The Massive Multimodal Biomedical Understanding Challenge

Ryan D'Cunha[1,2,*], Alejandro Lozano[1*], Akira Nishii[2*], Darshan Kalola[3], Maximilian Rokuss[7], Xiaoxiao Sun[1], Jeya Maria Jose V.[5], MingYu Lu[9], Cliff Wong[5], Timothy Ossowski[5], Ryan Poplin[6], Min Woo Sun[1], Josiah Aklilu[4], Anas Zafar[8], Kyle Harrington[12], Alan Lowe[12], Kavita Kulkarni[12], Shalin Mehta[12], Ryutaro Tanno[4], Natsuki Kataoka[10], Qin Liu[11], Yuhui Zhang[1], Paola Avila Robayo[1], Fiona Cai[1], and Serena Yeung-Levy[1]

[1] Stanford University, Stanford, CA, USA
[2] GXL, Palo Alto, CA, USA
[3] Anthropic, San Francisco, CA, USA
[4] Google DeepMind, Mountain View, CA, USA
[5] Microsoft AI, Redmond, WA, USA
[6] Apple, Cupertino, CA, USA
[7] University of Heidelberg, Germany
[8] The University of Texas MD Anderson Cancer Center, Houston, TX, USA
[9] University of Washington, Seattle, WA, USA
[10] Highlanders, Tokyo, Japan
[11] University of California, Davis, CA, USA
[12] Biohub, San Francisco, CA, USA

Abstract. By integrating rich visual information across diverse medical imaging modalities, anatomical scales, and disease contexts with complementary molecular, clinical, and scientific knowledge, multimodal large language models (MLLMs) hold immense potential to transform scientific discovery and precision medicine. MLLMs could, for example, jointly interpret visual and textual information to connect population-specific patient observations with the broader landscape of biomedical knowledge, thereby uncovering novel molecular phenotypes, prognostic biomarkers, and previously unrecognized associations [5].

Realizing this potential, however, requires biomedical MLLMs to possess two fundamental capabilities: perception (accurately recognizing and interpreting visual features) and cognition (integrating visual observations with prior knowledge to derive complex conclusions) [4]. While recent advances have emphasized reasoning (a subset of cognition involving structured thinking), emerging evidence suggests that **visual perception remains the primary bottleneck for reliable downstream reasoning** [1–3, 11]. How, then, can MLLMs deliver reliable value in downstream applications when they cannot reliably perceive and interpret biomedical images?

To address this problem, we launch the MMBU Challenge, a large-scale initiative to advance the visual perception capabilities of MLLMs across biomedicine and invite researchers worldwide to participate in this collaborative effort. Using a large-scale peer-reviewed biomedical dataset

curated by domain experts, we evaluate the open-ended visual question answering (VQA) capabilities of MLLMs. Specifically, the MMBU challenge pushes MLLMs to recognize, localize, and characterize visual features across diverse biomedical images spanning imaging modalities, biological scales, anatomical regions, and clinical contexts [2].

Motivation

The pace at which biomedical data are being generated is accelerating rapidly. Advances in imaging and measurement technologies are producing biomedical data faster than we can systematically analyze and synthesize it. In the healthcare industry alone, approximately 700 million medical imaging examinations are performed worldwide each year [6], alongside an estimated 6.4 billion glass slides generated annually in anatomic pathology. In modern biology workflows, electron microscopy can produce petabytes of data, while high-throughput and spatial imaging technologies can generate billions of images and measure thousands of molecular features simultaneously. This unprecedented scale and complexity presents an enormous opportunity: to uncover the vast, untapped latent knowledge **already** embedded in biomedical data that is generated and stored, yet remains largely unexplored, and transform it into scientific insight.

Artificial intelligence (AI) can help overcome this bottleneck by enabling biomedical data to be processed and analyzed at scale. GigaTIME, for example, is among the first works to demonstrate this potential, using AI to enable population-scale analysis of the tumor microenvironment and uncovering more than 1,000 statistically significant associations between protein activity and clinical attributes [9].

