LLMAide: Language-Assisted Neural Solver for Vehicle Routing Problems

Extended Abstract

Manuj Malik*
Singapore Management University
Singapore, Singapore
manujm@smu.edu.sg

Jianan Zhou
Nanyang Technological University
Singapore, Singapore
jianan004@e.ntu.edu.sg

Yan Jin
Huazhong University of Science and Technology
Wuhan, China
yuandong@hust.edu.cn

Zhiguang Cao
Singapore Management University
Singapore, Singapore
zgcao@smu.edu.sg

ABSTRACT

We propose a novel approach to Vehicle Routing Problems (VRPs) that integrates the strengths of neural spatial embeddings with the semantic understanding of large language models (LLMs). While traditional neural methods excel in structured routing, they typically "feel" constraints through rigid mechanisms like feasibility masking, limiting their expressiveness. Conversely, LLMs demonstrate robust semantic capabilities but lack the spatial reasoning essential for precise routing optimization. To address these gaps, we introduce a hierarchical, multi-scale fusion architecture that integrates LLM-derived semantics with spatial routing features. Our contributions include: (1) a scale-aware decomposition aligning LLM features with spatial representations, (2) a bidirectional cross-modal attention module enabling interaction between linguistic and spatial domains, and (3) a progressive refinement pathway ensuring semantic and spatial fidelity. Through theoretical analysis, we prove the convergence properties of our fusion mechanism and its effectiveness in preserving critical information from both modalities. Evaluations across 16 VRP variants demonstrate competitive performance, highlighting the benefits of embedding semantic understanding into VRP optimization.

KEYWORDS

Large Language Model; Vehicle Routing Problem; Combinatorial Optimization; Learning to Optimize

ACM Reference Format:
Manuj Malik*, Jianan Zhou, Yan Jin, and Zhiguang Cao. 2026. LLMAide: Language-Assisted Neural Solver for Vehicle Routing Problems: Extended Abstract. In *Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), Paphos, Cyprus, May 25 – 29, 2026*, IFAAMAS, 3 pages. https://doi.org/10.65109/

* Corresponding author.

Proc. of the 25th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2026), C. Amato, L. Dennis, V. Mascardi, J. Thangarajah (eds.), May 25 – 29, 2026, Paphos, Cyprus. © 2026 International Foundation for Autonomous Agents and Multiagent Systems (www.ifaamas.org). https://doi.org/10.65109/

1 INTRODUCTION

The fair and efficient optimization of Vehicle Routing Problems (VRPs) is one of the most fundamental challenges in combinatorial optimization [1], with wide-ranging applications in logistics, transportation, and supply chain management [3, 10]. Recently, neural combinatorial optimization (NCO) has emerged as a promising alternative, where neural solvers [5, 7] learn heuristics in a data-driven manner. Multi-task VRP solvers [2, 4, 6, 13] are designed to simultaneously tackle multiple problem variants, aiming to learn more generalizable representations.

Despite these advances, current neural solvers exhibit a key limitation: they rely on feasibility masking to enforce constraints, i.e., by blocking invalid actions during decoding. Although this guarantees valid outputs, it offers no semantic insight into the constraints themselves. Conversely, LLMs offer semantic understanding and can interpret unstructured, language-based constraints, but lack inductive biases for spatial optimization. This creates a modal mismatch: neural solvers excel at spatial reasoning but are semantically blind, while LLMs are linguistically rich but spatially limited.

In this paper, we propose a hierarchical framework for multi-task vehicle routing that seamlessly integrates LLM-derived linguistic features with spatial routing by extending existing neural solver architectures [2]. Our approach employs scale-aware decomposition and bidirectional cross-modal attention to facilitate interaction between language and spatial domains. Moreover, a progressive fusion strategy with residual connections ensures that both semantic context and spatial precision are maintained. We conduct extensive experiments on 16 VRP variants, demonstrating the competitive performance of our method and confirming that embedding semantic understanding improves routing optimization.

2 PRELIMINARY

VRP Formulation. Let $G = (V, A)$ be a directed graph where $V = \{0, 1, \ldots, n\}$ represents nodes (with 0 as the depot and $1, \ldots, n$ as customers) and $A = \{(i, j) : i, j \in V, i \neq j\}$ represents arcs. Each customer i has a demand $q_i \geq 0$, and each arc (i, j) has a travel cost $c_{ij} \geq 0$. A fleet of m vehicles, each with capacity Q, is stationed at the depot. The objective is to minimize total travel cost $\sum_{k=1}^{m} \sum_{(i,j)\in A} c_{ij} x_{ijk}$, where $x_{ijk} \in \{0, 1\}$ indicates whether vehicle k traverses arc (i, j), subject to capacity, visit-once, and flow

