IAES International Journal of Artificial Intelligence (IJ-AI)
Vol. 15, No. 1, February 2026, pp. 129∼139
ISSN: 2252-8938, DOI: 10.11591/ijai.v15.i1.pp129-139

Automated data exploration with mutual information in natural language to visualization

Hue Luong-Thi-Minh[1], Vinh-The Nguyen[1], Van-Viet Nguyen[1], Kim-Son Nguyen[1], Huu-Khanh Nguyen[2]
[1]Faculty of Information Technology, Thai Nguyen University of Information and Communication Technology, Thai Nguyen, Viet Nam
[2]Distance Learning Center, Thai Nguyen University, Thai Nguyen, Viet Nam

Article Info

Article history:

Received Sep 22, 2025
Revised Nov 13, 2025
Accepted Jan 10, 2026

Keywords:

Evaluation and benchmarking
Feature selection
Information theory
Mutual information
Natural language to visualization

ABSTRACT

Transcribing natural language to visualization (NL2VIS) has been investigated for years but still suffer from several fundamental limitations (e.g., feature selection). Although large language models (LLMs) are good candidates but they incur computation cost and hard to trace their made decisions. To alleviate this problem, we introduced an alternative information-theoretic framework that utilized mutual information (MI) to quantify the statistical relationship between utterances and database features. In our approach, kernel density estimation (KDE) and neural estimation techniques were utilized to estimate MI, and to optimize a diversity-promoting objective balancing feature relevance and redundancy. We also introduced the information coverage ratio (ICR) to quantify the amount of information content preserved in feature selection decisions. In our experiments, we found that the proposed approach improved information-theoretic metrics, with F1-score of 0.863 and an ICR of 0.891. We observed that these improvements did not come at the cost of traditional benchmarks: validity reached 88.9%, legality 85.2%, and chart-type accuracy 87.6%. Moreover, significance tests ($p < 0.001$) and large effect sizes (Cohen's $d > 0.8$) further supported that these improvements were meaningful for feature selection. Thus, this study provides a mathematical framework for applications requiring analytical validity that extends beyond NL2VIS to other machine learning contexts.

This is an open access article under the CC BY-SA license.

Corresponding Author:

Vinh The Nguyen
Faculty of Information Technology, Thai Nguyen University of Information and Communication Technology
Thai Nguyen, Viet Nam
Email: vinhnt@ictu.edu.vn

1. INTRODUCTION

In the era of big data, consuming a large amount of information plays a crucial role in the decision-making process, and data visualization (VIS) is a viable solution [1]–[3]. Traditional VIS tools relied on rules, heuristics and probability, creating a barrier for non-technical users [4]–[6]. Recently, natural language for data visualization (NL2VIS) has emerged as one of the most promising approaches that allows users to generate visualizations (e.g., bar charts, line graphs, scatter plots, and heat maps) using only simple conversational utterances [7], [8]. For instance, instead of writing a computer language such as "SELECT region, SUM(revenue) FROM sales WHERE date >= '2023-01-01' GROUP BY region ORDER BY SUM(revenue) DESC", a user may use natural language "show me total sales by region this year". The system then interprets the request and builds a corresponding visualization, as illustrated in Figure 1, which

presents an example of the NL2VIS problem. Thus, this idea fundamentally shortens the gap between domain experts and normal users in data analysis workflows [2], [5].

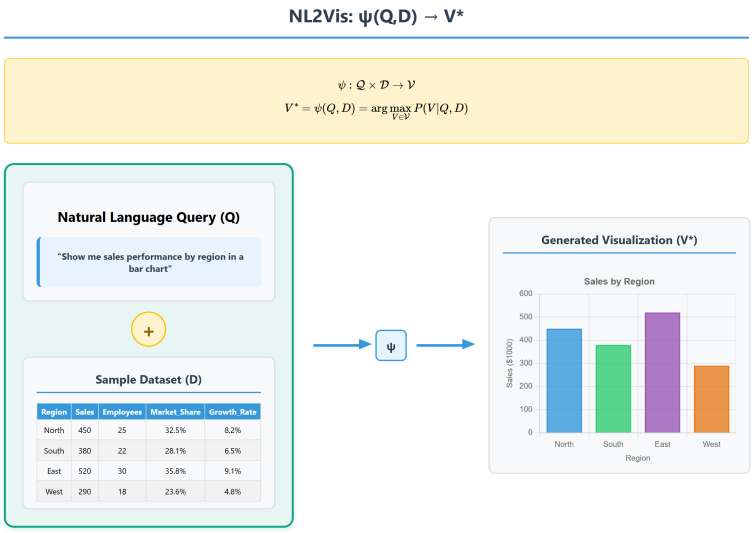

Figure 1. Example of NL2VIS problem

In formal terms, the NL2VIS problem can be seen as a transformation $\psi : (\mathcal{Q}, \mathcal{D}) \rightarrow \mathcal{V}$, where a natural-language query Q and dataset D are transformed into an appropriate visualization $V^* = \arg\max_V P(V|Q, D)$. In practice, this mapping is rarely straightforward. The core challenge lies in feature selection which identify the subset of dataset attributes $\mathcal{F}^* \subseteq \mathcal{F}$ that best expresses the user's analytical intent $I(Q)$. Earlier approaches mostly treated this as a similarity-matching problem. However, such approaches often failed to capture the probability distribution $P(\mathcal{F}|Q)$, which describes how relevant each feature is to a given query. As an illustration, traditional dependency models relied on Pearson's coefficient to approximate statistical relationships as indicated in (1).

