AutoTitle: An Interactive Title Generator for Visualizations

Can Liu, Yuhan Guo, Xiaoru Yuan

Abstract—We propose AutoTitle, an interactive visualization title generator satisfying multifarious user requirements. Factors making a good title, namely, the feature importance, coverage, preciseness, general information richness, conciseness, and non-technicality, are summarized based on the feedback from user interviews. Visualization authors need to trade off among these factors to fit specific scenarios, resulting in a wide design space of visualization titles. AutoTitle generates various titles through the process of visualization facts traversing, deep learning-based fact-to-title generation, and quantitative evaluation of the six factors. AutoTitle also provides users with an interactive interface to explore the desired titles by filtering the metrics. We conduct a user study to validate the quality of generated titles as well as the rationality and helpfulness of these metrics.

Index Terms—Visualization title, natural language, deep learning, large language model

1 Introduction

Title is an essential component of visualization, which helps authors convey information and readers comprehend visualization contents. Visualization titles speed up the acquisition of information [1] by illuminating data features and play a critical role in helping participants recall the content of the visualization [2]. An inappropriate title would mislead readers, as it can easily influence their understanding of the visualization. Kong et al. [3] showed that led by different titles, users could have opposite recognition for the same visualization content. However, creating a proper title is a non-trivial task, as different scenarios require different titles. The length, readers' literacy, and importance of titles are often considered, but can not be easily satisfied at the same time. Different usage scenarios of visualizations also influence the requirements of titles. For example, the visualizations designed for the general public pursue titles with low technicality, while titles for academics should be more accurate.

In recent years, many approaches focused on automating visualization processes, such as recommending visualizations [4], describing visualizations [5], and answering questions on visualizations [6], [7]. These processes can make data and visualizations more accessible to a wider range of users. However, the titles, serving as a crucial component of visualization, have not received sufficient research focus in the context of automated visualization processes. Generating effective visualization titles can be a challenging task, given the vast space of potential titles. While some titles may require only a few words to describe the visualization, others may involve technical terminology or specific data features, depending on the intended usage scenarios. Even for experts, it can be difficult to identify the key components that make a visualization title successful. As illustrated in Figure 1, various title choices may be suitable for different scenarios. We propose a generative approach to modeling the design space of titles and automatically generating effective visualization titles.

To investigate the design space of titles, we conducted a formative user study with 54 participants to figure out the factors that contribute to good visualization titles. In the study, participants were asked to write good titles for several visualizations and were interviewed about the characteristics of effective visualization titles. Based on their feedback, we identified six factors for good visualization titles: feature importance, content coverage, general information richness, preciseness, conciseness, and non-technicality. According to the formative study, most visualization experts tend to prefer titles with data features, while participants with news backgrounds mentioned that if targeting the general public, the use of accurate numbers in the titles should be avoided. This demonstrates that different audiences have distinct expectations and preferences for particular factors. It is difficult for a single title to excel in all factors. Prioritizing one aspect, such as brevity, may result in sacrificing coverage. However, finding the right balance between these factors can be challenging since the design space for generating visualization titles is vast.While individual creators may find it challenging to comprehensively consider all relevant factors, machine-generated judgments with quantifying metrics can provide a helpful and reliable perspective. Therefore, using a title generator can help represent the space of titles and enable creators to gain a more in-depth understanding of the title space. Ultimately, this understanding will allow them to select the most appropriate title for their visualizations.

AutoTitle generates a variety of titles for different requirements. The system comprises four modules: fact extraction, title generation from facts, metrics for quantifying titles, and an interactive system. The system extracts underlying data from an input visualization and traverses the multi-level facts of the visualization to extract and organize the hierarchical facts associated with it. In the fact-to-title generation process, we employ a large-scale natural language transformer as the base model, which is fine-tuned to generate fluent and diverse natural language titles from hierarchical facts. Additionally, we introduce metrics to quantify the quality of the generated titles for different factors, such as

- *Can Liu, Yuhan Guo, and Xiaoru Yuan are with Key Laboratory of Machine Perception (Ministry of Education), School of Intelligence Science and Technology, Peking University. E-mail: {can.liu, yuhan.guo, xiaoru.yuan}@pku.edu.cn.*
- *Xiaoru Yuan is also with National Engineering Laboratory for Big Data Analysis and Application, Peking University.*
- *Xiaoru Yuan is the corresponding author.*

Manuscript received November xx, 2022; revised November xx, 22082022.

