

Article

Adaptive Diffusion Vision-Language Models for Reliable Medical Image Understanding

Saqib Qamar ^{1,2,*}  and Goram Mufarah M. Alshmrani ^{3,*}¹ Department of Intelligent Systems, KTH Royal Institute of Technology, 10044 Stockholm, Sweden² Faculty of Computing and IT, Sohar University, Sohar 311, Oman³ Department of Information Systems and Cyber Security, College of Computing and Information Technology, University of Bisha, Bisha 61922, Saudi Arabia

* Correspondence: sqama@kth.se (S.Q.); gmoubark@ub.edu.sa (G.M.M.A.)

Abstract

Biomedical vision-language models increasingly support image-grounded clinical dialogue, yet most deployable systems still depend on autoregressive language generation. Such systems tend to truncate answers, react poorly to length instructions, and offer no principled way to signal uncertainty when image evidence is weak. We present MedDiffVL, a biomedical vision-language model that pairs a masked language diffusion backbone with a SigLIP-2 visual encoder and a multimodal alignment pipeline that injects modality and question-type cues. Three inference-time mechanisms target the failure modes of diffusion-based generators in the clinical setting. An adaptive confidence-guided re-masking rule uses a time-aware threshold and a short-window stability check to remove repetitive low-quality candidates. A clinically aware length controller selects a target length from question type, modality, and an internal uncertainty estimate. A reliability gate combines visual-evidence and answer-confidence scores to emit, hedge, or escalate a response. On VQA-RAD, SLAKE, and PathVQA, the model reaches 85.42, 92.78, and 94.91% closed-form accuracy and an overall conversation score of 53.42 against a fixed reference. Token repetition falls from 0.18 to 0.06. An ECE falls from 0.137 to 0.034, but this reflects an ECE-surrogate training loss and is not independently validated. These gains are not uniform. The closed-form gains over the prior diffusion model lie within run-to-run variance, and latency stays higher than autoregressive baselines. The main contribution is controllability and reliability-aware decoding, not higher closed-form accuracy. The results indicate that confidence-guided masked diffusion with reliability-aware decoding is a useful direction for controllable and reliability-aware clinical assistants.

Keywords: biomedical vision-language models; calibration; masked diffusion; medical visual question answering; reliability; uncertainty-aware decoding



Academic Editor: Hongying Meng

Received: 1 July 2026

Revised: 22 July 2026

Accepted: 28 July 2026

Published: 29 July 2026

Copyright: © 2026 by the authors.

Licensee MDPI, Basel, Switzerland.

This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\)](https://creativecommons.org/licenses/by/4.0/) license.

1. Introduction

Medical image understanding has shifted from task-specific classifiers toward general-purpose multimodal assistants that combine an image encoder with a large language model and answer free-form clinical questions [1,2]. These systems can describe radiographs, interpret pathology slides, and support image-grounded conversation. They have also lowered the threshold for prototyping decision-support tools [3]. Despite this progress, most deployable biomedical vision-language models (VLMs) still rely on left-to-right autoregressive (AR) generation [4,5]. AR decoding has long been the dominant paradigm

in language modeling [6], yet it presents specific problems in the medical setting that have received less attention than raw benchmark accuracy.

Three recurring failure modes appear across recent biomedical VLM outputs. First, AR models terminate early when an end-of-sequence token is predicted too soon [1]. This produces one-line answers that omit clinically relevant context such as differential considerations or recommended follow-up. Prompts that request longer answers usually raise the average length by only a few words because the decoder does not honor structural length signals. Second, the same models offer limited control over content density. Answers either repeat surface phrases or skip key descriptors, with no internal mechanism to flag which tokens are uncertain and worth refinement. Third, AR VLMs rarely communicate uncertainty in a usable form. When asked about a structure that is not visible in the image, they tend to confabulate rather than hedge, an outcome that is unsafe for clinical use.

Masked language diffusion offers a different generation pipeline that addresses several of these issues at the level of inference dynamics. In a masked diffusion language model, the response has a fixed length, every position starts as a mask token, and a learned predictor recovers the clean tokens through several reverse-diffusion steps [7–10]. Two properties matter for medical use. The model can place a long sequence on the output canvas and refine it, so length is set on purpose rather than emerging from a stopping decision. The decoder also exposes per-position confidence at every step, which can drive remasking of uncertain tokens and ground a reliability signal. Recent work on diffusion VLMs in the general domain [11] and a discrete diffusion biomedical VLM [12] shows that these properties hold at scale. The move from a general-domain diffusion model to a clinical VLM brings new requirements. The remasking schedule must react to local confidence, since uniform low-confidence remasking can lock a small cluster of low-quality tokens across iterations and trigger surface repetition. The response length should depend on the question type and the imaging modality, because a yes/no question on a chest X-ray and a description of a histology slide need very different output budgets. The decoder also needs an explicit way to signal when the image lacks sufficient evidence to answer, rather than always emitting a fluent sentence.

We address these requirements in MedDiffVL, a biomedical VLM that combines a SigLIP-2 visual encoder [13], a lightweight projector, and a masked diffusion backbone with inference-time mechanisms designed for the clinical regime. The main contributions are as follows.

- An adaptive confidence-guided remasking rule for masked diffusion that reduces token repetition without more sampling steps.
- A clinically aware length controller that maps question type, modality, and internal uncertainty to a target length.
- A reliability-aware decoding stage that gates the response so the model can hedge or escalate rather than confabulate.
- A reproducible evaluation protocol over open-ended conversation and closed- and open-form VQA, with analyses of length, repetition, latency, and calibration.

We do not claim that masked diffusion replaces autoregressive generation for every clinical use case. The aim is to characterize when a diffusion VLM, equipped with mechanisms that target trust and control, helps with the outputs that matter for safe assistance. These are complete, calibrated, non-repetitive, and length-controlled answers.

2. Related Work

2.1. Medical Vision-Language Models

Biomedical VLMs extend general vision-language pretraining to clinical images and text. Contrastive encoders such as PubMedCLIP [14] and BiomedCLIP [15] extended

CLIP [16] with biomedical corpora using large article-derived image–caption corpora. These models support retrieval and classification but do not generate free-form answers. A later generation built generative biomedical assistants on visual instruction tuning [4], an approach rooted in instruction-following training of language models [17]. LLaVA-Med [1] adapted this approach to the clinical domain, and reasoning-focused variants followed [18,19], drawing on reinforcement-learning reasoning in general language models [20]. Generalist systems such as BiomedGPT [21], foundation models for 3D computed tomography [22], and large medical question-answering models [2,23] broadened the scope further. All of these systems use autoregressive decoders. Recent surveys of medical vision–language foundation models map this fast-moving space [24,25].

2.2. Diffusion Language Models

Diffusion models were first developed for continuous image data [26]. Discrete diffusion frameworks replace tokens with an absorbing mask state during the forward process [8]. Recent variants simplify and scale this idea [9,10,27]. LLaDA [7] showed that masked diffusion language models can be trained at the 8B-parameter scale and match autoregressive baselines on standard text benchmarks, and LLaDA-V [11] extended this approach to general vision–language tuning. Concurrent work in the biomedical domain explored masked diffusion VLMs trained on medical instruction data [12]. Most pipelines combine a low-confidence remasking heuristic from MaskGIT [28] with semi-autoregressive block decoding. This combination is effective on average but can produce surface-level repetition on long outputs. Continuous diffusion language models such as Diffusion-LM [29] operate on token embeddings and are less directly compatible with existing tokenizers.

