LeGend: Lesion-Guided Longitudinal CXR Generation via Gaussian-Biased Causal Attention

Yiran Wang[1], Xiaoyu Yue[1], Haimei Zhao[1], Xinyu Chen[1], and Luping Zhou[1*]

The University of Sydney, Sydney, Australia

Abstract. Generating follow-up chest X-rays (CXRs) conditioned on a reference study and progression description can support controllable longitudinal visualization and data completion, and recent autoregressive (AR) models have shown promising fidelity for this task. However, they often display limited lesion-centric spatial awareness, leading to disease-inconsistent follow-up synthesis. In this work, we identify and quantify a previously unrecognized corner/edge attention bias in AR generation, where follow-up queries disproportionately attend to peripheral, non-lesional reference regions. We further introduce LeGend, a lesion-guided framework for longitudinal CXR generation built on Gaussian-Biased Causal Attention (GBCA), a lightweight, causally aligned, plug-and-play correction that injects a lesion-conditioned 2D Gaussian prior into causal self-attention logits to provide sample-specific guidance without modifying the backbone architecture. The prior is obtained from sparse lesion coordinates predicted by an offline vision–language model and smoothed into a disease-relevant salience map to softly steer decoding toward lesion-relevant regions. Experiments show that LeGend improves lesion-relevant attention, image fidelity, and classifier-assessed disease consistency, while yielding more interpretable attention patterns with negligible computational overhead. Our code are available here:`https://github.com/yiranhi/Legend.git`.

Keywords: Follow-up CXR Generation · Longitudinal CXR Synthesis · Lesion-guided Attention.

1 Introduction

Follow-up chest radiography (CXR) is routinely used to monitor disease progression, treatment response, and post-operative recovery. Generating follow-up CXRs conditioned on a reference CXR and a progression description can support controllable longitudinal visualization and data completion, yet disease-consistent synthesis remains challenging because pathological changes are subtle and spatially localized. Recent generative models have improved CXR synthesis and editing (e.g., PIE [20], BioMedJourney [21], CXR-IRGen [25]), while autoregressive (AR) transformers enable high-fidelity multimodal generation at

* Corresponding author. Email: `luping.zhou@sydney.edu.au`

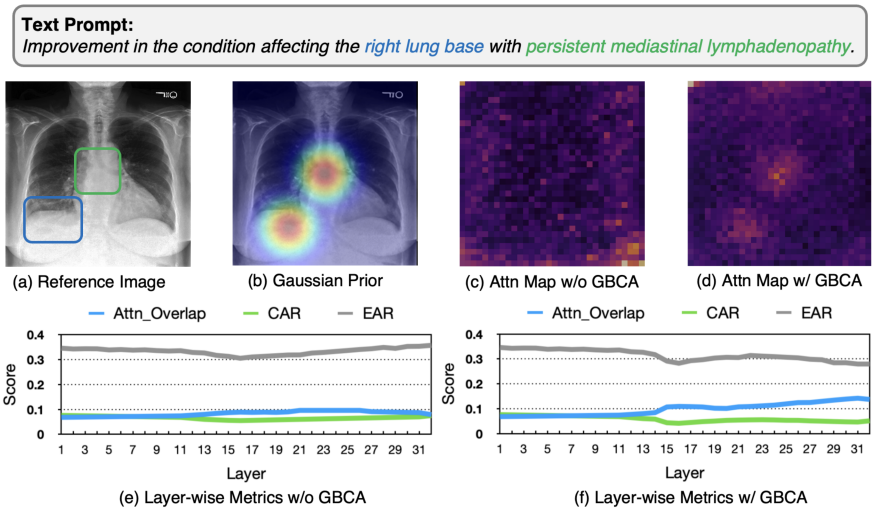

Fig. 1. (a) Reference CXR with a prompt-specified region (box). (b) Lesion-centric Gaussian prior. (c,d) Without GBCA, last-layer attention concentrates on the periphery; GBCA (L14–16) redirects it to the lesion region. (e,f) Layer-wise attention metrics (Attn-Overlap, CAR, EAR) before and after injecting GBCA into middle layers.

scale (e.g., Emu3 [12]) and have been adapted to longitudinal prediction (e.g., ProgEmu [22]). However, these methods rely on implicitly learned attention and often fail to consistently localize or evolve lesions.

