THE DIRECTION CONVENTION THAT TOUR COST HIDES IN NEURAL TSP SOLVERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Neural solvers for the traveling salesman problem are compared by the cost of their tours, but one of their choices is invisible to that cost: the direction of travel. A tour and its reverse cost the same, yet on our 1,024 test instances with 100 cities, greedy POMO constructs every tour clockwise and Sym-NCO every tour counterclockwise. We call the direction in which a model constructs its tours its *convention*. We give a model the first cities of an optimal tour, a prefix, traveled clockwise or counterclockwise, and let it complete the tour. With 15 cities left, POMO and Sym-NCO complete it almost optimally with their convention and, against it, worse than when they construct whole tours. LEHD and BQ-NCO, whose tours cost within a fraction of a percent of POMO's, have no convention and complete closer still to optimal in both directions. One part of the score with which POMO and Sym-NCO pick the next city, the priority, ranks the cities left largely in the order of each model's own tours. On a given input, the priority depends only on which cities are left, so it ranks them the same from either endpoint of the prefix. Against the convention, where removing the priority changes the first choice, the priority usually favors the city whose optimal completion costs more. The priority thus shows one place where, in models that encode the cities once, the convention enters a choice. Models that are handed a prefix should therefore be tested in both directions.

1 INTRODUCTION

Many neural solvers for the traveling salesman problem (TSP) construct a tour one city at a time (Kool et al., 2019; Kwon et al., 2020). At each step, the model scores the cities left and, under greedy decoding, moves to the one with the highest score. Models are compared by the cost of their tours, and probes and sparse autoencoders read what their embeddings contain (Zhang et al., 2025; Narad et al., 2025). Neither splits the score with which a model chooses one city over another.

Such a split could show, choice by choice, a property that tour cost cannot: a tour's direction. On the symmetric Euclidean TSP, a tour and its reverse have equal cost. Yet three released models trained by reinforcement learning each construct their tours in one direction. Decoded greedily, POMO (Kwon et al., 2020) constructs all of its uniform 100-city test tours clockwise. Sym-NCO (Kim et al., 2022) and the Attention Model (AM) (Kool et al., 2019) construct all of theirs counterclockwise, AM from the start city it chooses itself. We call this direction the model's *convention*.

The convention matters when a model does not choose the direction itself, as in methods that keep part of a tour and let a model rebuild the rest. Some of these methods try both directions (Ye et al., 2024; Cheng et al., 2023), whereas others, including LEHD, pick one at random (Luo et al., 2023; 2025). Although the authors of LEHD note that a model could build better solutions in one direction (Luo et al., 2023), none of these works measures how much worse a model completes a tour against its convention, or how the convention shapes its choice of the next city (Appendix A).

We therefore ask three questions about models that have a convention. (i) How well does a model complete a tour that was started against its convention (Section 3)? (ii) Does the convention appear in a part of the model's score for the next city (Section 4)? (iii) When removing that part moves a choice to a city with a cheaper optimal completion, is the model's tour then shorter (Section 5)?

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

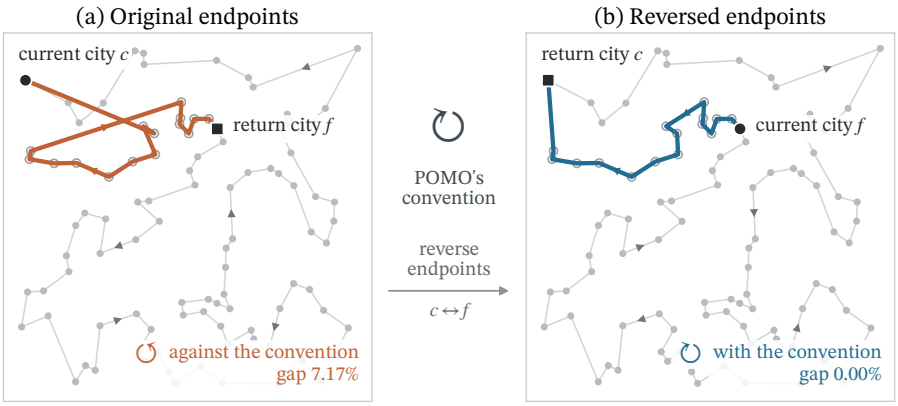

Figure 1: POMO completing one optimal-tour prefix, with 15 cities left, from its original (a) and reversed (b) endpoints. Gray: prefix; colored: completion. The instance, of 1,024, whose change in gap is nearest the median.

To answer these questions, we give a model the first cities of an optimal tour, a *prefix* (a partial tour), and let it complete the tour. To test the same prefix both with the model's convention and against it, we reverse its endpoints (Figure 1): the model completes the tour from the city where the prefix started and must return to the city where it ended. Coordinates, distances and the cities left all stay the same. As a reference, we use LEHD (Luo et al., 2023) and BQ-NCO (Drakulic et al., 2023), which imitate optimal completions and have no convention. A prefix also replaces the start city that AM would choose itself, and AM completes far from optimally in either direction (Section 3).

POMO, Sym-NCO and AM compute city embeddings once (Huang et al., 2025), and each candidate's raw score (before clipping) splits exactly into parts (Section 4.1). On a given input, one part depends only on the cities left, so it ranks them the same from either endpoint of the prefix; we call it the priority. We remove the priority, replacing each candidate's priority by the candidates' mean (a mean ablation), and see which choices change. For (iii), where the choice after removal has the cheaper optimal completion, we force it once and let the model, priority included, complete the tour.

Our contributions are as follows.

- **Against their convention, POMO and Sym-NCO complete a tour worse than when they construct whole tours** (Section 3). With it and 15 cities left, they complete almost optimally, while LEHD and BQ-NCO complete almost exactly either way: tour costs within a fraction of a percent can hide completion gaps that differ by orders of magnitude.
- **For most pairs of cities left, the priority favors the one the model's greedy tour visits first** (Section 4). Against the convention, where removing it changes POMO's or Sym-NCO's first choice, the choice after removal usually has the cheaper optimal completion.
- **Forcing a city with a cheaper optimal completion shortens the model's tour far more often with the convention than against it** (Section 5). Against it, POMO's and Sym-NCO's tours are shorter, on optimal-tour prefixes, after only a slight majority of the forced choices (51.9 to 59.7% over both models and four numbers of cities left; with it, 84.6 to 88.7%), and the model's later choices cancel most of the saving on average.

Models that are handed a prefix should therefore be tested in both directions.

2 Setup

Models. The five models construct a TSP tour one city at a time. An encoder maps each city to an embedding. At each step, a decoder scores every candidate against a query that includes the current city and the return city, and greedy decoding moves to the candidate with the highest score. AM, POMO and Sym-NCO compute the embeddings once and learn by reinforcement learning from their own complete tours, whereas LEHD and BQ-NCO re-encode the cities left at every step and learn to imitate optimal completions. Each model is a publicly released checkpoint decoded greedily, so it

is a deterministic policy π. Unless stated otherwise, instances come from the TSP100 test set: 1,024 instances of $n = 100$ cities drawn uniformly from the unit square.

Completing a prefix. A prefix holds the first cities of a tour in their order of travel, and the model completes the tour by choosing the cities left one at a time. We cut prefixes from an optimal tour traveled either way from city 0, which gives one clockwise and one counterclockwise prefix of each length, and from each model's own greedy tour; some comparisons also use the nearest-neighbor tour. The prefix's endpoints are the return city f, where it begins and to which the completed tour returns, and the current city c, where it ends and the model chooses next. The model's completion is the path it adds from c through the u cities left, the set U, to f. An optimal completion is one of least cost, and a candidate's is the cheapest one that visits it first. Visited cities are masked, and the query uses the prefix only through c and f. On a given input, each choice therefore depends on the prefix only through the state $s = (c, U, f)$, as an optimal completion does.

With and against the convention. A tour's direction is the sign of the signed area of the polygon that it traces, negative for clockwise (CW) and positive for counterclockwise (CCW), and a prefix has the direction of the tour it is cut from, or the opposite one when its endpoints are reversed. A model's convention is the direction in which it constructs its tours. POMO constructs all 1,024 test tours CW. Sym-NCO constructs all of them CCW, and so does AM from its own start city (Appendix B.3). LEHD and BQ-NCO construct tours in both directions; having no convention, they serve as the reference in Section 3. A prefix is *with* a model's convention when, as the model receives it, it has the direction of the model's own tours, and *against* it otherwise, so a prefix of a model's own tour is with it unless its endpoints are reversed. In Figure 1a, for example, the prefix is cut from an optimal tour traveled CCW, so it is against POMO's CW convention.

Reversed endpoints. A prefix can be handed to a model from either endpoint. With its *original endpoints*, the model completes the tour from the city where the prefix ends. With *reversed endpoints*, the current city and the return city swap, so the state (c, U, f) becomes (f, U, c) and the model completes the tour from f (Figure 1b). No coordinate changes, so the input, every distance and the cities left stay as they were. Because distances are symmetric, each completion from f to c is a completion from c to f traveled backward, at the same cost, so an optimal completion also costs the same. What changes is the endpoint the model completes from, and with it the prefix's direction, so with and against swap: the prefix of Figure 1 is with POMO's convention on reversed endpoints.

Gap. The gap of a tour is the percentage by which its cost exceeds the instance's optimal tour cost; for a greedy tour, this is the usual optimality gap. Optimal costs come from the exact solver Concorde (Applegate et al., 2003) or from dynamic programming. Differences of gaps, and all other cost differences, are in percentage points (pp) of the optimal tour cost. Gaps and differences are computed per instance and then averaged, over instances or, when a result concerns a set of choices, over those choices (Appendix B.4). When no optimal tour keeps the prefix, part of the gap is a cost forced by the prefix (Appendix B.1). On an optimal-tour prefix, with original or reversed endpoints, the rest of the optimal tour is an optimal completion, so the gap comes from the model's completion alone.

3 The Cost of Completing Against the Convention

3.1 Similar construction, different completion

Do models whose tours cost nearly the same also complete prefixes equally well? In Table 1, each prefix holds the first 85 cities of an optimal tour traveled clockwise or counterclockwise from city 0 (its CW or CCW source), and the model completes the tour through the $u = 15$ cities left. In construction, POMO, Sym-NCO, LEHD and BQ-NCO reach gaps within a fraction of a percent of one another. AM's tours cost more, and AM, whose own start city is replaced by every prefix (Appendix B.3), completes prefixes far from optimally in either direction. LEHD and BQ-NCO complete them almost exactly in either direction. POMO and Sym-NCO leave a small gap with their convention but, against it, a larger gap than their greedy tours. With 50 or 70 cities left (Section 5), POMO's gap even with the convention exceeds that of its greedy tour (Appendix C.1). So the gap

Table 1: Mean gap (%) of construction (one greedy tour, Appendix B.3) and completion of optimal-tour prefixes, TSP100, $u = 15$. Completion, Greedy: no augmentation. With and against: on the original input and endpoints; parentheses: CW or CCW source (Both sources: their mean).

	Construction	Completion		Completion, best of the original input and		
Model	Greedy	Prefix	Greedy	3 rotations	4 reflections	the other 7
POMO	0.846	With (CW)	0.323	0.120	0.231	0.089
		Against (CCW)	8.065	6.630	0.110	0.099
Sym-NCO	0.926	With (CCW)	0.343	0.148	0.236	0.105
		Against (CW)	8.366	6.715	0.094	0.083
AM	4.346	With (CCW)	14.982	2.746	5.983	2.181
		Against (CW)	19.047	9.105	2.812	2.457
LEHD	0.582	Both sources	7.2×10^{-4}	1.9×10^{-4}	1.7×10^{-4}	1.6×10^{-4}
BQ-NCO	0.344	Both sources	7.8×10^{-4}	1.6×10^{-4}	2.0×10^{-4}	1.1×10^{-4}

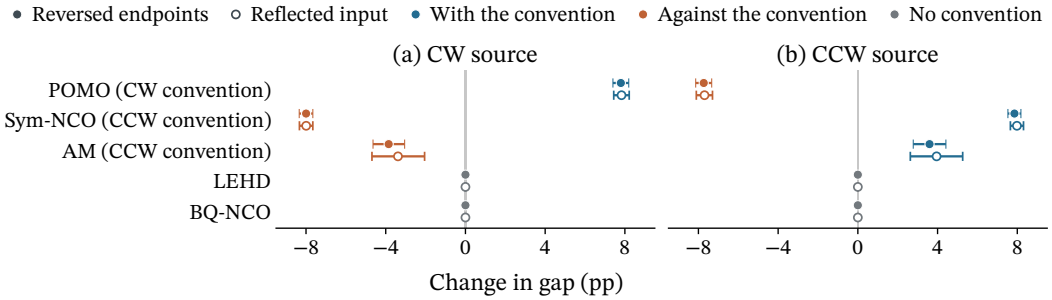

Figure 2: Change in gap under reversed endpoints and the reflected input (Table 1's optimal-tour prefixes, $u = 15$). With/against: on the original input and endpoints. Bars: multiplicity-adjusted 95% intervals (Appendix B.4).

of the greedy tour, measured without a prefix, cannot show how POMO and Sym-NCO complete prefixes against their convention.

3.2 Reversing the endpoints moves the gap with the direction

Table 1's prefixes with and against the convention hold different cities. Does the gap follow the prefix or its direction? Reversed endpoints (Figure 1) separate the two: they keep the input, the prefix's cities and the cities left, but flip the prefix's direction. If the gap follows the direction, reversing the endpoints should raise it for prefixes with a model's convention on the original endpoints, lower it for those against it, and leave LEHD's and BQ-NCO's unchanged. Figure 2 shows this for all five models: reversed endpoints change POMO's mean gap by 7.798 pp on those with its convention and by -7.733 pp on those against it, and LEHD's and BQ-NCO's by less than 0.001 pp (Table 6). For POMO and Sym-NCO, this change is about the difference between their gaps with and against the convention in Table 1.

The *reflected input*, $(x, y) \mapsto (1 - x, y)$ applied to the original input (the coordinates as given), flips the direction through the coordinates instead, keeping the prefix, the cities left and every distance, and moves the gap the same way and by about as much (POMO: 7.831 and -7.694 pp; Figure 2 shows both). This flip is the only change common to both (Appendix C.5), so POMO's, Sym-NCO's and AM's gaps follow the prefix's direction relative to the convention. The sign of this dependence, if not its size, carries over to nearly all of 51 planar TSPLIB instances (Reinelt, 1991) with 48 to 1,000 cities ($u = 15$). The change in gap under the reflected input is higher for prefixes with the convention on the original input than for those against it on all 51 instances for POMO, 49 for Sym-NCO and 47 for AM. On the 21 instances above 200 cities, which also differ in layout, this holds on 21, 20 and 18 of them, but POMO's mean changes shrink to 1.77 and -1.80 pp (Appendix C.2).

POMO's $\times 8$ augmentation keeps the cheapest completion over the eight rotations and reflections of the input, the symmetries of the unit square (the group D_4). A rotation keeps a prefix's direction and a reflection flips it, so a prefix against the convention on the original input is with it on every reflection. For POMO and Sym-NCO against the convention, the best over the original input and its three rotations leaves most of the gap, and the best over it and its four reflections almost none (Table 1). Augmentation over reflections thus hides the convention without changing it.

Yet reflections might help only by being more numerous, five inputs against four in Table 1. To compare equal numbers of inputs, we average the gap of the cheaper of two completions over all pairs of each kind. *Mixed pairs* (a rotation and a reflection) present the prefix in both directions, whereas *same-class pairs* (two rotations or two reflections) present it in one. For two completions with gaps e and e', the cheaper has gap $\min(e, e') = (e + e')/2 - |e - e'|/2$. Each of the eight inputs lies in four of the sixteen mixed pairs and in three of the twelve same-class pairs, so both kinds average $(e + e')/2$ to the same value, the mean of the eight gaps. Mixed pairs therefore do better, with a lower mean gap, exactly when they average a larger $|e - e'|$ than same-class pairs. For POMO and Sym-NCO, mixed pairs do better in all six TSP100 combinations of $u = 15$ or 50 and a prefix of the optimal tour traveled either way or of the nearest-neighbor tour (Appendix C.3). With optimal-tour prefixes and $u = 15$, POMO's mean gap is 0.277% with mixed pairs and 3.749% with same-class pairs. With equal numbers of inputs, POMO's and Sym-NCO's gaps thus differ more between directions than within one, on average. For AM, which also has a convention, same-class pairs do better in all six, so a convention alone does not imply that mixed pairs do better.

