

# DX-Mamba: Exploring State Space Model for Dog X-ray Report Generation

Jialu Li, M.S. in Artificial Intelligence

FACULTY MENTOR: Youshan Zhang, Ph.D.



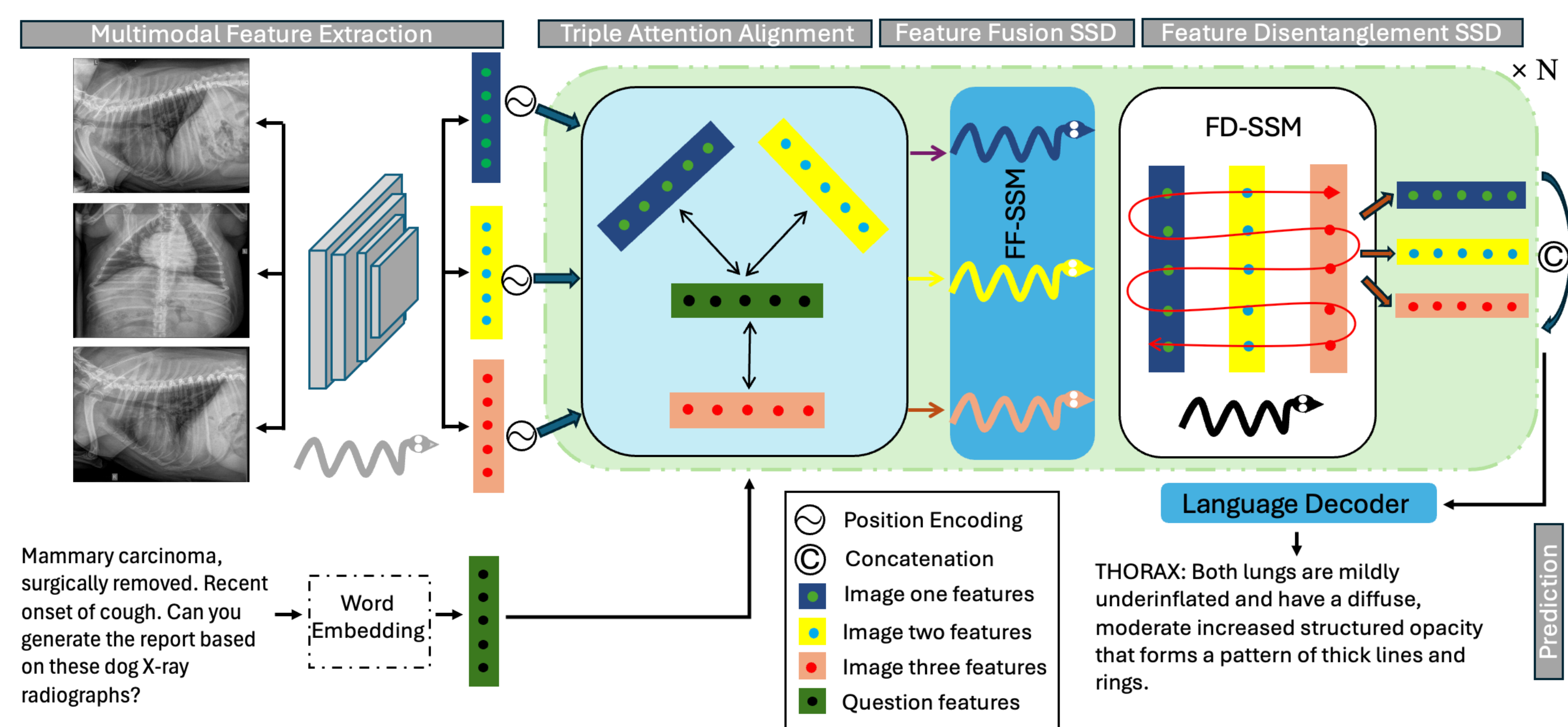
**Katz**  
Katz School  
of Science and Health

## Introduction

- Background:** Thoracic diseases are a common cause of dog death. The veterinary medicine field has been focusing on the early and accurate detection of canine thoracic diseases. However, writing diagnosis reports is time-consuming, requires expertise, and is error-prone.
- Problem:** Automatic medical report generation requires multi-modal knowledge learning to produce coherent and accurate reports (Liao et al., 2023). Current deep-learning models produce factually incomplete and inconsistent reports (Tanida et al., 2023). Other challenges include a lack of data for training models and comparative imaging data.
- Aim:** We aim to address the difficulty in manually producing medical reports in the veterinary field and achieve the goal of automatic report generation by improving the performance accuracy of deep learning models on this task.

