Pathology-Aware Adaptive Watermarking for Text-Driven Medical Image Synthesis

Chanyoung Kim*[1] Dayun Ju*[1] Jinyeong Kim[1] Woojung Han[1]
Roberto Alcover-Couso[2] Seong Jae Hwang†[1]

[1] Department of Artificial Intelligence, Yonsei University, Republic of Korea
[2] Amazon, Madrid, Spain‡
{chanyoung, juda0707, jinyeong1324, dnwjddl, seongjae}@yonsei.ac.kr,
ralcover@amazon.com

Abstract. As recent text-conditioned diffusion models have enabled the generation of high-quality images, concerns over their potential misuse have also grown. This issue is critical in the medical domain, where text-conditioned generated medical images could enable insurance fraud or falsified records, highlighting the urgent need for reliable safeguards against unethical use. While watermarking techniques have emerged as a promising solution in general image domains, their direct application to medical imaging presents significant challenges. A key challenge is preserving fine-grained disease manifestations, as even minor distortions from a watermark may lead to clinical misinterpretation, which compromises diagnostic integrity. To overcome this gap, we present MedSign, a deep learning-based watermarking framework specifically designed for text-to-medical image synthesis, which preserves pathologically significant regions by adaptively adjusting watermark strength. Specifically, we generate a pathology localization map using cross-attention between medical text tokens and the diffusion denoising network, aggregating token-wise attention across layers, heads, and time steps. Leveraging this map, we optimize the LDM decoder to incorporate watermarking during image synthesis, ensuring cohesive integration while minimizing interference in diagnostically critical regions. Experimental results show that our MedSign preserves diagnostic integrity while ensuring watermark robustness, achieving state-of-the-art performance in image quality and detection accuracy on MIMIC-CXR and OIA-ODIR datasets.

Keywords: Image Watermarking · Diffusion Models · Cross Attention

1 Introduction

Recent advancements in diffusion probabilistic models [11, 12, 22, 24] have enabled the generation of highly realistic synthetic images, benefiting applications such as content creation [23] and data augmentation [29]. However, as generative models become more powerful, concerns over their potential misuse have grown,

* Equal contribution † Corresponding Author ‡ Work done prior to joining Amazon

leading to the introduction of regulatory measures such as the EU AI Act [7], Korean AI Basic Act [21], and Chinese AI governance rules [28].

A widely accepted solution to mitigate such risks is watermarking, which embeds imperceptible yet robust signals into generated content. Traditional signal-based methods [1] are fragile to transformations, prompting the rise of deep learning-based approaches [8, 13, 26, 33] that embed in image or feature space. Recent semantic-aware strategies [3, 31] enhance durability by manipulating latent features, though they may affect image structure. As a result, general-purpose deep watermarking methods continue to gain traction for balancing robustness and fidelity.

This need becomes even more critical in the medical domain, where text-driven image generation [2, 10, 16, 19, 32] enables the synthesis of highly realistic images from clinical descriptions. While effective, such capabilities raise concerns, as misuse could directly impact clinical decisions. Any manipulation of medical images can compromise patient care, violate confidentiality, and lead to serious real-world consequences [20]. For instance, fabricated medical images could be exploited for insurance fraud by falsely indicating the presence of a disease. Moreover, synthetic medical images could be altered to create misleading medical records, falsely attributing medical conditions to individuals, potentially leading to legal ramifications and a breach of public trust. Hence, there is an urgent need for watermarking methods that can ensure the authenticity and integrity of *text-driven synthetic medical images*, safeguarding against potential misuse.

However, unlike general images, medical images exhibit distinct characteristics that make watermarking particularly challenging. A critical requirement is *the preservation of fine-grained anatomical details* (e.g., detailed tissue texture), as even subtle modifications can affect clinical interpretations [27]. For instance, in chest X-ray (CXR) images, minor alterations in opacity patterns may mimic or obscure signs of pulmonary edema, potentially leading to misdiagnosis. Therefore, a watermarking method for medical images must embed a watermark message while maintaining diagnostic integrity and ensuring security.

