

TACO: Token-Aware Context Optimization for Structured OpenStreetMap Query Generation

ZHUOYUE WAN, Hong Kong Polytechnic University, China

WENTAO HU, Hong Kong Polytechnic University, China

HWANHEE KIM, Hong Kong Polytechnic University, China

CHEN JASON ZHANG, Hong Kong Polytechnic University, China

SHUAIMIN LI, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, China

YUANFENG SONG*, Independent Researcher, China

MING DENG, Hong Kong Polytechnic University, China

XIAO-YONG WEI, Hong Kong Polytechnic University, China

RAYMOND CHI-WING WONG, Hong Kong University of Science and Technology, China

Generating executable structured queries for OpenStreetMap (OSM) from natural language remains a challenging task due to the rigid syntax of OverpassQL and the steep learning curve it presents to end-users. While steering Large Language Models (LLMs) via *demonstration contexts* offers a promising solution, existing retrieval-based approaches typically neglect the critical constraint of token budgets. This oversight often leads to prohibitive inference costs and ignores the combinatorial nature of context selection, resulting in suboptimal performance. Furthermore, the lack of reliable benchmarks hinders rigorous evaluation in this domain.

To address these challenges, we first introduce **OsmNL**, a robust benchmark curated via a human-in-the-loop pipeline to resolve alignment noise in prior datasets. Building upon this foundation, we formulate the construction of optimal demonstration contexts under constraints as a *Budgeted Example Subset Mining (BESM)* problem. We analyze BESM as an NP-hard optimization task that admits a submodular objective, theoretically balancing semantic relevance with structural diversity. Leveraging this property, we propose **TACO (Token-Aware Context Optimization)**, a principled framework that instantiates a family of budget-aware algorithms to navigate the trade-off between approximation guarantees and computational efficiency. Extensive experiments on OsmNL show that TACO establishes a new state-of-the-art, achieving superior accuracy compared to strong baselines while reducing token consumption by over 50%, thereby offering a robust and cost-efficient solution for large-scale geospatial query interfaces.

CCS Concepts: • **Computing methodologies** → *Natural language processing*; • **Information systems** → **Structured Query Language**.

Additional Key Words and Phrases: Text-to-overpassQL, budgeted submodular maximization, natural language interfaces, token-aware context optimization

*Corresponding author.

Authors' Contact Information: Zhuoyue Wan, Hong Kong Polytechnic University, Hong Kong SAR, China; Wentao Hu, Hong Kong Polytechnic University, Hong Kong SAR, China; Hwanhee Kim, Hong Kong Polytechnic University, Hong Kong SAR, China; Chen Jason Zhang, Hong Kong Polytechnic University, Hong Kong SAR, China; Shuaimin Li, Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China; Yuanfeng Song, Independent Researcher, Shanghai, China; Ming Deng, Hong Kong Polytechnic University, Hong Kong SAR, China; Xiao-Yong Wei, Hong Kong Polytechnic University, Hong Kong SAR, China; Raymond Chi-Wing Wong, Hong Kong University of Science and Technology, Hong Kong SAR, China.



This work is licensed under a Creative Commons Attribution 4.0 International License.

© 2026 Copyright held by the owner/author(s).

ACM 2836-6573/2026/9-ART288

<https://doi.org/10.1145/3837126>

ACM Reference Format:

Zhuoyue Wan, Wentao Hu, Hwanhee Kim, Chen Jason Zhang, Shuaimin Li, Yuanfeng Song, Ming Deng, Xiao-Yong Wei, and Raymond Chi-Wing Wong. 2026. TACO: Token-Aware Context Optimization for Structured OpenStreetMap Query Generation. *Proc. ACM Manag. Data* 4, 4 (SIGMOD), Article 288 (September 2026), 27 pages. <https://doi.org/10.1145/3837126>

1 Introduction

Structured query languages constitute the primary interaction layer for exploring, managing, and reasoning over large-scale structured data repositories. Enabling users to express such queries through natural language has therefore become a long-standing and central research problem in data management and information systems. This problem setting has given rise to a broad family of *natural language to structured query generation* tasks, which aim to translate free-form user intents into executable programs over structured backends. Representative instances of this paradigm span multiple data modalities and query backends, including Text-to-SQL [7–9, 12, 22, 23, 34] for relational databases, Text-to-Visualization [24, 25, 28, 39, 45] for declarative visualization specifications, and Text-to-OverpassQL [36, 38, 44] for querying large-scale geospatial data. While these tasks differ in their target query languages and execution environments, they are united by the goal of translating natural language inputs into precise and executable structured queries under domain-specific constraints.

Within the broader landscape of structured data management, geospatial databases have become a fundamental class of system infrastructure, as their spatially indexed data models, hierarchical attribute schemas, and specialized query operators enable large-scale spatial analytics, efficient evaluation of complex spatial predicates, and a wide range of location-aware decision processes that support downstream workloads such as geographic knowledge extraction, urban mobility analysis, and spatio-temporal prediction [5, 13, 29–31]. Among existing geospatial data resources, OpenStreetMap (OSM) has emerged as one of the most widely adopted platforms for large-scale spatial data analysis due to its rich tagging schema and continuously evolving, crowdsourced coverage; however, querying OSM typically relies on the Overpass Query Language (OverpassQL), a domain-specific structured language that supports expressive spatial filters, recursive traversals, and compositional constraints, but also introduces a steep learning curve for users unfamiliar with OSM tagging conventions or the syntactic and semantic rules of OverpassQL. To lower this barrier, Text-to-OverpassQL methods translate natural language questions into executable OverpassQL programs [36]. Existing approaches mainly follow two directions: retrieval-based prompting, which incorporates semantically related query pairs as in-context examples, and model-centric pre-training, which injects OSM tag knowledge into language models to improve their understanding of tag semantics and structured geospatial queries, as in OsmT [38].

Motivation. Despite recent progress, existing approaches for structured OSM query generation continue to exhibit several fundamental limitations. First, the widely used OverpassNL benchmark [36] contains substantial data noise, as many natural language questions are only weakly aligned with their corresponding OverpassQLs, which hampers the construction of reliable example pools for context-based query generation and weakens the robustness of empirical evaluation. Second, current methods largely overlook the practical cost of token usage. In real-world query interfaces such as Overpass Turbo¹, longer prompts directly degrade serving efficiency by increasing inference latency, memory footprint, and monetary cost. Importantly, this input-length dependence cannot be trivially resolved by system-level optimizations like KV caching, as the initial prefill stage still imposes a computational bottleneck proportional to the prompt length. Nevertheless,

¹<https://overpass-turbo.eu/>

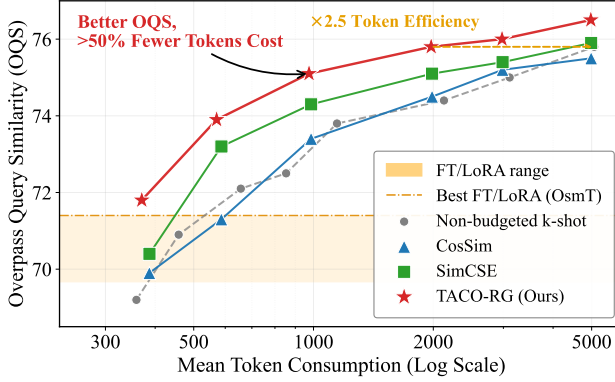


Fig. 1. Pareto frontier on the OsmNL test dataset, illustrating the relationship between mean token consumption (log scale) and Overpass Query Similarity (OQS) for structured OpenStreetMap query generation, and highlighting the practical trade-off between token cost and generation quality.

existing approaches typically retrieve a fixed number of examples without explicit consideration of token budgets, leaving users without a principled mechanism to balance contextual informativeness against cost constraints. As shown in Figure 1, this limitation leads to a pronounced trade-off between token cost and OverpassQL quality across different context selection strategies. Finally, prior techniques lack a solid theoretical foundation, as heuristic retrieval strategies and similarity-based selection methods do not provide stability guarantees or provable performance bounds under constrained context budgets. Consequently, their performance varies substantially across input distributions, dataset conditions, and budget settings, limiting their reliability in practical, large-scale geospatial query systems.

Proposed Solution. Within this concern, we construct **OsmNL** through a human-in-the-loop data curation process, where LLMs are employed to assist in candidate refinement, semantic consistency checking, and the identification of potentially misaligned example pairs in OverpassNL. This process is further complemented by the addition of newly manually annotated high-quality examples. More specifically, we propose a utility formulation that approximates OverpassQL generation quality by jointly modeling the relevance and diversity of the selected subset, and formulate context construction as a principled budgeted subset mining optimization problem, termed *Budgeted Example Subset Mining (BESM)*. We show that BESM is NP-hard and admits a submodular objective. To efficiently address this, we design **TACO (Token-Aware Context Optimization)**, a framework that leverages the submodularity of BESM to instantiate a family of algorithms with different optimality–efficiency trade-offs. Specifically, TACO-E provides a theoretical approximation guarantee through an efficiency-guided partial enumeration strategy, whereas TACO-G adopts a pure gain-guided greedy selection based solely on marginal utility. To further improve scalability, TACO-RG restricts gain evaluation to a relevance-filtered candidate subset, substantially reducing computational overhead while retaining strong empirical performance.

Contributions. We highlight our main contributions:

- To the best of our knowledge, we are the first to formulate token-constrained context construction as a principled optimization problem, termed *Budgeted Example Subset Mining (BESM)*. We provide a theoretical analysis proving that BESM is NP-hard while admitting a submodular objective.
- We propose **TACO**, a token-aware context optimization framework that exploits the submodularity of BESM to instantiate a family of budget-aware greedy algorithms, enabling flexible trade-offs between theoretical approximation guarantees and computational efficiency.

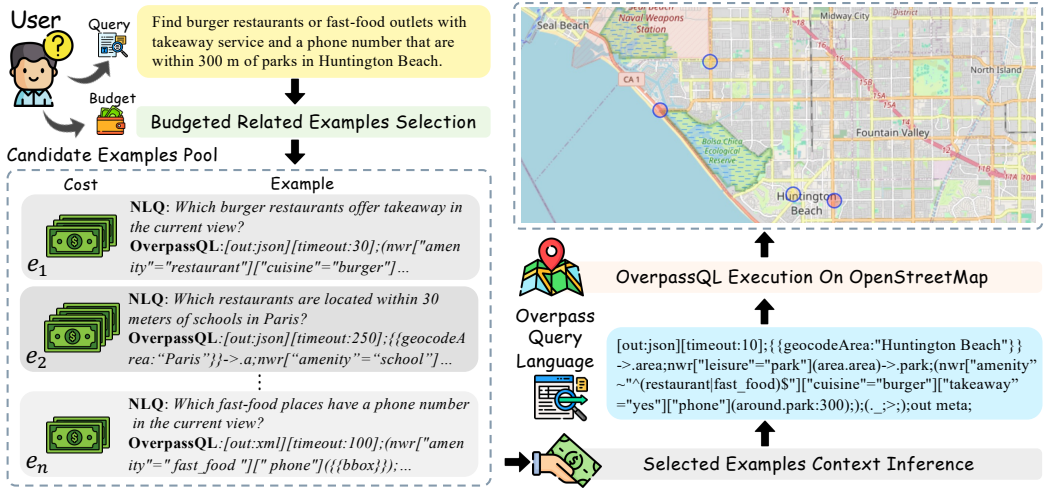


Fig. 2. Overall workflow of token-constrained context construction for Text-to-OverpassQL. A user query and a token budget drive budgeted example selection and context composition, which together guide OverpassQL generation.

- We construct **OsmNL**, a high-quality benchmark for Text-to-OverpassQL, curated via a human-in-the-loop pipeline that combines LLM-assisted refinement with systematic human verification and annotation.
- We conduct extensive experiments on OsmNL across diverse token budgets and evaluation settings, showing that compared with SOTA methods, **TACO** attains better OQS with over 50% fewer prompt tokens, demonstrating its effectiveness, efficiency, and robustness.

2 Preliminaries

2.1 Text-to-OverpassQL Task

The Text-to-OverpassQL task aims to translate a natural language query into an executable OverpassQL for retrieving geographic elements from the OpenStreetMap database. While existing approaches [36, 38] typically employ pre-training or standard in-context learning, they often face a critical trade-off between prompt length and semantic relevance. Naive retrieval strategies tend to introduce redundant or low-information demonstrations, resulting in inflated token costs without proportional gains in generation quality. This limitation becomes particularly evident in practical settings that require concise and information-dense prompts.