These advances, however, still rely heavily on human experts to interpret and validate the resulting insights. Multimodal language models (MLLMs), on the other hand, offer a compelling opportunity to break this paradigm. Rather than using models merely to classify individual images or reproduce expert performance on well-defined tasks, MLLM agents could operate over vast collections of biomedical observations: searching millions of images, comparing phenotypes across patients and experiments, connecting visual patterns to textual and molecular context, and uncovering relationships that would be difficult to discover through manual inspection alone. Offering all these benefits while providing a native natural language interface through which domain experts can inspect and validate findings [4].

Realizing this potential, however, requires MLLM to possess two fundamental capabilities: perception (the ability to recognize and interpret visual features accurately) and cognition (to integrate those perceptual observations with prior knowledge to reason and derive complex conclusions) [4]. While recent advances in MLLMs have increasingly emphasized reasoning capabilities, emerging evidence suggests that **deficiencies in visual perception remain the primary bottleneck** [1, 10]. Indeed, further analysis of frontier model behavior shows that, while these models can achieve moderate performance across a broad range of biomedical tasks, substantial room for improvement remains. In particular,

models often struggle with basic perceptual information, such as identifying in Figure 2.

To address this problem, we introduce the MMBU Challenge, a large-scale initiative to evaluate and advance the visual perception capabilities of M-LLMs systematically. The MMBU Challenge pushes models to recognize, localize, and characterize visual features across a broad spectrum of biomedical images, spanning diverse imaging modalities, biological scales, anatomical regions, and clinical contexts.

Benchmark Information

The MMBU Challenge is evaluated on the Massive Multimodal Biomedical Understanding (MMBU) benchmark [2], which spans a wide range of clinical datasets and imaging modalities. The four MMBU tasks differ in what the final answer asks for and, for object detection, in how that answer is judged. Example questions per task are show in Figure 1.

- **Ungrounded Classification:** Identifying broad pathologies, clinical conditions, or global image context. The judge scores semantic equivalence between the predicted answer and the reference answer.
- **Fine-Grained Classification (from Segmentation):** Classifying specific abnormalities or anatomical features, with segmentation masks supplying precise visual grounding. Judged by semantic equivalence.
- **Fine-Grained Classification (from Detection):** Categorizing localized clinical findings derived from object detection bounding boxes. Judged by semantic equivalence.
- **Object Detection:** Localizing key anatomical structures and clinical anomalies. The final answer must be a bounding box in the format `[x, y, width, height]` specified in the prompt, and is counted as correct only at IoU ≥ 0.5. Metadata fields are scored as in the other three tasks.

Two aspects of the challenge configuration differ from MMBU as published. First, the MMBU private held-out set contains **open-ended VQA only**; the public development set contains both open-ended and closed-ended VQA, so development-set performance will not be directly comparable to final rankings. Second, all non-commercially licensable datasets are held in the private set, which is why final scores cannot be reproduced locally and are released only after submissions close.

Tracks

Track 1: Open Frontier Performance Track

Objective Function: Best Overall Performance.

Participants are invited to push the frontier of biomedical perception by submitting their **best-performing model**. There are no restrictions on model size;