The gap shows that the convention matters, not which choices it affects; Section 4 asks which.

4 Where Does the Convention Enter the Next-City Choice?

We examine one term of the decoder's raw score, the priority, through its exact share of the margin between any two candidates and the choices that its removal changes. We find that it carries the convention into the choice: (1) it ranks most pairs of cities left in the model's visiting order; (2) removing it changes far more first choices against the convention than with it; and (3) against the convention, where removal changes POMO's or Sym-NCO's first choice, the city chosen instead usually has the cheaper optimal completion.

4.1 A score term that ranks cities the same from either endpoint

Reversing a prefix's endpoints keeps the input and the set U of cities left (Figure 1; Appendix C.5), and as distances are symmetric, an optimal completion, when unique, then visits U in reverse order. One ranking cannot match both an order and its reverse, so a score term that ranks the cities left strictly, and the same from either endpoint, can match the optimal order from at most one endpoint when $u \geq 2$. To find such a term, we write z_a for the raw score of a candidate $a \in U$ before the final clipping to $10 \tanh(z_a)$. At state s, decoder head ℓ attends to the cities left with weights $\alpha_{\ell j}(s)$ summing to one over U, and in POMO, Sym-NCO and AM, z_a is a fixed bias plus attention-weighted interactions per head:

$$z_a(s) = \beta_a + \sum_{\ell} \sum_{j \in U} \alpha_{\ell j}(s)\, G_\ell[j, a], \qquad \beta_a = \frac{h_a^\top b_o}{\sqrt{d}}, \qquad G_\ell[j, a] = \frac{h_a^\top W_{o,\ell} W_{v,\ell} h_j}{\sqrt{d}}, \quad (1)$$

where h_a is the d-dimensional embedding of city a, b_o the output bias, and $W_{v,\ell}$ and $W_{o,\ell}$ head ℓ's value projection and its block of the output projection (AM's variant: Appendix D.1). The interaction $G_\ell[j, a]$ is what head ℓ's attention to j adds to the raw score of a. As β_a and G_ℓ come from the embeddings, which the encoder computes once from the input, only the attention, whose query includes c and f, depends on the endpoints.

The priority is a ranking of the cities left, built in four steps. (1) A direction orders a pair of cities one way, so the priority uses only the antisymmetric part $A = (M - M^\top)/2$ of each interaction table M over all n cities: $A[j, a]$ compares what attention to j adds to a with what attention to a adds to j. (2) A global ranking, a table $q(a) - q(j)$, orders all cities the same whatever has been visited and so cannot follow the prefix; the cyclic part $\mathrm{cyc}(M)[j, a] = A[j, a] - q(a) + q(j)$ removes from A the global ranking closest to it in least squares, whose $q(a) = n^{-1} \sum_j A[j, a]$ is the column mean of A (Jiang et al., 2011). (3) A direction is unchanged under rotation and flips under reflection, so

the cyclic parts of G_ℓ^γ, computed on the input transformed by each $\gamma \in D_4$ (keeping city labels), are averaged with the sign $\det(\gamma)$, 1 for a rotation and -1 for a reflection, into K_ℓ: a part that changes like a direction enters all eight terms with the same sign and is kept, while one that is the same on all eight inputs, or changes in any other way under D_4, cancels. (4) The *priority* of a candidate a is the mean of its column of K_ℓ over the cities left, summed over heads, so that the attention cannot change its part of every margin:

$$K_\ell = \frac{1}{8} \sum_{\gamma \in D_4} \det(\gamma)\, \mathrm{cyc}\left(G_\ell^\gamma\right), \qquad p_U(a) = \sum_\ell p_{\ell,U}(a), \quad p_{\ell,U}(a) = \frac{1}{u} \sum_{j \in U} K_\ell[j,a]. \tag{2}$$

No tour or cost enters the priority, and K_ℓ is an exact additive part of $\mathrm{cyc}(G_\ell)$ (Appendix D.1). By construction, the priority changes sign under reflection and not under rotation; its ranking on a given input comes from the model's embeddings, and were these the same on all eight inputs, it would be zero (Appendix D.1). As K_ℓ has zero column sums over all n cities, $p_{\ell,U}(a) = -u^{-1} \sum_{j \notin U} K_\ell[j,a]$, so the ranking on U comes from the visited cities. A hand-set cyclic K on cities 0 to 5 placed clockwise on a circle (Appendix D.1) gives, after the prefix $0 \to 1$, $p_U = (\frac{1}{2}, \frac{1}{4}, -\frac{1}{4}, -\frac{1}{2})$ on cities $2, 3, 4, 5$, which ranks them in the optimal order from $c = 1$; on reversed endpoints, where the model completes from $f = 0$, the ranking is the same, but the optimal order is $5, 4, 3, 2$.

Like q in step (2), the column mean $p_{\ell,U}$ of step (4) is a least-squares ranking, now of K_ℓ on U: there, K_ℓ splits into $P_{\ell,U}[j,a] = p_{\ell,U}(a) - p_{\ell,U}(j)$ and a rest $K_\ell - P_{\ell,U}$ with zero column sums over U. Because each head's attention weights sum to one over U,

$$\sum_{j \in U} \alpha_{\ell j}(s)\, P_{\ell,U}[j,a] = p_{\ell,U}(a) - \sum_{j \in U} \alpha_{\ell j}(s)\, p_{\ell,U}(j),$$

so, summed over heads, $P_{\ell,U}$ adds $p_U(a)$ to the raw score of every candidate a, less a constant shared by all candidates at s, whatever the attention; the rest of K_ℓ adds terms that depend on the attention. *Removing the priority* subtracts this contribution from every raw score, that is, deletes $P_{\ell,U}$ from each G_ℓ in Equation 1: the raw score after removal, r_a, keeps β_a and every other part of the interactions, the rest of each K_ℓ included (Appendix D.1). As p_U has mean zero over U (K_ℓ is antisymmetric), removal replaces each candidate's priority by the candidates' mean (a mean ablation) and shifts all raw scores equally. The margin between two candidates a and b thus splits exactly, and only its first part depends on the attention:

$$z_a - z_b = \underbrace{(r_a - r_b)}_{\text{raw score after removal}} + \underbrace{p_U(a) - p_U(b)}_{\text{priority}}. \tag{3}$$

On a given input, the priority depends only on the cities left, so it ranks them the same from either endpoint of the prefix; original and reversed endpoints thus share its part of every margin, and any difference between their choices comes through r_a.

4.2 Choices changed by removing the priority

We test the priority on prefixes of each model's own greedy tour (own-tour prefixes; Section 2), which are with its convention on original endpoints and against it on reversed endpoints; each gives one state, that of the completion's first choice, and with the original and the reflected input of each of the 1,024 instances there are 2,048 states per model, u and endpoints. *Agreement* is the share of pairs of cities left that the priority ranks in the model's visiting order, the order in which its own tour visits them; a ranking unrelated to that order scores 50%. With $u = 10$, agreement on original endpoints is 89.9% for POMO and 88.9% for Sym-NCO, and AM's is also far above 50% (Table 2). As the ranking is the same on reversed endpoints, it agrees with the reverse of that visiting order only on the other pairs (there are no ties at $u = 10$). For eight released checkpoints (AM, Sym-NCO and six POMO checkpoints, which do not all share one convention), agreement likewise exceeds 50% on uniform instances, although for most of them it is lower on clustered and half-clustered ones (Appendix D.5); it also falls as more cities are left, to 78.9% for POMO at $u = 99$.

A simpler ranking, however, does nearly as well at $u = 70$: an angular sweep of the cities left around the centroid of all cities, starting toward the visited cities' centroid and turning in the model's convention, trails the priority by 0.6 percentage points for POMO and 0.6 for Sym-NCO, whereas at $u = 10$ no geometric ranking tested comes close (Appendix D.2).

Table 2: The priority's agreement with the model's visiting order, and the first choices its removal changes (own-tour prefixes, $u = 10$; Appendix D.2). Changes count only if the two choices' raw-score margin exceeds 10^{-7} before and after removal.

Model	Convention	Agreement (%) (original endpoints)	Changed choices (%) With (original endpoints)	Changed choices (%) Against (reversed endpoints)
POMO	CW	89.9	6.3	52.9
AM	CCW	85.3	14.1	46.7
Sym-NCO	CCW	88.9	7.7	58.4

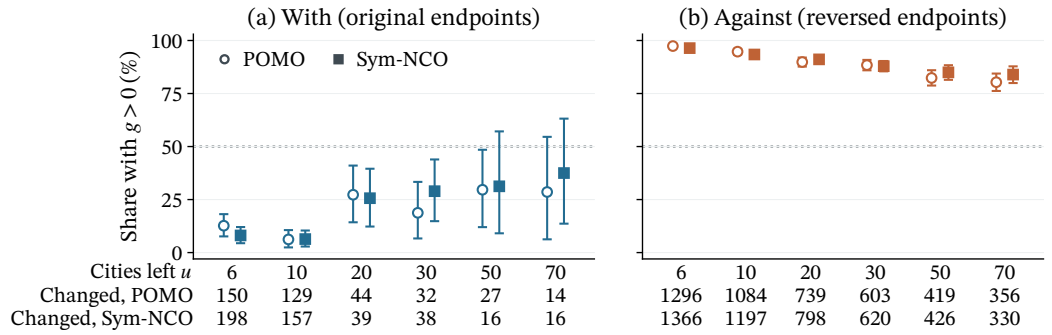

Figure 3: Share of first choices changed by removal that go to a cheaper city. Own-tour prefixes; changes within numerical tolerance are not counted (Appendix D.1); the counts of changed choices are out of 2,048 states per model, u and endpoints, and at $u = 10$ they give Table 2's percentages for POMO and Sym-NCO. Bars: pointwise 95% intervals.

With $u = 10$, removing the priority changes about half of each model's first choices against the convention but few with it (Table 2). As only r_a changes with the endpoints (Equation 3), the raw score after removal already picks the model's choice at most states with the convention, whereas against the convention it picks another city at about half of them, where the priority overrides it (Appendix D.1).

To see where these changed choices go, we write $Q^*(s, v)$, the optimal action value, for the cost of choosing city v at state s and completing the tour optimally (Appendix B.1). For the model's choice a and the choice after removal b, the saving $g = Q^*(s, a) - Q^*(s, b)$, in pp, is positive exactly when b is a *cheaper city*; we compute it for POMO and Sym-NCO only (Appendix B.3). Against the convention (Figure 3b), most changed first choices go to a cheaper city at every u tested, although the share falls as u grows; at $u = 70$, 286 of POMO's 356 do, and 277 of Sym-NCO's 330. Where removal changes the first choice against the convention, the priority thus usually favors the costlier of the two cities. With the convention (panel a), the estimated share of cheaper cities is below one half at every u; some intervals include one half at $u = 50$ and 70, and at $u = 50$ the mean of g is positive, with an interval that includes zero (Appendix D.4).

A *noise control* tests whether removal picks cheaper cities more often than independent Gaussian noise added to the raw scores, with one scale per model, u and endpoints, set on half of the instances to change the choice as often as removal does; only this rate is matched. On the other 512 instances, on prefixes against the convention with $u = 6$ or 10, at the states where removal changes the choice, the noise control's changes go to a cheaper city in 54.9 to 62.8%, against 93.2 to 97.9% for removal (POMO and Sym-NCO). Removing instead the rest of K_ℓ's contribution, which transforms like the priority under D_4, raises the mean Q^* of the chosen cities, in a comparison not matched in rate (Appendix D.3). On these prefixes, removal's changes to a cheaper city thus depend on the term removed, not on how many choices change.

These results concern single choices: they show neither that the priority is the only score term in which the convention appears (Appendix D.1), nor what share of Section 3's gap against the convention it accounts for, nor whether a cheaper city shortens the model's tour (Section 5).

5 DOES A CHEAPER CITY SHORTEN THE MODEL'S TOUR?

Forcing a cheaper city usually shortens the model's tour when the prefix is with the convention; against it, only a slight majority of forced cities do, and on average the model's later choices cancel most of the saving. At each of 10, 25, 50 and 70 cities left at the first choice, POMO and Sym-NCO each complete eight optimal-tour prefixes per instance, four with the convention and four against it (Appendix E.1). Every choice along a completion keeps the label of its prefix. At a state s, a is the model's choice, b the choice after removal (Section 4.1), and $g = Q^*(s, a) - Q^*(s, b)$ the saving of b under an optimal completion. We study the changed choices at which b is a cheaper city ($g > 0$); changes within numerical tolerance are not counted (Appendix D.1).

The noise control. As after own-tour prefixes (Section 4.2), on prefixes against the convention with 25 or 50 cities left, removal picks a cheaper first city far more often than the noise control does. As the noise control can pick any city, this comparison needs Q^* for every candidate, which we stored on every tenth instance, 103 of the 1,024; no such comparison exists with 10 or 70 cities left (Appendix E.2). The noise control is scaled on the other instances to change the choice as often as removal does, per model, number of cities left, step, and prefix with or against the convention. Of the first choices that removal changes, 86 to 91% go to a cheaper city; of all the first choices that the noise control changes, 39 to 43%. Each range covers both models and both numbers of cities left. Averaged over all first choices, with $g = 0$ for unchanged ones, removal's mean g exceeds the noise control's by 0.62 to 1.06 pp, with every multiplicity-adjusted 95% interval above zero. So the cheaper first cities that removal picks are specific to the priority. In the last third of a completion, by contrast, removal picks costlier cities on average, except on prefixes against the convention with ten cities left. Where compared, the noise control does no better than removal, so this sign alone does not show an effect specific to the priority (Appendix E.2).

How often the model's tour gets shorter. So far, an optimal completion priced each choice. From here on, the model completes the tour itself. At each changed choice to a cheaper city along a completion, a separate run forces b once and lets the model, priority included, complete the tour; forced choices are never combined along a completion (Appendix F.1). With $Q^\pi(s, v)$ the cost of choosing v at s and letting the model complete the tour, b saves $R = Q^\pi(s, a) - Q^\pi(s, b)$ under the model's completion, in pp, and $R > 0$ exactly when forcing b shortens the model's tour. We find that this happens far more often with the convention than against it (Figure 4a). With the convention, $R > 0$ after 84.6 to 88.7% of the changed choices to a cheaper city; against it, after only a slight majority, 51.9 to 59.7%, each range covering both models and all four numbers of cities left. The rates with the convention rest on far fewer such choices, 468 to 1,694 per model and number of cities left, than those against it, 5,606 to 7,443. As every forced city is cheaper under an optimal completion, a tour that does not get shorter means that the model's later choices cancelled the saving.

Where the saving goes. Against the convention, the mean R lies far below the mean g (Figure 4b). The model's *excess* after choosing v at s, $Q^\pi(s, v) - Q^*(s, v)$, is how much more its completion from the resulting state costs than an optimal one. Since R and g differ only in who completes the tour, their difference is a difference of excesses (Appendix F.2):

$$R = g - (\text{excess after } b - \text{excess after } a). \tag{4}$$

Against the convention, the model's mean excess is larger after b than after a. With ten cities left, the difference cancels most of the mean g. With 25 or more, the 95% interval of POMO's mean R includes zero; Sym-NCO's includes zero with 50 or 70 and lies below zero with 25, at -0.072 $[-0.141, -0.002]$ pp. Equation 4 locates where g is lost, not why completions with and against the convention differ: they pass through different states, which we did not compare beyond whether they start with or against it (Appendix F.2).

