
Discrete Diffusion Language Models for Interactive Radiology Report Drafting

Max Van Puyvelde^{*1,2}
maxvpuyv@stanford.edu

H. Ibrahim Gulluk^{*3}
gulluk@stanford.edu

Wim Van Criekinge^{†2}
wim.vancriekinge@ugent.be

Olivier Gevaert^{†1}
ogevaert@stanford.edu

¹Department of Biomedical Data Science, Stanford University School of Medicine

²Department of Mathematical Modelling, Statistics & Bioinformatics, Ghent University

³Department of Electrical Engineering, Stanford University

Abstract

Diffusion language models, which generate text by denoising a token canvas bidirectionally instead of emitting tokens left to right, have become competitive with autoregressive (AR) generation. Medical foundation models, however, remain almost entirely autoregressive. We adapt a mixture-of-experts diffusion language model, DiffusionGemma-26B, and benchmark it against its same-size AR sibling Gemma-4-26B under an identical LoRA recipe on medical visual question answering datasets, scored by a verbosity-robust LLM judge. Diffusion matches or exceeds AR on all of them, and the finetuned model (3.8B active) is competitive with frontier vision-language models; its decoding is also 3.5–4.4× faster. Beyond this parity, the diffusion model offers a drafting capability AR lacks: any-order infill. Because the canvas is denoised bidirectionally, a radiologist can fix report fragments and have the model fill the text between them, an operation inherent to diffusion but not to autoregression, which is subpar at it. This suits real reports, which are often terse or inconsistent across clinicians and institutions.

1 Introduction

Autoregressive (AR) generation, which produces text one token at a time from left to right, underlies nearly all large language and vision-language models. Discrete diffusion language models [1, 18, 19] are a recent alternative: they generate a sequence by iteratively denoising a fixed token canvas, with each position attending to the entire canvas rather than only to preceding tokens. On general text these models are competitive with autoregressive models of comparable size [18, 22], which makes them a plausible backbone for domains that have so far relied on autoregression. One open instance, DiffusionGemma-26B [6], couples this denoising decoder with a native multimodal encoder, and belongs to a model family that also includes a same-size autoregressive model, Gemma-4-26B [5]; the two share size, family, and lineage, and differ chiefly in their generative paradigm.

Existing medical foundation models, however, are almost exclusively autoregressive. Radiology report generation (RRG), the task of drafting a report from an image, is dominated by AR models [2, 7–9, 12, 25], as are medical vision-language assistants [15]. Whether a diffusion language model is viable as a medical foundation model, both accurate enough and useful in the clinical workflow, is largely untested. A few diffusion models already generate CXR reports [4, 17, 23], but produce complete reports only and do not address interactive drafting.

^{*} Joint first authors. [†] Joint senior authors.

We finetune both the diffusion model and its autoregressive sibling on paired image-text data from medical visual-question-answering datasets, under an identical LoRA recipe that varies only the generative paradigm (same backbone size, vision tower, LoRA targets, and data), and benchmark them against each other and frontier vision-language models with a verbosity-robust LLM judge.

Beyond accuracy, the two paradigms differ in what they can be conditioned on. Reporting practice varies: negative and normal findings are stated explicitly in some settings and omitted in others, and section conventions differ across institutions. A tool that completes or normalizes a report around content the radiologist has already entered, at arbitrary positions, is therefore a useful drafting operation. Because a diffusion decoder denoises the whole canvas bidirectionally, it can fill such a gap from the fixed text on both sides. An autoregressive model, conditioning each token only on preceding text, cannot: a fragment fixed after the gap cannot inform the text filled before it. We call this any-order infill.

We make three contributions. (i) **A diffusion language model is a competitive medical foundation model.** DiffusionGemma-26B equals or exceeds its autoregressive sibling on medical VQA and rivals frontier vision-language models while decoding $3.5\text{--}4.4\times$ faster (Section 4.1, Section 4.2), in a matched comparison that varies only the generative paradigm. (ii) **Any-order infill is a conditioning capability inherent to diffusion.** We cast infill as sampling a report given fragments fixed at arbitrary positions (Section 3.3) and show on MIMIC-CXR that the diffusion model exploits context on both sides of a gap far more effectively than its autoregressive sibling (Section 4.3). (iii) We release our code and finetuned checkpoints.¹

2 Related Work

Diffusion for medical RRG and infill. RRG is dominated by autoregressive models such as MAIRA [2, 12] and ReXrank [25]. Discrete diffusion [1, 18, 19] denoises a token canvas bidirectionally, and several systems apply it to CXR report generation: *AnchorDiff* [23] (vision-conditioned LLaDA-8B, claimed as the first masked diffusion for RRG), *MeDiM* [17] (unified any-to-any generation), and ECHO [4] (one-step distillation). All use bidirectionality only to improve full generation, and none isolate the paradigm against a matched autoregressive backbone or expose interactive infill. Generic diffusion infill is established [10, 22] but not framed as clinical drafting, and existing interactive report tools condition on a region [20] or a prefix [21], not on fragments fixed at arbitrary positions.

Medical VQA and LLM-as-judge. VQA-RAD [14], SLAKE [16], and VQA-Med [3] pair radiology images with short open- and closed-ended questions. Because exact-match scoring penalizes valid paraphrases, open-ended medical VQA is now evaluated with an LLM judge [15, 26], which we adopt (Section 4.1).

3 Method

3.1 Matched Backbones

We compare diffusion and autoregression with everything else held fixed. The diffusion model is DiffusionGemma-26B [6], a discrete diffusion language model, and its AR sibling is Gemma-4-26B [5]; both are 25.2B/3.8B-active mixture-of-experts models with a SigLIP-lineage [24] vision encoder (~ 280 image tokens). We adapt each backbone with low-rank adaptation (LoRA) [11]: rank-64 updates ($\alpha=128$) on the attention and shared-MLP projections, with the 128 experts, the router, and the vision tower frozen. The experts hold most of the weights, so adapting only the shared projections updates the model at a small fraction of the cost of a full finetune, and the identical recipe and data across both backbones leave the generative paradigm as the only deliberate variable. The optimizer is the lone exception: each paradigm keeps the AdamW settings established for its objective, since a shared one underfits one of the two losses. Full hyperparameters are in Section A.