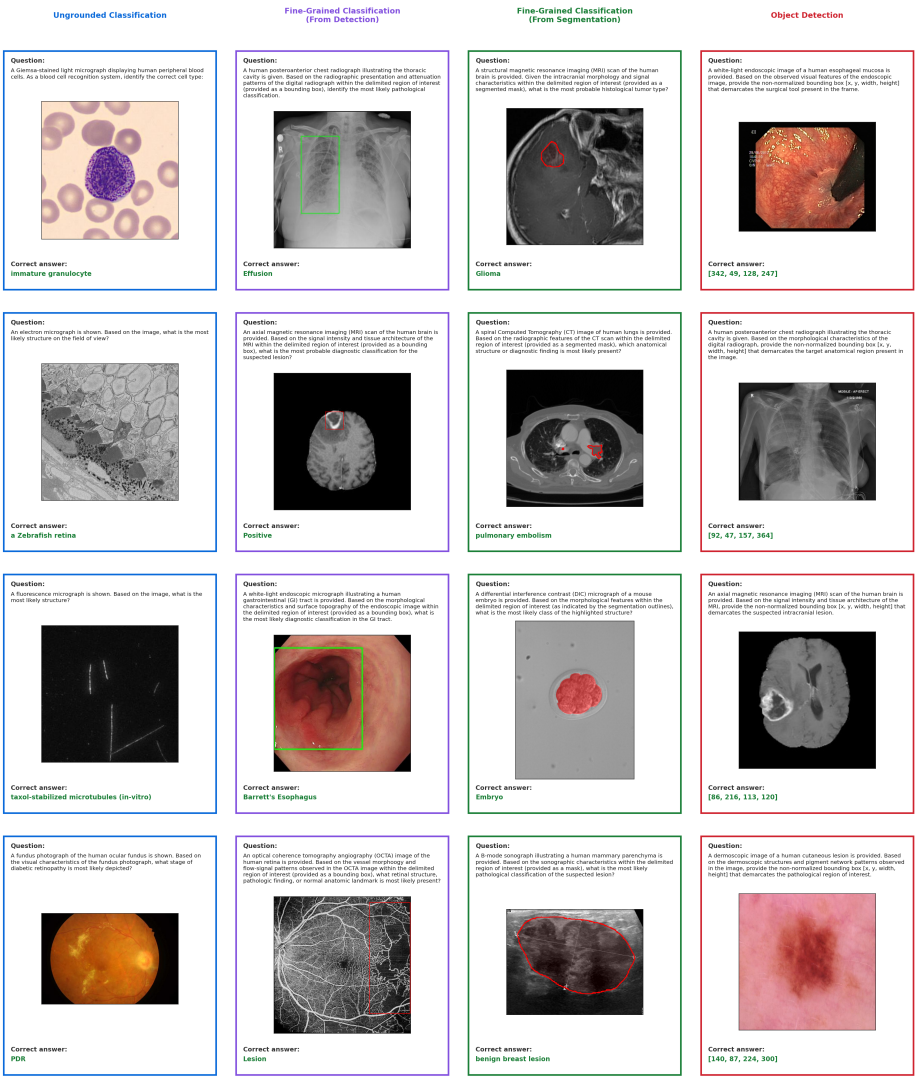

Fig. 1: Representative open-ended VQA examples from the four MMBU task types. Each panel shows the prompt, the input image, and the reference answer. Model predictions are not depicted. [2]. We recognize different clinicians may provide different annotations of findings and biomedical context within MMBU, but we put forward our best effort across a panel of clinicians to ensure quality.

however, models with more than 27B active parameters must be made available for evaluation through an **API key**.

