https://doi.org/10.65091/icicset.v2i1.11

Vision Language Model for Clinical Decision Support System

Mohan Bhandari
Department of CSIT
Samriddhi College
Bhaktapur, Nepal
mail2mohanbhandari@gmail.com

Smriti KC
Department. of CSIT
Padmakanya Multiple Campus
Kathmandu, Nepal
kcsmriti80@gmail.com

Anam Afaq
Asian Business School
Noida, India
anam24afaq@gmail.com

Loveleen Gaur
Graduate Business School
University of the South Pacific
Suva, Fiji
gaurloveleen@yahoo.com

***Abstract*—In the field of assisted reproductive technology, the differentiation of competent embryos ready to transfer is also founded on the Gardner grading system, which is an assessment of blastocysts that primarily focuses on morphology, but is also largely subject to inter- and intra-observer variability. This variability can decrease the integrity and continuity of the in-vitro fertilization (IVF) process based on embryo classification to determine viability and an improved probability of pregnancy. To solve this problem, we present the application of Bootstrapping Language Image Pre-training (BLIP), for blastocyst grading along Gardner, and from that grading proposes an easy classification scheme. Of 249 day-5 human blastocyst images and Gardner grades, 204 images were used to fine-tune BLIP where BLIP frames the grading task as medical image captioning. Select the number of unique grades to determine that the number of model outputs was greater than 10. The average training loss was 0.1010 and the fine-tuned model achieved a Recall-Oriented Understudy for Gisting Evaluation score of 0.7391, Hamming accuracy of 0.8913 and a Metric for Evaluation of Translation with Explicit ORdering of 0.3696 for the test set. These results demonstrate that a fine-tuned vision language model can accurately approximate complex morphological features in the blastocyst image and predict their assigned grade classification. Training and testing codes for the model developed in this study are available in https://github.com/MohanBhandari/GardnerGrade-VLM.**

***Index Terms*—Clinical decision support, Gardner Grades, Human Blastocyst, Image Captioning, Medical Images, Vision Language Model**

I. Introduction

IVF plays a vital role in reproductive medicine, helping millions of infertile people become parents [1]. More than 2.5 million IVF cycles worldwide produce about 500,000 births annually [2]. Success rates, based on live births per cycle, depend on factors such as age and health. Choosing the best embryo for transfer is crucial to increase pregnancy chances and reduce risks such as preterm birth or low birth weight [3]. For years, the Gardner system has been the benchmark for embryo selection, using bright-field microscopy to manually assess blastocysts [4]. It evaluates three key components: blastocoel expansion, inner cell mass (ICM) quality, which forms the fetus, and trophectoderm (TE) quality, which becomes the placenta. Although widely used, this method is subjective and leads to inconsistent results even among skilled embryologists. This variability can affect embryo selection and clinical outcomes. Additionally, static morphological grades of this system, although related to implantation, are not always reliable predictors of blastocyst viability or developmental potential.

Studies have effectively used Convolutional Neural Networks (CNNs), such as ResNet, Inception, and U-Net, to segment key components and standardize the evaluation of blastocyst expansion, ICM quality, and TE quality, often performing as well as or better than less-experienced embryologists [5–7] . Advanced models, particularly those using time-lapse imaging data, go further by directly predicting implantation potential. These AI systems provide objective, high-throughput analysis, but, they do not focus on automated gardner grade guide (GGG).

Vision-Language Models (VLMs) introduce a flexible and robust approach to embryo evaluation, surpassing traditional AI segmentation methods [8]. These models learn to connect visual data with natural language by training on image-text datasets, allowing them to interpret images alongside detailed descriptions. This makes them highly adaptable for clinical use, supporting tasks such as zero-shot classification, image-text retrieval, visual question answering, and automated report creation. Existing medical VLMs, such as MedCLIP or PMC-CLIP, are mainly trained in radiology (e.g., X-rays, CT scans) or pathology (e.g., histology slides), which differ significantly from blastocyst microscopy in visual features and terminology [9, 10]. Additionally, evaluating blastocysts requires detecting subtle details such as TE cell count or ICM compactness, unlike broader objects in general data sets such as ImageNet.

To overcome the limitations of existing AI approaches in embryology, a framework for human blastocyst Grading using bootstrapping language image pre-training (BLIP) is proposed. BLIP can be domain-adapted to bridge the gap between blastocyst morphology and descriptive grades of GGG. The architecture of BLIP includes modules for captioning and filtering and can be implemented to bootstrap its own learning from limited datasets [11].

