Supplementary Material—LLaDA-MedV: Exploring Large Language Diffusion Models for Biomedical Image Understanding

Xuanzhao Dong[1*] Wenhui Zhu[1*] Xiwen Chen[2,5*] Zhipeng Wang[3*] Peijie Qiu[4]
Shao Tang[3] Xin Li[1] Yalin Wang[1†]

[1] Arizona State University, AZ, USA, [2] Clemson University, SC, USA
[3] LinkedIn Corporation, CA, USA, [4] Washington University in St. Louis, MO, USA
[5] Morgan Stanley, NY, USA

A. Implementation Details

A.1. Training Configuration

The training process of LLaDA-MedV is structured into three stages. Following the approach of [5], the first two stages aim to establish semantic alignment between biomedical language and visual content, while also enabling the model to follow visual instructions within a biomedical context. To further improve performance in dataset-specific scenarios, we introduce a third stage involving supervised fine-tuning (SFT) on three biomedical VQA datasets.

Throughout all stages, we employ LLaDA-8B-Instruct [10] as the language backbone. The vision tower is based on SigLIP2 [14], and the vision-language projection module is implemented as a lightweight two-layer MLP with GELU activation. All training is conducted using four NVIDIA A100 GPUs (80GB each). Additional training details are provided in Tab. 1.

A.2. Inference Procedure

Inference Details. The response generation process simulates the reverse dynamics of mask diffusion. Starting from a fully masked response (i.e., $t = 1$), the learned mask predictor p_θ iteratively reconstructs the assistant's response. To better align with the reverse diffusion process, an appropriate remasking strategy is applied at each step. The detailed inference procedure is presented in Algorithm 1.

Following prior works [10, 15], we adopt the low-confidence remasking strategy [2], in which only tokens with low confidence (i.e., lower p_θ values) are remasked at each step (e.g., Step 8), rather than selecting tokens uniformly at random. In addition, we explore a semi-autoregressive generation strategy [10]. Specifically, when $B < L$, the response of length L is divided into L/B blocks, where B denotes the block length. Generation proceeds sequentially from left to right across blocks, with each block undergoing $Z \cdot B/L$ sampling steps (i.e., as noted in Algorithm. 1).

Algorithm 1 Inference Strategy in LLaDA-MedV

Require: Trained mask predictor p_θ, user visual input X_v, text prompt u_0, response length L, and total sampling steps Z
Ensure: Start with fully masked response r_1 of length L (i.e., $t = 1$)

1: **for** t from 1 down to $1/Z$ with step $1/Z$ **do**
2: Set $s = t - 1/Z$
3: Predict $r_0 = \arg\max p_\theta(r_0 \mid X_v, u_0, r_t)$
4: **for** $i = 1$ to L **do**
5: **if** $r_t^i \neq \mathrm{M}$ **then**
6: Set $r_0^i = r_t^i$ {Retain meaningful token}
7: **else**
8: Remask $r_0^i = \mathrm{M}$ with probability s/t
9: **end if**
10: **end for**
11: **end for**
12: **return** r_0

For the main open-ended biomedical conversation tasks, we apply this strategy with $L = 256$, $B = 64$, and $Z = 256$. However, for downstream biomedical VQA tasks, we disable the semi-autoregressive mechanism by setting $L = B = Z = 64$. Unless otherwise specified, this setting is used consistently in all detailed analyses.

Baseline Model Configuration. We compare LLaDA-MedV against nine baseline vision-language models, covering both general-domain and biomedical-specific systems. These models are evaluated using their default system prompts and a maximum token limit of 256. For fair-

†Corresponding author: ylwang@asu.edu
*These authors contributed equally to this paper.

Training stage	Alignment	MD-SFT	SD-SFT (VQA-RAD)	SD-SFT(SLAKE)	SD-SFT(PathVQA)
Vision tower	Siglip2-so400m-patch14-384 [14]				
Language tower	LLaDA-8B-Instruct [10]				
Projector	2-layer MLP with GELU				
Attention	Bidirectional attention				
DeepSpeed	ZeRO-3 [13]				
Optimizer	AdamW				
Scheduler	Cosine scheduler with 3% warmup				
Batch size (Global)	32	8	8	8	8
Model max length	8192	8192	8192	8192	8192
#Samples in training set	600K	60k	1797	4919	19755
LR of language tower	-	1×10^{-5}	2×10^{-6}	2×10^{-6}	2×10^{-6}
LR of projector	1×10^{-3}	1×10^{-5}	2×10^{-6}	2×10^{-6}	2×10^{-6}
Epoch	2	4	2	10	7

Table 1. Training configurations of LLaDA-MedV across three stages. **MD-SFT** denotes the multi-turn dialogue SFT, and **SD-SFT** represent the single-turn dialogue over training set of VQA-RAD, SLAKE and PathVQA, respectively.

ness, models with intermediate reasoning capabilities (e.g., MedVLM-R1) are evaluated using the full generated response.