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{1}$$

This measures only capture linear relationships and fail to detect complex, non-linear dependencies between query intent $I(Q)$ and features $\mathcal{F}$. Recent efforts [9]–[11] utilized machine learning approaches that learn non-linear relationships through neural networks. This formulation is expressed in (2).

$$P(\mathcal{F}^*|Q, D) = \text{softmax}(f_\theta(\mathbf{e_Q}, \mathbf{e_D})) \tag{2}$$

Where $f_\theta : \mathbb{R}^{d_Q + d_D} \rightarrow \mathbb{R}^{|\mathcal{F}|}$ is a neural network with parameters θ, $\mathbf{e_Q} \in \mathbb{R}^{d_Q}$ is the query embedding, and $\mathbf{e_D} \in \mathbb{R}^{d_D}$ is the dataset embedding. Due to the lack of data for training, especially understanding users' intention, this approach has been advanced by modern tools such as LIDA [12] or Vizagent [7], which employ large language models (LLMs) like GPT-4 to automate the visualization generation task. The primary limitation of utilizing LLMs is the ability to do sophisticated prompt engineering and consume extensive tokens, which consequently incurs substantial computational costs [13]. As such, it presents a significant barrier for researchers with constrained financial resources to iteratively conduct experiments [14]. Furthermore, LLMs offer more knowledge (trained on a vast amount of data on the internet) than needed in this problem, so the research question is "can we tackle the same issue with an affordable approach?". From the aforementioned pain points, there is a need for an alternative solution that could balance the learned capabilities of LLMs with computational efficiency and accessibility [15], [16]. The sparking idea is to leverage the state-of-the-art semantic understanding capabilities of pre-trained models while remaining lightweight and cost-effective of applications. This thought motivates us to develop methods that can capture complex query-feature dependencies through principled mathematical frameworks, without the overhead associated with large-scale language model deployment.

Thus, the current study proposed a unique information-theoretic approach for feature selection, particularly in NL2VIS systems. The proposed framework provided mathematically grounded principles that move beyond simple existing similarity measures. Building on prior surveys [3], [15], [16], we position NL2VIS as presented in Table 1 which reported in our experiments, and qualitative properties from prior work.

Table 1. Comparative positioning of NL2VIS approaches (taxonomy)

Criterion	Rule-based	Similarity-based	Neural ranking	LLM-based	MI-based (Ours)
Principle	Heuristic	Similarity	Learned similarity	Generative reasoning	Information-theoretic
Typical methods	Grammars / Rules	Cosine; TF-IDF + Corr	Contrastive ranking	GPT-4 prompting; LIDA; VizAgent	KDE + MINE
Interpretability	High	Medium	Low	Low–medium	High
Compute cost	Low	Low	Medium	High (token-dependent)	Medium (+31.7% time)
Accuracy	N/A (task-specific)	Val 82.3–85.7; Leg 74.1–78.9; F1 0.62–0.69; ICR 0.72–0.76	Val 87.8; Leg 84.1; F1 0.782; ICR 0.847	Val 93.4; Leg 89.7; F1 0.758; ICR 0.834	Val 88.9; Leg 85.2; F1 0.863; ICR 0.891
Notes	Transparent rules; brittle in open domains	Simple; struggles with non-linear intent	Learns non-linear patterns	Strong UX/aesthetics; higher cost	Principled, redundancy-aware selection

2. METHOD

2.1. Research design

To address the limitations identified in existing NL2VIS systems, we propose a unique application of mutual information (MI) theory to the NL2VIS domain. While MI is a well-established concept in information theory [17], [18], its systematic application in NL2VIS systems remains unexplored [15], [19]. For two discrete random variables X and Y, MI is expressed as (3).

$$I(X;Y) = \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x,y) \log \left(\frac{p(x,y)}{p(x)p(y)} \right) \tag{3}$$

Where $p(x,y)$ is the joint probability distribution of X and Y, and $p(x)$ and $p(y)$ are the marginal probability distributions. Alternatively, MI can also be expressed in terms of entropy as in (4).

$$I(X;Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) \tag{4}$$

Where $H(X) = -\sum_x p(x) \log p(x)$ is the Shannon entropy [20] of X, and the conditional entropy of X given Y is defined as (5).

$$H(X|Y) = -\sum_y p(y) \sum_x p(x|y) \log p(x|y) \tag{5}$$

Our framework consists of four main components: query intent extraction, feature representation, MI computation, and optimization-based feature selection. First, we transformed natural language queries (so called utterances) into higher dimensional spaces using pretrained language models. Mathematically, given a query $Q \in \mathcal{Q}$, we extracted its semantic representation as $\mathbf{v}_Q \in \mathbb{R}^d$ where d is the number of dimensions in the continuous embedding spaces. For feature representation, each feature in the database f_i is encoded as a multi-dimensional vector $\mathbf{v}_{f_i}$ and that vector contains semantic, statistical, and structural information. The semantic component utilizes some properties such as the feature name and metadata to create embeddings [21], [22], while the statistical component captures data distribution characteristics such as cardinality, skewness, and data type. The structural component encodes relational information, including primary/foreign key relationships and table hierarchies [6]. The ultimate purpose of the transformation is to let features interact with each other. Formally, we define as (6).