The following are the objectives of the study.

a. This study aims to fine-tune the pre-trained vision-language model, BLIP, for specialized task of generating human blastocyst grade captions. The core challenge is achieving an efficient domain transfer from the BLIP

pretrained to the medical features of blastocyst microscopy.

b. Evaluate BLIP performance using Recall-Oriented Understudy for Gisting Evaluation (ROUGE), Hamming accuracy and Metric for Evaluation of Translation with Explicit ORdering (METEOR).

II. Review of Literature

A. Transfer Learning in medical imaging

In fields such as medical imaging, where labeled data are often scarce, transfer learning has become a dominant strategy for training deep neural networks. Different studies have provided comprehensive applications of transfer learning in medical images [12]. Findings have been mixed; for example, Mustafa et al. [13] suggested that the transfer of natural images can be beneficial if applied at a sufficient scale. Other studies like those of Peng et al. [14] and Raghu et al.[15] have suggested that large and complex models do not invariably outperform simpler, more lightweight models. So, VLM can be a better alternative in medical imaging.

B. Vision-Language Models

VLM is a critical area of research for medical image analysis, particularly for the automatic generation of diagnostic captions and radiology reports. This task is essential to reduce clinician workload and standardize findings. Study by Alomar et al. [16] highlighted the transition from traditional CNN-LSTM architectures to more powerful transformer-based models, which are "noticeably better" at capturing the complex long-range dependencies required for clinically accurate text. Different studies contribute to modern VLP approaches, aiming to pre-train models on vast datasets before fine-tuning them on specific medical tasks. Transformer based models such as R2Gen [17, 18], designed to generate radiology reports by retrieving and modifying sentences from similar existing reports, are now gradually being outperformed by VLM based visual question answering and report generation [19].

VLM has been used to adapt successful frameworks in the general-domain. Zhang et al. [20], explored the contrastive learning paradigm for medical VLP, aiming to learn robust medical visual representations of the associated text. This concept was significantly advanced by MedCLIP [21], which was designed to build a robust domain-specific contrastive model from scratch by pre-training millions of image-report pairs to learn fine-grained medical semantics. The unified framework in MedViLL [22] had the objective of improving generalization for both understanding and generation tasks using a unique attention masking scheme within a single BERT-based model. This "fine-tuning" approach with pre-trained general models has proven highly effective. ClinicalBLIP by Ji et al. [23] is an excellent example, with the aim of demonstrating the effectiveness of fine-tuning a model based on InstructBLIP for the specific task of generating radiology reports. LLaVA-Med [24] is designed to create a biomedical conversational assistant through visual instruction-tuning, enabling it to generate highly detailed and context-sensitive image descriptions. These VLM techniques are also being applied to other medical specialties, such as pathology, where He et al. [25] developed a model with the objective of generating pathology captions by cleverly leveraging a VQA data set. New VLM frameworks are being developed to improve the accuracy of radiology reports by using a prompt-based retrieval-generation system.

III. Methodology

Figure 1 shows the overall workflow. The study begins with system configuration and data preparation, where the image-text pairs were loaded, cleaned, and filtered. A pre-trained BLIP model is then fine-tuned over 50 epochs. During training, the model's loss is monitored, and the best-performing checkpoint is saved and evaluated.

A. Configuration setup

For the experiment, the model configurations are defined to ensure reproducibility and optimize performance. The training data are sourced from a CSV file with the grades and the corresponding directory of blastocyst images. Model checkpoints are periodically saved to preserve training progress. The training process is set to run for a maximum of 50 epochs, using a physical batch size of 8 image-text pairs to fit within GPU memory constraints. To ensure training stability, this is combined with gradient accumulation over 4 steps, which simulates a larger, more effective batch size of 32. The fine-tuning learning rate of 5×10^{-5} is applied, along with a weight loss of 0.01 to provide regularization and prevent overfitting.

B. Dataset and preprocessing

In this study, the microscopic dataset of human blastocysts [27] is used. Each of the 249 images in this dataset has a mask annotation and a unique categorical grade that represents its quality according to GGG. The distribution of the grades is not balanced and some categories, such as "3AA" and "4AA", have a high frequency. In this study, only image grades with a frequency greater than 10 occurrences are included. A total of 204 images and their associated grades are considered for subsequent analytical robustness. Of 232 images, 80% are used for training and 20% for testing

C. BLIP

The BLIP framework (Figure 2) uses a single visual transformer as an image encoder and a flexible text transformer that operates in three modes, corresponding to its three main pre-training objectives

a. Image-Text Contrastive (ITC)
   ITC uses a unimodal encoder to align the visual features of the image with the text features, teaching the model to pull the representations of positive pairs closer.

b. Image-Text Matching (ITM)
   ITM uses an image-grounded text encoder to predict whether an image-text pair is a match. It learns fine-grained alignment between vision and language.