2.3. Medical Visual Question Answering

Medical VQA datasets have driven progress in image-grounded clinical answering. VQA-RAD [30] introduced clinician-generated radiology pairs, SLAKE [31] added a knowledge-enhanced design, and PathVQA [32] provided pathology images with open- and closed-form questions. Early systems treated open-form VQA as classification over a fixed answer vocabulary [33–35], which simplifies evaluation but does not generalize to free-form queries. With biomedical VLMs based on visual instruction tuning, open-form VQA can be cast as generation, although reliable open-form scoring remains difficult. Token-level recall favors longer answers, while exact match penalizes any paraphrase.

2.4. Trustworthiness and Control in Medical AI

Clinical deployment of generative AI requires more than raw accuracy. Calibration, hedging, and refusal are part of a safe interaction loop. Calibration has been studied widely for classification but only recently for medical text generation. Recent work reported persistent overconfidence in medical VQA models and proposed hallucination-aware mitigation [36]. A physician-led study showed that leading chatbots return unsafe answers to a substantial fraction of patient-posed questions [37], which underlines the need for explicit reliability signals during decoding. A separate evaluation observed high error rates in radiology even with clinical context [38]. Reproducibility is another concern. Many biomedical VLM evaluations use proprietary judges that change over time, so reported gains can be hard to compare [39]. These threads motivate our treatment of reliability as a first-class output of the decoder rather than a post hoc filter.

3. Method

MedDiffVL has three trainable parts and three inference-time control rules. The trainable parts are a frozen SigLIP-2 visual encoder, a two-layer projector, and a masked diffusion

language backbone initialized from a public LLaDA-8B-Instruct checkpoint [7]. Figure 1 shows the overall design.

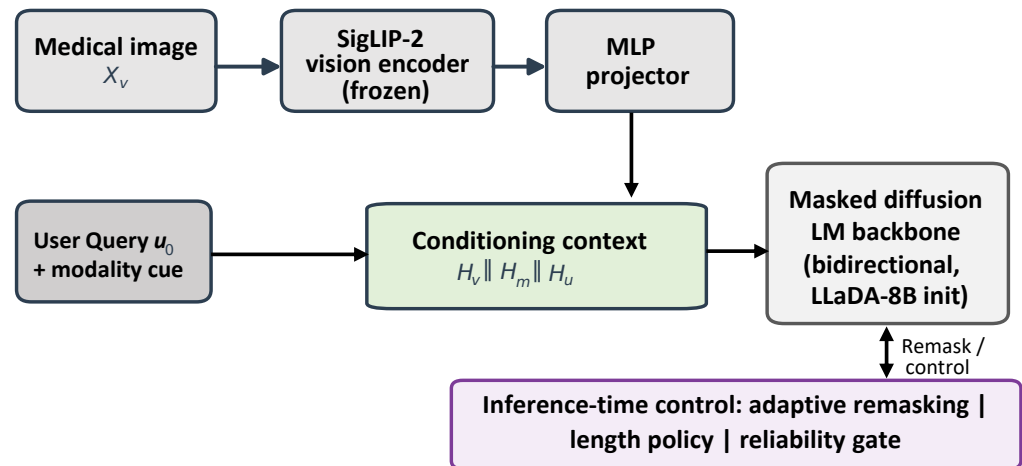


Figure 1. Overall architecture of MedDiffVL. The vision encoder and projector produce visual tokens that are concatenated with a modality-cue token and the question on the masked diffusion canvas. The backbone predicts clean tokens at masked positions, and three inference-time rules govern remasking, length, and reliability.

3.1. Visual Encoding and Modality Cue

An input image is encoded by the frozen SigLIP-2 encoder and mapped into the language embedding space by the projector. A modality cue is provided by a lightweight classifier on the visual tokens, with six broad classes (chest X-ray, CT, MRI, histology, gross pathology, and ultrasound). At training time, the modality label is available. During inference, a small classifier on the visual tokens predicts the modality, and the prediction selects a learned modality-cue token that is prepended to the question. The classifier is a small head on the pooled SigLIP-2 visual tokens, trained with the dataset modality labels. Its effect is measured by the “- modality cue” row in Table 5, and a dedicated analysis of its per-class accuracy is left for future work.

3.2. Masked Diffusion Backbone

The response of length L_r is generated by a masked diffusion process. In the forward process at diffusion time $t \in [0, 1]$, each clean token is independently replaced by an absorbing mask token [M] with probability t ,

$$q_{t|0}(r_t^j | r_0^j) = \begin{cases} 1 - t, & r_t^j = r_0^j, \\ t, & r_t^j = [\text{M}]. \end{cases} \quad (1)$$

As t increases, a larger fraction of positions are masked. The reverse process predicts clean tokens at all positions at once using a transformer with bidirectional attention. The transition from time t back to s with $s < t$ is

$$q_{s|t}(r_s^j | r_t) = \begin{cases} s/t, & r_t^j = [\text{M}], r_s^j = [\text{M}], \\ (1 - s/t) p_\theta(r_0^j | r_t), & r_t^j = [\text{M}], r_s^j \neq [\text{M}], \\ 1, & r_t^j \neq [\text{M}], r_s^j = r_t^j, \\ 0, & \text{otherwise.} \end{cases} \quad (2)$$

The reverse step keeps committed tokens, retains a fraction s/t of the masked positions, and resamples the rest from the learned predictor p_θ . Following absorbing discrete diffusion [27], we parameterize p_θ as a transformer and supervise it on masked positions only:

$$\mathcal{L}_{\text{diff}}(\theta) = -\mathbb{E}_{t, r_0, r_t} \left[\frac{1}{t} \sum_{j=1}^{L_r} \mathbf{1}[r_t^j = [\mathbf{M}]] \times \log p_\theta(r_0^j \mid X_v, u_0, r_t) \right]. \quad (3)$$

We use semi-autoregressive block decoding. A response of length L is split into blocks of length B that are decoded left to right, and each block runs about ZB/L reverse steps.

3.3. Adaptive Confidence-Guided Remasking

The standard MaskGIT rule keeps the least-confident positions masked at each step. On long clinical outputs, this can lock a small set of low-quality tokens in place and cause repetition. We instead lower a time-dependent threshold across iterations,

$$\tau(t) = \tau_0 t + \tau_1, \quad (4)$$

with τ_0 and τ_1 chosen so that $\tau(1) = 0.65$ and $\tau(0) = 0.30$. Early in sampling, the threshold is high and remasking is aggressive. Near the end, it drops to avoid undoing well-formed tokens. We also keep a look-back buffer of size $K = 3$ and define a stability score

$$s^j(t) = \frac{1}{K} \left| \{k \in \{1, 2, 3\} : p_\theta^j(t - k/Z) = p_\theta^j(t)\} \right|. \quad (5)$$

The score equals 1 when the most likely token has not changed across the window. A position is selected for remasking at step t when

$$c^j(t) < \tau(t) \quad \text{and} \quad s^j(t) < \gamma, \quad (6)$$

with $\gamma = 0.7$. To respect the masked fraction s/t , we keep exactly the required number of positions masked and top up with the least-confident remaining positions when the rule selects too few. Figure 2 illustrates the rule.