¹Code: https://github.com/mxvp/discrete_diffusion_RRG. Checkpoints: <https://huggingface.co/gevaertlab/diffusiongemma-radiology-vqa>.

3.2 Image-Conditioned Adaptation

We condition on the image and diffuse the text target; the image is never generated. Both paradigms are supervised only on the target tokens, with the image and prompt held fixed, and share the same target string: the report (Findings and Impression) for drafting and infill, or a short answer for VQA. A full report fits in one 256-token canvas, so intra-report attention is bidirectional end to end, which any-order infill requires.

Each paradigm is finetuned with its standard supervised objective, the only difference between the two runs: the diffusion model uses the uniform-state dLLM objective [1, 6] (a random fraction of the target tokens is replaced with uniform draws from the vocabulary, a random token rather than a [MASK] symbol, and the model is trained to recover them), and the AR sibling uses next-token cross-entropy on the same targets.

3.3 Any-Order Infill

Infill fixes part of the report and has the model fill the rest, conditioned on what is fixed. A radiologist who leaves a gap in a draft, for instance, fixes the text on either side of it. Let F be the fixed positions, \mathbf{a} the tokens placed there, and \bar{F} the positions left to fill; infill samples the free positions from the model’s conditional

$$\mathbf{x}_{\bar{F}} \sim p_{\theta}(\mathbf{x}_{\bar{F}} \mid \mathbf{x}_F=\mathbf{a}, c), \tag{1}$$

its report distribution restricted to outputs that carry \mathbf{a} at F . A diffusion decoder samples this conditional directly, with no retraining: at each denoising step we re-impose $\mathbf{x}_F=\mathbf{a}$, before the update so the model predicts the free positions while seeing the fixed ones, and after it so they survive the step’s re-randomization. Because attention within the canvas is bidirectional, a free position conditions on fixed tokens to its right as much as to its left, so the gap is filled from context on both sides.

An autoregressive model factors left to right, $p_{\theta}(\mathbf{x} \mid c) = \prod_i p_{\theta}(x_i \mid x_{<i}, c)$, and cannot sample this conditional: a token never sees the positions after it, so text fixed after the gap cannot shape the fill before it. The paradigms differ in what they can be conditioned on, not in writing quality; Section 4.3 measures it, and Section C gives the sampler.

4 Experiments

We evaluate the adapted backbones three ways: medical VQA accuracy, inference speed, and any-order infill.


4.1 Medical VQA

We compare diffusion and AR on medical VQA and place both against frontier vision-language models (VLMs). We evaluate on three medical-VQA datasets: VQA-RAD [14], SLAKE [16], and VQA-Med-2019 [3] (sizes in Section B), each pairing an image and question with a short open or closed answer.

We evaluate each backbone on every dataset both zero-shot (*base*) and after per-dataset finetuning (*finetuned*), and compare against three frontier VLMs (Gemini-3.5-Flash, GPT-4.1-mini, and Claude-Sonnet-4.6). Finetuning adapts the backbone with LoRA on the dataset (Section A). The frontier models are run zero-shot in a single forward pass, without extended reasoning, and every model answers the same 350 held-out questions per dataset.

We score with an LLM judge. Standard exact-match accuracy is unsuitable for a cross-model comparison here: base and frontier models answer in full sentences and score near zero regardless of correctness (Fig. 1). We therefore score semantic correctness: a fixed judge (Claude Sonnet 4.6) returns a binary semantic-equivalence verdict per (question, reference, answer) triple, allowing paraphrase and added explanation [15, 26], the standard for open-ended medical VQA.

Table 1 reports LLM-judge accuracy for all models, and Fig. 2 plots it.



Q. what is abnormal in the ct scan?
GT. pancreatic ductal adenocarcinoma

diff (base)	×	This axial CT scan of the abdomen shows a small, hyperdense (white) spot in the area of the common bile duct, which is consistent with a gallstone (choledocholithiasis). [...]
AR (base)	×	This axial CT scan of the abdomen shows a high-density (bright white) object in the region of the pancreatic head/distal common bile duct, which is consistent with a gallstone or biliary stone. [...]
diff (ft)	✓	pancreatic adenocarcinoma
AR (ft)	×	This patient has a primary renal malignancy, with ultrasound and ct showing a large solid renal mass. On ultrasound, the mass is hypoechoic. On ct, the mass is hyperenhancing. [...]
Gemini-3.5-Flash	×	Based on the CT scan, there is a calcified gallstone (a hyperdense, bright white calcification) located within the lumen of the gallbladder, [...]
GPT-4.1-mini	×	The CT scan shows a metallic foreign body in the region of the pancreas, suggestive of a pancreatic stent.
Sonnet-4.6	×	The CT scan shows peripancreatic fat stranding and fluid surrounding the pancreas, consistent with acute pancreatitis. [...]

Figure 1: **A medical-VQA example (VQA-Med).** Every model’s answer to “what is abnormal in the CT scan?” (reference: *pancreatic ductal adenocarcinoma*), with the LLM judge’s verdict (✓ correct, × incorrect). Base and frontier models reply in full sentences that exact-match scoring would reject regardless of content; here only the finetuned diffusion model answers correctly.

Table 1: LLM-judge accuracy (Claude Sonnet 4.6, semantic-equivalence), $n=350$ items per dataset. *diff* / *AR* are our two backbones (DiffusionGemini / Gemini-4), evaluated zero-shot (*base*) and after per-dataset finetuning. Frontier VLMs (Gemini-3.5-Flash, GPT-4.1-mini, Claude-Sonnet-4.6) are zero-shot, single forward pass. Bold: best per dataset, *separately* among our models and among the frontier VLMs. † Claude-Sonnet-4.6 is also the judge model.