1. **Maximum Model Size:** None

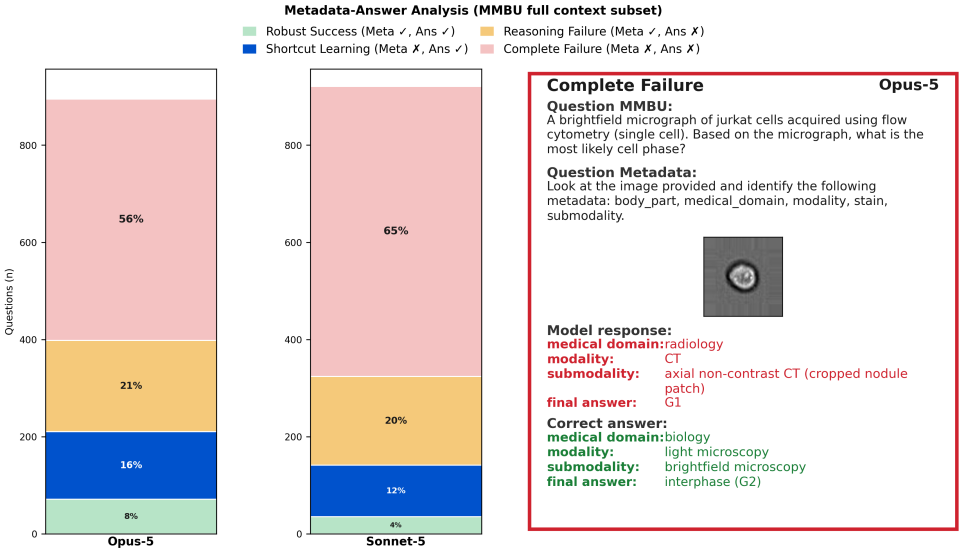

Fig. 2: Frontier models exhibit shortcut learning on biomedical VQA. Analysis of a subset of MMBU in open-ended VQA reveals that overall performance is low, but while frontier models can arrive at the correct answer, they fail to identify foundational image metadata (e.g., modality, submodality, and body part). This "shortcut learning" (Meta ×, Ans ✓) suggests a reliance on language priors or spurious correlations rather than robust visual perception. [2]. Image source: "We used image set BBBC048 Eulenberg et al. 2017, available from the Broad Bioimage Benchmark Collection [Ljosa et al., Nature Methods, 2012]."

2. **Model Restrictions:** Model must be pretrained, mid-trained, and post-trained by the same organization.

Track 2: Medical Domain Adaptation Track

Objective Function: Delta-based Adaptation Score.

Participants are invited to specifically adapt a pre-trained model towards the medical domain through mid-training or post-training. There are no restrictions on model size; however, models with more than 27B active parameters must be made available for evaluation through an **API key**.

1. **Maximum Model Size:** None. However, API access must be provided for models with more than 27B active parameters
2. **Model Restrictions:** The model must be adapted from a base model, which must be disclosed.

Track 3: Domain Adaptation and Efficient Inference Track

Objective Function: Maximum perception with minimal computation.

Participants are invited to either pre-train or adapt a small biomedical M-LLM with $\leq$ 4B active parameters. This track focuses on developing parameter-efficient models that maintain strong biomedical perception performance while minimizing computational and model-size requirements. To qualify for an award in this track, participants must disclose the number of active parameters in their model.

1. **Maximum Model Size:** $\leq$ 4B total **active** parameters and $\leq$ 12B total parameters (e.g., for MoE).
2. **Model Restrictions:** If the model is adapted from a base model, the base model must be disclosed.

Scoring

Every submission receives two scores, computed over the same N examples in the hidden evaluation set and reported separately on the leaderboard alongside the final track score:

Task Score (T) measures **task correctness**: whether the model produces the right answer to the clinical or scientific question posed by the example.

Biomedical Context Score (C) measures **biomedical context understanding**: whether the model can identify *what it is looking at* independently of the question it was asked. Concretely, C is computed over six metadata fields: *modality*, *submodality*, *specimen*, *body part*, *stain*, and *medical domain*. Not every field applies to every image, so each example is scored only on the fields annotated for it.

Every MMBU Challenge example is evaluated in **two independent inference passes**, and this separation is what makes the two scoring components measurable without confounding.

- **Pass 1 — Task.** The model receives the image and the full-context MMBU question, and returns a JSON object containing a final answer and a rationale. The LLM judge scores the final answer for semantic equivalence with the reference answer, contributing to the Task Score T.
- **Pass 2 — Biomedical context.** The model receives the same image and is asked to populate the metadata fields annotated for that example — modality, submodality, specimen, body part, stain, and medical domain — again in JSON, with a rationale. The judge scores each field, contributing to the Biomedical Context Score C.

The two passes are run independently: neither is conditioned on the other's output, and no state carries between them. A model therefore cannot infer the imaging modality from a clinical answer it has already committed to, nor steer its answer using metadata it has just enumerated. This is what allows T and C to be read jointly as evidence of shortcut learning — a model may answer

correctly in Pass 1 while failing in Pass 2 to identify what it was looking at, the behavior shown in Fig. 2 that this challenge targets. Malformed or unparsable JSON receives no credit for the affected example in the pass in which it occurs; because the passes are independent, a failure in one does not forfeit credit in the other.