$$\mathbf{v}_{f_i} = [\mathbf{v}_{sem}(f_i); \mathbf{v}_{stat}(f_i); \mathbf{v}_{struct}(f_i)] \tag{6}$$

where $[;]$ represents vector concatenation. The core idea of our approach is to compute MI between continuous vector representations. Originally, the MI is defined for discrete variables, we employ kernel density estimation

(KDE) to estimate probability densities for continuous vectors [23], [24]. For vectors $\mathbf{v}_Q$ and $\mathbf{v}_{f_i}$, we estimate their joint density $\hat{p}(\mathbf{v}_Q, \mathbf{v}_{f_i})$ and marginal densities $\hat{p}(\mathbf{v}_Q)$ and $\hat{p}(\mathbf{v}_{f_i})$ using Gaussian kernels as (7).

$$\hat{p}(\mathbf{v}) = \frac{1}{n}\sum_{j=1}^{n} K_h(\mathbf{v} - \mathbf{v}_j) \tag{7}$$

Where K_h is a Gaussian kernel with bandwidth h, and n is number of samples. The MI estimate becomes (8).

$$\hat{I}(\mathbf{v}_q; \mathbf{v}_{f_i}) = \int \hat{p}(\mathbf{v}_q, \mathbf{v}_{f_i}) \log\left(\frac{\hat{p}(\mathbf{v}_q, \mathbf{v}_{f_i})}{\hat{p}(\mathbf{v}_q)\hat{p}(\mathbf{v}_{f_i})}\right) d\mathbf{v}_q d\mathbf{v}_{f_i} \tag{8}$$

To handle the computational complexity of high-dimensional MI estimation, we also investigate neural estimation approaches. We employ the mutual information neural estimation (MINE) framework [25], which uses neural networks to approximate the Kullback-Leibler divergence between the joint and product distributions. The MINE estimator is defined as (9).

$$\hat{I}_{MINE}(X;Y) = \sup_{\theta} \mathbb{E}_{p(x,y)}[T_\theta(x,y)] - \log \mathbb{E}_{p(x)p(y)}[e^{T_\theta(x,y)}] \tag{9}$$

Where T_θ is a neural network parameterized by θ, and the supremum is taken over all possible network parameters. Our feature selection optimization objective aims to identify the subset of features $\mathcal{F}^*$ that maximizes the total MI with the query intent while maintaining diversity among selected features. We formulate this as (10).

$$\mathcal{F}^* = \arg\max_{\mathcal{S}\subseteq\mathcal{F}, |\mathcal{S}|\leq k} \sum_{f_i\in\mathcal{S}} I(\mathbf{v}_q; \mathbf{v}_{f_i}) - \lambda \sum_{f_i, f_j\in\mathcal{S}, i\neq j} I(\mathbf{v}_{f_i}; \mathbf{v}_{f_j}) \tag{10}$$

Where k is the maximum number of features to select, and λ is a regularization parameter that penalizes redundancy between selected features. The first term encourages selection of features highly relevant to the query, while the second term promotes diversity by penalizing features that are highly correlated with each other. Since this optimization problem is NP-hard, we employ a greedy approximation algorithm that iteratively selects features based on their marginal contribution to the objective function. At each step, we compute the incremental gain of adding each remaining feature and select the one that maximizes as in (11).

$$\Delta(f_i|\mathcal{S}) = I(\mathbf{v}_q; \mathbf{v}_{f_i}) - \lambda \sum_{f_j\in\mathcal{S}} I(\mathbf{v}_{f_i}; \mathbf{v}_{f_j}) \tag{11}$$

Where $\mathcal{S}$ is the current selected features set. Once the good candidate features were identified, we proceed to the next stage of generating visualization. First, appropriate chart types and encoding assignments were determined. For this task, we considered it as a classification problem, where the input features contains the selected features and query intent representation. For the encoding assignment, we used a constraint satisfaction approach that ensures visual encoding principles are respected while maximizing the utilization of the information content provided by the selected features. The whole pipeline of our proposed approach was presented in Figure 2.

2.2. Evaluation

To assess the effectiveness of our proposed approach, we conducted comprehensive experiments with the VisEval benchmark dataset [26]. We also compared the current method against state-of-the-art NL2VIS systems. In the domain of visualization, there is a scarcity of dataset. Thus, Microsoft research curated VisEval as a comprehensive benchmark for NL2VIS. In general, this dataset provided standardized items across diverse domains such as business intelligence, healthcare, social media, and financial analytics. Here, each domain covers challenges in feature complexity, semantic interpretation, and visualization requirements. The core value of this benchmark is that each item was curated and annotated by domain experts with ground truth feature selections and optimal visualization specifications. Overall, it assesses three critical dimensions: validity - whether the generated code can run and render figures, legality - if the rendered figure meets query requirements, and readability - whether the visualization can convey information to users. The first two metrics were computed by the program while the latter metric was conducted with 12 experts (rating the charts using

Likert-scale of 5). This standardized and curated benchmark has been widely used recently to evaluate the performance of the newly developed NL2VIS approach.

In terms of performance, we compared our proposed approach with several baseline methods that have been reported in the niche field of NL2VIS so far. The first baseline used the cosine similarity method to estimate the direct correlation between query embeddings and feature embeddings, as implemented in systems like Data2Vis [10]. The second baseline utilized term frequency – inverse document frequency (TF-IDF) weighted keyword matching combined with statistical feature ranking. This method was based on Pearson correlation coefficients. The third baseline used LLM (in this case, we used the current latest GPT-4) with carefully designed prompts to perform feature selection. Here, we reproduced the experiment of the existing LIDA framework [12] but using more advanced LLM model. The fourth baseline implemented a neural ranking model trained on query-feature pairs using contrastive learning. To the best of our knowledge, all of these baselines represent recent advances for feature selection in NL2VIS.