Fig. 1: When creating titles for visualization, individuals are faced with a variety of choices that are suitable for different scenarios. Some titles convey detailed information, while some are concise.

coverage and importance. The interactive system allows users to explore the design space of titles and select those that meet their requirements based on the metrics. We conducted a user study to evaluate the effectiveness of the AutoTitle system and found that it is capable of generating high-quality titles, and the metrics are both rational and useful.

The contributions of this work are summarized as follows:

1) We identi ed six important factors for evaluating visualization titles, and proposed quantitative metrics for each factor.
2) We developed an interactive title generator, AutoTitle, that not only generates visualization titles but also visualizes the title design space. This allows users to easily understand the metrics and make quantitative assessments of trade-offs. With this user-friendly interface, users can choose a suitable visualization title based on their preferences.

2 Related Work

Our approach relates visualization titles and controlled natural language generation. The quantifying metrics for the titles also relate to the quanti cation of semantic information.

2.1 Visualization Titles

There are different textual components in visualization, including titles, descriptions, and captions [5]. A title is typically a concise and attention-grabbing phrase or sentence that summarizes the visualization's content. It is commonly positioned above the visualization. The title helps users comprehend [1] and recall [8] the information presented in the visualization. Visualizations with titles are easier for users to understand [8] and require less mental effort to process [9], enabling faster acquisition of information [1]. Borkin et al. [2] found that people heavily rely on textual information such as titles, which signi cantly in uences their recognition of the content in the visualization. Visualizations with titles can lead users to believe the content they convey [8]. However, when the title creator incorporates subjective tendencies, bias may arise. Different slanted titles can make users recall even opposite messages for the same visualization [3]. Furthermore, Kong et al. [10] studied the in uence of different degrees of misalignment between the title and the visualization, and found that although contradictory titles can make users more likely to identify the bias between the two, they still tend to believe the content of the titles.

Despite the importance of titles for visualizations, many prevalent visualizations lack proper titles due to the complexity of the title space and the dif culty in natural language generation. To address this gap, we explored the design space of visualization titles and proposed a title generator for visualization.

2.2 Controlled Natural Language Generation

Natural language generation [11] is an emerging topic in deep learning elds. Natural language generation tasks are much more complex and uncontrollable than image processing due to the complex semantic structure of natural language. Traditional natural language tasks, namely machine translation [12] and image captioning [13], use supervised learning methods, which only learn the corresponding relationships between pairs of input and output from two domains by learning from a large corpus. Bowman et al. [14] aims to generate natural language from a continuous space using variational autoencoders (VAE) [15]. Further, Hu et al. [16] proposed the concept of controlled natural language generation, where users can manipulate the attributes, e.g., sentiment, of the generated text. Hu et al. [16] can generate more controllable results than Bowman et al.

In our task, we generate natural language based on the facts extracted from the visualization. To control the meaning of the titles, we focused on controlled natural language generation from structured data [17]. There are many applications for converting structured information to natural language, including the generation of weather reports [18] and sports reports [19]. Traditional natural language generation for structured data is mainly based on template-based methods [20]. Later, end-to-end deep-learning techniques [19], [21] were proposed. Recently, large-scale transformers [22], [23], [24], [25], [26] have shown the ability to transform to various tasks by ne-tuning. For example, T5 [26], short for text-to-text translation transformer, is able to handle various forms of translation tasks. Chen et al. [27] demonstrated that the pre-trained transformers could support the natural language generation from structured data [28]. In our method, we construct a fact-to-title dataset and ne-tune the T5 [26] to support generating titles from facts.

2.3 Semantic Information

Measuring visualization titles is non-trivial as there is no clear de nition for the amount of information in a title. We surveyed the papers that discussed the measuring of semantic information.