LLM judge. Both scores are assigned by an LLM judge: for T, by determining semantic equivalence between the predicted and reference answers, and for C, by scoring each annotated metadata field returned in the metadata pass. We evaluated Sonnet-5, Gemma-4, Qwen-3, and Qwen-3.5 as judges on a stratified sample of 1,000 MMBU examples spanning all four task types [7, 8, 12]. Mean pairwise inter-judge agreement was near-perfect (Cohen's $\kappa = 0.96$), indicating that the four judges are effectively interchangeable for the scoring decisions required by this challenge. The choice among them is not consequential for scoring. Sonnet-5 is used for challenge scoring on the basis of throughput and availability at evaluation scale. If a judge change becomes necessary before the submission deadline, it will be announced to all registered teams in advance. After the deadline, all final submissions are scored under a single judge.

Task Score. For each example i, the judge assigns an answer score $a_i \in [0, 1]$ based on whether the predicted answer is semantically equivalent to the reference answer. The next section specifies how a_i is determined for each of the four MMBU task types. The Task Score is the mean over all examples:

$$T = \frac{1}{N} \sum_{i=1}^{N} a_i$$

Biomedical Context Score. Let $m_{ik} \in [0, 1]$ denote the judge-assigned correctness of metadata field k for example i, and let

$$z_{ik} = \begin{cases} 1, & \text{if metadata field } k \text{ is annotated for example } i, \\ 0, & \text{otherwise.} \end{cases}$$

The context score for an individual example is the mean over its annotated fields,

$$c_i = \frac{\sum_k z_{ik} m_{ik}}{\sum_k z_{ik}},$$

and the Biomedical Context Score is the mean over all examples:

$$C = \frac{1}{N} \sum_{i=1}^{N} c_i.$$

Normalizing within each example ensures that examples carrying more annotations do not receive greater weight simply because more fields are available to score.

Combining the two scores. All three tracks share a common base score, the weighted arithmetic mean

$$0.8\,T + 0.2\,C.$$

Task performance carries the majority of the weight, while the context component rewards models that correctly recognize the biomedical setting underlying their predictions. The arithmetic mean prevents very poor performance on either component from being fully compensated by strong performance on the other. Each track then modifies this base score to reflect what that track is designed to reward.

Open Frontier Performance Track. The base score is used directly:

$$\boxed{S_{\text{Frontier}} = 0.8\,T + 0.2\,C}$$

Medical Domain Adaptation Track. This track rewards the value added by domain adaptation, not the strength of the foundation the team started from. The base score is therefore scaled by the change in final answer accuracy between the adapted model and its declared base model:

$$\boxed{S_{\text{Adapt}} = (0.8\,T + 0.2\,C) \times (\text{Acc}_{\text{med}} - \text{Acc}_{\text{base}})}$$

where Acc_{med} and Acc_{base} are the final answer accuracies of the adapted and base models respectively, both benchmarked by the MMBU Challenge organizers under identical conditions. Acc_{med} is the same as T. The multiplier ensures that a high S_{Adapt} requires demonstrated improvement over the base model rather than a strong starting checkpoint. **This quantity is signed**. A model that performs worse than its base model after adaptation receives a negative score and ranks below any model that achieved no improvement.

Domain Adaptation and Efficient Inference Track. This track additionally rewards strong performance from smaller models:

$$\boxed{S_{\text{Efficient}} = (0.8\,T + 0.2\,C)\left(\frac{P_{\max}}{P}\right)^{\alpha}}$$

where P is the submitted model's **active** parameter count, $P_{\max}$ is the maximum active parameter count permitted in this track, and α controls the strength of the efficiency bonus. The exponent is small by design, so that model quality remains the dominant factor while models achieving comparable performance with fewer parameters retain a meaningful advantage. **The value of α will be fixed and announced to all registered teams before the challenge end.**

Participation, Encouragements, and Benefits

Our goal is to make the MMBU Challenge easy to participate in and difficult to game. Participants are encouraged to experiment broadly with training data, architectures, and optimization strategies unless explicitly prohibited below. The restrictions primarily apply to hidden-test access, external information at evaluation time, model orchestration, and non-reproducible evaluation procedures. If a rule is ambiguous, teams are encouraged to ask during office hours. Clarifications affecting competition will be shared with all teams.