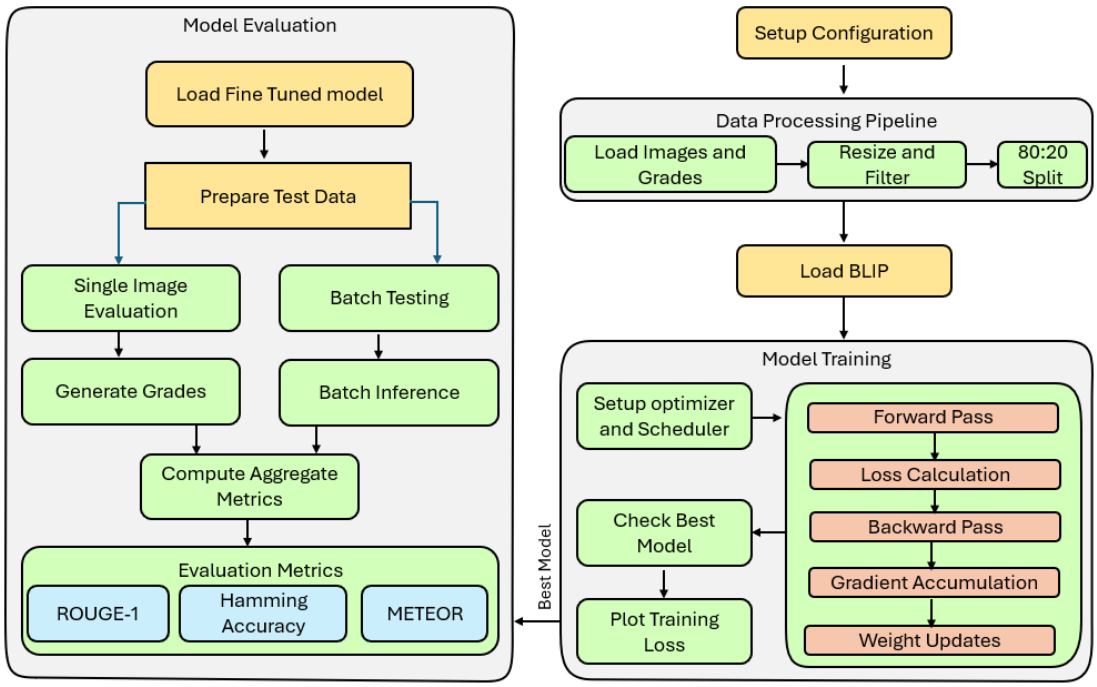


Fig. 1. Proposed methodology

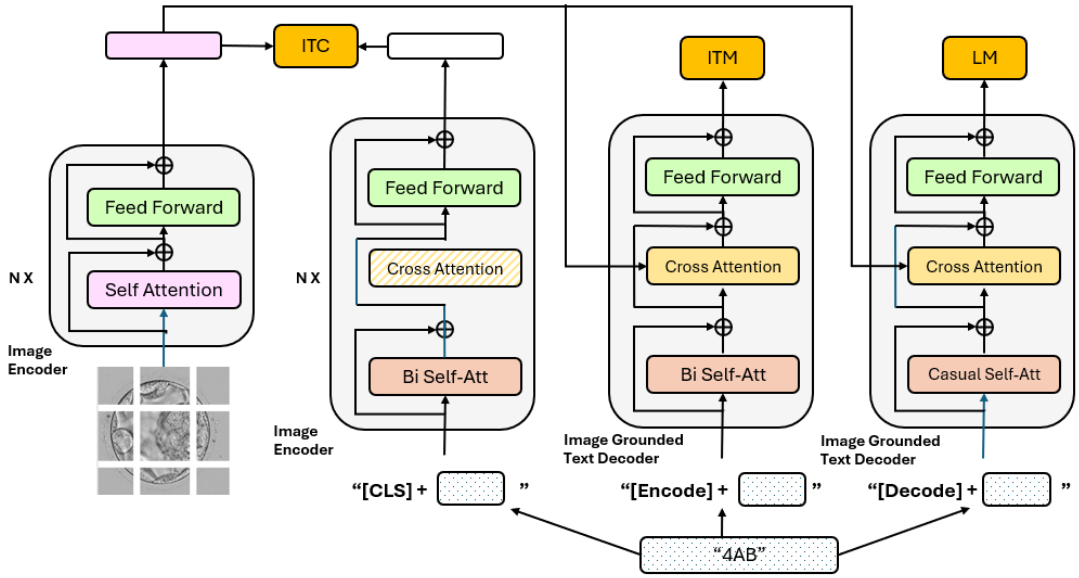


Fig. 2. Pre-training model architecture of BLIP [26]

c. Language Modeling (LM)
LM uses an image-grounded text decoder to generate captions. It learns to automatically predict text based on visual information.

D. Evaluation Metrics

1) ROUGE-1: Let $|S_{\text{true}}|$ be the set of unigrams in the reference (actual) grades $|S_{\text{gen}}|$ be the set of unigrams in the generated text. and $|S_{\text{true}} \cap S_{\text{gen}}|$ be the number of overlapping unigrams. The ROUGE-1 is calculated as

$$\text{ROUGE-1} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \tag{1}$$

where, Precision $= \frac{|S_{\text{true}} \cap S_{\text{gen}}|}{|S_{\text{gen}}|}$, and Recall $= \frac{|S_{\text{true}} \cap S_{\text{gen}}|}{|S_{\text{true}}|}$.

2) Hamming Accuracy: For N is the total number of positions and C is the number of matching positions, Hamming

accuracy calculated as

$$\text{Hamming Accuracy} = \frac{C}{N} \tag{2}$$