As previously mentioned, our evaluation followed the VisEval framework in terms of validity, legality, and readability. In addition, we also employed accuracy, F1-score to evaluate the structural accuracy of feature selection. Formally, F1-score is expressed as (12).

$$\mathrm{F1} = \frac{2\,\mathrm{TP}}{2\,\mathrm{TP} + \mathrm{FP} + \mathrm{FN}} \tag{12}$$

However, this metric considered all features equally no matter how individual informational contributes to the query intent. To alleviate this issue, we introduced a new metric called information coverage ratio (ICR). ICR quantifies the information-theoretic quality of feature selection decisions based on the (13).

$$\mathrm{ICR} = \frac{\sum_{f_i \in \mathcal{F}^*} I(\mathbf{v}_q; \mathbf{v}_{f_i})}{\sum_{f_j \in \mathcal{F}_{gt}} I(\mathbf{v}_q; \mathbf{v}_{f_j})} \tag{13}$$

Where $\mathcal{F}^*$ represents the predicted feature subset and $\mathcal{F}_{gt}$ denotes the ground truth feature subset.

Conceptually, ICR quantifies how much of the total query-related information content in the optimal feature selection is preserved by the predicted selection. Compared to the F1 score, which only reflects binary correctness, the ICR score takes into account a continuous and information-weighted assessment. That captures the degree of analytical completeness.

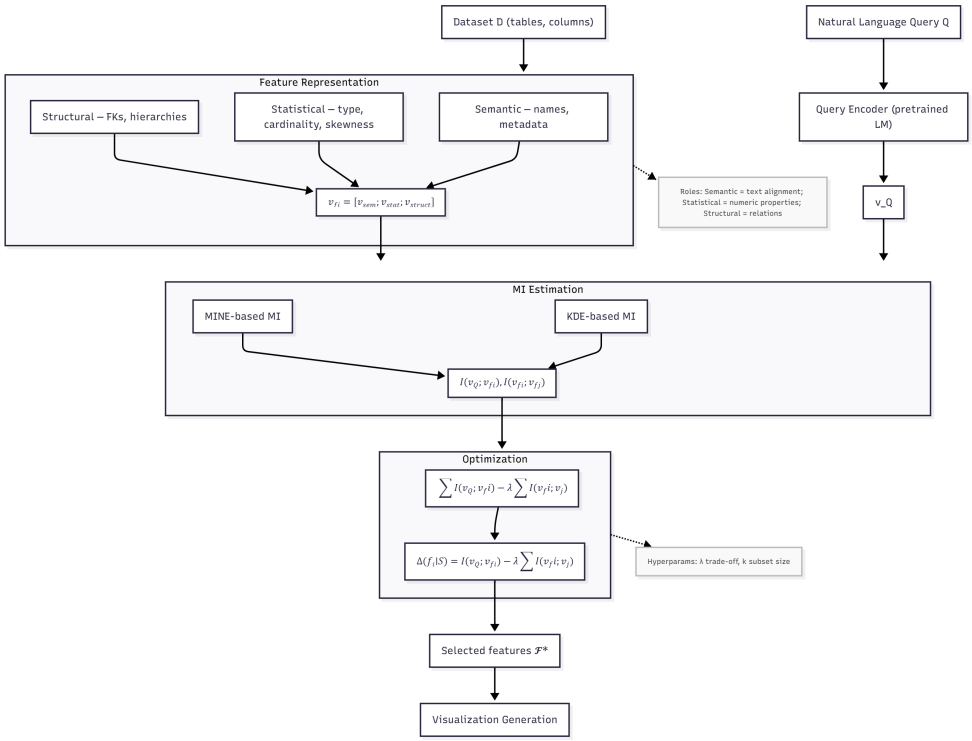

Figure 2. The pipeline of our proposed approach

3. RESULTS

Table 2 presents the snapshot of performance comparison across the two key VisEval evaluation dimensions, F1-score and ICR. Experimental results on the VisEval benchmark demonstrated the higher performance of our information-theoretic approach across all evaluation dimensions in terms of F1-score and ICR. Figure 3 depicts the approximate linear relationship between F1-score and ICR across all evaluated NL2VIS models. This near-linear trend conveys insight that while both metrics are aligned, ICR tends to yield higher values by weighting features due to their information contribution. This reinforce our assumption that the MI-based approach can capture not only structural correctness (as reflected by F1) but also the depth of analytical information preserved in the selected feature subsets.

