Generating Natural Questions About an Image

Grant Zhao, Jacob Fernandez

INTRODUCTION

- **Image captioning has focused on literal, surface-level descriptions**
- **This project proposes Visual Question Generation (VQG): automatically generating natural, engaging questions from an image.**

Reason for Study

- **While previous tasks focused on literal descriptions of images, VQG moves beyond that by exploring how questions address abstract events and commonsense inferences that objects in images evoke.**
- **A VQG task is designed to generate questions that are natural sounding, engaging, and prompt deeper thinking about the image.**

METHODOLOGY/MODEL ARCHITECTURE

GRU Cell Dynamics

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z) \quad \text{(update gate)} \tag{1}$$

$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r) \quad \text{(reset gate)} \tag{2}$$

$$\tilde{h}_t = \tanh(W_h x_t + U_h(r_t \circ h_{t-1}) + b_h) \quad \text{(candidate state)} \tag{3}$$

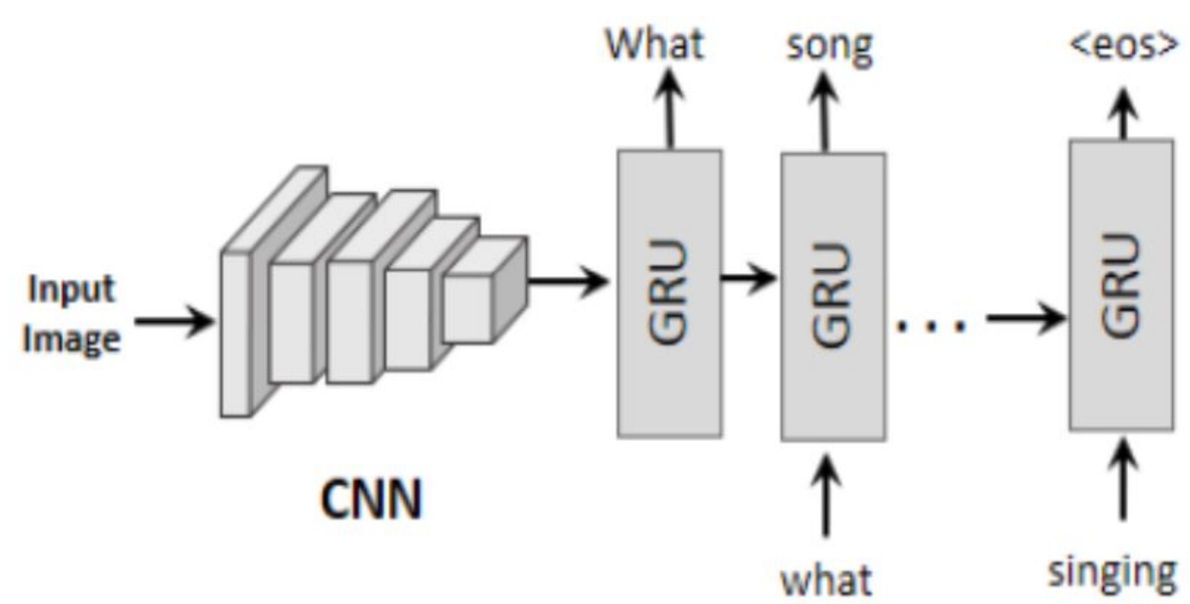
$$h_t = (1 - z_t) \circ h_{t-1} + z_t \circ \tilde{h}_t \quad \text{(new hidden state)} \tag{4}$$

Output Distribution

$$y_t = \text{softmax}(W_o h_t + b_o) \tag{5}$$

Training Loss (Negative Log-Likelihood)

$$\mathcal{L} = -\sum_{t=1}^{T} \log p(w_t \mid w_{<t}, I) \tag{6}$$



- 3 datasets, MS COCO, Flickr, and Bing
- COCO dataset limited in terms of concepts covered
- Flickr dataset images appear as middle of a photo album
- Bing dataset queried a search engine with 1,200 event-centric query terms
- 5,000 images per each dataset, total of 15,000 images and 75,000 questions