Removal at every first choice. Removing the priority at the first choice of every prefix with ten cities left, changed or not, makes the model's tours detectably shorter in none of the twelve combinations: eight of optimal-tour prefixes and four of nearest-neighbor prefixes (Appendix F.3). In all twelve, removal raises both models' mean excess after the first choice, the term of Equation 4 that cancelled the saving above. On the optimal-tour prefixes, removal lowers the mean Q^* of the first choice in all eight for both models, so the mean g is positive, and within the intervals the rise

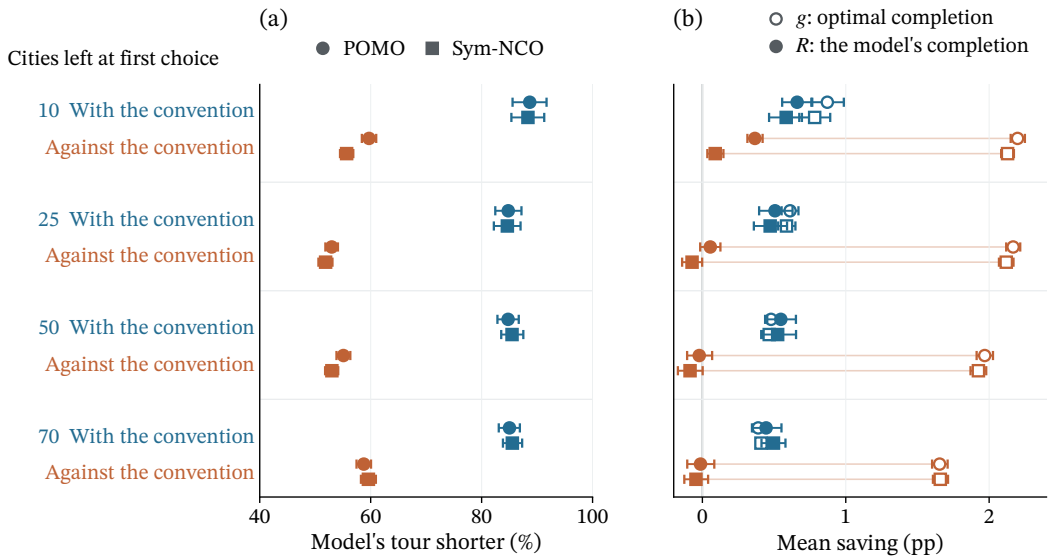

Figure 4: Savings from forcing b at changed choices to a cheaper city ($g > 0$) on optimal-tour prefixes. Bars: pointwise 95% intervals.

in excess offsets this saving: every multiplicity-adjusted interval of the change in POMO's tour cost includes zero, and Sym-NCO's tours cost more on average, with four of its eight adjusted intervals above zero. On nearest-neighbor prefixes, both models' tours cost detectably more in all four.

6 Discussion and Conclusion

We have shown that a tour's direction, invisible to tour cost, changes how well POMO and Sym-NCO complete a prefix. Given an optimal-tour prefix with 15 cities left, both complete almost optimally with their convention and, against it, worse than when they construct whole tours; reversing the prefix's endpoints, which moves no city, swaps these two outcomes.

The priority, which on a given input depends only on the cities left, ranks them the same from either endpoint, yet largely in the model's visiting order. Against the convention, removal usually moves the first choices it changes to a cheaper city, far more often than the noise control. The priority thus shows one place where the convention enters single choices; it shows neither that it is the only one nor what share of the gap it accounts for. This means that in models that encode the cities once, such as POMO, Sym-NCO and AM, a direction learned from whole tours can sit in a score term that no change of endpoints reaches. Models meant to complete tours from either end, as rebuilding methods require, can be checked for such a term with reversed endpoints; LEHD and BQ-NCO, which re-encode the cities left at every step, have no convention, although our results do not separate their architecture from their training. A cheaper city, however, does not yet make a shorter tour. On optimal-tour prefixes, forcing one shortens the model's tour in most cases with the convention but in only a slight majority against it, where the later choices cancel most of g on average; consistent with this, removal at every first choice does not detectably shorten tours. What causes the larger gap against the convention is open; retraining on own-tour prefixes with reversed endpoints would separate the direction itself from its rarity in training.

Methods that rebuild part of a tour, untested here, could inherit their model's convention, which differs among the six POMO checkpoints (Appendix D.5). Augmentation over reflections leaves almost none of the gap against the convention but hides the convention without changing it, so each checkpoint should be tested with original and reversed endpoints. Completing a prefix hands a model choices it did not make: a tour costs the same in either direction, but for POMO and Sym-NCO, completing one does not.

Reproducibility Statement

Appendix B specifies the tasks, reference costs, inference settings and statistical procedures. Appendix B.1 proves the cost decomposition, Appendix D.1 derives the score decomposition, and Appendices E.3 and F.2 prove the performance-difference sum along the model's completion and the relation between optimal and model action values after a single changed choice. Every evaluated checkpoint is publicly released (Appendices B.3 and D.5) and decoded greedily as Appendix B describes.

AI Use Statement

In this work, we used generative AI tools to help implement and test experiment code, run and check data analyses, check mathematical derivations and proofs, search for and summarize related literature, and draft and edit the manuscript, including its structure and wording. We did not use generative AI tools to generate data sets or to translate text. The research questions and the paper's claims originated with the authors. We have reviewed all AI-assisted work: every number in the paper is produced by scripts from recorded experiment outputs or checked against them, results were checked by independent audits against those records, and the authors verified all derivations and claims. We take responsibility for the final content of this work, including text, claims, code and artifacts produced with the aid of generative AI.

Ethics Statement

This study uses synthetic and public TSPLIB instances and involves no human subjects or operational deployment.

References

David Applegate, Robert Bixby, Vašek Chvátal, and William Cook. Concorde TSP Solver. Software, version 03.12.19, released December 19, 2003. URL https://www.math.uwaterloo.ca/tsp/concorde/.

Irwan Bello, Hieu Pham, Quoc V. Le, Mohammad Norouzi, and Samy Bengio. Neural combinatorial optimization with reinforcement learning. *arXiv preprint arXiv:1611.09940*, 2016. URL https://arxiv.org/abs/1611.09940.

Jieyi Bi, Yining Ma, Jiahai Wang, Zhiguang Cao, Jinbiao Chen, Yuan Sun, and Yeow Meng Chee. Learning generalizable models for vehicle routing problems via knowledge distillation. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://papers.nips.cc/paper_files/paper/2022/file/ca70528fb11dc8086c6a623da9f3fee6-Paper-Conference.pdf.

Kai-Wei Chang, Akshay Krishnamurthy, Alekh Agarwal, Hal Daumé III, and John Langford. Learning to search better than your teacher. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37, pp. 2058–2066. PMLR, 2015. URL https://proceedings.mlr.press/v37/changb15.html.

Hanni Cheng, Haosi Zheng, Ya Cong, Weihao Jiang, and Shiliang Pu. Select and optimize: Learning to solve large-scale TSP instances. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 1219–1231, 2023. URL https://proceedings.mlr.press/v206/cheng23a.html.

Darko Drakulic, Sofia Michel, Florian Mai, Arnaud Sors, and Jean-Marc Andreoli. BQ-NCO: Bisimulation quotienting for efficient neural combinatorial optimization. In *Advances in Neural Information Processing Systems*, volume 36, pp. 77416–77429, 2023. doi: 10.52202/075280-3385. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/f445ba15f0f05c26e1d24f908ea78d60-Abstract-Conference.html.

Olive Jean Dunn. Multiple comparisons among means. *Journal of the American Statistical Association*, 56(293):52–64, 1961. doi: 10.1080/01621459.1961.10482090.

Bradley Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1): 1–26, 1979. doi: 10.1214/aos/1176344552.

Richard Eglese and Sofoclis Zambirinis. Disruption management in vehicle routing and scheduling for road freight transport: a review. *TOP*, 26(1):1–17, 2018. doi: 10.1007/s11750-018-0469-4.

Billy E. Gillett and Leland R. Miller. A heuristic algorithm for the vehicle-dispatch problem. *Operations Research*, 22(2):340–349, 1974. doi: 10.1287/opre.22.2.340.

Nathan Grinsztajn, Daniel Furelos-Blanco, Shikha Surana, Clément Bonnet, and Thomas D. Barrett. Winner takes it all: Training performant RL populations for combinatorial optimization. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/97b983c974551153d20ddfabb62a5203-Abstract-Conference.html.

Michael Held and Richard M. Karp. A dynamic programming approach to sequencing problems. *Journal of the Society for Industrial and Applied Mathematics*, 10(1):196–210, 1962. doi: 10.1137/0110015.

André Hottung and Kevin Tierney. Neural large neighborhood search for the capacitated vehicle routing problem. In *Proceedings of the 24th European Conference on Artificial Intelligence (ECAI 2020)*, volume 325 of *Frontiers in Artificial Intelligence and Applications*, pp. 443–450. IOS Press, 2020. doi: 10.3233/FAIA200124. URL https://arxiv.org/abs/1911.09539.

André Hottung, Mridul Mahajan, and Kevin Tierney. PolyNet: Learning diverse solution strategies for neural combinatorial optimization. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/c7b9810b2e27bccdc8d605a3a54c8cd6-Abstract-Conference.html.

Ziwei Huang, Jianan Zhou, Zhiguang Cao, and Yixin Xu. Rethinking light decoder-based solvers for vehicle routing problems. In *International Conference on Learning Representations*, pp. 27471–27491, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/447ac93bf22099aa346a45577376492d-Abstract-Conference.html.

Xiaoye Jiang, Lek-Heng Lim, Yuan Yao, and Yinyu Ye. Statistical ranking and combinatorial Hodge theory. *Mathematical Programming*, 127(1):203–244, 2011. doi: 10.1007/s10107-010-0419-x.

Minsu Kim, Junyoung Park, and Jinkyoo Park. Sym-NCO: Leveraging symmetricity for neural combinatorial optimization. In *Advances in Neural Information Processing Systems*, volume 35, pp. 1936–1949, 2022. doi: 10.52202/068431-0141. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/0cddb777d3441326544e21b67f41bdc8-Abstract-Conference.html.

Wouter Kool, Herke van Hoof, and Max Welling. Attention, learn to solve routing problems! In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=ByxBFsRqYm.

Yeong-Dae Kwon, Jinho Choo, Byoungjip Kim, Iljoo Yoon, Youngjune Gwon, and Seungjai Min. POMO: Policy optimization with multiple optima for reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 33, pp. 21188–21198, 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/hash/f231f2107df69eab0a3862d50018a9b2-Abstract.html.

Sirui Li, Zhongxia Yan, and Cathy Wu. Learning to delegate for large-scale vehicle routing. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/dc9fa5f217a1e57b8a6adeb065560b38-Abstract.html.

Xiayang Li and Shihua Zhang. Learning-based TSP-solvers tend to be overly greedy. *Computers & Operations Research*, 192:107467, 2026. doi: 10.1016/j.cor.2026.107467.

Shen Lin. Computer solutions of the traveling salesman problem. *The Bell System Technical Journal*, 44(10):2245–2269, 1965. doi: 10.1002/j.1538-7305.1965.tb04146.x.

Shiyan Liu, Bohan Tan, Yaoxin Wu, and Yan Jin. MViewRouter: Internalizing geometric equivariance via multi-view alternating attention for combinatorial routing. *arXiv preprint arXiv:2606.01084*, 2026. URL https://arxiv.org/abs/2606.01084v1.

Fu Luo, Xi Lin, Fei Liu, Qingfu Zhang, and Zhenkun Wang. Neural combinatorial optimization with heavy decoder: Toward large scale generalization. In *Advances in Neural Information Processing Systems*, volume 36, pp. 8845–8864, 2023. doi: 10.52202/075280-0387. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/1c10d0c087c14689628124bbc8fa69f6-Abstract-Conference.html.

Fu Luo, Xi Lin, Yaoxin Wu, Zhenkun Wang, Xialiang Tong, Mingxuan Yuan, and Qingfu Zhang. Boosting neural combinatorial optimization for large-scale vehicle routing problems. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/264f2e10479c9370972847e96107db7f-Abstract-Conference.html.

Reuben Narad, Léonard Boussioux, and Michael Wagner. Mechanistic interpretability for neural TSP solvers. *arXiv preprint arXiv:2510.21693*, 2025. URL https://arxiv.org/abs/2510.21693.

Reuben Narad, Léonard Boussioux, and Michael Wagner. Beyond solving: Prescriptive probing for neural routing solvers. *arXiv preprint arXiv:2602.07216*, 2026. URL https://arxiv.org/abs/2602.07216.

Victor Pillac, Michel Gendreau, Christelle Guéret, and Andrés L. Medaglia. A review of dynamic vehicle routing problems. *European Journal of Operational Research*, 225(1):1–11, 2013. doi: 10.1016/j.ejor.2012.08.015.

Gerhard Reinelt. TSPLIB—a traveling salesman problem library. *ORSA Journal on Computing*, 3 (4):376–384, 1991. doi: 10.1287/ijoc.3.4.376.

Daniel J. Rosenkrantz, Richard E. Stearns, and Philip M. Lewis, II. An analysis of several heuristics for the traveling salesman problem. *SIAM Journal on Computing*, 6(3):563–581, 1977. doi: 10.1137/0206041.

Stéphane Ross and J. Andrew Bagnell. Reinforcement and imitation learning via interactive no-regret learning. *arXiv preprint arXiv:1406.5979*, 2014. URL https://arxiv.org/abs/1406.5979.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. Pointer networks. In *Advances in Neural Information Processing Systems*, volume 28, 2015. URL https://proceedings.neurips.cc/paper/2015/hash/29921001f2f04bd3baee84a12e98098f-Abstract.html.

Oriol Vinyals, Samy Bengio, and Manjunath Kudlur. Order matters: Sequence to sequence for sets. In *International Conference on Learning Representations*, 2016. URL https://arxiv.org/abs/1511.06391.

Haoran Ye, Jiarui Wang, Helan Liang, Zhiguang Cao, Yong Li, and Fanzhang Li. GLOP: Learning global partition and local construction for solving large-scale routing problems in real-time. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 20284–20292, 2024. doi: 10.1609/aaai.v38i18.30009.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations*, pp. 1651–1678, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/06a52a54c8ee03cd86771136bc91eb1f-Abstract-Conference.html.

Zhiqin Zhang, Yining Ma, Zhiguang Cao, and Hoong Chuin Lau. Probing neural combinatorial optimization models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 131824–131862, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/becf96dcee66f9df08169e4c96bbbd8c-Abstract-Conference.html.

Guide to the appendices. Appendix A relates this study to prior work, Appendices B to D follow Sections 2 to 4, and Appendices E and F support Section 5. Prefix, completion, the endpoints c and f, U, u, the state s, CW and CCW, convention, with and against, reversed endpoints, gap and pp (percentage points of the optimal tour cost) are defined in Section 2; the reflected input, D_4 and $\times 8$ augmentation in Section 3.2; raw score, interaction, priority, removal and r_a in Section 4.1; own-tour prefixes, agreement, changed choice, Q^*, g, cheaper city and noise control in Section 4.2; and Q^π, R and excess in Section 5, where every choice of a completion keeps its prefix's label, with or against. The appendices add u_0, the number of cities left at a completion's first choice, and a *condition*, one combination of prefix kind (for optimal-tour prefixes, also CW or CCW source or stored order; Appendix B.2), input, endpoints and number of cities left. Views are defined in Appendix C.3, and the tie filter and raw-margin rule for counting changed choices in Appendix D.1.

A Related Work

Constructing and selecting tours. Learned tour construction developed from supervised pointer networks, whose TSP targets followed one fixed direction because restricting the set of equivalent outputs helped learning (Vinyals et al., 2015; 2016), through reinforcement learning (Bello et al., 2016) to attention-based routing models (Kool et al., 2019). The POMO authors observe that learned agents choose cities in an orderly fashion, and POMO selects over start cities and coordinate views (Kwon et al., 2020). Poppy and PolyNet learn complementary construction strategies (Grinsztajn et al., 2023; Hottung et al., 2025). Sym-NCO adds symmetry-aware baselines and an invariance loss on a projection of the encoder output to existing solvers such as POMO (Kim et al., 2022), and MViewRouter attends across geometric views to target equivariance and reports POMO's complete-tour cost for each of the eight views (Liu et al., 2026). These uses of rotations and reflections concern complete tours; pipelines that rebuild path subproblems decode them from both endpoints and keep the result that costs less (Ye et al., 2024; Cheng et al., 2023). None of them compares reflections, which flip a tour's direction, with rotations, which preserve it, on prefixes of fixed direction, as Section 3.2 does.

Completing prefixes. LEHD learns by supervision to complete segments of optimal tours, of random length and direction, from a current city to a return city. It re-encodes the cities left at each step and motivates reconstruction in random directions by noting that a constructive model could build better solutions in one direction (Luo et al., 2023). BQ-NCO imitates optimal completions from states reduced to the current city, the cities left and the return city (Drakulic et al., 2023). Both train on completing optimal-tour prefixes. Within search, repair operators reconnect partial solutions (Hottung & Tierney, 2020), local reconstruction rebuilds segments between fixed endpoints in a random direction (Luo et al., 2025), and delegation re-solves selected subproblems (Li et al., 2021). Dynamic requests and disruptions require revising plans that keep the parts already executed (Pillac et al., 2013; Eglese & Zambirinis, 2018).

