ClinX: Multimodal De-Identification for Robust and Bias-Resilient Medical VLMMs

Anonymous ACL submission

Abstract

Visual large multimodal models (VLMMs) are increasingly being adopted for medical applications such as VQA in the clinical domain, but the use of such models on sensitive medical data raises significant concerns about privacy and generalization. To overcome the challenges associated with the use of such sensitive medical information, we present ClinX, a privacy-preserving medical multimodal inference system that de-identifies data while remaining Health Insurance Portability and Accountability Act (HIPAA) Safe Harbor compliant without compromising model diagnostic performance. The image de-identification component of the system uses a custom SPADE-based generative adversarial network (PP-GAN) approach to perform image redaction by inpainting regions marked by the OCR engine as PHI (e.g., names, timestamps) followed by lightweight mask aware postprocessing. The text de-identification component employs a three-step approach involving rule-based redaction, named entity recognition, and neural rewriting. We present experimental results using two medical multimodal VQA datasets, VQA-RAD and PathVQA, using a medical VLMM model, LLaVA-Med, and a large-scale general-purpose model, Llama, as a semantic judge. Experimental results show strong preservation of exact-match and semantic accuracy after de-identification. Furthermore, the proactive removal of textual overlays mitigates dataset-specific bias, in some cases even enhancing robustness by eliminating spurious textual shortcuts. These results validate ClinX as a practical and secure solution for privacy-conscious medical AI deployment.

1 Introduction

Recent successes of Large Language Models(LLMs) and Visual Large Multimodal Models (VLMMs) have helped achieve remarkable results in medical image comprehension tasks, especially for tasks like Visual Question Answering (VQA), clinical captioning, and radiology report generation tasks (Nazi and Peng, 2024; Nassiri and Akhloufi, 2024). These models excel in critical areas such as clinical decision support (Prabhod, 2023) and medical question-answering (Li et al., 2024a), providing early detection of medical conditions and delivering actionable insights for timely intervention (Gupta et al., 2024; Gao et al., 2024). Models like LLaVA have proven their ability to match the textual and visual modalities very effectively, placing them among the most promising solutions for decision-making in real-world clinical environments.

However, the use of these models within the healthcare industry is affected by two fundamental issues: privacy and generalization. Patient healthcare data, particularly in radiology images, tends to have PHI embedded within them in the form of patient name, date of birth, institutional ID, or timestamps (Truong et al., 2025). The HIPAA laws require the removal of all these identifiers prior to sharing the data or using them in the models. Notably, the overlaid texts in clinical images not only cause privacy issues but also work as spurious features that can easily be leveraged by multimodal models for dataset specific shortcut learning. Active removal of all overlaid texts, regardless of whether they contain explicit PHI, by ClinX enhances its robustness and reduces bias, as observed from its stable or improved performance on downstream VQA tasks.

To overcome these challenges, we introduce ClinX, which is a unified privacy-preserving framework for multimodal medical inference system that leverages the following two modules: (1) the visual de-identification module that is developed by using the SPADE-conditioned PP-GAN and includes the mask-aware post-processing step for the removal of GAN-induced artifacts within the regions covered by the PHI mask, and (2) the three-level text de-

identification module that uses rule-based redaction and the NER model with the rewriting transform for sanitizing the inputs, while preserving the clinical semantics.

We also benchmark our system, ClinX, on two standard datasets, VQA-RAD, and PathVQA, using medical VLMM for inference and a state-of-the-art language model for an LLM-based semantic correctness judge. We establish that despite extensive image and text redactions, our metric scores generally exhibit little-to-no degradation, or in certain instances, improve slightly for exact match and LLM-as-judge accuracy. Most notably, however, is that the proactive approach taken by ClinX to remove text overlays advances resilience and robustness of our system by addressing distributional bias issues related to learning shortcuts.

Our contributions are as follows: (1) We introduce ClinX, a HIPAA-compliant multimodal inference system for privacy-preserving clinical tasks. (2) We propose a PP-GAN based on SPADE and use mask-aware post-processing for OCR-supervised PHI inpainting inorder to improve the visual realism. (3) We develop a robust text de-identification pipeline that applies both symbolic and neural approaches for obfuscating sensitive language inputs. (4) We demonstrate that ClinX preserves or sometime improves downstream task performance across multiple accuracy and semantic evaluation metrics. ClinX opens a path to applying VLMMs in real-world medical scenarios, providing both privacy preservation and generalization.

2 Related Works

This section reviews the related works of medical large language models and privacy-preserving mechanisms used in medical LLMs.

2.1 Medical Large Language Models

Medical LLMs have shown strong efficacy in clinical decision support, medical question answering, and patient-interface applications, with real-world deployment of GPT-4-based medical assistants highlighting their potential impact (Haltaufderheide and Ranisch, 2024). The development of medical LLMs is propelled by extensive benchmarking initiatives such as Med-PaLM 2 on MedQA (Qian et al., 2024) or the Open Medical-LLM Leaderboard (Omiye et al., 2024), which systematically identify strengths and limitations of medical LLMs.

The clinical utility has also been enhanced by domain-specific pre-training on biomedical corpora such as PubMed (White, 2020) and MIMIC III (Johnson et al., 2016), and the use of electronic health record and hybrid AI pipelines (Li et al., 2024a; Goyal et al., 2024). Recent developments have included agent-based reasoning platforms like MedAgents (Tang et al., 2023), medical multi-service systems supporting multiple languages like Apollo (Wang et al., 2024b), and knowledge graph-based models that boost contextual understanding capabilities (Wang et al., 2024a). Although there have been considerable developments in medical LLMs, they also increases the concerns regarding the patient privacy , motivating the requirement of the privacy-aware framework design.

2.2 Privacy-Preserving Mechanisms in Medical LLMs

The issue of preserving sensitive medical data is a key challenge in medical LLMs. Differential privacy (DP) (Letafati and Otoum, 2023; Ziller et al., 2021) and homomorphic encryption (Kumar et al., 2020; Chirra, 2023) are some common methods that are currently being used to deal with risks associated with sensitive medical data leakages in medical AI models. DP training methods, specifically, have been identified to prevent medical information from being memorized by LLMs (Song et al., 2024; Salim et al., 2024), making them a foundational component of privacy-preserving medical model development.

Federated learning allows collaborative training of models within institutions without gathering the original patient data (Li et al., 2024b; Peng et al., 2023) in the central server. With the current advancements in the development of multimodal LLMs in medicine, which combine text, images, and structural biomedical data, strict privacy, particularly for cross-modal biomedical reasoning, is a challenging task from a technical perspective.

2.3 Image De-Identification in Medical AI

Image de-identification is a crucial step for HIPAA Safe Harbor compliance requirements on clinical data sharing (Johnson et al., 2020). Clinical and radiology images usually contain identifiers of patients as overlays. The basic approach to de-identification could potentially introduce artifacts, thus hampering downstream model performance. Generative models can offer a much-realistic choice in this area to replace regions. For

instance, DeepPrivacy generates a synthetic face to replace identifiable facial regions (Hukkelås et al., 2019).

Spatially-adaptive normalization techniques, for example, SPADE (Park et al., 2019), allow conditional GANs to conduct inpainting tasks guided by semantic masks, thereby facilitating structure-aware inpainting. These methods can decrease the domain shift problem while retaining the diagnostic utility, which makes them more applicable for privacy-preserving medical vision-language models like LLaVA.

2.4 Text De-Identification in Clinical NLP

Rule-based and regex-driven approaches to clinical de-identification were the earliest methods and tended to impact the coherence and readability of the clinical content (Johnson et al., 2020). More modern models based on the Transformer architecture and Named Entity Recognition (NER), such as BERT-NER (Liu et al., 2021), demonstrate a dramatic increase in PHI detection precision. More recent approaches now also leverage paraphrasing based on T5 architecture for successful removal of sensitive PHI with reduced impact on clinical semantics (Johnson et al., 2020). This layered strategy of using regex, NER and neural rewrite ensures the PHI removal while preserving clinical utility.

2.5 LLM-Based Evaluation and Semantic Judging

Standard evaluation metrics like exact match or F1 often fails to capture the similarity of semantically equivalent paraphrases. To combat these challenges, large language models, including GPT-4, are emerging as objective evaluators as a automated semantic judges (Zheng et al., 2023; Yamauchi et al., 2025). For the medical LLM performance evaluation, LLM-as-judge framework ensures more equitable assessment of the model prediction on the de-identified input, providing nuanced correctness assessments beyond surface-level token matching.