We uncover a systematic corner/edge attention bias in AR follow-up generation: causal self-attention disproportionately concentrates on peripheral reference tokens rather than lesion-relevant regions (Fig. 1(c)), reducing attention–lesion overlap and degrading disease fidelity. Layer-wise analysis shows this shortcut is structural: the bias emerges in early layers (higher edge/corner attention ratios (CAR/EAR)) that encode low-level textures, diminishes at mid-depth layers where semantic representations strengthen, and reappears in later layers that enforce global reconstruction (Fig. 1(e)).

To correct this issue, we introduce **LeGend**, a lesion-guided AR framework incorporating **Gaussian-Biased Causal Attention (GBCA)**—a principled, causally aligned bias-correction mechanism embedded directly within AR decoding. GBCA constructs lesion-centric spatial priors from multimodal evidence: an offline vision–language model (VLM) predicts sparse lesion coordinates, which are converted into a normalized 2D Gaussian salience field and injected as an additive bias to causal attention logits. This enables fine-grained modulation of token interactions while preserving autoregressive consistency and model stability. Layer-wise ablations show that mid-depth injection achieves the best trade-off between semantic alignment and image fidelity and produces the improved attention as in Fig. 1(d) and (f), whereas shallow layers lack meaningful lesion semantics and late-layer injection may over-constrain refinement.

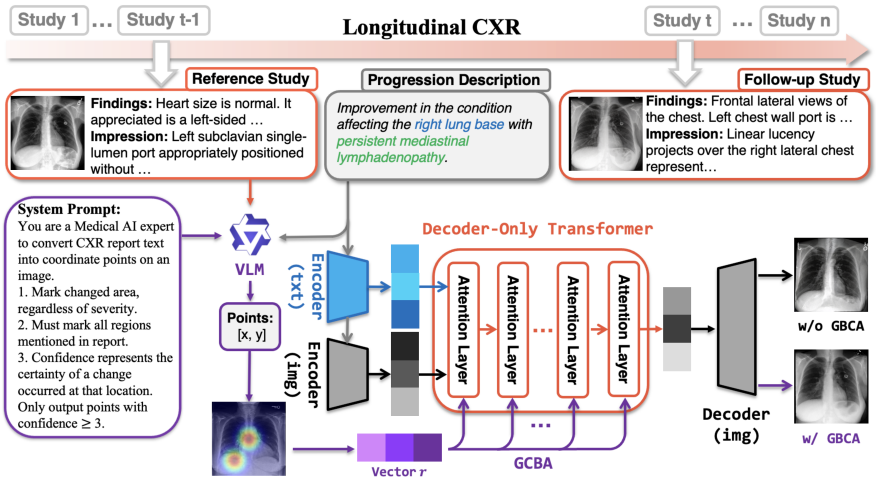

Fig. 2. LeGend overview. Given a reference CXR and a progression description, a VLM predicts sparse lesion coordinates that are then transformed into a normalized 2D Gaussian prior and added to causal self-attention logits (GBCA) in a decoder-only transformer, steering attention to lesion-relevant regions while preserving causality, thus improving follow-up synthesis with minimal overhead.