In the open-ended biomedical conversation setting, we use GPT-4.1 mini [11] to assess performance, as the original GPT-4 (0314) used in [5] is no longer publicly accessible. Additionally, Tab. 2 provides an overview of all baseline models included in this evaluation. For downstream VQA benchmarks, we report baseline results directly from [5] when available, and therefore omit redundant model-specific details for clarity. We kindly refer readers to [5] for comprehensive baseline configurations in the VQA setting.

A.3. Dataset

Training dataset. We adopt the training data introduced in [5] to equip LLaDA-MedV with biomedical visual understanding capabilities across three training stages. All images are processed using the SigLIP2 vision encoder [14].

- **Stage 1.** We use 600K aligned image-text pairs to train the vision-language projection module and learn robust cross-modal representations. Each training instance is structured as a single-turn dialogue, where the human prompt is removed except for a special `<image>` token. This setup ensures that the model learns to generate responses conditioned solely on image embeddings, without reliance on textual input.
- **Stage 2.** We leverage a 60K multi-turn dialogue dataset with inline entity mentions. Each instance contains a full conversation between a human and the assistant, where image embeddings are prepended only to the first human input. This configuration enables the model to follow visually grounded instructions across multiple dialogue turns.
- **Stage 3.** We perform supervised fine-tuning using the training sets from three biomedical VQA datasets: VQA-RAD [4], SLAKE [6], and PathVQA [3]. Each VQA sample is converted into a one-turn dialogue, where the question is treated as the human input and the ground-truth answer is used as the assistant's response. This training simulates the multi-turn dialogue structure introduced in Stage 2.

Open-end Biomedical Conversation Benchmark. We use the open-ended biomedical conversation benchmark proposed in [5] to evaluate model performance in realistic scenarios. Specifically, the dataset contains 193 novel questions paired with 50 corresponding images. Notably, these 50 image-caption pairs are entirely unseen during training. The questions fall into two categories: open-ended conversations and detailed descriptions. Conversation-style questions are typically broad biomedical queries related to the image content, while detailed description questions explicitly ask for fine-grained analysis of the visual information. As previously mentioned, ground-truth responses are generated by GPT-4 (i.e., GPT-4-0314) based solely on the associated figure captions. To evaluate the candidate model's response (e.g., from LLaDA-MedV), we use GPT-4.1 mini as the automatic evaluator. GPT-4.1 mini is provided with both the candidate response and the GPT-4 reference response, along with the original question, image caption, and relevant context. It is then prompted to rate each response across four dimensions, including helpfulness, relevance, accuracy, and level of detail, and to assign an overall score on a scale from 1 to 10, where higher scores indicate bet-

Table 2. Overview of baseline models evaluated on the open-ended biomedical conversation task. The **Domain** column indicates whether each model is general-purpose or biomedical-specific. The **Size** column refers to the model size, where specified. Detailed model variants are included when available.

Model	Domain	Size	Variant / Backbone
LLaMA [8]	General	11B	LLaMA-3.2-11B-Vision
LLaVA [7]	General	7B	llava-1.5-7B
LLaVA-Med [5]	Biomedical	7B	llava-med-v1.5-mistral-7B
Med-Flamingo [9]	Biomedical	9B	Med-Flamingo-9B
MedVLM-R1 [12]	Biomedical	-	-
Qwen-RL-Med [16]	Biomedical	-	-
Qwen-VL [1]	General	2B	Qwen2-VL-2B-Instruct
RetinalGPT [17]	Biomedical	-	-
LLaDA-V [15]	General	-	-

ter performance. Additionally, GPT-4.1 mini provides a detailed rationale for each evaluation, facilitating more transparent comparisons. The final score for the candidate model is normalized by the GPT-4 reference score to compute a relative performance metric.

Downstream Biomedical Visual Question Answering. Following [5], we use three biomedical visual question answering (VQA) benchmarks to further improve LLaDA-MedV performance in scenario that require high service quality. We outlined the detailed statistics in Tab. 3.