3 Methodology

We propose, ClinX, a privacy preserving multimodal framework for medical visual question answering. ClinX accepts an input of a medical image (example an X-ray or CT-scan), and an associated text (which could be a question, a prompt, or any text snippets) and then outputs a sanitized and anonymized text and images, which is then can be fed into a multimodal LLM to generate the output, which is shown in the figure 1. Then, the accuracy of the outputs produced by the LLM is compared using both metric computation (accuracy) and an LLM judge. In this section, we breakdown every-part of the ClinX in detail:

3.1 OCR-Guided Image Masking

The first step for the Image de-identification is to detect regions that potentially contains the PHI. So in order to identify text in images, we use an Optical Character Recognition (OCR) Engine, under the assumption that any readable text in a medical image could be an identifier (patient name, ID, date, hospital name, hospital location, etc.) unless it is proven otherwise. Here, We use a robust and open source OCR engine, EasyOCR, that gives us bounding boxes around detected text along with corresponding text. We don't attempt to semantically classify the text, and we take the conservative approach, that all detected text are treated as a PHI and targeted for removal. This approach aligns with a "Safe Harbor" practice and avoids the potential risk of leaking the patient's PHI as a normal word. To improve the analysis and reliability of the masking, we implement following enhancements: **Bounding Box Expansion:** A slight expansion of every discovered text area has been performed in order to include the entire text in it, inflating the bounding box by a small margin(5 pixels or a few percent of width/height) in all directions. A binary mask "m" for the image has been constructed, where all pixels in any of these bounding boxes are set to 1, and all other pixels are 0. **Feathering:**To avoid any hard boundaries or abrupt edges, feathering which is basically a Gaussian blur has been employed to soften this transition line between the masked part of the picture and the rest. **Heuristic Fallback Masking:** To avoid PHI leakage in case of OCR failures, a fallback mechanism has been incorporated. A large number of PHI components in scan images are known to lie in a certain location in predictable patterns, for example, in radiological images, near the top-or-bottom borders. Instead of text, we search for a low-contrast rectangle that may correspond to text overlapping. This involves threshold, morphology, and search for a low-contrast area likely to possess text overlays in the vicinity of boundaries. If present, it would be heuristically masked as PHI. We prioritize recall over precision, as false positives are seamlessly handled by our GAN inpainting. The output pro-

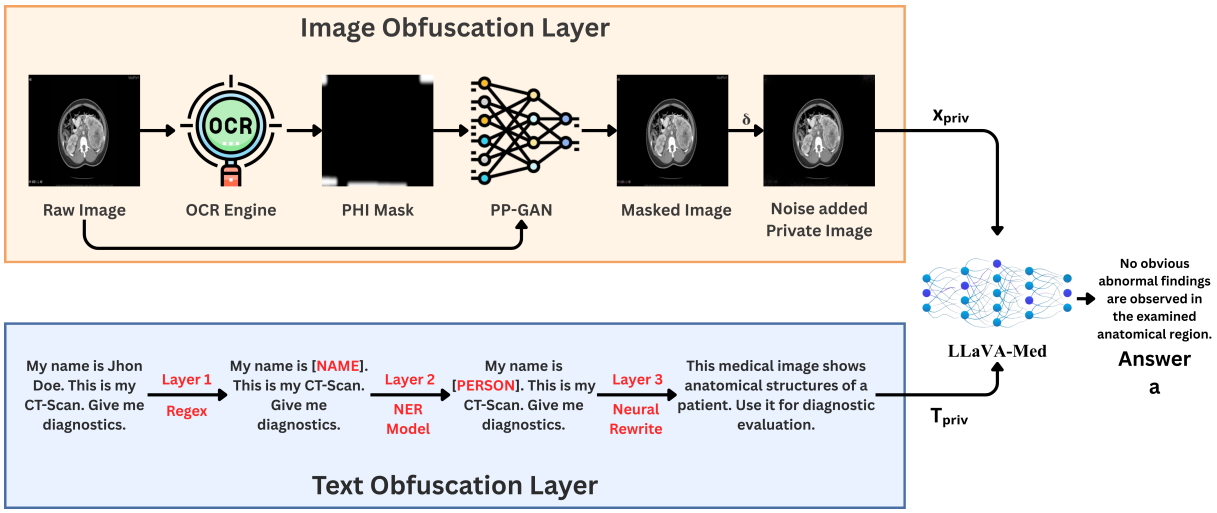


Figure 1: ClinX end-to-end workflow illustrating image and text obfuscation pipelines.

duced by the OCR model is a spatial mask denoted by $m \in [0,1]^{1\times H\times W}$, where $m(p) = 1$ indicates text, $m(p) = 0$ denotes background, and intermediate values represent feathered boundaries. A metric, mask_pct, tracks the total fraction of the image occluded. (See Appendix F for the corresponding inference-stage placement.)

3.2 PP-GAN: Privacy Preserving Image De-Identification with Inpainting

After having identified the areas to remove, we have OCR-generated mask m and our original picture x having the PHI information, and we then proceed to use our generative inpainting model, to fill the marked areas with content that appears natural. We term this model Privacy Preserving Generative Adversarial Network, or in short PP-GAN. Given an input medical image $x \in \mathbb{R}^{3\times H\times W}$ and a soft PHI mask $m \in [0,1]^{1\times H\times W}$ obtained via OCR-based text localization, PP-GAN generates a de-identified image $\hat{x}$ using a conditional generator:

$$\hat{x} = G(x, m). \tag{1}$$

The final privacy-preserved image is produced via mask-based compositing:

$$x_{\text{priv}} = \hat{x} \odot m + x \odot (1 - m), \tag{2}$$

ensuring that only masked PHI regions are modified while all unmasked anatomical content is preserved exactly. The PP-GAN architecture consists of two major components: generator G and a discriminator D.

3.2.1 Generator Architecture: U-Net with SPADE Conditioning

The generator of PP-GAN denoted by G and is inspired by a U-Net-like convolution encoder-decoder architecture, but without the skip connections, which prevents simply copying of the input pixels to output. To limit the synthesis only to the PHI regions m, we use Spatially-Adaptive Denormalization (SPADE)-based conditioning (Park et al., 2019). Particularly, the intermediate features f are modulated using spatially-adaptive normalization:

$$\text{SPADE}(f, m) = \text{BN}(f) \odot (1 + \gamma(m)) + \beta(m), \tag{3}$$

where $\gamma(m)$ and $\beta(m)$ are affine parameters learned from the mask m through a shallow CNN, and $\text{BN}(\cdot)$ denotes batch normalization without affine parameters. This enables the generator to focus its realistic modifications only on the masked zones, leaving the rest of the image structurally consistent.

3.2.2 PatchGAN Discriminator

The discriminator D is a CNN that has a PatchGAN architecture (Isola et al., 2018) that evaluates the image realism in the level of local small patches rather than giving a global authenticity score. Given an image $\tilde{x}$ and the corresponding PHI mask m, the discriminator predicts a two-dimensional grid of real/fake confidence scores/logits, where each entry corresponds to a fixed-size spatial patch(for example $70*70$ $pixels$). This encourages detailed textures in the reconstructed, as well as in the untouched, parts of the

image, supporting fine-grained adversarial training. Although the discriminator doesn't explicitly take the PHI mask m as an input, it is *implicitly conditioned* on masked regions during training. Specifically, unmasked regions are identical between real images x and generated images $\hat{x}$, forcing the discriminator to focus its discrimination capacity on the inpainted (masked) regions.

3.2.3 Adversarial Training Objective

The generator G and discriminator D are trained adversarially using a least-squares GAN (LSGAN) formulation. The discriminator loss is defined as:

$$\mathcal{L}_D = \mathbb{E}_{x\sim p_{\text{data}}}\left[(D(x)-1)^2\right] + \mathbb{E}_{\hat{x}\sim p_G}\left[D(\hat{x})^2\right], \tag{4}$$

while the generator's adversarial loss is:

$$\mathcal{L}_{\text{adv}} = \mathbb{E}_{\hat{x}\sim p_G}\left[(D(\hat{x})-1)^2\right], \tag{5}$$

where $x \sim p_{\text{data}}$ denotes samples from the real image distribution, and $\hat{x} = G(x, m) \sim p_G$ denotes images generated by the PP-GAN given an input image x and an OCR-derived PHI mask m.

The full generator objective combines adversarial and reconstruction-based terms:

$$\begin{aligned}\mathcal{L}_G = \mathcal{L}_{\text{adv}} + \lambda_{\text{L1}}\mathcal{L}_{\text{maskedL1}} + \lambda_{\text{id}}\mathcal{L}_{\text{identity}} \\ + \lambda_{\text{tv}}\mathcal{L}_{\text{TV}} + \lambda_{\text{perc}}\mathcal{L}_{\text{LPIPS}},\end{aligned} \tag{6}$$

where, $\mathcal{L}_{\text{maskedL1}} = \|(\hat{x} - x) \odot m\|_1$ encourages accurate synthesis within PHI-masked regions. $\mathcal{L}_{\text{identity}} = \|(\hat{x} - x) \odot (1 - m)\|_1$ preserves content outside the masked areas. $\mathcal{L}_{\text{TV}}$ denotes total variation loss for spatial smoothness. $\mathcal{L}_{\text{LPIPS}}$ is the perceptual similarity loss ensuring high-level structural fidelity.

3.2.4 Mask-Aware Postprocessing

To further prevent the deterministic inpainting artifacts and reduce the risk of re-identification of the masked PHI information, we further apply a lightweight mask-aware postprocessing module after image synthesis. This step simply applies low-amplitude stochastic perturbations limited only to masked and boundary regions, including spatially weighted Gaussian grain, localized color jitter, micro-blur, and optional light JPEG re-compression. These perturbations obscure potential synthetic fingerprints that may allow attackers to rebuild the original images while preserving diagnostic structure outside PHI regions.

In short, our framework of PP-GAN provides a robust mechanism for medical image de-identification, enabling downstream multi-modal reasoning models to operate on privacy-preserved inputs without sacrificing clinical accuracy and relevance.