Bar-Hillel and Carnap [29] rst pointed out that the traditional information theory [30], which treats the amount of information as a measure of the statistical rarity of a message, may not be suitable for semantic information scenarios. Evidence is that a contradictory statement is very informative in traditional information theory. Based on Bar-Hillel and Carnap's theory [29], the classic semantic information theory (CSI) is developed, where the information amount can be calculated using the set of possible states mentioned in a language. Floridi [31] proposed the theory of strong semantic information (TSSI), whose core idea is to measure the matching degree between a statement and the truth. CSI and TSSI differ in that when a statement is always true, the information amount is large in CSI, while the information amount is zero in TSSI due to no information gained. Based on CSI [29] and TSSI [31], D'Alfonso [32] quanti ed the information using atomic facts and showed several cases. However, previous works were on a conceptual level, which can not be applied to a real-world scenario. Montemurro and Zanette [33] proposed the quantifying method for the written natural language.

In this work, we propose a quantifying method for measuring the generated titles with multiple quantifying metrics. The method is based on atomic facts, inspired by semantic information works [29], [31], [32]. In the visualization title scenario, the design

space of visualization titles is multi-dimensional. Therefore, measuring the sentence with a single value, as previous works [29], [31], [32] did, is not enough. We measure the visualization titles using feature importance, preciseness, general information richness, and coverage.

3 Design Space of Visualization Titles

In this section, we discuss the design space of the visualization titles. First, we summarize the taxonomy of visualization titles through collected visualizations. We also conduct interviews with participants from different areas to identify the factors that make for a good visualization title.

3.1 The taxonomy of Visualization Titles

Visualization titles can be classi ed into two types: *generic titles*, which only contain generic information, and *informative titles*, which include data features [9]. Informative titles highlight the essential content of the visualization, improving its accessibility. We collected approximately 100 visualizations with titles from news websites and academic papers. Table 1 shows the design space of visualization titles. The components of titles can be classi ed into generic information and data features.

Generic information. A title typically presents generic information about the visualization, including data attributes, visualization types, task types, and feature types. Data attributes consist of names, granularity, and range. Some titles contain information on data attributes, such as “New cases in New York” with the quantitative attribute “new cases”. Other titles present the visualization types, tasks, and feature types, e.g., “the line chart of European countries' GDP” and “the trend of stock price”.

Data feature. A title may emphasize data features in the visualization, such as trend, comparison, aggregation, and proportion. For example, “Young people are drinking less” depicts the decreasing trend of young people's drinking habits.

TABLE 1: Design space of visualizations titles.

Dimension	Sub-dimension	Choice	Example
Generic Info	Data attributes	Name, range, granularity	Deaths by day.
	Visual encoding	Visualization type, mapping, feature type	Line chart of European countries' GDP.
	Visual task	Compare, distribute	COVID-19 Deaths Per 100,000 Inhabitants: A Comparison.
Data Feature	Trend	Trend type, degree, times, ratio, change amount	Young people are drinking less.
	Aggregate	Max, min, average, sum, range	Covid-19: Countries in Europe with the most deaths.
	Combine	Combine value, combine trend, etc.	Population gains among Asian, Latino, and multiracial children offset losses among white children.
	Compare	Compare-value, compare-trend, etc.	NHS has fewer staff than some counterparts.

3.2 Factors Making a Good Visualization Title

We conducted a formative study to identify the factors that contribute to a good visualization title. The users were recruited through questionnaires on university forums and other social media platforms, targeting students from various disciplines and professionals from diverse industries. The study began with collecting background information, including participants' professional backgrounds and visualization expertise. Next, participants were asked to write what they believed to be good titles for several visualizations, including two bar charts (one for categorical data and one for temporal data) and two line charts. Finally, participants were asked to identify the factors they consider important for a good visualization title. We received feedback from 54 participants with diverse backgrounds in elds such as computer science, engineering, law, journalism, and natural science. Of these participants, 11 self-identi ed as data analysis experts, 14 had experience using programming languages such as Python and R for data visualization, 17 had experience creating charts using software such as Excel, 11 had knowledge of visualization, and only 1 had never encountered visualization before. In total, we collected 204 titles, of which 42.16% were generic and 57.84% contained data features. The average length of these titles was 10.0 words, with titles written by experts having an average length of 11.27 words. We aimed to understand which factors would be frequently mentioned in the absence of any xed options, particularly by those with higher levels of expertise. Our formative study revealed participants' preference for visualization titles that embody three fundamental attributes: informativeness, non-technicality, and conciseness. While non-technicality and conciseness share similarities in their de nitions, informativeness exhibits multiple interpretations. Based on the participants' feedback, we categorized informativeness into feature importance, coverage, generic information richness, and preciseness. Ultimately, we identi ed six critical factors that are vital for effective visualization titles. These factors are listed by the frequency of their mention.