To address this, we formulate the task as a *Budgeted Example Subset Mining* problem. Figure 2 illustrates the overall workflow: given a target query and a token budget, the model retrieves relevant examples and composes them into a budget-feasible context that guides OverpassQL generation. Let $V = \{e_i\}_{i=1}^{|V|}$ denote a candidate pool, where each example e_i consists of a natural language query nl_i and its corresponding OverpassQL ovq_i . Given a target query q_i and a strictly defined token budget B , our goal is to select an optimal subset $S^* \subseteq V$ that maximizes the downstream generation quality $\mathcal{F}(S)$ while ensuring the total cost $C(S)$ remains within limits. This optimization problem is formally defined as:

$$S^* = \underset{S \subseteq V}{\operatorname{argmax}} \mathcal{F}(S) \quad \text{s.t.} \quad C(S) \leq B. \quad (1)$$

Since directly optimizing the black-box function $\mathcal{F}(S)$ is intractable during subset mining, we propose a specific utility formulation to approximate this objective, which we detail in Section 2.3.

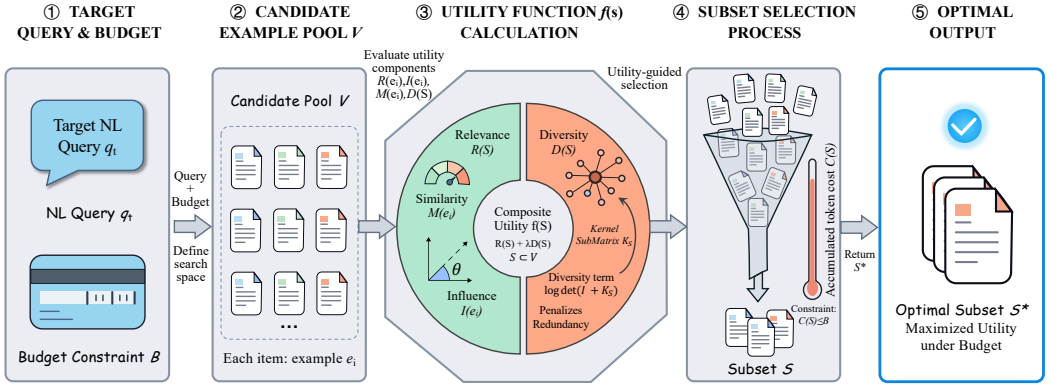


Fig. 3. The overall pipeline of the TACO framework for Budgeted Example Subset Mining.

2.2 Function Definition

We further formalize the token cost function C , the similarity and influence components ($\mathcal{M}$ and $\mathcal{I}$), the unified relevance score $\mathcal{R}$, and the diversity term $\mathcal{D}$.

Token Cost Function C . For a context containing $k = |S|$ examples, the total input-token cost is

$$C_{total}^k = \underbrace{C_{system} + C_{q_t} + C_{prefix}}_{\text{Fixed Context}} + \underbrace{\sum_{e_i \in S} (C_{base} + C_{nl}^{(i)} + C_{ovq}^{(i)})}_{\text{Per-Example Cost}}, \quad (2)$$

where $C_{nl}^{(i)}$ and $C_{ovq}^{(i)}$ are the token costs of nl_i and ovq_i , respectively.

Similarity Score Function $\mathcal{M}$. The similarity score captures coarse-grained semantic proximity between the example query nl_i and the target query q_t . We compute a normalized non-negative similarity using their distance in the embedding space:

$$\mathcal{M}(e_i) = 1 - \underbrace{\frac{d(\vec{nl}_i, \vec{q}_t) - d_{\min}}{d_{\max} - d_{\min}}}_{\text{Min-Max Norm}}, \quad (3)$$

where $\vec{nl}_i$ and $\vec{q}_t$ denote the vector embeddings of the example query nl_i and target query q_t , respectively, $d(\cdot, \cdot)$ denotes the Euclidean distance between a pair of embedding vectors, and

$$d_{\min} = \min_{e_j \in V} d(\vec{nl}_j, \vec{q}_t), \quad d_{\max} = \max_{e_j \in V} d(\vec{nl}_j, \vec{q}_t).$$

Notwithstanding the reliability of the similarity score $\mathcal{M}$ in measuring coarse-grained semantic proximity within a real embedding space, absolute distance-based metrics inevitably lose discriminative power in high-similarity saturation zones. [26]. Within the context of Text-to-OverpassQL generation, this introduces a fundamental limitation: subtle lexical variations that necessitate entirely distinct geospatial operators are frequently overshadowed by aggregate surface similarity. For example, “Find restaurants near schools” and “Find restaurants in schools” exhibit high surface similarity, yet require fundamentally different spatial constraints. Therefore, relying solely on distance-based similarity is insufficient for fully capturing fine-grained structural alignment.

Influence Score Function $\mathcal{I}$. To complement $\mathcal{M}$ and provide a fine-grained measure of structural and operational compatibility, we formulate an influence score $\mathcal{I}$ by projecting queries into a complex embedding space. This transformation maps standard semantic vectors to complex representations characterized by both magnitude and phase. While the real component absorbs the coarse surface similarity, the imaginary component models subtle relational variations. Consequently, by quantifying the phase discrepancy in the complex plane, we derive a robust and non-saturating gradient signal. This angle difference effectively captures the underlying directional and structural influence an example query nl_i exerts on the target query q_t during OverpassQL inference

Specifically, let $\tilde{nl}_i = a_i + b_i i$ and $\tilde{q}_t = c_t + d_t i$ denote the complex embeddings of nl_i and q_t , respectively, where $a_i, b_i, c_t, d_t \in \mathbb{R}^d$, and i denotes the imaginary unit. Following the complex division rule, the angular difference $\Delta\theta_{it}$ between $\tilde{nl}_i$ and $\tilde{q}_t$ is defined as

$$\Delta\theta_{it} = \arg \left(\frac{(a_i c_t + b_i d_t) + (b_i c_t - a_i d_t) i}{\sqrt{(a_i^2 + b_i^2)(c_t^2 + d_t^2)}} \right), \quad (4)$$

where $\arg(\cdot)$ denotes the principal argument of a complex number.

A smaller angular difference indicates stronger directional consistency in the complex embedding space. To convert this angular discrepancy into a scalar influence score while preserving its monotonicity, we first map $\Delta\theta_{it}$ to a non-negative angular deviation measure:

$$\delta_{it} = \sqrt{2 - 2 \cos(\Delta\theta_{it})}. \quad (5)$$

We then define a normalized non-negative influence score function:

$$\mathcal{I}(e_i) = 1 - \underbrace{\frac{\delta_{it} - \delta_{\min}}{\delta_{\max} - \delta_{\min}}}_{\text{Min-Max Norm}}, \quad (6)$$

where $\delta_{\min}$ and $\delta_{\max}$ denote the minimum and maximum angular deviation values computed over all candidate examples with respect to the target query.

Relevance Score Function $\mathcal{R}$. To select exemplars that are both semantically relevant and structurally informative, we unify them into a comprehensive relevance score via a convex combination:

$$\mathcal{R}(e_i) = \alpha \cdot \mathcal{M}(e_i) + (1 - \alpha) \cdot \mathcal{I}(e_i), \quad (7)$$

where $\alpha \in [0, 1]$ controls the trade-off between similarity and influence. For a selected subset S , the relevance function is defined as

$$\mathcal{R}(S) = \phi \left(\sum_{e_i \in S} \mathcal{R}(e_i) \right), \quad (8)$$

where $\phi(\cdot)$ is non-decreasing and concave, capturing diminishing returns with respect to the aggregate relevance mass.

Diversity Score Function $\mathcal{D}$. We first define a pairwise similarity kernel between candidate examples. For any two examples $e_i, e_j \in V$, the kernel value $\mathbf{k}_{ij}$ is defined as

$$\mathbf{k}_{ij} = 1 - \frac{d(\vec{nl}_i, \vec{nl}_j) - d_{\min}^{\text{ex}}}{d_{\max}^{\text{ex}} - d_{\min}^{\text{ex}}}, \quad (9)$$

where $d(\cdot, \cdot)$ denotes the Euclidean distance between embedding vectors, and $d_{\min}^{\text{ex}}$ and $d_{\max}^{\text{ex}}$ denote the minimum and maximum distances computed over all example pairs.

Given a selected subset $S \subseteq V$, we construct the kernel submatrix $\mathbf{K}_S \in \mathbb{R}^{|S| \times |S|}$ by restricting $\{\mathbf{k}_{ij}\}$ to indices in S . The diversity score is then defined using a determinantal point process (DPP)–based objective:

$$\mathcal{D}(S) = \log \det \left(\mathbf{I} + [\mathbf{k}_{ij}]_{e_i, e_j \in S} \right) = \log \det(\mathbf{I} + \mathbf{K}_S), \quad (10)$$

where $\mathbf{I}$ denotes the identity matrix. The log-determinant objective penalizes correlated features, thereby suppressing structurally redundant exemplars. Under tight token budgets, similarity-based retrieval in Text-to-OverpassQL is particularly prone to redundancy, as semantically similar queries often induce near-duplicate OverpassQL blocks (e.g., repeated geometry filters or recursion patterns). Under a DPP kernel, a candidate exemplar yields a smaller marginal gain in $\mathcal{D}(S)$ when its geometric or syntactic features overlap with the selected subset S , thereby favoring complementary exemplars that broaden coverage of OverpassQL grammar-defined structures.

2.3 BESM Problem Formulation

Recall that our primary objective in Eq. 1 is to maximize the downstream generation quality $\mathcal{F}(S)$, an intractable black-box functional induced by the generation behavior of the underlying LLM, whose reliable evaluation generally entails actual LLM execution together with downstream output assessment. This makes $\mathcal{F}(S)$ inaccessible during context construction and therefore rules out direct gradient-based or oracle-based optimization. To this end, the key challenge is not merely the optimization of a prescribed utility, but the construction of a tractable surrogate that remains faithful to the black-box objective under token constraints. We therefore introduce a surrogate objective $f(S)$ that is efficiently computable, submodular, and grounded in the semantic and structural signals that govern Text-to-OverpassQL generation quality. This formulation yields a tractable approximation to $\mathcal{F}(S)$ and admits a natural decomposition into relevance and diversity components.

Formally, let $S \subseteq V$ denote a candidate subset. We seek to identify a subset S such that the total token cost satisfies $C(S) \leq B$, while maximizing a composite utility function $f(S)$:

$$S^* = \operatorname{argmax}_{S \subseteq V} \underbrace{\left(\mathcal{R}(S) + \lambda \mathcal{D}(S) \right)}_{\text{utility } f(S)} \quad \text{s.t.} \quad \underbrace{C(S) \leq B}_{\text{budget constraint}}, \quad (11)$$

where $\lambda \geq 0$ controls the relative contribution of the diversity term. To make the objective explicit, the utility of a subset can be decomposed as

$$\begin{aligned} f(S) &= \mathcal{R}(S) + \lambda \mathcal{D}(S) \\ &= \phi \left(\sum_{e_i \in S} (\alpha \mathcal{M}(e_i) + (1 - \alpha) \mathcal{I}(e_i)) \right) + \lambda \log \det(\mathbf{I} + \mathbf{K}_S), \end{aligned} \quad (12)$$

where $\alpha \in [0, 1]$ controls the relative weighting between the similarity and influence components in the relevance score.

3 TACO Overview

Building on the preliminary scoring functions (Section 2.2) and problem formulation (Section 2.3), TACO models token-constrained subset mining along two complementary dimensions: *relevance modeling* and *diversity modeling*.

