Poster: Towards Network Model Generalization using *Strategic* Data Collection

Benjamin Hoffman
ETH Zürich
bhoffman@ethz.ch

Alexander Dietmüller
ETH Zürich
adietmue@ethz.ch

Laurent Vanbever
ETH Zürich
lvanbever@ethz.ch

ABSTRACT

Essential networking applications, such as video streaming, require accurate network models to estimate current and future network states (e.g., is the network congested?). Due to the complexity of today's networks and the subsequent difficulty of this modeling task, Machine Learning (ML)-based approaches have emerged as an alternative to first-principle modeling methods. However, proposed ML algorithms suffer from a generalization crisis: they often fail to perform in deployments outside of their training environment. Moreover, simple solutions such as naively training on *more data* do not guarantee improved generalization performance.

We propose an interpretable approach to improving model generalization by focusing on the *quality* of a dataset over sample quantity already *during* data collection. Notably, our approach's interpretability allows us to reason on which environments to prioritize at the data acquisition stage. To this end, we investigate the impact of dataset metrics such as Round Trip Time (RTT) and throughput on both in-distribution (ID) and out-of-distribution (OOD) model performance. Our results suggest that *strategically* performing data collection in environments with broader state-space coverage in areas of higher RTT and lower throughput is key to achieving improved model generalization and OOD performance.

CCS CONCEPTS

• **Networks → Network dynamics**; • **Computing methodologies → Machine learning**; **Artificial intelligence**.

KEYWORDS

Network Modeling, Machine Learning, Model Generalization

ACM Reference Format:

Benjamin Hoffman, Alexander Dietmüller, and Laurent Vanbever. 2025. Poster: Towards Network Model Generalization using *Strategic* Data Collection. In *ACM SIGCOMM 2025 Conference (SIGCOMM Posters and Demos '25), September 8–11, 2025, Coimbra, Portugal.* ACM, New York, NY, USA, 3 pages. https://doi.org/10.1145/3744969.3748407

1 INTRODUCTION

Accurately modeling a network's state is a fundamental challenge across many networking applications. For example, to provide robust performance, video streaming services require models for real-time estimation of the current network state, its propagation over time, and the subsequent *Transmission Time Prediction* of a video chunk. This modeling problem is inherently difficult in the networking context: we must contend with a high-dimensional modeling space that is successively growing ever more complex as new applications and protocols continue to emerge, while the space of observable signals has remained almost unchanged.

In response, research has increasingly shifted away from pursuing first-principle modeling methods in favor of using ML-based approaches in many applications, including video streaming [2, 3, 6, 7, 9, 10, 14], congestion control [1, 8, 13, 15], network traffic optimization [5], routing [12], and network simulation [16].

The Generalization Problem Applying learning-based methods has presented a new challenge: to perform well, ML models require training data or simulation environments that are representative of their real-world deployments. This is particularly challenging in the networking context due to the Internet's dynamic, heavy-tailed nature, limited centralized observability, and subsequent hampered access to representative datasets. As a result, learned models often fail to generalize and perform poorly outside of their training environment, in particular when trained on synthetic data.

Various attempts have been made to address the challenge of model generalization in the context of video streaming. On one hand, methods such as *Plume* [10], *CausalSim* [3] and *Memento* [6] focus on improving existing datasets and simulation environments, either through clever sampling or by learning a causal model to mitigate biases in collected traces. In contrast, *Puffer* [14] advocates for real-world learning *in situ* and against pursuing model generalization across different deployment environments.

Strategic Data Collection We argue that the challenge of model generalization should be addressed at the source, i.e., by already focusing on the *(i) quality* of a dataset *(ii) during* the data collection stage. Similarly to *Plume* and *Memento*, we observe that simply using more training data does not necessarily improve model performance or generalization. However, instead of reasoning on how to select representative samples from an existing dataset, we focus on how to select representative (real-world) environments for data collection. To this end, we investigate how the distribution of RTT and throughput of a training set impacts both ID and OOD model performance. To achieve broad diversity in our training sets, we collect real-world traffic data using servers and clients globally.