Dataset	base		finetuned		frontier VLMs		
	diff	AR	diff	AR	Gemini	GPT	Sonnet†
VQA-RAD	0.614	0.523	0.649	0.649	0.777	0.571	0.654
SLAKE	0.700	0.674	0.863	0.817	0.751	0.646	0.703
VQA-Med-2019	0.629	0.614	0.666	0.631	0.683	0.586	0.654

Finetuning. Finetuning improves LLM-judge accuracy for both paradigms, most on SLAKE: +0.163 diffusion (0.700→0.863) and +0.143 AR (0.674→0.817); VQA-RAD-AR gains +0.126, and the VQA-Med gains are marginal. Base diffusion already reaches 0.61–0.70.

Diffusion versus AR. Finetuned diffusion equals or exceeds finetuned AR on the judge metric for all three datasets, and base diffusion exceeds base AR on all three (Table 1). On per-item McNemar tests over the judge verdicts ($n=350$), the difference is significant on SLAKE finetuned (+0.046, $p=0.026$) and VQA-RAD base (+0.091, $p<0.001$); the other four diffusion-vs.-AR comparisons are not significant. The difference is concentrated on closed (yes/no) questions, where the answer format is irrelevant (e.g. on VQA-RAD finetuned, closed-question accuracy is 0.825 for diffusion vs. 0.757 for AR). That a uniform-state denoising model matches its next-token sibling at equal scale, on questions that turn on fine-grained image grounding, indicates the diffusion paradigm is a viable substrate for a medical foundation model, on which the infill capability of Section 4.3 builds.

Frontier VLMs. The finetuned 26B model (3.8B active) is competitive with the three frontier VLMs (Table 1; Fig. 2b). Finetuned diffusion has the highest judge accuracy on SLAKE (0.863); Gemini-3.5-Flash is highest on VQA-RAD (0.777) and VQA-Med (0.683). Finetuned diffusion exceeds GPT-4.1-mini on all three datasets; only Gemini-3.5-Flash clearly surpasses it, on VQA-RAD and VQA-Med, while the judge model itself edges it on VQA-RAD (0.654 vs. 0.649, within noise at $n=350$). Example per-model answers appear in Section D.1.

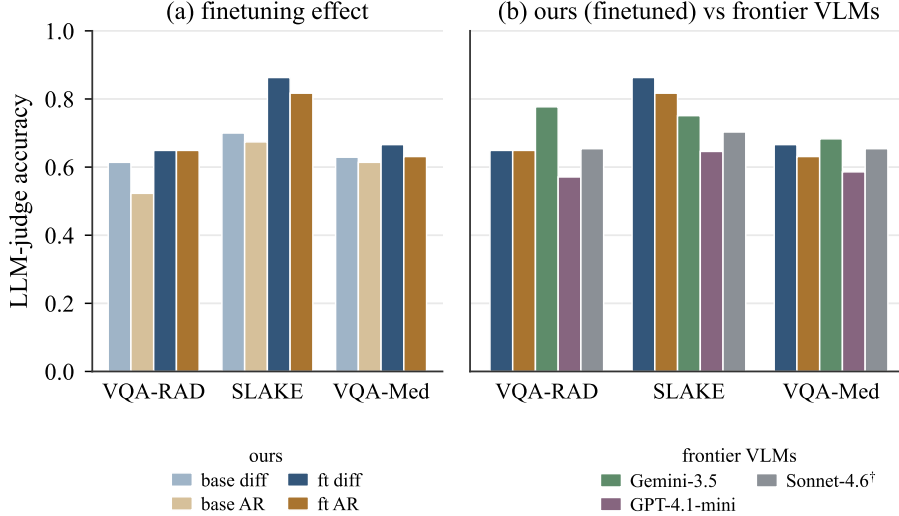


Figure 2: LLM-judge accuracy (Claude Sonnet 4.6). **(a)** base vs. finetuned, for diffusion and AR. **(b)** the finetuned 26B model (3.8B active) against three frontier non-reasoning VLMs. [†]Claude-Sonnet-4.6 is the judge model.

Table 2: Inference speed on one H100 (bf16, ~ 256 -token generation). AR is greedy decode; DiffusionGemma is swept over the denoising-step budget. Speedup is latency relative to AR.

Decoder	Latency (s)	Throughput (tok/s)	Speedup
Gemma-4 (AR), greedy	6.43	24.6	1.0 \times
DiffusionGemma, 16 steps	1.46	175.3	4.4 \times
DiffusionGemma, 32 steps	1.74	147.4	3.7 \times
DiffusionGemma, 48 steps	1.84	139.4	3.5 \times

4.2 Inference Speed

Latency matters for interactive drafting: the model must produce a draft fast enough to be regenerated as the radiologist works. We characterize inference speed for the two decoders on matched hardware.

The cost structures differ. AR cost scales with decoded tokens: each token is one sequential forward pass (with KV caching), so latency grows with report length. Diffusion cost is set by the denoising-step budget over the 256-token canvas: each step is one forward pass updating all unaccepted positions in parallel, independent of length. Because latency therefore depends on token count and step budget rather than on the report’s content, we measure a generic ~ 256 -token generation rather than a specific dataset.

DiffusionGemma-26B drafts 3.5–4.4 \times faster than its AR sibling and at 5.7–7.1 \times higher throughput (Table 2); AR is timed at its natural, shorter output while diffusion fills the full canvas, so the comparison is generous to AR.

4.3 Any-Order Infill

Section 3.3 cast infill as sampling the conditional of Eq. (1). Here we evaluate the capability it affords that autoregression lacks: filling a gap in the report from the fixed text on *both* sides. A radiologist editing one sentence of a draft, for instance, wants the surrounding text updated from both directions, where an AR model would regenerate only what follows the edit. Figure 3 contrasts the two paradigms.