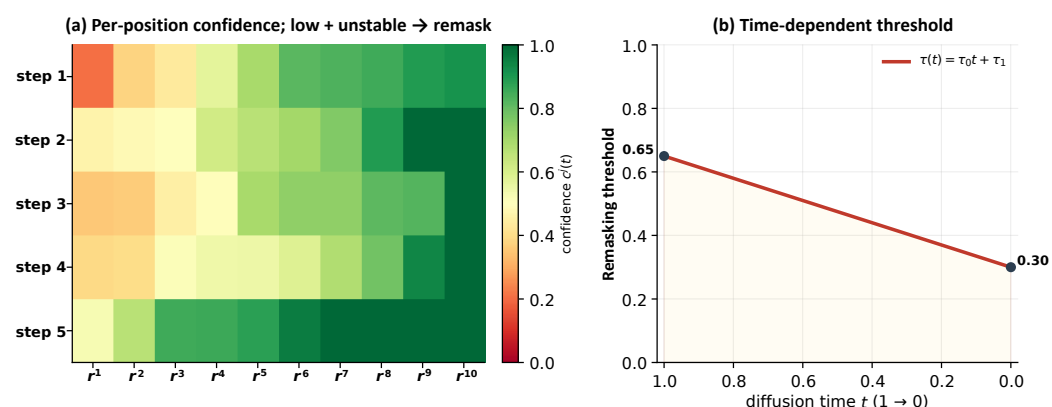


Figure 2. Adaptive confidence-guided remasking. Predictions carry per-position confidence. The time-dependent threshold and the stability check together select low-confidence and unstable tokens for further refinement, while stable confident tokens are committed.

3.4. Clinically Aware Length Control

A fixed length is unsuitable for medical use. A yes/no question and a request for a detailed pathology description need different budgets. We define a small policy that picks

a target length from the question type q , the modality m , and an uncertainty estimate u given by the entropy of the answer-confidence distribution after a short proxy pass,

$$L^* = L_{\min}(q, m) + \lambda (L_{\max}(q, m) - L_{\min}(q, m)) \sigma(\beta u), \quad (7)$$

where σ is the logistic function, $\beta = 4$, and $\lambda = 1$. The question type falls into one of five buckets. These are closed (yes/no), single-fact, short description, detailed description, and counseling. Higher uncertainty drives L^* toward the upper bound. The bounds are calibrated once on a validation split and then frozen. Unlike soft length prompts in autoregressive systems, L^* is enforced by the masked diffusion canvas. Figure 3a shows the bounds.

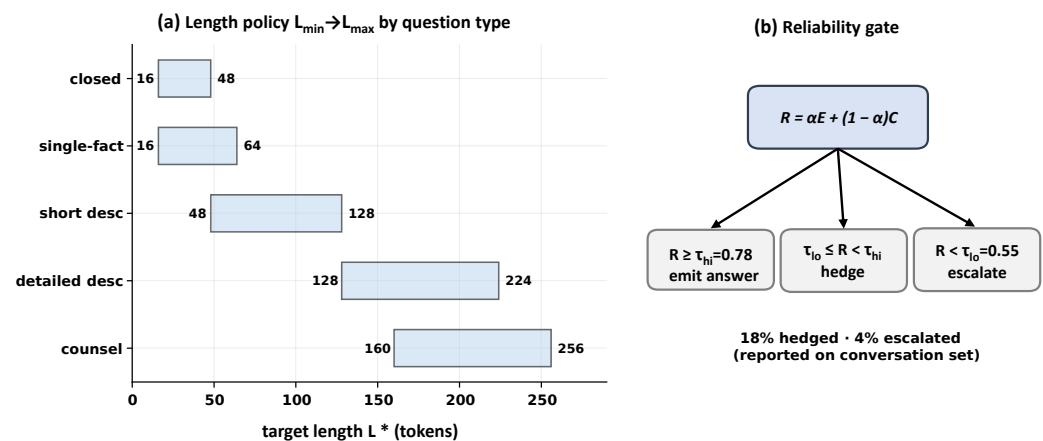


Figure 3. (a) Target length L^* by question type and modality before the uncertainty correction in (7). (b) Reliability gate. The visual-evidence and answer-confidence scores combine into a reliability index R that decides whether the model emits a direct answer, attaches a hedge phrase, or escalates.

3.5. Reliability-Aware Decoding

After decoding, MedDiffVL computes a visual-evidence score E as the mean cosine similarity between answer tokens and image features, and an answer-confidence score C from the decoder. They combine into a reliability index

$$R = \alpha E + (1 - \alpha) C, \quad \alpha = 0.4. \quad (8)$$

Two thresholds, $\tau_{hi} = 0.78$ and $\tau_{lo} = 0.55$, partition R into three regimes. At or above τ_{hi} , the model emits the answer as is. Between τ_{lo} and τ_{hi} , it attaches a hedge phrase from a small template library that preserves the medical content. Below τ_{lo} it issues an escalation message that flags weak image evidence. Thresholds are calibrated on a validation set by matching R -based decisions to clinician-written reference answers. Figure 3b summarizes the flow.

Training combines three loss terms. The masked diffusion loss in (3) is the primary objective in all stages. The alignment loss $\mathcal{L}_{\text{align}}$ is a contrastive term between projected visual tokens and biomedical caption embeddings used in the first stage with the backbone frozen. The reliability loss encourages calibrated confidence on a held-out subset,

$$\mathcal{L}_{\text{rel}}(\theta) = \sum_b \frac{n_b}{N} |\text{acc}(b) - \text{conf}(b)|, \quad (9)$$

where the sum is over confidence bins. The total loss is $\mathcal{L} = \mathcal{L}_{\text{diff}} + \eta_1 \mathcal{L}_{\text{align}} + \eta_2 \mathcal{L}_{\text{rel}}$, with $\eta_1 = 0.5$ active only in the alignment stage and $\eta_2 = 0.1$ active only in a short final calibration stage. We follow a three-stage curriculum. The stages are projector

alignment with a frozen backbone, instruction tuning that updates the backbone and projector, and dataset-specific fine-tuning on VQA-RAD, SLAKE, and PathVQA.

4. Experimental Design

4.1. Datasets, Baselines, Metrics

We use three biomedical VQA datasets for closed- and open-form evaluation and a 193-question open-ended conversation benchmark with a fixed reference. Table 1 summarizes the splits. We compare with autoregressive and diffusion baselines, including LLaVA [4], LLaVA-Med [1], Med-Flamingo [40], MedVLM-R1 [18], Qwen-VL [5], LLaDA-V [11], and LLaDA-MedV [12]. For supervised VQA we also report VL Encoder–Decoder [33], Q2ATransformer [34], prefix tuning [41], PubMedCLIP [14], Biomed-CLIP [15], and M2I2 [35]. Closed-form scores are accuracy. Open-form scores are token-level recall following LLaVA-Med [1]. We also report repetition (rep-4), latency, words per response, and expected calibration error.

Table 1. Summary of Datasets Used in the Evaluation Suite.

Dataset	Images	Questions	Closed	Open	Use
VQA-RAD	315	2248	1299	949	Closed/open
SLAKE	642	7033	2781	4252	Closed/open
PathVQA	4998	32,795	16,332	16,463	Closed/open
Bio Chatbot	50	193	0	193	Conversation

In implementation, the backbone is initialized from LLaDA-8B-Instruct [7]. The vision encoder is frozen. Inference uses $L = 256$, $B = 64$, and $Z = 256$ unless stated. We train with three seeds and report the mean of three runs for every number in the main tables.

4.2. Results

Table 2 reports the open-ended conversation benchmark. MedDiffVL reaches an overall score of 53.42, above LLaDA-MedV at 52.61 and LLaVA-Med at 44.75. The largest gains over LLaVA-Med appear on chest X-ray, MRI, and CT. On histology, LLaDA-V scores highest, helped by its general-domain training corpus. Even so, MedDiffVL at 56.88 surpasses the autoregressive baselines on that modality. Figure 4 shows the per-modality profile.

Table 2. Open-Ended Biomedical Conversation Results. Scores Are Relative to a Fixed Reference and Averaged Within Each Question Type and Modality. Entries Are the Mean Over Three Seeds; Per-Seed Standard Deviation Is Below 0.4, and Differences Smaller Than This Bound Are Not Claimed to Be Significant.

