Advancing Interdisciplinary Sequence Generation Research Through Natural Language Processing

My research develops long-form text generation algorithms with a focus on coherence, efficiency, and model transparency. One of the key strengths of my research is its general applicability to other domains that also exhibit a sequential nature, such as genomics, speech, and images. I have explored these connections through interdisciplinary research projects and have developed new methods for generating high-quality text and other forms of data in these fields.

Long-Form Coherence Applied to Genomics My research on long-form coherence aims at generating coherent documents with self-consistency [1, 2]. Coherence is important not only in text generation, but also in other domains where the outputs take the form of long discrete sequences. For example, in genome sequences, there are crucial long-range interactions among distant genes [3]. Modeling these interactions is challenging since genome sequences are extremely long. To this end, I applied my research on long-form coherence to Covid genome modeling [4] in collaboration with Nvidia and Argonne National Lab. Our proposed long-form model better captures high-level phenomena such as open reading frames (ORFs) (Figure 1 left). Furthermore, this model generates new sequences that cover different lineages including all existing variants (Figure 1 right). Our work has won an **ACM Gordon Bell Covid Prize**. This direction presents huge potential impacts on life science and healthcare. For example, a prediction of how the genome of a certain virus evolves will assist in vaccine development.

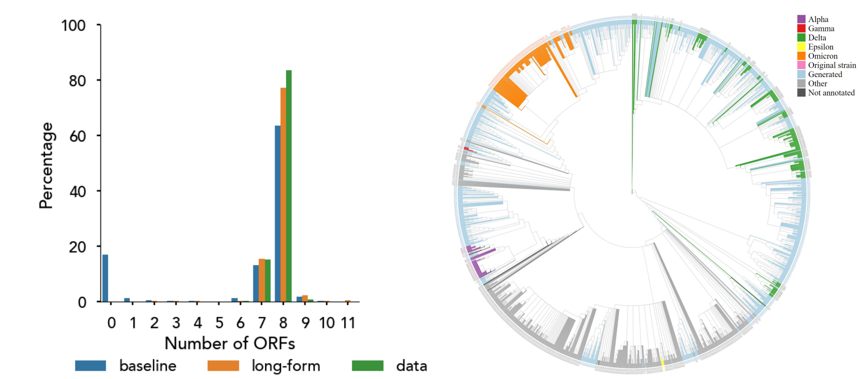

Figure 1: Long-form modeling of genomes improves the modeling of open reading frames (ORFs). Left: Comparison of statistics for the ORFs. The long-form model (orange) matches the data (green) much better than the baseline (blue). Right: The overlay of the generated sequences (light blue) with the real phylogenetic tree shows good coverage.

Efficient Systems Applied to Vision My research on efficient systems aims at handling the scale, complexity, and real-time requirements of long-form text generation [5, 6]. This line of research can be applied to large inputs of text as well as other modalities such as images.

In *Image-to-Markup Generation with Coarse-to-Fine Attention* [7], I proposed to use a cascade of attention mechanisms to gradually zoom into the input image. The first attention mechanism is used to identify the salient regions using a coarse-grained image, which are then passed to the second attention to generate the output using a fine-grained image (Figure 2). By applying fast coarse-to-fine inference, these models can scale to much larger inputs, such as large medical images.

Other Insights from Text Generation Insights gained from text generation research can be applied to other domains. For example, in *Markup-to-Image Diffusion Models with Scheduled Sampling* [8], I used a diffusion model to generate molecules and sheet music. Diffusion models generate images by applying a *sequence* of denoising operations to an initial noise image [9]. We found that it has similar issues as the *exposure bias* issue in text generation [10]: the model fails to correct its own

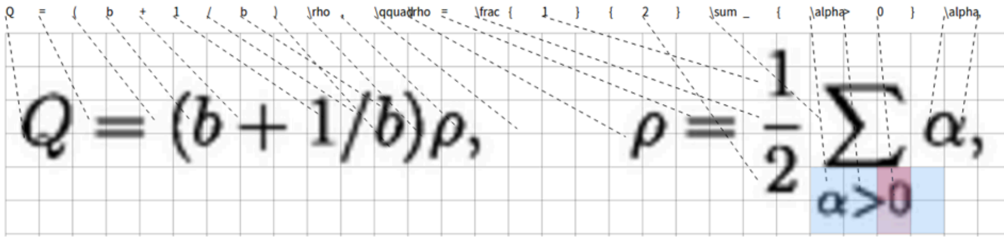

Figure 2: A coarse-to-fine model to process large input images. When producing each symbol, the model first selects a subset of the image (marked in blue) using a coarse-grained feature map, and then only uses finer-grained features inside the selected region.

mistakes during generation because it has never seen them during training. To mitigate these issues, I adapted the *scheduled sampling* algorithm, originally used for text generation and imitation learning [11, 12], to diffusion training. This approach leads to promising results in molecule and sheet music generation (Figure 3).