Our initial results show that training on data collected in real-world environments with a broader state-space coverage in areas of higher RTT and lower throughput leads to improved generalization across OOD environments, without sacrificing ID performance. This suggests that RTT and throughput can be interpreted as proxy metrics for the underlying diversity of an environment and that prioritizing such diverse environments during data collection is key to improving the generalization of network traffic models.

SIGCOMM Posters and Demos '25, September 8–11, 2025, Coimbra, Portugal

ACM ISBN 979-8-4007-2026-0/2025/09.
https://doi.org/10.1145/3744969.3748407

2 PRELIMINARY EVALUATION

With video streaming being the most prevalent internet application (65% of all traffic in 2023 [11]), we focus on ML-based models for chunk *Transmission Time Prediction* to demonstrate the effectiveness of our approach. As the transmission time is a function of the current network conditions, its prediction requires estimating the latent network state. This estimation, in turn, is at the core of many other learning-based methods across networking applications, lending relevance to our approach beyond video streaming.

Model and Data Collection We use an encoder to estimate the latent network state from history and a decoder to predict the chunk transmission time from the action and estimated latent state. For data collection, we customize and employ the video streaming data collection infrastructure from *Puffer*. To achieve a broad diversity of environments, we utilize *NetUnicorn* [4] to deploy our infrastructure to data centers worldwide and stream to real clients across the globe.

Evaluation We evaluate the effectiveness of our approach by training one model per dataset, each collected in a distinct environment: *Ohio* (server hosted in an AWS data center in Ohio, USA) and *Zurich* (server hosted at ETH Zürich, Switzerland). In both cases, we stream to real clients globally. We then investigate the differences in the dataset distributions, as well as their impact on ID and OOD model performance compared to simply increasing the dataset size, highlighting the importance of strategic data collection.

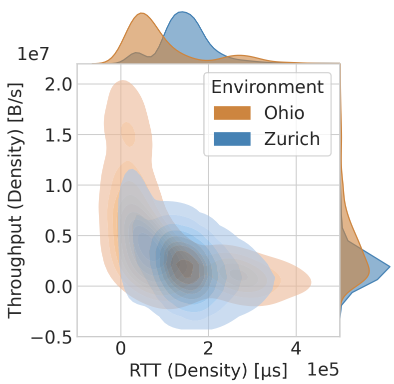

Figure 1: Joint and marginal Kernel Density Estimates (KDE) of RTT and throughput. The Zurich environment covers areas of higher RTT and lower throughput more broadly.

Environment Analysis We analyze the underlying conditions of our environments based on the distribution of RTT and throughput that their data exhibits. The joint and marginal Kernel Density Estimates (KDE) show that the Zurich environment covers areas of higher RTT and lower throughput more broadly (Figure 1). The RTT mode is shifted to the right (higher RTT) of the Ohio-environment, and the throughput coverage is shifted slightly below (lower throughput). In contrast, the Ohio-environment has either high throughput or high RTT, lacking broad coverage in between. These distinct distributions highlight how we can use RTT and throughput as proxy metrics to describe the underlying conditions of an environment, as well as the data we can collect in it.

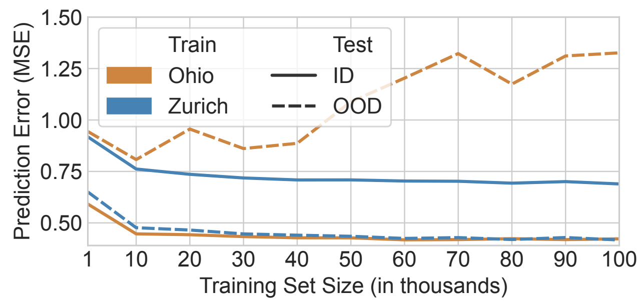

Figure 2: The Zurich-trained model (blue) generalizes well to the Ohio environment, while the Ohio-trained model (orange) fails to generalize to the Zurich environment.