We mask one complete sentence (deterministically, near the middle) of each held-out MIMIC-CXR [13] report ($n=249$, one canvas) and fill the resulting gap under two conditions, scoring each fill against the masked sentence by token-F1 and by the LLM judge of Section 4.1

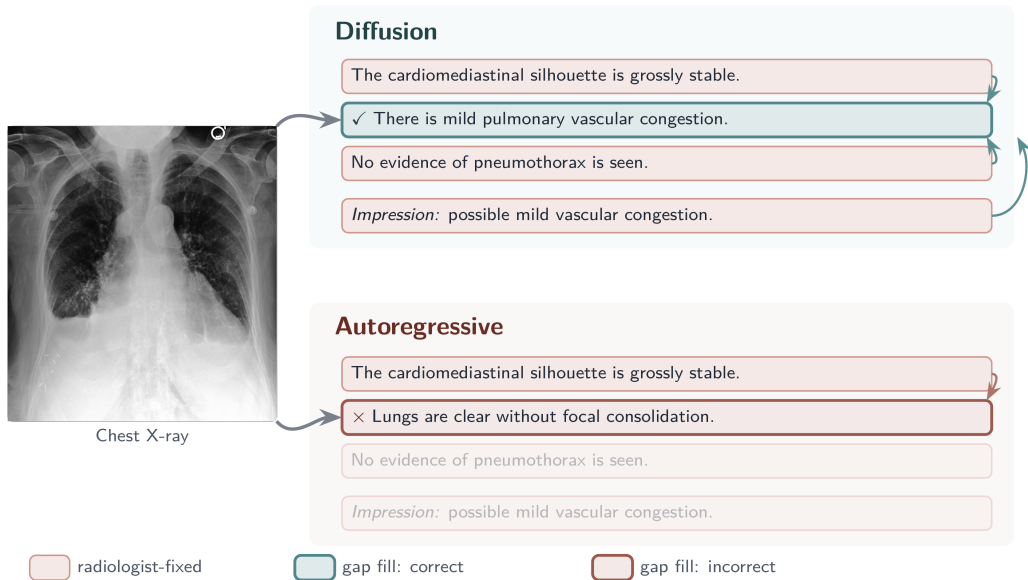


Figure 3: **Completing a gap from both sides.** One sentence of a chest X-ray report is masked (the gap) and filled from the surrounding fixed fragments. Top: the diffusion model draws on fragments on either side and recovers the sentence correctly. Bottom: the autoregressive sibling sees only the fragments before it, the rest greyed out, and reconstructs it incorrectly. Real MIMIC-CXR example.

Table 3: Infill ablation on MIMIC-CXR ($n=249$): fill a masked sentence with context on both sides vs. left only. Token-F1 and LLM-judge accuracy of the fill against the masked sentence; Δ is the gain from adding the right-side context (paired test). AR *bidir.* is given both sides in its prompt. $*p < 10^{-3}$; n.s.: not significant.

	token-F1			LLM-judge		
	bidir.	left	Δ	bidir.	left	Δ
diffusion	0.320	0.211	+0.109*	0.285	0.157	+0.129*
AR (prompted)	0.275	0.244	+0.031 ^{n.s.}	0.257	0.273	-0.016 ^{n.s.}

(semantic equivalence to the reference sentence). The *bidirectional* condition clamps the fragments on both sides of the gap; *left-only* clamps only the left, emulating an AR view. We run this for the diffusion model and for AR; in the AR *bidirectional* condition the right-side context is supplied in the prompt, the only way an autoregressive model can condition on it. The within-model gain from adding the right context measures bidirectional exploitation, and the model \times context interaction measures the capability asymmetry. Both models are the released (base) checkpoints, so the result reflects architecture rather than finetuning.

Table 3 reports the 2×2 . The diffusion model uses the right-side context strongly: adding it raises token-F1 by +0.109 (paired t , $p < 10^{-10}$, 95% CI [+0.077, +0.141]) and judge accuracy by +0.129 ($p = 2 \times 10^{-5}$). AR does not: even when prompted with both sides, the right context does not significantly help it (+0.031 token-F1, $p = 0.08$; -0.016 judge, n.s.). The model \times context interaction is significant on both metrics (+0.078 token-F1, $p = 2 \times 10^{-4}$; +0.145 judge, $p = 3 \times 10^{-4}$): diffusion benefits about $3.5 \times$ more from bidirectional context. Example fills are shown in Section D.2.

5 Conclusion

We studied discrete diffusion versus autoregression for chest X-ray report drafting with two same-size, same-family backbones, so the generative paradigm is the only variable. On a matched medical-VQA benchmark scored by a verbosity-robust LLM judge, the diffusion model matches or exceeds its AR sibling and is competitive with frontier vision-language models, while decoding $3.5\text{--}4.4 \times$ faster.

Beyond this, it adds a drafting capability AR lacks, any-order infill: a sampler modification lets a radiologist fix report fragments and have the diffusion model fill the gaps between them. On MIMIC-CXR it exploits context on both sides of a gap (+0.109 token-F1, +0.129 judge accuracy) while AR does not, even when the right-side context is in its prompt. We release our code and finetuned checkpoints.