To address these challenges, we introduce MedSign, a pioneering deep watermarking framework for text-to-medical image synthesis. To preserve diagnostic integrity, we leverage the diffusion model's internal cross-attention representation [18, 30] to identify critical regions, as these attention maps highlight clinically significant features. Leveraging this pathology-aware guidance, we fine-tune the variational autoencoder (VAE) decoder within the latent diffusion model (LDM) [24] to regulate watermark placement, ensuring that embedding occurs only in non-critical areas while preserving diagnostically relevant structures. This enables MedSign to achieve a balance between robust watermarking and clinical utility, securing generated medical images without compromising diagnostic reliability. Notably, MedSign inherently integrates watermarking into the image synthesis process without modifying the LDM's architecture or requiring additional post-processing of the generated image. Experiments on CXRs [14] and fundus images [17] confirm that MedSign effectively directs the watermark to non-critical regions while maintaining diagnostic fidelity. Furthermore, our MedSign

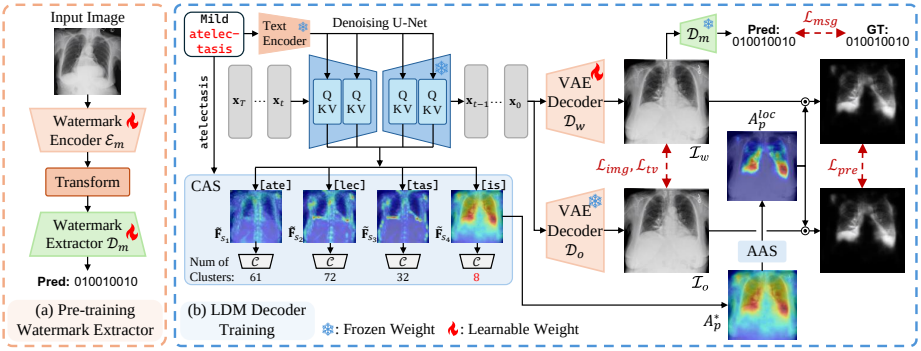

Fig. 1. Detailed illustration of our pipeline. (a) We first train the watermark encoder $\mathcal{E}_m$ and extractor $\mathcal{D}_m$ to embed and extract watermarks from transformed images. (b) We then train the LDM's VAE decoder $\mathcal{D}_w$ to embed watermarks while preserving pathology by aligning its output with the original VAE decoder $\mathcal{D}_o$, using a pathology localization map A_p^{loc} in the loss function to regulate watermark placement.

outperforms watermarking techniques designed for general images, achieving superior image quality and watermarking performance in the medical domain.

2 Methods

As illustrated in Fig. 1, we propose MedSign, a deep watermarking framework for text-driven medical image synthesis, ensuring high-fidelity preservation of pathological features. Our approach consists of (1) pre-training a watermark extractor (Sec. 2.1) and (2) fine-tuning the LDM decoder to integrate watermarks while minimizing interference with diagnostically significant regions (Sec. 2.2).

2.1 Pre-training the watermark extractor

We train the watermark extractor using the HiDDeN architecture [34] as illustrated in Fig. 1(a). This architecture consists of: (1) a watermark encoder $\mathcal{E}_m$ embedding a k-bit message into an image, (2) a transformation layer, and (3) a watermark extractor $\mathcal{D}_m$ that retrieves the binary message. The watermark encoder $\mathcal{E}_m$ embeds a k-bit message $m \in \{0,1\}^k$ into an image $\mathcal{I}$ to produce a watermarked image $\tilde{\mathcal{I}}$. After applying image transformation T (e.g., cropping, rotation) to $\tilde{\mathcal{I}}$, extractor $\mathcal{D}_m$ recovers the message, formally represented as $\tilde{m} = \mathcal{D}_m(T(\tilde{\mathcal{I}}))$. The model is trained by minimizing the binary cross-entropy loss $\mathcal{L}_{msg} = -\sum_{i=1}^{k} [m_i \log \sigma(\tilde{m}_i) + (1 - m_i) \log(1 - \sigma(\tilde{m}_i))]$. Similar to [8], image quality is not optimized (since $\mathcal{E}_m$ is discarded); instead, PCA whitening is applied after training to remove bias and decorrelate outputs. Feature vectors from $\mathcal{D}_m$ are centered using $\mu = \mathbb{E}[\mathcal{D}_m]$, and the covariance matrix $\Sigma = U \Lambda U^T$ is eigendecomposed. The whitening transform $\Lambda^{-1/2} U^T$ is then applied, with the resulting bias $-\Lambda^{-1/2} U^T \mu$ and weight $\Lambda^{-1/2} U^T$ appended as a linear layer.