3.2.5 Bias Reduction Rationale

Textual overlays in medical images, such as patient names, time stamps, or hospital names, can represent not only a risk of privacy but also a problem in model behavior. For example, a vision-language model might learn a shortcut based on particular text entries in a dataset, such as the name of a hospital ("Hospital A"), to make predictions, which can negatively affect model generalization in another setting. To address this problem, ClinX chooses to remove all such text using both OCR masking and PP-GAN image inpainting.

Extended architectural diagrams and training dynamics are provided in Appendix A and Appendix B.

3.3 Privacy-Preserving Text Obfuscation

A three-tiered approach is utilized by the ClinX system for ensuring that the sharing of clinical texts is HIPAA compliant by removing sensitive information from texts that support medical images. This is achieved through the combination of accurate rule-based matching, Named Entity Recognition (NER) models, as well as neural text rewriting techniques for completely removing personally identifiable information, or PHI.

3.3.1 Rule-Based Redaction

This initial phase focuses on finding structural and obvious entities of PHI via high-precision regex or dictionary search. Regex expressions for dates (e.g., "01/05/2023" or "January 5th, 2023"), social security numbers, contact details, or medical record numbers are found and replaced with placeholders like `[DATE]` or `[ID]`. Following HIPAA Safe Harbor provisions, ages with a value of 89 or more are transformed into generic expressions like '90+'. Name dictionaries with capitalization-based heuristic techniques are also employed to spot and remove name or initials like 'John D.' to the placeholder `[NAME]`. This phase is very accurate but does not spot hidden or context-based identifiers.

3.3.2 NER-Based PHI Detection

A transformer-based Named Entity Recognition (NER) is utilized in the second stage for the

finer task of medical de-identification inorder to improve recall. Having access to models like BioBERT or ClinicalBERT, it is able to identify the tokens corresponding to the type of PHI: PERSON, HOSPITAL, LOCATION, ID, among others. These tokens are replaced with their respective type placeholders (e.g., MASSACHUSETTS GENERAL HOSPITAL → `[HOSP]`). False positives are decreased by matching the entities with medical vocabularies, such that crucial terms in medicine are maintained. This phase helps in identifying other unique identifiers found in more natural language.

3.3.3 Neural Text Rewriting

Finally, in the last stage, we conduct a fluency-enhancing rewrite via a T5 structure based text-to-text transformer model trained on synthetic data related to the injection/removal of PHI. When fed a censored input sentence, it generates a fluent response with equivalent meaning that does not include any private information. This is achieved through "[NAME] is a 54-year-old female with chest pain since [DATE]" becoming "The patient is a 54-year-old female with chest pain since early 2023." Moreover, it is limited to generate accurate information that is relevant for downstream applications like medical VQA.

This multi-tier approach of precise redactioning, statistical entity recognition, and neural rewriting is utilized for robust and fluent de-identification. This anonymized text (T_{priv}) is privacy-preserving as well as clinically informative for multimodal LLM applications. Implementation details of the NER and rewriting components are described in Appendix C.

3.4 Multimodal LLM Inference and Evaluation

ClinX employs a multimodal large language model, which is visual large language model (VLLM) for downstream inference tasks like VQA, Diagnosis Reasoning, and Case Understanding. At the point of inference, this model takes the pair of input: the de-identified medical image and the obfuscated clinical prompt. This way, the PHI is not revealed at any point during query processing.

To understand the effect of de-identification with respect to the utility of models, both answer correctness and semantic alignment are evaluated. In these experiments, progress is measured using both exact match (EM) and accuracy scores. However, such scores could be detrimental in determining valid paraphrasing or stylistic differences. To overcome such challenges, the study utilises a large language model-as-a-judge (LLM-J) framework proposed by Zheng et al. (Zheng et al., 2023). In this study, a separate LLM, such as GPT-4, is employed to measure semantic alignments between ground truth answers and answers generated by models. This helps the study effectively evaluate the effect of models that could potentially impair clinical reasoning or understanding through privacy-preserving transformations.

ClinX applies OCR-guided image masking and PP-GAN inpainting in parallel with multi-stage text de-identification before multimodal inference; a visual overview of the end-to-end pipeline is provided in Appendix F, and a step-by-step procedural description is given in Appendix G.

4 Experimental Setup

4.1 Hardware and Infrastructure

A high performance GPU cluster with 8 NVIDIA Tesla V100 Tensor Core GPUs (32 GB of VRAM each) with Intel Xeon CPU E5-2698 v4@2.20GHz processors and 512 GB system RAM was used for all experiments.

4.2 Datasets

Our framework were tested on two popular medical VQA datasets: VQA-RAD (`flaviagiammarino/vqa-rad`) containing 1,793 training and 451 test samples, and PathVQA (`flaviagiammarino/path-vqa`) containing approximately 19,700 training, 6,260 validation, and 6,720 test samples. The official *training splits* were used to train the PP-GAN image de-identification model, and the *test splits* were used solely for evaluation of downstream VQA performance. No test images were seen during PP-GAN training.

4.3 PP-GAN Training Setup

To train the PP-GAN model, masked medical image samples were drawn from the training portion of the VQA-RAD and PathVQA datasets, using the OCR-guided masking and heuristic fallback strategy described in Section 3.1. All input images were resized to 256×256 resolution. Training of the PP-GAN employed the Adam optimizer with the standard GAN hyperparameters($\beta_1 = 0.5$, $\beta_2 = 0.999$), a learning rate of 2×10^{-4}, and a batch size of 8 along with least-squares GAN

(LSGAN) objective for improved stability. The generator was optimized using a combination of adversarial loss, masked L_1 reconstruction loss, identity preservation loss for unmasked regions, total variation regularization, and perceptual similarity loss (LPIPS). PP-GAN model was trained for 50 epochs, after which the final generator checkpoint was used for all image de-identification and downstream evaluation experiments.

4.4 Multimodal VQA Model

For visual question answering tasks, we used LLaVA-Med v1.5 (Mistral-7B) through a Hugging Face-compatible checkpoint available at `chaoyinshe/llava-med-v1.5-mistral-7b-hf`. The model is an exact drop-in replacement for the original Microsoft LLaVA-Med model release. Experiments were conducted using both original images and privacy-preserving images generated from our pipeline. The same set of questions and deterministic decoding settings were used for the both original images and texts, and private image and texts generated by our framework, ClinX.

4.5 LLM-as-Judge Evaluation

To check if an answer was semantically correct without relying solely on exact word matches, we followed an LLM-as-Judge evaluation setup. To do this, we used LLaMA-3.1-8B-Instruct to compare model-generated responses with the actual responses. The judge evaluation setup includes a direct instruction to make the LLM function as a harsh medical consultant, prioritizing medical validity over surface level lexical similarity to produce a binary correctness score. The prompt provided to the LLM-judge enforces it to produce structured JSON output to allow the automatic aggregation of the evaluation result. Details of the judge model, prompt design, and deterministic decoding are provided in Appendix E.

4.6 Other Evaluation Metrics

Our other evaluation covers standard VQA metrics such as accuracy and token-level F1 scores. To gain a better insight into performance on these tasks, these measurements were computed for yes/no queries and open-ended queries separately. Our evaluation was conducted on non-private inputs and inputs preserved with privacy to understand the effect of de-identification on accuracy.

5 Results and Analysis

Our framework, ClinX was evaluated on two publicly available medical VQA datasets, VQA-RAD and PathVQA. Although these two datasets do not possess any actual Protected Health Information (PHI), a realistic level of de-identified medical content is simulated using our OCR-based masking and generative inpainting to all of the text-embedded region in the images and text obfuscation technique.

All experiments use model named ‘`chaoyinshe/llava-med-v1.5-mistral-7b-hf`’ for multimodal inference. In addition to standard string-based metrics, we assess semantic correctness using ‘`meta-llama/Llama-3.1-8B-Instruct`‘ as a judgment model. We report: (i) **Strict Accuracy (S)**, requiring an exact normalized match; (ii) **Relaxed Accuracy (R)**, which considers a prediction correct if it either exactly matches the ground truth or achieves a content-token Jaccard overlap above a fixed threshold; and (iii) **LLM-Judge Score (J)**, measuring semantic agreement.

5.1 Main Results and Ablation Study

Table 1 compares three settings: the *Raw Baseline* (original images), an *OCR-only* ablation that masks detected text regions without reconstruction, and the proposed *ClinX* pipeline. Figure 2 complements the tabular results by illustrating the performance degradation under OCR-only masking and the consistent recovery achieved by ClinX across strict, relaxed, and semantic evaluation metrics.

Effect of Generative Inpainting: The OCR-only ablation study showcases a considerable degradation in performance, including the strict accuracy reduction of 10.1% on VQA-RAD, thus indicating that hard occlusion of the text regions with just a black mask disrupts clinically relevant visual information. In contrast to this result, ClinX demonstrates its ability to recover this loss and performs at a level comparable to and in some instances better than the baseline. These findings verify that inpainting and generation algorithms are necessary to maintain multi-modal inference performance while still applying visual de-identification.