Understanding learned choices. Learned TSP solvers perform worse on instances whose optimal tours join fewer cities to a nearest neighbor, which Li & Zhang (2026) read as a tendency to be overly greedy. ReLD argues that a light decoder must serve every later choice from one set of static embeddings (Huang et al., 2025); the priority is an exact term of such a decoder's raw score that, on a given input, ranks the cities left the same from either endpoint of a prefix. Probing NCO fits linear probes to the city embeddings of AM, POMO and LEHD, including one that separates optimal-tour edges from nearest-neighbor edges with labels checked by constrained exact re-solves, and ablates the dimensions it identifies in LEHD's final layer (Zhang et al., 2025). Sparse autoencoders expose geometric features (Narad et al., 2025), and prescriptive probing predicts the value of changes to an instance from frozen representations (Narad et al., 2026). We instead split the decoder's raw scores into additive parts, select one by how it transforms under rotations and reflections of the input, using the cyclic part of an interaction table (Jiang et al., 2011), remove it at a single state, and price the model's choice and the choice after removal by their optimal action values. No probe, label or cost defines the priority.

Evaluating changed choices. The value of a choice depends on the choices that follow it. Learning to search completes one-step deviations from states of the learned policy either with a reference policy or with the learned policy itself, and the two completions can favor different deviations

(Chang et al., 2015). Section 5 follows this design, with optimal completion as the reference and the model as the learned policy. Interpreting an intervention also requires naming both the change and the measured outcome (Zhang & Nanda, 2024). Here the change removes the priority at a single state, and the outcomes are the next choice, its optimal action value and the completed tour.

B Completion Setup and Evaluation

B.1 Completion costs and excess

The arguments below use neither Euclidean coordinates nor the triangle inequality, so we state them for a finite complete graph on the cities with symmetric, nonnegative edge costs $W_{ij} = W_{ji}$. States $s = (c, U, f)$ and completions are those of Section 2; the endpoints lie outside U and coincide only when the prefix consists of one city. For a permutation $\omega = (\omega_1, \ldots, \omega_u)$ of U, the completion that visits U in this order costs $L_s(\omega)$, and the optimal completion cost $V^*(s)$ is its minimum:

$$L_s(\omega) = W_{c\omega_1} + \sum_{i=1}^{u-1} W_{\omega_i\omega_{i+1}} + W_{\omega_u f}, \qquad V^*(s) = \min_{\omega \in \mathrm{Perm}(U)} L_s(\omega). \tag{5}$$

When U is empty, the completion is the return edge alone, so $V^*(c, \varnothing, f) = W_{cf}$; as W is symmetric, reversing ω gives $V^*(c, U, f) = V^*(f, U, c)$. Choosing $v \in U$ leads to the state $s_v = (v, U \setminus \{v\}, f)$, so the optimal action value of Section 4.2 is $Q^*(s, v) = W_{cv} + V^*(s_v)$. Every completion from a state with nonempty U is the edge to some $v \in U$ followed by a completion of s_v, so $V^*(s) = \min_{v \in U} Q^*(s, v)$. A deterministic feasible policy π completes s at a cost $V^\pi(s)$, and its excess at s, the regret of π from that state, is $E_\pi(s) = V^\pi(s) - V^*(s) \geq 0$. Choosing v at s adds the *step excess* $Q^*(s, v) - V^*(s) \geq 0$, and along a completion the step excesses sum to the excess (Proposition E.1).

Cost forced by the prefix and excess. Let $W(\tau)$ be the cost of the edges of a prefix τ, s the state at its end and C^* the optimal tour cost of the instance. Every tour that keeps τ consists of these edges and one completion of s, so the cheapest such tour costs $C^*(\tau) = W(\tau) + V^*(s)$ and the tour that π completes costs $C_\pi(\tau) = W(\tau) + V^\pi(s)$. Adding and subtracting $C^*(\tau)$ in $C_\pi(\tau) - C^*$, and cancelling $W(\tau)$ in $C_\pi(\tau) - C^*(\tau)$, proves

$$C_\pi(\tau) - C^* = \underbrace{C^*(\tau) - C^*}_{\text{cost forced by the prefix}} + \underbrace{E_\pi(s)}_{\text{excess}}. \tag{6}$$

The first term is nonnegative because the tours that keep τ form a subset of all tours, and the second because no feasible completion is cheaper than $V^*(s)$. The first is zero on a prefix of an optimal tour, which that tour keeps, with original or reversed endpoints, and in construction, where fixing at most the start city excludes no tour; in both cases the gap in percent equals the excess in pp. For two policies π and π' on the same prefix, $W(\tau)$ cancels in the difference of their tour costs and $V^*(s)$ in the difference of their excesses, so $C_\pi(\tau) - C_{\pi'}(\tau) = E_\pi(s) - E_{\pi'}(s)$. Paired completion differences therefore compare what π and π' could still avoid, even when no optimal tour keeps the shared prefix.

B.2 Test sets and optimal costs

For each of the sizes 20, 50 and 100, the test set holds 1,024 independent instances with cities drawn uniformly from the unit square. A prefix's *source* is the tour it is cut from: an optimal tour, the nearest-neighbor (NN) tour, which always moves to the nearest city left (Rosenkrantz et al., 1977), or the model's own greedy tour. Each instance's optimal tour is stored in two orders from city 0, Order 1 and its reverse, Order 2; on each instance one is traveled clockwise and is the CW source, and the other is the CCW source. Both orders and the NN tour start at city 0, so every prefix cut from them returns to city 0 unless its endpoints are reversed. Unless a result is split into CW and CCW sources, means over optimal-tour prefixes average the two orders within each instance. Figures 5 and 12, the six combinations of Section 3.2 and the removal at every first choice in Section 5 instead keep the stored orders, so each order is the CW source on some instances and the CCW source on others. All tours, on the original input and its rotations and reflections, are scored on one shared

integer distance matrix, $W_{ij} = \lfloor 10^7 \|x_i - x_j\|_2 + 0.5 \rfloor$ for city positions x_i, so that cost differences between inputs come from the tours selected alone.

Concorde (Applegate et al., 2003) solves each instance on W, confirms optimality and returns a feasible tour of matching cost. On a prefix of an optimal tour, the rest of that tour is an optimal completion, since a cheaper one would contradict the tour's optimality. For other states with few cities left, subset dynamic programming (Held & Karp, 1962) evaluates the recurrence $V^*(s) = \min_{v \in U} Q^*(s, v)$ of Appendix B.1 down to an empty U. For the rest, Concorde solves $U \cup \{c, f\}$ plus an auxiliary city joined to c and f at zero cost and to every other city at a cost above that of any completion, and deleting the auxiliary city from an optimal tour leaves an optimal completion, which is recosted on W. The optimal action value $Q^*(s, v)$ also requires $V^*(s_v)$, which equals $V^*(s) - W_{cv}$ when v begins a known optimal completion of s and is solved separately otherwise.

B.3 Models, checkpoints and decoding

All models use their frozen released weights: POMO and AM size-specific TSP20/50/100 checkpoints, LEHD and BQ-NCO one set of TSP weights across sizes, and Sym-NCO its TSP100 checkpoint. Optimal action values and forced choices use the POMO and Sym-NCO TSP100 checkpoints on the TSP100 test set, the controls of Appendix D.3 other than the noise control add POMO's TSP20 and TSP50 checkpoints on their test sets, and the checkpoints and instance sets of Appendix D.5 measure the convention, agreement and changed choices, not costs.

Construction yields one greedy tour, which POMO, Sym-NCO, LEHD and BQ-NCO start at city 0 and which LEHD constructs without its reconstruction step. AM instead starts from its own start city, the city it chooses first, because starting it at city 0 changes its result sharply: its greedy tours have gaps of 0.278%, 1.614% and 4.346% at TSP20/50/100 from its own start city and 27.012%, 28.411% and 66.717% from city 0. Optimal-tour and NN prefixes also fix AM's start city, city 0 unless their endpoints are reversed, so AM's completion gaps may include an effect of this start that the gaps from city 0 do not estimate; contrasts that change only the input hold this start fixed.

A prefix is completed by replaying it whole and then applying the released greedy scorer and tie rule, without multi-start search, beam search or reconstruction. For the forced choices of Section 5 and Appendix F, each forced run reuses the embeddings of the model's own run, including the mean city embedding in Sym-NCO's query, so forcing changes only the choice. If the model's choice fails, for example through a nonfinite score, nearest neighbor completes the tour, which stays in the evaluation; Section 3 and Appendix C never required this fallback.

B.4 Intervals and means

The independent unit under the fixed released weights is the instance, so each bootstrap draw (Efron, 1979) keeps an instance's CW and CCW sources, inputs, numbers of cities left, completion steps and candidate permutations (Appendix D.3) together, and reusing a test set across analyses provides no independent replication. Each cost difference is in pp, percentage points of its instance's optimal tour cost. A mean over completions weights instances equally and includes completions without changed choices, whereas a mean over changed choices, or over those to a cheaper city, weights each choice equally.

Multiplicity-adjusted 95% intervals use 50,000 instance draws and the percentile tails $0.025/m$ and $1 - 0.025/m$ for a family of m contrasts, a Bonferroni allocation (Dunn, 1961) that relies on bootstrap coverage approximations; showing only part of a family does not reduce m. Pointwise 95% intervals use the 2.5th and 97.5th percentiles of 2,000 draws (50,000 for Sym-NCO's means before and after removing the priority or the remainder, Appendix D.3) and do not form simultaneous bands. The levels of Tables 1, 3 and 4, the results of Appendix D.5 and the rows of Figure 8 other than removal of the priority are means without intervals. A subset empty in some bootstrap draw reports no interval, only its estimate and count.

In Figure 2, the reflected-minus-original gaps of POMO, AM, LEHD and BQ-NCO on CW and CCW sources form one family of eight contrasts; POMO's intervals are $[7.441, 8.217]$ and $[-8.105, -7.294]$ pp. Sym-NCO's two contrasts come from a separate run on the same test set and form their own family with $m = 2$: -7.993 $[-8.330, -7.658]$ pp on CW sources and 7.983 $[7.654, 8.312]$ on CCW sources, where each draw resamples an instance's two prefixes together; its

reflected-minus-original gap has the sign of its mean on 993 of the 1,024 instances on CW sources and on 998 on CCW sources. The reversed-endpoints contrasts of Figure 2, like each column of Table 6, form families of the same two kinds.

C Direction and Augmentation over Rotations and Reflections

C.1 Other sizes, numbers of cities left and prefix kinds

Tables 3 and 4 repeat the comparison of Table 1 on TSP20 and TSP50 and, on TSP100, with $u = 50$ and on NN prefixes. At every size and with $u = 50$, POMO constructs short tours but completes optimal-tour prefixes with substantial gaps, whereas LEHD and BQ-NCO leave far smaller ones. On the completions of Appendix E.1, which also use the reflected input and reversed endpoints, POMO's mean gap with the convention grows with the number of cities left, to 1.06% $[0.98, 1.15]$ at $u_0 = 50$ and 1.27% $[1.18, 1.37]$ at $u_0 = 70$, above the gap of its greedy tour (Table 1); against the convention, it is 15.31% $[15.02, 15.62]$ and 14.34% $[14.09, 14.58]$. NN prefixes force extra cost, but on a shared prefix the difference between two models' gaps equals the difference between their excesses (Appendix B.1); on NN prefixes, POMO's excess exceeds LEHD's by 4.12 pp with $u = 15$ and by 7.95 pp with $u = 50$.

Table 3: Mean gap (%) of construction and of completion of optimal-tour prefixes (CW and CCW sources averaged), and mean excess (pp) on NN prefixes; TSP20/50, $u = 15$, original input.

	TSP20			TSP50		
Model	Construction	Optimal-tour prefix	NN prefix (excess)	Construction	Optimal-tour prefix	NN prefix (excess)
POMO	0.077	9.097	5.490	0.398	7.634	6.839
AM	0.278	22.357	18.499	1.614	18.916	17.520
LEHD	0.881	0.445	1.023	0.459	0.001	1.352
BQ-NCO	0.417	0.205	0.550	0.270	0.002	0.972

Table 4: Table 1 extended to $u = 50$ and NN prefixes (TSP100, original input, mean gap in %; optimal-tour prefixes: CW and CCW sources averaged). NN gaps include the cost forced by the prefix; Appendix C.1 compares models on them.

	Gap (%)				
		Optimal-tour prefix		NN prefix	
Model	Construction	$u = 15$	$u = 50$	$u = 15$	$u = 50$
POMO	0.846	4.194	8.110	25.519	18.603
Sym-NCO	0.926	4.355	8.059	25.516	18.189
AM	4.346	17.015	78.266	45.147	80.888
LEHD	0.582	0.001	0.025	21.395	10.656
BQ-NCO	0.344	0.001	0.021	20.896	10.457

C.2 The convention on TSPLIB instances

The construction and reflection comparisons of Section 3 are repeated on all 51 TSPLIB instances (Reinelt, 1991) with planar coordinates and 48 to 1,000 cities, excluding linhp318, a copy of lin318. Coordinates are scaled into the unit square by the longer side, optima are recomputed with Concorde on the integer objective of Appendix B.2, and the TSP100 checkpoints run unchanged. Each of eight start cities per instance gives a CW and a CCW source of the optimal tour, cut at $u = 15$. Greedy tours are constructed on all eight rotations and reflections of the input, from each of the eight start cities for POMO and Sym-NCO, from its own start city for AM (Appendix B.3), and from the first for LEHD and BQ-NCO. Both comparisons and their predicted signs were fixed before the run, after a smoke test on kroA100 had printed the tours' directions; all models run on CPU in float32.

Both predicted patterns appear for POMO, Sym-NCO and AM, and neither appears for LEHD and BQ-NCO (Table 5). In the coordinates each model receives, 98.9% of POMO's greedy tours are CW, and 99.4% of Sym-NCO's and 99.3% of AM's are CCW. The CW-minus-CCW difference, the reflected-minus-original gap on CW sources minus that on CCW sources, is positive for POMO on 51 of 51 instances and negative for Sym-NCO and AM on 49 and 47 (exact sign tests with p-values below 10^{-9}). The difference shrinks with size, which varies together with layout on these instances, so the size bands are descriptive: on the 21 instances with more than 200 cities, POMO's mean reflected-minus-original gaps are 1.77 $[1.12, 2.47]$ pp on CW sources and -1.80 $[-2.63, -1.08]$ pp on CCW sources, yet the difference still has the predicted sign on 21, 20 and 18 of them for POMO, Sym-NCO and AM. Without augmentation, POMO's completions have mean gaps of 0.85% $[0.65, 1.11]$ on CW sources and 5.75% $[4.83, 6.70]$ on CCW sources over all instances. The score and Q^* analyses use the TSP100 test set.

Table 5: TSPLIB check on 51 planar instances ($u = 15$). CW greedy tours: share of greedy tours constructed CW. Reflected − original gap: mean on CW or on CCW sources (pp), with pointwise 95% intervals; 0.00 marks a magnitude below 0.005. CW − CCW difference: numbers of instances on which it is positive, negative or zero.

Model	CW greedy tours (%)	Reflected − original gap, CW source	Reflected − original gap, CCW source	CW − CCW difference: > 0 / < 0 / $= 0$
POMO	98.9	4.96 [3.83, 6.18]	-5.00 [-6.13, -3.95]	51 / 0 / 0
Sym-NCO	0.6	-4.98 [-6.14, -3.89]	4.99 [3.97, 6.12]	2 / 49 / 0
AM	0.7	-4.23 [-6.52, -2.61]	4.32 [2.79, 5.96]	4 / 47 / 0
LEHD	50.2	0.00 [0.00, 0.00]	0.00 [0.00, 0.01]	11 / 9 / 31
BQ-NCO	49.5	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]	9 / 11 / 31

C.3 Augmentation over views and two-view selection

Augmentation and source direction. A *view* γ is the input transformed by a rotation or reflection $\gamma \in D_4$ (Section 3.2); the original input is the identity view id. The class $\mathcal{G}_+$ of four rotations, which include id, keeps a tour's direction, and the class $\mathcal{G}_-$ of four reflections flips it; one of them, $\rho_x : (x, y) \mapsto (1 - x, y)$, gives the reflected input. With C_γ the cost of the tour completed on view γ, the completion columns of Table 1 report C_{id}, $\min_{\gamma \in \mathcal{G}_+} C_\gamma$, $\min\{C_{\mathrm{id}}, \min_{\gamma \in \mathcal{G}_-} C_\gamma\}$ and $\min_{\gamma \in \mathcal{G}_+ \cup \mathcal{G}_-} C_\gamma$. Augmentation needs neither a direction label nor an optimum, and as every set contains id, it never costs more than C_{id}. Table 1 splits these columns by CW and CCW source for POMO, Sym-NCO and AM and averages the two for LEHD and BQ-NCO, whose mean gaps stay below 0.002% in every column on each source.