- **VQA-RAD [4]** consists of 3,515 clinician-generated QA pairs grounded in 315 radiology images, which are evenly distributed across three anatomical regions: head, chest, and abdomen. Each image is paired with multiple questions, covering a diverse range of clinical inquiries. The questions are categorized into 11 distinct types, including abnormality, attribute, modality, organ system, color, counting, object or condition presence, size, imaging plane, positional reasoning, and others. Approximately half of the answers are closed-ended (i.e., yes/no), while the remaining are open-ended, typically expressed as a single word or short phrase. It is worth noting that we use the cleaned version of the dataset in all experiments, resulting in a total of 2,248 QA pairs.
- **SLAKE [6]** is a Semantically-Labeled Knowledge-Enhanced benchmark designed for medical visual question answering. It contains 642 radiology images and over 7,000 diverse question-answer pairs, all annotated by experienced physicians. Many of the questions require external medical knowledge, which is supported by an accompanying medical knowledge graph. The dataset also provides rich visual annotations, including semantic segmentation masks and object detection bounding boxes. In addition, SLAKE encompasses a broader range of imaging modalities and anatomical regions compared to existing datasets, covering areas such as the brain, neck, chest, abdomen, and pelvic cavity. Here, we only consider the English subset throughout the experiments.
- **PathVQA [3]** is a VQA benchmark consisting of pathology images. It includes 4,998 images accompanied by 32,799 question-answer pairs. Each image is associated with multiple questions addressing various aspects such as location, shape, color, and appearance. The questions are categorized into two main types: open-form (e.g., "what," "why," "how," "where") and closed-form, encompassing a wide range of question styles relevant to pathology interpretation.

We evaluate the performance of LLaDA-MedV and report baseline results outlined in [5] across all three VQA benchmarks, where available. However, for more detailed analysis, we focus exclusively on VQA-RAD and SLAKE to ensure evaluation efficiency.

Dataset Details	VQA-RAD		SLAKE			PathVQA		
	Train	Test	Train	Val	Test	Train	Val	Test
# Images	313	203	450	96	96	2599	858	858
# QA	1797	451	4919	1053	1061	19755	6279	6761
# Closed-form	1027	272	1943	422	416	9806	3135	3391
# Open-form	770	179	2976	631	645	9949	3144	3370

Table 3. Statistics of the downstream VQA benchmarks. # denotes the number of samples in each split.

B. Visualization

B.1. Open-end Biomedical Conversation

Due to space limitations in the main text, we provide additional visualizations for open-ended biomedical conversations in Fig.1, Fig.2, and Fig. 3. As shown, LLaDA-MedV consistently produces relatively longer responses enriched with more contextual information. For all visualized examples, we use the following generation settings (as used in main setting): $L = 256$, $B = 64$, and $Z = 256$. For clarity, we omit GPT-4 reference answer.

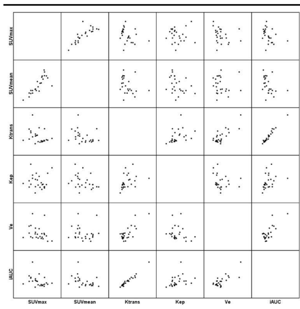


Figure 1. Illustration of LLaDA-MedV response in open-end biomedical conversation task. Image, question and corresponding caption comes from [5]

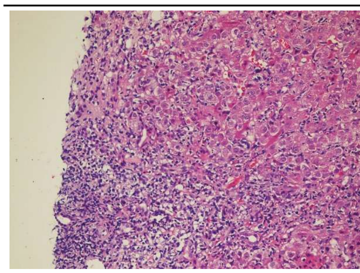


Figure 2. Illustration of LLaDA-MedV response in open-end biomedical conversation task. Image, question and corresponding caption comes from [5]

B.2. Inference-Time Key Factors

In our investigation of inference-time factors influencing LLaDA-MedV, we systematically evaluate the effects of response length (L), block length (B), and the number of sampling steps (Z) on final generation quality. This section is dedicated to providing additional visualizations to facilitate a deeper understanding of their impact.

Specifically, We observe that variations in the generation length L impact overall performance in open-ended biomedical conversation tasks, as shown in Fig. 4. This discrepancy may raise initial concerns about the fairness of

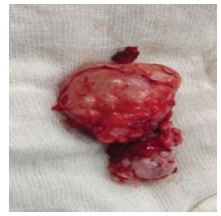
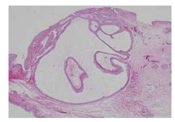


Figure 3. Illustration of LLaDA-MedV response in open-end biomedical conversation task. Image, question and corresponding caption comes from [5]



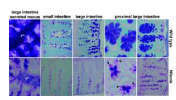
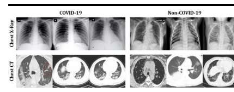
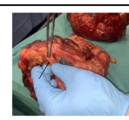


Figure 4. Illustration of LLaDA-MedV responses on open-end biomedical conversation as response length L changes. Image, question and corresponding caption comes from [5]. We omit GPT-4 reference answer for clarity.