5.2 Question-Type Analysis

To evaluate the robustness across different task categories, the results was further categorized into *Yes/No* (closed-ended), and *Open-ended*, as shown in the below table 2. ClinX performs considerably

Table 1: Comparison of Raw Baseline, OCR-only masking, and ClinX. **S**: Strict Accuracy, **R**: Relaxed Accuracy, **J**: LLM-Judge Score.

Method	VQA-RAD S	VQA-RAD R	VQA-RAD J	PathVQA S	PathVQA R	PathVQA J
Raw Baseline	40.35	47.23	47.45	**31.92**	**33.13**	41.94
OCR-only	30.16	30.20	41.69	29.14	29.16	35.72
ClinX (Ours)	**41.24**	**48.11**	**47.45**	31.82	32.99	**42.19**

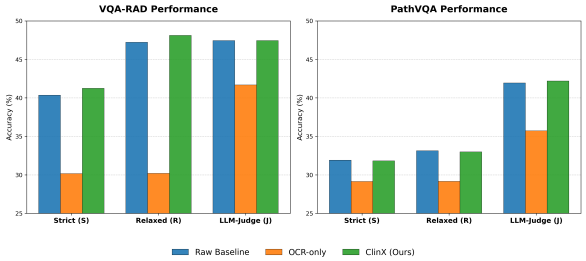


Figure 2: **Performance Comparison across Metrics.** Grouped bar chart comparing the Raw Baseline (Blue), OCR-only ablation (Orange), and our proposed ClinX framework (Green) on VQA-RAD (left) and PathVQA (right).

well with minimal improvement in VQA-RAD. These results can be ascribed to the lack of superficial textual information and cues, which promotes the reliance solely on the anatomical and structural visuals features.

Table 2: Accuracy breakdown by question type (Raw vs. ClinX).

Dataset	Type	Raw (%)	ClinX (%)	Δ
VQA-RAD	Yes/No	54.18	54.98	+0.80
	Other	23.00	24.00	+1.00
PathVQA	Yes/No	59.43	59.22	–0.21
	Other	4.38	4.38	+0.00

5.3 Robustness and Mask Severity Analysis

In order to demonstrate the robustness of ClinX, the results were tested by analyzing performance based on the level of privacy masking (detailed in the appendix in Section D.1). From Table 3, we can see the datasets exhibit distinct distribution characteristics, where VQA-RAD contains significantly heavier text overlays compared to PathVQA.

Nonetheless, ClinX is remarkably stable in spite of this occlusion. Under low-to-moderate levels of masking ($< 30\%$), which represent the majority of the dataset, the performance gap between Raw and de-identified datasets is negligible. Notably, there is a surprising improvement in the semantic accuracy of the model for the tail section in the high-severity category ($> 30\%$ masked), where the semantic accuracy of the model increases for the PathVQA model as shown in Appendix Figure 8.

Table 3: Mask severity statistics. **Avg Mask Area** denotes the average percentage of pixels occluded per image. **High Severity Samples** counts images where $> 30\%$ of the pixels are masked.

Dataset	Avg Mask Area	High Severity Samples
VQA-RAD	12.95%	$N = 36$
PathVQA	5.19%	$N = 24$

Additional analyses on mask coverage and qualitative failure cases are provided in Appendix D.

6 Conclusion

We introduced ClinX, a privacy-aware inference framework to facilitate HIPAA compliance through the de-identification process of medical images as well as related textual content. ClinX is a combination of OCR-based image masking with the SPADE-conditioned PP-GAN inpainting algorithm, along with the mask aware postprocesing in order to remove all of the overlaid identifiers while minimizing the reconstruction artifacts. Apart from privacy protection, another key takeaway from our findings is that removing of the overlaid texts, even if it doesn't explicitly contains the PHI, can help reduce biases in datasets. By removing the surface-level textual information from the images, the ClinX encourages models to look at meaningful visual evidence rather than the overlaid textual information, improving robustness and generalization. Across two VQA medical datasets and various performance evaluation criteria, from strict and relaxed string comparisons to semantic assessments via LLMs, ClinX retains downstream utility within $\sim 1\%$ of the baseline while ensuring effective de-identification. These results imply privacy-preserving reconstruction is not only compatible with multimodal clinical reasoning, it can actually enable more robust and generalizable medical AI.

Limitations

Although ClinX maintains an exemplary level of privacy preservation with negligible utility loss, it has some drawbacks by design. First of all, due to the ethical unavailability of public datasets containing real PHI, the evaluation relies on the simulated masking of all of the textual information that appears on the images. Second, our evaluation focused on 2D static images from VQA-RAD (radiology) and PathVQA (pathology); while these represent the two primary pillars of medical imaging, future work must extend validation to volumetric 3D data and broader multi-domain benchmarks (e.g., SLAKE). Third, despite the stability of our GAN-based inpainting, minute artifacts or non-native phrasing may persist in out-of-distribution samples. Lastly, although we have optimized the text and image obfuscation layer to be as light as possible, but because our approach relies on the use of semi-powerful LLaVA-Med for inference, it leads to increased computational complexity, making it slightly difficult for on-device implementation in limited-resource clinical setups.

Use of AI Assistants: We utilized large language model, particularly ChatGPT-4o, and Gemini, to assist with generating boilerplate code for data processing and for grammatical editing of the manuscript. All of the codes and the manuscript has been manually verified by the authors.

Ethical Considerations

This work aims to further improve privacy preservation for multi-modal medical inference models. The system ClinX is developed to address the risk of revealing confidential patient information by removing and reconstructing textual regions embedded within medical images and blurring possible textual information that could delineate identities. Though the tested datasets VQA-RAD and PathVQA do not involve Protected Heath Information directly, our testing simulates a realistic scenario that explores the tradeoff of privacy with the usefulness of the system in a clinical environment.

ClinX is not meant for replacing clinical expertise or being used as a standalone diagnostic solution. It should be followed by suitable clinical validation, regulatory approval, and human assessment in any further use of privacy-preserving multimodal models. Moreover, though the use of generative inpainting reduces reliance on dataset-specific textual shortcuts, it can generate subtle visual cues, which, when inappropriately used, can lead to misinterpretations.

Lastly, this research aims to promote the ethical sharing of data as well as the development of models by making the learning process privacy-aware without the need to have direct access to identifiable patients information, thereby reducing barriers to responsible medical AI research..

References

Bharadwaja Reddy Chirra. 2023. Enhancing healthcare data security with homomorphic encryption: A case study on electronic health records (ehr) systems. *Revista de Inteligencia Artificial en Medicina*, 14(1):549–59.

Yulan Gao, Ziqiang Ye, Ming Xiao, Yue Xiao, and Dong In Kim. 2024. Guiding iot-based healthcare alert systems with large language models. *arXiv preprint arXiv:2408.13071*.

Sagar Goyal, Eti Rastogi, Sree Prasanna Rajagopal, Dong Yuan, Fen Zhao, Jai Chintagunta, Gautam Naik, and Jeff Ward. 2024. Healai: A healthcare llm for effective medical documentation. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, pages 1167–1168.

Gaurav Kumar Gupta, Aditi Singh, Sijo Valayakkad Manikandan, and Abul Ehtesham. 2024. Digital diagnostics: The potential of large language models in recognizing symptoms of common illnesses. *arXiv preprint arXiv:2405.06712*.

Joschka Haltaufderheide and Robert Ranisch. 2024. The ethics of chatgpt in medicine and healthcare: a systematic review on large language models (llms). *NPJ digital medicine*, 7(1):183.

Håkon Hukkelås, Rudolf Mester, and Frank Lindseth. 2019. Deepprivacy: A generative adversarial network for face anonymization. *CoRR*, abs/1909.04538.

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. 2018. Image-to-image translation with conditional adversarial networks. *Preprint*, arXiv:1611.07004.

Alistair E W Johnson, Lucas Bulgarelli, and Tom J Pollard. 2020. Deidentification of free-text medical records using pre-trained bidirectional transformers. In *Proceedings of the ACM Conference on Health, Inference, and Learning*, New York, NY, USA. ACM.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

A Vijaya Kumar, Mogalapalli Sai Sujith, Kosuri Tarun Sai, Galla Rajesh, and Devulapalli Jagannadha Sriram Yashwanth. 2020. Secure multiparty computation enabled e-healthcare system with homomorphic encryption. In *IOP Conference Series: Materials Science and Engineering*, volume 981, page 022079. IOP Publishing.

Mehdi Letafati and Safa Otoum. 2023. Global differential privacy for distributed metaverse healthcare systems. In *2023 International Conference on Intelligent Metaverse Technologies & Applications (iMETA)*, pages 01–08. IEEE.

Shuyue Stella Li, Vidhisha Balachandran, Shangbin Feng, Jonathan Ilgen, Emma Pierson, Pang Wei Koh, and Yulia Tsvetkov. 2024a. Mediq: Question-asking llms for adaptive and reliable medical reasoning. *arXiv preprint arXiv:2406.00922*.

Xingyu Li, Lu Peng, Yuping Wang, and Weihua Zhang. 2024b. Open challenges and opportunities in federated foundation models towards biomedical healthcare. *arXiv preprint arXiv:2405.06784*.

Zihan Liu, Feijun Jiang, Yuxiang Hu, Chen Shi, and Pascale Fung. 2021. NER-BERT: A pre-trained model for low-resource entity tagging. *CoRR*, abs/2112.00405.