Our approach is complementary to existing attention-control paradigms. Unlike input-agnostic positional biases (e.g., T5/ALiBi [14,16]) and auxiliary control branches that add training pathways (e.g., ControlNet [24]), it injects a sample-specific lesion prior into causal logits. Unlike post-hoc attention editing that operates after generation and may break autoregressive consistency (e.g., Attend-and-Excite [23]), GBCA guides decoding online, maintaining token-level causality. Our contributions are summarized as follows:

1. We identify and quantify a corner-/edge-biased attention phenomenon in autoregressive longitudinal CXR generation, where fine-tuned transformers over-attend to peripheral non-lesional regions.
2. We propose **LeGend** with **GBCA**, a lightweight and interpretable mechanism that injects a sample-specific lesion prior into causal self-attention logits with negligible overhead, which can be plugged into different AR backbones.
3. We demonstrate consistent gains in semantic alignment and disease fidelity, improving lesion-focused attention, image quality, and downstream disease-classification performance across multiple baselines.

2 Method

Given a reference CXR image I_A and a progression description T, our goal is to generate a follow-up image $\hat{I}_B$ that (i) preserves patient-specific anatomy, and (ii)

localizes and evolves lesions consistent with T. We adopt an autoregressive (AR) decoder-only transformer that models the follow-up visual-token sequence X_B via next-token prediction conditioned on the concatenated context $\{X_A, X_T\}$, where X_A and X_B are the visual-token sequences of I_A and I_B, and X_T denotes text tokens of T. Despite strong global realism, we observe that the AR decoder exhibits a systematic peripheral shortcut: when generating X_B, follow-up queries assign disproportionate attention mass to *corner/edge reference tokens* rather than lesion-relevant regions, leading to lesion misplacement (Fig. 1).

We mitigate this corner/edge attention bias by injecting a lesion-centric spatial prior directly into causal self-attention logits. Specifically, we (i) profile the bias layer-by-layer using attention statistics that explicitly measure how X_B attends to reference keys in X_A (Sec. 2.1); (ii) obtain sparse lesion coordinates from an offline VLM conditioned on (I_A, T) and convert them into a normalized 2D Gaussian prior on the reference-token grid (Sec. 2.2); and (iii) inject this prior as an additive logit bias into selected decoder layers, yielding Gaussian-Biased Causal Attention (GBCA) (Sec. 2.3). Fig. 2 summarizes the pipeline. $(H_{\text{px}}, W_{\text{px}})$ denotes the pixel resolution of the reference image, while (H, W) denotes the decoder's visual-token grid resolution ($H = W = 32$). Grid locations are indexed by (i, j), and pixel locations are indexed by (u, v).

2.1 Layer-wise Attention Bias Profiling

Autoregressive fine-tuning can generate plausible follow-ups but may misplace lesions. To precisely characterize the observed peripheral shortcut, we analyze causal self-attention under teacher forcing on the mixed sequence $\{X_A, X_T, X_B\}$ of length S. Our analysis focuses on the cross attention from follow-up queries (Q_B; indices of tokens in X_B) to reference keys (K_A; indices of tokens in X_A), because this term governs how the decoder reuses reference evidence during follow-up synthesis. We hypothesize this peripheral shortcut arises because border/corner regions often provide stable, low-variance anatomical anchors under registration noise and subtle lesion changes, making them an easy-to-exploit conditioning cue for next-token prediction.

Let $L^{(\ell)} = \frac{Q^{(\ell)}(K^{(\ell)})^\top}{\sqrt{d_k}} + M \in R^{S\times S}$ denote the masked pre-softmax attention logits at layer ℓ for $\{X_A, X_T, X_B\}$, where M is the causal mask. We extract the sub-block $L^{(\ell)}[Q_B, K_A]$, normalize only over reference keys $k \in K_A$, and average over follow-up queries to obtain a reference-grid heatmap:

$$A^{(\ell)}_{i,j} = \frac{1}{|Q_B|} \sum_{q\in Q_B} \left[\text{Softmax}_{k\in K_A}\left(L^{(\ell)}_{q,k}\right)\right]_{k=g_A(i,j)}, \tag{1}$$

where $g_A(i, j)$ maps a grid location (i, j) to its corresponding reference-token index in K_A. By construction, $\sum_{i=1}^{H}\sum_{j=1}^{W} A^{(\ell)}_{i,j} = 1$ over the reference grid.