- **Broad Experimentation:** Teams have full freedom to explore innovative training recipes, architectural designs, and optimization techniques.
- **Development Set:** Teams will receive a portion of MMBU released publicly with answers to understand benchmark design and for local evaluation.
- **Dedicated Support:** Organizers hold 30-minute weekly office hours to answer questions, resolve ambiguities, and provide official clarifications.
- **Recognition and Publication:** The top-performing team in each track will be invited to contribute to the MMBU Challenge technical report.

Rules and Restrictions

1. **Model Availability and Track Eligibility:** For the Domain Adaptation and Efficient Inference Track and all models from other tracks below 27B, models must be publicly available on **Hugging Face** and downloadable without gating or access restrictions. Models must satisfy the parameter limit for their selected track. Parameter counts include all **active** components used at inference, including vision, language, multi-modal projection, and adapter modules. For all models, unless previously disclosed and agreed upon by MMBU Challenge organizers, teams must report total and active parameters and be willing to upload to Hugging Face. For Medical Domain Adaptation Track, base model must be readily available on Hugging Face or API rules following the same rules previously specified.
2. **One Model per Track:** A model checkpoint may be entered into only **one track**. However, teams may compete in multiple tracks using distinct eligible checkpoints if registration is approved.
3. **Single End-to-End VLM:** The end goal of this challenge is to improve native perception capabilities of VLMs. Submissions must use a **single end-to-end VLM**. Ensembles, cascades, auxiliary models, external OCR, retrieval systems, learned preprocessing/postprocessing modules, or other model orchestration are **prohibited**.
4. **External Tools:** Models may not access web search, external APIs, databases, retrieval systems, calculators, code execution environments, or other external information sources during evaluation. Predictions must rely only on the submitted model and benchmark input.

5. **Inference Procedure:** All models will be evaluated using the official MMBU Challenge evaluation harness and organizer-specified inference settings. Each example may be processed only once. Repeated sampling, self-consistency, generation ensembling, iterative prompting, test-time adaptation, and weight modification during evaluation are prohibited.
6. **Training Data Disclosure and Contamination Restrictions:** Teams must disclose, to the best of their knowledge, major pretraining, continued-pretraining, fine-tuning, and synthetic-data sources, including any known overlap with datasets contributing to MMBU [2]. Teams are strictly prohibited from intentionally training, fine-tuning, selecting checkpoints, or otherwise optimizing using MMBU evaluation examples, labels, annotations, or derivatives. Unknown overlap inherited from third-party pretraining will be reviewed at the organizers' discretion.
7. **Development and Final Evaluation:** A development set will be released on Hugging Face for local evaluation. Final rankings will use a separate hidden evaluation set and metrics described below. Our evaluation prompt will ask for a json schema formatted answer with our criteria. Models will be run with **BF16 or FP16 precision**. Before the submission deadline, each team must designate and upload **one frozen final model per track**; failure to do so will result in no final score. Hidden-set scores will be released only after final submissions close.
8. **Scoring and Compute:** Final rankings will use the official MMBU Challenge metric for each track. Missing, refused, malformed, or unparsable predictions receive no credit for the affected example in the pass which it occurs. Submissions must also satisfy organizer-specified hardware, memory, runtime, and inference constraints.
9. **Failures and Reproducibility:** Organizer-side infrastructure failures may be rerun at the organizers' discretion. Participant-side failures, including incompatible code, model-loading errors, out-of-memory errors, malformed outputs, or compute-limit violations, do not automatically qualify for reruns. Top-performing teams may be required to provide model revisions, inference code, prompts/templates, dependencies, and hardware requirements sufficient to reproduce their results. Non-reproducible submissions may be removed from the final ranking.
10. **Tie Breaking:** In the case of a tie in a given track, the model with the highest final answer performance, T, will be the winner.
11. **Team Eligibility:** Teams must **apply and be accepted** before participating. Each participant may belong to only one competing team unless approved by the organizers. Team membership must be finalized by the registration deadline. Individuals with access to hidden evaluation labels or other non-public information providing a competitive advantage are ineligible for official ranking or awards.
12. **Official Communications and Rule Changes:** Official clarifications and material rule changes will be communicated to all registered teams. The organizers may amend rules when necessary to address ambiguity, evaluation errors, security concerns, or unforeseen circumstances.

