**Reviewer Comment 1:**

Thank you very much for your review and feedback.

Regarding model complexity, it is indeed true that a network combining ConViT, Transformer, and multiple parallel LSTMs demands significant computational resources.

Concerning the baseline, we acknowledge a descriptive oversight on our part. The baseline model comprises ViT, Transformer, and a single LSTM, and we conducted ablation studies by sequentially replacing corresponding components. Notably, ConViT has been enhanced by incorporating a weight-balancing operator. Additionally, we connected ConViT and the Transformer with an LSTM to serialize graphical features. We introduced multiple LSTMs later in the process—one of which focuses on pre-fusion images and confidence comparison. This adjustment has, indeed, impacted both speed and computational complexity. We prioritized accuracy over efficiency in this first-generation model.

Furthermore, your suggestion led us to recognize that our complex model might lead to overfitting on small datasets, a notion that subsequent experiments have confirmed. We plan to address these biases in our future work.

Finally, I apologize for my limited writing skills, as this is my first attempt, and I will refine my writing with your comments in mind.

**Reviewer Comment 2:**

Thank you for your review.

Due to the 8-page limit, we could not include future work in the initial submission. We have actually been developing Vision-Language Models (VLMs) in parallel with this model. This first-generation model is specifically designed for smaller institutions that lack high-performance GPUs.

We also generated VLM results using GPT-4. For instance, Figure 5’s first image produced the following report: “There are no acute cardiopulmonary abnormalities. The heart size and mediastinal contours are within normal limits, and pulmonary vascularity appears normal. There is no evidence of focal consolidation, suspicious pulmonary opacity, pneumothorax, or significant pleural effusion. The visualized osseous structures remain intact.”

The second image report reads: “There are no acute cardiopulmonary abnormalities. The mediastinal contours are normal, and the lungs appear clear. No pneumothorax or significant pleural effusion is present.”

Our model achieved higher BLEU scores than fine-tuned GPT-4 results. We will add these details in the final version.

We introduced Tri-LSTM because one LSTM is responsible for pre-fusion image processing and confidence comparison, whereas previous models used separate classifiers. Our approach integrates image features before text classification, which improves accuracy at the cost of speed and computational complexity. In the paper “A ConvNet for the 2020s,” it is mentioned that with appropriate training strategies and by leveraging dataset-specific statistics, CNNs can outperform transformers, though they require longer training. We designed a complex model to capture effective features amidst label redundancy, incorporating a weight-balancing operator for feature enhancement. Our results confirm the accuracy gains, albeit with reduced speed. Reviewer 3’s insights have prompted further optimization, including a single memory-bound bridge to accelerate FFN and improve efficiency.

We also collaborated with clinicians, who conducted clinical assessments of the generated reports. I will add VLMs and manual references in the camera-ready version.

Thanks especially to Reviewer 2, as this feedback inspired additional improvements.

**Reviewer Comment 3:**

Thank you very much for your review.

I am delighted to share that we have conducted comparisons with large models like GPT-4. For instance, the results for Figure 5’s first image were “There are no acute cardiopulmonary abnormalities. The heart size and mediastinal contours are within normal limits, and pulmonary vascularity appears normal. There is no evidence of focal consolidation, suspicious pulmonary opacity, pneumothorax, or significant pleural effusion. The visualized osseous structures remain intact.”

For the second image, the output was: “There are no acute cardiopulmonary abnormalities. The mediastinal contours are normal, and the lungs appear clear. No pneumothorax or significant pleural effusion is present.”

We excluded this from the paper because we plan to release these results in conjunction with a newly proposed large dataset.

If accepted, we will include large-model results in the camera-ready version. Our model’s BLEU score surpasses that of larger models. Moreover, our collaborators include clinicians, so in addition to VLM evaluations, we also have clinical validation. We can further calibrate the model using physician feedback in human-AI interactions, and we will include VLMs and clinical assessments in the camera-ready version.

I also greatly appreciate your second suggestion, which helped us identify some inefficiencies in ConViT related to memory usage. In response, we developed a single memory-bound bridge to accelerate FFN.

We also observed redundancies among attention heads, which consumes unnecessary resources. This insight presents a new challenge.

The first resolved point will be mentioned in the methodology section, and the rest will be added to future work. I apologize for not including future work initially due to the page limit.

**Regarding Your First Concern:**

Due to previous page-limit constraints (the paper was rejected initially for exceeding the page limit), we omitted a detailed list of hyperparameters.

Here are a few key hyperparameters:

* Learning Rate: 0.0005
* Epochs: 10
* LSTM step size: 512
* Attention mechanism with double weighting, etc.

For validation, we applied cross-validation and compared the impact of different hyperparameters on the model.

We experimented with adjusting the loss function weights, and here are a few sample results (IU-X-ray):

* (a=3): BLEU-1 = 0.497, BLEU-2 = 0.389, BLEU-3 = 0.291, BLEU-4 = 0.184, CIDEr = 0.410, ROUGE = 0.427
* (a=4): BLEU-1 = 0.509, BLEU-2 = 0.356, BLEU-3 = 0.279, BLEU-4 = 0.193, CIDEr = 0.413, ROUGE = 0.414
* (a=5): BLEU-1 = 0.493, BLEU-2 = 0.378, BLEU-3 = 0.277, BLEU-4 = 0.201, CIDEr = 0.383, ROUGE = 0.441

While we could illustrate trend changes with graphs, the page limit restricted us from including them. I also recognize the need to improve my writing skills.

We will include these details in the camera-ready version.

**Regarding Your Third Concern:**

Our intention was to optimize the model based on dataset-specific statistics. The weight-balancing operator addresses the uneven sample feature distribution within small datasets (as seen in Figure 3). The operator harmonizes features based on these statistical characteristics, enhancing the model’s specialization, though requiring substantial computational resources.

For AENSI, they utilized standalone classifiers, while we integrated an image-feature classifier to address the challenge of extracting features from closely annotated datasets. In iterative calculations, features can diminish, so we reinforced image features for greater accuracy at the cost of longer training time.