CONCLUSION

		$Human_{consensus}$	$Human_{random}$	$GRNN_X$	$GRNN_{all}$
Human Evaluation					
	Bing	2.50	2.36	1.38	**1.81**
	COCO	2.50	2.40	1.62	**1.97**
	Flickr	2.33	2.28	1.27	**1.58**
BLEU	Bing	87.3	83.6	**12.4**	11.0
	COCO	86.1	83.8	**13.8**	14.3
	Flickr	84.5	83.4	**10.0**	9.8
MET.	Bing	62.0	59.0	**16.0**	15.6
	COCO	60.7	58.5	**18.2**	18.3
	Flickr	59.5	58.0	**14.1**	14.0
ΔBLEU	Bing	63.0	57.5	**11.5**	10.7
	COCO	61.0	56.9	**12.3**	12.4
	Flickr	62.0	57.2	**9.4**	9.2

Our evaluation results for the GRNN model using BLEU 1-4 metrics n-gram overlap

BLEU Scores	Bing	COCO	Flickr
	12.1	13.6	10.2

RESEARCH CHALLENGES

- **Corrupted image dataset, removal and data augmentation**
- **Dataset limitations, creating datasets with truly natural questions**
- **Question diversity - pursuing more complex event-centric questions**
- **Text Augmentation - use back translation**
- **Evaluation metrics, how do we evaluate the quality of a generated question?**

RESULTS

			
HUMAN	- How long did it take to make that ice sculpture?	- Is the dog looking to take a shower?	- Was this explosion an accident?
GRNN	- Where was this picture taken ?	- Why is this dog in a bathroom ?	- What caused this explosion ?

NEXT STEPS

- **Question generation within a conversation system?**
- **While our models learn to generate promising questions, large gap to match humans still exists**

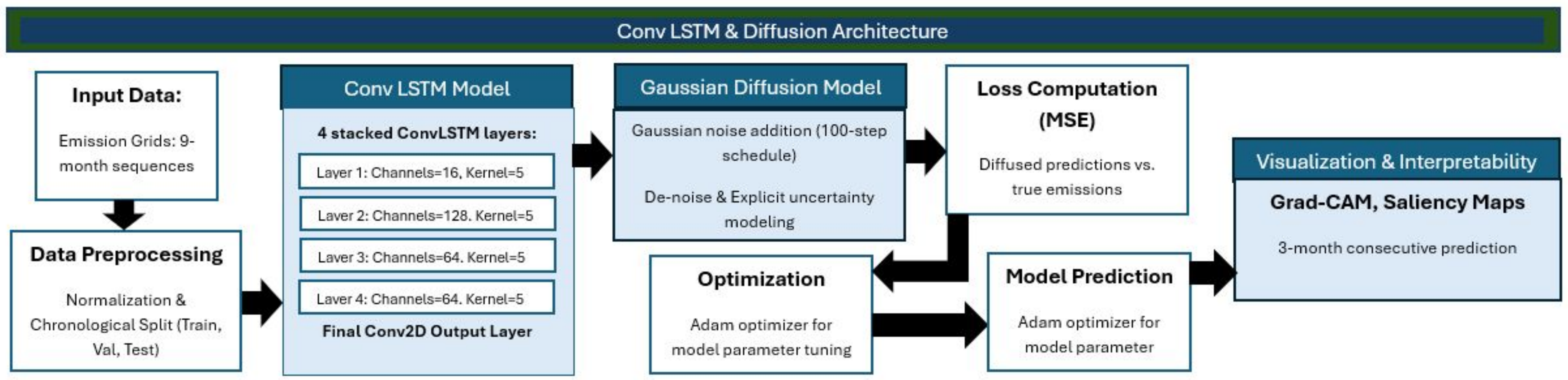
ACKNOWLEDGEMENTS

- Mostafazadeh, N. et al. (2016) Generating natural questions about an image, arXiv.org. Available at: https://arxiv.org/abs/1603.06059 (Accessed: 05 May 2025).