Relevance Modeling. We first instantiate relevance using the most direct cue, language-level similarity, by measuring the Euclidean proximity between the target query q_t and each candidate example query nl_i in the embedding space, yielding the similarity score $\mathcal{M}$. While $\mathcal{M}$ captures coarse-grained semantic relatedness and supports efficient relevance filtering, it does not fully

Algorithm 1 Overall TACO Framework

Require: Target query q_t ; candidate example pool V ; budget B ; cost function C ; inference model LLM ;

Ensure: Selected subset S and generated OverpassQL ovq_t

```

1:  $S \leftarrow \emptyset$ 
2:  $B_r \leftarrow B - (C_{\text{system}} + C_{q_t} + C_{\text{prefix}})$  ▷ remaining budget
3: if  $B_r \leq 0$  then
4:   Construct zero-shot context prompt  $T(q_t, S)$ 
5:    $ovq_t \leftarrow LLM(T(q_t, S))$ 
6:   return  $S, ovq_t$ 
7: end if
8: while  $B_r > 0$  do
9:    $\Omega \leftarrow \{e \in V \setminus S : \Delta_C(e) \leq B_r\}$  ▷ candidates feasible under the remaining budget
10:  if  $\Omega = \emptyset$  then
11:    break
12:  end if
13:  Select
      
$$e^* = \underset{e \in \Omega}{\text{SUBSETMINING}}(e \mid S, q_t)$$

      ▷ instantiated by TACO-E / TACO-G / TACO-RG
14:   $S \leftarrow S \cup \{e^*\}$ 
15:   $B_r \leftarrow B_r - \Delta_C(e^*)$ 
16: end while
17: Construct context prompt  $T(q_t, S)$ 
18:  $ovq_t \leftarrow LLM(T(q_t, S))$ 
19: return  $S, ovq_t$ 

```

characterize structural compatibility for Text-to-OverpassQL task. In this task, queries that are close in natural language space may still induce different geospatial operators, spatial constraints, and program compositions. As a result, absolute distance-based similarity becomes less discriminative in high-similarity regions, where surface-level proximity can obscure fine-grained structural differences. This motivates an additional influence signal that captures not only whether an example is semantically relevant, but also whether it exerts a compatible directional effect on the generation of the target OverpassQL structure.

To address this limitation, we introduce a directional influence score based on an angle-aware objective inspired by [26]. Rather than using the angle signal only for semantic discrimination, we adapt it to capture compatibility with the target OverpassQL structure and execution behavior. Specifically, for each query–example pair (q, e) , we construct a continuous supervision label $y(q, e) \in [0, 1]$ by aggregating three complementary signals: (i) tag-level correspondence between OverpassQL programs, (ii) query skeleton alignment, and (iii) execution consistency measured from one-shot generation results. This supervision is then encoded in a complex-valued representation, where the real component preserves surface-level semantic variation and the imaginary component captures latent relational factors associated with structural and execution consistency. Further details are provided in Section 6.

We then combine semantic similarity and directional influence into a unified relevance score through a mixing parameter α , and aggregate relevance at the set level via $\mathcal{R}(S)$. A linear summation of relevance would over-prefer many individually relevant yet highly overlapping examples, which

leads to diminishing practical returns. To model this saturation effect in the aggregate relevance mass, we introduce a non-decreasing concave transformation $\phi(\cdot)$.

Diversity Modeling. Under strict token budgets in the BESM setting, selecting redundant or highly similar examples fails to fully utilize the context space, thereby degrading downstream generalization performance. As established in Section 2.2, this vulnerability is uniquely amplified in Text-to-OverpassQL task. To mitigate this, we introduce a diversity term modeled via a DPP kernel. We instantiate this as a log-determinant objective $\mathcal{D}(S)$, which favors subsets whose elements are mutually dissimilar across both geometric and syntactic feature spaces. Under this formulation, as the structural features of a candidate exemplar overlap with the already-selected subset S , its marginal contribution to $\mathcal{D}(S)$ diminishes, effectively suppressing syntactic redundancy.

This diversity objective complements the concave transformation ϕ employed in relevance aggregation. While ϕ regulates saturation in relevance magnitudes, $\mathcal{D}(S)$ explicitly captures higher-order interactions among selected demonstrations. Together, they enforce TACO to construct a compact, budget-feasible context that maximizes relevance to the target query q_t while covering a broad spectrum of OverpassQL structural paradigms, preventing repetitive syntax blocks from dominating the prompt.

Framework Overview. Algorithm 1 formalizes the end-to-end TACO pipeline for token-constrained context construction in Text-to-OverpassQL. As illustrated in Figure 3, the framework takes a target natural language query q_t together with a strict token budget B , and searches over a candidate pool V to mine a budget-feasible context subset S by maximizing a composite utility function $f(S) = \mathcal{R}(S) + \lambda\mathcal{D}(S)$. The relevance component $\mathcal{R}(S)$ is derived from similarity $\mathcal{M}$ and directional influence $\mathcal{I}$, while diversity is captured by a redundancy-aware term $\mathcal{D}(S)$ instantiated via a log-determinant objective. The selection procedure is implemented using a family of token-aware utility-guided greedy algorithms, namely **TACO-E**, **TACO-G**, and **TACO-RG**. These variants offer configurable trade-offs between optimality and computational efficiency while strictly respecting the remaining token budget. A detailed description of the algorithms is provided in Section 5.

4 BESM Problem Analysis

This section presents the theoretical analysis of the *Budgeted Example Subset Mining* problem. Specifically, we prove its NP-hardness via reductions from well-known NP-hard problems and establish that the objective utility function is submodular.

4.1 NP-Hardness of BESM

We analyze the computational complexity of the BESM problem. Specifically, we establish its **NP-hardness** by demonstrating that it subsumes two classical NP-hard problems: the 0/1 KNAPSACK PROBLEM and the MAXIMUM ENTROPY SAMPLING PROBLEM.

THEOREM 4.1. *The BESM problem is NP-hard.*

PROOF. We prove the NP-hardness by restriction. That is, we show that specific instances of BESM are equivalent to known NP-hard problems.

Membership in NP. We first formulate the decision version of the BESM problem. Given a candidate set V , a budget B , and a utility threshold τ , the decision problem asks whether there exists a subset $S \subseteq V$ satisfying both the budget constraint $C(S) \leq B$ and the utility constraint $f(S) \geq \tau$. For any candidate subset S , checking the budget constraint $C(S) \leq B$ takes $O(|S|)$ time. Computing the objective $f(S) = \mathcal{R}(S) + \lambda\mathcal{D}(S)$ is dominated by the log-determinant computation in $\mathcal{D}(S)$, which requires $O(|S|^3)$ time via Cholesky decomposition. Since the validity of a proposed solution can be verified in polynomial time, BESM is in NP.

Reduction from 0/1 Knapsack. When the diversity term is disregarded (i.e., $\lambda = 0$), the objective function simplifies to $f(S) = \phi(\sum_{e_i \in S} \mathcal{R}(e_i))$. Given that $\phi(\cdot)$ is monotonically non-decreasing, maximizing $f(S)$ is strictly equivalent to maximizing the total relevance $\sum_{e_i \in S} \mathcal{R}(e_i)$ subject to the budget constraint. We construct a direct reduction from the 0/1 KNAPSACK PROBLEM. Consider a Knapsack instance where items have values v_i , weights w_i , and total capacity W . We map this to BESM by defining $\mathcal{R}(e_i) = v_i$, $C(e_i) = w_i$, and $B = W$. The optimal subset S^* for this BESM instance corresponds exactly to the optimal solution of the Knapsack problem. Since the 0/1 KNAPSACK PROBLEM is NP-hard, it follows that BESM is also NP-hard.

Reduction from MESP. Conversely, consider the scenario where relevance is disregarded (i.e., $\mathcal{R}(e_i) = 0$) and costs are uniform (i.e., $C(e_i) = 1$) with a budget $B = k$. The optimization objective simplifies to maximizing the diversity term subject to a cardinality constraint $|S| \leq k$:

$$\max_{S \subseteq V, |S| \leq k} \lambda \log \det(\mathbf{I} + \mathbf{K}_S).$$

Given that the log-determinant is monotonic with respect to set inclusion for positive definite kernels, the inequality constraint binds as an equality $|S| = k$. This formulation is strictly equivalent to the MAXIMUM ENTROPY SAMPLING PROBLEM (MESP), which entails selecting a size- k principal submatrix to maximize the determinant. As MESP has been established as NP-hard [18], it follows that the diversity component alone renders BESM NP-hard.

Conclusion. Since the BESM framework subsumes these NP-hard problems as special cases (through specific parameter restrictions), the general optimization problem is inherently NP-hard. $\square$

4.2 Submodularity of the Objective

We analyze the submodularity of the BESM objective. We formally show that the composite utility function $f(S)$ is *monotone submodular*.

THEOREM 4.2. *The utility function $f(S) = \mathcal{R}(S) + \lambda \mathcal{D}(S)$ is monotone submodular, provided that $\lambda \geq 0$.*

PROOF. Since the class of submodular functions is closed under non-negative linear combinations, it suffices to prove that both the relevance term $\mathcal{R}(S)$ and the diversity term $\mathcal{D}(S)$ are monotone submodular individually.

Relevance Term. Let $L(S) = \sum_{e_i \in S} \mathcal{R}(e_i)$. Since the individual relevance scores are non-negative, $L(S)$ is a monotone *modular* function. The relevance objective is defined as $\mathcal{R}(S) = \phi(L(S))$, where $\phi(\cdot)$ is a non-decreasing concave function. It is a standard result that the composition of a non-decreasing concave function with a non-negative modular function yields a submodular function. Specifically, for any $A \subseteq B \subseteq V$ and an element $x \in V \setminus B$, let $\Delta = \mathcal{R}(x)$. Due to the concavity of ϕ , we have:

$$\phi(L(A) + \Delta) - \phi(L(A)) \geq \phi(L(B) + \Delta) - \phi(L(B)).$$

This inequality confirms the diminishing returns property. Furthermore, since both L and ϕ are non-decreasing, their composition $\mathcal{R}(S)$ is monotone.

Diversity Term. We formulate the diversity term as $\mathcal{D}(S) = \log \det(\mathbf{P}_S)$, where $\mathbf{P}_S$ denotes the principal submatrix of the regularized kernel matrix $\mathbf{P} = \mathbf{I}_V + \mathbf{K}_V$ corresponding to the subset S . Since the kernel submatrix $\mathbf{K}_S$ is positive semi-definite (PSD), the regularized submatrix $\mathbf{P}_S$ is strictly positive definite (PD). The submodularity of the log-determinant function is a fundamental property in combinatorial optimization; thus, $\mathcal{D}(S)$ is submodular.

Algorithm 2 Classical Efficiency-Guided Greedy Algorithm**Require:** Candidate set V ; budget B ; cost function C ; utility function f **Ensure:** Selected subset S

```

1:  $S \leftarrow \emptyset$ 
2:  $B_r \leftarrow B - (C_{system} + C_{q_t} + C_{prefix})$  ▷ remaining budget
3: while true do
4:    $\Omega \leftarrow \{e \in V \setminus S : \Delta_C(e) \leq B_r\}$ 
5:   if  $\Omega = \emptyset$  then
6:     break
7:   end if
8:   Select  $e^* = \operatorname{argmax}_{e \in \Omega} \frac{\Delta_f(e | S)}{\Delta_C(e)}$ 
9:    $S \leftarrow S \cup \{e^*\}$ 
10:   $B_r \leftarrow B_r - \Delta_C(e^*)$ 
11: end while
12: return  $S$ 

```

To establish monotonicity, we consider the marginal gain of adding an element $x \notin S$:

$$\mathcal{D}(S \cup \{x\}) - \mathcal{D}(S) = \log \left(\frac{\det(\mathbf{P}_{S \cup \{x\}})}{\det(\mathbf{P}_S)} \right) = \log(\sigma_{x|S}^2),$$

where $\sigma_{x|S}^2$ denotes the conditional variance of x given S , derived via the Schur complement. We can interpret $\mathbf{P}$ as the covariance matrix of a noisy Gaussian process $Y = Z + \epsilon$, where $Z \sim \mathcal{N}(0, \mathbf{K})$ is the underlying signal and $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ is independent Gaussian noise. The conditional variance can then be written as $\sigma_{x|S}^2 = \text{Var}(Z_{x|S}) + 1 \geq 1$ due to the irreducible noise component. Consequently, the marginal gain is always non-negative ($\log(\sigma_{x|S}^2) \geq 0$), implying that $\mathcal{D}(S)$ is monotone increasing.