Khalid Nassiri and Moulay A Akhloufi. 2024. Recent advances in large language models for healthcare. *BioMedInformatics*, 4(2):1097–1143.

Zabir Al Nazi and Wei Peng. 2024. Large language models in healthcare and medical domain: A review. In *Informatics*, volume 11, page 57. MDPI.

Jesutofunmi A Omiye, Haiwen Gui, Shawheen J Rezaei, James Zou, and Roxana Daneshjou. 2024. Large language models in medicine: the potentials and pitfalls: a narrative review. *Annals of Internal Medicine*, 177(2):210–220.

Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. 2019. Semantic image synthesis with spatially-adaptive normalization. *Preprint*, arXiv:1903.07291.

Le Peng, Gaoxiang Luo, Sicheng Zhou, Jiandong Chen, Ziyue Xu, Rui Zhang, and Ju Sun. 2023. An in-depth evaluation of federated learning on biomedical natural language processing. *medRxiv*, pages 2023–11.

Kummaragunta Joel Prabhod. 2023. Integrating large language models for enhanced clinical decision support systems in modern healthcare. *Journal of Machine Learning for Healthcare Decision Support*, 3(1):18–62.

Jili Qian, Zhengyu Jin, Quan Zhang, Guoqing Cai, and Beichang Liu. 2024. A liver cancer question-answering system based on next-generation intelligence and the large model med-palm 2. *International Journal of Computer Science and Information Technology*, 2(1):28–35.

Mikail Mohammed Salim, Xianjun Deng, and Jong Hyuk Park. 2024. A privacy-preserving local differential privacy-based federated learning model to secure llm from adversarial attacks. *Human-centric Computing and Information Sciences*, 14(57).

Yiping Song, Juhua Zhang, Zhiliang Tian, Yuxin Yang, Minlie Huang, and Dongsheng Li. 2024. Llm-based privacy data augmentation guided by knowledge distillation with a distribution tutor for medical text classification. *arXiv preprint arXiv:2402.16515*.

Xiangru Tang, Anni Zou, Zhuosheng Zhang, Ziming Li, Yilun Zhao, Xingyao Zhang, Arman Cohan, and Mark Gerstein. 2023. Medagents: Large language models as collaborators for zero-shot medical reasoning. *arXiv preprint arXiv:2311.10537*.

Tuan Truong, Ivo M Baltruschat, Mark Klemens, Grit Werner, and Matthias Lenga. 2025. Exploring AI-based system design for pixel-level protected health information detection in medical images. *J. Imaging Inform. Med.*

Junda Wang, Zhichao Yang, Zonghai Yao, and Hong Yu. 2024a. Jmlr: Joint medical llm and retrieval training for enhancing reasoning and professional question answering capability. *arXiv preprint arXiv:2402.17887*.

Xidong Wang, Nuo Chen, Junyin Chen, Yan Hu, Yidong Wang, Xiangbo Wu, Anningzhe Gao, Xiang Wan, Haizhou Li, and Benyou Wang. 2024b. Apollo: Lightweight multilingual medical llms towards democratizing medical ai to 6b people. *arXiv preprint arXiv:2403.03640*.

Jacob White. 2020. Pubmed 2.0. *Medical reference services quarterly*, 39(4):382–387.

Yusuke Yamauchi, Taro Yano, and Masafumi Oyamada. 2025. An empirical study of llm-as-a-judge: How design choices impact evaluation reliability. *Preprint*, arXiv:2506.13639.

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. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

Alexander Ziller, Dmitrii Usynin, Rickmer Braren, Marcus Makowski, Daniel Rueckert, and Georgios Kaissis. 2021. Medical imaging deep learning with differential privacy. *Scientific Reports*, 11(1):13524.

Appendix Contents

A PP-GAN Architecture and Design **12**
A.1 Generator Architecture 12
A.1.1 SPADE Conditioning Mechanism 12
A.2 PatchGAN Discriminator Architecture 13
A.3 Design Summary and Consistency Notes 13

B PP-GAN Training and Inference Workflows **14**
B.1 Adversarial Training Loop 14
B.2 Training Dynamics and Convergence 14
B.3 Inference Workflow 14

C Textual Privacy Implementation Details **14**
C.1 NER Architecture (Token-Level PHI Detection) 15
C.2 Text Rewriting and Paraphrasing . 15
C.3 Deployment Behavior and Design Rationale 16

D Extended Evaluation and Robustness Analysis **16**
D.1 Accuracy under Increasing Mask Severity 16
D.2 Dataset Bias and Mask Distribution Effects 17
D.3 Alignment between Automatic Metrics and LLM-as-Judge 17
D.4 Failure Modes and Qualitative Observations 18
D.5 Summary of Appendix Findings . 18

E LLM-as-Judge Evaluation **18**
E.1 Judge Model and Evaluation Setup 18
E.2 Judge Prompt 18
E.3 Output Parsing and Robustness . . 19
E.4 Why LLM-as-Judge is Used . . . 19

F Full System Inference and Evaluation Pipeline **19**

G Algorithmic Details **19**

A PP-GAN Architecture and Design

This appendix supplies additional architectural and design explanation for the image de-identification component PP-GAN that is part of the proposed system ClinX, and aims at explaining some of the critical details that are not mentioned in the main paper but have an important impact on privacy and the multimodal reasoning task.

A.1 Generator Architecture

The generator in the PP-GAN model has an encoder-decoder structure without skip connections, focusing on avoiding any direct pixel copying of protected health information, as illustrated in figure 3. In contrast to U-Net models, this structure bypasses any addition of encoder information to the decoder, which forces the model to reconstruct the masked regions through learned synthesis instead of copying features.

The encoder contains a series of convolutional layers which downsample the image progressively in order to capture the global anatomical context. It contains a SPADE conditioned residual block at bottle neck, which adds spatial guidance for the soft PHI mask. The decoder upsamples the latent spatial representation back to its original image size by utilizing transposed convolutions. The final SPADE-based refinement block applies mask-conditioned modulation in order to produce the RGB output.

The mask conditioning is done using a SPADE-like normalization module with a internal resizing of the soft PHI mask to match the spatial resolutions of the feature maps. The final activation function is hyperbolic tangent to ensure the values fall within the range of $[-1, 1]$. Table 4 gives the architecture of the generator network employed in the PP-GAN.

A.1.1 SPADE Conditioning Mechanism

The generator consists of a normalization module that is like SPADE inorder to condition feature activations on the soft PHI mask. For the input feature tensor x and the mask m, SPADE Normalization essentially applies parameter-free batch normalization to x and then uses the mask m to compute the values of the scaling and shifting parameters. The mask is resized by bilinear interpolation to the same size as the feature map before every SPADE application. This enables spatially localized modulation while maintaining a single global mask representation. SPADE is used only at: (1) The bottleneck residual block, which allows global context modulation, and (2) The final refinement block, which does mask-aware synthesis on the output. Adding this block helps make the model more spatially aware. This design balances spatial control with

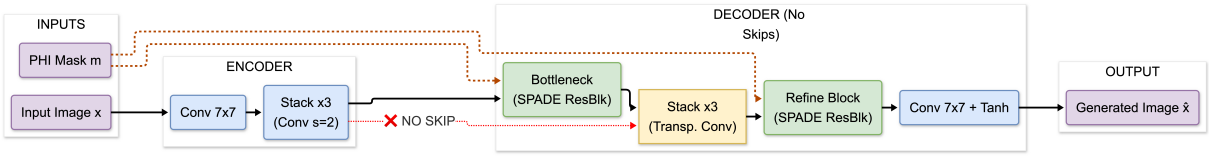


Figure 3: PP-GAN generator overview. The generator uses an encoder–decoder design with no skip connections for privacy. SPADE conditioning uses a multi-scale resized mask to localize synthesis to masked regions.

Stage	Operation	Channels	Notes
Encoder-1	Conv 7×7 + ReLU	64	Initial feature extraction
Encoder-2	Conv 4×4, s=2 + ReLU	128	Downsampling
Encoder-3	Conv 4×4, s=2 + ReLU	256	Downsampling
Encoder-4	Conv 4×4, s=2 + ReLU	512	Downsampling
Bottleneck	SPADE ResNet Block	512	Mask-conditioned context aggregation
Decoder-1	Transposed Conv 4×4, s=2 + ReLU	256	Upsampling
Decoder-2	Transposed Conv 4×4, s=2 + ReLU	128	Upsampling
Decoder-3	Transposed Conv 4×4, s=2 + ReLU	64	Upsampling
Refinement	SPADE ResNet Block	64	Mask-conditioned refinement
Output	Conv 7×7 + Tanh	3	RGB output image

Table 4: PP-GAN generator architecture aligned with the released implementation. SPADE conditioning is applied at the bottleneck and final refinement stages only. No skip connections are used.

computational efficiency and training stability.

A.2 PatchGAN Discriminator Architecture

The discriminator in PP-GAN is based on a lightweight PatchGAN architecture, which instead of rating the realism of an image as a whole, rates it for small patches of the image, as is shown in Figure 4. The discriminator takes an input image of either x (real) and $\hat{x}$ (generated) and uses a series of strided convolution layers to generate a spatial map for the realism score.