On the 2,048 optimal-tour prefixes with $u = 15$, the original or the reflected input, whichever presents the complete source tour in the model's convention, obtains 93.8% $[92.6, 94.9]$ of POMO's and 93.8% $[92.7, 94.9]$ of Sym-NCO's gain of $\times 8$ augmentation in mean gap over the original input (pointwise 95% intervals), almost all of it on prefixes against the convention on the original input, although this choice is unavailable at inference, because it depends on the direction of the complete source tour, including the part still to be completed, and the comparison does not separate the direction of a tour from its familiarity to the model.

The rotation and reflection columns of Table 1 take the best of four and of five views. *Two-view selection* keeps the cheaper completion of two views; averaged over the twelve *same-class pairs* (two rotations or two reflections) or over the sixteen *mixed pairs* (one of each), it uses every view equally often and so removes this difference in size:

$$\bar{C}_{\mathrm{same}} = \frac{1}{12} \sum_{\{\gamma, \gamma'\} \subset \mathcal{G}_+ \text{ or } \mathcal{G}_-} \min(C_\gamma, C_{\gamma'}), \qquad \bar{C}_{\mathrm{mixed}} = \frac{1}{16} \sum_{\gamma \in \mathcal{G}_+, \gamma' \in \mathcal{G}_-} \min(C_\gamma, C_{\gamma'}).$$

A positive $\bar{C}_{\mathrm{same}} - \bar{C}_{\mathrm{mixed}}$ means that mixed pairs do better. On optimal-tour prefixes with $u = 15$, averaged over CW and CCW sources, POMO's mean gap is 0.277% with mixed pairs and 3.749% with same-class pairs. Mixed pairs do better for POMO and Sym-NCO in all six TSP100 conditions, which combine Order 1, Order 2 and NN prefixes with $u = 15$ and $u = 50$, whereas same-class

pairs do better for AM in all six, and for LEHD and BQ-NCO the two kinds differ by less than 0.02 pp (Figure 5a for $u = 15$). With equal numbers of views, POMO trails LEHD by less with mixed pairs than with same-class pairs on optimal-tour and NN prefixes (Figure 5b).

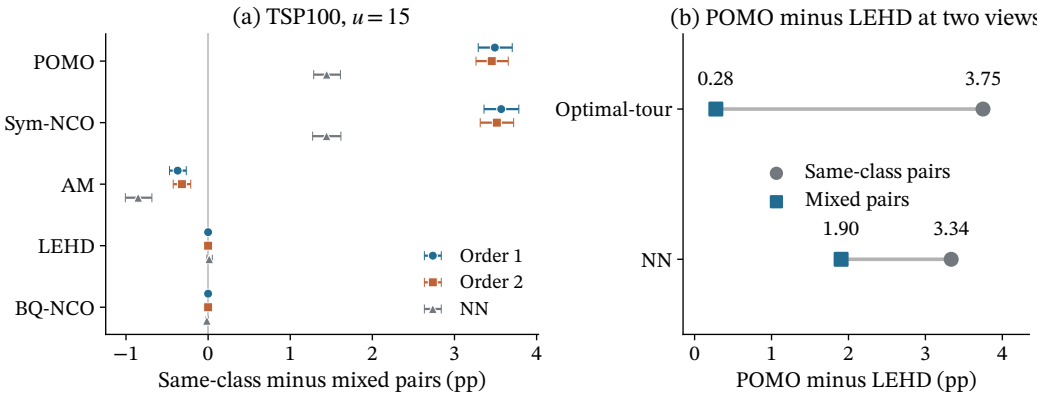

Figure 5: Two-view selection on TSP100 with same-class or mixed pairs. (a) Same-class minus mixed-pair mean gap at $u = 15$ (Order 1 and 2 as in Appendix B.2); multiplicity-adjusted 95% intervals over 42 contrasts, which also include those at $u = 50$ and POMO's and AM's at TSP20/50, not plotted but summarized in the text. (b) POMO's minus LEHD's mean gap for each kind of pair at $u = 15$, without intervals.

At TSP20 and TSP50, mixed pairs instead do better for AM in all six conditions (two sizes, Order 1, Order 2 and NN prefixes), as for POMO at every size, with positive multiplicity-adjusted intervals for both; AM's better kind of pair thus changes between its size-specific checkpoints. On TSP50 with NN prefixes, AM's mean gain from mixed pairs is 0.619 pp, but its median is -0.056 pp, and only 495 of 1,024 instances gain.

Two-view selection and cost disagreement. Each view lies in three of the twelve same-class pairs and in four of the sixteen mixed pairs, so both kinds average $(C_\gamma + C_{\gamma'})/2$ to the same value, $\frac{1}{8}\sum_\gamma C_\gamma$. Let Δ_{same} and Δ_{mixed} be the mean of $|C_\gamma - C_{\gamma'}|$ over the pairs of each kind. Applying $\min(C_\gamma, C_{\gamma'}) = (C_\gamma + C_{\gamma'} - |C_\gamma - C_{\gamma'}|)/2$ to every pair gives

$$\bar{C}_{\text{same}} = \tfrac{1}{8}\sum_\gamma C_\gamma - \tfrac{1}{2}\Delta_{\text{same}}, \qquad \bar{C}_{\text{mixed}} = \tfrac{1}{8}\sum_\gamma C_\gamma - \tfrac{1}{2}\Delta_{\text{mixed}},$$

so $\bar{C}_{\text{same}} - \bar{C}_{\text{mixed}} = \frac{1}{2}(\Delta_{\text{mixed}} - \Delta_{\text{same}})$. The identity holds for every instance and survives positive cost normalization and averaging. Same-class pairs present the source tour twice in one direction and mixed pairs in both, so mixed pairs do better on average exactly when costs disagree more, on average, across directions than within one.

A gain from mixed pairs alone does not reveal a convention. With costs $(0, 0, 10, 10)$ in one class and $(5, 5, 5, 5)$ in the other, both class means equal 5, yet $\bar{C}_{\text{same}} = 10/3$ exceeds $\bar{C}_{\text{mixed}} = 5/2$; the signed reflected-minus-original comparison of Section 3.2 is needed as well. A convention and a gain from covering both directions are therefore different properties, as AM shows by having a convention and doing better with same-class pairs at TSP100.

C.4 A BOUNDARY BEYOND TSP: CAPACITATED ROUTING

For the capacitated vehicle routing problem (CVRP), a separate set of 1,024 instances has 100 customers and a depot, coordinates uniform on the unit square, demands uniform on $\{1, \ldots, 9\}$ and capacity 50. Shared prefixes are cut from capacity-feasible NN tours and polar-angle sweep tours (Gillett & Miller, 1974) improved by a deterministic best-improvement feasible descent over 2-opt (Lin, 1965) and relocate moves (at most 100 accepted moves), either at a fixed number of customers left, 15 or 50, or at the nonterminal depot return nearest to 50 or 85 served customers; each prefix retains its visits and its remaining capacity. Cost differences are $100\,\Delta C/B$, in percentage points of B (pp of B), with $B = (C_{\text{NN}} + C_{\text{sweep}})/2$ the mean of the NN cost and the sweep cost before

improvement. Mixed pairs do better for POMO in every condition, and same-class pairs for AM and, by less than 0.03 pp of B with every interval below zero, for BQ-NCO (Figure 6); so on CVRP, too, the value of pairing a rotation with a reflection depends on the model. The direction counts of greedy tours, the reversed-endpoints and reflected-input comparisons of Section 3.2 and the Q^* analyses use TSP.

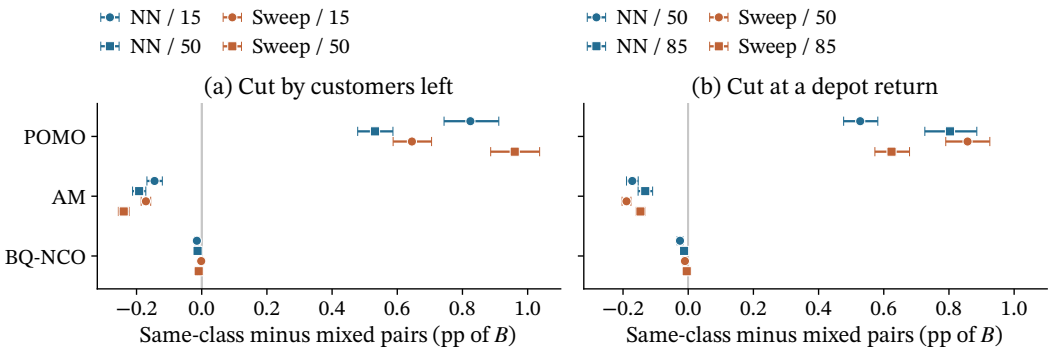

Figure 6: Same-class minus mixed-pair cost on CVRP100 (pp of B), all 1,024 instances in every condition. Bars: multiplicity-adjusted 95% intervals over the twelve contrasts of each panel.

C.5 Reflected input, reversed endpoints and CW and CCW sources change different things

Reversed endpoints (Section 2) keep the input and the cities left and swap the current city and the return city; the reflected input (Section 3.2) keeps the prefix's cities in order and the cities left and changes the coordinates. Both flip the prefix's direction, the only change they share. Cutting the prefix from the CCW source instead of the CW source, as between the rows of Table 1, also flips its direction; it keeps the return city but changes the current city, the prefix's edges and the cities left.

The same change of source applies to the model's own tours. From a prefix of its greedy tour, POMO reproduces that tour; from a prefix cut at the same start city from that tour traveled in the opposite direction, it completes a tour of higher mean cost in all 15 combinations of three sizes and five prefix lengths (Figure 7). The plotted difference bounds POMO's excess (Appendix B.1) from below: the greedy tour, of cost C_{own}, keeps both prefixes, so C_{own} is at least the cost $C^*(\tau)$ of the cheapest

Table 6: Reversed endpoints and the reflected input on TSP100 optimal-tour prefixes ($u = 15$, 1,024 instances): mean change in gap (pp). Columns: reflected minus original input and reversed minus original endpoints on the original input (the two contrasts of Figure 2), reversed minus original endpoints on the reflected input, and both operations minus neither. On the original input, POMO's CW sources and Sym-NCO's and AM's CCW sources are with the convention. Brackets: multiplicity-adjusted 95% intervals as in Figure 2, per column; 0.00 marks a magnitude below 0.005.

Model	Source	Reflected input	Reversed endpoints	Reversed endpoints, reflected input	Both operations
POMO	CW	7.83 [7.44, 8.22]	7.80 [7.40, 8.19]	-7.80 [-8.20, -7.40]	0.03 [-0.06, 0.13]
	CCW	-7.69 [-8.11, -7.29]	-7.73 [-8.14, -7.33]	7.76 [7.36, 8.15]	0.06 [-0.19, 0.31]
Sym-NCO	CW	-7.99 [-8.33, -7.66]	-8.00 [-8.34, -7.66]	7.81 [7.48, 8.15]	-0.18 [-0.39, 0.03]
	CCW	7.98 [7.65, 8.31]	7.86 [7.53, 8.18]	-8.02 [-8.35, -7.69]	-0.03 [-0.11, 0.04]
AM	CW	-3.39 [-4.69, -2.05]	-3.85 [-4.64, -3.05]	3.54 [2.72, 4.36]	0.16 [-0.97, 1.28]
	CCW	3.95 [2.63, 5.26]	3.60 [2.78, 4.42]	-3.49 [-4.29, -2.68]	0.46 [-1.02, 1.92]
LEHD	CW	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]
	CCW	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]
BQ-NCO	CW	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]
	CCW	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]	0.00 [0.00, 0.00]

tour that keeps the second prefix τ, and POMO's tour completed from τ, of cost $C_\pi(\tau)$, has excess $C_\pi(\tau) - C^*(\tau) = [C_\pi(\tau) - C_{\text{own}}] + [C_{\text{own}} - C^*(\tau)] \geq C_\pi(\tau) - C_{\text{own}}$. Like a change between CW and CCW sources, this test changes more than the direction; only the reflected input keeps the prefix's cities in order.

Reversed endpoints on the original and the reflected input. Table 6 crosses the two operations that keep the cities left. Reversing the endpoints of an optimal-tour prefix keeps its edges, and the rest of that tour still completes it, so the cheapest tour that keeps it still costs C^*, and every column is a difference of excesses (Appendix B.1). With original endpoints, the completions reproduce those behind Figure 2's reflected-input values exactly. AM's values come from a CUDA run on an L4 GPU, whose tie-breaking they reflect; CPU decoding breaks AM's frequent score ties differently but moves each AM contrast by at most 0.067 pp, with no sign change. For POMO, Sym-NCO and AM, reversed endpoints on the original input change the mean gap by about as much as the reflected input does and with the same sign, raising it for prefixes with the convention on the original input and endpoints and lowering it for those against it. On the reflected input, where with and against swap, the signs flip, and the two operations together leave the gap unchanged within intervals that cover zero. For LEHD and BQ-NCO, reversed endpoints, alone or with the reflected input, change the mean gap by less than 0.001 pp. As in Section 3.2, the gap follows the prefix's direction relative to the convention.

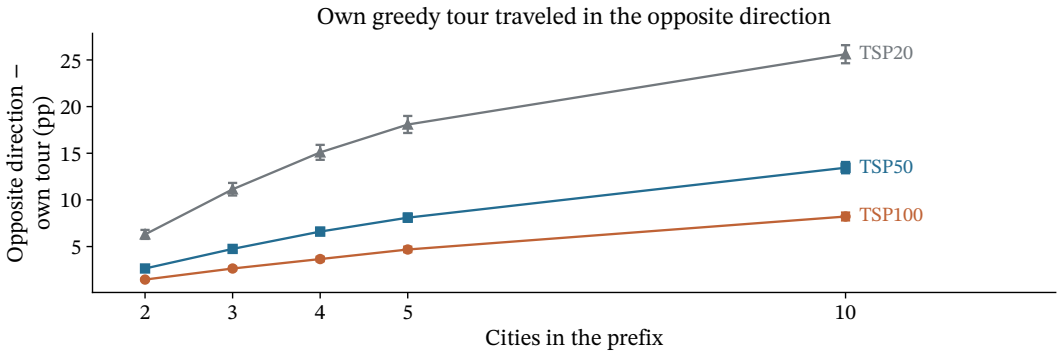

Figure 7: Prefixes of POMO's greedy tour traveled in the opposite direction: mean cost of the completed tour minus that from prefixes of the greedy tour itself (pp; same start city, original input). Bars: multiplicity-adjusted 95% intervals over the 15 combinations of size and prefix length.

D The Priority and Optimal Action Values

Prefixes of a model's own greedy tour (own-tour prefixes) are with its convention on original endpoints and against it on reversed endpoints (Section 2).

D.1 Derivations and the tie filter

Definitions. At state $s = (c, U, f)$, head ℓ attends over the cities left with weights $\alpha_{\ell j}(s)$, which sum to one over U and are computed from a query that contains c and f and, for AM and Sym-NCO, the mean of the city embeddings. The decoder adds the output bias b_o to the combined head outputs and computes each candidate's raw score as a scaled inner product with its embedding,

$$z_a(s) = \frac{h_a^\top}{\sqrt{d}}\Big(b_o + \sum_\ell W_{o,\ell} \sum_{j \in U} \alpha_{\ell j}(s) W_{v,\ell} h_j\Big),$$

which is Equation 1 with β_a and G_ℓ as defined there. For AM, the candidate's h_a in G_ℓ becomes its logit key $W_K h_a$, and $\beta_a = 0$ because AM's output projection has no bias. The cyclic part $\text{cyc}(M)$ of Section 4.1 is antisymmetric with zero row and column sums (Jiang et al., 2011), and the sign $\det(\gamma)$ of a rotation or reflection $\gamma \in D_4$ is the determinant of its linear part.