In this study $N = 3$.

3) Metric for Evaluation of Translation with Explicit ORdering: METEOR combines precision and recall into an F-score, specifically the weighted harmonic mean, to balance the two as:

$$\text{METEOR} = \frac{\text{Precision} \cdot \text{Recall}}{\alpha \cdot \text{Precision} + (1-\alpha) \cdot \text{Recall}} \tag{3}$$

The default value of the weighting parameter ($\alpha = 0.5$) is used [28].

IV. Result Analysis

The BLIP model trained in 50 epochs shows a training loss of 0.1010 as shown in Figure 3.

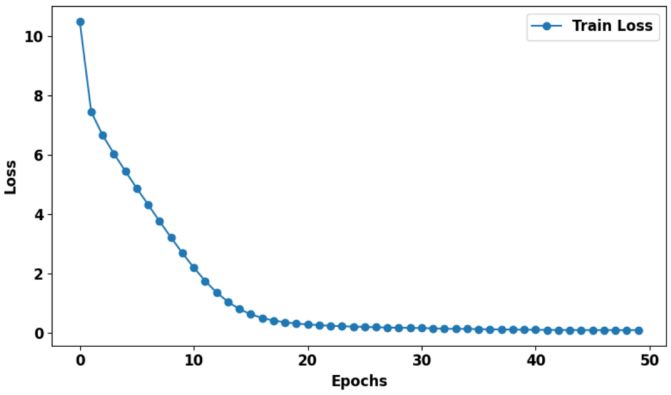
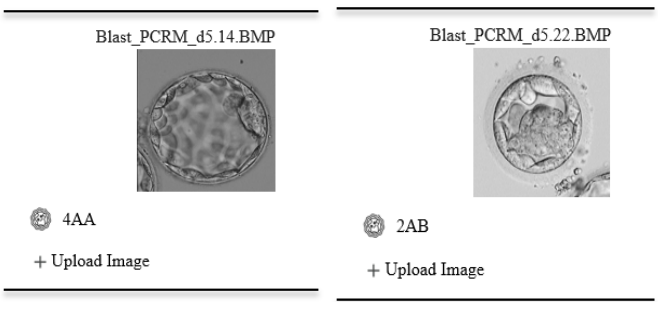


Fig. 3. Training Loss

The top-performing model, carefully adjusted and tracked by using loss during training, was saved to generate the gardner grades. Its performance was tested on randomly selected dataset images, as shown in Figure 4 (a), where it predicted a "4AA" grade for a blastocyst image with a true "4AB" grade, and in Figure 4 (b), where it accurately identified the correct gardner grade.