In particular, the discriminator architecture comprises three convolution layers with 4×4 kernels and stride 2, and gradually enhances the channels from 64 to 256. After each convolution operation, a Leaky ReLU activation function with a slope of 0.2 is appended. A final 4×4 convolution with stride 1 is further utilized to obtain a single-channel feature map with raw realism logits.

Note that, compared to its deeper PatchGAN counterparts, this discriminator does not have an intermediate 512-channel convolutional layer. Thus, each element in its outputs represents a receptive field of about 46×46 pixels within the image. This design provides localized adversarial supervision while remaining computationally efficient.

There is no sigmoid activation function in the output layer since the discriminator in this case relies on least-squares GAN (LSGAN) loss/objective.

A.3 Design Summary and Consistency Notes

The PP-GAN model used in the proposed system, ClinX, is carefully designed to address all issues regarding privacy preservation, realism, and computational complexity. The Generator PPGAN follows an encoder-decoder framework, sans skip connections, eliminating direct pixel-level leakage of PHI. Mask conditioning is achieved through SPADE inspired normalization modules, acting on the bottleneck and refinement modules, enabling spatially aware synthesis tasks that utilize a soft PHI mask.

The discriminator employs a concise PatchGAN architecture to impose a realistic detail level on local image patches. Although the proposed discriminator is not as deep as a standard 70×70 PatchGAN architecture, it is still able to provide a local adversarial detail cue that is adequate for the image inpainting task.

Importantly, the discriminator does not receive the PHI mask as a explicit input. Since the unmasked region is the same for real and generated images, the adversarial training method automatically emphasizes the masked region. Every description of architecture, table, and image in the appendix follows the implementation as it is released.

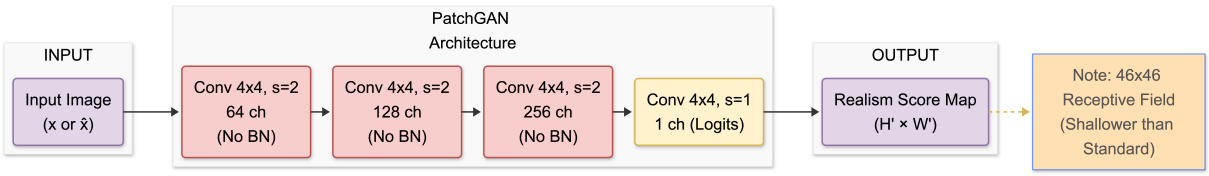


Figure 4: PatchGAN discriminator architecture used in PP-GAN. A sequence of 4×4 convolutions produces a spatial patch-logit map (approximately 70×70 receptive field per output cell). No sigmoid is used under LSGAN.

Layer	Kernel / Stride	Channels
Conv-1	4×4 / 2	64 + LeakyReLU
Conv-2	4×4 / 2	128 + LeakyReLU
Conv-3	4×4 / 2	256 + LeakyReLU
Output	4×4 / 1	1 (patch logits)

Table 5: Patch discriminator architecture aligned with our implementation. The output is a spatial patch-level score map (no sigmoid; used with LSGAN).

B PP-GAN Training and Inference Workflows

This appendix section details the training/optimization flow, objective function calculation and the deployment strategies used for the PP-GAN framework.

B.1 Adversarial Training Loop

The training process of the model can be seen in Figure 5. In this adversarial process or cycle, the Discriminator (D) and the Generator (G) are trained and updated in a synchronized and sequential way. In fact, the Discriminator completes two different forward passes in every training iteration, one for real samples (x) and other for generated samples ($\hat{x}$), inorder to compute the LSGAN objective.

B.2 Training Dynamics and Convergence

To optimize storage and efficiency, we saved only the final converged generator weights (`g.pt`). Nonetheless, stability during training was monitored through **high-frequency visual sampling** with a period of 25 iterations.

In Fig. 6, we can see the fast convergence of the PP-GAN generator which is described as follow:

- **Early Iterations (Epoch 1):** The generator is in the stochastic initialization stage. Although SPADE Normalization Layer starts to enforce the global spatial constraints imposed by the PHI mask immediately, there is no structured content inside the generated image adn it lacks semantic coherence.
- **Mid-Training (Epoch 2):** The adversarial loss compels the generator to resemble the distribution of the target domain in terms of color (H&E stain). Structural boundaries emerge clearly, with global geometry being learned by our model. However, the microscopic textures are seen to be blurred or smoothed.
- ***Final Convergence (Epoch 5):*** The refinement network/block performs the synthesis step on the high-frequency details successfully. The cell nuclei, stroma, and complex textures are fully reconstructed, and the inpainted areas are imperceptible compared to the ground truth.

B.3 Inference Workflow

When deploying for clinical use, we opt to discard the states of the Discriminator and the optimizer while retaining only the final weights generated for the generators (`g.pt`). The inference pipeline that produces privacy-preserved output from raw medical images using 3 stage process is shown below in Figure 7.

C Textual Privacy Implementation Details

To facilitate the underlying conceptual framework introduced in Subsection 3.3, this appendix section documents the implementation nuances of the PHI sanitizing entities related to text that are incorporated into the ClinX framework. These entities are part of the overall system. However, due to the curated nature of the standard medical VQA benchmarks, the textual input consist of little to no explicit PHI. We describe the essential design

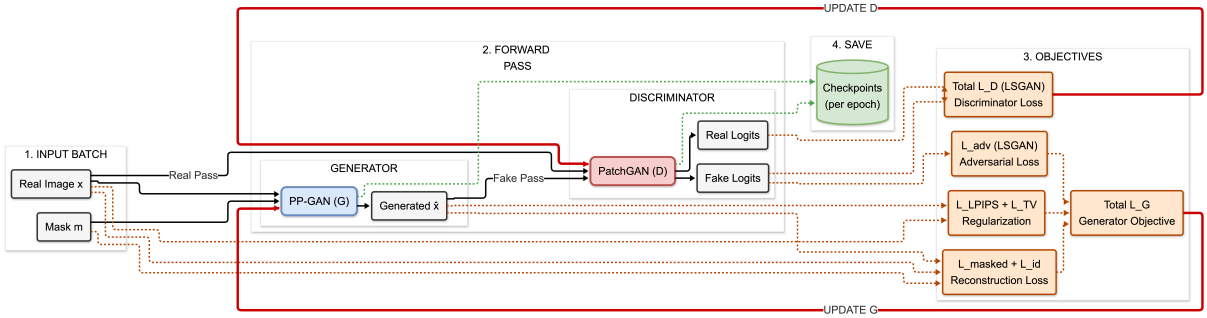
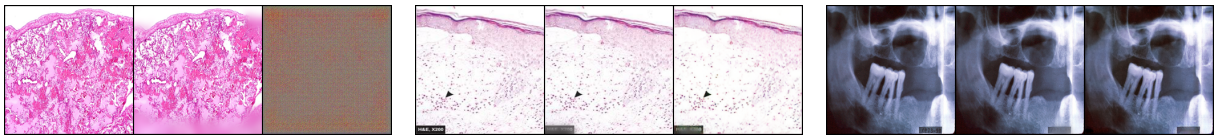


Figure 5: **Complete PP-GAN training loop.** The generator (G) synthesizes de-identified images ($\hat{x}$) from real inputs (x) and PHI masks (m). Solid arrows indicate tensor flow, while dotted arrows denote objective dependencies. The masked reconstruction loss is applied strictly within PHI regions, whereas the identity loss ensures content preservation in unmasked regions. Models are optimized via backpropagation (solid red arrows), and checkpoints are saved per epoch.



Epoch 1 Epoch 2 Epoch 5 (Final)

Figure 6: **Training progression of the PP-GAN Inpainter.** Each panel displays a triplet: **(Left) Ground Truth** real image; **(Middle) Masked Input** seen by the network; **(Right) Generated Output**. At Epoch 1, the output is stochastic noise (Right). By Epoch 2, structural features emerge, and by Epoch 5, the generator successfully reconstructs the masked regions to match the Ground Truth.

here which has been omitted in the main section, for completeness and to demonstrate real-world applicability.

C.1 NER Architecture (Token-Level PHI Detection)

ClinX contains a token-level Named Entity Recognition (NER) subsystem for the stochastic identification of PHI spans/region in texts. The NER subsystem operates on a standard token classification formulation, mapping an input tokens sequence:

$$X = \{x_1, x_2, \ldots, x_n\} \tag{7}$$

to a sequence of contextual representations

$$H = \{h_1, h_2, \ldots, h_n\}, \quad h_i \in \mathbb{R}^{768}. \tag{8}$$

Each token representation is projected to a discrete PHI label space $\mathcal{Y}$ (e.g., `NAME`, `DATE`, `ID`, `O`) via a linear classification head:

$$P(y_i \mid X) = \text{softmax}(W_{\text{cls}} h_i + b_{\text{cls}}), \quad W_{\text{cls}} \in \mathbb{R}^{|\mathcal{Y}| \times 768}. \tag{9}$$

The tokens labeled to the PHI classes are then grouped into a contiguous range and replaced by typed substitutes (for example, `[NAME]`, `[DATE]`). Notice that in the VQA datasets used for the experiment assessment, the question string is usually brief and specific to the task at hand (for example, "Is there pneumonia?" and so on), and no PHI was identified by the NER component. This means that the original input is left intact.