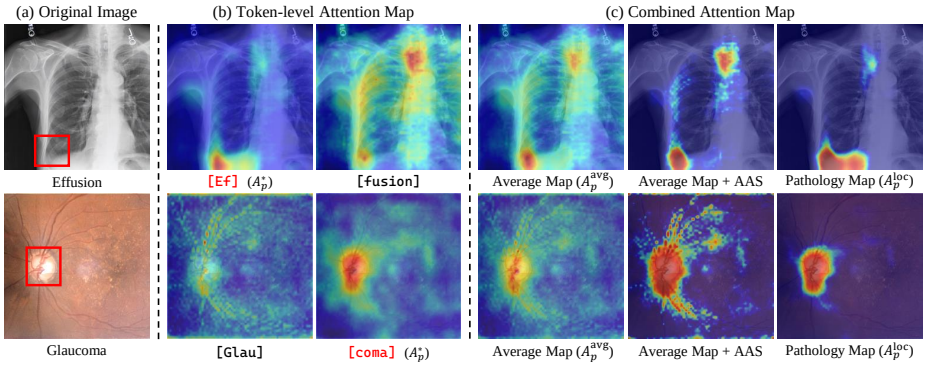

Fig. 2. Visualization of token-level and combined attention maps with pathology localization map. (a) Original CXR images with pathology regions in red boxes. (b) Token-level attention maps: individual maps for tokens composing pathology-related words. The red-highlighted token's attention map is the one selected through CAS. (c) Combined attention maps: (left) the average attention map of all tokens, (middle) its refinement using AAS, and (right) the final pathology localization map.

2.2 Fine-tuning the LDM decoder

We fine-tune the LDM's VAE decoder $\mathcal{D}_w$ to integrate watermarking directly into image generating process as shown in Fig. 1(b). To preserve diagnostically critical regions, we use a pathology localization map from cross-attention in the diffusion process. This map guides the loss function, enabling adaptive watermarking while maintaining image quality.

Pathology Localization Map Generation. To localize pathologically significant regions, we leverage cross-attention mechanisms [24, 25] that connect the diffusion denoising U-Net's spatial layout with medical text tokens.

Cross-Attention for Pathology Localization. We build on [18] to aggregate attention maps from diffusion models, highlighting pathological regions in medical images. Specifically, a U-Net-based [25] denoising network [24] $\epsilon_\theta(\mathbf{x}, t; \mathbf{y})$ maps noisy image features $\mathbf{x}_t \in \mathbb{R}^{h \times w}$ at step t to a denoised output, conditioned on an input medical report $\mathbf{y}$. Here, $\mathbf{y}$ comprises medical words, each word p consisting of n tokens as $p = \{s_1, s_2, \ldots, s_n\}$. To incorporate textual guidance, cross-attention [24, 30] injects token-level semantic embeddings from the medical report into the denoising network. This U-Net-based denoising network progressively refines the noisy input $\mathbf{x}_t$, generating token-specific attention maps $\mathbf{F}_s$, which correlate each token's semantic content with its spatial layout (see Fig. 2(b)). Then, each $\mathbf{F}_s$ is upsampled to $\widetilde{\mathbf{F}}_s$ to match the image resolution (h, w), and the token-level maps are then aggregated across layers, attention heads, and time steps as $A_p^{\text{avg}}(x, y) = \frac{1}{n} \sum_{i=1}^{n} \sum_{l,h,t} \widetilde{\mathbf{F}}_{s_i}^{(l,h,t)}(x, y)$, where $\widetilde{\mathbf{F}}_{s_i}^{(l,h,t)}$ is the upsampled attention map for token s_i at layer l, attention head h, and time

step t. This aggregation captures the collective spatial influence of all tokens in p, aligning attention with pathological regions in the medical image.