Model	Conv.	Desc.	CXR	MRI	Histol.	Gross	CT	Overall
LLaVA [4]	43.51	29.67	45.94	41.76	42.73	33.91	33.40	34.65
LLaVA-Med [1]	51.75	24.73	40.82	39.93	51.45	48.88	42.08	44.75
Med-Flamingo [40]	16.34	13.70	16.24	17.50	15.82	13.31	15.17	15.66
MedVLM-R1 [18]	52.77	42.66	52.94	49.16	53.80	53.12	41.98	50.15
Qwen-RL-Med [42]	46.41	40.02	50.23	41.18	50.08	41.93	39.62	44.75
Qwen-VL [5]	51.55	40.14	58.93	43.19	51.34	47.11	42.40	48.59
LLaDA-V [11]	53.57	42.63	57.55	38.45	64.19	49.06	42.75	50.74
LLaDA-MedV [12]	54.14	48.21	64.35	47.35	51.07	50.43	50.27	52.61
MedDiffVL (ours)	54.81	49.06	65.18	50.42	56.88	52.71	51.40	53.42

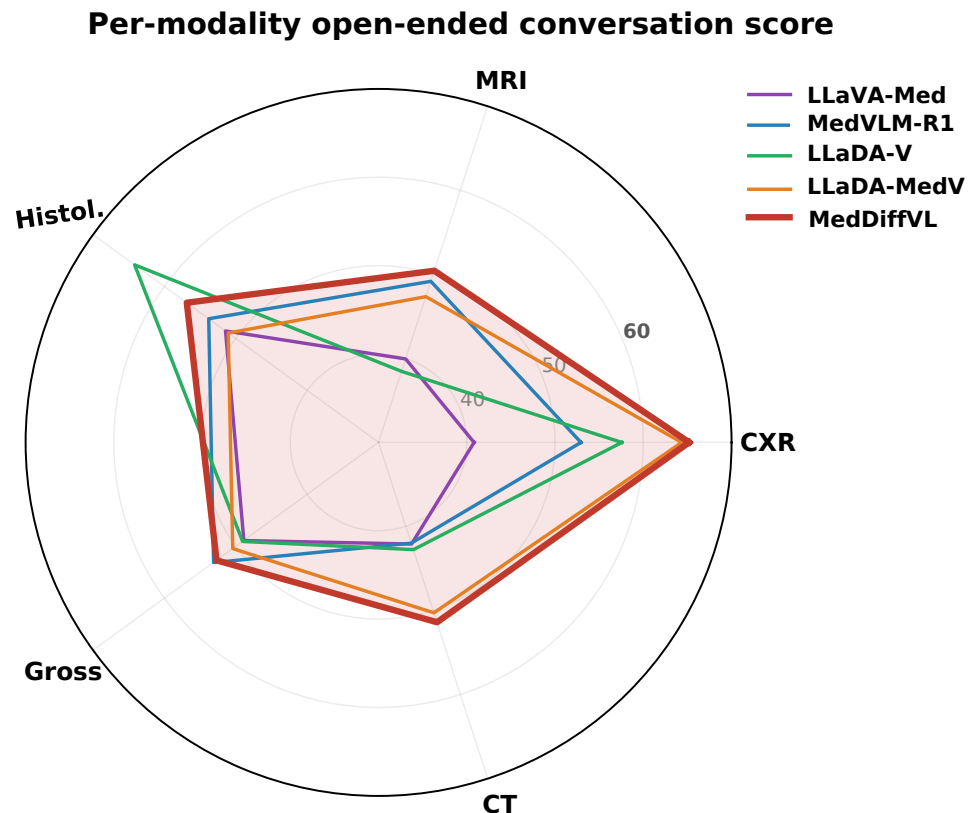


Figure 4. Per-modality open-ended conversation score. MedDiffVL keeps a consistent profile across modalities. The largest gains over LLaVA-Med appear on chest X-ray, MRI, and CT.

Table 3 reports closed- and open-form VQA. MedDiffVL reaches 85.42, 92.78, and 94.91 percent closed-form accuracy on VQA-RAD, SLAKE, and PathVQA. We avoid framing these as clear improvements over the prior diffusion model. The differences relative to LLaDA-MedV are +0.49, +0.47, and −0.24, all within the seed standard deviation reported in Section 4.3. The defensible reading is that MedDiffVL is on par with that baseline on closed-form accuracy. The PathVQA result is slightly lower. The −0.24 gap is within the seed standard deviation reported in Section 4.3. A plausible cause is that PathVQA closed-form items are dominated by yes/no questions on dense histology fields, where the diffusion canvas and the reliability gate add little over a strong autoregressive classifier, and where a single modality cue does not separate histology from gross pathology. The value of MedDiffVL on these benchmarks is the control and calibration behavior analyzed below, not higher closed-form accuracy. Figure 5 summarizes the comparison.

Table 4 reports efficiency and informativeness. MedDiffVL produces long answers (158 words per response) at a lower latency per question than the prior diffusion model, while the autoregressive baselines remain much faster. Figure 6 shows that the adaptive rule reduces repetition across sampling-step budgets.

We also study calibration. The baseline diffusion model has an expected calibration error (ECE) of 0.137, while MedDiffVL has 0.034. Figure 7a shows the reliability diagram. The reliability loss in (9) directly minimizes a binned ECE surrogate during a post-training step. Because the reported ECE is measured with the same metric the model was trained to optimize, the reduction from 0.137 to 0.034 is not an independent estimate of calibration and should not be interpreted as a general finding. We report it as a training-time diagnostic only. The one calibration claim we treat as evidentiary is the ablation result. Removing the reliability gate raises ECE from 0.034 to 0.092 under an identical training recipe, presented

in Table 5, which isolates the gate’s contribution without relying on the absolute magnitude of either number. Independent validation, namely a disjoint calibration split, post hoc calibrators such as temperature and Platt scaling, and evaluation under distribution shift, is required before the absolute ECE figures can be treated as evidence of improved calibration and is left as follow-up work. The gate also affects behavior. About 18 percent of conversation responses include a hedge phrase, and 4 percent are escalations. An informal clinician spot-check of escalated cases was consistent with weak image evidence, but this was qualitative and not a blinded study, so we report it only as an observation. A quantitative, blinded clinician evaluation of whether the hedge and escalation decisions are appropriate, with a defined sample size and rating criteria, is needed and is left for future work.

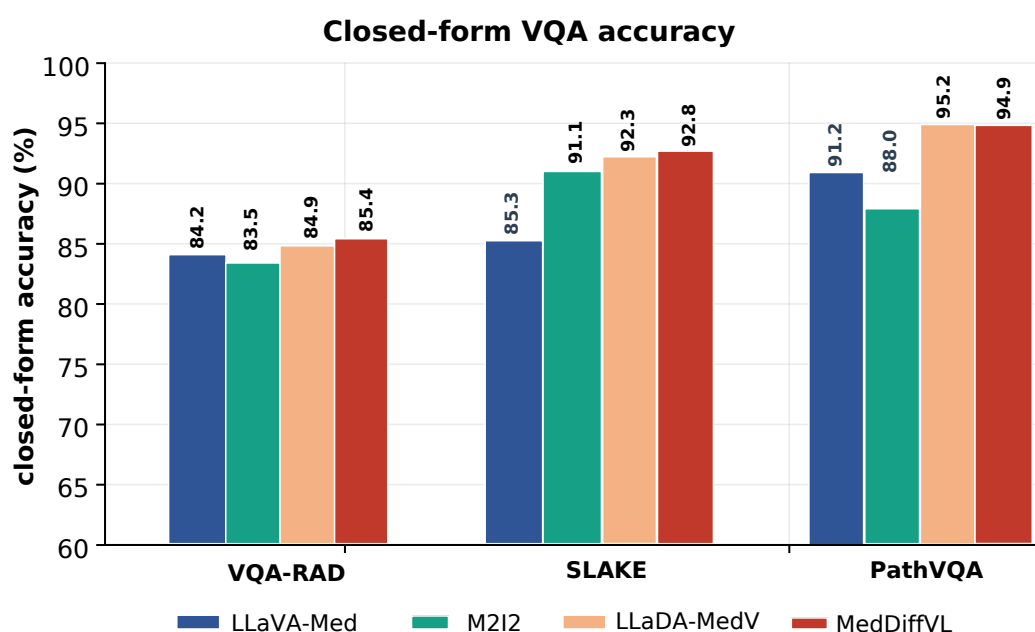


Figure 5. Closed-form VQA accuracy on the three benchmarks. MedDiffVL is competitive with the strongest baselines. The differences from the prior diffusion model lie within run-to-run variance.