Table 2. Performance comparison on VisEval benchmark

Method	Validity (%)	Legality (%)	F1 Score	ICR
Cosine similarity	82.3	74.1	0.623	0.721
TF-IDF + correlation	85.7	78.9	0.691	0.759
GPT-4 prompting	**93.4**	**89.7**	0.758	0.834
Neural ranking	87.8	84.1	0.782	0.847
Our method	88.9	85.2	**0.863**	**0.891**

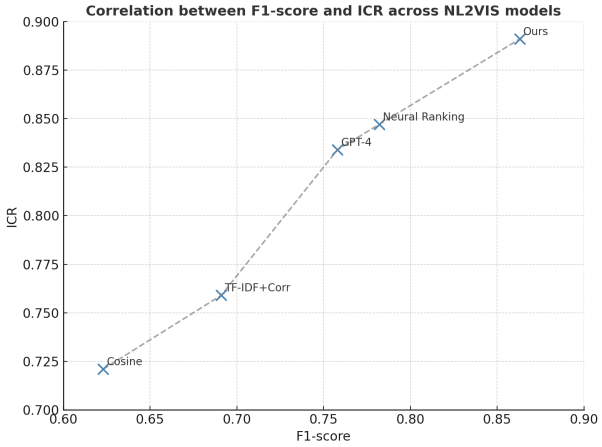

Figure 3. Correlation between F1-score and ICR across NL2VIS models

Back to Table 2, some interesting patterns were revealed between our information-theoretic approach and other methods. In terms of validity and legality, GPT-4 prompting achieved the highest performance compared to other methods with 93.4% and 89.7% respectively. This is not uncommon because this recent model was trained on the vast amount of data including code, thus, not surprisingly, demonstrates superior natural language understanding capabilities for interpreting user intent and producing code for generating visualizations. On the other hand, our method excelled in specialized information-theoretic measures: F1-score of 0.863 and ICR of 0.891. This performance pattern revealed a fundamental distinction: GPT-4 demonstrated superior semantic comprehension and visualization generation, but our approach provided more mathematically principled and statistically sound feature selection that ensures analytical correctness and explainable.

Figure 4 compared the performance of different methods with respect to readability and chart accuracy. In terms of chart accuracy (compared to ground truth), results revealed that GPT-4 prompting took the lead with 91.2% compared to our method of 87.6%. This indicated that when pretrained on mass data, it can capture the relationship between user intent and commonly used charts accordingly. However, our approach also demonstrated competitive performance while offering advantages in analytical solution and computational cost-effectiveness. Readability scores showed GPT-4 achieving 4.35/5.0 compared to our 4.18/5.0 which indicated that while advanced language models produce more visually intuitive visualizations, our approach maintains high user satisfaction levels while providing stronger guarantees of analytical correctness.

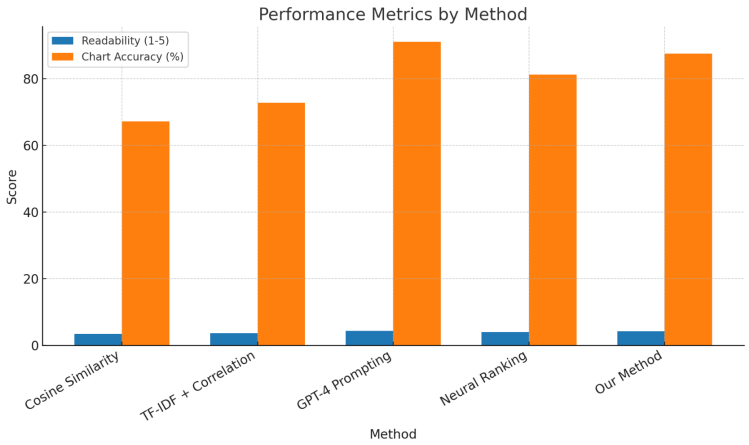

Figure 4. Comparison of chart accuracy and readability across NL2VIS methods

Table 3 reports encoding appropriateness, processing time (sec) and time overhead of our proposed method with the best baseline performed on our device. When normalized the score, the encoding appropriateness achieved 0.824. This implied that MI demonstrated superior technical quality in systematic feature-to-encoding mappings that ensure statistical validity. In terms of computational efficiency, result showed that our approach incurred a computational trade-off for improved feature selection quality. That is, the average processing time for feature selection ranges from 3.4 to 11.2 seconds per query compared to 2.1 to 8.9 seconds for the best baseline methods which resulted in a +31.7% time overhead. This computational cost reflects the mathematical complexity of MI estimation but delivers encoding appropriateness (0.824 vs 0.789). The KDE-based MI estimation accounted for approximately 60% of the total computation time, while our neural MINE estimation approach reduced this overhead by 35% with minimal accuracy trade-offs (average F1-score reduction of 0.023).

We also employed some statistical significance testings to quantify our deeper understanding of the metrics. Result from t-test showed that there is a statistical difference between F1-score and ICR metric ($p < 0.001$ for F1-score and ICR metrics). While traditional metrics show mixed results, with GPT-4 prompting leading in user experience measures, our information-theoretic metrics demonstrated substantial gains. Effect size analysis using Cohen's d revealed large effect sizes ($d > 0.8$) for F1-score and ICR comparisons. This indicated that improvements in feature selection quality are practically meaningful for real-world NL2VIS applications. These consistent performance gains in the benchmark scenarios demonstrated the robustness of our MI approach for analytical feature selection.

In addition, we also performed ablation studies which investigated the contribution of different components. When we removed the diversity regularization term (setting $\lambda = 0$ in equation 10) results in average F1-score decreases of 0.067. This implies the importance of promoting feature diversity. When we replaced the multi-modal feature representation with purely semantic embeddings, the performance was reduced by 0.089 F1-score. This suggests the value of incorporating statistical and structural information. When we used only KDE-based MI estimation without the neural MINE alternative. The computation time was increased by 43% without improving accuracy, and thus it supports our hybrid approach. Error analysis revealed that the most challenging cases for our method involved queries with highly ambiguous intent or domains with unconventional feature naming conventions. For instance, user utterance such as "show me interesting patterns in the data" lack specific analytical direction, making it difficult for any automated method to identify relevant features. Similarly, datasets with cryptic column names (e.g., "col_A", "var_123") that provided no semantic information pose challenges for the semantic component of our feature representation.