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z) \quad \text{(update gate)} \quad (1)$$
$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r) \quad \text{(reset gate)} \quad (2)$$
$$\tilde{h}_t = \tanh(W_h x_t + U_h(r_t \circ h_{t-1}) + b_h) \quad \text{(candidate state)} \quad (3)$$
$$h_t = (1 - z_t) \circ h_{t-1} + z_t \circ \tilde{h}_t \quad \text{(new hidden state)} \quad (4)$$

(5)

(6)



(5)

GRU Cell Dynamics:

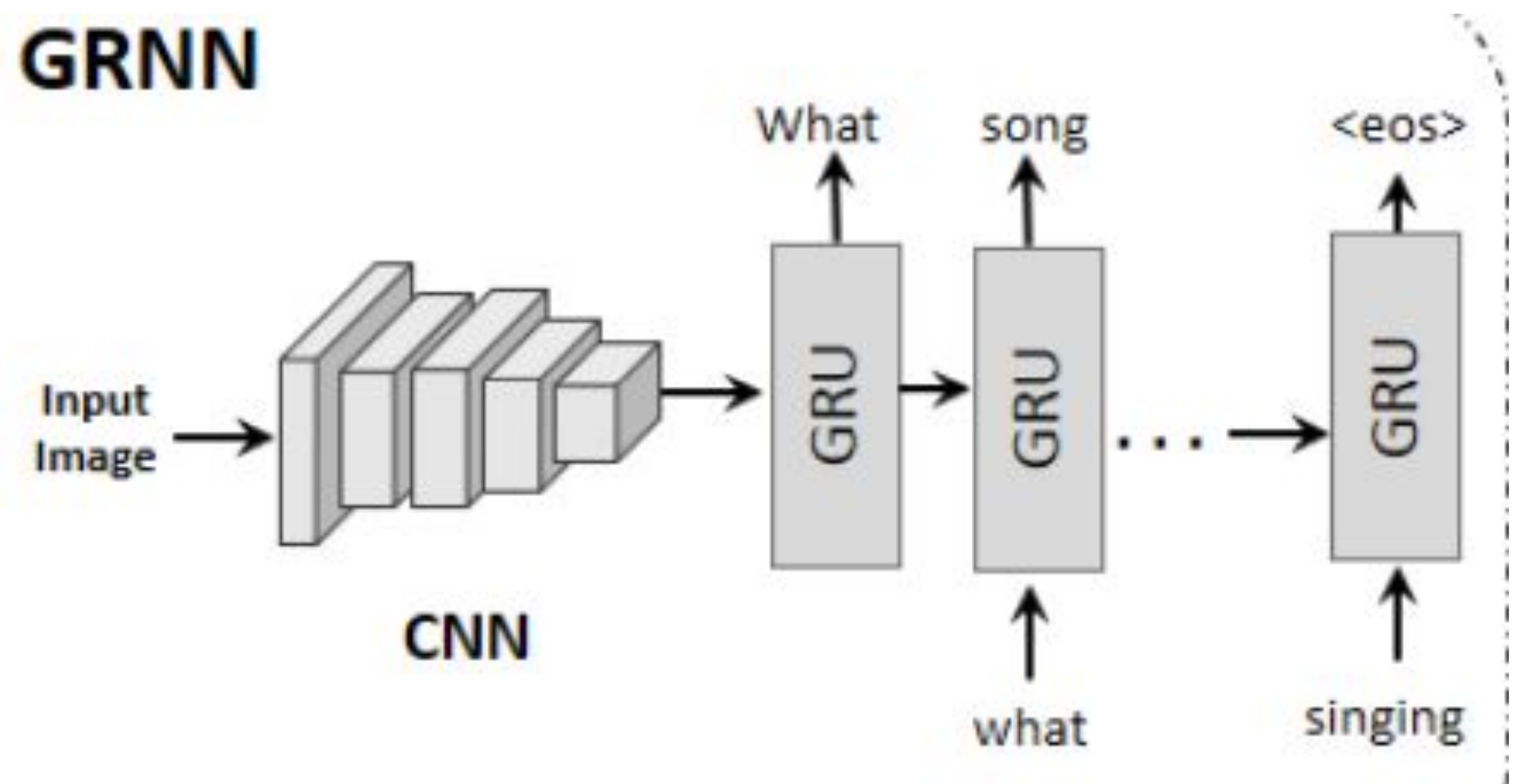
$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z) \quad \text{(update gate)} \quad (1)$$
$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r) \quad \text{(reset gate)} \quad (2)$$
$$\tilde{h}_t = \tanh(W_h x_t + U_h(r_t \circ h_{t-1}) + b_h) \quad \text{(candidate state)} \quad (3)$$
$$h_t = (1 - z_t) \circ h_{t-1} + z_t \circ \tilde{h}_t \quad \text{(new hidden state)} \quad (4)$$