References

- [1] Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. doi: 10.48550/arXiv.2107.03006.
- [2] Shruthi Bannur, Kenza Bouzid, Daniel C. Castro, Anton Schwaighofer, Anja Thieme, Sam Bond-Taylor, Maximilian Ilse, Fernando Pérez-García, Valentina Salvatelli, Harshita Sharma, et al. MAIRA-2: Grounded radiology report generation, 2024. arXiv preprint arXiv:2406.04449.
- [3] Asma Ben Abacha, Sadid A. Hasan, Vivek V. Datla, Joey Liu, Dina Demner-Fushman, and Henning Müller. VQA-Med: Overview of the medical visual question answering task at ImageCLEF 2019. In *CLEF 2019 Working Notes, CEUR Workshop Proceedings*, 2019.
- [4] Lifeng Chen, Tianqi You, Hao Liu, Zhimin Bao, Jile Jiao, Xiao Han, Zhicai Ou, Tao Sun, Xiaofeng Mou, Xiaojie Jin, and Yi Xu. ECHO: Efficient chest x-ray report generation with one-step block diffusion, 2026. arXiv preprint arXiv:2604.09450.
- [5] Gemma Team, Google DeepMind. Gemma 4: Open multimodal models. Model card, <https://huggingface.co/google/gemma-4-26B-A4B-it>, 2026.
- [6] Google DeepMind. DiffusionGemma: Block discrete-diffusion language models. Model card, <https://huggingface.co/google/diffusiongemma-26B-A4B-it>, 2026.
- [7] Halil Ibrahim Gulluk and Olivier Gevaert. SemEnrich: Self-supervised semantic enrichment of radiology reports for vision-language learning. *arXiv preprint arXiv:2604.09887*, 2026. doi: 10.48550/arXiv.2604.09887.
- [8] Halil Ibrahim Gulluk, Max Van Puyvelde, and Olivier Gevaert. OpenMedQ: Broad open pretraining for medical vision-language models. *arXiv preprint arXiv:2606.12953*, 2026. doi: 10.48550/arXiv.2606.12953.
- [9] Halil Ibrahim Gulluk, Max Van Puyvelde, Wim Van Criekinge, and Olivier Gevaert. SDR: Set-distance rewards for radiology report generation, 2026. arXiv preprint arXiv:2606.00440.
- [10] HKU NLP Group. DreamOn: Diffusion language models for code infilling beyond fixed-size canvas, 2026. arXiv preprint arXiv:2602.01326.
- [11] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*, 2022. doi: 10.48550/arXiv.2106.09685.
- [12] Stephanie L. Hyland, Shruthi Bannur, Kenza Bouzid, Daniel C. Castro, Mercy Ranjit, Anton Schwaighofer, Fernando Pérez-García, Valentina Salvatelli, Shaury Srivastav, Anja Thieme, et al. MAIRA-1: A specialised large multimodal model for radiology report generation, 2023. arXiv preprint arXiv:2311.13668.
- [13] Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathaniel R. Greenbaum, Matthew P. Lungren, Chih-ying Deng, Roger G. Mark, and Steven Horng. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6(1):317, 2019. doi: 10.1038/s41597-019-0322-0.
- [14] 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): 180251, 2018. doi: 10.1038/sdata.2018.251.

- [15] 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. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks*, 2023. doi: 10.48550/arXiv.2306.00890.
- [16] 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 *IEEE International Symposium on Biomedical Imaging (ISBI)*, 2021. doi: 10.1109/ISBI48211.2021.9434010.
- [17] Jiawei Mao, Yuhan Wang, Lifeng Chen, Can Zhao, Yucheng Tang, Dong Yang, Liangqiong Qu, Daguang Xu, and Yuyin Zhou. Discrete diffusion models with MLLMs for unified medical multimodal generation, 2025. arXiv preprint arXiv:2510.06131.
- [18] 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. doi: 10.48550/arXiv.2502.09992.
- [19] Subham Sekhar Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T. Chiu, Alexander Rush, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. doi: 10.48550/arXiv.2406.07524.
- [20] Tim Tanida, Philip Müller, Georgios Kaissis, and Daniel Rueckert. Interactive and explainable region-guided radiology report generation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. doi: 10.1109/CVPR52729.2023.00718.
- [21] Sheng Wang et al. CopilotCAD: Empowering radiologists with report completion models and quantitative evidence from medical image foundation models, 2024. arXiv preprint arXiv:2404.07424.
- [22] Jiacheng Ye, Zhihui Xie, Lin Zheng, Jiahui Gao, Zirui Wu, Xin Jiang, Zhenguo Li, and Lingpeng Kong. Dream 7b: Diffusion large language models. *arXiv preprint arXiv:2508.15487*, 2025. doi: 10.48550/arXiv.2508.15487.
- [23] Shiyong Yu, Jielei Wang, and Guoming Lu. AnchorDiff: Topology-aware masked diffusion with confidence-based rewriting for radiology report generation, 2026. arXiv preprint arXiv:2605.17071.
- [24] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *International Conference on Computer Vision (ICCV)*, 2023. doi: 10.48550/arXiv.2303.15343.
- [25] Xiaoman Zhang, Hong-Yu Zhou, Xiaoli Yang, et al. ReXrank: A public leaderboard for ai-powered radiology report generation, 2024. arXiv preprint arXiv:2411.15122.
- [26] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2023. doi: 10.48550/arXiv.2306.05685.

A Backbones and Adaptation Recipe

Table A1 lists the two backbones and their (identical) adaptation recipe; only the generative paradigm and its established optimizer differ.

B Datasets

Table A2 lists the three medical-VQA datasets and their sizes.

Table A1: The two backbones and their adaptation. Same family, size, vision tower, LoRA targets, and data (“same” = identical to the diffusion column); only the generative paradigm and its optimizer differ. Vision is frozen for both.

	DiffusionGemma-26B-it	Gemma-4-26B-it
paradigm	uniform-state discrete diffusion	autoregressive
size	25.2B / 3.8B active MoE	same
vision	native <code>gemma4_vision</code> (~280 tok)	same
infill	yes (bidirectional)	no (structural)
LoRA	$r64/\alpha128, \{q, k, v, o, mlp\}$	same
optimizer	AdamW $\beta=(0.9, 0.95)$	AdamW $\beta=(0.9, 0.999)$
learning rate	1×10^{-4} , cosine, warmup 100	1×10^{-4} , constant
batch (grad-accum)	16	24
diffusion settings	$\epsilon=0.1$, canvas $L=256$	—
precision	bf16, single H100	same

Table A2: Medical-VQA datasets. Sizes are train/test QA pairs; evaluation uses a fixed random subset of $n=350$ test items per dataset.

Dataset	Train/Test	Answers
VQA-RAD [14]	1,817 / 431	yes/no, open (~1 word)
SLAKE [16]	4,919 / 1,053	yes/no, open
VQA-Med-2019 [3]	~12.8k / 2.0k	open: modality, plane, organ

C Infill Sampler

We patch the uncompiled outer denoising step (the compiled inner sampler is reassigned as an instance attribute and shadows a class-level patch), clamping the fixed positions on the incoming canvas and on both outgoing canvases at each step (Fig. A1).