Density Clustering-Driven Attention Selection. While the aggregated map A_p^{avg} captures pathological regions, certain tokens within p can produce noise, shifting attention to irrelevant areas (Fig. 2(c) left), which may prevent watermarking in permissible regions and degrade watermarking performance. To address this, we adopt a density clustering-driven [6] attention selection mechanism (CAS) that suppresses dispersed attention signals and enhances pathological localization. Given a medical word p consisting of tokens $s_1, \ldots, s_n$, we seek the most precise attention map by assuming that well-defined maps exhibit dense, localized distributions. Thus, for each token s_i, we define a clustering function $\mathcal{C}$ that estimates the number of dense clusters k_i in the spatial attention distribution of $\widetilde{\mathbf{F}}_{s_i}$ using a density-based clustering algorithm, formulated as $k_i = \mathcal{C}(\widetilde{\mathbf{F}}_{s_i})$. To select the most localized attention map, we identify the token whose attention distribution exhibits the minimum cluster count as $s^* = \arg\min_{s_i} k_i$. Hence, the final localized attention map for p is defined as $A_p^*(x, y) = \widetilde{\mathbf{F}}_{s^*}(x, y)$. By selecting the attention map with the most compact distribution, this approach effectively mitigates noise and ensures precise pathological localization.

Adaptive Attention Scaling. To improve the clarity of localization boundaries, we apply an Adaptive Attention Scaling (AAS) mechanism that refines the attention distribution. We first normalize the selected attention map A_p^* using z-score transformation, followed by smooth sigmoid scaling and global min-max normalization. To emphasize salient regions, we define an adaptive threshold τ as the 0.7 quantile of A_p^*, a value empirically determined to provide a favorable trade-off between watermark robustness and the preservation of critical diagnostic information. We then define our final pathology localization map as

$$A_p^{\text{loc}}(x, y) \;=\; \sigma\left(A_p^*(x, y) \;-\; \tau\right), \tag{1}$$

where $\sigma(\cdot)$ is the sigmoid function. As shown in Fig. 2(c), this process enhances the contrast between pathological regions and the background, leading to sharper and more well-defined localization boundaries.

Pathology-Aware Adaptive Loss. We train the LDM decoder $\mathcal{D}_w$ with a loss function comprising image perceptual loss ($\mathcal{L}_{\text{img}}$), message loss ($\mathcal{L}_{\text{msg}}$), total variation regularization ($\mathcal{L}_{\text{tv}}$), and pathology preservation loss ($\mathcal{L}_{\text{pre}}$).

To preserve perceptual consistency, we use Watson-VGG perceptual loss [5] for $\mathcal{L}_{\text{img}}$, which minimizes visual distortions while keeping the watermark imperceptible. $\mathcal{L}_{\text{msg}}$ is computed via binary cross-entropy between the ground-truth message m and the predicted message $\hat{m}$, ensuring robust message recovery. To achieve this, we use the watermark extractor $\mathcal{D}_m$ trained in Sec. 2.1 by freezing its parameters during training. To encourage smooth watermark blending and suppress high-frequency artifacts, we incorporate $\mathcal{L}_{\text{tv}}$, which penalizes sharp watermark variations. Meanwhile, the pathology preservation loss ($\mathcal{L}_{\text{pre}}$) prevents

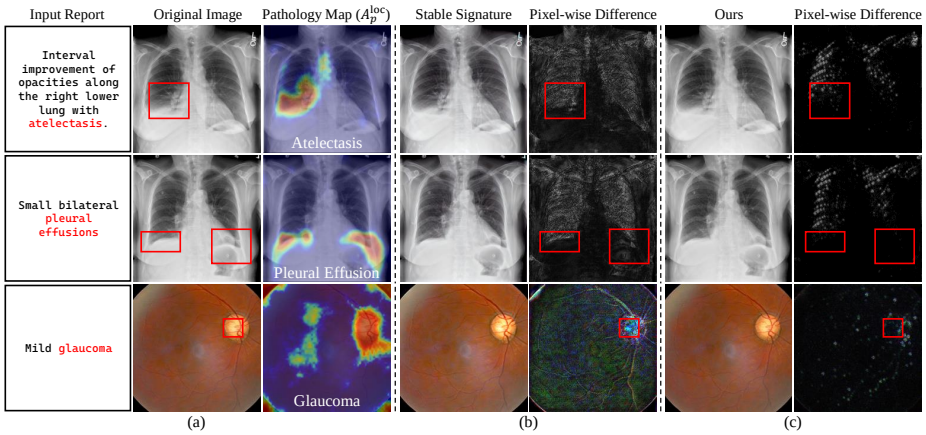

Fig. 3. Qualitative comparison of Stable Signature [8] and ours on MIMIC-CXR-JPG [14] and OIA-ODIR [17]. (a): Original image with pathology localization map. (b) & (c): Watermarked image with pixel-wise difference maps for Stable Signature and ours, respectively. Unlike Stable Signature, which embeds watermarks in critical regions (red box), our method preserves diagnostic areas.