Output Distribution:

$$y_t = \mathrm{softmax}(W_o h_t + b_o) \quad (5)$$

Training Loss (Negative Log-Likelihood):

$$\mathcal{L} = -\sum_{t=1}^{T} \log p(w_t \mid w_{<t}, I) \quad (6)$$



$$i_t = \sigma\left(W_{xi} * X_t + W_{hi} * H_{t-1} + W_{ci} \circ C_{t-1} + b_i\right)$$
$$f_t = \sigma\left(W_{xf} * X_t + W_{hf} * H_{t-1} + W_{cf} \circ C_{t-1} + b_f\right)$$
$$\tilde{C}_t = \tanh\left(W_{x\tilde{C}} * X_t + W_{h\tilde{C}} * H_{t-1} + b_{\tilde{C}}\right)$$
$$C_t = f_t \circ C_{t-1} + i_t \circ \tilde{C}_t$$
$$o_t = \sigma\left(W_{xo} * X_t + W_{ho} * H_{t-1} + W_{co} \circ C_t + b_o\right)$$
$$H_t = o_t \circ \tanh(C_t)$$

- Is the dog looking to take a shower?

	$Human_{consensus}$	$Human_{random}$	$GRNN_X$	$GRNN_{all}$
Human Evaluation				
Bing	2.50	2.36	1.38	**1.81**
COCO	2.50	2.40	1.62	**1.97**
Flickr	2.33	2.28	1.27	**1.58**
Bing	87.3	83.6	**12.4**	11.0
COCO	86.1	83.8	**13.8**	14.3
Flickr	84.5	83.4	**10.0**	9.8
Bing	62.0	59.0	**16.0**	15.6
COCO	60.7	58.5	**18.2**	18.3
Flickr	59.5	58.0	**14.1**	14.0
Bing	63.0	57.5	**11.5**	10.7
COCO	61.0	56.9	**12.3**	12.4
Flickr	62.0	57.2	**9.4**	9.2

BLEU

MET.

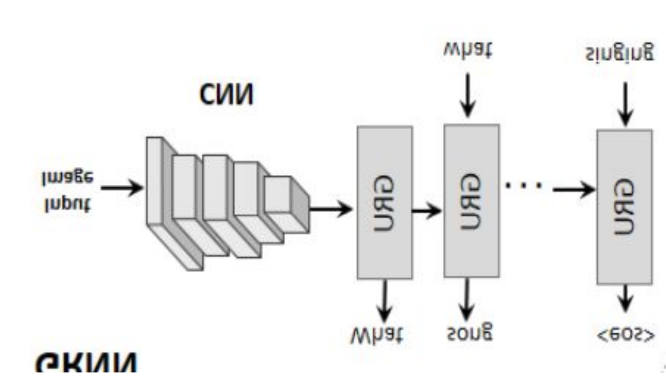
ΔBLEU

HUMAN

- How long did it take to make that ice sculpture?
- Is the dog looking to take a shower?
- Was this explosion an accident?

GRNN

- Where was this picture taken ?
- Why is this dog in a bathroom ?
- What caused this explosion ?



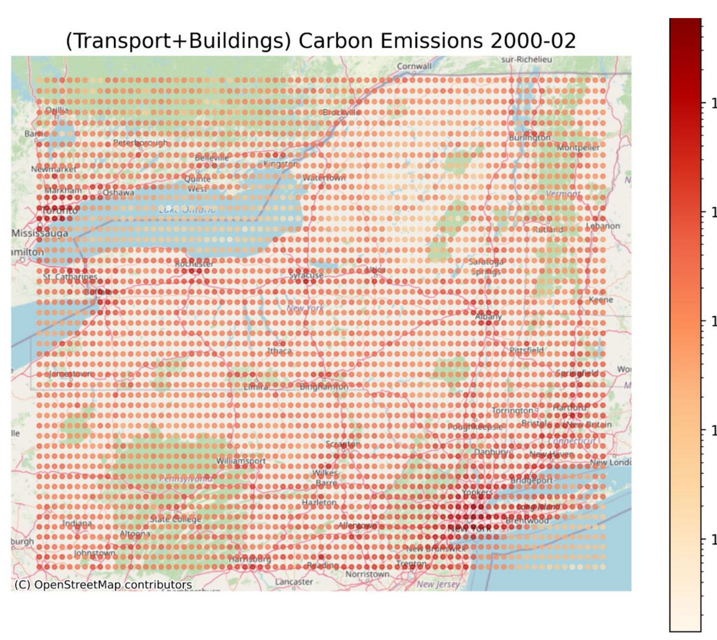
Spatiotemporal ConvLSTM Modeling for Carbon Emission Prediction in Digital Twin Cities