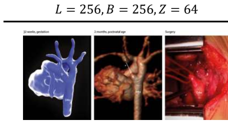
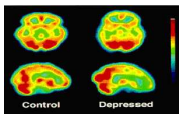


Figure 5. Illustration of LLaDA-MedV responses on open-end biomedical conversation as sampling steps Z changes. Image, question and corresponding caption comes from [5]. We omit GPT-4 reference answer for clarity.

our comparisons, since longer responses typically provide richer context to GPT-based evaluators, potentially leading to higher perceived answer quality. However, we emphasize that all evaluations are conducted under a controlled setting in which models are encouraged to generate responses



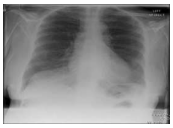
$L = 256, B = 32, Z = 256$

Question: What does hypofrontality indicate?

Caption: Single photon emission computed tomography (SPECT) images from a depressed patient showing characteristic hypofrontality relative to a healthy control subject

LLaDA-MedV: Hypofrontality is a term used to describe a pattern in brain imaging, such as functional magnetic resonance imaging (fMRI), where there is a decrease in the activity of the frontal regions of the brain. The frontal regions of the brain include areas such as the prefrontal cortex, orbitofrontal cortex, and cingulate cortex, which are involved in various cognitive functions, including decision-making, problem-solving, and emotional regulation. Hypofrontality can be associated with various neurological and psychiatric conditions. Some of these include: 1. Depression: A common mood disorder characterized by a loss of interest or pleasure in activities, worthlessness, and difficulty concentrating. 2. Autism spectrum disorders: A group of disorders that affect the way a person interacts with others and experiences the world around them. 3. Parkinson's disease: A progressive neurodegenerative disorder that affects movement and can lead to cognitive decline. 4. Huntington's disease: A genetic disorder that causes progressive cognitive decline and psychiatric symptoms. It is important to note that the presence of hypofrontality alone does not necessarily indicate a specific diagnosis. Further evaluation, including clinical history, symptoms, and other diagnostic tests, is needed to determine the underlying cause of the hypofrontality.

Figure 6. Illustration of LLaDA-MedV responses on open-end biomedical conversation as block length B changes. Image, question and corresponding caption comes from [5]. We omit GPT-4 reference answer for clarity.

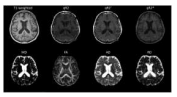


$L = 256, B = 128, Z = 256$

Question: Are there any signs of lymph node involvement?

Caption: Chest X-ray: Normal with no signs of metastatic lesions

LLaDA-MedV: The chest X-ray does not show any signs of lymph node involvement. However, it is important to remember that a chest X-ray is just one diagnostic tool, and further evaluation, such as a CT scan or MRI, may be necessary to rule out any potential lymph node involvement or other underlying conditions.



$L = 256, B = 256, Z = 256$

Question: Describe the following image in detail

Caption: Selected slice from a multi-modal MRI dataset of a patient with Parkinson's disease.

LLaDA-MedV: The image is an MRI (Magnetic Resonance Imaging) scan of the brain. It includes three types of sequences: T1-weighted, T2-weighted, and diffusion-weighted imaging (DWI). 1. T1-weighted images: These are the most common type of MRI sequence and provide good contrast between different tissues. They typically show bright areas of water (nerves) and dark areas of fat (muscle). 2. T2-weighted images: These images are similar to T1-weighted images but provide different contrast between different tissues. They typically show dark areas of water (nerves) and bright areas of fat (muscle). 3. Diffusion-weighted imaging (DWI): DWI is a specialized MRI technique that measures the diffusion of water molecules within tissues. It is particularly useful for detecting changes in tissue microstructure, such as those caused by inflammation, tumors, or other pathological processes. In this particular image, the T1-weighted, T2-weighted, and DWI sequences are used to visualize different aspects of the brain, such as the distribution of water, the presence of fat, and the movement of water molecules within the tissue. This can help healthcare professionals diagnose and monitor various neurological conditions and assess the effectiveness of treatments.