C.2 Text Rewriting and Paraphrasing

After the PHI masking, the model contains another phase that involves rewriting the text with the intention of restoring its fluency by eliminating any remaining stylistic clues when redactions take place. This phase works with the cleaned text that uses placeholder tokens and reformulates the question while preserving its clinical intent.

Conceptually, this is represented by a conditional transformation:

- **Input:** `[NAME] presented with [SYMPTOM] on [DATE].`
- **Output:** `The patient presented with [SYMPTOM] in late 2023.`

In practice, because explicit PHI is rarely expressed within the text of the benchmark question, there was no redaction-induced fragmentation to

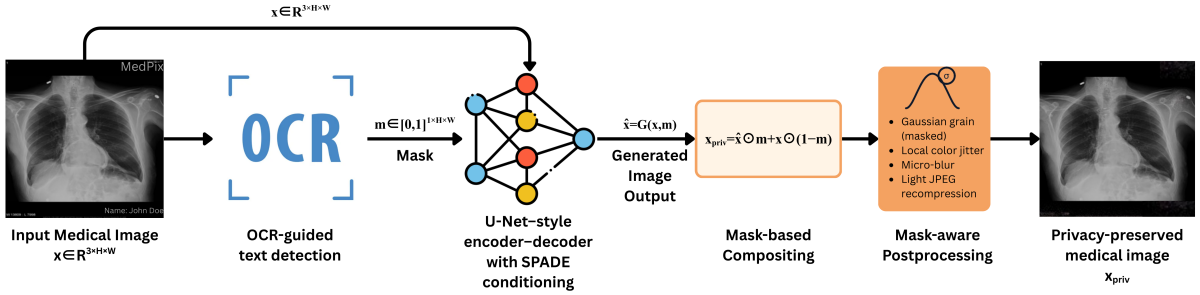


Figure 7: **End-to-End Inference Pipeline.** The deployment workflow automates privacy preservation by cascading three stages: 1) **Detection:** An OCR module identifies sensitive text to generate the binary mask m. 2) **Synthesis:** The pre-trained PP-GAN generator predicts the de-identified content $\hat{x}$. 3) **Compositing:** The prediction is merged with the original image and post-processed to ensure consistent sensor noise characteristics.

trigger the rewriting step. As such, this ensured that there was no alteration to the text of the question asked in all experiments reported.

C.3 Deployment Behavior and Design Rationale

The textual privacy module is designed to be "always on" in the ClinX processing stream, though it is oblivious to PHI when the sensetive information is absent. This is formalized by the definition of the set of PHI spans $\mathcal{D}(X)$, which is the result of the detectoion of the PHI words by NER module on the input sequence X: The sanitization operator $S(X)$ is a conditional function:

$$S(X) = \begin{cases} \text{T5}_\phi(\text{Mask}_{\mathcal{D}(X)}(X)) & \text{if } \mathcal{D}(X) \neq \emptyset \\ X & \text{if } \mathcal{D}(X) = \emptyset \end{cases} \tag{10}$$

where $\text{Mask}_{\mathcal{D}(X)}$ represents the replacement of detected spans with placeholders, and T5_ϕ represents the generative rewriting step.

In the overwhelming majority of cases for evaluation samples for both VQA-RAD and PathVQA, the expression $\mathcal{D}(X) = \emptyset$ is true, and the system collapses to the identity function $S(X) = X$. This architecture choice is made for a reason: ensuring privacy preservation is mandatory in practical settings, where user inquiries might potentially leak some identifiers, while the performance on the task should not be affected in the standard evaluation settings.

D Extended Evaluation and Robustness Analysis

In this appendix, we will include some additional quantitative and qualitative assessments to support the robustness claims discussed in the main paper. Specifically, we will investigate performance characteristics when varying levels of image obfuscation are considered, study dataset-specific phenomena, as well as provide additional metrics to conclude the results obtained.

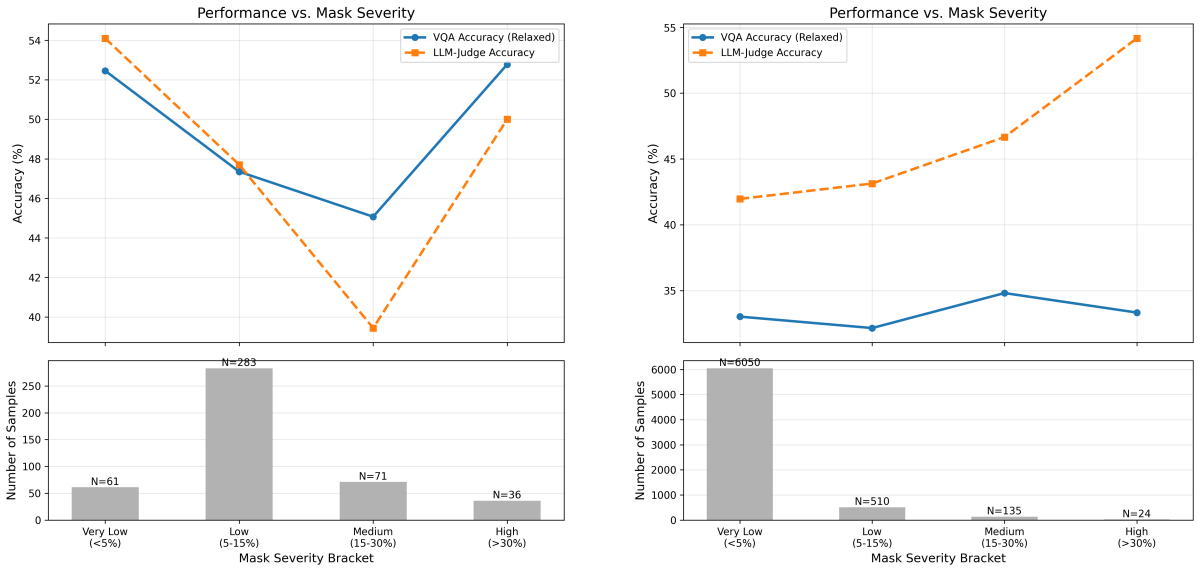
D.1 Accuracy under Increasing Mask Severity

Figure 8 illustrates the relaxed accuracy of the VQA task, along with LLM-as-judge accuracy stratified by PHI mask severity over both evaluation sets, with mask severity measured by the percentage of image pixels overlaid by PHI masks created by the OCR engine before PP-GAN inpainting, including the number of samples (N) per severity bin.

Observed trends. In both datasets, the accuracy holds constant in the low/moderate mask settings (below 15%), which in fact represents the most common scenario. This suggests that PHI removal does indeed preserve enough semantics in the inpainted regions through the PP-GAN process for the following VQA tasks.

On PathVQA (Fig. 8b), the standard relaxed accuracy (blue line) remains stable across mask regimes, fluctuating near the baseline. Notably, however, the LLM-Judge accuracy (orange dashed line) diverges in the high-severity tail ($> 30\%$), rising significantly to $\sim 54\%$. This suggests that while exact lexical matches do not improve, the model's generated responses maintain high clinical validity according to the semantic judge, even when substantial visual context is occluded. However, given the extreme sparsity of the high-mask tail ($N = 24$), this apparent performance gain should be interpreted with caution.

In contrast, there is larger variance in the high-



(a) VQA-RAD **(b) PathVQA**

Figure 8: **Relaxed VQA accuracy and LLM-Judge scores as a function of PHI mask severity.** Bars (bottom) indicate sample counts per severity bin. (a) **VQA-RAD** exhibits higher variance in high-severity bins due to limited sample size ($N = 36$). (b) **PathVQA** demonstrates stable relaxed accuracy (blue) across mask regimes, while LLM-Judge scores (orange) diverge positively in the high-mask tail, suggesting preserved semantic validity despite occlusion.

severity bin in VQA-RAD (Fig. 8a). This seeming gain in performance may not really genuine, especially when one considers that there are only $N = 36$ examples in that bin. In these low-sample regimes, mean accuracy metrics become extremely case-dependent, especially when one considers that yes/no questions in VQA may be answered from relatively limited visual cues.

Implications for clinical utility. Importantly, the relative gap between the accuracy of the RAW and privacy-preserved (PRIV) inputs remains small in the low and moderate mask bins for both datasets. This suggests that performance differences are primarily driven by information removal due to masking, rather than artifacts introduced by the PP-GAN generator during the masking process. From a practical standpoint, this supports the use of ClinX as a preprocessing layer for multimodal clinical inference under realistic PHI distributions.

D.2 Dataset Bias and Mask Distribution Effects

The two data sets have quite disparate distribution characteristics for mask severity. PathVQA has a strong propensity for small masks ($< 5\%$) to denote localized PHI like dates and specimens. VQA-RAD has more moderate masks (5–15%) with overlaid headers and acquisition data, which is quite common in radiographic images.

This difference in the distribution of values across these two groups explains both the increased stability seen in PathVQA and the larger variance in the tail bins in VQA-RAD. This, in turn, emphasizes the value of reporting sample count with accuracy statistics to assess the robustness to extreme masking conditions.

D.3 Alignment between Automatic Metrics and LLM-as-Judge

To complement relaxed accuracy evaluation, we employ an LLM-as-Judge to assess semantic correctness of model predictions, with the ground truth. Table 6 reports overall judge accuracy and agreement between judge decisions and relaxed automatic metrics.