Nikita Dahiya, Jonathan Zhang, Aolei Cao, Jacob Fernandez, Department of ORIE & Department of Systems Engineering, Cornell University

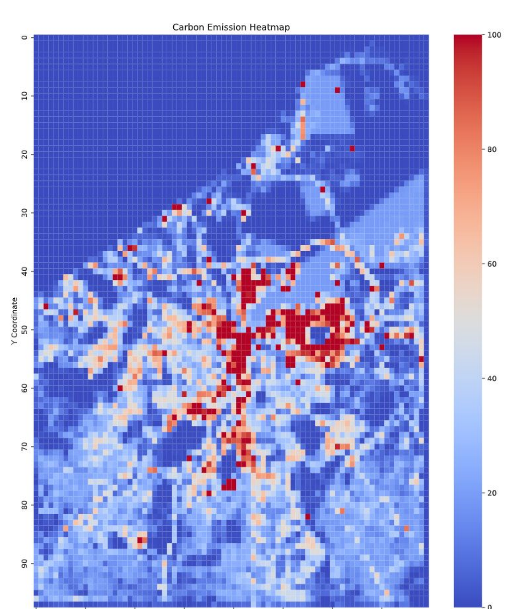
INTRODUCTION

Cities contribute significantly to global carbon emissions, resulting in environmental challenges. Urban planners and policymakers need accurate, real-time emission forecasts (nowcasting) to support sustainable decision-making, but traditional models often fail to capture the interactions between spatial and temporal emission patterns.

SAMPLE DATA



Grid Data NYC Dataset



Heatmap NYC Dataset

RESEARCH CHALLENGES

- **Capturing complex spatial-temporal interactions.**
  - Ensuring the network effectively learns highly localized spikes, such as rush hour peaks, while retaining long-range dependencies is a challenge.
- **Modeling nonlinear interactions among various influencing factors.**
  - Emissions are driven by a combination of meteorological, land use, and socioeconomic factors that interact in nonlinear ways.
- **Enhancing model interpretability through visualization techniques to build stakeholder trust.**
  - Interpreting how specific input features or regions drive forecast outcomes is difficult.

PRIOR MODELS

Conv-LSTM

$$i_t = \sigma\big(W_{xi} * X_t + W_{hi} * H_{t-1} + W_{ci} \circ C_{t-1} + b_i\big)$$
$$f_t = \sigma\big(W_{xf} * X_t + W_{hf} * H_{t-1} + W_{cf} \circ C_{t-1} + b_f\big)$$
$$\tilde{C}_t = \tanh\big(W_{x\tilde{C}} * X_t + W_{h\tilde{C}} * H_{t-1} + b_{\tilde{C}}\big)$$
$$C_t = f_t \circ C_{t-1} + i_t \circ \tilde{C}_t$$
$$o_t = \sigma\big(W_{xo} * X_t + W_{ho} * H_{t-1} + W_{co} \circ C_t + b_o\big)$$
$$H_t = o_t \circ \tanh(C_t)$$

- **Tensors:** X_t (input), H_{t-1}, H_t (hidden), C_{t-1}, C_t (cell)
- **Gates:** i_t (input), f_t (forget), o_t (output), $\tilde{C}_t$ (candidate)
- **Weights & biases:** W_{x*}, W_{h*} (conv kernels), W_{c*} (peepholes), b_* (biases)
- **Activations & ops:** σ (sigmoid), $\tanh$; "$*$" (convolution), "$\circ$" (Hadamard)