## Method

Figure 1. Workflow of the DX-Mamba model



We propose a novel DX-Mamba model for automatic dog X-ray image report generation, including three key structures, as shown in Figure 1:

- Triple Attention Alignment (TAA):** aligning the text features with each image to get shared information between them.

$$\{F_T(B_i)\}_{n=1}^3 = TAA(\{B_{fi}\}_{i=1}^3, W(R)).$$

- Feature Fusion SSM (FF-SSM):** enabling complementary feature learning from input features.

$$F_F(B_i) = \Psi(F_T^a(B_i) \times B_c + F_T^b(B_i) \times B_c) + F_T(B_i)$$

- Features Disentanglement SSM (FD-SSM):** facilitating distinct features.

$$F_D(B_1) = F_D(B)[:, : T_p, :], F_D(B_2) = F_D(B)[:, T_p : 2T_p, :] F_D(B_3) = F_D(B)[:, 2T_p :, :]$$

A transformer decoder is used for the report generation task. We also include a combination of objective functions:

- Intra-class loss: calculates distances between class samples and the center.
- Inter-class loss: calculates distances between different class centers.
- Cross-entropy loss.

## Results

We conducted extensive experiments with our Dog-Xray dataset and refer to the following evaluation metrics:

- BLEU-4: Measures how many n-grams in the generated text match with reference text.
- CIDEr: Measures consensus of n-grams among multiple references.
- METEOR: Evaluates precision, recall, stemming, synonyms, and word order.
- ROUGE-L: Measures sentence-level structure similarity.

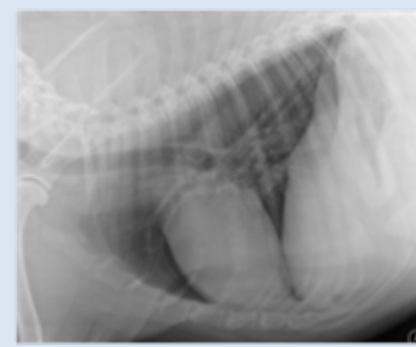

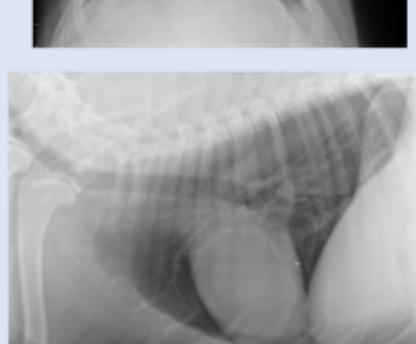
As shown in Table 1, our proposed DX-Mamba model outperforms state-of-the-art visual-language models and RAG models on all metrics in the automatic report-generation task using the Dog-Xray test sets. In particular, we can see great improvements in BLEU and CIDEr scores.

Table 1. Results comparisons of different methods on the Dog-Xray dataset test set.

Methods	Test						
	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
RAG.raw	0.257	0.188	0.156	0.137	0.363	0.309	0.109
RAG.Mistral (Kurtev & van den Berg, 2003)	0.214	0.143	0.112	0.095	0.370	0.291	0.011
RAG.LLaMA (Touvron et al., 2023)	0.219	0.143	0.115	0.099	0.299	0.360	0.004
RAG.BART (Lewis, 2019)	0.120	0.0807	0.066	0.058	0.153	0.188	0.0214
RAG.GPT2 (Sufi, 2024)	0.184	0.123	0.097	0.082	0.282	0.365	0.024
LLaMA-3.2-3B (Dubey et al., 2024)	0.284	0.219	0.185	0.163	0.337	0.390	0.085
LLaVA-1.5-7B (Liu et al., 2024c)	0.226	0.161	0.131	0.112	0.265	0.353	0.035
DCL (Li et al., 2023a)	0.233	0.171	0.142	0.126	0.271	0.359	0.088
R2GenCSR-Qwen (Wang et al., 2024)	0.258	0.179	0.149	0.132	0.286	0.361	0.980
R2Gen-Mamba (Sun et al., 2024)	0.298	0.189	0.157	0.149	0.326	0.366	0.980
DX-Mamba	<b>0.456</b>	<b>0.350</b>	<b>0.291</b>	<b>0.252</b>	<b>0.396</b>	<b>0.399</b>	<b>0.143</b>

We can also see from Figure 2, the prediction samples comparisons with an existing model, R2Gen-Mamba, that the prediction from the DX-Mamba model is highly similar to the sample ground truth, while the R2Gen-Mamba model gives additional irrelevant information.