Feature Importance: More than half (28/54) of the participants mentioned that the visualization title should convey key features. This proportion was higher among 11 visualization experts, where almost all (10/11) emphasized the importance of conveying key features. In addition, participants with less visualization experience also mentioned the importance of the content of the visualization, which may refer to the key features. The importance of features is in uenced by the signi cance of fact types and the signi cance of individual facts. The signi cance of fact type relates to the type of visualization used, for instance, trends are more important in a line chart. The salience of fact is also important, such as sudden changes in trends.

Conciseness: Nearly half of the participants (24) emphasized the importance of concision in a title. A concise title allows users to quickly understand the content of the visualization.

Preciseness: The importance of accuracy and correctness in titles was emphasized by many participants. Accuracy can be understood as a combination of correctness and preciseness, with the former referring to the title's adherence to the real data and the latter to its degree of closeness to the truth. As a title should always be correct, preciseness is a useful measure of how closely and accurately a title represents the real data.

Generic Information Richness: Nine participants mentioned the importance of including generic information such as data attributes and data range in the visualization title. These content types fall under the category of generic information content as discussed in subsection 3.1.

Coverage: 10 participants stressed the importance of including comprehensive information in visualization titles, without omitting any details. For example, for a multi-line chart, the title should cover the entire time range and summarize different categories. Additionally, recent studies [3], [10] have demonstrated that missing information can lead to bias.

Fig. 2: The pipeline of the generation process. The underlying data are extracted from the given chart, and the atomic facts are traversed. The multi-level facts are composed according to the basic facts and computation methods. These facts are converted to natural language titles by the language generator. The quanti ed metrics are calculated for the generated titles.

Non-Technicality: Non-technicality refers to the degree of specialized knowledge or background required to understand a visualization's title, including the use of technical terms and numbers. In our formative study, a participant majoring in journalism suggested avoiding technical terms and numbers when targeting a general audience. However, some participants also showed a preference for precise numbers, which may be more technical. Therefore, the technical or non-technical nature of a title may vary depending on the target audience and the purpose of the title. The number of words and the complexity of words is the most important factors for the non-technicality of visualization titles.

The space for exploration in visualization titles is vast, and trade-offs are necessary because it is dif cult for a title to possess all desirable factors. For instance, when preciseness is prioritized, a precise number may be included in the title, which can decrease its non-technicality. To illustrate this point, consider the following two sentences that describe a change in a company's market share: (A) “The company's market share has grown from 12.5% in 2008 to 26.7% in 2020, with an increase of 1.136 times,” and (B) “The company doubled its market share.” While sentence (A) is more precise, it has higher technicality (lower non-technicality) compared to sentence (B). Additionally, it can be dif cult to achieve both conciseness and high coverage and preciseness in a title. Some participants in our study felt that the title should focus on a speci c point, while others felt that no information should be omitted. Overall, the emphasis on different factors in a visualization title may vary depending on the user's requirements.

4 Title Generation for Visualization

We propose an interactive system, AutoTitle, for generating titles. The system enables users to explore the space of possible titles and select their desired titles interactively to satisfy their requirements on factors such as the importance of features, coverage, preciseness, conciseness, and non-technicality. The system consists of four modules: fact extraction, title generation, quanti cation of factors, and an interactive interface. Figure 2 illustrates how the titles are generated from the input chart and displays the corresponding metrics.

4.1 Extracting Facts from Visualization

Given a chart as input, the rst step is to extract the underlying data using reverse-engineering methods. The atomic facts are simple descriptions of the value of each data item in the underlying data. Based on these atomic facts, higher-level facts can be calculated using fact calculation operations. The task of generating facts has been widely discussed in several previous works [34], [35], [36], [37]. DataShot [35] and Calliope [36] provide an identi cation of a breakdown space, within which various derived values and facts are generated. Our fact-generation method also subdivides the space and generates several different derived values. The main difference lies in the emphasis on nested computations in a bottom-up manner, with the output of one level serving as the input for the next level. For example, while previous works generate a single trend for the whole subset of the dataset, our method provides trends with different levels of precision through different combinations of sub-trends.

