Implementation Details

Hyperparameter Settings The hyperparameters for our EPR strategy are detailed in Table 7. Baseline configurations for the various POMO variants adhere to the default settings specified in their respective original publications. We set the guidance data proportion to $P = 0.1$, and the path pattern length to $K = 20$ for TSP and $K = 3$ for CVRP. This choice of K is informed by the distinct structural properties of the two problems. We hypothesize that decision-making in CVRP is inherently more local due to capacity constraints, making shorter patterns (e.g., $K = 3$) sufficient to capture critical local structures. In contrast, TSP involves a single, uncapacitated tour, rendering its decision process more global; an ostensibly optimal local choice can incur substantial costs later. Therefore, longer patterns (e.g., $K = 20$) are employed to help the model learn larger-scale structural information and foster more globally-aware decisions. The sensitivity of these key hyperparameters, P and K, is further analyzed in Section Hyperparameter Sensitivity Analysis, confirming the model's robustness to reasonable variations.

The heuristic solvers used for guidance were configured to balance quality and speed. For HGS, the parameters were set as follows: `TimeLimit=10s`, `TargetFeasible=0.01`, and `NbElite=75`. For LKH3, the parameters were set as follows: `maxTrials=3000`, and `runs=20`, ensuring that the generation of a single elite solution is strictly capped at one second. Consequently, the total time consumed by the guidance strategy during the entire training process remains minimal and well within acceptable limits.

Computational Environment All experiments were conducted using PyTorch. The batch size was set to 80 (60) for TSP and 60 (50) for CVRP, determined by GPU memory limitations. The main results (Tables 2, 3, and 4) were obtained on a single NVIDIA RTX 2080Ti GPU with 22GB of VRAM. Due to their extensive computational requirements, the remaining experiments (Tables 1, 5, and 6) were conducted on a server with four NVIDIA V100 GPUs, each with 16GB of VRAM.

Hyperparameter	**TSP**	**CVRP**
Guidance Proportion (P)	0.1	0.1
Pattern Length (K)	20	3
Epoch Size	250,000	350,000
Batch Size (Tables 2, 3, and 4)	80	60
Batch Size (Tables 1, 5, and 6)	60	50
Learning Rate	1×10^{-4}	1×10^{-4}
Weight Decay	1×10^{-6}	1×10^{-6}

Table 7: Hyperparameter settings for the EPR strategy.

Implementation Details in Training

Following the experimental setup from, we generate node coordinates $\{(x_{ni}, y_{ni}) \mid i \in [N]\}$ randomly with uniform distribution, set vehicle capacity $Q = 50$, and generate demands $\{d_i \mid i \in [N]\}$ also from uniform distribution. To handle varying objective function scales across instances, we normalize the advantage terms in REINFORCE as follows:

$$A_i = R_i - \frac{1}{N}\sum_{i=1}^{N} R_i, \quad i \in [N], \tag{1}$$

$$\tilde{A}_i = \frac{A_i}{\max\{A_i \mid i \in [N]\}}. \tag{2}$$