Conclusion. We have established that both $\mathcal{R}(S)$ and $\mathcal{D}(S)$ are monotone submodular. As submodularity is closed under non-negative linear combinations, the composite objective $f(S)$ is guaranteed to be monotone submodular for any $\lambda \geq 0$. $\square$

5 TACO Greedy Algorithms

Building on the NP-hardness of BESM and the monotone submodularity of $f(\cdot)$ established in Section 4, we propose a family of greedy algorithms to solve the problem, including the efficiency-guided greedy variant **TACO-E**, the pure gain-guided greedy variant **TACO-G**, and the relevance-guided gain greedy variant **TACO-RG**. Before introducing these variants, we first revisit a classical efficiency-guided greedy rule and examine its suitability in the token-constrained setting, as shown in Algorithm 2. For a current subset S and a candidate example e , the *marginal gain* of adding e is defined as

$$\Delta_f(e | S) = f(S \cup \{e\}) - f(S),$$

while the corresponding *marginal cost* is

$$\Delta_C(e) = C(S \cup \{e\}) - C(S) = C_{base} + C_{nl}^{(e)} + C_{ovq}^{(e)}.$$

The efficiency of example e at state S is then measured by the marginal utility density

$$\eta(e | S) = \frac{\Delta_f(e | S)}{\Delta_C(e)}.$$

Algorithm 3 TACO-E: Efficiency-Greedy Algorithm

Require: Candidate set V ; budget B ; target query q_t ; relevance score $\mathcal{R}$; sample size m ; enumeration depth d ; cost function C ; utility function f

Ensure: Selected subset S

```

1:  $\mathcal{W} \leftarrow \text{TOPFILTER}(V, \mathcal{R}(\cdot \mid q_t), m)$   $\triangleright |\mathcal{W}| = m$ 
2:  $B_r \leftarrow B - (C_{\text{system}} + C_{q_t} + C_{\text{prefix}})$   $\triangleright$  remaining budget
3: if  $B_r \leq 0$  then
4:   return  $\emptyset$ 
5: end if
6:  $\triangleright$  Best feasible solution with  $|S| < d$  over  $\mathcal{W}$ 
7:  $S_1 \leftarrow \underset{S \subseteq \mathcal{W} : C(S) \leq B_r, |S| < d}{\text{argmax}} f(S)$ 
8:  $\triangleright$  Best greedy-extended solution seeded by a size- $d$  subset in  $\mathcal{W}$ 
9:  $S_2 \leftarrow \emptyset$ 
10: for all  $U \subseteq \mathcal{W}$  such that  $|U| = d$  and  $C(U) \leq B_r$  do
11:    $\mathcal{W}' \leftarrow \mathcal{W} \setminus U$ 
12:    $S \leftarrow \text{ALGORITHM 2}(\mathcal{W}', B, C, f)$  with initialization  $S \leftarrow U$  and  $B_{\text{remains}} \leftarrow B_r - C(U)$ 
13:   if  $f(S) > f(S_2)$  then
14:      $S_2 \leftarrow S$ 
15:   end if
16: end for
17: if  $f(S_1) \geq f(S_2)$  then
18:   return  $S_1$ 
19: else
20:   return  $S_2$ 
21: end if

```

At each iteration, the efficiency-guided greedy algorithm selects a feasible example that maximizes $\eta(e \mid S)$. This density criterion is well motivated under budget constraints, as selecting examples solely based on marginal gain can be suboptimal when high-utility demonstrations also incur large token costs, potentially crowding out combinations with higher overall utility within the budget.

However, while effective in classical knapsack settings, pure density-greedy strategies can become myopic under strict token budgets. They tend to favor examples with small token costs even when such examples exhibit weak relevance to the target query, which leads to the inclusion of uninformative demonstrations. This failure mode is particularly detrimental in LLM prompting. Indiscriminately injecting irrelevant context into the prompt can significantly degrade model performance, even when individual examples are syntactically correct, as LLMs are known to be easily distracted by irrelevant information [35].

5.1 TACO-E: Efficiency-Guided Greedy

Motivated by the failure of pure density-greedy strategies under token constraints, we design TACO-E (Algorithm 3) with the goal of improving solution quality while retaining provable guarantees. The key idea is to prevent early commitment to low-impact demonstrations and to better capture high-utility combinations that may be missed by myopic greedy selection.

To this end, TACO-E first restricts the search space by ranking candidates according to their relevance to the target query and retaining only the top-ranked examples as a pre-filtered candidate pool $\mathcal{W}$ of size m . Within this relevance-restricted pool, TACO-E explores two complementary

Algorithm 4 TACO-G: Gain-Guided Greedy Algorithm**Require:** Candidate set V ; budget B ; target query q_t ; cost function C ; utility function f **Ensure:** Selected subset S

```

1:  $S \leftarrow \emptyset$ 
2:  $B_r \leftarrow B - (C_{system} + C_{q_t} + C_{prefix})$  ▷ remaining budget
3: while true do
4:    $\Omega \leftarrow \{e \in V \setminus S : \Delta_C(e) \leq B_r\}$ 
5:   if  $\Omega = \emptyset$  then
6:     break
7:   end if
8:   Select  $e^* = \operatorname{argmax}_{e \in \Omega} \Delta_f(e \mid S)$ 
9:    $S \leftarrow S \cup \{e^*\}$ 
10:   $B_r \leftarrow B_r - \Delta_C(e^*)$ 
11: end while
12: return  $S$ 

```

solution types. The first considers all feasible subsets with cardinality smaller than a predefined enumeration depth d , capturing cases where a small number of demonstrations already yields high utility. The second performs seeded greedy completion by enumerating all feasible size- d seeds $U \subseteq \mathcal{W}$ and greedily extending each seed on the residual pool under the remaining budget using the same efficiency criterion. The final output is selected as the better solution between the best small-set candidate and the best greedy-extended seeded candidate.

5.2 TACO-G: Gain-Guided Greedy

Despite its theoretical appeal, the efficiency-guided greedy rule introduces an additional normalization by marginal cost, which can bias selection toward short but weakly informative examples. As a complementary perspective, we introduce a pure gain-guided greedy strategy that omits cost normalization and selects the example with the largest marginal utility at each iteration. We refer to this variant as **TACO-G**, which serves as a simple yet informative heuristic baseline. As shown in Algorithm 4, at each iteration TACO-G selects the feasible example that yields the largest marginal increase in utility.

5.3 TACO-RG: Relevance-Guided Gain Greedy

Although **TACO-G** provides a simple yet effective greedy strategy, its selection step still requires evaluating $\Delta_f(e \mid S)$ over all feasible candidates in V . This can be computationally expensive when the candidate pool is large and the utility function involves non-trivial diversity computations. Motivated by the need to improve efficiency without compromising relevance, we develop **TACO-RG**, which accelerates gain-guided greedy selection by pruning the search space using relevance as a prerequisite.

As shown in Algorithm 5, following the relevance-guided filtering strategy introduced earlier, TACO-RG first constructs a reduced candidate pool $\mathcal{W}$ of size m by retaining the top-ranked examples with respect to the target query. It then applies the same pure gain-guided greedy rule as **TACO-G**, but restricts the feasible set to $\mathcal{W}$, selecting the example that yields the largest marginal increase in utility under the remaining budget B_r .

Algorithm 5 TACO-RG: Relevance-Restricted Gain-Guided Greedy Algorithm

Require: Candidate set V ; budget B ; target query q_t ; relevance score $\mathcal{R}$; sample size m ; cost function C ; utility function f

Ensure: Selected subset S

```

1:  $\mathcal{W} \leftarrow \text{TOPFILTER}(V, \mathcal{R}(\cdot \mid q_t), m)$   $\triangleright |\mathcal{W}| = m$ 
2:  $S \leftarrow \emptyset$ 
3:  $B_r \leftarrow B - (C_{\text{system}} + C_{q_t} + C_{\text{prefix}})$   $\triangleright$  remaining budget
4: while true do
5:    $\Omega \leftarrow \{e \in \mathcal{W} \setminus S : \Delta_C(e) \leq B_r\}$ 
6:   if  $\Omega = \emptyset$  then
7:     break
8:   end if
9:   Select  $e^* = \operatorname{argmax}_{e \in \Omega} \Delta_f(e \mid S)$ 
10:   $S \leftarrow S \cup \{e^*\}$ 
11:   $B_r \leftarrow B_r - \Delta_C(e^*)$ 
12: end while
13: return  $S$ 

```

5.4 Approximation Analysis

We now analyze the approximation behavior of the proposed TACO algorithms. Among the three variants, only **TACO-E** admits a provable approximation guarantee, owing to its partial-enumeration design combined with efficiency-guided greedy completion. The remaining variants, **TACO-G** and **TACO-RG**, are heuristic algorithms without formal approximation bounds and are primarily motivated by computational efficiency and empirical effectiveness. Before proceeding, we recall a classical result for the efficiency-guided greedy algorithm (Algorithm 2), which serves as the analytical foundation for the theoretical guarantee of **TACO-E**.

LEMMA 5.1. *Let $f : 2^V \rightarrow \mathbb{R}_{\geq 0}$ be monotone submodular, and let $c(S) = \sum_{e \in S} c(e)$ with $0 < c(e) \leq B$ for every $e \in V$. Define $S_0 = \emptyset$ and, for $i \geq 1$,*

$$e_i \in \operatorname{argmax}_{e \in V \setminus S_{i-1}} \frac{\Delta_f(e \mid S_{i-1})}{c(e)}, \quad S_i = S_{i-1} \cup \{e_i\}.$$

Then, for every i for which e_i is defined,

$$f(S_i) \geq \left(1 - \exp\left(-\frac{c(S_i)}{B}\right)\right) f(S^*),$$

where

$$S^* \in \operatorname{argmax}_{S \subseteq V : c(S) \leq B} f(S).$$

The sets S_i are analytical greedy prefixes.

LEMMA 5.2. *Let $\mathcal{W}$ be a candidate pool and let S_E be the solution returned by **TACO-E** (i.e., partial enumeration with depth $d = 3$ followed by efficiency-guided greedy completion within $\mathcal{W}$). Then*

$$f(S_E) \geq \left(1 - \frac{1}{e}\right) f(S_{\mathcal{W}}^*),$$

where $S_{\mathcal{W}}^$ is an optimal feasible solution to the BESM problem restricted to $\mathcal{W}$ under budget B .*

Following the classical analysis of monotone submodular maximization under a knapsack constraint [37], Lemma 5.1 and Lemma 5.2 follow from the density-greedy recurrence and the depth-three partial-enumeration argument. We next characterize the approximation behavior of TACO-E under relevance-based filtering. Intuitively, if the relevance-filtered candidate pool still contains all elements of an optimal solution, then TACO-E behaves like the idealized partial-enumeration scheme analyzed in Lemma 5.2 and inherits its $(1 - 1/e)$ guarantee. The additional error comes from the possibility that the filtering step removes some critical elements of the optimal solution.

THEOREM 5.3. *Let S_E denote the solution returned by **TACO-E**. For enumeration depth $d = 3$, with probability at least $1 - \varepsilon$, it holds that*

$$f(S_E) \geq \left(1 - \frac{1}{e} - \varepsilon\right) f(S^*),$$

where S^* denotes an optimal solution to the BESM problem under budget B . Here ε depends on the size m of the relevance-filtered candidate pool: as m increases, the probability that the filter discards critical elements of S^* decreases, and thus ε monotonically decreases.

PROOF SKETCH. Let $\mathcal{W}$ denote the relevance-filtered candidate pool used by TACO-E. Following the standard reasoning for greedy optimization over reduced ground sets [32], the filtering step can be viewed as inducing a randomized restriction of the original ground set V to $\mathcal{W}$. Let $\mathcal{E}$ be the event that the optimal solution S^* for the original instance is fully contained in $\mathcal{W}$. By construction of the filter and the choice of m , we assume that

$$\Pr[\mathcal{E}] \geq 1 - \varepsilon.$$

Conditioned on event $\mathcal{E}$, the restricted instance over $\mathcal{W}$ still admits S^* as an optimal feasible solution under budget B . In this case, TACO-E coincides with the idealized partial-enumeration scheme analyzed in Lemma 5.2, and we obtain

$$f(S_E) \geq \left(1 - \frac{1}{e}\right) f(S^*).$$

Since $1 - \frac{1}{e} \geq 1 - \frac{1}{e} - \varepsilon$ for any $\varepsilon \geq 0$, the weaker inequality

$$f(S_E) \geq \left(1 - \frac{1}{e} - \varepsilon\right) f(S^*)$$

also holds whenever $\mathcal{E}$ occurs. Therefore, with probability at least $\Pr[\mathcal{E}] \geq 1 - \varepsilon$, the stated bound is satisfied. $\square$

5.5 Algorithmic Complexity

Table 1 compares the space and time complexities of the baseline and proposed algorithms. Let N denote the total number of candidates, M the size of the relevance-filtered subset ($M \ll N$), and K the maximum number of greedy iterations enforced by the token budget. As illustrated, classical algorithms [17, 37] operate directly on the entire candidate pool, incurring high-degree polynomial time complexities (e.g., $O(N^5)$, $O(N^2)$, and $O(N^{d+2})$) that scale poorly for large-scale retrieval. In contrast, TACO variants alleviate this computational bottleneck. By first pruning the search space to a concise subset ($M \ll N$) and subsequently bounding the optimization process via K , TACO effectively decouples the dominant complexity terms from the full candidate size N . For the TACO-E variant, the combinatorial overhead is governed by the enumeration depth d , where E_1 and E_2 represent the number of feasible small subsets ($|S| < d$) explored during initialization and the valid size- d seed subsets used for subsequent greedy completion. Detailed empirical evaluations regarding execution time and performance trade-offs are provided in Section 7.3 and Table 5.