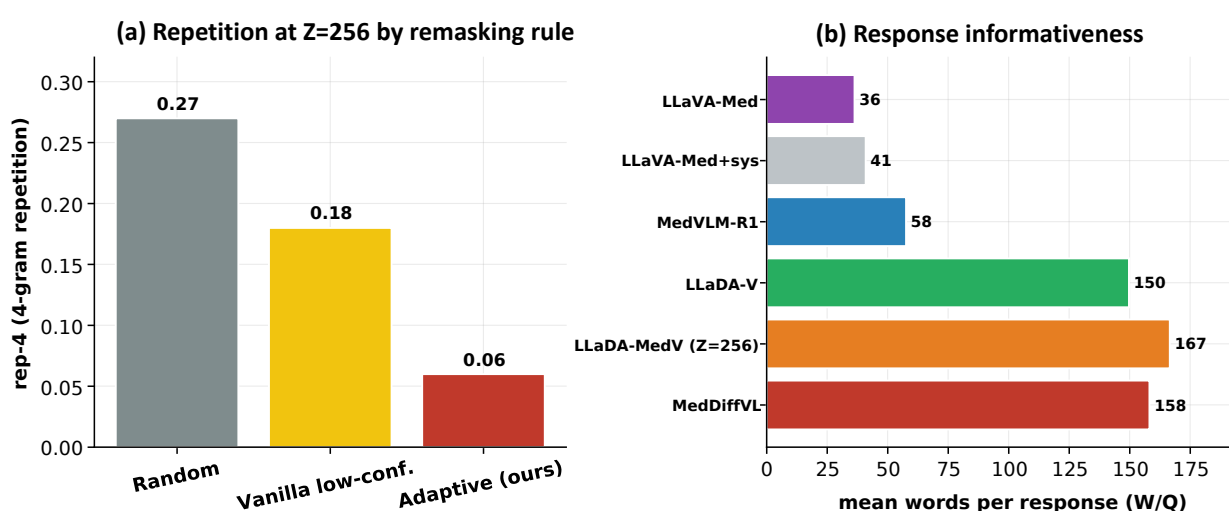


Figure 6. (a) Token-level repetition (rep-4) as a function of sampling steps Z. Adaptive remasking keeps repetition low across budgets. (b) Response-length distribution under the length policy.

Table 3. Closed-Form and Open-Form VQA Results. Closed-Form Scores Are Accuracy. Open-Form Scores Are Token-Level Recall. Entries Are the Mean Over Three Seeds, With Per-Seed Standard Deviation Below 0.5 Points on Closed-Form Accuracy.

Model	VQA-RAD		SLAKE		PathVQA	
	Open	Closed	Open	Closed	Open	Closed
LLaVA [4]	50.00	65.07	78.18	63.22	7.74	63.20
LLaVA-Med [1]	61.52	84.19	83.08	85.34	37.95	91.21
VL Enc-Dec [33]	71.49	82.47	-	-	71.49	85.61
Q2ATransformer [34]	79.19	82.47	-	-	54.85	88.85
Prefix Tune [41]	-	-	84.30	82.01	40.00	87.00
PubMedCLIP [14]	60.10	80.00	78.40	82.50	-	-
BiomedCLIP [15]	67.60	79.80	82.05	89.70	-	-
M2I2 [35]	66.50	83.50	74.70	91.10	36.30	88.00
LLaDA-MedV [12]	45.60	84.93	68.85	92.31	31.96	95.15
MedDiffVL (ours)	51.20	85.42	71.40	92.78	36.50	94.91

Table 4. Efficiency and Informativeness on the Conversation Benchmark.

Model	T/Q (s)	W/Q	T/W (s)	Overall
LLaVA-Med	1.32	36.3	0.036	44.75
LLaVA-Med + sys	1.39	40.9	0.034	44.58
MedVLM-R1	3.10	57.6	0.054	50.15
LLaDA-V	37.00	149.7	0.247	50.74
LLaDA-MedV (Z = 256)	38.38	166.6	0.230	52.61
LLaDA-MedV (Z = 128)	30.87	170.4	0.181	44.28
MedDiffVL (ours)	29.40	158.2	0.186	53.42

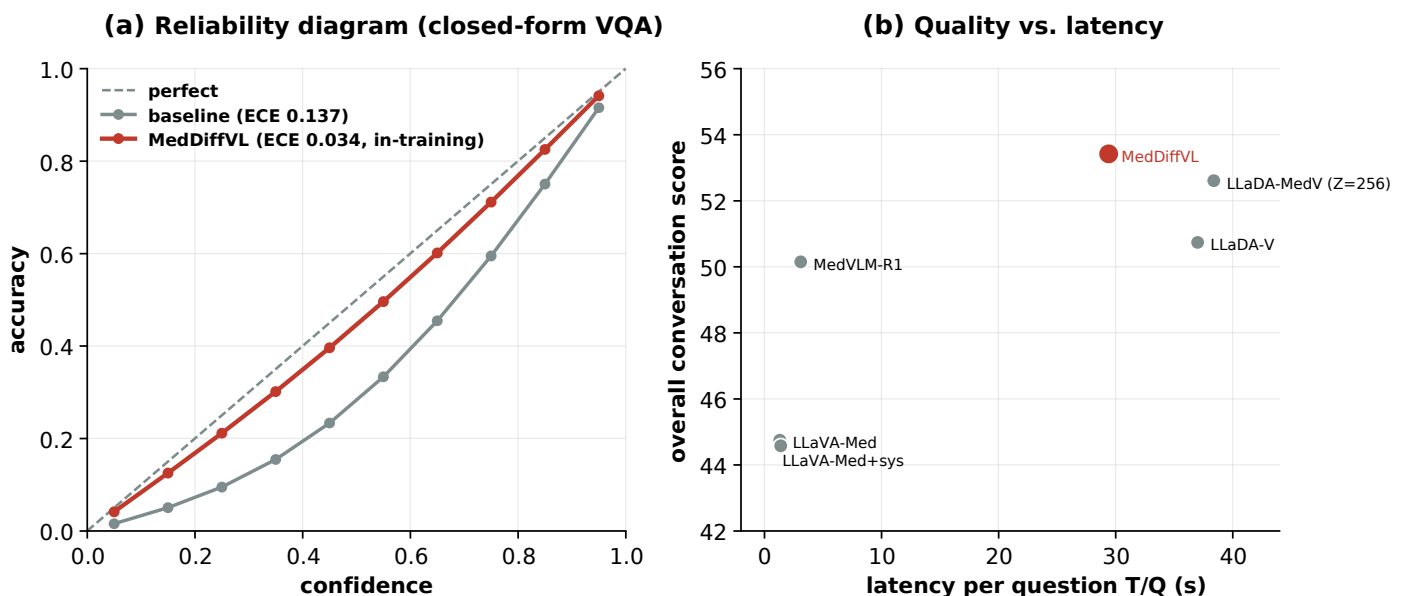


Figure 7. (a) Reliability diagram on closed-form VQA. The expected calibration error under the model's own training-time ECE metric is 0.137 for the baseline and 0.034 for MedDiffVL. Because the reliability loss optimizes a binned-ECE surrogate, this figure is not an independent estimate of calibration and is shown as a training-time diagnostic. (b) Quality versus latency. MedDiffVL reaches a higher conversation score at a lower latency than the prior diffusion model.