watermark insertion in critical regions by leveraging the pathology localization map A_p^{loc}, preserving important areas while allowing embedding in less critical regions. The total objective $\mathcal{L}_{\text{total}}$ integrates these components as

$$\begin{aligned}\mathcal{L}_{\text{total}} = \lambda_{\text{img}} \underbrace{\sum_l \delta_l \left\| \left(\phi_l(\mathcal{I}_w) - \phi_l(\mathcal{I}_o)\right) \right\|_2^2}_{\mathcal{L}_{\text{img}}} + \lambda_{\text{msg}} \underbrace{\sum_{i=1}^{k} \text{BCE}(m_i, \hat{m}_i)}_{\mathcal{L}_{\text{msg}}} \\ + \lambda_{\text{tv}} \underbrace{\text{TV}(|\mathcal{I}_w - \mathcal{I}_o|)}_{\mathcal{L}_{\text{tv}}} + \lambda_{\text{pre}} \underbrace{\mathbb{E}\left[\|(\mathcal{I}_w - \mathcal{I}_o) \odot A_p^{\text{loc}}\|_2^2\right]}_{\mathcal{L}_{\text{pre}}},\end{aligned} \tag{2}$$

where the watermarked image is defined as $\mathcal{I}_w = \mathcal{D}_w(\mathbf{x}_0)$, while the original (non-watermarked) image $\mathcal{I}_o$ is decoded from the frozen original LDM decoder $\mathcal{D}_o$ as $\mathcal{I}_o = \mathcal{D}_o(\mathbf{x}_0)$. The VGG layer-$l$ features, denoted as $\phi_l(\cdot)$, are weighted by δ_l to guide perceptual consistency.

3 Experiments

Datasets. We use the impression section of reports from MIMIC-CXR-JPG [14] for CXR images. For training, we randomly select 6,000 unique samples from the official train split after removing duplicates. For testing, we use 1,903 unique reports from the test set, also excluding duplicates. For fundus data, we use unique reports from the OIA-ODIR dataset [17]. As images are generated from 290 distinct reports with different random seeds, no predefined split exists. Thus, we randomly assign 3,361 samples for training and 2,000 for testing.

Table 1. Quantitative evaluation of watermarking performance on the MIMIC-CXR-JPG dataset [14] and OIA-ODIR dataset [17]. Evaluated attack scenarios include None (no attack), Brt (brightness ×2), Crp (50% center crop), JPG (JPEG compression 50%), Rot (25° rotation), and Res (70% scaling). Method with † uses 32-bit message encoding, while others use 48-bit. We highlight the **best** and second-best results.

	Method	PSNR	SSIM	Bit Accuracy						
				None	Brt	Crp	JPG	Rot	Res	Avg
	MIMIC-CXR-JPG									
Post Gen	DCT-DWT [1]	35.17	0.89	**1.00**	0.51	0.50	0.52	0.51	0.53	0.59
	SSL Watermark [9]	36.20	0.89	0.93	0.66	0.84	0.71	0.85	0.87	0.81
	WAM† [26]	35.12	0.93	**1.00**	**1.00**	0.99	0.57	0.46	**1.00**	0.83
In Gen	Stable Signature [8]	33.24	0.92	0.99	0.99	0.99	0.70	0.71	0.95	0.89
	Ours (MedSign)	**39.71**	**0.96**	**1.00**	**1.00**	**1.00**	**0.74**	**0.96**	0.97	**0.95**
	OIA-ODIR									
Post Gen	DCT-DWT [1]	35.64	0.90	0.81	0.49	0.49	0.51	0.50	0.80	0.60
	SSL Watermark [9]	36.37	0.89	0.92	0.84	0.84	0.71	0.86	0.89	0.84
	WAM† [26]	31.35	**0.94**	**1.00**	**1.00**	0.79	0.72	0.45	**1.00**	0.85
In Gen	Stable Signature [8]	30.20	0.79	0.97	0.97	0.95	0.71	0.66	0.91	0.86
	Ours (MedSign)	**40.81**	**0.94**	**1.00**	0.95	**0.99**	**0.78**	**0.95**	0.97	**0.94**