Table 1. Space and Time Complexity Analysis.

Algorithm	Space Complexity	Time Complexity
Classical Algorithms		
Sviridenko [37]	$O(N)$	$O(N^5)$
Khuller (Alg-1) [17]	$O(N)$	$O(N^2)$
Khuller (Alg-2) [17]	$O(N)$	$O(N^{d+2})$
Our Algorithms		
TACO-E	$O(M)$	$O(N \log M + E_1 d^3 + E_2 \cdot MK^3)$
TACO-G	$O(N)$	$O(NK^3)$
TACO-RG	$O(M)$	$O(N \log M + MK^3)$

6 Influence Modeling

6.1 Supervision Dataset Construction

To equip the influence model with the ability to discern both structural validity and execution correctness, we construct a supervision dataset where eligible query–example pairs (q, e) are assigned a continuous relevance label $y(q, e) \in [0, 1]$. This label is synthesized from three distinct signal sources: tag-based overlap, skeleton alignment, and execution fidelity.

Structural Alignment Signals. We measure the static structural alignment between the target query q and a candidate e via two complementary signals: fine-grained tag matching and abstract skeleton agreement.

Tag-based Similarity (s_{tag}). Let $T(q)$ and $T(e)$ denote the sets of full tags (key-relation-value triples), keys, and values extracted from the queries. We define a heuristic scoring function $s_{\text{tag}}(q, e)$ to quantify their overlap:

$$s_{\text{tag}}(q, e) = \begin{cases} 1.0, & \text{if } T(q) \equiv T(e) \wedge T(q) \neq \emptyset, \\ 0.95, & \text{if } T(q) \cap T(e) \text{ contains a full tag,} \\ 0.8, & \text{if } T(q) \cap T(e) \text{ shares both keys and values,} \\ 0.7, & \text{if } T(q) \cap T(e) \neq \emptyset. \end{cases}$$

Skeleton Similarity (s_{skel}). To capture high-level structure, we normalize queries into skeletons. If the skeletons of q and e are valid and identical, we assign a fixed bonus $s_{\text{skel}}(q, e) = 0.85$.

These two metrics are aggregated into a unified structural score s_{struct} based on signal availability:

$$s_{\text{struct}}(q, e) = \begin{cases} \max(s_{\text{tag}}, s_{\text{skel}}), & \text{if both signals exist,} \\ s_{\text{tag}}, & \text{if only } s_{\text{tag}} \text{ exists,} \\ s_{\text{skel}}, & \text{if only } s_{\text{skel}} \text{ exists.} \end{cases}$$

Execution Fidelity (s_{exec}). Beyond static analysis, we incorporate dynamic feedback via the Overpass Query Similarity (OQS). We discretize the raw $\text{OQS}(q, e)$ score to filter noise and emphasize high-quality execution matches:

$$s_{\text{exec}}(q, e) = \begin{cases} \text{OQS}, & \text{if } \text{OQS} > 0.8, \\ 0.5, & \text{if } 0.4 \leq \text{OQS} \leq 0.8, \\ 0.1, & \text{if } 0 < \text{OQS} < 0.4, \\ 0.0, & \text{if } \text{OQS} = 0. \end{cases}$$

Label Aggregation. The final continuous supervision label $y(q, e)$ combines the structural and execution signals. While the vast majority of supervision labels are automatically generated from these signals, we incorporate 133 manually labeled, high-confidence skeleton pairs. To prioritize this expert knowledge, these manually curated pairs are strictly assigned $y(q, e) = 1.0$. For all other

pairs possessing valid structural or execution signals, we employ a weighted aggregation controlled by a coefficient β (empirically 0.65):

$$y(q, e) = \begin{cases} 1.0, & \text{if pair is manually curated,} \\ \beta \cdot s_{\text{struct}} + (1 - \beta) \cdot s_{\text{exec}}, & \text{if both signals exist,} \\ s_{\text{struct}}, & \text{if only } s_{\text{struct}} \text{ exists,} \\ s_{\text{exec}}, & \text{if only } s_{\text{exec}} \text{ exists.} \end{cases}$$

Pairs lacking both structural and execution relevance are naturally excluded from the training set.

6.2 Angle-Based Influence Learning

To better capture fine-grained semantic relevance that is critical for example selection under strict token budgets, we train the influence model using an angle-based objective that emphasizes *directional semantics* in the embedding space. Unlike magnitude-based or global distance objectives, the angle-based formulation directly optimizes the angular alignment between examples and the target query, enabling more discriminative relevance estimation in regimes where conventional similarity measures may saturate.

Complex Text Embeddings. To improve discriminability in regions where conventional cosine similarity becomes less informative, we encode each input text as a complex-valued embedding, in which the real and imaginary components capture complementary semantic signals. Given an input text x , the embedding layer produces

$$\mathbf{E} = \text{Emb}(x) \in \mathbb{R}^{n \times 2d},$$

which is then split into real and imaginary parts and fed into a Transformer encoder. After pooling, we obtain a text-level complex representation

$$\mathbf{X} = \mathbf{X}^{\text{re}} + i\mathbf{X}^{\text{im}} \in \mathbb{C}^d, \quad \mathbf{X}^{\text{re}}, \mathbf{X}^{\text{im}} \in \mathbb{R}^d.$$

Angle-Based Objective. While angle-based metrics excel at capturing fine-grained, utility-oriented directional semantics, they omit magnitude information essential for basic relevance. Conversely, cosine similarity provides a robust signal for global semantic proximity. We thus propose a hybrid training objective that synergizes both metrics for a holistic relevance measurement.

Formally, given the complex embeddings of a candidate query q_i and a target query q_t , we quantify their directional discrepancy via the angular difference $\Delta\theta_{it}$ (as defined in Eq. (4)). Simultaneously, to ensure global semantic consistency, we compute the standard cosine similarity on the magnitude-normalized real components. The influence model is therefore trained by jointly optimizing an angle-based ranking loss alongside a cosine-based regularization term. The angle-specific component is formulated as:

$$\mathcal{L}_{\text{angle}} = \log \left(1 + \sum_{s_{ij} > s_{mn}} \exp \left(\frac{\Delta\theta_{ij} - \Delta\theta_{mn}}{\tau} \right) \right),$$

where τ represents a temperature hyperparameter controlling the sharpness of the distribution, and pairs (i, j) versus (m, n) represent relative ranking constraints derived from the supervision labels. The final objective combines $\mathcal{L}_{\text{angle}}$ with its cosine counterpart through a weighted summation, ensuring the encoder captures both coarse semantic alignment and subtle directional relevance.

7 Experiments and Discussion

7.1 Experimental Setup

OsmNL Dataset The existing OverpassNL dataset [36] suffers from substantial noise due to weakly aligned Text-OverpassQL pairs. To address this, we conducted a rigorous five-month human-in-the-loop refinement campaign. LLM first flagged 575 inconsistent instances, which were subsequently rewritten manually and verified via execution-based checking to guarantee syntactic and OSM tagging compliance. As detailed in Table 2, OsmNL contains 8,180 verified question-program pairs spanning 5,333 unique OSM tags and 1,429 geographic regions, split into 6,180 training, 1,000 development, and 1,000 test examples.

Table 2. Statistics of the **OsmNL** dataset.

Split	Samples	Unique Tags	Unique Regions
Train	6,180	4,517	1,146
Dev	1,000	1,181	351
Test	1,000	1,212	371
Total	8,180	5,333	1,429

Budget Setting. We set the default token budgets B to 400/600/1000 tokens, enabling a controlled comparison with standard 3/5/10-shot In-Context Learning (ICL), which empirically consume 453.6, 648.0, and 1136.9 prompt tokens on average, respectively. This setting allows us to assess whether TACO can improve generation quality under comparable or even stricter token constraints than standard ICL. The lower bound of 400 tokens is chosen for feasibility, as the mandatory system prefix and long user queries may leave little room for demonstrations under smaller budgets. The upper bound of 1000 tokens reflects realistic deployment constraints, where longer prompts incur higher API cost and serving overhead.

Backbones and Baselines. We adopt DeepSeek-V3.2 [3] as the LLM backbone and use Qwen3-Embedding-0.6B [33] as the shared embedding model across all settings. Token consumption is measured separately using DeepSeek-llm-7B-chat. We evaluate TACO against three baseline families: (i) *Fine-tuned models*: OverpassT5-base [36] and OsmT-base [38]; (ii) *LoRA-adapted LLMs*: Qwen2.5-Coder-7B [40], Qwen2.5-7B [40], and Llama-3.1-8B [11]; (iii) *ICL*: Fixed- k shots ($k \in \{3, 5, \dots, 100\}$), Cosine Similarity (CosSim) selection, and SimCSE [10] selection adapted on dataset (Section 6.1).

Evaluation Metrics. Following [36, 39], we evaluate generation quality along three primary axes: Exact Match (**EM**), execution fidelity, and *Overpass Query Similarity* (**OQS**). OQS is a composite metric that averages three complementary dimensions: surface fidelity measured by **chrF**, semantic correctness measured by **KVS**, and structural alignment measured by **TreeS**. Formally, $OQS = \frac{1}{3}(\text{chrF} + \text{KVS} + \text{TreeS})$. For execution fidelity, we report standard Execution Accuracy (**EX**), Soft Execution Accuracy (**EX_{soft}**), and the number of Execution Failures (**EN**). Additionally, prompt token statistics (min/max/mean) are tracked to monitor inference costs.

Implementation Details The influence model is trained using a jointly optimized Angle and Cosine objective with equal weighting, with the contrastive temperature fixed at $\tau = 0.5$. For TACO-E, we set the enumeration depth to $d = 3$. All remaining hyperparameters are tuned on the OsmNL development split, and the configuration achieving the highest mean OQS is selected. For the fine-tuned variants, LoRA adapters are applied to all linear projections of the BF16 backbone with rank $r=128$ and scaling factor $\alpha=256$. All generated OverpassQL programs are executed against the same OpenStreetMap snapshot used by OverpassNL [36], and fine-tuned results are reported as the mean and standard deviation over three random seeds.

Table 3. Evaluation results on OsmNL dev set. **Bold** and **blue** indicate the best result in each setting and overall.