Table 3. Computational performance analysis

Metric	Our method	Best baseline
Encoding appropriateness	0.824	0.789
Processing time (sec)	3.4–11.2	2.1–8.9
Time overhead	+31.7%	

4. DISCUSSION

The comprehensive evaluation on the VisEval benchmark demonstrated that our proposed solution yielded a reasonable solution in addressing fundamental challenges in NL2VIS systems. Similar to prior research [3], [15], we found that analytical frameworks can outperform LLMs on domain-specific analytical correctness. The consistent performance improvements in specialized information-theoretic measures (F1-score of 0.863 and ICR of 0.891) suggested that MI provided a explainable mathematical foundation for quantifying the relationship between user intent and data features. The ICR complements precision–recall metrics by directly measuring information content preservation in feature selection [25].

In our experiment, GPT-4 stilled achieved higher scores in many facets (93.4% validity and 89.7% legality compared to our method's 88.9% and 85.2%, respectively). This is explainable because it was trained on a mass amount of data. In the previous GPT versions, they were mainly trained on text from internet and only covered a small portion of code, thus their performances were not expected. However they were still achieved higher score than conventional approaches. Recently, GPT-4 was trained more on code, thus in a recent benchmark for text-to-visualization, GPT-4 achieved the highest pass rate [26].

For casual users without prompting techniques, GPT-4 acts like a blackbox because the result is inconsistent – meaning that same query may give different charts. Therefore, this gap highlights a fundamental trade-off: while GPT-4 demonstrates good understanding of user intent and produces runnable code (4.35/5.0 readability and 91.2% chart type accuracy) [19], our approach emphasizes on explainable decision via mathematical rigor and analytical correctness in feature selection. This distinction represents an insight for the field — that is the choice between conversational fluency and statistical soundness which depends on the specific application requirements. Our approach is similar to the Kolmogorov-Arnold networks idea [27] where we sacrificed computation for explainable ones.

In a broader context, the current study contributes beyond the immediate application to NL2VIS systems. The framework for computing MI between real/float numbers posits a fundamental challenge in high-dimensional data [25], while maintaining the accuracy of MI estimation. In addition, the optimization function with a regularization parameter enables for a principled selection of the feature set, and thus, it ensures both query relevance and avoiding duplication of information between features. This intuition can be extended to many other feature selection problems [28]. Finally, we attempted to use as much information as possible in the given dataset to represent a feature (combining semantic, statistical, and structural information). This representation can also be useful for other domains that require a deeper understanding of the relationship between data structure and semantic meaning [29].

There are several limitations in the current work that should be acknowledged. First, we relied only a single LLM (GPT-4) to perform the experiment. This is due to computational cost incurred when using proprietary API. This implies that interested searchers could reproduce the work with different models. Second, our method utilized small pre-trained models such as bidirectional encoder representations from transformers (BERT) or robustly optimized BERT pretraining approach (RoBERTa), which sometimes do not capture domain-specific terms. Thus, unlike GPT-4, our approach was constrained by the semantic boundaries of the selected embedding models. Furthermore, while GPT-4 can handle ambiguous queries such as "show me something interesting" through creative interpretation that embedded in the model, our framework required more specific analytical direction to perform effective feature selection. Despite these limitations, we hope the MI framework remains useful beyond NL2VIS, particularly in explainable AI and biomedical analytics, where understanding feature relevance and redundancy is essential for transparent decision-making.

In another facet, our method performed efficiently on datasets with small numbers of features. However, the computation cost of MI estimation would grows rapidly at larger scales. As showed in the result section, the runtime increased by roughly 31.7% compared with baseline methods. Furthermore, the current framework is dedicated for structured data, thus, it is less flexible than GPT-4 when dealing with diverse data types or visualization settings. Even so, the MI-based selection mechanism would be promising beyond NL2VIS as it can help identify key sensor signals (continuous) for monitoring or fault detection. In addition, its lightweight computations would fit well with embedded or edge-level dashboards.

5. CONCLUSION

In summary, the current study introduced an MI framework for NL2VIS systems. The proposed method relied on MI to select features and improved analytical accuracy. Specifically, the model achieved an

F1-score of 0.863 and an ICR of 0.891. It also maintained visualization quality with 87.6% chart-type accuracy. Moreover, the approach emphasized mathematical rigor instead of conversational fluency. We defined a new metric to measure information coverage and optimize feature diversity. In addition, the current work can be extended to other machine-learning domains that required transparent feature selection. The computational cost remained a practical concern. Finally, future research planned to balance analytical precision with aesthetic quality through hybrid models that combined explainable AI and LLMs.

ACKNOWLEDGMENTS

This research was supported by the DH2025-TN07-05 project conducted at the Thai Nguyen University of Information and Communication Technology, Thai Nguyen, Vietnam, with additional support from the AI&SE Lab.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Hue Luong-Thi-Minh	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Vinh-The Nguyen	✓	✓			✓	✓		✓	✓	✓	✓	✓		✓
Van-Viet Nguyen				✓	✓	✓		✓	✓	✓				
Kim-Son Nguyen			✓	✓	✓	✓		✓	✓	✓				
Huu-Khanh Nguyen			✓	✓	✓	✓		✓	✓	✓				

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal Analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject Administration
Fu : **Fu**nding Acquisition

CONFLICT OF INTEREST STATEMENT

The authors have no financial, personal, or professional relationships that could inappropriately influence the research presented in this paper. The authors state no conflict of interest.

