collection of data from 2300 patients. This diverse set encompasses a broad range of ages and includes both male and female subjects. The patients represented in this cohort cover a wide array of patient types, ensuring a natural and realistic distribution reflective of the general patient population encountered in clinical settings.

Imaging modality(ies)

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

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).

NA

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

We will provide essential metadata, including de-identified information on gender and age, to accompany the dataset.

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.

The image data originates from positive cases (patients diagnosed with diseases) from two hospitals. This data specifically includes 2300 CT scans of these patients, encompassing a wide array of medical conditions.

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.

The primary focus of the algorithms in this task is to analyze and interpret CT scan images to identify key medical findings. These include but are not limited to the detection and characterization of tumors, lesions, and other significant abnormalities within various organs captured in the abdominal CT scans. The algorithms are designed to synthesize this information into comprehensive medical reports that accurately reflect the patient's condition as depicted in the scans. In AMOS-MM, the task is primarily on generating the 'impression' section of medical reports.

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)).

The ability of the algorithm to correctly identify and interpret medical information from CT scans is paramount. This includes accurate detection and description of abnormalities, lesions, or other significant findings.

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).

Aquilion ONE; Brilliance16; Somatom Force; Optima CT660; Optima CT540

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

The imaging data for this project was acquired following standard, routine clinical protocols, which are commonly employed in everyday medical practice. Despite the use of various imaging devices across different hospitals, each device adhered to these standard protocols.

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.

Longgang Central Hospital of Shenzhen, China;
Longgang People Hospital of Shenzhen, China;

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).

The data acquisition process was carried out by qualified radiology personnel who have undergone rigorous training and passed necessary examinations to be officially employed in their roles.

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.

A case in this task refers to a set of abdominal CT scan images for a single patient. Each case includes the CT images along with any relevant clinical information that has been de-identified. The desired algorithm output for each case is a comprehensive medical report that accurately reflects the findings observed in the CT images. This report should include key details such as the presence, location, and characteristics of any abnormalities or noteworthy features.

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

Training Cases: 1500
Validation Cases: 300
Test Cases: 500

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

The choice of the total number of cases and the specific proportion of training (65%), validation (13%), and test cases (22%) was made to strike a balance between effective algorithm training, rigorous evaluation.

d) 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 further characterized by strict gender and age stratification during data partitioning to align with real-world distribution.

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.

To establish reference annotations for the medical report generation task in the AMOS-MM Challenge, we first collected corresponding abdominal CT images and their medical reports from a hospital, ensuring all personal information was anonymized. Next, three junior radiologists were tasked with translating these reports from Chinese to English, focusing on maintaining the accuracy of medical terminology and the integrity of the original reports. Finally, two senior radiologists reviewed and corrected these translations, ensuring the translated reports were accurate both medically and linguistically. This process not only provides high-quality, expert-verified reference annotations for training the algorithm, but also ensures that the model's output is clinically relevant and precise.

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.

For the first task in the AMOS-MM Challenge, annotators were given specific instructions to ensure precise and standard translation of medical terms from Chinese to English. This included a focus on maintaining the integrity and accuracy of the medical information, with special emphasis on correctly translating complex medical terminologies. They were provided with training materials and examples of accurately translated reports to understand the expected standards. The annotators were also briefed on the importance of confidentiality and ethical handling of patient data.

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.

For the first task of the AMOS-MM Challenge, the annotation was meticulously handled by a team of medically trained professionals, comprising three junior radiologists and two senior radiologists. The junior radiologists, each with an average of 3 years of professional experience and equipped with medical school training and valid practicing licenses, were primarily responsible for the initial translation of medical reports. The two senior radiologists, bringing an average of 8 years of experience and similar medical qualifications, played a crucial role in reviewing and correcting these translations.

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.

N/A

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.

Personally identifiable information has been carefully removed to ensure privacy and security.

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.