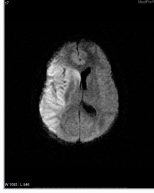
```
# clamp user-fixed tokens at fixed positions every denoising step
# (patch the uncompiled outer _denoising_step, not the compiled inner sampler)
def patched_step(self, *a, current_canvas=..., **kw):
    cc = where(fixed_mask, fixed_tokens, current_canvas) # condition on fixed
    cur, argmax, *rest = orig_step(self, current_canvas=cc, ...)
    cur = where(fixed_mask, fixed_tokens, cur) # survive renoise
    argmax = where(fixed_mask, fixed_tokens, argmax)
    return (cur, argmax, *rest)
```

Figure A1: The any-order infill procedure (abridged). `fixed_tokens / fixed_mask` are $[B, L]$; the wrapper clamps the fixed positions on the incoming canvas and both outgoing canvases at each step.

D Example Outputs

D.1 Medical VQA

Held-out items from each dataset, with every model’s answer and the LLM judge’s verdict (✓ correct, × incorrect). Base and frontier models answer in full sentences, which exact-match scoring penalizes regardless of correctness; the judge scores meaning (Section 4.1). Long answers are abbreviated with [...]; one sample per dataset is shown in full to illustrate this verbosity.



Q. Is there ventricular enlargement?

GT.

- diff (base) × Based on the provided MRI image, there is evidence of enlargement of the right lateral ventricle (the space on the left side of the image). The image shows a large area of high signal intensity (bright area) in the right hemisphere, which is consistent with a large acute ischemic stroke. The enlargement of the adjacent ventricle is likely due to "ex vacuo" effect, where the ventricle expands because of tissue loss (atrophy) from the stroke. [...]
- AR (base) × This image shows a Diffusion Weighted Imaging (DWI) MRI sequence of the brain. It is not possible to determine if there is ventricular enlargement from this single image. A formal evaluation of ventricular size requires several different types of MRI sequences (such as T1-weighted images) and a full examination of all slices in a study. The most striking finding on this image is a large, bright area in the right hemisphere (the left side of the image). In DWI, this bright signal is characteristic of an acute ischemic stroke affecting a large territory of the brain. [...]
- diff (ft) ✓ No
- AR (ft) × Yes
- Gemini-3.5-Flash × Based on the provided MRI scan, there is no generalized ventricular enlargement. However, there is asymmetric compression and effacement of the right lateral ventricle (shown on the left side of the image) due to mass effect from the large, hyperintense acute ischemic stroke in the right middle cerebral artery (MCA) territory.
- GPT-4.1-mini × Yes, there is ventricular enlargement.
- Sonnet-4.6 × Yes, there is ventricular enlargement (hydrocephalus) visible, with the right lateral ventricle appearing dilated. Additionally, there is a large area of restricted diffusion (bright signal on DWI) in the right hemisphere, consistent with an acute ischemic stroke involving the right MCA territory.



Q. is there a mass

GT.

- diff (base) ✓ This chest X-ray shows a rounded opacity (a white area) in the upper left lung. [...]
- AR (base) ✓ This chest X-ray shows a well-defined opacity (a "spot" or density) in the upper part of the left lung. [...]
- diff (ft) ✓ yes
- AR (ft) ✓ Yes
- Gemini-3.5-Flash × Based on the provided chest X-ray, there is no evidence of a definitive pulmonary mass.
- GPT-4.1-mini × No, there is no visible mass in this chest X-ray.
- Sonnet-4.6 × No, there is no discrete mass visible on this chest X-ray. [...]



Q. Describe the borders of the lesion.

GT. Smooth

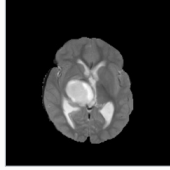
- | | | |
|------------------|---|---|
| diff (base) | × | The provided chest X-ray does not show any clearly defined lesions. The lung fields appear clear, and the heart size and shape are within normal limits. [...] |
| AR (base) | × | I cannot provide a description of a lesion because no lesion is visible in this chest X-ray. The image appears to show normal anatomy for a chest radiograph. [...] |
| diff (ft) | ✓ | Well-defined |
| AR (ft) | ✓ | The lesion has well-defined borders. |
| Gemini-3.5-Flash | × | Based on the provided chest radiograph, there is no visible lung lesion or abnormal mass. [...] |
| GPT-4.1-mini | × | The borders of the lesion are irregular and poorly defined. |
| Sonnet-4.6 | × | The lesion (located in the right hilum/perihilar region) has well-defined, lobulated borders, suggesting a discrete mass rather than an infiltrative process. [...] |



Q. From what structure is the large hypodense mass emanating from?

GT. The base of the cecum

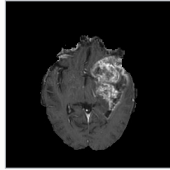
- | | | |
|------------------|---|--|
| diff (base) | × | Based on the provided CT scan of the abdomen, the large hypodense mass appears to be emanating from the small bowel. [...] |
| AR (base) | × | This single axial CT image shows a large, hypodense (darker) mass in the central abdomen. [...] |
| diff (ft) | ✓ | cecum |
| AR (ft) | × | The mass is seen as an extension of the ascending colon. |
| Gemini-3.5-Flash | × | Based on the axial CT image, the large hypodense mass in the lower abdomen/pelvis is emanating from the appendix (consistent with an appendicitis [...]) |
| GPT-4.1-mini | × | The large hypodense mass is emanating from the abdominal aorta. |
| Sonnet-4.6 | × | The large hypodense mass is emanating from the mesentery (mesenteric origin). |