INFORMED CONSENT

We have obtained informed consent from all individuals included in this study.

ETHICAL APPROVAL

This research does not require ethical approval as it does not involve human participants, animal subjects, or sensitive data.

DATA AVAILABILITY

No new data were created or analyzed in this study. Results are based on the publicly available VisEval benchmark dataset. Implementation code is available upon reasonable request and will be released publicly after 24 months from publication, subject to project policies.

REFERENCES

[1] V. T. Nguyen, K. Jung, and V. Gupta, "Examining data visualization pitfalls in scientific publications," *Visual Computing for Industry, Biomedicine, and Art*, vol. 4, no. 1, Dec. 2021, doi: 10.1186/s42492-021-00092-y.

[2] S. Park, B. Bekemeier, A. Flaxman, and M. Schultz, "Impact of data visualization on decision-making and its implications for public health practice: a systematic literature review," *Informatics for Health and Social Care*, vol. 47, no. 2, pp. 175–193, Apr. 2022, doi: 10.1080/17538157.2021.1982949.

[3] A. Wu *et al*., "AI4VIS: Survey on artificial intelligence approaches for data visualization," *IEEE Transactions on Visualization and Computer Graphics*, vol. 28, no. 12, pp. 5049–5070, Dec. 2022, doi: 10.1109/TVCG.2021.3099002.

[4] T.-V. Nguyen and T.-N. Phung, "Enhanced literature review visualization: a novel sorted stream graphs with integrated word elements," in *Advances in Information and Communication Technology (ICTA 2024)*, Cham, Switzerland: Springer, 2024, pp. 159–168, doi: 10.1007/978-3-031-80943-9_17.

[5] E. Hoque and M. S. Islam, "Natural language generation for visualizations: state of the art, challenges and future directions," *Computer Graphics Forum*, vol. 44, no. 1, Feb. 2025, doi: 10.1111/cgf.15266.

[6] K. Zhou, Z. Liu, R. Chen, L. Li, S.-H. Choi, and X. Hu, "Table2Graph: transforming tabular data to unified weighted graph," in *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence*, 2022, pp. 2420–2426, doi: 10.24963/ijcai.2022/336.

[7] H. L. T. Minh, V. N. The, and T. Q. Xuan, "VizAgent: towards an intelligent and versatile data visualization framework powered by large language models," in *Advances in Information and Communication Technology (ICTA 2024)*, Cham, Switzerland: Springer, 2024, pp. 89–97, doi: 10.1007/978-3-031-80943-9_10.

[8] N. V. Viet *et al*., "Revolutionizing education: an extensive analysis of large language models integration," *International Research Journal of Science, Technology, Education, and Management*, vol. 4, no. 4, pp. 10-21, 2024, doi: 10.5281/zenodo.00000000.

[9] Y. Luo, X. Qin, N. Tang, and G. Li, "DeepEye: towards automatic data visualization," in *2018 IEEE 34th International Conference on Data Engineering (ICDE)*, Paris: IEEE, Apr. 2018, pp. 101–112, doi: 10.1109/ICDE.2018.00019.

[10] V. Dibia and C. Demiralp, "Data2Vis: Automatic generation of data visualizations using sequence-to-sequence recurrent neural networks," *IEEE Computer Graphics and Applications*, vol. 39, no. 5, pp. 33–46, Sep. 2019, doi: 10.1109/MCG.2019.2924636.

[11] R. Tabalba *et al*., "Articulate+: an always-listening natural language interface for creating data visualizations," in *Proceedings of the 4th Conference on Conversational User Interfaces*, 2022, pp. 1–6, doi: 10.1145/3543829.3544534.

[12] V. Dibia, "LIDA: a tool for automatic generation of grammar-agnostic visualizations and infographics using large language models," *arXiv:2303.02927*, 2023.

[13] G. Kusano, K. Akimoto, and K. Takeoka, "Revisiting prompt engineering: a comprehensive evaluation for LLM-based personalized recommendation," in *Proceedings of the Nineteenth ACM Conference on Recommender Systems*, Prague Czech Republic: ACM, Sep. 2025, pp. 832–841, doi: 10.1145/3705328.3748159.

[14] B. Chen, Z. Zhang, N. Langrené, and S. Zhu, "Unleashing the potential of prompt engineering for large language models," *Patterns*, vol. 6, no. 6, Jun. 2025, doi: 10.1016/j.patter.2025.101260.

[15] L. Shen *et al*., "Towards natural language interfaces for data visualization: a survey," *IEEE Transactions on Visualization and Computer Graphics*, vol. 29, no. 6, pp. 3121–3144, Jun. 2023, doi: 10.1109/TVCG.2022.3148007.

[16] W. Yang, M. Liu, Z. Wang, and S. Liu, "Foundation models meet visualizations: challenges and opportunities," *Computational Visual Media*, vol. 10, no. 3, pp. 399–424, Jun. 2024, doi: 10.1007/s41095-023-0393-x.

[17] S. Liu and M. Motani, "Improving mutual information based feature selection by boosting unique relevance," *Journal of Artificial Intelligence Research*, vol. 82, pp. 1267–1292, Mar. 2025, doi: 10.1613/jair.1.17219.

[18] J. Tang, Y. Luo, M. Ouzzani, G. Li, and H. Chen, "Sevi: speech-to-visualization through neural machine translation," in *Proceedings of the 2022 International Conference on Management of Data*, 2022, pp. 2353–2356, doi: 10.1145/3514221.3520150.