The main sources of error in image annotation stemmed from the translation process conducted by junior radiologists. These errors primarily included translation inaccuracies and occasional omissions of content. Although quantifying the exact range of these errors is challenging, they were substantially addressed and corrected in the review phase by senior radiologists, whose expertise significantly reduced the risk of inaccuracies

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

Reviewer errors

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)

To assess the performance of algorithms in medical report generation, three metrics are commonly used: BLEU, for linguistic accuracy; Rouge-L, focusing on recall by evaluating the longest common subsequence with reference texts; and METEOR, aligning both precision and recall, considering synonyms and grammar. For ranking the algorithms, the average values of these three metrics are employed, providing a balanced evaluation that encompasses various aspects of an algorithm's output, including accuracy, coherence, and overall quality. This method ensures a comprehensive assessment of each algorithm's capability to generate medically accurate and linguistically coherent reports.

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

The choice of metrics (BLEU, Rouge-L, METEOR) for the first task in the AMOS-MM Challenge is justified by their widespread use in evaluating text generation tasks, including their application in previous studies like MIMIC-CXR. While these metrics have limitations in fully capturing clinical correctness or the depth of medical insights in the generated reports, evaluating clinical accuracy is inherently complex. To balance the challenge's scope and feasibility, these metrics are considered sufficient. They provide a reliable measure of linguistic alignment with ground truth texts, which is crucial in a biomedical context where accuracy and coherence of reports are paramount.

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.

For the task1, the method involves aggregating results across test cases using the selected metrics. Specifically, the average value for each metric (BLEU, Rouge-L, METEOR) is calculated across all cases. Then, a ranking is assigned to each algorithm based on these average values for each metric. The ranks for each metric are summed up, and the algorithm with the lowest total rank across all three metrics is deemed the winner.

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

If any submission fails to provide a result for a specific case, the metric for that case will be set to zero.

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

For the task1, the method involves aggregating results across test cases using the selected metrics. Specifically, the average value for each metric (BLEU, Rouge-L, METEOR) is calculated across all cases. Then, a ranking is assigned to each algorithm based on these average values for each metric. The ranks for each metric are summed up, and the algorithm with the lowest total rank across all three metrics is deemed the winner.

Statistical analyses

a) Provide details for the statistical methods used in the scope of the challenge analysis. This may include

- description of the missing data handling,
- details about the assessment of variability of rankings,
- description of any method used to assess whether the data met the assumptions, required for the particular statistical approach, or
- indication of any software product that was used for all data analysis methods.

In the AMOS-MM Challenge, our statistical analysis approach involves excluding participants who don't report on the entire test set. We use mean and standard deviation for metrics used. Significance testing is conducted with the Wilcoxon test, using p-values to determine if top-performing algorithms are significantly better than others. Additionally, we assess variability using variance, squared deviation, average absolute deviation, and inter-quartile range. This comprehensive methodology ensures a robust and detailed evaluation of the algorithms' performances.

b) Justify why the described statistical method(s) was/were used.

The mean value of metrices produced by the algorithms are indicators of their overall performance. The standard deviation measures the performance stability of the algorithms.

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.

N/A

TASK 2: Medical Visual Question Answering

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.

The Medical Visual Question Answering (VQA) task, a cornerstone of the AMOS-MM Challenge, ventures into uncharted territory by combining computer vision, natural language processing, and clinical acumen. This task is specifically aimed at advancing AI's proficiency in interpreting and responding to complex clinical inquiries based on abdominal CT scans.

We are proud to introduce a unique dataset featuring 2300 abdominal CT scans. Each scan is paired with a series of visual questions, carefully curated to represent a broad spectrum of clinical situations. These questions are designed to test the algorithms' ability to understand visual content and provide precise, contextually appropriate answers.

The questions fall into two primary categories, each addressing a critical aspect of medical imaging analysis:

Imaging Perception:

1. Anatomical Identification: Focusing on identifying and naming organs, bones, or anatomical structures in the image.
2. Abnormality Detection: Aimed at spotting visible abnormalities such as tumors, fractures, or lesions, without delving into their clinical implications.
3. Sign Recognition: Concentrating on identifying patterns within the image, like tissue textures, fluid distributions, or specific radiological signs.
4. Quantitative Assessment: Involving measurement or quantification of visible features, such as organ size, fracture length, or degree of a noticeable change.