Implementation Details. For diffusion model initialization, we use pretrained weights from RoentGen [2] for CXR images and DERETFound [15] for fundus images. For density-based clustering, we employ DBSCAN [6] with default settings. The hyperparameters are set as follows: for CXR images, $\lambda_{\text{img}} = 2.1$, $\lambda_{\text{msg}} = 7.5$, $\lambda_{tv} = 3.0$ and $\lambda_{\text{pre}} = 975.0$; and for fundus images, $\lambda_{\text{img}} = 1.9$, $\lambda_{\text{msg}} = 6.0$, $\lambda_{\text{tv}} = 10.0$ and $\lambda_{\text{pre}} = 800.0$, ensuring a balanced scale across them. Training runs for 2 epochs using AdamW with a 5×10^{-4} learning rate.

3.1 Evaluation Results

Qualitative Analysis. We qualitatively compare our watermarked images with the in-generation watermarking method [8] in Fig. 3. The pixel-wise difference map, computed as $|\mathcal{I}_w - \mathcal{I}_o| \times 10$, highlights watermarking effects. (1) *Pathology Localization Map*: Our method reduces misplaced alterations and artifacts by guiding watermark placement away from clinically significant areas using A_p^{loc}. (2) *Preservation of Critical Regions*: Integrating A_p^{loc} into training enhances watermark embedding in non-critical areas while preserving diagnostic integrity. Compared to the baseline, our approach minimizes interference in pathological regions (see red box in Fig. 3(b) and (c)), ensuring minimal clinical impact.

Quantitative Evaluation on Watermarking Performance. Table 1 evaluates watermarking performance on MIMIC-CXR-JPG [14] and OIA-ODIR [17], considering image quality (PSNR, SSIM) and watermarking robustness (Bit Accuracy) under perturbations. Methods are classified into *Post-Generation Water-*

Table 2. Comparative analysis on MIMIC-CXR-JPG [14]. (a) Diagnostic confidence: Absolute confidence score differences for six diseases. (b) Robustness and image quality: Average bit accuracy and PSNR, assessing the general performance of different models.

Method	(a)							(b)	
	Pna.	Pmtx.	Atel.	Ede.	Cmgl.	Effu.	Avg (↓)	AvgAcc (↑)	PSNR (↑)
Stable Signature [8]	13.18	2.76	1.99	7.87	2.99	1.70	5.08	0.89	33.24
Ours (MedSign)	**1.83**	**0.36**	**0.33**	**0.85**	**0.02**	**0.15**	**0.59**	**0.95**	**39.71**
(1) Ours w/o AAS	1.68	0.45	0.23	0.73	0.06	0.13	0.55	0.94	38.65
(2) Ours w/o CAS	1.75	0.38	0.25	0.75	0.06	0.16	0.56	0.92	39.51
(3) Ours w/o $\mathcal{L}_{\text{tv}}$	1.79	0.47	0.37	0.84	0.11	0.14	0.62	0.95	37.75
(4) Ours w/o $\mathcal{L}_{\text{pre}}$	4.20	1.91	2.27	2.31	1.23	1.89	2.30	0.94	35.02

marking, where watermarks are added after image generation, and *In-Generation Watermarking*, where they are embedded during synthesis for direct comparison with ours. DCT-DWT [1], used in Stable Diffusion [24], ensures perceptual fidelity but lacks robustness. SSL Watermark [9] and WAM [26] enhance robustness but remain vulnerable to specific attacks, while Stable Signature [8] balances detection at the cost of image fidelity. In contrast, MedSign ensures high-fidelity representation and resists attacks across CXR and fundus imaging, ensuring robustness across organs and modalities in medical watermarking.

Quantitative Evaluation on Diagnostic Impact. To assess the diagnostic impact, Table 2(a) reports absolute differences in confidence scores (%p) between watermarked and original images (i.e., $|P_{\text{watermarked}} - P_{\text{original}}| \times 100$) across six pulmonary diseases. Scores are obtained using a state-of-the-art CXR classifier [4]. A non-medical-specific watermarking model [8] introduces an average confidence difference of 5.08, reaching 13.18 for Pneumonia, which could misrepresent clinical findings. In contrast, ours achieves a significantly lower average difference, minimizing diagnostic impact while preserving medical integrity.