Reverse-engineering takes a visualization as input and extracts the underlying data, which has been well-studied and discussed in previous works (e.g., Poco et al. [38] and ReVision [39]). In this work, while the reverse-engineering part is not our main contribution, we begin with the SVG format visualization to extract the underlying data. To accomplish this, we adopt the methods outlined by Poco et al. [38], which involve classifying the text role (e.g., axis ticks, legends) based on text position. Axes are classi ed into temporal, categorical, or quantitative types based on the text content. The attributes of visual elements can then be extracted based on the axes and the text. For the bitmap-format chart, ReVision [39] can extract the underlying data.

Atomic facts are simple descriptions of individual data items. Given the underlying data, we can traverse each data item to extract all atomic facts. A fact consists of two parts: the reference name and the content. The subject of the natural language sentence is the reference name, and the rest of the sentence is the content. For example, in the sentence “India's population in 2010 was over one billion,” the reference name is “India's population in 2010,” and the content is “was over one billion.” The reference name can take different forms, but it typically includes a quantitative attribute measuring a data item, such as area, price, or GDP. The name of a categorical attribute (e.g., India's) or a temporal attribute (e.g., in 2010) may also be included in the reference name. An

atomic fact is denoted as a *value fact*, and its basic structure is "[reference name] is [value]."

Fig. 3: Multi-level facts structure of visualization. High-level facts are composed of atomic facts. The calculation operation includes aggregating, computing trends, combining, and so on.

Computing high-level facts. We constructed a multi-level fact structure, as shown in Figure 3 based on the atomic facts. The higher-level facts are computed using several operations based on lower-level facts, including aggregation, comparison, trend computation, combination, and merging. A formal de nition of high-level facts calculation is presented in Table 2, which uses data attributes S_i to represent the input and output of an operation. Outputs are generated through various operations such as aggregation, trend computation, comparison, combination, and merge. We enable the representation of output facts in a value format with derived dimensions and values that can be used as input for next-level operations.

Aggregate operation computes the aggregation for a list of value-type facts with only one different attribute. Common types of aggregation include maximum, minimum, average, and summation. The structure of an aggregated fact is *aggregate* <obj, aggregate type, [range], value> .

Compare operation is a computation operation of two counterparts of data facts. Two data facts should have the same measure but with different categorical attributes or different temporal ranges. The sign of the comparison is larger, smaller, and similar. The degree of comparison is expressed in various ways, including ratios, percentages of differences, times, and differences amounts. The comparison result for data facts is *compare* <ref1, ref2, sign, [degree]> .

Trend operation takes a list of facts with continuous temporal attributes as input. A trend fact has a trend type and degree. The trend type can be "increase", "decrease", or "stay stable". The degree of the trend is described using adverbs (e.g., quickly, slowly, signi cantly), percentages (e.g., by 15%), multiples (by 3-folds), and change values (by 20 dollars). The trend fact is *trend* <ref, range, trend type, [degree]> .

Merge accepts a list of facts with different reference names and the same content, which produces a new fact by merging the reference name. For example, the merging result of "China's population increased" and "India's population increased" is "China and India's population increased." A merge is *merge* <merged obj, content> .

Combination operation accepts a list of facts with different contents and ranges. Combining two or more trends with the same reference name and different ranges can produce a complex trend fact. For example, the fact about stock price increasing rst and then decreasing is the output of combining two simple facts. A combination is *combine* <ref, [content 1, content 2]> .

Overview accepts all atomic facts as input and generates the general facts corresponding to generic titles. General facts only involve overall information, e.g., attribute names and feature type. A general fact is *overview* <attribute name, [feature type], [range]> .

The input facts of an operation are denoted as sub-facts of the output fact. The output fact can be used as the input for the next level, which builds a multi-level fact structure. For example, we can compare the difference in trends such as "India's population has increased faster than China's in the last decade," or calculate the trend of the summation of all Asian countries with the statement "The total population of Asian countries is on the rise." However, as the number of atomic facts and computation levels increases, the calculation of facts will increase exponentially. To address this, we have implemented methods to limit the size of the facts, such as setting maximum levels and avoiding unimportant calculations.

Set maximum calculation levels. It is rare for a natural language sentence to present with more than three calculation levels. Therefore, the fact calculation stops at the third level.