Clinical Reasoning

1. Disease Diagnosis: This category is about identifying specific diseases or medical conditions from the imaging. It involves analyzing image features to diagnose conditions accurately, based on their unique radiological signatures.
2. Etiological Analysis: This involves understanding the cause behind a visible abnormality or condition in the image. It's about linking radiological findings to their underlying pathological processes.
3. Therapeutic Suggestion: Here, the focus is on recommending appropriate treatments or interventions based on the diagnosed condition. This requires knowledge of treatment protocols and their suitability in different medical scenarios as suggested by the imaging.
4. Risk Evaluation: This category is about assessing potential risks or complications associated with the medical condition observed in the imaging. It involves evaluating the image to predict adverse events or complications

that might arise from the condition.

Participants are challenged to develop algorithms that not only bridge the gap between visual data and textual inquiries but also embody a deep understanding of both the visual and clinical aspects. The algorithms' performance will be evaluated on their accuracy in answering questions that draw upon both the visual information in CT scans and integrated medical knowledge.

Keywords

List the primary keywords that characterize the task.

Multi-Modality, Medical VQA

ORGANIZATION

Organizers

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

Yuanfeng Ji ; The University of Hongkong
Chongjian Ge ; The University of Hongkong
Ruijiang Li; Stanford University
Ping Luo; The University of Hongkong

b) Provide information on the primary contact person.

Yuanfeng Ji; u30008013@connect.hku.hk

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

Open call (used data will always be available to the community after the competition is over)

Challenge venue and platform

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

This challenge will be hosted as part of a half-day thematic event on Abdominal analysis, together with two additional challenges:
24: CURVAS: Calibration and Uncertainty for multiRater Volume Assessment in multiorgan Segmentation (Meritxell Riera-Marin)
101: Abdominal Circumference Operator-agnostic UltraSound measurement in Low-Income Countries using

Artificial Intelligence (María-Sofía Sappia)

During this event, each challenge will be assigned a timeslot for each organizer to present an overview of the task(s) and dataset involved, and for top-performing teams to present their solutions. Additionally, an opening and closing session hosted by the organizers of all three challenges will be held. The proposed program for this event is as follows:

00:00 - 00:10: Opening session (all challenges)
00:10 - 01:10: ACOUSLIC-AI
01:10 - 01:30: Break
01:30 - 02:30: CURVAS
02:30 - 02:50: Break
02:50 - 03:50: AMOS-MM
03:50 - 04:00: Closing session (all challenges)

b) Report the platform (e.g. grand-challenge.org) used to run the challenge.

CodaLab https://competitions.codalab.org/

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

NA, will be publiced later

Participation policies

a) Define the allowed user interaction of the algorithms assessed (e.g. only (semi-) automatic methods allowed).

only automatic method allowed

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 to publicly available data including (open) pre-trained nets.

Publicly available data is allowed. However, teams using additional datasets need to publish/share out the additional data during competition

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.

Can participate and be listed in the leaderboard. The final data used in the competition is only accessible to Yuanfeng Ji and Chongjian Ge.

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

Prizes will be awarded to the top three teams, the top five teams will be invited to co-author a paper, and cash prizes are being sought.

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 10 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).

Participants from each of the top five groups, with a maximum of two individuals per group, may receive an invitation to co-author a paper. Additionally, all participants are encouraged to publish their papers at their convenience.

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.

Due to the challenges posed by the storage and computational demands of the large-scale CT dataset, we will deviate from the AMOS22 docker submission process. AMOS-MM will be divided into two distinct phases: online validation and online testing. In both phases, participants will perform metric computations by submitting their prediction results to our platform (hosted in CodaLab platform). The online validation phase will become publicly accessible following the release of the training data. As for the online testing phase, it will open to participants 48 hours after final phase started. During this phase, teams are required to complete their analysis and submit their results within the specified deadline.

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.