A reflection flips the sign of K_ℓ. Below, K and Y^γ denote either one head's K_ℓ and $\mathrm{cyc}(G_\ell^\gamma)$ or their sums over heads; every statement below is linear in the tables, so it holds per head and for the sum. Because every $\gamma \in D_4$ keeps the city labels, reindexing the sum over D_4 gives $K^\gamma = \det(\gamma)K$, where K^γ is K computed on the input transformed by γ: rotations preserve K, and reflections flip its sign. If Y^γ is the same for every γ, as it is when the embeddings are the same on every rotated and reflected input, then $K = 0$ because $\sum_{\gamma \in D_4} \det(\gamma) = 0$. As functions of γ, the eight tables Y^γ of an instance split into five D_4 *components* by how they transform under D_4, one for each of its four homomorphisms to $\{1, -1\}$ and one for its two-dimensional irreducible representation. The component for $\det$, evaluated at γ, is $\det(\gamma)K = K^\gamma$, so K^γ is an exact additive part of the interactions that the model uses on the input transformed by γ. The component that Figure 8 removes as a control, which flips under $(1 - x, y)$ and under quarter turns, is the one-dimensional component with signs $(1, -1, 1, -1, -1, 1, -1, 1)$ over $(\mathrm{id}, \rho_x, \rho_{180}, \rho_y, \rho_{90}, \rho_d, \rho_{270}, \rho_{ad})$, where $\rho_x : (x, y) \mapsto (1-x, y)$ gives the reflected input, $\rho_y : (x, y) \mapsto (x, 1-y)$, ρ_{90}, ρ_{180} and ρ_{270} rotate counterclockwise about $(1/2, 1/2)$, $\rho_d : (x, y) \mapsto (y, x)$ and $\rho_{ad} : (x, y) \mapsto (1-y, 1-x)$.

How the priority enters the raw score. As K_ℓ is antisymmetric, $\sum_{a \in U} \sum_{j \in U} K_\ell[j, a] = 0$, so $p_{\ell,U}$ sums to zero over U, and the rest $K_\ell - P_{\ell,U}$ of Section 4.1, here called the *remainder*, has column sums $\sum_{j \in U} K_\ell[j, a] - u\, p_{\ell,U}(a) + \sum_{j \in U} p_{\ell,U}(j) = 0$ over U. Being antisymmetric with zero column sums over U, the remainder is orthogonal to every table $\phi(a) - \phi(j)$ on U, so $P_{\ell,U}$ is the least-squares fit of K_ℓ on U by such tables, the global rankings of U (Jiang et al., 2011). Splitting each G_ℓ on U into $P_{\ell,U}$ and $G_\ell - P_{\ell,U}$ in Equation 1 and applying the identity of Section 4.1 to $P_{\ell,U}$ gives

$$z_a(s) = d_P(a; s) + r_a(s), \qquad r_a(s) = \beta_a + \sum_\ell \sum_{j \in U} \alpha_{\ell j}(s) \left(G_\ell - P_{\ell,U}\right)[j, a],$$

where $d_P(a; s) = p_U(a) - \bar{p}(s)$ is the contribution that removal subtracts and the offset $\bar{p}(s) = \sum_\ell \sum_{j \in U} \alpha_{\ell j}(s)\, p_{\ell,U}(j)$ is shared by all candidates. The raw score after removal, r_a, thus keeps β_a and every part of each G_ℓ except $P_{\ell,U}$, the remainder included.

A six-city example. Place cities 0 to 5 clockwise on a ring and let a hand-set table $K[j, a]$ be 1 if a lies one or two places before j clockwise, -1 if one or two places after, and 0 otherwise, an antisymmetric table with zero column sums that a reflection negates. After the prefix $0 \to 1$, with $c = 1$ and $f = 0$, the column mean of Equation 2 gives $p_U = (\frac{1}{2}, \frac{1}{4}, -\frac{1}{4}, -\frac{1}{2})$ on cities $2, 3, 4, 5$, negated on the reflected input, and so ranks them in the order of the clockwise completion from $c = 1$. With reversed endpoints, the model completes from $f = 0$ back to $c = 1$; U, and hence this ranking, is unchanged, whereas the optimal completion from $f = 0$ visits $5, 4, 3, 2$: the priority matches the optimal order from the original endpoints only.

Reversed endpoints leave the priority's margins unchanged. In $d_P(a; s) = p_U(a) - \bar{p}(s)$, the offset $\bar{p}(s)$ is shared by all candidates, so the priority's part of the margin between two candidates a and b is $p_U(a) - p_U(b)$, which, on a given input, depends on U alone; this proves that the priority's part of Equation 3 is the same from either endpoint, although the attention weights depend on the endpoints. The priority does change as cities are visited: with $K = \sum_\ell K_\ell$, $p_U(a) = -u^{-1} \sum_{j \notin U} K[j, a]$ (Section 4.1), and for $v \in U$ with $u > 1$ and $a \in U \setminus \{v\}$, visiting v gives $p_{U \setminus \{v\}}(a) = (u\, p_U(a) - K[v, a])/(u - 1)$. Conversely, only per-candidate values give margins that no attention distribution over U can change. If an interaction table M contributes the same margin to each pair $a, b \in U$ under every distribution with positive weights, as a softmax gives, then, as margins are linear in the weights, it does so under the distribution on a single city j as well, which shows that $M[j, a] - M[j, b]$ does not depend on j, and fixing b at a reference candidate b^* gives $M[j, a] = \mu(a) + \nu(j)$ on U. If M is antisymmetric, $M[a, a] = 0$ gives $\nu = -\mu$, so M has no cyclic part on U. Such margins therefore come only from per-candidate terms: the bias β_a and, for each fixed interaction table, a term $\mu(a)$ whose least-squares choice, with a residual table of zero column sums over U, is its column mean over U. Like the priority's, these margins do not change when the endpoints are reversed. The column means split into those of the symmetric part and the global ranking of each G_ℓ and those of the D_4 components of Y^γ. The priority is the one drawn from the component for $\det$, which flips sign under every reflection, and the analyses here do not show that it is the only part of the raw score in which the convention appears.

The tie filter and the raw-margin rule. Removing the priority subtracts the full d_P, offset included, before the clipping to $10\tanh$ and the masking of visited cities, a mean ablation across the candidates of a state (Section 4.1) rather than across a dataset. The offset cannot change which candidate has the highest raw score, but it shifts the raw scores before the clipping and can therefore create or resolve ties after it. Let a be the model's choice, b the choice after removal, and e_s the largest discrepancy at s between the model's raw scores and those reconstructed from the decomposition. A change passes the *tie filter* only if $a \neq b$, each is the unique winner of its own clipped scores, and a leads b in raw score before removal and b leads a after removal, both by more than $10^{-6} + 2e_s$. Because a has the highest raw score and b the highest raw score after removal, the priority then overrides the ranking of a and b by r_a (Equation 3). The tie filter uses scores alone and selects no sign of g. The *raw-margin rule* instead counts a change when a leads b by more than 10^{-7} in raw score before removal and trails it by more than 10^{-7} after. Table 2, Appendix D.5 and every count for AM use the raw-margin rule; all other counts of changed choices use the tie filter. For POMO and Sym-NCO on own-tour prefixes at $u = 10$, the two rules give identical counts (Tables 2 and 9).

D.2 Agreement with the model's visiting order

Agreement is the fraction of pairs $i < j$ of the model's visiting order $(v_1, \ldots, v_u)$ with $p_U(v_i) - p_U(v_j) > 10^{-6}$, so that a pair whose priorities differ by less is a tie and does not count as agreement (Sym-NCO's agreement across u and Tables 2 and 10 use 10^{-10}); an unrelated ranking without ties scores 50%. It is averaged over the original and reflected inputs of each of 1,024 instances per size, at every u from 2 to $n-1$, for POMO at all three sizes and for Sym-NCO at TSP100, and the values are means over instances.

On TSP100, POMO's agreement falls from 87.3% at $u = 20$ to 84.0% at $u = 70$ and 78.9% at $u = 99$, and Sym-NCO's is 78.9% at $u = 99$, the lowest value for both models. No POMO state with $u = n - 1$ has agreement 100% at any tested size, but with few cities left the priority orders them far more tightly than chance. Over the 4,096 states per size at $u = 6, 10$ (two inputs and two values of u per instance), the city POMO visits next is the unique top-priority candidate in 3471, 3126 and 2909 states at TSP20, TSP50 and TSP100, and agreement is 100% in 1617, 1199 and 1006 states, whereas a random ranking meets these conditions with probabilities $1/u$ and $1/u!$.

Geometric rankings. The geometric rankings tested do not explain the ranking with few cities left and explain it only in part with many. On the same TSP100 states at $u \in \{6, 10, 20, 30, 50, 70\}$, we compare p_U with five rankings, the only ones tested: an area ranking, which orders each candidate a by the mean over $j \in U$ of $\det(x_j - \bar{x},\ x_a - \bar{x})$, with x_j the position of city j and $\bar{x}$ the centroid of all cities, and four angular sweeps, which combine a center (the centroid of all cities or of those left) with a ray (toward the visited cities' centroid or the current city). All five are oriented by the model's convention; oriented against it, none exceeds 45.8% agreement. The priority agrees on more pairs than each of them at every u, with every pointwise interval of the difference above zero. At $u = 10$, they agree on at most 71.1% of pairs for POMO and 71.5% for Sym-NCO, against 89.9% and 88.9% for the priority. At $u = 70$, however, the sweep of Section 4.2 comes within 0.6 $[0.2, 1.1]$ percentage points of the priority for POMO and 0.6 $[0.2, 1.0]$ for Sym-NCO. On pairs led by one of the first five cities of the visiting order, the priority agrees on 90.6 to 91.9% at every u, against at most 85.9%, but at $u = 50$ and 70 the sweep around the same center, started toward the current city, ranks the model's next city first 9.3 to 11.3 percentage points more often than the priority. The comparison concerns rankings, not what the network computes.

D.3 Controls for removing the priority

On own-tour prefixes on reversed endpoints with $u = 6$ or 10, removal lowers the mean Q^* of the choice, and three controls do not reproduce this: removing another part of the raw score, giving the same values to other candidates (the contrast Ξ below), and the noise control, which changes choices as often. Each keeps some properties of removal, so together they are specific comparisons, not a null for every score change.

The part removed matters. On POMO's own-tour prefixes on reversed endpoints, with 1,024 instances at each of TSP20, TSP50 and TSP100 (Figure 8), removal lowers the mean Q^* of the

choice at every size, and subtracting half of d_P lowers it less. Removing the remainder $K_\ell - P_{\ell,U}$, which changes sign under every reflection as the priority does, raises it instead, and removing the one-dimensional D_4 component of Appendix D.1 that flips sign under $(x, y) \mapsto (1 - x, y)$, as K_ℓ does, but also under quarter turns leaves it nearly unchanged. For Sym-NCO at $u = 10$ on the original input, the mean step excess $Q^*(s, v) - V^*(s)$ of the chosen city v is 2.53 [2.39, 2.67] for the model's choice, 0.87 [0.79, 0.95] after removal of the priority and 3.54 [3.32, 3.76] after removal of the remainder (pp, pointwise 95% intervals), with the same sign of change at $u = 6$ and on the reflected input.

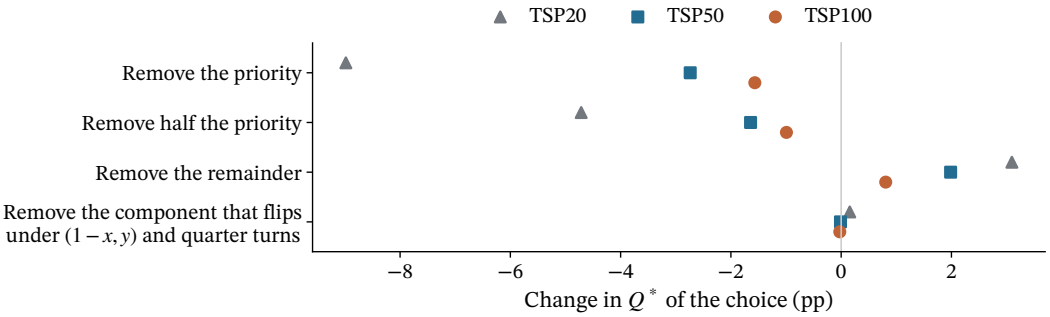

Figure 8: Four removals from the raw score (Appendix D.1) on POMO's own-tour prefixes on reversed endpoints: mean change in Q^* of the choice over the original and reflected inputs and $u = 6, 10$, unchanged choices counting zero. Only removal of the priority has intervals (Table 7).

The assignment matters. The difference between original and reversed endpoints depends on which candidate receives each value, not on the values alone. To test this, we replace the priority in a recipient view γ by its value in a donor view γ' and keep the recipient's attention and the rest of its raw score. By the sign law $K^\gamma = \det(\gamma)K$ (Appendix D.1) over one instance's views, with the same U,

$$z_{\gamma \leftarrow \gamma'} = z^\gamma - d_P^\gamma + \sum_\ell (\alpha_\ell^\gamma)^\top P_{\ell,U}^{\gamma'} = z^\gamma + [\det(\gamma')\det(\gamma) - 1] d_P^\gamma,$$

so a donor of the recipient's class changes nothing, and a donor of the opposite class flips the sign of the priority by adding $-2d_P^\gamma$. Here $\gamma, \gamma' \in \{\mathrm{id}, \rho_x\}$, the original and the reflected input. For an own tour constructed on input k, with $\bar{k}$ the other input, let $\eta_{k,o}^{\gamma,\gamma'}$ be the step excess of the chosen city under recipient γ and donor γ', with $o = \mathrm{orig}$ for a prefix on original endpoints and $o = \mathrm{rev}$ for one on reversed endpoints. The contrast

$$\Xi_k = \tfrac{1}{2} \sum_\gamma \left[(\eta_{k,\mathrm{rev}}^{\gamma,k} - \eta_{k,\mathrm{rev}}^{\gamma,\bar{k}}) - (\eta_{k,\mathrm{orig}}^{\gamma,k} - \eta_{k,\mathrm{orig}}^{\gamma,\bar{k}}) \right]$$

is positive when flipping the sign of the priority of the tour's input saves more on reversed endpoints than on original endpoints. Replacing the sign-flipping perturbation $-2d_P^\gamma$ by 64 permutations of the same vector over the cities left, shared across recipients and across original and reversed endpoints within each instance, source and u, keeps its values before clipping, shift, range and norm but not the assignment. For POMO at TSP20, TSP50 and TSP100 and for Sym-NCO at TSP100, the actual assignment gives a larger Ξ than these shuffles, with multiplicity-adjusted intervals above zero over 1,024 instances (Table 7).

Matching how often choices change. The noise control changes the choice as often as removal, but to a cheaper city far less often. It adds independent Gaussian noise to the raw scores of the candidates before clipping, with 64 draws per state and one scale per model, u and endpoints ($\sigma = 0.78$ to 1.60), set from the scores and chosen cities on half of the instances. The estimates use the other 512 instances, both inputs and the tie filter. On reversed endpoints, 93.2 to 97.9% of removal's changed choices go to a cheaper city, and 38.3 to 47.6% of the noise control's (Table 8). At the 550 to 697 states per model and u where removal changes the choice, the noise control's changes go to a cheaper city in 54.9 to 62.8%, about as often as the runner-up and a random other candidate, and removal exceeds it by 33.2 to 38.4 percentage points. Averaged over all states, removal lowers the

Table 7: Removal and the contrast Ξ on own-tour prefixes, $u = 6$ and 10 pooled (pp; own tours constructed on the original and reflected inputs unless a row names one). Removal rows: mean change in Q^* of the choice over all states, unchanged choices counting zero. Brackets: multiplicity-adjusted 95% intervals over 18 contrasts for POMO (15 shown; the other three, one per size, use a cost bound not reported here) and five for Sym-NCO.