13. **Evaluation Integrity:** Participants must not attempt to infer, extract, reconstruct, or otherwise exploit information about the hidden test set or evaluation procedure. Prohibited activities include, but are not limited to, **test-set discovery, prompt injection, evaluator exploitation,** and **metric exploitation**. Any such attempt by an individual participant or member of a team will result in the **automatic disqualification of the entire team** and all of its submissions.
14. **Anti-Collusion:** Participants must not share non-public information relevant to the challenge, including dev-set analysis beyond public leaderboard results or coordinated submission timing or strategy, with members of another registered team.

The organizers reserve the right to verify submissions and disqualify teams that violate these rules, exploit evaluation-set leakage, LLM judging, misrepresent their system, exceed challenge restrictions, or submit results that cannot be reproduced.

Sponsorship and Partnership

The MMBU Challenge is conducted in collaboration with sponsors and partners across various domains. Here, we clearly define each partnership category, as well as the specific institutions and companies that support this initiative.

Organizers: The MMBU Challenge is officially organized by Stanford University, specifically within the Stanford Medical AI and Computer Vision Lab (Stanford MARVL) under the guidance of Dr. Serena Yeung-Levy, Assistant professor of Biomedical Data Science and, by courtesy, of Computer Science and of Electrical Engineering at Stanford University. The primary challenge organizers are Ryan D'Cunha, Alejandro Lozano, Akira Nishii, and Maximillian Rokuss, with full affiliations listed above. Only the organizers have access to the final held out private MMBU set allowing sponsors and collaborators to participate in the challenge.

Sponsorship: All sponsors have committed monetary support to the MMBU Challenge and are formally recognized in the technical report. Sponsors may also choose to participate as co-authors. Sponsors also assist in publicizing the MMBU Challenge and serve as the primary corporate and institutional entities associated with the challenge. The official sponsors of the MMBU Challenge are **Stanford AI Lab**, **Anthropic**, **GXL**, **Highlanders**, and **Biohub**. Sponsors may also opt in to attend office hour sessions to receive direct feedback from participants about their use of sponsor-provided tools.

Collaborators: Building upon the foundation established by the MMBU benchmark [2], we work alongside numerous collaborators to advance the goal of addressing the perception bottleneck in multimodal models. Collaborators offer **no**

monetary assistance to the MMBU Challenge and are strictly distinct from sponsors. Instead, collaborators contribute to challenge development by providing feedback on challenge rules, evaluation metrics, track recommendations, and experimental design, as well as supporting outreach and working alongside the organizers to address the technical challenges. Collaborators are individual researchers rather than institutions, spanning academia and industry, with affiliations including Stanford University, the University of Cambridge, the University of Washington, the University of Texas MD Anderson Cancer Center, Microsoft AI, Google DeepMind, and Apple. Affiliations are listed for identification only and do not imply institutional endorsement or partnership. Collaborators may be included in the technical report either as authors or in the acknowledgments section, solely based on their contributions. **In all public-facing materials — including this document, the challenge website, LinkedIn posts, the award ceremony, and any challenge-related publicity — collaborator institutions are acknowledged in text only. No collaborator logos, wordmarks, or other brand assets are used in any public-facing material.**

MMBU Authors: We also acknowledge all authors of MMBU [2], without whom this challenge would not be possible with. Ryan D'Cunha, Alejandro Lozano, Xiaoxiao Sun, Daniel Vela Jarquin, Min Woo Sun, Josiah Aklilu, James Burgess, Yuhui Zhang, Ryan Nayebi, Paola Avila Robayo, Jin Ye, Ming Hu, Zhongying Deng, Junjun He, Xin Chen, Yue Yao, Robert Tibshirani, Jeffrey J. Nirschl, and Serena Yeung-Levy. Full affiliations for authors are shown in the paper.