Using $A^{(\ell)}$, we quantify peripheral over-focus by the Corner Attention Ratio (CAR) and Edge Attention Ratio (EAR), and assess clinical relevance by Attention-Lesion overlap (Attn-Overlap) under teacher forcing. Attn-Overlap measures the probability mass assigned to the lesion-relevant region Ω.

$$\text{CAR} = \sum_{(i,j)\in C} A_{i,j}^{(\ell)}, \quad \text{EAR} = \sum_{(i,j)\in E} A_{i,j}^{(\ell)}, \quad \text{Attn-Overlap} = \sum_{(i,j)\in \Omega} A_{i,j}^{(\ell)}, \quad (2)$$

where C denotes the union of four corner squares with side length αH and E the border band of width βH (we use $\alpha = 0.12$, $\beta = 0.08$). The lesion-relevant region is defined as $\Omega = (i,j) \mid R_{i,j} \geq \tau$ with $\tau = 0.5$, using the same saliency prior R as in GBCA (Sec. 2.2) to keep bias profiling and injection consistent. We compute layer-wise attention statistics under teacher forcing, while GBCA is applied during autoregressive decoding (with KV-cache) to steer generation.

2.2 2D Gaussian Prior from an Offline VLM

We obtain lesion-relevant spatial cues from an offline vision–language model (VLM) conditioned on (I_A, T). The VLM outputs K lesion-related pixel coordinates $\{p_k = (u_k, v_k)\}_{k=1}^{K}$. We convert these sparse points into a dense salience prior by placing an isotropic 2D Gaussian at each coordinate and aggregating them with point-wise maximum:

$$\tilde{R}(u,v) = \max_k \exp\left(-\frac{(u-u_k)^2 + (v-v_k)^2}{2\sigma^2}\right), \quad (3)$$

where (u, v) denotes pixel locations and σ controls the spatial spread; we set $\sigma = 0.1 \cdot \min(H_{\text{px}}, W_{\text{px}})$ and use a point-wise maximum over coordinates to keep the prior in $(0, 1]$ without scaling with K. The pixel-space prior is then downsampled to the token grid, $R = \text{Down}(\tilde{R}) \in R^{H\times W}$, and flattened to $r = \text{vec}(R) \in R^{HW}$ such that $r_{g_A(i,j)} = R_{i,j}$.

2.3 Gaussian-Biased Causal Attention

For causal self-attention at layer ℓ, given logits $L^{(\ell)}$ and the prior r (Sec. 2.2), we apply the key-position bias only to reference-key indices K_A (text and follow-up keys receive zero bias). Broadcasting this 1D bias across heads and queries yields B_{gaussian}, injected at the logit level:

$$\hat{L}^{(\ell)} = L^{(\ell)} + s^{(\ell)} \cdot B_{\text{gaussian}}, \quad (4)$$

where $s^{(\ell)}$ is a learnable scalar produced by a lightweight two-layer MLP (one per layer) to adapt the bias strength. We add the bias before Softmax to preserve numerical stability and to allow controlled gradient flow through the attention distribution. GBCA is plug-and-play and can be applied to selected layers without modifying the backbone architecture.

Table 1. Comparison of image generation quality and downstream classification.

Method	Generation Quality				Classifier Quality	
	FID↓	CLIP-T↑	MS-SSIM↑	PSNR↑	AUC↑	F1↑
BioMedJourney [21]	47.5307	35.01	0.4479	12.09	0.6166	0.7423
PIE [20]	49.0129	35.13	0.4982	12.92	0.6479	0.7764
CXR-IRGen [25]	44.1238	32.38	0.4760	12.26	0.6230	0.7544
ProgEmu [22]	35.0192	35.45	0.4884	13.22	0.7153	0.8132
LeGend (Ours)	**26.1247**	**38.27**	**0.6617**	**15.87**	**0.7559**	**0.8402**
EditAR [37]	39.5823	35.57	0.5973	14.08	0.7521	0.7875
EditAR w/ GBCA	33.5722	36.21	0.6313	14.74	0.7872	0.8040