Results based on last submission

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

Training Data Release & Registration Opening:
Date: May 1, 2024
Release of training cases with URL on the official challenge website.
Registration for the challenge opens on the same day.

Online Validation Phase:
Start Date: June 1, 2024
Following the release of the training data, this phase opens for participants to begin submitting their results for validation.

Online Test Phase & Submission of Final Results:
Start Date: September 1, 2024
Duration: 48 hours
This critical phase starts 48 hours after the release of the test cases. Teams must complete their analysis and submit the results within this timeframe.

Release of Challenge Results:
Date: September 10, 2024
Final challenge results will be announced, a month prior to MICCAI 2024.

MICCAI 2024 Conference:
Date: October 6-10, 2024
The conference will possibly include associated challenge days, details of which will be confirmed closer to the event.

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).

The data utilized for this challenge has already undergone and received ethics approval during the AMOS22 project.

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)

We follow the CC BY-SA (Attribution-ShareAlike) adopted by AMOS22.

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.

We plan to simultaneously release the evaluation code alongside the training data using the GitHub platform. This will provide clear guidance on the input and output formats required for the challenge.

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

We intend to systematically gather the codes submitted by participants and, with their consent, make them available for communal sharing and access.

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.

Funding for the data labeling was secured through academic grants from the University of Hong Kong. Exclusive access to the labels, which were critical for the final ranking calculations, was limited to Yuanfengji and Chongjiange.

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, Education

Task category(ies)

State the task category(ies)

Examples:

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

Prediction

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.

The target cohort for our task comprises patients who have undergone CT scans at two distinct hospitals. This diverse group has been carefully selected to ensure a wide representation in the final biomedical application. The data from these patients were captured using five different imaging devices, offering a variety of imaging conditions and technical specifications. This selection aims to reflect the real-world variability and complexity encountered in clinical settings, thereby enhancing the applicability and robustness of the participating algorithms. The inclusion criteria for this cohort do not specify restrictions based on sex or age, allowing for a comprehensive representation of the general patient population typically undergoing CT scans for various medical reasons.

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

The challenge cohort consists of a meticulously curated collection of data from 2300 patients. This diverse set encompasses a broad range of ages and includes both male and female subjects. The patients represented in this cohort cover a wide array of patient types, ensuring a natural and realistic distribution reflective of the general patient population encountered in clinical settings.

Imaging modality(ies)

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

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).

NA

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

We will provide essential metadata, including de-identified information on gender and age, to accompany the dataset

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.

The image data originates from positive cases (patients diagnosed with diseases) from two hospitals. This data specifically includes 2300 CT scans of these patients, encompassing a wide array of medical conditions.

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.

For the Medical VQA task, the algorithms are targeted at effectively responding to specific clinical queries based on the visual data from abdominal CT scans. The focus here is on accurately interpreting the CT images to answer questions related to anatomical structures, pathological findings, and potential diagnostic interpretations. This includes identifying and understanding the significance of various features such as organ size and shape, presence of any anomalies, and correlating these observations with potential clinical conditions.

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)).

The foremost priority is the algorithm's ability to accurately answer clinical questions based on the CT images. This means correctly interpreting the visual data and providing precise, correct answers to the posed queries.

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).

Aquilion ONE; Brilliance16; Somatom Force; Optima CT660; Optima CT540

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

The imaging data for this project was acquired following standard, routine clinical protocols, which are commonly employed in everyday medical practice. Despite the use of various imaging devices across different hospitals, each device adhered to these standard protocols.

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.

Longgang Central Hospital of Shenzhen, China;
Longgang People Hospital of Shenzhen, China;

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).

The data acquisition process was carried out by qualified radiology personnel who have undergone rigorous training and passed necessary examinations to be officially employed in their roles.

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.