References

1. Burgess, J., Nirschl, J.J., Bravo-Sánchez, L., Lozano, A., Gupte, S.R., Galaz-Montoya, J.G., Zhang, Y., Su, Y., Bhowmik, D., Coman, Z., Hasan, S.M., Johannesson, A., Leineweber, W.D., Nair, M.G., Yarlagadda, R., Zuraski, C., Chiu, W., Cohen, S., Hansen, J.N., Leonetti, M.D., Liu, C., Lundberg, E., Yeung-Levy, S.: Microvqa: A multimodal reasoning benchmark for microscopy-based scientific research (2025), https://arxiv.org/abs/2503.13399
2. D'Cunha, R., Lozano, A., Sun, X., Jarquin, D.V., Sun, M.W., Aklilu, J., Burgess, J., Zhang, Y., Nayebi, R., Avila, P., et al.: Mmbu: A massive multi-modal biomedical understanding benchmark to probe the perception capabilities of vision-language models. arXiv preprint arXiv:2606.06696 (2026)
3. Fu, S., Bonnen, T., Guillory, D., Darrell, T.: Hidden in plain sight: Vlms overlook their visual representations, 2025. URL https://arxiv. org/abs/2506.08008
4. Lozano, A., Nirschl, J., Burgess, J., Gupte, S.R., Zhang, Y., Unell, A., Yeung-Levy, S.: Micro-bench: a vision-language benchmark for microscopy understanding. In: Proceedings of the 38th International Conference on Neural Information Processing Systems. pp. 30670–30685 (2024)
5. Lozano, A., Sun, M.W., Burgess, J., Chen, L., Nirschl, J.J., Gu, J., Lopez, I., Aklilu, J., Rau, A., Katzer, A.W., et al.: Biomedica: An open biomedical image-caption archive, dataset, and vision-language models derived from scientific litera-

ture. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 19724–19735 (2025)
6. Mahesh, M., Ansari, A.J., Mettler Jr, F.A.: Patient exposure from radiologic and nuclear medicine procedures in the united states and worldwide: 2009–2018. Radiology **307**(1), e221263 (2022)
7. Qwen Team: Qwen3.5: Towards native multimodal agents (February 2026), https://qwen.ai/blog?id=qwen3.5
8. Team, G., Abd, S.E., Aggarwal, V., Algayres, R., Andreev, A., Bachem, O., Ballantyne, I., Brick, C., Cărbune, V., Casbon, M., et al.: Gemma 4 technical report. arXiv preprint arXiv:2607.02770 (2026)
9. Valanarasu, J.M.J., Xu, H., Usuyama, N., Kim, C., Wong, C., Argaw, P., Shimol, R.B., Crabtree, A., Matlock, K., Bartlett, A.Q., et al.: Multimodal ai generates virtual population for tumor microenvironment modeling. Cell **189**(2), 386–400 (2026)
10. Wang, W., Lv, Q., Yu, W., Hong, W., Qi, J., Wang, Y., Ji, J., Yang, Z., Zhao, L., Song, X., et al.: Cogvlm: Visual expert for pretrained language models. Advances in Neural Information Processing Systems **37**, 121475–121499 (2024)
11. Wang, X., Huang, J., Zhang, X., Wang, T., Ma, J.W.: Your reasoning benchmark may not test reasoning: Revealing perception bottleneck in abstract reasoning benchmarks. In: Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 18113–18124 (2026)
12. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Hu, F., Ge, H., Wei, H., Lin, H., Tang, J., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Zhou, J., Lin, J., Dang, K., Bao, K., Yang, K., Yu, L., Deng, L., Li, M., Xue, M., Li, M., Zhang, P., Wang, P., Zhu, Q., Men, R., Gao, R., Liu, S., Luo, S., Li, T., Tang, T., Yin, W., Ren, X., Wang, X., Zhang, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Zhang, Y., Wan, Y., Liu, Y., Wang, Z., Cui, Z., Zhang, Z., Zhou, Z., Qiu, Z.: Qwen3 technical report (2025), https://arxiv.org/abs/2505.09388