	Solver	$n = 50$			$n = 100$				Solver	$n = 50$			$n = 100$		
		Obj.	Gap	Time	Obj.	Gap	Time			Obj.	Gap	Time	Obj.	Gap	Time
CVRP	HGS-PyVRP	10.372	*	10.4m	15.628	*	20.8m	VRPTW	HGS-PyVRP	16.031	*	10.4m	25.423	*	20.8m
	RF-TE	10.511	1.288%	3s	15.863	1.505%	9s		RF-TE	16.344	1.927%	3s	**26.268**	**3.178%**	9s
	LA-TE	**10.496**	**1.176%**	3s	**15.853**	**1.423%**	9s		LA-TE	**16.338**	**1.897%**	3s	26.301	3.431%	9s
OVRP	HGS-PyVRP	6.507	*	10.4m	9.725	*	20.8m	VRPB	HGS-PyVRP	9.687	*	10.4m	14.377	*	20.8m
	RF-TE	6.674	2.543%	2s	10.126	4.110%	8s		RF-TE	9.972	2.931%	1s	14.941	3.952%	8s
	LA-TE	**6.667**	**2.428%**	2s	**10.120**	**4.055%**	8s		LA-TE	**9.966**	**2.876%**	2s	**14.925**	**3.864%**	8s
VRPL	HGS-PyVRP	10.587	*	10.4m	15.766	*	20.8m	VRPBL	HGS-PyVRP	10.186	*	10.4m	14.779	*	20.8m
	RF-TE	10.748	1.499%	2s	16.059	1.819%	9s		RF-TE	10.548	3.500%	2s	15.497	4.837%	8s
	LA-TE	**10.747**	**1.492%**	2s	**16.054**	**1.835%**	9s		LA-TE	**10.541**	**3.480%**	2s	**15.494**	**4.827%**	8s

Table 1: Selected results on 6 of 16 VRP variants (1000 test instances each). * denotes best-known solutions. Full results for all 16 variants and all solver configurations are available in the full version.

conservation constraints. We consider 16 VRP variants derived by combining subsets of constraints $\{O, B, L, TW\}$ (Open routes, Backhauls, Duration Limits, Time Windows) with the base capacity constraint C.

Problem Formulation. We integrate LLM-derived semantic embeddings with route embeddings. Consider route embeddings $\mathbf{R} \in \mathbb{R}^{B\times N\times D}$ and language model embeddings $\mathbf{L} \in \mathbb{R}^{B\times N\times D}$, where B is the batch size, N is the number of nodes, and D is the embedding dimension. LLMs are prompted with task-specific templates that convert routing constraints into natural language. Our objective is to learn an optimal fusion function $f(\mathbf{R}, \mathbf{L})$ that combines these modalities while preserving their complementary information.

3 PROPOSED METHOD

Our LLMAide builds on a transformer-based encoder-decoder architecture similar to RouteFinder [2]. The encoder consists of standard transformer blocks; the decoder autoregressively constructs routes by selecting the next node at each step, constrained by dynamic feasibility masks.

Multi-Scale Decomposition. We decompose each modality into S distinct semantic scales through learned transformations:

$$\mathbf{R}^s = \phi_s(\mathbf{R}) + \mathbf{R}^{s-1}, \quad \mathbf{L}^s = \psi_s(\mathbf{L}) + \mathbf{L}^{s-1}, \tag{1}$$

where ϕ_s and ψ_s are scale-specific neural networks:

$$\phi_s(\mathbf{x}) = \text{ReLU}(\text{LayerNorm}(\mathbf{W}_s^R\mathbf{x} + \mathbf{b}_s^R)),$$
$$\psi_s(\mathbf{x}) = \text{ReLU}(\text{LayerNorm}(\mathbf{W}_s^L\mathbf{x} + \mathbf{b}_s^L)).$$

Cross-Modal Attention. For each scale s, we employ bidirectional cross-modal attention:

$$\mathbf{A}_{RL}^s = \text{MultiHead}(\mathbf{R}^s, \mathbf{L}^s, \mathbf{L}^s), \tag{2}$$
$$\mathbf{A}_{LR}^s = \text{MultiHead}(\mathbf{L}^s, \mathbf{R}^s, \mathbf{R}^s). \tag{3}$$

Scale Importance Learning. We use an adaptive scale weighting mechanism:

$$\boldsymbol{\alpha} = \text{softmax}\Big(\mathbf{W}_\alpha\Big[\,\Big\|_{s=1}^{S} \mu\Big(\tfrac{\mathbf{A}_{RL}^s + \mathbf{A}_{LR}^s}{2}\Big)\Big] + \mathbf{b}_\alpha\Big), \tag{4}$$

where $\mu(\cdot)$ is global average pooling and $\|$ denotes concatenation.