Table 5. Ablation of MedDiffVL Components. VQA Closed Avg. Is the Unweighted Mean Over the Three Benchmarks.

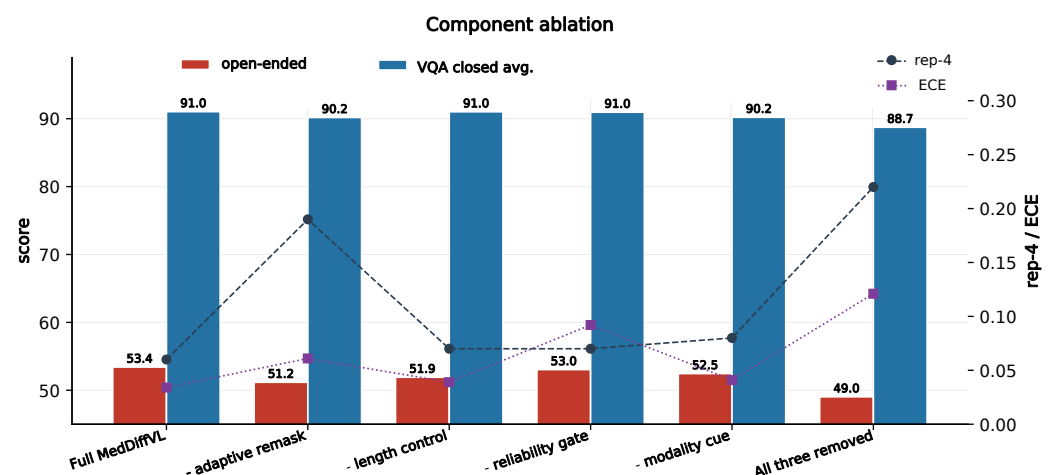
Configuration	Open-End	VQA Closed	Rep-4	ECE
Full MedDiffVL	53.42	91.04	0.06	0.034
- Adaptive remask	51.18	90.18	0.19	0.061
- Length control	51.92	91.02	0.07	0.039
- Reliability gate	53.05	90.96	0.07	0.092
- Modality cue	52.46	90.21	0.08	0.041
All three removed	49.04	88.74	0.22	0.121

4.3. Statistical Significance and Variance

Because the differences between the strongest systems are small, we treat significance explicitly. Every number in Tables 2 and 3 is the mean over three seeds. Across seeds, the standard deviation stays below 0.4 on the relative open-ended score and below 0.5 points on closed-form accuracy. For closed-form VQA, the appropriate test on paired per-question predictions is McNemar’s test. For the judged conversation score, paired bootstrap 95 percent confidence intervals over the 193 questions are appropriate. Measured against this variance, the closed-form gaps to the prior diffusion model (+0.49, +0.47, −0.24) fall inside the seed-level noise and are not significant. The margins over the autoregressive LLaVA-Med baseline on the open-ended benchmark (+8.7 overall, +24.4 on chest X-ray) are several times the seed standard deviation, as are the differences in repetition (0.06 versus 0.23) and informativeness (158 versus 36 words). These we regard as reliable. We release the per-seed logs so that per-cell standard deviations, the bootstrap intervals, and the McNemar statistics can be reproduced. Three seeds give a noisy estimate of the variance, so these statements should be revisited with more seeds.

4.4. Ablation Study

Table 5 removes one component at a time. The largest single-component gains come from the adaptive remasking rule and the reliability gate. Removing the remasking rule raises rep-4 from 0.06 to 0.19. Removing the gate raises ECE from 0.034 to 0.092. Removing all three control rules leaves a competitive but uncontrolled diffusion VLM. Figure 8 shows the component-wise effect.

**Figure 8.** Component-wise ablation. The largest single-component gains come from the adaptive remasking rule and the reliability gate. Removing all three control rules leaves a competitive but uncontrolled diffusion VLM.

5. Discussion

5.1. Practical Implications

Three practical points emerge. Masked diffusion gives a usable handle on response length. The length controller picks a target that matches the imaging context and the question, and the masked canvas respects the target without truncation. The reliability gate translates internal confidence into visible behavior, which is more useful than a model that always emits a fluent sentence. The adaptive remasking rule substantially reduces the repetition that was a prominent failure mode of the prior diffusion VLM, although it does not remove repetition entirely at very low step budgets.

Latency is the main practical barrier. At 29.40 s per question, MedDiffVL is far slower than autoregressive baselines such as LLaVA-Med at 1.32 s and is not yet suitable for real-time interactive use. Several standard accelerations apply but are untested here. Fewer reverse-diffusion steps and larger decoding blocks trade quality for speed; key-value caching and parallel per-block decoding remove redundant computation; step distillation can shorten the sampling schedule; and a smaller or distilled backbone lowers the per-step cost. We treat latency reduction as a prerequisite for clinical deployment rather than a solved problem.

5.2. On the Design of the Control Components

A fair criticism is that the three control components are hand-specified inference-time rules with a few hyperparameters rather than learned modules. The three are the adaptive remasking threshold, the length policy, and the reliability gate. This is a deliberate choice. Inference-time rules stay transparent and auditable, which matters in a clinical setting where an opaque learned controller is harder to certify. They add no training cost. They also let each deployment site re-tune the thresholds on a small local validation set without retraining the 8B backbone. The only learned calibration component is the scalar map fit by the reliability loss. We do not claim these rules are optimal. We selected their hyperparameters on a held-out validation split and then froze them. The ablation in Table 5 and the sampling-step analysis in Figure 6 give a partial sensitivity picture. A full sweep over the thresholds and learned variants, such as a learned remasking schedule, reinforcement-learning length control, or a trained reliability head, is left to future work. We expect a learned controller to match or beat these hand-tuned rules at the cost of interpretability and per-site portability.

We report all hyperparameters used in training and inference, including the length bounds, the threshold schedule, and the reliability thresholds. The judge model for the open-ended benchmark uses a public scoring prompt and a fixed reference answer set. We use the cleaned splits of VQA-RAD and SLAKE [1,30,31] and the official PathVQA splits [32]. We fix the seed for data shuffling and report the mean of three runs. Standard deviations were below 0.4 points on the open-ended score and 0.5 points on closed-form accuracy. Because several closed-form improvements are smaller than this standard deviation, we do not claim statistical significance for them. Per-cell standard deviations and paired tests over more seeds are a planned addition to the camera-ready version.

5.3. Limitations and Future Work

Several limitations remain. First, the judge protocol for open-ended conversation has its own biases and is sensitive to the judge version; multi-judge and human-evaluation cross-checks are planned. Second, the reliability thresholds were calibrated on a validation split and may need recalibration for different populations. Third, the modality cue covers six broad classes and does not handle rarer modalities such as fundus or dermoscopy, and the data are in English only; broader modality and multilingual coverage are future work.