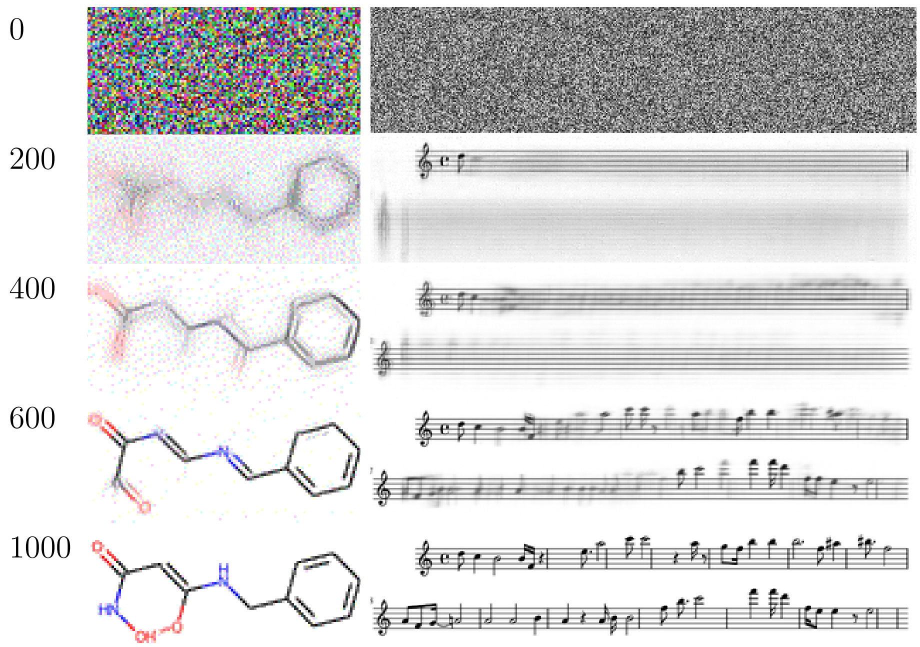

Figure 3: The generation process of diffusion models with scheduled sampling. The numbers are the number of diffusion steps. Left: generating a molecule. Right: generating sheet music.

Workshop Proposal In light of the significant advances in large language models, I propose to organize a workshop, *Sequence Generation for Science and Beyond*. This workshop will bring together experts in machine learning, NLP, computational biology, chemistry, speech, music, economics, etc., to stimulate vibrant discussions about applying sequence generation research to science and other domains. My past experience organizing the *Efficient Natural Language and Multimodal Models Workshop* will help me organize this workshop successfully. I believe that the *Sequence Generation for Science and Beyond* workshop will provide a platform for researchers to engage in interdisciplinary collaborations and foster the development of novel applications of sequence generation research in various domains.

Future Plans As a professor, I will continue my research on long-form generation, with a focus on developing coherent, efficient, and transparent models that enable many exciting new applications in different domains. I am particularly interested in collaborating with other groups at the University of British Columbia on interdisciplinary research that combines natural language processing with other fields. For example, my work on sequence generation can be used to solve problems in biology, economics, linguistics, speech, and education. I am excited to be a part of the University of British Columbia community and contribute to the advancement of natural language processing and related fields through my work.

References

[1] **Yuntian Deng**, Volodymyr Kuleshov, and Alexander Rush. 2022. Model criticism for long-form text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, Abu Dhabi, UAE. Association for Computational Linguistics.

[2] **Yuntian Deng**, Anton Bakhtin, Myle Ott, Arthur Szlam, and Marc'Aurelio Ranzato. 2020. Residual energy-based models for text generation. In *International Conference on Learning Representations*.

[3] Amartya Sanyal, Bryan R Lajoie, Gaurav Jain, and Job Dekker. 2012. The long-range interaction landscape of gene promoters. *Nature*, 489(7414):109–113.

[4] Maxim Zvyagin*, Alexander Brace*, Kyle Hippe*, **Yuntian Deng***, Bin Zhang, Cindy Orozco Bohorquez, Austin Clyde, Bharat Kale, Danilo Perez-Rivera, Heng Ma, Carla M. Mann, Michael Irvin, J. Gregory Pauloski, Logan Ward, Valerie Hayot, Murali Emani, Sam Foreman, Zhen Xie, Diangen Lin, Maulik Shukla, Weili Nie, Josh Romero, Christian Dallago, Arash Vahdat, Chaowei Xiao, Thomas Gibbs, Ian Foster, James J. Davis, Michael E. Papka, Thomas Brettin, Rick Stevens, Anima Anandkumar, Venkatram Vishwanath, and Arvind Ramanathan. 2022. Genslms: Genome-scale language models reveal sars-cov-2 evolutionary dynamics. *bioRxiv*. *Equal Contribution. **ACM Gordon Bell Covid Prize**.

[5] **Yuntian Deng** and Alexander Rush. 2020. Cascaded text generation with markov transformers. In *Advances in Neural Information Processing Systems*, volume 33, pages 170–181. Curran Associates, Inc.

[6] **Yuntian Deng** and Alexander Rush. 2021. Sequence-to-lattice models for fast translation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3765–3772, Punta Cana, Dominican Republic. Association for Computational Linguistics.

[7] **Yuntian Deng**, Anssi Kanervisto, Jeffrey Ling, and Alexander M. Rush. 2017. Image-to-markup generation with coarse-to-fine attention. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 980–989. PMLR.

[8] **Yuntian Deng**, Noriyuki Kojima, and Alexander M. Rush. 2022. Markup-to-image diffusion models with scheduled sampling. *arXiv preprint*, arXiv:1710.03740.

[9] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc.

[10] Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2015. Sequence level training with recurrent neural networks. *arXiv preprint arXiv:1511.06732*.

[11] Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

[12] Stephane Ross, Geoffrey Gordon, and Drew Bagnell. 2011. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, volume 15 of *Proceedings of Machine Learning Research*, pages 627–635, Fort Lauderdale, FL, USA. PMLR.