Q. How many kinds of abnormalities are there in this image?

GT. 3

-
- | | | |
|------------------|---|---|
| diff (base) | × | Based on the single image provided, there is one primary area of abnormality visible: a large, hyperintense (bright) lesion located in the right hemisphere, specifically involving the thalamus/basal ganglia region. A definitive medical diagnosis requires a review of the entire imaging study, multiple sequences, and the patient's clinical history by a qualified professional. If this image belongs to you or someone you know, you should consult with a physician for a formal interpretation and clinical plan. |
| AR (base) | × | Based on the provided image, which appears to be an axial MRI scan of a brain, there is a large, prominent abnormality visible. The image shows a significant area of high signal intensity (bright white area) located in the central/right region of the brain. This represents a single major type of abnormality, which could indicate several different medical conditions such as a tumor, hemorrhage, or an area of edema/infarction. Important Medical Disclaimer: I am an AI, not a doctor. [...] |
| diff (ft) | ✓ | 3 |
| AR (ft) | ✓ | There are 3 kinds of abnormalities in this image. |
| Gemini-3.5-Flash | × | Based on the provided brain MRI, there are two main types of abnormalities visible: 1. A large intracranial mass/tumor (the well-defined, hyperintense lesion in the deep right hemisphere/thalamic region). 2. Obstructive ventriculomegaly/hydrocephalus (abnormal enlargement and dilation of the lateral ventricles, particularly the occipital and frontal horns, caused by mass effect and CSF obstruction). |
| GPT-4.1-mini | × | There are two kinds of abnormalities in this image. |
| Sonnet-4.6 | × | Looking at this brain MRI (DWI sequence), I can identify 2 kinds of abnormalities: 1. Restricted diffusion (bright signal) — indicating acute ischemia/infarction in the posterior fossa/cerebellum region 2. Mass effect/enlarged structure — suggesting associated edema or structural enlargement in that area |



Q. How many kinds of abnormalities are there in this image?

GT. 3

-
- | | | |
|------------------|---|--|
| diff (base) | × | Based on the image provided, there is one large, primary abnormality: a significant, heterogeneous mass located in the left hemisphere of the brain. [...] |
| AR (base) | × | This image is a single medical scan (an axial slice of a brain MRI). [...] |
| diff (ft) | ✓ | 3 |
| AR (ft) | × | There is 1 kind of abnormality in this image. |
| Gemini-3.5-Flash | × | Based on the provided brain MRI, there is 1 major kind of abnormality (a large, heterogeneous, enhancing intra-axial mass lesion in the left [...]) |
| GPT-4.1-mini | × | There is one kind of abnormality visible in this image. |
| Sonnet-4.6 | × | Looking at this brain MRI image, I can identify 2 kinds of abnormalities: 1. Hemorrhage (hyperdense/bright area indicating blood) 2. [...] |



Q. How many organs are there in this image?

GT. 2

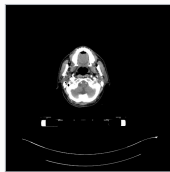
-
- | | | |
|------------------|---|--|
| diff (base) | × | I cannot count the number of organs in this image. A chest X-ray is a medical imaging tool used to visualize internal structures, but it is not used to count organs. [...] |
| AR (base) | × | I cannot provide a medical interpretation or diagnosis from this image. [...] |
| diff (ft) | ✓ | 2 |
| AR (ft) | ✓ | 2 |
| Gemini-3.5-Flash | × | Based on this chest X-ray, we can identify 3 major visible organs: the lungs (left and right), the heart, and the stomach (visible via the g [...]) |
| GPT-4.1-mini | × | There are two organs visible in this image: the lungs. |
| Sonnet-4.6 | × | In this chest X-ray, there are multiple organs visible, but the main identifiable ones are: 1. Lungs (2 — left and right) 2. Heart 3. Liver (partially visible at the right lower border) 4. [...] |



Q. How many organs are there in this image?

GT. 6

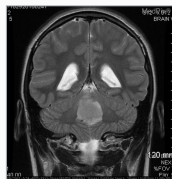
-
- | | | |
|------------------|---|--|
| diff (base) | × | I cannot count the number of organs from a single CT scan slice. A CT scan consists of many sequential images that are necessary to fully identify and evaluate all internal structures. [...] |
| AR (base) | × | I cannot provide a count of organs in this image. Interpreting medical imaging requires specialized training and context from a complete scan, which a single image cannot provide. [...] |
| diff (ft) | ✓ | 6 |
| AR (ft) | ✓ | There are 6 organs in this image. |
| Gemini-3.5-Flash | × | Based on this abdominal CT slice, there are 4 major distinct organs visualized: 1. Spleen (on the right side of the image / anatomical left) 2. [...] |
| GPT-4.1-mini | × | There are four organs visible in this image: the liver, both kidneys, and portions of the intestines. |
| Sonnet-4.6 | × | In this abdominal CT scan, the visible organs include: 1. Kidneys (bilateral) — 2 2. Bowel loops (small/large intestine) 3. Vertebral body (bone structure) 4. [...] |



Q. Is this a study of the neck?

GT. No

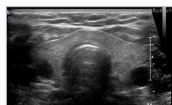
-
- | | | |
|------------------|---|--|
| diff (base) | × | Yes, this is an axial CT scan of the lower head and upper neck region. [...] |
| AR (base) | × | This single image is an axial CT slice through the level of the upper neck/base of the skull, showing structures such as the mandible, pharynx, and the base of the brain. [...] |
| diff (ft) | ✓ | No |
| AR (ft) | ✓ | No |
| Gemini-3.5-Flash | × | Yes. |
| GPT-4.1-mini | × | Yes. |
| Sonnet-4.6 | × | Yes, this is a CT study of the neck/head junction region — specifically an axial CT image at the level of the skull base/upper neck, showi [...] |



Q. what is the primary abnormality in this image?

GT. juvenile pilocytic astrocytoma

- diff (base) ✓ This T2-weighted coronal MRI of the brain shows a large, well-defined hyperintense (bright) area located in the cerebellum. The appearance and location of this lesion are concerning for a cerebellar tumor. [...]
- AR (base) ✗ This image is a coronal T2-weighted MRI of the brain. The primary abnormality is a large, well-circumscribed, hyperintense (bright) lesion located in the midline of the cerebellum (the posterior fossa). [...]
- diff (ft) ✓ pillary astrocytoma
- AR (ft) ✓ The primary abnormality in this image is a juvenile pilocytic astrocytoma (jpa). Juvenile pilocytic astrocytoma is a benign, slow-growing neoplasm of the central nervous system. [...]
- Gemini-3.5-Flash ✗ Based on the provided coronal T2-weighted brain MRI, the primary abnormality is a large, hyperintense mass in the brainstem (specifically wi [...])
- GPT-4.1-mini ✗ The primary abnormality in this image is bilateral symmetrical hyperintense lesions in the basal ganglia, indicative of a metabolic or toxic [...]
- Sonnet-4.6 ✗ The primary abnormality in this image is bilateral enlargement of the lateral ventricles (hydrocephalus). [...]