Contrast	POMO20	POMO50	POMO100	Sym-NCO100
Removal, reversed endpoints	-8.984 [-9.559, -8.421]	-2.738 [-2.946, -2.538]	-1.566 [-1.675, -1.460]	-1.680 [-1.783, -1.576]
Removal, reversed − original endpoints	-9.679 [-10.304, -9.061]	-2.916 [-3.136, -2.702]	-1.739 [-1.864, -1.616]	-1.959 [-2.091, -1.832]
Ξ minus shuffled Ξ	26.407 [25.568, 27.251]	10.512 [10.100, 10.921]	5.244 [5.026, 5.465]	5.195 [5.004, 5.385]
Ξ, original input	29.506 [28.152, 30.852]	10.930 [10.266, 11.603]	5.423 [5.068, 5.787]	5.764 [5.452, 6.077]
Ξ, reflected input	29.527 [28.193, 30.870]	11.101 [10.445, 11.776]	5.615 [5.267, 5.967]	5.508 [5.189, 5.832]

Table 8: Removal, the noise control and two alternatives on own-tour prefixes on reversed endpoints (the 512 instances not used for calibration, both inputs, tie filter). Same states: those at which removal changes the choice; runner-up and random other candidate: the raw-score runner-up and a uniformly drawn other candidate there. Brackets: pointwise 95% intervals.

	POMO		Sym-NCO	
Changed choices to a cheaper city (%)	$u = 6$	$u = 10$	$u = 6$	$u = 10$
Same states (count)	656	550	697	603
Removal	97.9 [96.7, 98.9]	93.8 [91.6, 95.8]	96.0 [94.2, 97.5]	93.2 [91.2, 95.2]
Noise control, all states	45.3 [43.0, 47.5]	38.3 [36.1, 40.6]	47.6 [45.4, 49.8]	40.4 [38.3, 42.7]
Noise control, same states	62.1 [60.0, 64.2]	55.6 [53.0, 58.2]	62.8 [60.6, 65.0]	54.9 [52.5, 57.3]
Runner-up, same states	62.7 [59.1, 66.3]	61.5 [57.3, 65.5]	61.4 [57.6, 64.9]	60.0 [56.2, 63.8]
Random other candidate, same states	64.5 [62.5, 66.5]	49.2 [47.1, 51.3]	65.6 [63.6, 67.5]	50.2 [48.2, 52.3]
Removal minus noise control, same states (percentage points)	35.7 [33.6, 37.9]	38.2 [35.5, 40.8]	33.2 [31.2, 35.2]	38.4 [35.7, 40.9]

mean Q^* of the choice by 1.40 to 1.84 pp, whereas the noise control raises it by 0.07 to 0.40, with pointwise intervals above zero except for Sym-NCO at $u = 6$. On original endpoints, removal raises it by 0.07 to 0.28 pp more than the noise control does, with pointwise intervals above zero for both models at both values of u. The shuffles keep the perturbation's values but not how often choices change, and the noise control keeps how often but not the assignment; neither keeps the spacing of the scores after clipping.

D.4 Changed choices and g across u

Table 9 extends Table 2's changed choices to u from 6 to 70, and Figure 3 gives the share of them with $g > 0$. On reversed endpoints, the share of changed choices falls as u grows, but the mean g is positive with intervals above zero at every u, and the share with $g > 0$ stays above one half: at $u = 70$, $g > 0$ for 286 of POMO's 356 and 277 of Sym-NCO's 330 changed choices. On original endpoints, the mean g has intervals below zero at every u except $u = 50$, where only 8 of 27 POMO and 5 of 16 Sym-NCO changed choices have $g > 0$, yet their mean g is positive with intervals that include zero: the few changes to a cheaper city outweigh the many to a costlier one. At $u = 70$,

Table 9: Changed choices on own-tour prefixes (% of the 2,048 states per model, u and endpoints; tie filter) and their mean g (pp). Brackets: pointwise 95% intervals.

		POMO		Sym-NCO	
u	Endpoints	Changed (%)	Mean g	Changed (%)	Mean g
6	Original	7.3	-1.607	9.7	-2.119
		[6.3, 8.3]	[-1.918, -1.315]	[8.4, 11.0]	[-2.433, -1.811]
10	Original	6.3	-3.620	7.7	-4.619
		[5.2, 7.4]	[-4.359, -2.932]	[6.5, 8.8]	[-5.175, -4.094]
20	Original	2.1	-0.341	1.9	-0.882
		[1.6, 2.8]	[-0.773, -0.092]	[1.4, 2.5]	[-1.578, -0.300]
30	Original	1.6	-0.144	1.9	-0.153
		[1.1, 2.1]	[-0.241, -0.066]	[1.3, 2.4]	[-0.284, -0.043]
50	Original	1.3	0.102	0.8	0.035
		[0.8, 1.9]	[-0.138, 0.537]	[0.4, 1.2]	[-0.084, 0.184]
70	Original	0.7	-0.287	0.8	-0.181
		[0.3, 1.1]	[-0.538, -0.063]	[0.4, 1.2]	[-0.343, -0.029]
6	Reversed	63.3	2.654	66.7	2.624
		[61.1, 65.4]	[2.567, 2.748]	[64.6, 68.9]	[2.525, 2.723]
10	Reversed	52.9	2.745	58.4	2.754
		[50.7, 55.2]	[2.614, 2.876]	[56.1, 60.7]	[2.621, 2.882]
20	Reversed	36.1	2.742	39.0	2.834
		[33.9, 38.3]	[2.566, 2.937]	[36.9, 41.2]	[2.653, 3.015]
30	Reversed	29.4	3.084	30.3	2.710
		[27.4, 31.3]	[2.836, 3.358]	[28.2, 32.3]	[2.501, 2.932]
50	Reversed	20.5	2.851	20.8	2.898
		[18.8, 22.2]	[2.488, 3.238]	[19.0, 22.7]	[2.585, 3.227]
70	Reversed	17.4	2.452	16.1	2.680
		[15.8, 19.0]	[2.114, 2.777]	[14.5, 17.7]	[2.359, 3.032]

only 4 of 14 POMO and 6 of 16 Sym-NCO changed choices on original endpoints have $g > 0$, so the shares on original endpoints at $u = 50$ and 70 rest on few changes and some of their intervals include one half.

D.5 Other checkpoints and instance sets

For every released checkpoint tested, the priority ranks most pairs in the model's visiting order. Table 10 compares eight checkpoints on four sets of 1,024 TSP100 instances: the test set and new uniform, clustered (three clusters with standard deviation 0.07) and half-clustered (half clustered, half uniform) sets. The six POMO checkpoints are POMO's long- and standard-training weights, the first of which is used elsewhere, Sym-NCO's POMO baseline and three AMDKD teachers (Bi et al., 2022). AM constructs from its own start city, the others from city 0. On the original input of all four sets, POMO-long, POMO-standard, POMO-baseline and AMDKD-uniform construct CW, and AMDKD-cluster, AMDKD-mixed and Sym-NCO CCW. For these seven checkpoints on every set, agreement at $u = 10$ lies between 81.3% and 90.2%, and removal changes 5.3% to 10.3% of the choices on original endpoints against 39.0% to 58.4% on reversed endpoints. Agreement is lower for most checkpoints on the clustered and half-clustered sets, and AM reaches only 63.0% on the clustered set. Shared NN prefixes, whose direction no model sets, narrow the difference between original and reversed endpoints: on the test set, removal changes 29.5% of POMO's choices on original endpoints and 37.0% on reversed endpoints, against the own-tour shares of Table 2. All values here are at $u = 10$, with changes counted by the raw-margin rule of Appendix D.1.

Table 10: The priority's agreement (%) at $u = 10$ for eight released checkpoints on four sets of TSP100 instances; means without intervals. Shares of changed choices at $u = 10$ are given in the text.

Checkpoint	Test	New uniform	Clustered	Half-clustered
POMO-long	89.9	90.2	82.8	83.7
POMO-standard	89.9	90.0	83.1	84.1
POMO-baseline	89.1	89.5	83.2	84.0
AMDKD-uniform	88.8	89.6	82.3	82.9
AMDKD-cluster	87.5	87.4	87.3	81.3
AMDKD-mixed	84.3	85.2	82.2	86.1
AM	85.3	85.3	63.0	82.3
Sym-NCO	88.9	89.5	82.1	84.3

E The Priority Along the Model's Own Completions

E.1 Completions from optimal-tour prefixes

For each number $u_0 = 10, 25, 50, 70$ of cities left at the first choice, each of the 1,024 TSP100 test instances supplies eight prefixes: those cut from its CW and CCW sources (Appendix B.2), on the original and the reflected input, each with original and with reversed endpoints. Reflecting the input and reversing the endpoints each flip a prefix's direction, so either operation moves a prefix between with and against the convention; each instance supplies four prefixes with each model's convention and four against it, and POMO and Sym-NCO each complete 8,192 prefixes per u_0. With and against refer to the prefix a completion starts from; later states are not reclassified. At every state, the model's choice and the choice after removal are priced by their optimal action values Q^*; the model follows only its own choice.

E.2 The priority's mean effect along a completion

To follow removal's effect without conditioning on a changed choice, every completion enters the mean at every step. At the state s_t of step t, let a_t be the model's choice and b_t the choice after removal, let $\mathbf{1}_t$ indicate a changed choice that passes the tie filter (Appendix D.1), and let $g_t = Q^*(s_t, a_t) - Q^*(s_t, b_t)$ be the saving g of Section 4.2 before division by the optimal tour cost. For the completions with the convention, and separately for those against it, the mean g per step is

$$\bar{g}_t = \frac{1}{N} \sum_{i=1}^{N} \frac{1}{4} \sum_{w=1}^{4} \frac{100\, \mathbf{1}_{iwt}\, g_{iwt}}{C_i^*}, \qquad N = 1{,}024, \tag{7}$$

where C_i^* is the optimal tour cost of instance i and w indexes its four completions with (or against) the convention. Positive values mean that removal lowers Q^* of the choice on average. Steps without a changed choice contribute zero, so $\bar{g}_t$ combines how often removal changes the choice with the g of those changes. At $t = u_0 - 1$ only one city is left; the earlier steps are the *steps with a choice*, and those after the first choice form three consecutive thirds of nearly equal size, early, middle and late, with any extra steps in the earlier thirds.

At $u_0 = 70$ (Figure 9), the mean against the convention at the first choice is 0.379 $[0.335, 0.423]$ pp for POMO and 0.427 $[0.380, 0.478]$ for Sym-NCO, and it falls sharply over the next few steps, whereas completions with the convention show little early effect, with a first-choice mean of -0.008 $[-0.014, -0.003]$ for Sym-NCO. At $u_0 = 25$, 50 and 70, the mean against the convention is positive at the first choice, and the mean over the late third is negative with and against the convention and for both models, from -0.205 to -0.023 pp per step, with pointwise intervals below zero. The mean per step is lowest with 7 to 10 cities left at all three values of u_0, so the minimum falls at a similar number of cities left rather than at a similar step. At $u_0 = 25$ and 50, the mean against the convention is positive again, with pointwise intervals above zero, at one or more of the last three steps with a choice. At $u_0 = 10$, the mean against the convention remains positive at the first choice and in each third, with a late-third mean of 0.058 $[0.045, 0.071]$ pp per step for POMO and 0.057

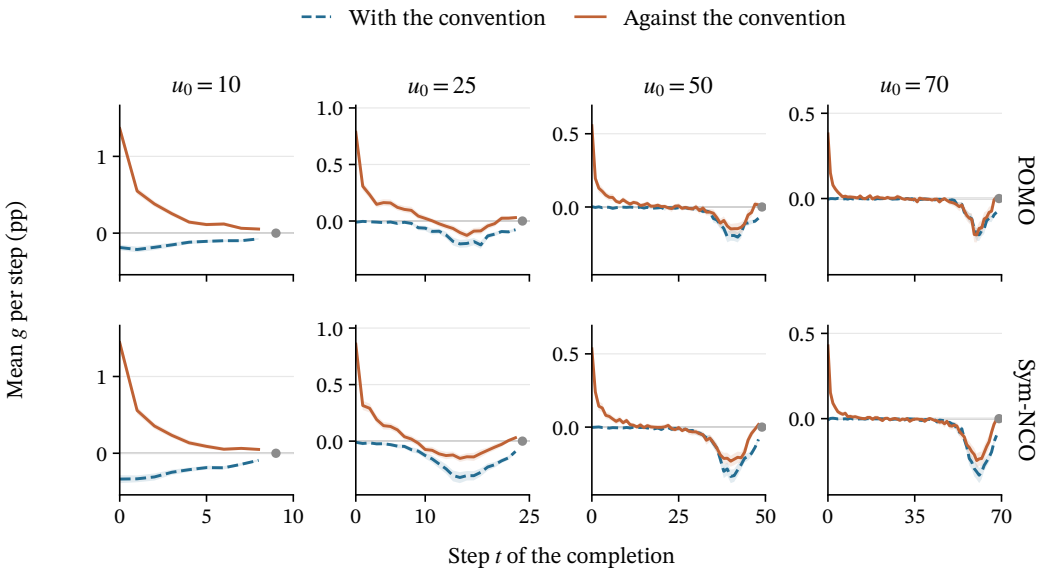

Figure 9: Mean g per step, $\bar{g}_t$ (Equation 7), along completions of optimal-tour prefixes at all four u_0. Grey dot: the last step, with one city left. Bands: pointwise 95% intervals; models share a scale per u_0.

$[0.045, 0.069]$ for Sym-NCO, whereas with the convention the mean at the same u_0 is negative from the first choice.

Comparison with the noise control. At the first choice, removal lowers Q^* of the choice more than the noise control does; in the late third, it is not distinguished from the noise control at $u_0 = 50$ and does better at $u_0 = 25$, where the noise control raises Q^* more than removal does. A negative mean in the late third need not reflect the priority, because wherever the model's choice is optimal, every other city costs at least as much. We therefore compare removal with the noise control of Appendix D.3 on the four completions against the convention of each of the 103 instances, every tenth, for which Q^* of every candidate was stored at $u_0 = 25$ and 50. The noise control is calibrated on the other instances, with one scale per model, u_0, step, and prefix with or against the convention, so that it changes the choice as often as removal does. At the first choice, removal's mean g exceeds the noise control's in all four combinations of model and u_0, by 0.62 $[0.43, 0.83]$ to 1.06 $[0.82, 1.35]$ pp (multiplicity-adjusted 95% intervals), and 86 to 91% of the first choices that removal changes go to a cheaper city, against 39 to 43% of those that the noise control changes. In the late third, where the model's choice is optimal at 79 to 85% of states, the mean g per step of removal minus that of the noise control is not distinguished from zero at $u_0 = 50$, -0.01 $[-0.06, 0.04]$ pp per step for POMO and -0.06 $[-0.13, 0.01]$ for Sym-NCO, and positive at $u_0 = 25$, 0.15 $[0.08, 0.23]$ and 0.09 $[0.02, 0.16]$. The first-choice mean minus the late-third mean exceeds that of the noise control in all four combinations, by 0.63 $[0.45, 0.82]$ to 0.97 $[0.75, 1.23]$ pp (multiplicity-adjusted). The intervals condition on the 64 noise draws per state, and no noise-control comparison exists at $u_0 = 10$ or 70.

Changes of both signs. A small mean can combine changes of both signs. Replacing $\mathbf{1}_t g_t$ in Equation 7 by $\mathbf{1}_t[g_t]_+$ or $\mathbf{1}_t[-g_t]_+$, with $[x]_+ = \max(x, 0)$, splits $\bar{g}_t$ by the sign of g_t. In the middle third (steps 24 to 46) of completions against the convention at $u_0 = 70$, 41.7% of POMO's changed choices have $g > 0$ and 58.3% have $g < 0$, contributing 0.012 and 0.015 pp per step (shares and terms without intervals; net -0.0028 $[-0.0052, -0.0004]$), and Sym-NCO's shares are 38.2% and 61.8%, with a net mean of -0.0064 $[-0.0084, -0.0043]$ pp per step.

E.3 Why a later comparison cannot undo excess already incurred

Whether the priority favors the cheaper of two cities at a state differs from whether that state still admits a completion as cheap as one that an earlier choice excluded. The deterministic finite-horizon

case of the performance-difference lemma (Ross & Bagnell, 2014, Lemma 4.3), with optimal completion as the reference and the return edge in the terminal value, separates the two.