Vision Transformer

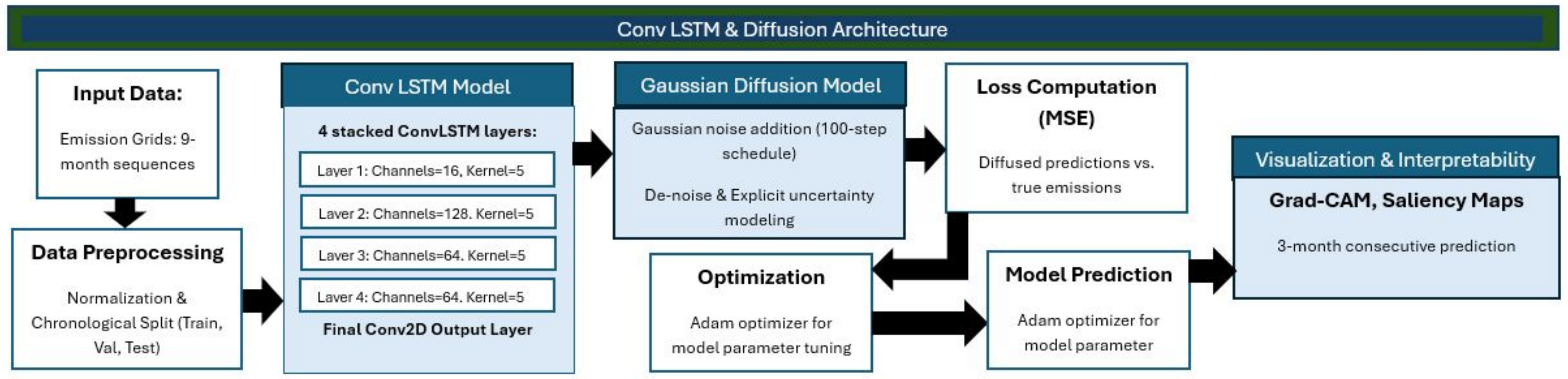
$$\mathrm{Attention}(Q,K,V) = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V$$
$$\mathrm{head}_i = \mathrm{Attention}(QW_{Q_i},\ KW_{K_i},\ VW_{V_i}) \quad (i = 1,\dots,h)$$
$$\mathrm{MHA}(Q,K,V) = \mathrm{Concat}(\mathrm{head}_1,\dots,\mathrm{head}_h)\,W_O$$
$$X_l' = X_{l-1} + \mathrm{MHA}\big(\mathrm{LN}(X_{l-1}),\ \mathrm{LN}(X_{l-1}),\ \mathrm{LN}(X_{l-1})\big)$$
$$X_l = X_l' + \mathrm{MLP}\big(\mathrm{LN}(X_l')\big)$$

- **Layer I/O:** X_{l-1} (input), X_l' (post-attn), X_l (output)
- **Attention inputs:** $Q, K, V \in \mathbb{R}^{N\times d}$; d (per-head dim), h (# heads)
- **Projections:** $W_{Q_i}, W_{K_i}, W_{V_i}$ (heads), W_O (output)
- **Components:** LN (layer-norm), MLP (feed-forward), $\mathrm{softmax}$, Concat

Self-Attention

$$\mathrm{Attention}(Q,K,V) = \mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V$$

- **Inputs:** $Q, K \in \mathbb{R}^{N\times d}$, $V \in \mathbb{R}^{N\times d_v}$
- **Dims:** d (q/k dim), d_v (v dim)
- **Core op:** $\mathrm{softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V$ with scale $1/\sqrt{d}$
- **Multi-head:** Concatenate per-head outputs via Concat