Model Performance We evaluate the chunk *Transmission Time Prediction* performance of our trained models on test sets from their own training environment (ID) and the respective other environment (OOD) across increasing training set sizes (Figure 2). The difference in OOD performance highlights the importance of *strategically* collecting data instead of simply collecting *more* data. The performance of the Zurich-trained model (blue) improves with increasing training set size, both in the ID and OOD environments. Notably, it rapidly converges to the performance of the Ohio-trained model in the Ohio environment, i.e., it generalizes well. In contrast, the performance of the Ohio-trained model (orange) *decreases* in the OOD environment, i.e., the model fails to generalize.

While we can see diminishing returns across all experiments, this significant difference in OOD performance suggests that the underlying differences we observed in the training sets impact how well their respective model generalizes. More particularly, it indicates that the broader coverage in areas of higher RTT and lower throughput in the Zurich dataset leads to improved generalization of a model trained on it. Our findings suggest that higher RTT and lower throughput are the results of a variety of variables in our network exhibiting higher diversity. With this diversity in the underlying network conditions largely lacking in the Ohio environment, adding more training samples only reinforces the model's ID performance, while decreasing its OOD performance, i.e., the model subsequently overfits and fails to generalize.

3 CONCLUSION AND FUTURE WORK

We have presented an interpretable approach to improving model generalization through strategic data collection. To address this challenge, we focus on how to select representative (real-world) environments for data collection using networking-specific metrics and domain knowledge. In future work, we will further explore the potential and limitations of our approach. In particular, we will investigate whether RTT and throughput remain meaningful proxy metrics for choosing training data across other learning-based prediction tasks and environments. Furthermore, we will examine how our approach's performance compares to methods focused on sampling representative training data from existing datasets, and lastly, whether our insights allow us to generate representative *synthetic* datasets that yield equal generalization performance.

REFERENCES

[1] Soheil Abbasloo, Chen-Yu Yen, and H. Jonathan Chao. 2020. Classic Meets Modern: a Pragmatic Learning-Based Congestion Control for the Internet. In *Proceedings of the Annual Conference of the ACM Special Interest Group on Data Communication on the Applications, Technologies, Architectures, and Protocols for Computer Communication (SIGCOMM '20)*. Association for Computing Machinery, New York, NY, USA, 16 pages. https://doi.org/10.1145/3387514.3405892

[2] Zahaib Akhtar, Yun Seong Nam, Ramesh Govindan, Sanjay Rao, Jessica Chen, Ethan Katz-Bassett, Bruno Ribeiro, Jibin Zhan, and Hui Zhang. 2018. Oboe: auto-tuning video ABR algorithms to network conditions. In *Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication (SIGCOMM '18)*. Association for Computing Machinery, New York, NY, USA, 44–58. https://doi.org/10.1145/3230543.3230558

[3] Abdullah Alomar, Pouya Hamadanian, Arash Nasr-Esfahany, Anish Agarwal, Mohammad Alizadeh, and Devavrat Shah. 2023. CausalSim: A Causal Framework for Unbiased Trace-Driven Simulation. In *20th USENIX Symposium on Networked Systems Design and Implementation (NSDI 23)*. USENIX Association, Boston, MA, 1115–1147. https://www.usenix.org/conference/nsdi23/presentation/alomar

[4] Roman Beltiukov, Wenbo Guo, Arpit Gupta, and Walter Willinger. 2023. In Search of NetUnicorn: A Data-Collection Platform to Develop Generalizable ML Models for Network Security Problems. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security (CCS '23)*. Association for Computing Machinery, New York, NY, USA, 2217–2231. https://doi.org/10.1145/3576915.3623075

[5] Li Chen, Justinas Lingys, Kai Chen, and Feng Liu. 2018. AuTO: scaling deep reinforcement learning for datacenter-scale automatic traffic optimization. In *Proceedings of the 2018 Conference of the ACM Special Interest Group on Data Communication (SIGCOMM '18)*. Association for Computing Machinery, New York, NY, USA, 191–205. https://doi.org/10.1145/3230543.3230551