(a) 4AB predicted to 4AA (b) 2AB predicted to 2AB

Fig. 4. Real time implementation of fine-tuned BLIP model

Table I presents the performance metrics for BLIP with four key metrics: average test loss, ROUGE-1, METEOR, and Hamming Accuracy for the test set. The average test loss is 0.1801. ROUGE-1, with a value of 0.7391, is relatively high, indicating strong alignment with the reference grades. METEOR, at 0.3696, suggests reasonable semantic alignment. Hamming Accuracy, at 0.8913, indicates high accuracy in classification tasks.

TABLE I
Performance Metrics

Metric	Value
Average Test Loss	0.1801
ROUGE-1	0.7391
METEOR	0.3696
Hamming Accuracy	0.8913

V. Conclusion

This study shows the feasibility of automating the GGG using a fine-tuned VLM. The variability of manual morphological evaluation presents a significant challenge in standardizing embryo selection during IVF. By framing blastocyst grading as an image captioning task, the fine-tuned BLIP model effectively interprets complex embryological features and generates the corresponding Gardner grade classifications. The strong performance of the model, evidenced by a high ROUGE-1 score of 0.7391 and a Hamming precision of 0.8913, indicates its ability to accurately describe blastocyst morphology according to established clinical guidelines. These findings suggest that VLMs offer a flexible alternative to traditional CNN-based approaches, capable of providing a comprehensive descriptive assessment rather than just a numerical classification. The implementation of BLIP improves the consistency, objectivity, and efficiency of the embryo selection process in clinical IVF workflows.

Future work should focus on expanding the dataset with more diverse grades and incorporating temporal data from time-lapse imaging to further improve robustness and potentially correlate generated captions with ultimate implantation potential, moving towards a more predictive model of embryo viability.

References

[1] G. Mendizabal-Ruiz, O. Paredes, E. Borrayo, and A. Chavez-Badiola, "Fertility care in low-and middle-income countries: The future use of ai to improve accessibility of assisted reproductive technology in low-and middle-income countries," *Reproduction and Fertility*, vol. 6, no. 3, 2025.

[2] Z. Shoham, A. Weissman, and Y. Yaron, "Global ethics in ivf: Harmonizing regulation, ensuring access, and governing innovation," *Journal of IVF-Worldwide*, vol. 3, no. 3, pp. 62–80, 2025.

[3] B. Li, D. He, H. Zhu, Y. Li, W. Ye, D. Guo, M. Yu, Y. Wu, J. Cai, L. Ji *et al.*, "Factors influencing the birth

weight of art-conceived offspring," *Journal of Assisted Reproduction and Genetics*, pp. 1–22, 2025.

[4] G. Mrugacz, I. Bołkun, T. Magoń, I. Korowaj, B. Golka, T. Pluta, O. Fedak, P. Cieśla, J. Zowczak, and E. Skórka, "Time-lapse imaging in ivf: Bridging the gap between promises and clinical realities," *International Journal of Molecular Sciences*, 2025.

[5] W. S. Miled, R. Ghali, S. Chtourou, and M. A. Akhloufi, "Semantic segmentation of human blastocyst images using deep cnns and vision transformers," *Network Modeling Analysis in Health Informatics and Bioinformatics*, vol. 14, no. 1, p. 63, 2025.

[6] K. Chen, Z. Li, G. Huang, and J. Guo, "Embryo-net: A blastocyst image segmentation network based on spatial modeling to resolve the challenge of textural similarity between the te and icm," *Neurocomputing*, vol. 638, p. 130153, 2025.

[7] R. AlSaad, L. Abusarhan, N. Odeh, A. Abd-Alrazaq, F. Choucair, R. Zegour, A. Ahmed, S. Aziz, and J. Sheikh, "Deep learning applications for human embryo assessment using time-lapse imaging: scoping review," *Frontiers in Reproductive Health*, vol. 7, p. 1549642, 2025.

[8] C. Liu, Y. Jin, Z. Guan, T. Li, Y. Qin, B. Qian, Z. Jiang, Y. Wu, X. Wang, Y. F. Zheng *et al.*, "Visual–language foundation models in medicine," *The Visual Computer*, vol. 41, no. 4, pp. 2953–2972, 2025.

[9] H. Lin, C. Xu, and J. Qin, "Taming vision-language models for medical image analysis: A comprehensive review," *arXiv preprint arXiv:2506.18378*, 2025.

[10] Z. Shui, J. Zhang, W. Cao, S. Wang, R. Guo, L. Lu, L. Yang, X. Ye, T. Liang, Q. Zhang *et al.*, "Large-scale and fine-grained vision-language pre-training for enhanced ct image understanding," *arXiv preprint arXiv:2501.14548*, 2025.