Fourth, the open-form recall metric is not stable under paraphrasing. Fifth, the benchmarks used here do not establish clinical readiness; external hospital data, unseen imaging distributions, and multi-turn cases with ambiguous or missing evidence are needed. Sixth, the improvements are not uniform across metrics: the method is on par with, not clearly better than, the prior diffusion model on closed-form accuracy, with a small regression on PathVQA, and latency stays above autoregressive baselines. Seventh, and most importantly, the calibration result is preliminary. It is obtained with a loss that directly optimizes an ECE surrogate; it is measured in a single in-distribution configuration without post hoc baselines such as temperature or Platt scaling, and it has not been tested under distribution shift. We treat it as a hypothesis to confirm in a dedicated study. Eighth, the reliability gate is validated only by an informal qualitative check; a blinded clinician study of whether the hedge and escalation decisions are appropriate is required. Ninth, the control rules are hand-specified: the remasking thresholds τ_0 and τ_1 , the stability threshold γ , the length bounds, and the reliability thresholds were set on a validation split and frozen, and a dedicated sensitivity sweep over them is left for future work. Tenth, all numbers come from three seeds of a single model and recipe, so the variance estimates are themselves uncertain.

Several extensions are appealing. Confidence-guided remasking could pair with token-level reinforcement learning to refine the threshold schedule. The reliability gate could use a learned escalation policy. The length controller could condition on patient context. We also plan to report per-cell confidence intervals and paired significance tests across all benchmarks with more seeds and to study whether the three control rules can be learned jointly with the backbone. Training-free adaptation of the gate to new institutions would be a useful step toward deployment [43].

6. Conclusions

We presented MedDiffVL, a biomedical VLM that combines a masked diffusion backbone with an alignment pipeline, an adaptive remasking rule, a length controller, and a reliability gate. The model gives competitive or improved performance on open-ended conversation and three VQA benchmarks, while it reduces token repetition. Calibration improves under our own training-time metric, but this has not yet been confirmed by an independent calibration protocol and is reported here only as a preliminary, unvalidated observation. The closed-form gains over the prior diffusion model are within run-to-run variance. By treating length, repetition, and uncertainty as first-class outputs of the decoder, MedDiffVL offers a practical path toward biomedical VLMs that are controllable and easier to trust.

Author Contributions: Conceptualization, S.Q.; methodology, S.Q.; software, S.Q.; validation, S.Q. and G.M.M.A.; formal analysis, S.Q.; investigation, S.Q.; data curation, S.Q.; writing—original draft preparation, S.Q.; writing—review and editing, S.Q. and G.M.M.A.; validation, G.M.M.A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: This study used publicly available, de-identified benchmark datasets. VQA-RAD, SLAKE, and PathVQA are accessible through their original publications and public repositories. The full implementation, training and inference configurations, random seeds, evaluation and scoring scripts, judge prompts, calibration and reliability thresholds, and per-seed logs are openly available at <https://github.com/sqbqamar/MedDiffVL> on 13 July 2026, which serves as the reproduction guide. No new patient data were collected for this work.

Acknowledgments: The authors thank the creators of the VQA-RAD, SLAKE, and PathVQA datasets for making their data publicly available and the developers of the open-source LLaDA and LLaVA-Med projects on which this work builds.

Conflicts of Interest: The authors declare no conflicts of interest.

References

- Li, C.; Wong, C.; Zhang, S.; Usuyama, N.; Liu, H.; Yang, J.; Naumann, T.; Poon, H.; Gao, J. LLaVA-Med: Training a large language-and-vision assistant for biomedicine in one day. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 28541–28564. [\[CrossRef\]](#)
- Tu, T.; Azizi, S.; Driess, D.; Schaeckermann, M.; Amin, M.; Chang, P.C.; Carroll, A.; Lau, C.; Tanno, R.; Ktena, I.; et al. Towards generalist biomedical AI. *NEJM AI* **2024**, *1*, A10a2300138. [\[CrossRef\]](#)
- Wang, S.; Zhao, Z.; Ouyang, X.; Liu, T.; Wang, Q.; Shen, D. Interactive computer-aided diagnosis on medical image using large language models. *Commun. Eng.* **2024**, *3*, 133. [\[CrossRef\]](#) [\[PubMed\]](#)
- Liu, H.; Li, C.; Wu, Q.; Lee, Y.J. Visual instruction tuning. *Adv. Neural Inf. Process. Syst.* **2023**, *36*, 34892–34916. [\[CrossRef\]](#)
- Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; Zhou, J. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv* **2023**, arXiv:2308.12966.
- Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. The Llama 3 herd of models. *arXiv* **2024**, arXiv:2407.21783.
- Nie, S.; Zhu, F.; You, Z.; Zhang, X.; Ou, J.; Hu, J.; Zhou, J.; Lin, Y.; Wen, J.-R.; Li, C. Large language diffusion models. *arXiv* **2025**, arXiv:2502.09992.
- Lou, A.; Meng, C.; Ermon, S. Discrete diffusion modeling by estimating the ratios of the data distribution. In Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria, 21–27 July 2024; Volume 235, pp. 32819–32848.
- Sahoo, S.S.; Arriola, M.; Schiff, Y.; Gokaslan, A.; Marroquin, E.; Chiu, J.T.; Rush, A.; Kuleshov, V. Simple and effective masked diffusion language models. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 130136–130184. [\[CrossRef\]](#)
- Shi, J.; Han, K.; Wang, Z.; Doucet, A.; Titsias, M.K. Simplified and generalized masked diffusion for discrete data. *Adv. Neural Inf. Process. Syst.* **2024**, *37*, 103131–103167. [\[CrossRef\]](#)
- You, Z.; Nie, S.; Zhang, X.; Zhou, J.; Lu, Z.; Wen, J.-R.; Li, C. LLaDA-V: Large language diffusion models with visual instruction tuning. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Denver, Colorado, USA, 3–7 June 2026; pp. 10093–10105.
- Dong, X.; Zhu, W.; Chen, X.; Wang, Z.; Qiu, P.; Tang, S.; Li, X.; Wang, Y. LLaDA-MedV: Exploring large language diffusion models for biomedical image understanding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Denver, CO, USA, 3–7 June 2026; pp. 22773–22783.
- Tschannen, M.; Gritsenko, A.; Wang, X.; Naeem, M.F.; Alabdulmohsin, I.; Parthasarathy, N.; Evans, T.; Beyer, L.; Xia, Y.; Mustafa, B.; et al. SigLIP 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv* **2025**, arXiv:2502.14786.
- Eslami, S.; Meinel, C.; de Melo, G. PubMedCLIP: How much does CLIP benefit visual question answering in the medical domain? In *Findings of the Association for Computational Linguistics: EACL 2023*; ACL Anthology: Stroudsburg, PA, USA, 2023; pp. 1181–1193. [\[CrossRef\]](#)
- Zhang, S.; Xu, Y.; Usuyama, N.; Bagga, J.; Tinn, R.; Preston, S.; Rao, R.; Wei, M.; Valluri, N.; Wong, C.; et al. Large-scale domain-specific pretraining for biomedical vision-language processing. *arXiv* **2023**, arXiv:2303.00915.
- Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning transferable visual models from natural language supervision. In Proceedings of the 38th International Conference on Machine Learning, Virtual Event, 18–24 July 2021; Volume 139, pp. 8748–8763.
- Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. Training language models to follow instructions with human feedback. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 27730–27744. [\[CrossRef\]](#)
- Pan, J.; Liu, C.; Wu, J.; Liu, F.; Zhu, J.; Li, H.B.; Chen, C.; Ouyang, C.; Rueckert, D. MedVLM-R1: Incentivizing medical reasoning capability of vision-language models (VLMs) via reinforcement learning. In *Medical Image Computing and Computer Assisted Intervention—MICCAI 2025*; Lecture Notes in Computer Science; Springer: Berlin/Heidelberg, Germany, 2025; Volume 15966, pp. 337–347. [\[CrossRef\]](#)
- Lai, Y.; Zhong, J.; Li, M.; Zhao, S.; Yang, X. Med-R1: Reinforcement learning for generalizable medical reasoning in vision-language models. *arXiv* **2025**, arXiv:2503.13939.
- Guo, D.; Yang, D.; Zhang, H.; Song, J.; Wang, P.; Zhu, Q.; Xu, R.; Zhang, R.; Ma, S.; Bi, X.; et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature* **2025**, *645*, 633–638. [\[CrossRef\]](#) [\[PubMed\]](#)
- Zhang, K.; Zhou, R.; Adhikarla, E.; Yan, Z.; Liu, Y.; Yu, J.; Liu, Z.; Chen, X.; Davison, B.D.; Ren, H.; et al. A generalist vision-language foundation model for diverse biomedical tasks. *Nat. Med.* **2024**, *30*, 3129–3141. [\[CrossRef\]](#) [\[PubMed\]](#)