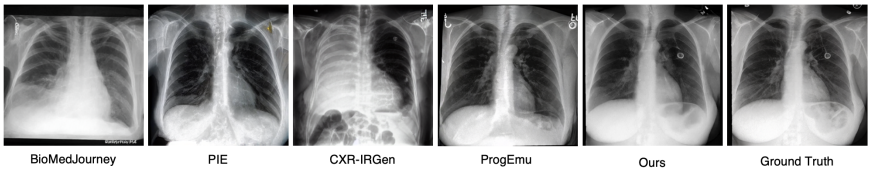

Fig. 3. Visual comparison of follow-up CXR generation.

3 Experiments

3.1 Dataset

We evaluate on ICG-CXR [22], a longitudinal, report-conditioned CXR benchmark curated from MIMIC-CXR [27] and CheXpertPlus [28]. It contains 11,439 PA-view reference–follow-up pairs from 7,388 patients and uses the official patient-disjoint split (10,679/760 train/test). Each sample provides a registered reference and real follow-up CXR pair, the corresponding reports, and an LLM-generated progression description used as the text condition.

3.2 Experimental Setup

Evaluation metrics. We measure attention alignment with Attn-Overlap and peripheral bias with CAR/EAR (Eq. 2). Generation quality is evaluated by FID, CLIP-T (using BioMedCLIP [30]), MS-SSIM, and PSNR, and classifier-assessed disease consistency by AUC/F1 of a pretrained disease classifier [29]. These classifier-based metrics are used as proxy measures of disease-relevant consistency and do not replace radiologist-based clinical validation. All methods use the same test split and identical registration/resolution. For attention-based metrics (Attn-Overlap, CAR, EAR), unless stated otherwise, we report statistics computed from the last decoder layer under teacher forcing, consistent with Fig. 1(c–f). **Training and inference.** GBCA is used in both training and inference. We fine-tune the autoregressive decoder with teacher forcing and next-token cross-entropy on X_B conditioned on (X_A, X_T), where the Gaussian prior r

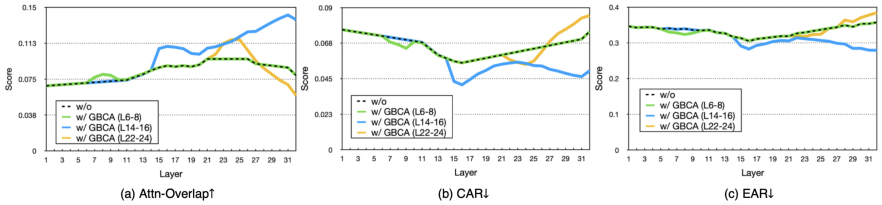

Fig. 4. Attention profiling across layers shows that GBCA injected in mid-layer improves lesion focus most effectively.

is built from VLM-predicted coordinates given (I_A, T) and injected into selected layers. At inference, we construct r in the same way and apply the identical bias during autoregressive decoding with KV-cache. **Implementation details.** We fine-tune Emu3 with LoRA on all self-attention projections (q/k/v/o), using r=32, α=64, dropout= 0.05, and train for 6k steps on 2× RTX A6000 with gradient checkpointing. We use batch size 16 with gradient accumulation, AdamW, cosine LR ($1\times10^{-5} \rightarrow 1\times10^{-6}$), and gradient clipping ($\|g\|_2 \leq 5$).