Progressive Hierarchical Fusion. We propose a progressive fusion scheme with M levels:

$$\mathbf{H}_0 = \mathbf{R}, \quad \mathbf{H}_m = G_m(\mathbf{H}_{m-1}, F_m(\mathbf{R}, \mathbf{L})) + \beta\mathbf{H}_{m-1}, \tag{5}$$

where F_m is the fusion function at level m:

$$F_m(\mathbf{R}, \mathbf{L}) = \sum_s \boldsymbol{\alpha}_s^m \cdot (\mathbf{A}_{RL}^{s,m} + \mathbf{A}_{LR}^{s,m})/2,$$

and G_m is a refinement network. The residual term $\beta\mathbf{H}_{m-1}$ ensures information preservation from both modalities across fusion levels.

4 EXPERIMENTS

We conducted experiments to evaluate LLMAide across 16 VRP variants. Synthetic instances were generated with $n = 50$ and $n = 100$ nodes. We compared our method with traditional solvers (HGS-PyVRP [11, 12], OR-Tools [8]) and neural baselines built on RouteFinder [2]: RF-POMO, RF-MoE, and RF-TE. We evaluate three configurations of LLMAide: LA-POMO, LA-MoE, and LA-TE. Our approach uses Qwen2-0.5B [9] for LLM embeddings.

We show selected experimental results in Table 1. Across most VRP variants, our methods show leading performance. LA-TE outperforms RF-TE in 13 out of 16 variants for $n = 50$ and 11 out of 16 for $n = 100$. Ablation studies confirm that: (i) three scales are optimal for decomposition, (ii) bidirectional cross-modal attention outperforms self-attention and unidirectional variants, (iii) two fusion levels are sufficient, and (iv) replacing LLM embeddings with random vectors significantly degrades performance, confirming the value of semantic context.

5 CONCLUSION

We proposed LLMAide, a hybrid framework that unifies the semantic reasoning capabilities of LLMs with the geometric precision of neural routing embeddings for solving VRPs. Our approach introduces multi-scale decomposition, bidirectional cross-modal attention, and progressive hierarchical fusion to integrate language and spatial modalities. Experimental results across 16 VRP variants demonstrate competitive performance, confirming that embedding semantic understanding improves routing optimization. The code is available at https://github.com/ra-MANUJ-an/llmaide-code.

ACKNOWLEDGMENTS

This research is supported by the National Research Foundation, Singapore under the AI Singapore Programme (AISG Award No: AISG3-RPGV-2025-017), and the Singapore Ministry of Education (MOE) Academic Research Fund (AcRF) Tier 1 grant.

REFERENCES

[1] Yoshua Bengio, Andrea Lodi, and Antoine Prouvost. 2021. Machine learning for combinatorial optimization: a methodological tour d'horizon. *European Journal of Operational Research* 290, 2 (2021), 405–421.

[2] Federico Berto, Chuanbo Hua, Nayeli Gast Zepeda, André Hottung, Niels A Wouda, Leon Lan, Junyoung Park, Kevin Tierney, and Jinkyoo Park. 2025. RouteFinder: Towards Foundation Models for Vehicle Routing Problems. *Transactions on Machine Learning Research* 2025 (2025).

[3] K. Braekers, K. Ramaekers, and I. Van Nieuwenhuyse. 2016. The vehicle routing problem: State of the art classification and review. *Computers & Industrial Engineering* 99 (2016), 300–313.

[4] Darko Drakulic, Sofia Michel, and Jean-Marc Andreoli. 2024. Goal: A generalist combinatorial optimization agent learner. *arXiv preprint arXiv:2406.15079* (2024).

[5] Wouter Kool, Herke van Hoof, and Max Welling. [n.d.]. Attention, Learn to Solve Routing Problems!. In *International Conference on Learning Representations*.

[6] Fei Liu, Xi Lin, Zhenkun Wang, Qingfu Zhang, Tong Xialiang, and Mingxuan Yuan. 2024. Multi-task learning for routing problem with cross-problem zero-shot generalization. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 1898–1908.

[7] M. Nazari, A. Oroojlooy, L. Snyder, and M. Takác. 2018. Reinforcement Learning for Solving the Vehicle Routing Problem. In *Advances in Neural Information Processing Systems*, Vol. 31. 9839–9849.

[8] Laurent Perron and Vincent Furnon. 2023. OR-Tools. Google.

[9] Qwen Team. 2024. Qwen Technical Report. *arXiv preprint arXiv:2309.16609* (2024).

[10] P. Toth and D. Vigo. 2014. *Vehicle Routing: Problems, Methods, and Applications* (2nd ed.). SIAM.

[11] Thibaut Vidal. 2022. Hybrid genetic search for the CVRP: Open-source implementation and SWAP* neighborhood. *Computers & Operations Research* 140 (2022), 105643.

[12] Niels A Wouda, Leon Lan, and Wouter Kool. 2024. PyVRP: A high-performance VRP solver package. *INFORMS Journal on Computing* (2024).

[13] Jianan Zhou, Zhiguang Cao, Yaoxin Wu, Wen Song, Yining Ma, Jie Zhang, and Chi Xu. 2024. MVMoE: Multi-Task Vehicle Routing Solver with Mixture-of-Experts. In *International Conference on Machine Learning*.