Method	EM $\uparrow$	chrF $\uparrow$	KVS $\uparrow$	TreeS $\uparrow$	OQS $\uparrow$	Tokens (Min) $\downarrow$	Tokens (Max) $\downarrow$	Tokens (Mean) $\downarrow$
<i>Finetuned Model</i>								
CodeT5+ (220m)	15.1 $\pm$ 1.4	73.6 $\pm$ 2.6	64.0 $\pm$ 2.5	71.7 $\pm$ 2.0	69.8 $\pm$ 2.3	-	-	-
ByT5-small	16.9 $\pm$ 0.6	75.3 $\pm$ 0.3	66.4 $\pm$ 0.8	71.7 $\pm$ 0.4	71.1 $\pm$ 0.5	-	-	-
ByT5-base	18.1 $\pm$ 1.2	75.1 $\pm$ 0.9	66.0 $\pm$ 1.6	72.0 $\pm$ 0.4	71.1 $\pm$ 1.0	-	-	-
OverpassT5-small [36]	17.7 $\pm$ 0.9	74.8 $\pm$ 0.6	66.0 $\pm$ 0.8	71.9 $\pm$ 0.2	70.9 $\pm$ 0.4	-	-	-
OverpassT5-base [36]	18.6 $\pm$ 0.4	75.9 $\pm$ 0.5	67.3 $\pm$ 0.8	72.6 $\pm$ 0.3	72.0 $\pm$ 0.4	-	-	-
OSMT-base [38]	18.7 $\pm$ 0.4	76.4$\pm$0.0	67.9 $\pm$ 0.2	72.7 $\pm$ 0.2	72.4$\pm$0.1	-	-	-
LoRA-Qwen2.5-Coder-7B-Instruct	16.1 $\pm$ 0.4	74.0 $\pm$ 0.6	65.3 $\pm$ 0.9	71.5 $\pm$ 0.2	70.3 $\pm$ 0.5	-	-	-
LoRA-Qwen2.5-7B-Instruct	18.8$\pm$0.4	75.2 $\pm$ 0.7	68.0$\pm$1.0	72.8$\pm$0.4	72.0 $\pm$ 0.7	-	-	-
LoRA-Llama-3.1-8B-Instruct	18.6 $\pm$ 1.8	75.3 $\pm$ 1.1	67.5 $\pm$ 1.5	72.3 $\pm$ 0.5	71.7 $\pm$ 1.0	-	-	-
<i>k-Shot CosSim In-Context Learning within DeepSeek-V3.2</i>								
3-Shot	20.2	75.1	69.6	69.8	71.5	325.0	1365.0	453.6
5-Shot	21.2	76.5	71.0	71.0	72.9	453.0	1915.0	648.0
10-Shot	22.8	77.7	72.4	73.5	74.6	777.0	3357.0	1136.9
20-Shot	23.4	78.3	73.4	74.5	75.4	1419.0	4669.0	2118.5
50-Shot	23.9	78.8	74.4	75.0	76.1	3456.0	8699.0	5051.2
100-Shot	24.2	79.4	74.9	75.5	76.6	6978.0	15352.0	9921.5
<i>Budget = 400 In-Context Learning within DeepSeek-V3.2</i>								
CosSim	18.8	74.0	68.5	58.8	70.7	357.0	400	385.6
SimCSE	21.0	76.0	71.0	70.7	72.6	357.0	400	386.4
TACO-E (Ours)	21.0	76.5	72.5	71.9	73.7	241.0	400	377.5
TACO-G (Ours)	21.3	77.4	73.5	72.2	74.4	357.0	400	386.2
TACO-RG (Ours)	21.5	77.5	74.0	72.5	74.7	285.0	400	374.9
<i>Budget = 600 In-Context Learning within DeepSeek-V3.2</i>								
CosSim	21.3	76.1	70.6	71.5	72.7	557.0	600.0	586.4
SimCSE	22.7	77.7	72.5	73.8	74.7	557.0	600.0	587.3
TACO-E (Ours)	23.5	78.4	73.8	75.2	75.8	465.0	600.0	580.6
TACO-G (Ours)	23.8	78.6	74.6	74.8	76.0	557.0	600.0	586.7
TACO-RG (Ours)	23.9	78.8	74.5	74.7	76.0	465.0	600.0	566.5
<i>Budget = 1000 In-Context Learning within DeepSeek-V3.2</i>								
CosSim	22.1	77.1	72.0	73.1	74.1	957.0	1000.0	986.4
SimCSE	24.2	78.6	73.8	75.0	75.8	957.0	1000.0	986.4
TACO-E (Ours)	24.3	79.3	75.2	75.8	76.8	757.0	1000.0	943.1
TACO-G (Ours)	24.7	79.4	75.5	75.6	76.8	957.0	1000.0	986.9
TACO-RG (Ours)	25.2	79.4	75.9	75.5	76.9	757.0	1000.0	940.2

7.2 Main Results

Table 3 shows that the proposed TACO family consistently delivers the strongest in-context generation quality under explicit token constraints. Across $B=400/600/1000$, TACO-RG achieves the best or tied-best OQS (74.7 at $B=400$ and 76.9 at $B=1000$), outperforming CosSim and SimCSE retrieval. TACO also yields substantially better token efficiency than naive k -shot scaling: 100-shot ICL reaches OQS 76.6 with 9921.5 mean prompt tokens, whereas TACO-RG attains OQS 76.9 using only 940.2 mean tokens at $B=1000$ (over $10\times$ fewer tokens). Among variants, TACO-E is the most budget-frugal, TACO-G remains competitive at larger budgets, and TACO-RG provides the best overall accuracy–cost balance.

Table 4 further validates these findings on the test split with execution-based metrics (EX/EX_{soft}) and the execution error count (EN). Across all budgets, TACO consistently improves OQS and execution correctness over CosSim/SimCSE retrieval. At $B=600$ and $B=1000$, TACO-RG delivers the strongest overall results, achieving the best or tied-best OQS while also leading on EX/EX_{soft} and EM. At $B=2000$, TACO-RG matches the best OQS achieved by large k -shot prompting with only about 20% of the prompt tokens required by 100-shot, and it also attains the best execution scores with the lowest EN. This comparison also indicates diminishing returns of naive scaling, where increasing from 50-shot to 100-shot fails to improve performance and can slightly degrade it.

7.3 Efficiency and Scalability Analysis

Complexity and Performance Analysis. As shown in Table 5, classical $O(N^5)$ algorithms (Sviridenko [37] and Khuller et al. [17] Algorithm-2) are practically intractable due to the scalability

Table 4. Performance comparison on the OsmNL **test** set. The best result in each setting is shown in **bold**. The overall **best-performing** method is marked in **Blue**. * indicating that the input may exceed the budget when the query q_t itself consumes most tokens, leaving no room for selecting examples, in which case the method falls back to a basic prompt for zero-shot inference.

Method	EM $\uparrow$	OQS $\uparrow$	EX $\uparrow$	EX _{soft} $\uparrow$	EN $\downarrow$	Tokens (Min) $\downarrow$	Tokens (Max) $\downarrow$	Tokens (Mean) $\downarrow$
<i>Finetuned Model</i>								
OverpassT5-base [36]	18.4 $\pm$ 1.9	70.6 $\pm$ 1.2	37.1 $\pm$ 1.7	49.8 $\pm$ 1.7	47 $\pm$ 8	-	-	-
OsmT-base [38]	18.5 $\pm$ 0.7	71.4$\pm$0.2	38.0 $\pm$ 1.2	50.1 $\pm$ 0.7	44$\pm$2	-	-	-
LoRA-Qwen2.5-Coder-7B-Instruct	15.8 $\pm$ 0.3	69.7 $\pm$ 0.7	34.2 $\pm$ 1.3	43.7 $\pm$ 2.5	335 $\pm$ 50	-	-	-
LoRA-Qwen2.5-7B-Instruct	18.3 $\pm$ 0.7	71.1 $\pm$ 0.7	40.4$\pm$1.6	53.3 $\pm$ 1.3	73 $\pm$ 22	-	-	-
LoRA-Llama-3.1-8B-Instruct	18.8$\pm$1.0	71.0 $\pm$ 1.2	40.2 $\pm$ 1.6	53.7$\pm$1.8	45 $\pm$ 11	-	-	-
<i>Budget = 400 vs Non-budgeted k-shot</i>								
3-Shot	18.6	70.9	41.8	56.5	43	320.0	1342.0	458.3
CosSim	16.8	69.9	41.4	55.7	44	357.0	465.0*	386.8
SimCSE	19.1	70.4	43.6	56.6	37	357.0	465.0*	387.0
TACO-E (Ours)	20.1	71.5	43.6	58.1	39	231.0	465.0*	378.2
TACO-G (Ours)	20.2	71.5	44.9	59.2	37	357.0	465.0*	387.1
TACO-RG (Ours)	20.2	71.8	44.1	58.3	29	286.0	465.0*	370.2
<i>Budget = 600 vs Non-budgeted k-shot</i>								
5-Shot	20.0	72.1	44.4	59.5	38	428.0	1944.0	656.6
10-Shot	21.2	73.8	46.2	61.5	32	755.0	3271.0	1145.6
CosSim	19.5	71.3	42.3	56.9	47	557.0	600.0	586.7
SimCSE	21.7	73.2	45.9	59.8	33	557.0	600.0	586.9
TACO-E (Ours)	21.9	73.2	46.0	60.7	33	463.0	600.0	579.5
TACO-G (Ours)	21.4	73.2	46.4	62.2	29	557.0	600.0	587.8
TACO-RG (Ours)	22.2	73.9	47.1	61.9	31	461.0	600.0	570.8
<i>Budget = 1000 vs Non-budgeted k-shot</i>								
20-Shot	21.7	74.4	47.2	63.1	24	1448.0	4891.0	2127.4
30-Shot	22.5	75.0	48.3	64.1	27	2088.0	5999.0	3111.2
CosSim	21.0	73.4	46.1	61.1	38	957.0	1000.0	986.1
SimCSE	22.3	74.3	47.4	61.7	31	957.0	1000.0	986.3
TACO-E (Ours)	22.1	74.1	49.0	63.3	32	739.0	1000.0	944.9
TACO-G (Ours)	22.5	75.2	49.0	64.1	23	957.0	1000.0	987.2
TACO-RG (Ours)	22.7	75.1	49.0	64.7	25	912.0	1000.0	974.5
<i>Budget = 2000 vs Non-budgeted k-shot</i>								
50-Shot	23.7	75.8	50.2	65.3	24	3457.0	8212.0	5075.1
100-Shot	23.1	75.7	49.8	65.2	25	6891.0	14406.0	9961.5
CosSim	21.8	74.5	46.9	62.6	27	1957.0	2000.0	1986.0
SimCSE	23.3	75.1	49.1	64.0	27	1957.0	2000.0	1985.7
TACO-RG (Ours)	23.8	75.8	50.4	65.5	20	1908.0	2000.0	1975.5

Table 5. Approximate computational cost and OQS on OsmNL dev dataset ($B = 400$, $d = 3$, $N = 6180$).

Algorithm	Time Complexity	Approx. Cost $\downarrow$	OQS $\uparrow$
<i>Classical Algorithms</i>			
Sviridenko [37]	$O(N^5)$	9.01×10^{18}	-
Khuller (Algorithm-1) [17]	$O(N^2)$	3.82×10^7	61.41
Khuller (Algorithm-2) [17]	$O(N^5)$	9.01×10^{18}	-
<i>Our Algorithms</i>			
TACO-E	$O(N \log M + M^4 K^3)$	2.20×10^5	73.67
TACO-G	$O(NK^3)$	1.26×10^5	74.36
TACO-RG	$O(N \log M + MK^3)$	3.57×10^4	74.68

bottleneck imposed by the large candidate pool N , failing to yield results within a reasonable timeframe. While Khuller et al. [17] (Algorithm-1) is executable under $O(N^2)$ complexity, it yields a low OQS of 61.41. In contrast, our proposed TACO variants achieve substantial improvements in both efficiency and quality. Specifically, all TACO variants reduce the computational cost by 2 to 3 orders of magnitude compared to Khuller et al. (Algorithm-1) while improving OQS by over 12 points. TACO-RG delivers the best performance, achieving the highest OQS (74.68) with the minimum empirical cost (3.57×10^4), validating its efficacy in scaling to large-scale scenarios.

Cross-Backbone Pareto Frontier Analysis. To verify the cross-backbone generalization and robustness of TACO, we contrast the performance-efficiency trade-off on GPT-4o-mini (Figure 4)

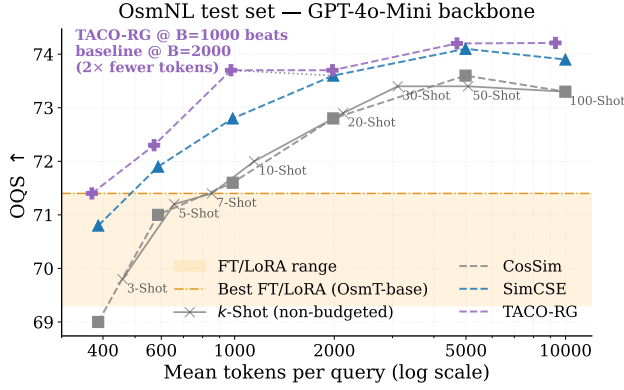


Fig. 4. Pareto frontier on the GPT-4o-Mini backbone.

Table 6. **Utility-component ablation on OsmNL dev dataset with TACO-G.** Ablating $\mathcal{M}$, $\mathcal{I}$, and $\mathcal{D}$ across budgets.

Setting	$\mathcal{M}$	$\mathcal{I}$	$\mathcal{D}$	OQS ↑		
				$B = 400$	$B = 600$	$B = 1000$
TACO-G	✓	✓	✓	74.4	76.0	76.8
w/o Diversity	✓	✓	✗	74.1	75.7	76.3
w/o Influence	✓	✗	✓	70.8	72.9	74.4
w/o Similarity	✗	✓	✓	73.9	76.0	76.0
Only Similarity	✓	✗	✗	70.7	72.7	74.1
Only Influence	✗	✓	✗	71.3	75.5	76.0

with that on DeepSeek-V3.2 (Figure 1). The empirical results indicate that TACO-RG consistently defines the optimal upper-left Pareto frontier across both backbones, exhibiting an invariant dominance over all baseline selectors irrespective of the model architecture openness. Notably, on the proprietary GPT-4o-mini backbone, TACO-RG yields over 2× token efficiency gains while maintaining superior or matching generation quality (OQS) compared to the strongest adapted embedding baseline (SimCSE) at the maximum budget ($B = 5000$). This consistent behavior highlights that the effectiveness of TACO-RG in computational context optimization is backbone-agnostic and robust under stringent token constraints.

