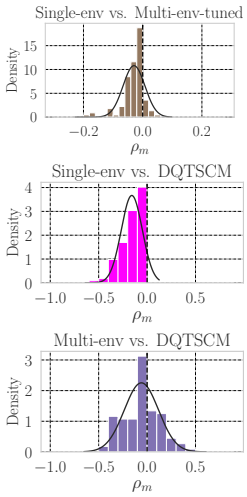
Thank you for carefully reviewing the manuscript and finding our idea *"novel"* and *"interesting"*.

Major Changes: (i) As suggested by R#3, we have added a new method in multi-env regime, which fine-tunes the policy on the fly. In this method, we start with the multi-env policy and improve with a very few training steps to calibrate the policy for every specific intersection. Following that, the maximum and average of multi-env gap compared to single-env decreased significantly after 200 training episodes (instead of 100,000 episodes when trained from scratch) such that we got to 5% gap in average with respect to the single-env regime. After 1000 training steps this gap decreased to 3%. (See the first row of the figures). (ii) We added a new benchmark, DQTSC-M, (R#1 and R#4). As it is shown in the second row of the figure, single-env model obtains 15% smaller ATT in average and outperforms DQTSC-M in all 112 cases. Compared to the multi-env regime (the third row of the figures), AttendLight obtains 4% smaller ATT compared to DQTSC-M. (iii) Added the result of AttendLight with mean-state instead of state-attention model (R#2 and R#3).

R#1: Thank you for carefully reviewing the paper and finding it "novel".



- "*too application focused*": There is a growing interest in application-focused papers in NeurIPS as ML/RL is becoming common in real-world. There is an "application" subject area where $\sim$ 20% of accepted papers fall under this category [1] in the NeurIPS 2019. Besides, relevant papers on traffic control are published in NeurIPS. For example, see [2]. Moreover, AttendLight framework by itself is of independent interest and can be applied to various domains such as matching, routing, etc. We will add a discussion about this.
- "*transfer learning baselines*": In the current literature of TSCP, transfer learning is used for reducing the training time and it usually results in the same or even worse performance. For example, FRAP uses transfer learning to train policies for new intersections which achieves the same quality solution in comparison to the training from scratch (see, Figure 10 in [30]). Instead, we have compared AttendLight with the existing fully trained from scratch RL-based alternatives and we don't expect that adding transfer learning will affect the solution quality of the baselines. We will verify the validity of this claim in our camera-ready version.
- "*training costs*": Some training details are explained in appendix A.3. We will add more details in the final version.
- "*learn generalizable intersection control efficiently*": We added a fine-tuning mechanism to further improve the generalizability of the multi-env results. Following this makes the multi-env regime quite efficient in terms of training-time. See the Major change comment.

R#2: Thank you for your positive feedback and great suggestions.

- "*ablation studies*": We added one ablation study to show the effect of state-attention (as suggested by R#3). The idea behind having mean query was to learn the importance of each lane-traffic compared to the average traffic per lane. We will clarify these points in final revision and add experiments to better justify the role of components.
- "*distribution of ρ_m* ": As it can be seen from Figure 4, there is no noticeable difference between these two groups. We will add the separated figures of ρ_m for training and testing to the appendix.
- "*pattern in the state*": This is a great suggestion. We will try to add a section about this to the final version.

R#3: Thank you for carefully reviewing the paper and your great suggestions.

- "*CO2 emissions*": CityFlow does not provide CO2 statistics, but we will construct a CO2 emission metric based on the traffic flow, and will add the reduction of CO2 emission to the appendix.
- "*fine-tuning*": This is a great suggestion. After fine-tuning for 200 episodes, the average gap drops to 3% (from 13%) and the worst-case gap is decreased to 21%. See the major changes.
- "*ablation Study*": We re-ran the single-env regime for all cases with the mean-state. The results show that mean-state obtains 5% larger ATT (in average) than that of with state-attention. We will add this result to the appendix.

R#4: We appreciate your constructive feedback and finding the paper "novel".

- "*primary motivation*": To reduce the degradation observed in mutli-env regime, we added a fine-tuning which helps the multi-env regime to fine-tune the policy quickly. See major changes.
- "*conflicting statements...*": In average, multi-env performs worse than single-env; although, there are some special cases that the multi-env model outperforms the single-env model. We will make sure to remove the confusion.
- "*less restrictive baseline*": Thanks for introducing these papers. We added DQTSC-M. See the major change above.
- "*why single-env AttendLight outperforms FRAP*": First, we would like to mention that AttendLight supports both single-env and multi-env by design. We believe that the superior performance of single-env model compared to FRAP originates from two attention model.
- "*city-wide control performance...*" After using the quick fine-tuning, the ATT gap of the intersection that you mentioned is now 5%, after 200 training-episodes. Further training to 1000 episodes decreases the gap to 3%.

[1] What we learned from neurips 2019 data. `https://medium.com/@NeurIPSConf/what-we-learned-from-neurips-2019-data-111ab996462c`. Accessed: 2020-08-12.

[2] Silvia Richter, Douglas Aberdeen, and Jin Yu. Natural actor-critic for road traffic optimisation. In *Advances in neural information processing systems*, pages 1169–1176, 2007.