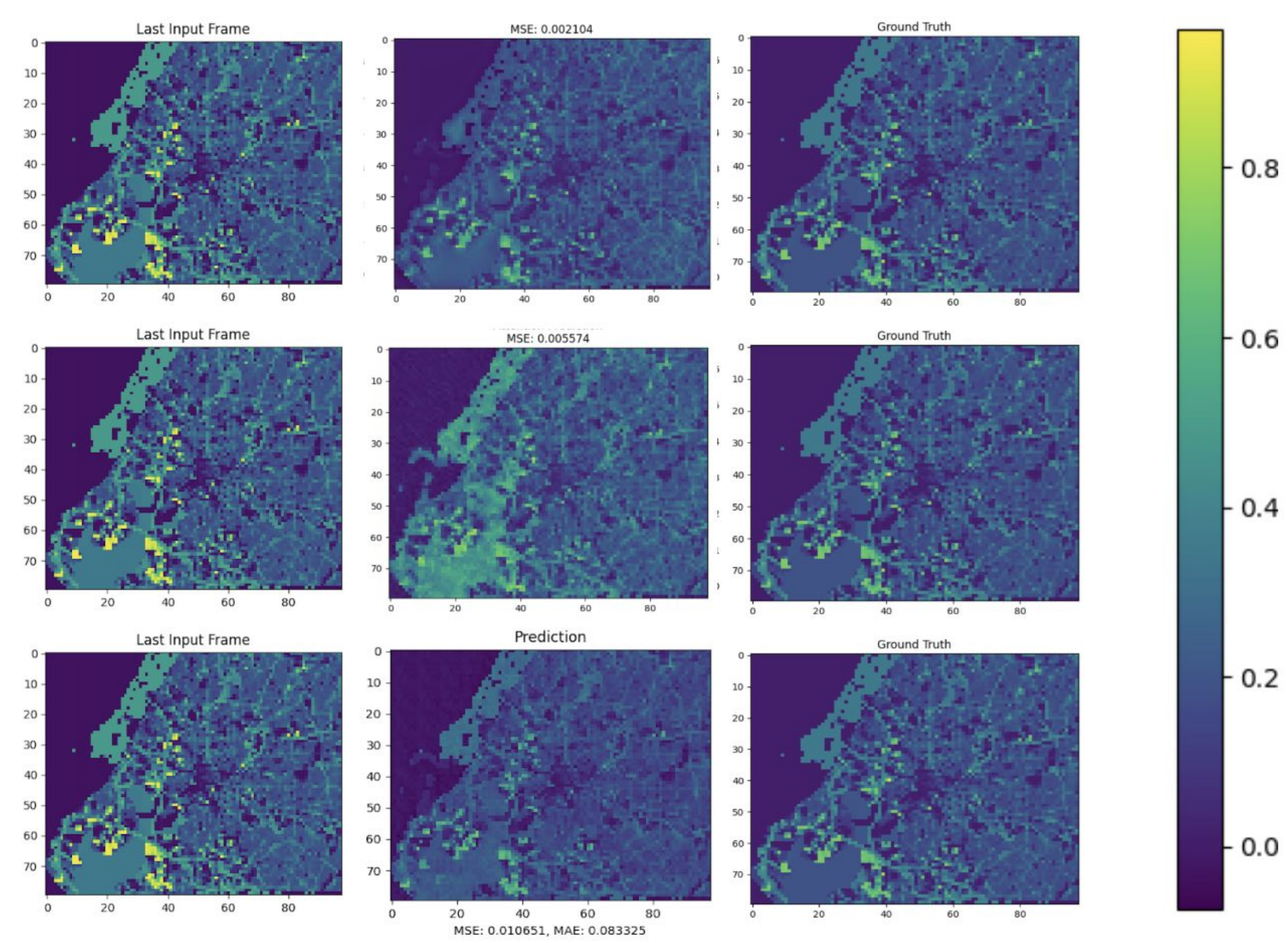
RESULTS

LAST INPUT | PREDICTED | REAL

ConvLSTM

Attention

ViT



CONCLUSION

Model	Dataset	Average MSE
VIT	NYC	.0281
VIT	Baltimore	.0106
Attention	NYC	.0077
Attention	Baltimore	.0055
ConvLSTM	NYC	.0193
ConvLSTM	Baltimore	.0021

Our comparative analysis reveals that the **vanilla ConvLSTM model** on the **Baltimore dataset** achieved superior performance with the lowest average MSE of 0.0021, outperforming the Vision Transformer and ConvLSTM + Attention models.

NEXT STEPS

- **Construct Gaussian Diffusion model using each of the prior models**
- **Conditional latent Diffusion model**
  - Using VAE to get latent space
  - Using cross-attention to realize conditional diffusion

ACKNOWLEDGEMENTS

We would like to thank Professor Gao and Yishuo Jiang for their guidance and support throughout the project's timeline.