[6] Alexander Dietmüller, Romain Jacob, and Laurent Vanbever. 2024. On Sample Selection for Continual Learning: A Video Streaming Case Study. *SIGCOMM Comput. Commun. Rev.* 54, 2 (Aug. 2024), 10–35. https://doi.org/10.1145/3687234.3687237

[7] Alexander Dietmüller, Siddhant Ray, Romain Jacob, and Laurent Vanbever. 2022. A new hope for network model generalization. In *Proceedings of the 21st ACM Workshop on Hot Topics in Networks (HotNets '22)*. Association for Computing Machinery, New York, NY, USA, 152–159. https://doi.org/10.1145/3563766.3564104

[8] Nathan Jay, Noga Rotman, Brighten Godfrey, Michael Schapira, and Aviv Tamar. 2019. A Deep Reinforcement Learning Perspective on Internet Congestion Control. In *Proceedings of the 36th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 97)*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.). PMLR, 3050–3059. https://proceedings.mlr.press/v97/jay19a.html

[9] Hongzi Mao, Ravi Netravali, and Mohammad Alizadeh. 2017. Neural Adaptive Video Streaming with Pensieve. In *Proceedings of the Conference of the ACM Special Interest Group on Data Communication (SIGCOMM '17)*. Association for Computing Machinery, New York, NY, USA, 197–210. https://doi.org/10.1145/3098822.3098843

[10] Sagar Patel, Junyang Zhang, Nina Narodystka, and Sangeetha Abdu Jyothi. 2024. Practically High Performant Neural Adaptive Video Streaming. *Proc. ACM Netw.* 2, CoNEXT4, Article 30 (Nov. 2024), 23 pages. https://doi.org/10.1145/3696401

[11] Sandvine Corporation. 2023. Video Permeates, Streaming Dominates. *The Global Internet Phenomena Report* January (2023), 14–15. https://www.sandvine.com/hubfs/Sandvine_Redesign_2019/Downloads/2023/reports/SandvineGIPR2023.pdf

[12] Asaf Valadarsky, Michael Schapira, Dafna Shahaf, and Aviv Tamar. 2017. Learning to Route. In *Proceedings of the 16th ACM Workshop on Hot Topics in Networks (HotNets '17)*. Association for Computing Machinery, New York, NY, USA, 185–191. https://doi.org/10.1145/3152434.3152441

[13] Keith Winstein and Hari Balakrishnan. 2013. TCP ex machina: computer-generated congestion control. In *Proceedings of the ACM SIGCOMM 2013 Conference on SIGCOMM (SIGCOMM '13)*. Association for Computing Machinery, New York, NY, USA, 123–134. https://doi.org/10.1145/2486001.2486020

[14] Francis Y. Yan, Hudson Ayers, Chenzhi Zhu, Sadjad Fouladi, James Hong, Keyi Zhang, Philip Levis, and Keith Winstein. 2020. Learning in situ: a randomized experiment in video streaming. In *17th USENIX Symposium on Networked Systems Design and Implementation (NSDI 20)*. USENIX Association, Santa Clara, CA, 495–511. https://www.usenix.org/conference/nsdi20/presentation/yan

[15] Lei Zhang, Kewei Zhu, Junchen Pan, Hang Shi, Yong Jiang, and Yong Cui. 2020. Reinforcement Learning Based Congestion Control in a Real Environment. In *2020 29th International Conference on Computer Communications and Networks (ICCCN)*. 1–9. https://doi.org/10.1109/ICCCN49398.2020.9209750

[16] Qizhen Zhang, Kelvin K. W. Ng, Charles Kazer, Shen Yan, João Sedoc, and Vincent Liu. 2021. MimicNet: fast performance estimates for data center networks with machine learning. In *Proceedings of the 2021 ACM SIGCOMM 2021 Conference (SIGCOMM '21)*. Association for Computing Machinery, New York, NY, USA, 287–304. https://doi.org/10.1145/3452296.3472926