[11] Y. Jian, C. Gao, and S. Vosoughi, "Bootstrapping vision-language learning with decoupled language pre-training," *Advances in Neural Information Processing Systems*, vol. 36, pp. 57–72, 2023.

[12] H. E. Kim, A. Cosa-Linan, N. Santhanam, M. Jannesari, M. E. Maros, and T. Ganslandt, "Transfer learning for medical image classification: a literature review," *BMC medical imaging*, vol. 22, no. 1, p. 69, 2022.

[13] B. Mustafa, A. Loh, J. Freyberg, P. MacWilliams, M. Wilson, S. M. McKinney, M. Sieniek, J. Winkens, Y. Liu, P. Bui *et al.*, "Supervised transfer learning at scale for medical imaging," *arXiv preprint arXiv:2101.05913*, 2021.

[14] L. Peng, H. Liang, G. Luo, T. Li, and J. Sun, "Rethinking transfer learning for medical image classification," *arXiv preprint arXiv:2106.05152*, 2021.

[15] M. Raghu, C. Zhang, J. Kleinberg, and S. Bengio, "Transfusion: Understanding transfer learning for medical imaging," *Advances in neural information processing systems*, vol. 32, 2019.

[16] K. Alomar, H. I. Aysel, and X. Cai, "Cnns, rnns and transformers in human action recognition: a survey and a hybrid model," *Artificial Intelligence Review*, vol. 58, no. 12, pp. 1–44, 2025.

[17] S. Allaberdiev, A. Khan, S. Mamarasulov, and X. Chen, "Chestxgen: Dynamic memory-augmented vision-language transformer with context-aware gating for radiology report generation," *Journal of Artificial Intelligence and Soft Computing Research*, vol. 16, no. 1, pp. 55–72, 2026.

[18] I. Shahzadi, T. M. Madni, U. I. Janjua, G. Batool, B. Naz, and M. Q. Ali, "Csamdt: Conditional self attention memory-driven transformers for radiology report generation from chest x-ray," *Journal of Imaging Informatics in Medicine*, vol. 37, no. 6, pp. 2825–2837, 2024.

[19] I. Hartsock and G. Rasool, "Vision-language models for medical report generation and visual question answering: A review," *Frontiers in artificial intelligence*, vol. 7, p. 1430984, 2024.

[20] Y. Zhang, H. Jiang, Y. Miura, C. D. Manning, and C. P. Langlotz, "Contrastive learning of medical visual representations from paired images and text," in *Machine learning for healthcare conference*. PMLR, 2022, pp. 2–25.

[21] Z. Wang, Z. Wu, D. Agarwal, and J. Sun, "Medclip: Contrastive learning from unpaired medical images and text," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, vol. 2022, 2022, p. 3876.

[22] J. H. Moon, H. Lee, W. Shin, Y.-H. Kim, and E. Choi, "Multi-modal understanding and generation for medical images and text via vision-language pre-training," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 12, pp. 6070–6080, 2022.

[23] J. Ji, Y. Hou, X. Chen, Y. Pan, and Y. Xiang, "Vision-language model for generating textual descriptions from clinical images: model development and validation study," *JMIR Formative Research*, vol. 8, p. e32690, 2024.

[24] C. Li, C. Wong, S. Zhang, N. Usuyama, H. Liu, J. Yang, T. Naumann, H. Poon, and J. Gao, "Llava-med: Training a large language-and-vision assistant for biomedicine in one day," *Advances in Neural Information Processing Systems*, vol. 36, pp. 28 541–28 564, 2023.

[25] X. He, Y. Zhang, L. Mou, E. Xing, and P. Xie, "Pathvqa: 30000+ questions for medical visual question answering," 2020. [Online]. Available: https://arxiv.org/abs/2003.10286

[26] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International conference on machine learning*. PMLR, 2022, pp. 12 888–12 900.

[27] P. Saeedi, D. Yee, J. Au, and J. Havelock, "Automatic identification of human blastocyst components via texture," *IEEE Transactions on Biomedical Engineering*,

vol. 64, no. 12, pp. 2968–2978, 2017.

[28] J. Baradia, S. Gupta, and S. Kundu, “Mirror minds: An empirical study on detecting llm-generated text via llms,” in *Proceedings of the 1stWorkshop on GenAI Content Detection (GenAIDetect)*, 2025, pp. 59–67.