3.3 Results

Comparison with existing methods. We compare LeGend with diffusion baselines (BioMedJourney [21], PIE [20], CXR-IRGen [25]) and the autoregressive baseline ProgEmu [22] on ICG-CXR under a unified protocol (same resolution, preprocessing, and train split). All baselines are fine-tuned on the ICG-CXR training set. As shown in Tab. 1, LeGend achieves the best overall performance, with the lowest FID and consistently higher CLIP-T, MS-SSIM, and PSNR, indicating improved realism, semantic alignment, and structural fidelity. **Disease identifiability.** LeGend also attains the best AUC and F1 under a pretrained thoracic disease classifier [29] (Tab. 1), suggesting improved disease-relevant consistency under this proxy evaluation. **Qualitative results.** Fig. 3 shows that diffusion baselines often yield visually plausible but less faithful follow-ups, while ProgEmu better preserves global anatomy yet may drift in lesion localization. We attribute LeGend's gains to lesion-conditioned attention steering: GBCA increases the probability mass assigned to lesion-relevant reference tokens during decoding, which improves the reuse of patient-specific local evidence and suppresses peripheral shortcuts, thereby producing sharper anatomy and more accurately localized progression consistent with the text condition. **Generality across autoregressive backbones.** We transfer GBCA to a different autoregressive backbone, EditAR [37]. We first fine-tune EditAR on ICG-CXR, then freeze all backbone parameters and train only the lightweight GBCA module (injected at the middle transformer layer). As shown in Tab. 1, GBCA consistently improves both generation quality and downstream pathology classification on EditAR, supporting GBCA's generality across AR architectures.

Table 2. Ablation of GBCA injection depth and mid-depth layer selection.

Method	FID↓	CLIP-T↑	MS-SSIM↑	PSNR↑	AUC↑	F1↑
(A) Injection depth (coarse)						
Ours (L6–8)	34.7979	35.50	0.4910	13.30	0.7160	0.8140
Ours (L14–16)	**26.1247**	**38.27**	**0.6617**	**15.87**	**0.7559**	**0.8402**
Ours (L22–24)	40.2545	33.20	0.4710	12.60	0.6127	0.7250
Ours (L14–24)	41.8016	32.49	0.4592	12.88	0.5721	0.6817
(B) Mid-depth layer selection (fine, within L14–16)						
Ours (L14)	28.6369	36.86	0.6131	14.24	0.7436	0.8270
Ours (L15)	28.1581	37.69	0.6509	14.78	0.7543	0.8385
Ours (L16)	29.1533	36.43	0.5928	13.92	0.7452	0.8287
Ours (L14–15)	27.0684	38.18	0.6581	15.77	0.7554	0.8396
Ours (L15–16)	27.3550	38.12	0.6557	15.31	0.7550	0.8393

Table 3. The ablation of VLM

Coordinate Source	FID↓	CLIP-T↑	MS-SSIM↑	PSNR↑	AUC↑	F1↑
Qwen2.5-VL	**26.1247**	**38.27**	**0.6617**	**15.87**	0.7559	**0.8402**
Llama-3.2-vision	29.0563	37.68	0.6396	15.54	**0.7742**	0.8370
Random Coordinate	44.3268	32.58	0.4702	12.56	0.6084	0.7211

3.4 Ablation Study

Injection depth and layer selection. We ablate where to inject the Gaussian prior in the 32-layer autoregressive decoder (Tab. 2). Mid-layer injection (L14–16) achieves the best overall trade-off, yielding the lowest FID while consistently improving CLIP-T, MS-SSIM, PSNR, and the clinical proxy metrics (AUC/F1). In contrast, shallow injection (L6–8) provides only marginal gains, whereas late injection (L22–24) or an overly wide window (L14–24) substantially degrades performance, suggesting that biases injected too early are overwritten and biases injected too late (or too repeatedly) can disrupt late-stage refinement. (Tab. 2A and Fig. 4). We further conduct a fine-grained study within L14–16 (Tab. 2B). Single-layer injection already improves over the baseline, with L15 as the strongest choice; injecting two adjacent layers (L14–15 or L15–16) further stabilizes the guidance, and the compact block L14–16 performs best overall.