Figure 7. Illustration of LLaDA-MedV responses on open-end biomedical conversation as block length B changes. Image, question and corresponding caption comes from [5]. We omit GPT-4 reference answer for clarity.

of fixed length ($L = 256$). For ARMs such as LLaVA-Med, this constraint is enforced by setting the maximum token length to 256. In contrast, LLaDA-MedV leverages a masked prediction mechanism, which inherently encourages the generation of more complete and informative responses within the same length constraint. Additionally, differences in benchmark design and evaluation protocols may also contribute to the observed variation in performance trends. We further find that the number of sampling steps Z plays an important role in controlling response diversity. As shown in Fig. 5, when generating long responses, an insufficient number of steps can lead to noticeable token repetition. Lastly, when using semi-autoregressive generation, the choice of block length B must be made with care. To aid understanding, we provide qualitative visualizations in Fig. 6 and Fig. 7.

References

[1] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023. 3

[2] Huiwen Chang, Han Zhang, Lu Jiang, Ce Liu, and William T Freeman. Maskgit: Masked generative image transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11315–11325, 2022. 1

[3] Xuehai He, Yichen Zhang, Luntian Mou, Eric Xing, and Pengtao Xie. Pathvqa: 30000+ questions for medical visual question answering. *arXiv preprint arXiv:2003.10286*, 2020. 2, 3

[4] Jason J Lau, Soumya Gayen, Asma Ben Abacha, and Dina Demner-Fushman. A dataset of clinically generated visual questions and answers about radiology images. *Scientific data*, 5(1):1–10, 2018. 2, 3

[5] Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36:28541–28564, 2023. 1, 2, 3, 4, 5

[6] Bo Liu, Li-Ming Zhan, Li Xu, Lin Ma, Yan Yang, and Xiao-Ming Wu. Slake: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In *2021 IEEE 18th international symposium on biomedical imaging (ISBI)*, pages 1650–1654. IEEE, 2021. 2, 3

[7] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 3

[8] Meta AI. Llama-3.2-11B-Vision: A Multimodal Vision–Language LLM. Model card via Meta AI, 2024. Version released September 25, 2024; instruction-tuned for image reasoning, captioning, and VQA with 10.6B parameters. 3

[9] Michael Moor, Qian Huang, Shirley Wu, Michihiro Yasunaga, Yash Dalmia, Jure Leskovec, Cyril Zakka, Eduardo Pontes Reis, and Pranav Rajpurkar. Med-flamingo: a multimodal medical few-shot learner. In *Machine Learning for Health (ML4H)*, pages 353–367. PMLR, 2023. 3

[10] Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models. *arXiv preprint arXiv:2502.09992*, 2025. 1, 2

[11] OpenAI. Introducing gpt-4.1 in the api. https://openai.com/index/gpt-4-1/, 2025. Includes GPT-4.1, GPT-4.1 mini, and GPT-4.1 nano. 2

[12] Jiazhen Pan, Che Liu, Junde Wu, Fenglin Liu, Jiayuan Zhu, Hongwei Bran Li, Chen Chen, Cheng Ouyang, and Daniel Rueckert. Medvlm-r1: Incentivizing medical reasoning capability of vision-language models (vlms) via reinforcement learning. *arXiv preprint arXiv:2502.19634*, 2025. 3

[13] Samyam Rajbhandari, Olatunji Ruwase, Jeff Rasley, Shaden Smith, and Yuxiong He. Zero-infinity: Breaking the gpu memory wall for extreme scale deep learning. In *Proceedings of the international conference for high performance*

computing, networking, storage and analysis, pages 1–14, 2021. 2

[14] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025. 1, 2

[15] Zebin You, Shen Nie, Xiaolu Zhang, Jun Hu, Jun Zhou, Zhiwu Lu, Ji-Rong Wen, and Chongxuan Li. Llada-v: Large language diffusion models with visual instruction tuning. *arXiv preprint arXiv:2505.16933*, 2025. 1, 3

[16] Wenhui Zhu, Xuanzhao Dong, Xin Li, Peijie Qiu, Xiwen Chen, Abolfazl Razi, Aris Sotiras, Yi Su, and Yalin Wang. Toward effective reinforcement learning fine-tuning for medical vqa in vision-language models. *arXiv e-prints*, pages arXiv–2505, 2025. 3

[17] Wenhui Zhu, Xin Li, Xiwen Chen, Peijie Qiu, Vamsi Krishna Vasa, Xuanzhao Dong, Yanxi Chen, Natasha Lepore, Oana Dumitrascu, Yi Su, et al. Retinalgpt: A retinal clinical preference conversational assistant powered by large vision-language models. *arXiv e-prints*, pages arXiv–2503, 2025. 3