7.4 Ablation Study

Utility Function Decomposition. As shown in Table 6, we perform a component-wise ablation of the utility function using TACO-G to isolate the impact of each term without pre-filtering interference. The full configuration ($\mathcal{M} + \mathcal{I} + \mathcal{D}$) establishes the performance upper bound across all token budgets. Most prominently, the exclusion of the influence term $\mathcal{I}$ precipitates the most substantial drop in OQS, demonstrating that directional influence delivers the vital discriminative signal for effective context selection. Conversely, removing the diversity term $\mathcal{D}$ causes a smaller, stable performance decrease, which confirms the efficacy of log-determinant regularization in mitigating demonstration redundancy.

Embedding Type Alternatives. Table 7 presents the ablation results regarding task-specific adaptation and embedding complexity across various token budgets. The empirical outcomes reveal two primary insights. First, removing the complex embedding component (degrading from TACO-G to SimCSE) incurs a consistent performance drop across all budgets, particularly decreasing the OQS by 1.8 points at $B = 400$. Second, completely disabling task-specific adaptation (as represented

Table 7. Ablation of task-specific fine-tuning and complex embeddings on OsmNL dev dataset across budgets.

Setting	Adapted	Complex	OQS $\uparrow$		
			$B = 400$	$B = 600$	$B = 1000$
TACO-G	✓	✓	74.4	76.0	76.8
SimCSE [10]	✓	✗	72.6	74.7	75.8
CosSim	✗	✗	70.7	72.7	74.1

Table 8. Ablation study on ϕ function on OsmNL dev set.

ϕ Function	OQS $\uparrow$		
	$B = 400$	$B = 600$	$B = 1000$
$\phi(x) = \log(1+x)/\log(2)$	74.4	76.0	76.8
$\phi(x) = \sqrt{x}$	74.1	75.8	76.6
$\phi(x) = x$	74.3	75.9	76.5

Table 9. Ablation study on Influence modeling.

Setting	Spearman $\uparrow$	Pearson $\uparrow$	MSE $\downarrow$
Angle & Cosine Objective	0.5984	0.5864	0.1657
Only Angle Objective	0.5814	0.5737	0.1688
Only Cosine Objective	0.5764	0.5676	0.1716

by CosSim) further degrades the generation quality, yielding the lowest OQS across the board. The substantial performance margin between the full TACO-G configuration and the ablated variants demonstrates that both task-specific fine-tuning and advanced embedding formulations are indispensable for effective context optimization under tight token constraints.

Concave Transformation ϕ Analysis. Table 8 presents a sensitivity study of TACO-G regarding the functional form of ϕ . The empirical outcomes highlight the architectural robustness of the framework, with different variants delivering near-identical OQS values across all budgets. In particular, $\phi(x) = \log(1+x)/\log(2)$ maintains a slight but stable lead, surpassing $\sqrt{x}$ and the identity mapping by a slim margin of 0.1 to 0.3 points. This finding suggests that although the introduction of ϕ is necessary to model diminishing returns during relevance aggregation, the optimization quality depends primarily on the core utility design instead of the specific curvature of the transformation.

Influence Model Training Objectives. As detailed in Table 9, we investigate the performance impact of different training objectives on influence modeling. The dual formulation combining angle and cosine objectives consistently establishes the performance upper bound, delivering the strongest correlations (0.598 Spearman and 0.586 Pearson) and the lowest MSE (0.166). In contrast, employing either objective in isolation yields substantially weaker alignment. This behavior underscores the necessity of the angle term as a scale-invariant directional constraint; when integrated with cosine similarity, it provides a stable influence signal that enhances the accuracy of demonstration ranking.

Hyperparameter Sensitivity. Figure 5 shows TACO-G’s sensitivity to the relevance mixing factor α and diversity weight λ on the dev set. Performance is robust across a broad range and peaks at intermediate values. Specifically, $\alpha = 0.2$ achieves the best overall OQS (e.g., 76.7 at $B = 1000$); exceeding this threshold increasingly down-weights complementary influence signals, slightly degrading results. Similarly, a balanced diversity strength is optimal, with $\lambda = 0.5$ yielding the peak OQS (76.8 at $B = 1000$), whereas deviating from this value incurs minor losses. We thus deploy this dev-optimal configuration in subsequent experiments.

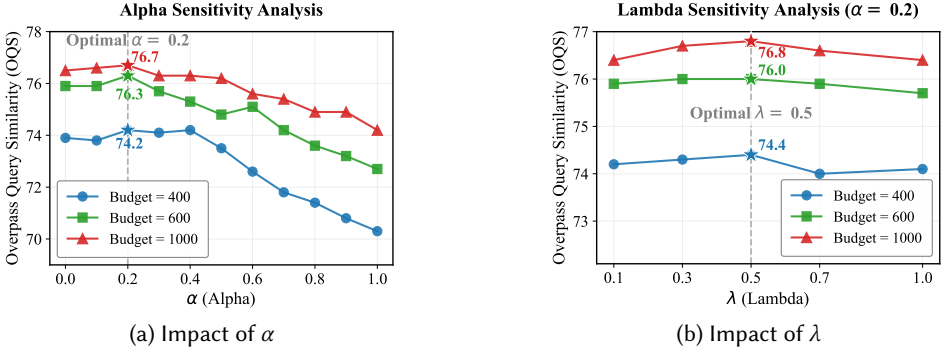


Fig. 5. Sensitivity analysis of TACO-G on the OsmNL development set under different budget settings with respect to (a) the relevance factor $\alpha \in \{0, 0.1, \dots, 1.0\}$ and (b) the diversity weight $\lambda \in \{0.1, 0.3, 0.5, 0.7, 1.0\}$.

Table 10. Cross-domain Text-to-SQL performance on Spider.

Selection	$B=400$	$B=600$	$B=900$	$B=1200$
Question Similarity [12]	0.771	0.796	0.805	0.810
Masked Question Similarity [12]	0.781	0.813	0.825	0.828
DAIL [12]	0.789	0.809	0.823	0.830
TACO-RG (Ours)	0.796	0.817	0.829	0.841

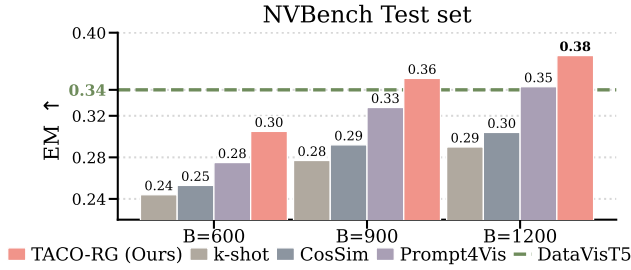


Fig. 6. Cross-domain generalization performance on NVBench test set for Text-to-Visualization task.

7.5 Cross-Domain Generalization

While TACO focuses on Text-to-OverpassQL, the BESM formulation captures a general context-selection problem for structured query generation from natural language, where the goal is to select a compact and informative set of demonstrations under an explicit token budget. We therefore evaluate TACO-RG on two additional tasks, Text-to-SQL and Text-to-Visualization, to assess whether its effectiveness extends beyond OverpassQL. Table 10 presents the Text-to-SQL results on Spider [41] using DeepSeek-V4-flash [4]. Compared with three SQL-specific example selectors from DAIL-SQL [12], TACO-RG consistently yields the highest accuracy across all tested budgets, achieving 0.796 at $B = 400$ and 0.841 at $B = 1200$. Figure 6 reports the Text-to-Visualization results on NVBench [28] using GPT-4o-mini. TACO-RG consistently outperforms basic selection baselines and Prompt4Vis [24] across all tested budgets. At larger budgets, TACO-RG also surpasses the fine-tuned DataVisT5 baseline [39], achieving an EM accuracy of 0.38 compared with 0.34 for DataVisT5. These results provide additional evidence that the proposed budget-aware selection objective can transfer beyond OverpassQL-specific syntax. Although the primary empirical focus remains geospatial query generation, TACO-RG yields effective contexts for other structured generation tasks without task-specific optimization.

8 Related Work

Structured Query Generation. Translating natural language into executable structured queries is a long-standing challenge in database research [15, 22, 27], with representative settings including Text-to-SQL [8, 9, 12, 22], Text-to-Visualization [24, 45], and recently Text-to-OverpassQL [36, 38]. In the geospatial domain, early semantic parsing studies mapped questions to logical forms over small-scale databases, as exemplified by *GeoQuery* [16, 42, 43]. With the rise of OpenStreetMap as a large-scale geospatial knowledge base, subsequent work introduced intermediate machine-readable languages grounded in OSM tags, such as *NLmaps* [20, 21], before recent efforts moved toward direct OverpassQL generation. While these studies improve the accuracy and executability of geospatial query generation, they rarely address how demonstration contexts should be selected under explicit token budgets. This gap becomes critical in retrieval-augmented and agentic LLM settings, where prompt length affects latency, serving cost, and the amount of useful context that can be included. Motivated by this limitation, we formulate budgeted demonstration selection for Text-to-OverpassQL as BESM and instantiate it through the TACO framework.

Data Subset Mining and Optimization. Data subset mining aims to identify a compact set of informative instances from large-scale data collections. It is commonly studied through submodular optimization, since diminishing returns provide a natural way to model redundancy among selected instances. Existing work has applied subset optimization to robust training, where clean samples are mined from noisy or mislabeled datasets to improve efficiency and resilience [6]. Representative subset mining has also been used for interpretability and attribution, including black-box attribution in computer vision and analyses of data points that contribute strongly to model performance [1, 2]. In addition, submodular selection has been adopted in targeted active learning for rare-class mining [19] and combined with Shapley value analysis to estimate modality importance in multimodal learning [14]. Our work differs by studying subset mining for token-budgeted demonstration construction in structured query generation, where utility and prompt feasibility must be optimized jointly.

9 Conclusion

This paper presents **TACO**, a token-aware context optimization framework for cost-efficient Text-to-OverpassQL generation. We introduced the high-quality OsmNL benchmark and formalized context construction as the BESM problem. By analyzing this NP-hard task under a submodular objective, we derived a family of budget-aware algorithms that theoretically balance semantic relevance with structural diversity, effectively navigating the trade-off between approximation guarantees and computational efficiency. Extensive evaluations demonstrated that TACO establishes a new state-of-the-art in query accuracy while reducing token consumption by over 50% compared to strong baselines. Ultimately, TACO bridges the gap between high-performance language modeling and resource-constrained geospatial data systems, offering a robust and cost-efficient foundation for large-scale query interfaces.

Acknowledgments

This work was partially supported by the NSFC/RGC Joint Research Scheme (N_PolyU5179/25), the Research Grants Council of Hong Kong (PolyU25600624), the Innovation and Technology Fund (ITS/052/23MX, PRP/009/22FX, and PRP/004/25FX), and the RCDTT Project (P0058790, RAG4Tourism: Advancing Transparent and Collaborative Travel Planning with LLM).