In the medical VQA task, a case consists of a set of abdominal CT scan images accompanied by a series of clinical questions related to these images. These questions are designed to evaluate the algorithm's ability to accurately interpret images and provide contextually appropriate answers. The desired output for each case is a set of answers that accurately answer the question posed, demonstrating the algorithm's understanding of the medical content depicted in the CT scan. In AMOS-MM, we take the form of multiple choice questions and therefore expect the model to output one of the options

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

Training Cases: 1500
Validation Cases: 300
Test Cases: 500
Beside, for VQA, we will expect 4 questions per case

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

The choice of the total number of cases and the specific proportion of training (65%), validation (13%), and test cases (22%) was made to strike a balance between effective algorithm training, rigorous evaluation.

d) 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 further characterized by strict gender and age stratification during data partitioning to align with real-world distribution.

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.

In the Medical Visual Question Answering (VQA) task of the AMOS-MM Challenge, we aim to minimise manual annotation while ensuring rigorous verification. Using the refined medical reports from Task 1, we use GPT-4 to generate a series of questions required, with a particular focus on explaining image features. For each question, GPT-4 also provides corresponding answers and rationale. This innovative approach significantly reduces the need for extensive human annotation. The validity and clinical accuracy of these question-answer-reasoning sets are rigorously reviewed by a senior radiologist and a senior clinical physician. Only sets that pass this rigorous review and rely solely on image content for answers are selected. This process effectively combines the efficiency of AI with expert medical oversight, ensuring a high standard of accuracy with reduced manual effort.

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.

For the second task of the AMOS-MM Challenge, annotators received clear instructions emphasizing two critical aspects. First, they were to ensure that each question generated is fundamentally based on understanding the medical images. This meant filtering out any questions that could be answered without image reference, thereby maintaining the qa task relies on image-based analysis. Second, a rigorous review process was mandated for each question-answer-reasoning set. Annotators were tasked with meticulously evaluating these sets, discarding any with inaccuracies or errors in either the question, answer, or reasoning.

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.

In the second task of the AMOS-MM Challenge, GPT-4 was tasked with designing questions, answers, and reasoning processes, while a senior radiologist with 8 years of experience and a clinical doctor were responsible for rigorously reviewing these questions, ensuring their medical accuracy and relevance to the imaging data

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.

N/A

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.

Personally identifiable information has been carefully removed to ensure privacy and security.

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.

In the second task, the primary source of error was attributed to the hallucination issues of ChatGPT in generating questions, answers, and reasoning. These hallucinations refer to instances where the AI model generates incorrect or nonsensical information. Such errors, however, were effectively addressed and rectified in the subsequent verification process, ensuring the accuracy and relevance of the final outputs for the task.

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

Reviewer errors

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)

For the second task in the AMOS-MM Challenge, where the questions are in a multiple-choice format, the evaluation is straightforwardly based on accuracy. This metric measures the percentage of questions where the algorithm's selected answer matches the correct answer.

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

For the second task in the AMOS-MM Challenge, the chosen metric is accuracy, which is particularly suitable for the multiple-choice question format of this task. This metric was selected because it directly quantifies how often the algorithm correctly identifies the right answer from the given options, a clear and objective measure of performance.

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.

For the task2, the performance ranking of algorithms is determined by computing the accuracy across all test cases.

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

If any submission fails to provide a result for a specific case, the metric for that case will be set to zero.

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

For the task2, the performance ranking of algorithms is determined by computing the accuracy across all test cases.

Statistical analyses

a) Provide details for the statistical methods used in the scope of the challenge analysis. This may include

- description of the missing data handling,
- details about the assessment of variability of rankings,
- description of any method used to assess whether the data met the assumptions, required for the particular statistical approach, or
- indication of any software product that was used for all data analysis methods.

In the AMOS-MM Challenge, our statistical analysis approach involves excluding participants who don't report on the entire test set. We use mean and standard deviation for metrics used. Significance testing is conducted with the Wilcoxon test, using p-values to determine if top-performing algorithms are significantly better than others. Additionally, we assess variability using variance, squared deviation, average absolute deviation, and inter-quartile range. This comprehensive methodology ensures a robust and detailed evaluation of the algorithms' performances.

b) Justify why the described statistical method(s) was/were used.

The mean value of metrices produced by the algorithms are indicators of their overall performance. The standard deviation measures the performance stability of the algorithms.

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.

N/A

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.

N/A