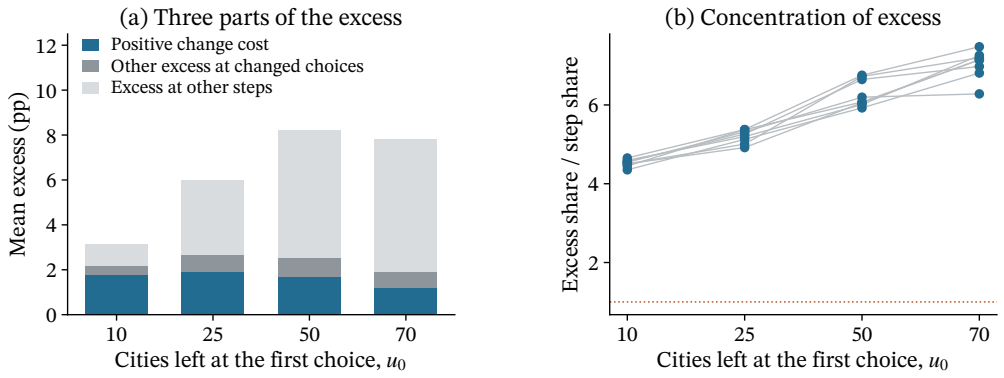

Figure 10: POMO's excess along completions of optimal-tour prefixes. (a) The three parts of Equation 8, averaged over the eight optimal-tour conditions. (b) Concentration ratio of changed choices per optimal-tour condition; dotted line: proportionality.

Proposition E.1 (Excess as a sum of step excesses)**.** *For the states s_t and choices a_t of a feasible completion by the model from s_0, the step excesses $\eta_t = Q^*(s_t, a_t) - V^*(s_t)$ are nonnegative and sum to the excess (Appendix B.1), $E_\pi(s_0) = \sum_{t=0}^{u_0-1} \eta_t$, where the last term, with one city left, is zero.*

Proof. At every step, including the last, $Q^*(s_t, a_t) = W_{c_t a_t} + V^*(s_{t+1})$, where c_t is the current city at s_t, so $\eta_t = W_{c_t a_t} + V^*(s_{t+1}) - V^*(s_t)$, and summing over the completion cancels the intermediate values. After the last choice, no city is left and $V^*(c, \varnothing, f) = W_{cf}$ is the return edge that the model also pays, so the sum equals $E_\pi(s_0)$. Each term is nonnegative because $V^*(s_t) = \min_v Q^*(s_t, v)$. □

Over the first T steps, the same cancellation gives the excess already incurred,

$$\sum_{t<T} \eta_t = \sum_{t<T} W_{c_t a_t} + V^*(s_T) - V^*(s_0),$$

which cannot decrease as T grows, because every η_t is nonnegative. The model's choice can therefore beat the choice after removal, $g_t < 0$, and still add excess, $\eta_t > 0$, when both are costlier than the best city at that state; such a comparison subtracts nothing from the excess already incurred.

Locating excess at changed choices. Let $\kappa_t = \mathbf{1}_t[g_t]_+$ denote the positive change cost, which is nonzero only at a changed choice where the priority favors the costlier city. Because the choice after removal cannot beat the optimum, $[Q^*(s_t, a_t) - Q^*(s_t, b_t)]_+ \le Q^*(s_t, a_t) - V^*(s_t) = \eta_t$, so $0 \le \kappa_t \le \mathbf{1}_t \eta_t$, which gives $\eta_t \ge g_t$ at a changed choice with $g_t > 0$, and

$$\eta_t = \underbrace{\kappa_t}_{\text{positive change cost}} + \underbrace{(\mathbf{1}_t \eta_t - \kappa_t)}_{\text{other excess at changed choices}} + \underbrace{(1 - \mathbf{1}_t)\eta_t}_{\text{excess at other steps}}, \tag{8}$$

with three nonnegative parts and $\mathbf{1}_t = 0$ when fewer than two cities are left. At a changed choice with $g_t > 0$, the middle part is the step excess $Q^*(s_t, b_t) - V^*(s_t)$ that the choice after removal, a cheaper city, still carries. Summed along a completion, the three parts give the excess (Proposition E.1; Figure 10a). The partition locates excess but assigns no causal share of it to the priority: negative values of g_t cannot cancel its parts, and the completions after forced choices (Appendix F.1) do not enter it.

The parts are summed along each completion in pp of the instance's optimal tour cost and then averaged, with step shares over the steps with a choice. The eight optimal-tour conditions per

u_0 combine the stored order of the optimal tour (Order 1 or 2, Appendix B.2), the input and the endpoints, so each pools prefixes with and against the convention. For POMO at $u_0 = 70$, changed choices make up 3.3 to 3.7% of the steps with a choice but carry 21 to 25% of the excess in each condition, a concentration ratio (excess share over step share) of 6.28 to 7.48 (Figure 10b); at $u_0 = 10$, they make up 15.1 to 16.3% of the steps with a choice and carry 69 to 71%. Completions of NN prefixes on the same instances (four conditions per u_0) give ratios of 4.57 to 5.41 at $u_0 = 70$. These values are ranges across conditions, not intervals. Changed choices thus carry several times their share of the excess, yet on long completions most of the excess arises at steps without a changed choice. No concentration ratio is reported for Sym-NCO.

F Cheaper Cities and the Model's Own Completion

F.1 Forced cheaper cities along the model's completions

Each forced run starts separately at a changed choice (tie filter, Appendix D.1) along the completions of Appendix E.1, so forced choices are never combined and their completions supply no new states. Over the four values of u_0, these completions contain 67,312 changed choices for POMO and 79,910 for Sym-NCO, of which Figure 4 keeps those to a cheaper city, 30,101 and 31,588; the selection uses Q^*, not the cost of the model's tour. Per model and u_0, completions with the convention supply 468 to 1,694 forced choices from 287 to 749 instances, and those against it 5,606 to 7,443 from at least 1,017 instances.

Figure 4 gives intervals for the rates of Section 5 and both models' mean g and R at every u_0. At $u_0 = 70$, forced choices against the convention have mean g 1.655 and mean R -0.013 $[-0.106, 0.083]$ pp for POMO and 1.659 and -0.045 $[-0.126, 0.040]$ for Sym-NCO. Frequent shortening does not imply a positive mean R: for Sym-NCO at $u_0 = 25$, 51.9% of forced choices against the convention shorten the model's tour, yet their mean R is -0.072 pp $[-0.141, -0.002]$.

Figure 11 divides the forced choices into the first choice and the early, middle and late thirds (Appendix E.2). In each, for every u_0 and model, a larger share shortens the model's tour with the convention than against it, with separated pointwise intervals except at POMO's first choice with $u_0 = 50$, both models' first choice with $u_0 = 70$ and POMO's late third with $u_0 = 10$, although the mean g is larger against the convention. Against the convention at $u_0 = 70$, early-third forced choices have mean g of 1.779 pp for POMO and 1.835 for Sym-NCO but mean R of -0.351 $[-0.534, -0.176]$ and -0.337 $[-0.507, -0.177]$. In the middle third, 62.3% of POMO's and 64.1% of Sym-NCO's forced choices shorten the model's tour, yet their mean R is -0.205 $[-0.406, -0.007]$ and -0.243 $[-0.443, -0.042]$. For every u_0, late-third forced choices against the convention shorten the model's tour more often than early-third ones, with positive point estimates of mean R for both models, so the near-zero means of R over whole completions at $u_0 = 70$ combine negative early and middle means with positive late ones.

F.2 Where the saving of a cheaper city goes

Proposition F.1 (Savings under optimal and model completion)**.** *Force a or b once at the same state s and let the model complete the tour; $Q^\pi(s, v)$ includes the forced edge to v and the model's completion from the state s_v after v. With $g = Q^*(s, a) - Q^*(s, b)$, $R = Q^\pi(s, a) - Q^\pi(s, b)$ and $E_\pi(s) = V^\pi(s) - V^*(s)$ (Appendix B.1), Equation 4 holds with the excess after v equal to $E_\pi(s_v)$: $R = g - [E_\pi(s_b) - E_\pi(s_a)]$.*

Proof. For $v \in \{a, b\}$, $Q^\pi(s, v) = W_{cv} + V^\pi(s_v) = Q^*(s, v) + E_\pi(s_v)$, because $Q^*(s, v) = W_{cv} + V^*(s_v)$ and $E_\pi(s_v) = V^\pi(s_v) - V^*(s_v)$; the edge to v belongs to $Q^*(s, v)$, and both completions retain the return edge. So the excess after v, $Q^\pi(s, v) - Q^*(s, v)$, is $E_\pi(s_v)$, and subtracting the case b from the case a gives the identity; with whole tour costs, the common prefix cancels. □

Because b departs once from the model's choice, s_b is generally a state the model would not otherwise reach. With the convention at $u_0 = 70$, the estimated mean of $R - g$ is positive for both models, 0.055 $[-0.038, 0.152]$ pp for POMO and 0.088 $[0.013, 0.160]$ for Sym-NCO (pointwise 95% intervals of the paired difference), but only Sym-NCO's interval excludes zero, so only for Sym-NCO is

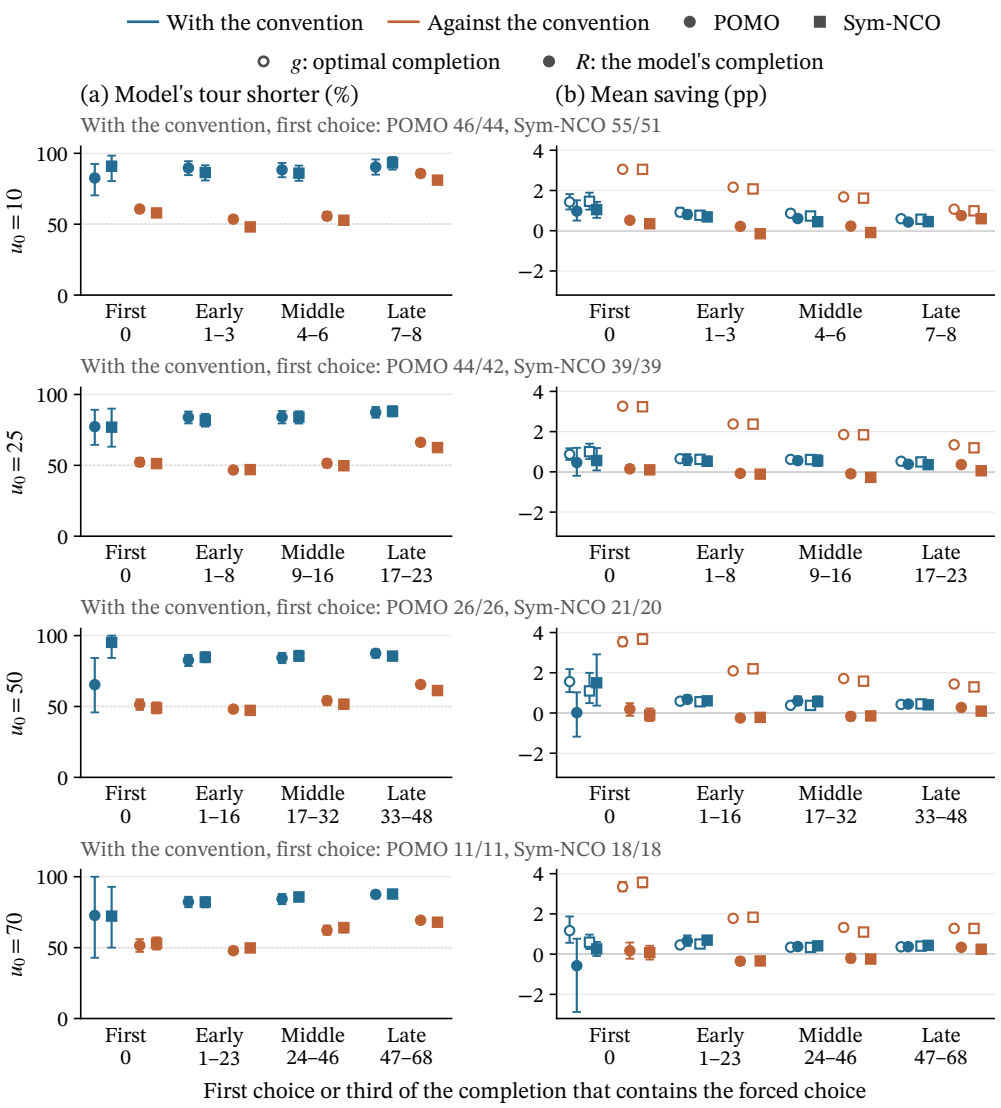

Figure 11: Forcing a cheaper city ($g > 0$) on optimal-tour prefixes, at the first choice and in each third of the completion (Appendix E.2). Bars: pointwise 95% intervals. Annotations: forced choices/instances at the first choice with the convention.

the lower mean excess after b than after a detectable. The identity locates where g is lost, not why completions with and against the convention differ: these are labeled only by their prefix, and the familiarity, endpoint compatibility and other properties of their states were not measured.

F.3 Own-tour prefixes and removal at every first choice

Own-tour prefixes. With the forced choices and the $g > 0$ selection of Appendix F.1, own-tour prefixes with reversed endpoints and $u = 70$ show no detectably positive mean R. Of the 286 POMO and 277 Sym-NCO changed choices to a cheaper city on these prefixes (Appendix D.4), 140 and 124 also shorten the model's tour and 146 and 153 make it cost more, shares of 49.0% $[43.0, 55.1]$ and 44.8% $[38.7, 50.6]$. The signs of g and R also disagree the other way: 22 of POMO's 70 and 13 of Sym-NCO's 53 changed choices to a costlier city shorten the model's tour. With original endpoints, all 4 POMO and 6 Sym-NCO changed choices to a cheaper city, rare among 2,048 states, shorten the model's tour, without intervals because some resamples contain none.

Every first choice, including unchanged ones. Removing the priority at the first choice of every prefix with ten cities left, including those where the choice after removal is a costlier city, raises both models' mean excess after the first choice in all twelve conditions, and in none of them does it make the model's tour detectably shorter. Each of the twelve conditions of Figure 12 uses all 1,024 instances. The changes after minus before removal in Q^* of the first choice, $-g$, and in the model's tour cost, $-R$, are positive when removal is worse, and unchanged choices contribute zero. By Equation 4, their difference $g - R = E_\pi(s_b) - E_\pi(s_a)$ is the change in excess after the first choice, computed within each instance, with intervals of the paired difference. On optimal-tour prefixes, removal lowers the mean Q^* of the first choice in all eight conditions for both models (Figure 12a for POMO; for Sym-NCO, $-g$ lies between -0.646 and -0.443 pp with every multiplicity-adjusted interval below zero), and $g - R$ is positive in all twelve conditions (for Sym-NCO between 0.495 and 0.750 pp, every multiplicity-adjusted interval excluding zero). For the model's tour cost, POMO's intervals in the optimal-tour conditions include zero, whereas Sym-NCO's eight means there are positive, four with multiplicity-adjusted intervals that exclude zero, and in the four NN conditions both models' intervals lie above zero.

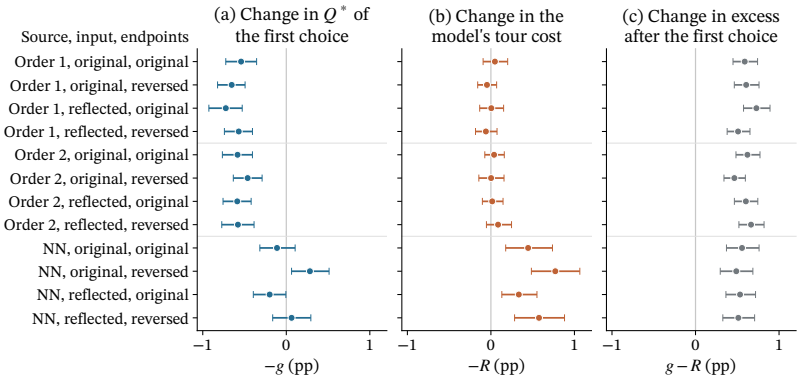

Figure 12: Removing POMO's priority at the first choice of every prefix with $u_0 = 10$ (after minus before removal). Bars: multiplicity-adjusted 95% intervals over 24 contrasts in (a) and (b) together and twelve in (c). Rows: source (Order 1 or 2, Appendix B.2, or NN), original or reflected input, original or reversed endpoints.