Figure 2: Predicted reports comparison on a sample image pair 12088 using DX-Mamba and R2Gen-Mamba models

Input Lung Images 12088	Ground truth	Prediction: DX-Mamba	Prediction: R2Gen-Mamba
	<b>No abnormality is detected.</b> SUMMARY: 1) Clinically normal thoracic radiographs CONCLUSION: The results of this evaluation are <b>negative for thoracic metastatic neoplasia</b> . The conclusions described in this report are based on the available information. If additional images (or other information) are obtained, then these conclusions might change.	<b>No abnormality is detected.</b> SUMMARY: Clinically normal thoracic radiographs CONCLUSION: The results of this examination are <b>negative for pulmonary metastasis</b> from reported mast cell tumor. The conclusions described in this report are based on the available information. If additional images (or other information) are obtained, then these conclusions might change.	THORAX: The lungs are <b>well inflated with normal opacity</b> . No abnormality is detected. <u>The liver is mildly enlarged with rounded margins.</u> SUMMARY: 1. Clinically normal thoracic radiographs 2. Mild, diffuse hepatomegaly CONCLUSION: The results of this evaluation are <b>negative for pulmonary metastasis</b> . The conclusions described in this report are based on the available information. If additional images (or other information) are obtained, then these conclusions might change.
			
			

We also compared the performance of DX-Mamba with state-of-the-art models using benchmark datasets, IU-Xray. Our model also displayed promising performance on the report generation task, indicating that our model has great potential to be generalized from the veterinary field to the human medical field.

**Results show that our proposed DX-Mamba model can achieve state-of-the-art performance in the automatic report generation task.**

## Results, Cont.

Table 2. Ablation studies on the three major components TAA, FF-SSM, and FD-SSM using the Dog-Xray validation set.

Methods	TAA	FF-SSM	FD-SSM	BLEU-1	BLEU-2	BLEU-3	BLEU-4	METEOR	ROUGE-L	CIDEr
Baseline	×	×	×	0.424	0.310	0.269	0.230	0.298	0.380	0.091
TAA	✓	×	×	0.425	0.313	0.271	0.232	0.302	0.380	0.097
FF-SSM	×	✓	×	0.427	0.314	0.272	0.232	0.317	0.386	0.103
FD-SSM	×	×	✓	0.430	0.316	0.273	0.233	0.338	0.386	0.114
TAA + FF-SSM	✓	✓	×	0.432	0.321	0.275	0.239	0.351	0.395	0.128
TAA + FD-SSM	✓	×	✓	0.432	0.324	0.275	0.239	0.372	0.398	0.128
FF-SSM + FD-SSM	×	✓	✓	0.434	0.336	0.281	0.246	0.381	0.403	0.129
DX-Mamba	✓	✓	✓	<b>0.447</b>	<b>0.344</b>	<b>0.286</b>	<b>0.249</b>	<b>0.394</b>	<b>0.398</b>	<b>0.137</b>

Table 3. Ablation studies on proposed loss functions on the Dog-Xray validation set (1, 2, 3, and 4 refer to BLEU-1, BLEU-2, BLEU-3, and BLEU-4).

Methods	1	2	3	4	METEOR	ROUGE-L	CIDEr
CE	0.434	0.336	0.281	0.246	0.227	0.383	0.129
Intra	0.439	0.340	0.281	0.247	0.298	0.384	0.122
Inter	0.439	0.339	0.282	0.248	0.301	0.393	0.137
Intra + Inter	<b>0.447</b>	<b>0.344</b>	<b>0.286</b>	<b>0.249</b>	<b>0.394</b>	<b>0.398</b>	<b>0.137</b>

Additional ablation studies in Tables 2 and 3, focusing on testing the function of different model components and losses, also prove that our proposed model architecture and loss functions are effective in improving our model's performance in producing more accurate and semantically relevant medical reports.

## Conclusions

Extensive experiment results showed our DX-Mamba achieves state-of-the-art performance on automatic report generation for both veterinary and human medical fields (human tables not included here), indicating the generalizability of our model.

**Future plans:** We plan to modify our model to make it a fully Mamba-based model, further leveraging the superior capacity of Vision Mamba on text and image feature learning and processing.

## Acknowledgments

I would like to express my sincere gratitude to my mentor, Professor Youshan Zhang, Ph.D., who guided me through the whole research, and to the Katz School of Science and Health community, which provided sufficient research resources for my experiments.

## References

- Liao, Y., Liu, H., & Spasić, I. (2023). Deep learning approaches to automatic radiology report generation: A systematic review. *Informatics in Medicine Unlocked*, 39, 101273.
- Tanida, T., Müller, P., Kaissis, G., & Rueckert, D. (2023). Interactive and explainable region-guided radiology report generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 7433-7442).
- Gu, A., & Dao, T. (2023). Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*.
- Liu, X., Zhang, C., & Zhang, L. (2024). Vision mamba: A comprehensive survey and taxonomy. *arXiv preprint arXiv:2405.04404*.
- Liu, C., Chen, K., Chen, B., Zhang, H., Zou, Z., & Shi, Z. (2024). Rscama: Remote sensing image change captioning with state space model. *IEEE Geoscience and Remote Sensing Letters*.