Agreement analysis: We notice that there is a strong correlation between LLM-judge and relaxed accuracy on both datasets, suggesting that relaxed accuracy is a good proxy for semantic validity. The cases where there are discrepancies are those involving answered questions that are paraphrased and partially correct, which are marked incorrect under strict matching but are valid answers according to the judge.

This alignment of functions lends credence to the usage of relaxed accuracy throughout the paper and affirms that privacy-preserving transformations have little effect on semantic results.

D.4 Failure Modes and Qualitative Observations

We conducted manual analysis of the failure examples to find the common error patterns. Three error patterns were found:

- **Extreme occlusion:** Large masks cover the diagnostically important areas of the image and cause certain open-ended questions to have no possible answer despite of the inpainting quality.
- **OCR miss-detections:** Rare OCR failures can leave residual text artifacts, leading to unusual, noisy and inconsistent semantic cues.
- **Ambiguous questions:** Certain questions may have multiple plausible answers depending on the incompletely represented visual data, resulting in a disagreement between the auto metrics and the LLM judge.

Although this model has these shortcomings, it is worth noting that systematic hallucinations or unstable behaviors were not found in the results that preserved privacy. The errors were more likely due to information loss rather than artifacts introduced by the de-identification pipeline.

D.5 Summary of Appendix Findings

These extended analyses endorse the main results of this paper: (i) ClinX achieves constant downstream functionality even in realistic PHI masking scenarios, (ii) The degradation with extreme masking is gradual and predictable, and (iii) automatic relaxed metrics show very strong correspondence with LLM-based semantic assessments.

Combining these findings offers further proof that the integration of privacy preserving image and text obfuscation techniques into multimodal medical VQA pipelines does not have to negatively impact their utility.

E LLM-as-Judge Evaluation

Automatic metrics evaluating exact matches might overlook correctness in medical VQA because valid answers may involve the use of phrases or synonyms (e.g., "no pleural effusion" vs. "absent effusion"). In addition to using the metric evaluating matches involving relaxed substring matching, we test the semantic correctness of answers by using an LLM-as-Judge protocol.

E.1 Judge Model and Evaluation Setup

To apply our LLM judge, we employ a HuggingFace chat model through the library `transformers`. Unless otherwise explicitly stated,, we employ `meta-llama/Meta-Llama-3.1-8B-Instruct` acting as our LLM judge. This judge function serves on both the RAW results and the privacy-preserving (PRIV) results from our ClinX framework.

The judge scripts accept per-sample JSONL outputs (the same schema as our LLaVA-med evaluation logs). For each sample, we supply the tuple: question Q, ground-truth answer GT, and model prediction A. The judge responds with a:

$$\begin{aligned}&\{\texttt{"score"}\colon\ s \in \{0,1\},\\ &\ \texttt{"reason"}\colon\ \texttt{"<short explanation>"}\}\end{aligned} \tag{11}$$

and judge accuracy is computed as the mean of s across all evaluated samples. When there is access to both RAW and PRIV outputs for the same sample ID, we can calculate the judge accuracies on the corresponding subset, to ensure fair comparision.

Deterministic inference. To mitigate the noise in the evaluation, we evaluate the judge with greedy decoding (`do_sample=False`, `temperature=0.0`) and with the output size fixed to 128 (`max_new_tokens=128`). This will give us stable, reproducible binary judgement for same input triple.

E.2 Judge Prompt

We apply a rigid judging rubric that focuses on medical accuracy, permitting slight variations in wording. The final judge prompt involves a system instruction and a user message that includes the (Q, GT, A) tuple. Then, the judge must print just one JSON object.

System message.

> You are a strict medical VQA evaluator. Always respond ONLY with a single JSON object.

User message template.

> Given the following medical visual question answering triple:
>
> Q (Question): {question}
> GT (Ground Truth short answer): {gt}
> A (Model answer): {answer}

Table 6: **Alignment between Automatic Metrics and LLM-as-Judge.** Comparison of Relaxed VQA Accuracy and LLM-Judge Utility Scores. **Agreement** indicates the percentage of samples where the relaxed automatic metric and the LLM judge reached the same verdict (Correct/Incorrect). High agreement (>90% for radiology) validates the use of relaxed accuracy as a reliable proxy.

Dataset	Condition	Relaxed Acc (%)	LLM-Judge Acc (%)	Agreement (%)
VQA-RAD	RAW	47.23	47.45	91.3
	PRIV	48.11	47.45	90.6
PathVQA	RAW	33.13	41.94	82.1†
	PRIV	32.99	42.19	81.4†

†Lower agreement in PathVQA reflects the open-ended nature of pathology QA, where the Judge accepts valid synonyms rejected by the parser.

> Goal: - Decide if A is clinically and semantically consistent with GT. - Ignore minor wording or stylistic differences. - If A is mostly correct and captures the clinical meaning of GT, score 1. - If A is incorrect, contradicts GT, or misses the main point, score 0.
>
> Important: - Do NOT be overly lenient. - Focus on medical correctness and whether the key concept in GT is correctly conveyed.
>
> Now respond ONLY with a single JSON object, no extra text. Example: {"score": 1, "reason": "Matches the diagnosis and key finding."}

E.3 Output Parsing and Robustness

The judge output is then parsed by extracting the first JSON-like substring from the generated text, which is then decoded with `json.loads`. In case of a parsing error or missing valid judge outputs with a JSON object, the script will then set `score=0` and store the raw text in the `reason` field, tagged as a parse error. This conservative policy prevents malformed judge outputs from inflating reported performance.

For transparency and error analysis, we log per-sample fields including the judge decision (`judge_raw_score`, `judge_priv_score`), the explanation (`judge_raw_reason`, `judge_priv_reason`), and the full raw generation (`judge_raw_raw_output`, `judge_priv_raw_output`).

E.4 Why LLM-as-Judge is Used

The assessment of LLM-as-Judge has two functions: (i) it offers a semantic correctness check where the answers are lexically different but clinically equivalent, and (ii) it helps validate that privacy-preserving transformations do not introduce semantic drift in downstream VQA outputs. In our experiments, the accuracy of scores assigned by judges follows the relaxed accuracy quite closely, with most disputes taking place in borderline areas (partial answers, underspecified truth, clinical synonyms).

F Full System Inference and Evaluation Pipeline

Figure 9 shows the full end-to-end process for the ClinX system. In the pipeline, the ClinX system includes the image de-identification module using the PP-GAN technique introduced in Appendix B, the text obfuscation component introduced in Appendix C, followed by the multimodal model named LLaVA-Med for the clinical reasoning task, where the output is evaluated by the evaluation framework which named as LLM-as-Judge.

G Algorithmic Details

This section provides a concise procedural description of the ClinX inference pipeline for completeness and reproducibility. The Algorithm1 summarizes the deterministic execution order of image de-identification, text obfuscation, and multi-modal reasoning described in Section 3 and Appendix F. All components correspond directly to modules introduced in the main paper and do not introduce additional modeling assumptions.

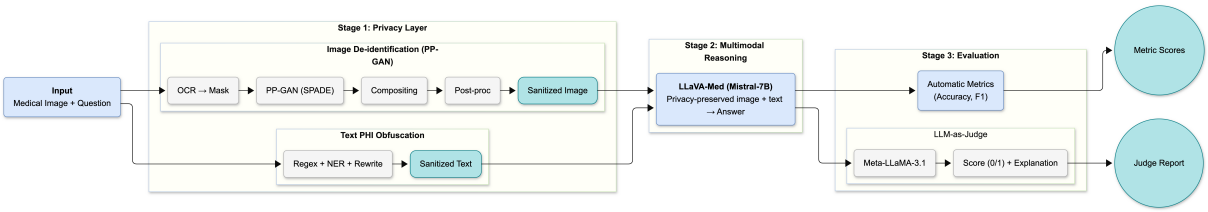


Figure 9: End-to-end inference pipeline for ClinX, illustrating the integration of the Privacy Layer (Stage 1), Multimodal Reasoning (Stage 2), and Evaluation (Stage 3).

Algorithm 1 ClinX: Privacy-Aware Multimodal Inference Pipeline

Input: Medical image x, clinical text T

Output: De-identified inputs $(x_{\text{priv}}, T_{\text{priv}})$ and answer a

Image De-Identification

Extract PHI mask m from OCR on x

Generate inpainted image: $\hat{x} \leftarrow G(x, m)$ {SPADE-based PP-GAN}

Composite result: $x_{\text{comp}} \leftarrow \hat{x} \odot m + x \odot (1 - m)$

Apply mask-aware postprocessing: $x_{\text{priv}} \leftarrow \text{PostProcess}(x_{\text{comp}}, m)$

Text De-Identification

Apply regex-based redaction on $T \rightarrow T_1$

Perform NER-based PHI masking: $T_1 \rightarrow T_2$

Rewrite fluently with neural model: $T_{\text{priv}} \leftarrow \text{Rewrite}(T_2)$

Multimodal Inference

Query LMM with de-identified inputs: $a \leftarrow \text{LLM}(x_{\text{priv}}, T_{\text{priv}})$

return $(x_{\text{priv}}, T_{\text{priv}}, a)$