3.2 Ablation Analysis

(1) & (2): Effect of AAS and CAS. We analyze the impact of AAS and CAS by removing each from MedSign in Table 2. Here, (2) refers to the use of A_p^{avg} instead of A_p^* in Eq. (1). Diagnostic confidence remains stable since the removal enforces watermarking within a narrower region than our proposed pathology localization map. However, this constraint reduces average bit accuracy, highlighting the trade-off between diagnostic integrity and watermark robustness.

(3): Effect of $\mathcal{L}_{\text{tv}}$. Total variation regularization $\mathcal{L}_{\text{tv}}$ enhances the smoothness of watermarking, preserving image fidelity while maintaining diagnostic consistency as shown in Table 2(3). Its removal reduces PSNR and introduces visual artifacts, though diagnostic confidence remains largely unaffected.

(4): Effect of $\mathcal{L}_{\text{pre}}$. As shown in Table 2(4), removing pathology preservation loss $\mathcal{L}_{\text{pre}}$ leads to a significant drop in PSNR and increased alterations in diagnos-

tic confidence, while average bit accuracy remains nearly unchanged. Without $\mathcal{L}_{\text{pre}}$, watermarking is applied globally, preserving bit accuracy but compromising diagnostic integrity and degrading overall image quality. This suggests that $\mathcal{L}_{\text{pre}}$ is essential for restricting watermarking to non-critical regions, maintaining diagnostic integrity, and preserving clinically significant features.

4 Conclusion and Discussion

We present a pathology-aware watermarking framework for text-to-medical image generation, ensuring robust embedding while preserving diagnostically critical regions. By leveraging a carefully tailored cross-attention map, our method precisely localizes pathologies and regulates watermark placement to minimize interference. Experiments show that our approach balances image fidelity and watermark resilience, surpassing existing methods in robustness while minimizing clinical interpretation. This study highlights the potential of pathology-aware watermarking to authenticate text-driven AI-generated medical images without compromising diagnostic utility. Future work includes extending our approach to other modalities, such as CT and MRI, for broader medical applicability.

Disclosure of Interests. The authors have no competing interests.

Acknowledgments. This work was supported in part by the IITP RS-2024-00457882 (AI Research Hub Project), IITP 2020-II201361, NRF RS-2024-00345806, NRF RS-2023-00219019, and NRF RS-2023-002620.

References

1. Al-Haj, A.: Combined dwt-dct digital image watermarking. Journal of computer science (2007)
2. Bluethgen, C., Chambon, P., Delbrouck, J.B., van der Sluijs, R., Połacin, M., Zambrano Chaves, J.M., Abraham, T.M., Purohit, S., Langlotz, C.P., Chaudhari, A.S.: A vision–language foundation model for the generation of realistic chest x-ray images. Nature Biomedical Engineering (2024)
3. Ci, H., Yang, P., Song, Y., Shou, M.Z.: Ringid: Rethinking tree-ring watermarking for enhanced multi-key identification. In: ECCV (2024)
4. Cohen, J.P., Viviano, J.D., Bertin, P., Morrison, P., Torabian, P., Guarrera, M., Lungren, M.P., Chaudhari, A., Brooks, R., Hashir, M., Bertrand, H.: TorchXRayVision: A library of chest X-ray datasets and models. In: Medical Imaging with Deep Learning (2022)
5. Czolbe, S., Krause, O., Cox, I., Igel, C.: A loss function for generative neural networks based on watson's perceptual model. In: NeurIPS (2020)
6. Ester, M., Kriegel, H.P., Sander, J., Xu, X., et al.: A density-based algorithm for discovering clusters in large spatial databases with noise. In: KDD (1996)
7. European Parliament and Council: Regulation (EU) 2024/1689 of the European Parliament and of the Council on Artificial Intelligence (2024), `https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32024R1689`