[19] P. Maddigan and T. Susnjak, "Chat2VIS: generating data visualizations via natural language using ChatGPT, Codex and GPT-3 large language models," *IEEE Access*, vol. 11, pp. 45181–45193, 2023, doi: 10.1109/ACCESS.2023.3274199.

[20] P. Saraiva, "On Shannon entropy and its applications," *Kuwait Journal of Science*, vol. 50, no. 3, pp. 194–199, Jul. 2023, doi: 10.1016/j.kjs.2023.05.004.

[21] N. M. Gardazi, A. Daud, M. K. Malik, A. Bukhari, T. Alsahfi, and B. Alshemaimri, "BERT applications in natural language processing: a review," *Artificial Intelligence Review*, vol. 58, no. 6, Mar. 2025, doi: 10.1007/s10462-025-11162-5.

[22] H. Man, N. T. Ngo, V. D. Lai, R. A. Rossi, F. Dernoncourt, and T. H. Nguyen, "LUSIFER: language universal space integration for enhanced representation in multilingual text embedding models," in *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, Padua, Italy: ACM, Jul. 2025, pp. 1360–1370, doi: 10.1145/3726302.3730029.

[23] Y. Ning *et al*., "A mutual information theory-based approach for assessing uncertainties in deterministic multi-category precipitation forecasts," *Water Resources Research*, vol. 58, no. 11, Nov. 2022, doi: 10.1029/2022WR032631.

[24] A. Moreo, P. González, and J. J. D. Coz, "Kernel density estimation for multiclass quantification," *Machine Learning*, vol. 114, no. 4, Apr. 2025, doi: 10.1007/s10994-024-06726-5.

[25] M. I. Belghazi *et al*., "Mutual information neural estimation," in *Proceedings of the 35th International Conference on Machine Learning*, PMLR, Jul. 2018, pp. 531–540.

[26] N. Chen, Y. Zhang, J. Xu, K. Ren, and Y. Yang, "VisEval: a benchmark for data visualization in the era of large language models," *IEEE Transactions on Visualization and Computer Graphics*, vol. 31, no. 1, pp. 1301–1311, Jan. 2025, doi: 10.1109/TVCG.2024.3456320.

[27] S. Somvanshi, S. A. Javed, M. M. Islam, D. Pandit, and S. Das, "A survey on Kolmogorov-Arnold network," *ACM Computing Surveys*, vol. 58, no. 2, pp. 1–35, Jan. 2026, doi: 10.1145/3743128.

[28] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 27, no. 8, pp. 1226–1238, Aug. 2005, doi: 10.1109/TPAMI.2005.159.

[29] K. Hu, M. A. Bakker, S. Li, T. Kraska, and C. Hidalgo, "Vizml: a machine learning approach to visualization recommendation," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, 2019, pp. 1–12, doi: 10.1145/3290605.3300358.

BIOGRAPHIES OF AUTHORS

Hue Luong-Thi-Minh received her Bachelor of Information Technology from Thai Nguyen University of Information Technology in 2010 and her Master of Information Technology from Manuel S. Enverga University, Philippines, in 2013. She has been a lecturer at the Faculty of Information Technology, Thai Nguyen University of Information Technology since 2010. Her research interests include computer science, artificial intelligence, and communication technology. For this research, she contributed to conceptualization, investigation, validation, and supervision of the project. She has experience in managing interdisciplinary research projects combining AI and geospatial technologies. She can be contacted at email: lmhue@ictu.edu.vn.

Vinh-The Nguyen is currently a senior lecturer at the Faculty of Information Technology, Thai Nguyen University of Information and Communication Technology. He graduated with a master's degree in Information Systems Management from Oklahoma State University, USA (under scholarship 322). He completed his Ph.D. program under project 911 in 2020 at Texas Tech University, USA. His main research interests are computer vision, computer visualization, and computer in human behavior. He has authored or coauthored more than 50 publications with 16 H-index and more than 900 citations. He can be contacted at email: vinhnt@ictu.edu.vn.

Van-Viet Nguyen is a researcher and Ph.D. student at the Thai Nguyen University of Information and Communication Technology, Thai Nguyen, Vietnam. He received a bachelor's at Thai Nguyen University of Information and Communication Technology, Vietnam in 2009. He got a master's degree on Information Technology at Manuel S. Enverga University, Philippines in 2012. He research interests include artificial intelligence, machine learning, and generative AI. He can be contacted at email: nvviet@ictu.edu.vn.

Kim-Son Nguyen received her Bachelor of Information Technology from Thai Nguyen University of Information and Communication Technology in 2009 and her Master of Information Technology from Manuel S. Enverga University, Philippines, in 2012. Currently, she is a lecturer at the Thai Nguyen University of Information and Communication Technology, Vietnam. Her main research interests are artificial intelligence, large language models, and geographic information systems. In the current study, she contributed to conceptualization, methodology, software development, and original draft preparation. She is currently leading a research project on the integration of open-source language models with GIS applications for resource-constrained environments. She can be contacted at email: nkson@ictu.edu.vn.

Huu-Khanh Nguyen has graduated with a master's degree in Computer Science from the Thai Nguyen University of Information and Communication Technology since 2022 and is currently a Ph.D. student here since 2023. His main research interests are computer science, natural language processing, generative AI, and computer vision. He can be contacted at email: khanhnh@tnu.edu.vn.