CLAIMS:

1. A computer-implemented method for determining routing using reinforcement learning (RL), comprising:

initializing a state of an RL model based on a routing solution comprising one or more routes for routing one or more vehicles through a plurality of locations, wherein the RL model comprises a plurality of improvement actions for applying to the state according to a policy to reduce a cost of the routing solution;

applying one or more of the plurality of improvement actions to the state according to the policy to reduce the cost of the routing solution and to obtain updated routing solutions until a predetermined condition is satisfied;

in response to the predetermined condition being satisfied, applying a perturbation action to obtain a perturbed routing solution and feeding the perturbed routing solution back to the RL model for the RL model to perform the applying one or more of the plurality of improvement actions according to the policy; and

determining a routing solution with a minimum cost from the updated routing solutions.

2. The method of claim 1, wherein the plurality of improvement actions for applying to the state according to the policy to reduce the cost of the routing solution comprise one or more of the following:

changing an order of at least two of the plurality of locations in one of the one or more routes; and

moving a location from one of the one or more routes to another one of the one or more routes.

3. The method of any of claims 1-2, wherein applying the perturbation action comprises one or more of the following:

shuffling at least two of the plurality of locations among at least two of the routes;

concatenating at least two of the routes; and

breaking at least one of the routes into multiple routes.

4. The method of any of claims 1-3, wherein each of the updated routing solutions is subject to one or more constraints, the constraints comprising one or more of the following:

time constraint;

travel distance constraint;

vehicle capacity constraint; and

power expense constraint.

5. The method of any of claims 1-4, wherein each of the plurality of locations is associated with one or more features comprising one or more of the following:

position information;

demand information;

position information of one or more neighboring locations along a corresponding route of the one or more routes; and

a remaining capacity of a vehicle leaving the each of the plurality of locations.

6. The method of claim 5, wherein the state comprises the one or more features associated with the plurality of locations.

7. The method of any of claims 1-6, wherein the state comprises information of one or more improvement actions previously applied to the state.

8. The method of any of claims 1-7, further comprising:

adjusting the policy based on a reward received by applying one or more of the plurality of improvement actions, wherein the reward is calculated based on cost changes after applying the one or more improvement actions.

9. The method of claim 8, wherein the reward for applying the one or more improvement actions is a predetermined positive number if the one or more improvement actions reduce the cost of the routing solution or a predetermined negative number if the one or more improvement actions do not reduce the cost of the routing solution.

10. The method of any of claims 8-9, wherein:

applying the one or more of the plurality of improvement actions to the state before applying the perturbation action corresponds to a first iteration; and

the reward for applying the one or more improvement actions in an *i-th* iteration comprises a difference between a prior total cost reduction of a prior iteration and a total cost reduction of the *i-th* iteration.

11. The method of any of claims 1-10, wherein the improvement action comprises an operation or a no-operation.

12. The method of any of claims 1-11, wherein the predetermined condition comprises that the cost of the routing solution fails to reduce after a threshold number of consecutive improvement actions are applied to the state.

13. The method of any of claims 1-12, wherein:

initializing the state of the RL model comprises: for each of a plurality of RL models comprising the RL model, initializing a state of the each RL model based on a routing solution, wherein the each RL model comprises a plurality of improvement actions for applying to the state according to a policy to reduce a cost of the routing solution;

applying one or more of the plurality of improvement actions to the state comprises: for each of the plurality of RL models, applying one or more of the plurality of improvement actions to the state according to the policy to reduce the cost of the routing solution and to obtain updated routing solutions until a predetermined condition is satisfied, wherein the state comprises information of a number of improvement actions previously applied to the state, and wherein the number of improvement actions is different for each of the plurality of RL models;

applying the perturbation action to obtain the perturbed routing solution and feeding the perturbed routing solution back to the RL model comprises: for each of the plurality of RL models, applying a perturbation action to obtain a perturbed routing solution and feeding the perturbed routing solution back to the each RL model for the each RL model to perform the applying one or more of the plurality of improvement actions according to the policy; and

determining the routing solution with the minimum cost from the updated routing solutions comprises: determining the routing solution with the minimum cost from the updated routing solutions obtained for the plurality of RL models.

14. The method of claim 13, wherein:

the plurality of RL models comprise *N* RL models each referred to as a *j-th* model, with *j* being 1, 2, 3, … , or *N*; and

the state of the *j-th* RL model comprises information of *(j-1)* improvement actions previously applied to the state.

15. The method of any of claims 1-14, wherein the method comprising iteratively performing the applying one or more of the plurality of improvement actions to the state, and in response to the predetermined condition being satisfied, the applying the perturbation action to obtain the perturbed routing solution and feeding the perturbed routing solution back to the RL model until an exit condition is met.

16. A system for determining routing using reinforcement learning (RL) comprising one or more processors and one or more non-transitory computer-readable memories coupled

to the one or more processors and configured with instructions executable by the one or more processors to cause the system to perform operations comprising:

initializing a state of an RL model based on a routing solution comprising one or more routes for routing one or more vehicles through a plurality of locations, wherein the RL model comprises a plurality of improvement actions for applying to the state according to a policy to reduce a cost of the routing solution;

applying one or more of the plurality of improvement actions to the state according to the policy to reduce the cost of the routing solution and to obtain updated routing solutions until a predetermined condition is satisfied;

in response to the predetermined condition being satisfied, applying a perturbation action to obtain a perturbed routing solution and feeding the perturbed routing solution back to the RL model for the RL model to perform the applying one or more of the plurality of improvement actions according to the policy; and

determining a routing solution with a minimum cost from the updated routing solutions.

17. The system of claim 16, wherein the plurality of improvement actions for applying to the state according to the policy to reduce the cost of the routing solution comprise one or more of the following:

changing an order of at least two of the plurality of locations in one of the one or more routes; and

moving a location from one of the one or more routes to another one of the one or more routes.

18. The system of claim 16, wherein applying the perturbation action comprises one or more of the following:

shuffling at least two of the plurality of locations among at least two of the routes; concatenating at least two of the routes; and

breaking at least one of the routes into multiple routes.

19. A non-transitory computer-readable storage medium for determining routing using reinforcement learning (RL) configured with instructions executable by one or more processors to cause the one or more processors to perform operations comprising:

initializing a state of an RL model based on a routing solution comprising one or more routes for routing one or more vehicles through a plurality of locations, wherein the RL model comprises a plurality of improvement actions for applying to the state according to a policy to reduce a cost of the routing solution;

applying one or more of the plurality of improvement actions to the state according to the policy to reduce the cost of the routing solution and to obtain updated routing solutions until a predetermined condition is satisfied;

in response to the predetermined condition being satisfied, applying a perturbation action to obtain a perturbed routing solution and feeding the perturbed routing solution back to the RL model for the RL model to perform the applying one or more of the plurality of improvement actions according to the policy; and

determining a routing solution with a minimum cost from the updated routing solutions.

20. The storage medium of claim 19, wherein the plurality of improvement actions for applying to the state according to the policy to reduce the cost of the routing solution comprise one or more of the following:

changing an order of at least two of the plurality of locations in one of the one or more routes; and

moving a location from one of the one or more routes to another one of the one or more routes.