8. Fernandez, P., Couairon, G., Jégou, H., Douze, M., Furon, T.: The stable signature: Rooting watermarks in latent diffusion models. ICCV (2023)
9. Fernandez, P., Sablayrolles, A., Furon, T., Jégou, H., Douze, M.: Watermarking images in self-supervised latent spaces. In: International Conference on Acoustics, Speech, and Signal Processing (2022)
10. Han, W., Kim, C., Ju, D., Shim, Y., Hwang, S.J.: Advancing text-driven chest x-ray generation with policy-based reinforcement learning. In: MICCAI (2024)
11. Han, W., Lee, Y., Kim, C., Park, K., Hwang, S.J.: Spatial transport optimization by repositioning attention map for training-free text-to-image synthesis. In: CVPR (2025)
12. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. NeurIPS (2020)
13. Jang, Y., Lee, D.I., Jang, M., Kim, J.W., Yang, F., Kim, S.: Waterf: Robust watermarks in radiance fields for protection of copyrights. In: CVPR (2024)
14. Johnson, A.E., Pollard, T.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Peng, Y., Lu, Z., Mark, R.G., Berkowitz, S.J., Horng, S.: Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. arXiv preprint arXiv:1901.07042 (2019)
15. Jonlysun: Deretfound: Deep embedded representation for fundus image analysis. `https://github.com/Jonlysun/DERETFound` (2023)
16. Lee, S., Kim, W.J., Chang, J., Ye, J.C.: LLM-CXR: Instruction-finetuned LLM for CXR image understanding and generation. In: ICLR (2024)
17. Li, N., Li, T., Hu, C., Wang, K., Kang, H.: A benchmark of ocular disease intelligent recognition: One shot for multi-disease detection. In: Benchmarking, Measuring, and Optimizing: Third BenchCouncil International Symposium, Bench 2020, Virtual Event, November 15–16, 2020, Revised Selected Papers 3. Springer (2021)
18. Marcos-Manchón, P., Alcover-Couso, R., SanMiguel, J.C., Martínez, J.M.: Open-vocabulary attention maps with token optimization for semantic segmentation in diffusion models. In: CVPR (2024)
19. Medghalchi, Y., Zakariaei, N., Rahmim, A., Hacihaliloglu, I.: Meddap: Medical dataset enhancement via diversified augmentation pipeline. arXiv preprint arXiv:2403.16335 (2024)
20. Mirsky, Y., Mahler, T., Shelef, I., Elovici, Y.: {CT-GAN}: Malicious tampering of 3d medical imagery using deep learning. In: 28th USENIX Security Symposium (USENIX Security 19) (2019)
21. National Assembly of the Republic of Korea: Act No. 20676 (2024), `https://www.law.go.kr/LSW/lsInfoP.do?lsiSeq=268543&viewCls=lsRvsDocInfoR#`
22. Nichol, A., Dhariwal, P., Ramesh, A., Shyam, P., Mishkin, P., McGrew, B., Sutskever, I., Chen, M.: Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741 (2021)
23. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., Chen, M.: Hierarchical text-conditional image generation with clip latents. arXiv preprint arXiv:2204.06125 (2022)
24. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022)
25. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: MICCAI (2015)
26. Sander, T., Fernandez, P., Durmus, A., Furon, T., Douze, M.: Watermark anything with localized messages. In: ICLR (2025)

27. Siracusano, G., La Corte, A., Nucera, A.G., Gaeta, M., Chiappini, M., Finocchio, G.: Effective processing pipeline pace 2.0 for enhancing chest x-ray contrast and diagnostic interpretability. Scientific Reports (2023)
28. State Council of the People's Republic of China: New Generation Artificial Intelligence Development Plan (2023), https://www.gov.cn/zhengce/content/202306/content_6884925.htm
29. Trabucco, B., Doherty, K., Gurinas, M.A., Salakhutdinov, R.: Effective data augmentation with diffusion models. In: ICLR (2024)
30. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: NeurIPS (2017)
31. Wen, Y., Kirchenbauer, J., Geiping, J., Goldstein, T.: Tree-rings watermarks: Invisible fingerprints for diffusion images. NeurIPS (2023)
32. Yellapragada, S., Graikos, A., Prasanna, P., Kurc, T., Saltz, J., Samaras, D.: Pathldm: Text conditioned latent diffusion model for histopathology. In: WACV (2024)
33. Zhang, X., Li, R., Yu, J., Xu, Y., Li, W., Zhang, J.: Editguard: Versatile image watermarking for tamper localization and copyright protection. In: CVPR (2024)
34. Zhu, J., Kaplan, R., Johnson, J., Fei-Fei, L.: Hidden: Hiding data with deep networks. In: ECCV (2018)