Q. what is most alarming about this ultrasound?

GT. papillary carcinoma of the thyroid

- diff (base) ✗ This image shows an ultrasound of the thyroid gland. The most alarming feature in this image is the presence of a large, well-defined nodule in the right thyroid lobe (the darker/hypoechoic circular area on the left side of the image). While many thyroid nodules are benign, certain features on ultrasound can be associated with an increased risk of malignancy (cancer). These features, which radiologists look for, include: Hypoechoogenicity: The nodule appearing darker than the surrounding thyroid tissue. Irregular margins: The edges of the nodule not being smooth or sharp. [...]
- AR (base) ✗ An ultrasound image cannot be used to provide a diagnosis. A complete medical evaluation requires a review of the original high-resolution imaging, a physical examination, and a discussion of your medical history and symptoms. The image you provided is a single, low-resolution frame from a thyroid ultrasound. Interpreting such an image is limited because the most important features (such as the exact texture, borders, and vascularity of a potential nodule) require real-time scanning and specific measurement tools. [...]
- diff (ft) ✓ thyroid carcinoma
- AR (ft) ✗ The ultrasound shows a normal thyroid gland, which is a reassuring finding in a patient with stable hypothyroidism.
- Gemini-3.5-Flash ✗ Based on the provided thyroid/neck ultrasound, the most alarming finding is the absence of the right thyroid lobe (thyroid hemiagenesis). The image shows a normal-appearing left thyroid lobe and isthmus crossing anterior to the trachea, but the right thyroid bed (on the left side of the image) is empty, with no identifiable right thyroid lobe tissue lateral to the trachea.
- GPT-4.1-mini ✗ The most alarming feature in this ultrasound is the presence of a large, well-defined, possibly cystic mass in the thyroid region, which could indicate a thyroid nodule or tumor requiring further evaluation.
- Sonnet-4.6 ✗ The most alarming finding in this ultrasound is the irregular, hypoechoic solid mass with posterior acoustic shadowing, which raises concern for malignancy (such as breast carcinoma). The mass appears to have irregular margins and heterogeneous echotexture, features that are suspicious for a malignant lesion rather than a benign cyst or fibroadenoma.

D.2 Any-Order Infill

Four held-out MIMIC-CXR reports, each with one sentence masked. Its position is marked (....) and the masked sentence is shown in the teal box below; the four fills form a {diffusion, AR} × {bidirectional, left-only} grid. *Bidirectional* supplies the fixed text on both sides of the gap (for AR, in its prompt), *left-only* only the left. Only the bidirectional diffusion fill reconstructs the masked sentence; the others, including AR with both sides in its prompt, cannot condition on the right-side context.

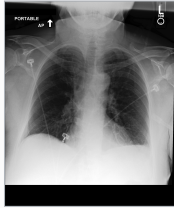


Report. The heart is normal in size. The mediastinal and hilar contours appear within normal limits. ... The lungs appear clear. Bony structures are unremarkable. No evidence of acute disease.

masked sentence:

There is no pleural effusion or pneumothorax.

	bidirectional	left-only
diffusion	✓ There is no pleural effusion or pneumothorax.	✗ The lungs are clear, with no infiltrates,
AR	✗ The pleural spaces are clear.	✗ The trachea is midline and the retrosternal clear space is preserved.



Report. ... No airspace opacification. No pneumothorax. No pulmonary edema. Mild density seen in the left costophrenic angle which may represent atelectasis or a small pleural effusion. Suture material projecting over the right hilar area. Narrowing of the subglottic trachea is probably due to recent intubation. The cardiomeastinal shadow is unchanged. No airspace opacification. No pneumothorax. No pulmonary edema. Mild density seen in the left costophrenic angle which may represent atelectasis or a small pleural effusion. Suture material projecting over the right hilar area. Narrowing of the subglottic trachea is probably due to recent intubation.

masked sentence:

The cardiomeastinal shadow is unchanged.

	bidirectional	left-only
diffusion	✓ The cardiomeastinal shadow is normal.	✗ Impression: 1. No acute pulmonary process
AR	✗ Lungs are clear.	✗ Bones and soft tissues are unremarkable.



Report. There are moderate bilateral pleural effusions with overlying atelectasis, underlying basilar consolidation is not excluded in the appropriate clinical setting. The cardiomeastinal silhouette is grossly stable. ... No evidence of pneumothorax is seen. Bilateral pleural effusions with overlying atelectasis, underlying consolidation not excluded. Possible mild vascular congestion.

masked sentence:

There is minimal pulmonary vascular congestion.

	bidirectional	left-only
diffusion	✓ There is mild pulmonary vascular congestion.	✗ There are surgical clips in the right
AR	✗ Lungs are clear without focal consolidation.	✗ The lungs are otherwise clear without focal consolidation, suspicious pulmonary nodules, or pneumothorax.



Report. The previous bilateral pleural effusions have resolved. Substantial apical thickening bilaterally and lung scarring, the sequela of likely radiation therapy is unchanged. No evidence of pneumonia. Resolution of pleural effusions. Sequela of radiation induced changes including biapical scarring and fibrosis.

masked sentence:

The cardiac size is normal.

bidirectional

left-only

diffusion

✓ The heart size is normal.

✗ There is no acute pulmonary consolidation

AR

✗ There is no new airspace opacity.

✗ There is no new focal consolidation, pneumothorax, or pleural effusion.