References

- [1] Ruoyu Chen, Siyuan Liang, Jingzhi Li, Shiming Liu, Li Liu, Hua Zhang, and Xiaochun Cao. 2025. Less is More: Efficient Black-box Attribution via Minimal Interpretable Subset Selection. arXiv:2504.00470 [cs.LG] <https://arxiv.org/abs/2504.00470>

00470

- [2] Anshuman Chhabra, Peizhao Li, Prasant Mohapatra, and Hongfu Liu. 2024. "What Data Benefits My Classifier?" Enhancing Model Performance and Interpretability through Influence-Based Data Selection. In *The Twelfth International Conference on Learning Representations*. OpenReview.net, Vienna, Austria, 32 pages. <https://openreview.net/forum?id=HE9eUQLAvo>
- [3] DeepSeek-AI. 2024. DeepSeek-V3 Technical Report. arXiv:2412.19437 [cs.CL] <https://arxiv.org/abs/2412.19437>
- [4] DeepSeek-AI. 2026. DeepSeek-V4: Towards Highly Efficient Million-Token Context Intelligence.
- [5] Cheng Deng, Tianhang Zhang, Zhongmou He, Yi Xu, Qiyuan Chen, Yuanyuan Shi, Luoyi Fu, Weinan Zhang, Xinbing Wang, Chenghu Zhou, Zhouhan Lin, and Junxian He. 2023. K2: A Foundation Language Model for Geoscience Knowledge Understanding and Utilization. arXiv:2306.05064 [cs.CL] <https://arxiv.org/abs/2306.05064>
- [6] Yuhao Deng, Chengliang Chai, Kaisen Jin, Linan Zheng, Lei Cao, Ye Yuan, and Guoren Wang. 2025. Two Birds with One Stone: Efficient Deep Learning over Mislabeled Data through Subset Selection. *Proc. ACM Manag. Data* 3, 3, Article 231 (June 2025), 28 pages. doi:10.1145/3728289
- [7] Yuankai Fan, Zhenying He, Tonghui Ren, Can Huang, Yinan Jing, Kai Zhang, and X. Sean Wang. 2024. Metasql: A Generate-Then-Rank Framework for Natural Language to SQL Translation. In *2024 IEEE 40th International Conference on Data Engineering (ICDE)*. IEEE Computer Society, Los Alamitos, CA, USA, 1765–1778. doi:10.1109/ICDE60146.2024.00143
- [8] Han Fu, Chang Liu, Bin Wu, Feifei Li, Jian Tan, and Jianling Sun. 2023. CatSQL: Towards Real World Natural Language to SQL Applications. *Proc. VLDB Endow.* 16, 6 (Feb. 2023), 1534–1547. doi:10.14778/3583140.3583165
- [9] Dawei Gao, Haibin Wang, Yaliang Li, Xiuyu Sun, Yichen Qian, Bolin Ding, and Jingren Zhou. 2024. Text-to-SQL Empowered by Large Language Models: A Benchmark Evaluation. *Proc. VLDB Endow.* 17, 5 (Jan. 2024), 1132–1145. doi:10.14778/3641204.3641221
- [10] Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple Contrastive Learning of Sentence Embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 6894–6910. doi:10.18653/v1/2021.emnlp-main.552
- [11] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, and Abhishek Kadian et al. 2024. The Llama 3 Herd of Models. arXiv:2407.21783 [cs.AI] <https://arxiv.org/abs/2407.21783>
- [12] Zihui Gu, Ju Fan, Nan Tang, Lei Cao, Bowen Jia, Sam Madden, and Xiaoyong Du. 2023. Few-shot Text-to-SQL Translation using Structure and Content Prompt Learning. *Proc. ACM Manag. Data* 1, 2, Article 147 (June 2023), 28 pages. doi:10.1145/3589292
- [13] Junlin He, Tong Nie, and Wei Ma. 2025. Geolocation representation from large language models are generic enhancers for spatio-temporal learning. *Proceedings of the AAAI Conference on Artificial Intelligence* 39, 16 (2025), 17094–17104. doi:10.1609/aaai.v39i16.33879
- [14] Yifei He, Runxiang Cheng, Gargi Balasubramaniam, Yao-Hung Hubert Tsai, and Han Zhao. 2024. Efficient Modality Selection in Multimodal Learning. *Journal of Machine Learning Research* 25, 47 (2024), 1–39. <http://jmlr.org/papers/v25/23-0439.html>
- [15] Zijin Hong, Zheng Yuan, Qinggang Zhang, Hao Chen, Junnan Dong, Feiran Huang, and Xiao Huang. 2025. Next-Generation Database Interfaces: A Survey of LLM-based Text-to-SQL. arXiv:2406.08426 [cs.CL] <https://arxiv.org/abs/2406.08426>
- [16] Rohit Kate, Yuk Wong, and Raymond Mooney. 2005. Learning to transform natural to formal languages. In *Proceedings of the Twentieth National Conference on Artificial Intelligence (AAAI-05)*. AAAI Press, Pittsburgh, PA, USA, 1062–1068.
- [17] Samir Khuller, Anna Moss, and Joseph (Seffi) Naor. 1999. The budgeted maximum coverage problem. *Inform. Process. Lett.* 70, 1 (1999), 39–45.
- [18] Chun-Wa Ko, Jon Lee, and Maurice Queyranne. 1995. An exact algorithm for maximum entropy sampling. *Operations Research* 43, 4 (1995), 684–691.
- [19] Suraj Kothawade, Saikat Ghosh, Sumit Shekhar, Yu Xiang, and Rishabh Iyer. 2022. TALISMAN: Targeted Active Learning for Object Detection with Rare Classes and Slices using Submodular Mutual Information. arXiv:2112.00166 [cs.CV] <https://arxiv.org/abs/2112.00166>
- [20] Carolin Lawrence and Stefan Riezler. 2016. NLmaps: A Natural Language Interface to Query OpenStreetMap. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, Hideo Watanabe (Ed.). The COLING 2016 Organizing Committee, Osaka, Japan, 6–10. <https://aclanthology.org/C16-2002>
- [21] Carolin Lawrence and Stefan Riezler. 2018. Improving a Neural Semantic Parser by Counterfactual Learning from Human Bandit Feedback. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Iryna Gurevych and Yusuke Miyao (Eds.). Association for Computational Linguistics, Melbourne, Australia, 1820–1830. doi:10.18653/v1/P18-1169
- [22] Boyan Li, Yuyu Luo, Chengliang Chai, Guoliang Li, and Nan Tang. 2024. The Dawn of Natural Language to SQL: Are We Fully Ready? *Proc. VLDB Endow.* 17, 11 (July 2024), 3318–3331. doi:10.14778/3681954.3682003

- [23] Haoyang Li, Jing Zhang, Hanbing Liu, Ju Fan, Xiaokang Zhang, Jun Zhu, Renjie Wei, Hongyan Pan, Cuiping Li, and Hong Chen. 2024. CodeS: Towards Building Open-source Language Models for Text-to-SQL. *Proc. ACM Manag. Data* 2, 3, Article 127 (May 2024), 28 pages. doi:10.1145/3654930
- [24] Shuaimin Li, Xuanang Chen, Yuanfeng Song, Yunze Song, Chen Jason Zhang, Fei Hao, and Lei Chen. 2025. Prompt4Vis: prompting large language models with example mining for tabular data visualization. *The VLDB Journal* 34, Article 38 (2025), 26 pages. doi:10.1007/s00778-025-00912-0
- [25] Shuaimin Li, Yuanfeng Song, Xuanang Chen, Anni Peng, Zhuoyue Wan, Chen Jason Zhang, and Raymond Chi-Wing Wong. 2024. Mind Your Questions! Towards Backdoor Attacks on Text-to-Visualization Models. arXiv:2410.06782 [cs.CR] <https://arxiv.org/abs/2410.06782>
- [26] Xianming Li and Jing Li. 2024. AoE: Angle-optimized Embeddings for Semantic Textual Similarity. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 1825–1839. doi:10.18653/v1/2024.acl-long.101
- [27] Xinyu Liu, Shuyu Shen, Boyan Li, Peixian Ma, Runzhi Jiang, Yuxin Zhang, Ju Fan, Guoliang Li, Nan Tang, and Yuyu Luo. 2025. A Survey of NL2SQL with Large Language Models: Where are we, and where are we going? arXiv:2408.05109 [cs.DB] <https://arxiv.org/abs/2408.05109>
- [28] Yuyu Luo, Nan Tang, Guoliang Li, Chengliang Chai, Wenbo Li, and Xuedi Qin. 2021. Synthesizing Natural Language to Visualization (NL2VIS) Benchmarks from NL2SQL Benchmarks. In *Proceedings of the 2021 International Conference on Management of Data*. Association for Computing Machinery, Virtual Event, China, 1235–1247. doi:10.1145/3448016.3457261
- [29] Yan Luo, Zhuoyue Wan, Yuzhong Chen, Gengchen Mai, Fu lai Chung, and Kent Larson. 2024. TransFlower: An Explainable Transformer-Based Model with Flow-to-Flow Attention for Commuting Flow Prediction. arXiv:2402.15398 [cs.LG] <https://arxiv.org/abs/2402.15398>
- [30] Gengchen Mai, Weiming Huang, Jin Sun, Suhang Song, Deepak Mishra, Ninghao Liu, Song Gao, Tianming Liu, Gao Cong, Yingjie Hu, Chris Cundy, Ziyuan Li, Rui Zhu, and Ni Lao. 2023. On the Opportunities and Challenges of Foundation Models for Geospatial Artificial Intelligence. arXiv:2304.06798 [cs.AI] <https://arxiv.org/abs/2304.06798>
- [31] Rohin Manvi, Samar Khanna, Gengchen Mai, Marshall Burke, David Lobell, and Stefano Ermon. 2024. GeoLLM: Extracting Geospatial Knowledge from Large Language Models. arXiv:2310.06213 [cs.CL] <https://arxiv.org/abs/2310.06213>
- [32] Baharan Mirzasoleiman, Ashwinkumar Badanidiyuru, Amin Karbasi, Jan Vondrak, and Andreas Krause. 2014. Lazier Than Lazy Greedy. arXiv:1409.7938 [cs.LG] <https://arxiv.org/abs/1409.7938>
- [33] Qwen. 2025. Qwen3 Technical Report. arXiv:2505.09388 [cs.CL] <https://arxiv.org/abs/2505.09388>
- [34] Tonghui Ren, Yuankai Fan, Zhenying He, Ren Huang, Jiaqi Dai, Can Huang, Yanan Jing, Kai Zhang, Yifan Yang, and X. Sean Wang. 2024. PURPLE: Making a Large Language Model a Better SQL Writer. In *2024 IEEE 40th International Conference on Data Engineering (ICDE)*. IEEE Computer Society, Los Alamitos, CA, USA, 15–28. doi:10.1109/ICDE60146.2024.00009
- [35] Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Scharli, and Denny Zhou. 2023. Large Language Models Can Be Easily Distracted by Irrelevant Context. In *Proceedings of the 40th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 202)*. PMLR, Honolulu, Hawaii, USA, 31210–31227. <https://proceedings.mlr.press/v202/shi23a.html>
- [36] Michael Staniek, Raphael Schumann, Maike Züfle, and Stefan Riezler. 2024. Text-to-OverpassQL: A Natural Language Interface for Complex Geodata Querying of OpenStreetMap. *Transactions of the Association for Computational Linguistics* 12 (2024), 562–575. doi:10.1162/tacl_a_00654
- [37] Maxim Sviridenko. 2004. A note on maximizing a submodular set function subject to a knapsack constraint. *Operations Research Letters* 32, 1 (2004), 41–43.
- [38] Zhuoyue Wan, Wentao Hu, Chen Jason Zhang, Yuanfeng Song, Shuaimin Li, Ruiqiang Xiao, Xiao-Yong Wei, and Raymond Chi-Wing Wong. 2025. OsmT: Bridging OpenStreetMap Queries and Natural Language with Open-source Tag-aware Language Models. arXiv:2512.04738 [cs.CL] <https://arxiv.org/abs/2512.04738>
- [39] Zhuoyue Wan, Yuanfeng Song, Shuaimin Li, Chen Jason Zhang, and Raymond Chi-Wing Wong. 2025. DataVisT5: A Pre-Trained Language Model for Jointly Understanding Text and Data Visualization. In *2025 IEEE 41st International Conference on Data Engineering (ICDE)*. IEEE Computer Society, Los Alamitos, CA, USA, 1704–1717. doi:10.1109/ICDE65448.2025.00131
- [40] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024. Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115* (2024).

- [41] Tao Yu, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, et al. 2018. Spider: A Large-Scale Human-Labeled Dataset for Complex and Cross-Domain Semantic Parsing and Text-to-SQL Task. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 3911–3921. doi:10.18653/v1/D18-1425
- [42] John M. Zelle and Raymond J. Mooney. 1996. Learning to parse database queries using inductive logic programming. In *Proceedings of the Thirteenth National Conference on Artificial Intelligence - Volume 2* (Portland, Oregon) (AAAI'96). AAAI Press/MIT Press, Portland, OR, USA, 1050–1055.
- [43] Luke S. Zettlemoyer and Michael Collins. 2005. Learning to map sentences to logical form: Structured classification with probabilistic categorial grammars. In *Proceedings of the Twenty-First Conference on Uncertainty in Artificial Intelligence (UAI'05)*. AUAI Press, Arlington, Virginia, USA, 658–666.
- [44] Haodi Zhang, Xinrui Zhu, Xuan Wang, Hanlin Gu, Yuanfeng Song, and Lin Ma. 2025. Natural Language to Overpass Query: A Multi-Step Approach Using Task Decomposition and Key-Value Correction. In *Proceedings of the 26th IEEE International Conference on Mobile Data Management (MDM)*. IEEE, IEEE, Irvine, CA, USA, 1–11.
- [45] Weixu Zhang, Yifei Wang, Yuanfeng Song, Victor Junqiu Wei, Yuxing Tian, Yiyan Qi, Jonathan H. Chan, Raymond Chi-Wing Wong, and Haiqin Yang. 2024. Natural Language Interfaces for Tabular Data Querying and Visualization: A Survey. *IEEE Trans. on Knowl. and Data Eng.* 36, 11 (Nov. 2024), 6699–6718. doi:10.1109/TKDE.2024.3400824

Received 17 January 2026; revised 19 May 2026; accepted 12 June 2026