22. Hamamci, I.E.; Er, S.; Wang, C.; Almas, F.; Simsek, A.G.; Esirgun, S.N.; Dogan, I.; Durugol, O.F.; Hou, B.; Shit, S.; et al. Generalist foundation models from a multimodal dataset for 3D computed tomography. *Nat. Biomed. Eng.* **2026**, 1–19. [\[CrossRef\]](#) [\[PubMed\]](#)
23. Singhal, K.; Tu, T.; Gottweis, J.; Sayres, R.; Wulczyn, E.; Amin, M.; Hou, L.; Clark, K.; Pfohl, S.R.; Cole-Lewis, H.; et al. Toward expert-level medical question answering with large language models. *Nat. Med.* **2025**, *31*, 943–950. [\[CrossRef\]](#) [\[PubMed\]](#)
24. Ryu, J.S.; Kang, H.; Chu, Y.; Yang, S. Vision-language foundation models for medical imaging: A review of current practices and innovations. *Biomed. Eng. Lett.* **2025**, *15*, 809–830. [\[CrossRef\]](#) [\[PubMed\]](#)
25. Qamar, S. Language-Guided Segmentation of Medical Images: A Review of Foundation Models. *Bioengineering* **2026**, *13*, 803. [\[CrossRef\]](#) [\[PubMed\]](#)
26. Ho, J.; Jain, A.; Abbeel, P. Denoising diffusion probabilistic models. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 6840–6851.
27. Ou, J.; Nie, S.; Xue, K.; Zhu, F.; Sun, J.; Li, Z.; Li, C. Your absorbing discrete diffusion secretly models the conditional distributions of clean data. In Proceedings of the International Conference on Learning Representations, Singapore, 24–28 April 2025.
28. Chang, H.; Zhang, H.; Jiang, L.; Liu, C.; Freeman, W.T. MaskGIT: Masked generative image transformer. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition New Orleans, LA, USA, 18–24 June 2022; pp. 11315–11325.
29. Li, X.L.; Thakur, J.; Gulrajani, I.; Liang, P.S.; Hashimoto, T.B. Diffusion-LM improves controllable text generation. *Adv. Neural Inf. Process. Syst.* **2022**, *35*, 4328–4343. [\[CrossRef\]](#)
30. Lau, J.J.; Gayen, S.; Abacha, A.B.; Demner-Fushman, D. A dataset of clinically generated visual questions and answers about radiology images. *Sci. Data* **2018**, *5*, 180251. [\[CrossRef\]](#) [\[PubMed\]](#)
31. Liu, B.; Zhan, L.-M.; Xu, L.; Ma, L.; Yang, Y.; Wu, X.-M. SLAKE: A semantically-labeled knowledge-enhanced dataset for medical visual question answering. In Proceedings of the 18th IEEE International Symposium on Biomedical Imaging, Nice, France, 13–16 April 2021; pp. 1650–1654. [\[CrossRef\]](#)
32. He, X.; Cai, Z.; Wei, W.; Zhang, Y.; Mou, L.; Xing, E.; Xie, P. Towards visual question answering on pathology images. In *59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing; Volume 2: Short Papers*; ACL Anthology: Stroudsburg, PA, USA, 2021; pp. 708–718. [\[CrossRef\]](#)
33. Bazi, Y.; Rahhal, M.M.A.; Bashmal, L.; Zuair, M. Vision-language model for visual question answering in medical imagery. *Bioengineering* **2023**, *10*, 380. [\[CrossRef\]](#) [\[PubMed\]](#)
34. Liu, Y.; Wang, Z.; Xu, D.; Zhou, L. Q2ATransformer: Improving medical VQA via an answer querying decoder. In *Information Processing in Medical Imaging–IPMI 2023; Lecture Notes in Computer Science*; Springer: Berlin/Heidelberg, Germany, 2023; Volume 13939, pp. 445–456. [\[CrossRef\]](#)
35. Li, P.; Liu, G.; Tan, L.; Liao, J.; Zhong, S. Self-supervised vision-language pretraining for medial visual question answering. In Proceedings of the 20th IEEE International Symposium on Biomedical Imaging, Cartagena, Colombia, 18–21 April 2023; pp. 1–5. [\[CrossRef\]](#)
36. Byun, J.Y.; Park, Y.-J.; Corbeil, J.-P.; Abacha, A.B. Overconfidence and calibration in medical VQA: Empirical findings and hallucination-aware mitigation. *arXiv* **2026**, arXiv:2604.02543.
37. Draelos, R.L.; Afreen, S.; Blasko, B.; Brazile, T.L.; Chase, N.; Desai, D.P.; Evert, J.; Gardner, H.L.; Herrmann, L.; House, A.V.; et al. Large language models provide unsafe answers to patient-posed medical questions. *npj Digit. Med.* **2026**, *9*, 241. [\[CrossRef\]](#) [\[PubMed\]](#)
38. von der Stüch, M.S.; Vuskov, R.; Westfechtel, S.; Siepmann, R.; Kuhl, C.; Truhn, D.; Nebelung, S. Visual large language models in radiology: A systematic multimodel evaluation of diagnostic accuracy and hallucinations. *Life* **2026**, *16*, 66. [\[CrossRef\]](#) [\[PubMed\]](#)
39. Zhou, S.; Xie, W.; Li, J.; Zhan, Z.; Song, M.; Yang, H.; Espinoza, C.; Welton, L.; Mai, X.; Jin, Y.; et al. Automating expert-level medical reasoning evaluation of large language models. *npj Digit. Med.* **2026**, *9*, 34. [\[CrossRef\]](#) [\[PubMed\]](#)
40. Moor, M.; Huang, Q.; Wu, S.; Yasunaga, M.; Dalmia, Y.; Leskovec, J.; Zakka, C.; Reis, E.P.; Rajpurkar, P. Med-Flamingo: A multimodal medical few-shot learner. In Proceedings of the 3rd Machine Learning for Health Symposium, New Orleans, LA, USA, 10 December 2023; Volume 225, pp. 353–367.
41. van Sonsbeek, T.; Derakhshani, M.M.; Najdenkoska, I.; Snoek, C.G.M.; Worring, M. Open-ended medical visual question answering through prefix tuning of language models. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2023; Lecture Notes in Computer Science*; Springer: Berlin/Heidelberg, Germany, 2023; Volume 14224, pp. 726–736. [\[CrossRef\]](#)
42. Zhu, W.; Dong, X.; Li, X.; Qiu, P.; Chen, X.; Razi, A.; Sotiras, A.; Su, Y.; Wang, Y. Toward effective reinforcement learning fine-tuning for medical VQA in vision-language models. *arXiv* **2025**, arXiv:2505.13973.
43. Qamar, S. Continual Learning in Medical Imaging: Methods, Foundation Models, Challenges, and Clinical Translation. 2026. Available online: <https://www.preprints.org/manuscript/202607.1301> (accessed on 17 July 2026).

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.