Coordinate Ablation for Gaussian prior. Tab. 3 compares GBCA using lesion coordinates from different VLMs. The Qwen and Llama point sets differ by $\overline{d}$ =41.54px (σ = 18.36px) after minimum-cost matching. Despite noticeable localization variance, GBCA yields consistent gains with either source. This suggests GBCA does not require pixel-level accurate points; coarse, prompt-relevant localization is sufficient to provide an effective lesion-centric bias in attention logits. Crucially, replacing VLM coordinates with random points (while K, σ, and injection layers are identical) significantly degrades both generation quality and clinical proxy performance. This sanity check rules out a trivial

regularization explanation and confirms that GBCA benefits from patient- and prompt-specific localization rather than an arbitrary spatial bias. Nevertheless, GBCA remains dependent on the quality of the VLM-derived prior: wrong-side, outside-lung, missing, or hallucinated coordinates may weaken or misdirect the attention bias. Since the prior is injected softly and the model remains conditioned on the reference CXR and progression text, such errors do not impose a hard editing constraint, but improving lesion localization remains an important direction. When available, radiologist annotations can directly replace VLM points for higher-precision priors.

4 Conclusion

We identify a consistent peripheral-attention bias in autoregressive follow-up CXR generation and show that correcting this issue improves lesion-focused attention and classifier-assessed disease consistency. LeGend addresses this by softly steering attention toward lesion-relevant regions through a lightweight, lesion-informed prior, resulting in stronger lesion alignment, more meaningful attention behavior, and improved disease-classification performance. These findings highlight the value of targeted attention correction in enhancing the effectiveness of autoregressive medical image synthesis.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Wani, Manzoor, et al. "Beyond Pneumonia: A Dual-Cycle Initiative to Improve Imaging Compliance and Uncover Early Malignancy." Cureus 17.8 (2025).
2. Little, Brent P., et al. "Outcome of recommendations for radiographic follow-up of pneumonia on outpatient chest radiography." American Journal of Roentgenology 202.1 (2014): 54-59
3. Nehvi, Aabid, et al. "Enhancing adherence to British Thoracic Society guidelines for follow-up chest radiographs in community-acquired pneumonia: a quality improvement project." Cureus 17.9 (2025).
4. Amorosa, Judith K., et al. "ACR appropriateness criteria routine chest radiographs in intensive care unit patients." Journal of the American College of Radiology 10.3 (2013): 170-174.
5. Al Shahrani, Abdullah, and Khaled Al-Surimi. "Daily routine versus on-demand chest radiograph policy and practice in adult ICU patients-clinicians' perspective." BMC Medical Imaging 18.1 (2018): 4.
6. Khan, Ali Nawaz, et al. "Reading chest radiographs in the critically ill (Part I): Normal chest radiographic appearance, instrumentation and complications from instrumentation." Annals of Thoracic Medicine 4.2 (2009): 75-87.
7. Chandrashekar, Mayanka, et al. "Domain shift analysis in chest radiographs classification in a veterans healthcare administration population." Journal of Imaging Informatics in Medicine (2025): 1-16.

8. Musa, Aminu, Rajesh Prasad, and Monica Hernandez. "Addressing cross-population domain shift in chest X-ray classification through supervised adversarial domain adaptation." Scientific Reports 15.1 (2025): 11383.
9. Cohen, Joseph Paul, et al. "On the limits of cross-domain generalization in automated X-ray prediction." Medical Imaging with Deep Learning. PMLR, 2020.
10. Xue, Zhiyun, et al. "Cross dataset analysis of domain shift in cxr lung region detection." Diagnostics 13.6 (2023): 1068.
11. Zech, John R., et al. "Confounding variables can degrade generalization performance of radiological deep learning models." arXiv preprint arXiv:1807.00431 (2018).
12. Wang, Xinlong, et al. "Emu3: Next-token prediction is all you need." arXiv preprint arXiv:2409.18869 (2024).
13. Geirhos, Robert, et al. "Shortcut learning in deep neural networks." Nature Machine Intelligence 2.11 (2020): 665-673.
14. Raffel, Colin, et al. "Exploring the limits of transfer learning with a unified text-to-text transformer." Journal of machine learning research 21.140 (2020): 1-67.
15. Liu, Ze, et al. "Swin transformer: Hierarchical vision transformer using shifted windows." Proceedings of the IEEE/CVF international conference on computer vision. 2021.
16. Press, Ofir, Noah A. Smith, and Mike Lewis. "Train short, test long: Attention with linear biases enables input length extrapolation." arXiv preprint arXiv:2108.12409 (2021).
17. Kim, Bum Jun, et al. "Understanding gaussian attention bias of vision transformers using effective receptive fields." arXiv preprint arXiv:2305.04722 (2023).
18. Hertz, Amir, et al. "Prompt-to-prompt image editing with cross attention control." arXiv preprint arXiv:2208.01626 (2022).
19. Hong, Susung, et al. "Improving sample quality of diffusion models using self-attention guidance." Proceedings of the IEEE/CVF International Conference on Computer Vision. 2023.
20. Liang, Kaizhao, et al. "Pie: Simulating disease progression via progressive image editing." arXiv preprint arXiv:2309.11745 (2023).
21. Gu, Yu, et al. "Biomedjourney: Counterfactual biomedical image generation by instruction-learning from multimodal patient journeys." arXiv preprint arXiv:2310.10765 (2023).
22. Ma, Chenglong, et al. "Towards interpretable counterfactual generation via multimodal autoregression." International Conference on Medical Image Computing and Computer-Assisted Intervention. Cham: Springer Nature Switzerland, 2025.
23. Chefer, Hila, et al. "Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models." ACM transactions on Graphics (TOG) 42.4 (2023): 1-10.
24. Zhang, Lvmin, Anyi Rao, and Maneesh Agrawala. "Adding conditional control to text-to-image diffusion models." Proceedings of the IEEE/CVF international conference on computer vision. 2023.
25. Shentu, Junjie, and Noura Al Moubayed. "Cxr-irgen: An integrated vision and language model for the generation of clinically accurate chest x-ray image-report pairs." Proceedings of the IEEE/CVF winter conference on applications of computer vision. 2024.
26. Rombach, Robin, et al. "High-resolution image synthesis with latent diffusion models." Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2022.

27. Johnson, Alistair EW, et al. "MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports." Scientific data 6.1 (2019): 317.
28. Chambon, Pierre, et al. "Chexpert plus: Augmenting a large chest x-ray dataset with text radiology reports, patient demographics and additional image formats." arXiv preprint arXiv:2405.19538 (2024).
29. Cohen, Joseph Paul, et al. "On the limits of cross-domain generalization in automated X-ray prediction." Medical Imaging with Deep Learning. PMLR, 2020.
30. Zhang, Sheng, et al. "Biomedclip: a multimodal biomedical foundation model pretrained from fifteen million scientific image-text pairs." arXiv preprint arXiv:2303.00915 (2023).
31. Heusel, Martin, et al. "Gans trained by a two time-scale update rule converge to a local nash equilibrium." Advances in neural information processing systems 30 (2017).
32. Wang, Zhou, Eero P. Simoncelli, and Alan C. Bovik. "Multiscale structural similarity for image quality assessment." The thrity-seventh asilomar conference on signals, systems computers, 2003. Vol. 2. Ieee, 2003.
33. Bradley, Andrew P. "The use of the area under the ROC curve in the evaluation of machine learning algorithms." Pattern recognition 30.7 (1997): 1145-1159.
34. Wu, Chenfei, et al. "Qwen-image technical report." arXiv preprint arXiv:2508.02324 (2025).
35. Grattafiori, Aaron, et al. "The llama 3 herd of models." arXiv preprint arXiv:2407.21783 (2024).
36. Bolya, Daniel, et al. "Perception encoder: The best visual embeddings are not at the output of the network." arXiv preprint arXiv:2504.13181 (2025).
37. Mu, Jiteng, Nuno Vasconcelos, and Xiaolong Wang. "Editar: Unified conditional generation with autoregressive models." Proceedings of the Computer Vision and Pattern Recognition Conference. 2025.