Avoid unimportant calculation. Calculating all possible combinations of sub-facts to determine all higher-level facts is almost impossible. Therefore, we employ a heuristic algorithm to avoid the computation of unimportant facts to limit the size. For instance, to identify trend features, we rst apply smoothing techniques to the data and then calculate the change points with the most signi cant slope changes. We use these points to segment the trends, and by combining these sub-trends, we can describe the trends at various levels. For example, we can describe the trend as a whole or divide it into two or three segments. In the comparison operation, we focus only on the most crucial elements, such as items with the maximum or minimum values within a subset or breakdown.

4.2 Fact-to-Title Generation

The goal of the fact-to-title process is to generate uent natural language from a given fact. The fact-to-title generation task is a structure-to-text problem [17]. Traditional methods for achieving this use a template-based approach, where the fact is directly fed into a sentence template to produce semantically correct but potentially repetitive and rigid sentences. Inspired by the success of deep learning models in natural language translation [11], [40], we propose using a deep learning-based approach to generate diverse natural language from data facts. Deep learning methods typically require a large amount of training data. Recent advances in pre-trained models [22] have shown that they can perform well on various tasks with minimal ne-tuning using small datasets [41]. In this work, we construct a fact-to-title dataset and ne-tune an large language model for the fact-to-title generation task.

Requirements. It is necessary for the training dataset to have diversity in data attributes, fact types, and natural language in order for the model to generate uent natural language from facts in various domains.

R1. Diversity in data attributes. Data attributes from various domains have different expressions when involving human idioms. For example, the word presents the meaning of "greater" for the measure "area" is "larger", while it is "higher" for "temperature". To enable the trained model to generate natural sentences that are suitable for various scenarios, the dataset should contain data attributes from a wide range of domains.

R2. Diversity in fact types. Data facts should cover a wide range of atomic facts and their high-level combinations. We extracted templates with various expressions from collected natural language for each fact type within three levels.

TABLE 2: Operations for calculating facts. The input facts have two key attributes, namely, D_1 and D_2. D_1 represents a dimension that remains constant and can represent a combination of multiple dimensions. D_2, on the other hand, is a dimension that varies across multiple input facts. D_3 represents the value dimension. The dimensions can be categorical (C), temporal (T), or quantitative (Q). For instance, the trend operation accepts a set of facts with different time values and generates a new fact that represents the change in these facts over time. Several trends can be combined or compared to create a higher-level fact.

Operation	Input Fact	Condition	Output Fact	Value format	Example
Aggregate	f value, {$D_1 = v_1$, $D_2 = v_{21}$}, $D_3 = v_{31}$g f value, {$D_1 = v_1$, $D_2 = v_{22}$}, $D_3 = v_{32}$g . . . , f value, {$D_1 = v_1$, $D_2 = v_{2n}$}, $D_3 = v_{3n}$g		f aggregate, {$D_1 = v_1$, $D_2 = \text{range}(v_{21}, v_{2n})$}, agg_type = agg_value[, degree]g	f value, {$D_1 = v_1$, $D_2 = \text{range}(v_{21}, v_{2n})$}, "agg_type of D_3" = agg_valueg	The Largest Mammal: Blue Whale.
Trend	f value, {$D_1 = v_1$, $D_2 = v_{21}$}, $D_3 = v_{31}$g f value, {$D_1 = v_1$, $D_2 = v_{22}$}, $D_3 = v_{32}$g . . . , f value, {$D_1 = v_1$, $D_2 = v_{2n}$}, $D_3 = v_{3n}$g	type(D_2) = T $v_{2i} < v_{2j}$ if $i < j$ type(D_3) = Q	f trend, {$D_1 = v_1$, $D_2 = \text{range}(v_{21}, v_{2n})$}, sign = sign_value[, degree_type = degree_value]g	f value, {$D_1 = v_1$, $D_2 = \text{range}(v_{21}, v_{2n})$}, "trend of D_3" = sign_valueg f value, {$D_1 = v_1$, $D_2 = \text{range}(v_{21}, v_{2n})$}, "sign_value degree_type of D_3" = degree_valueg	Global Food Production Keeps Increasing for Nearly a Century.
Compare	f value, {$D_1 = v_1$, $D_2 = v_{21}$}, $D_3 = v_{31}$g f value, {$D_1 = v_1$, $D_2 = v_{22}$}, $D_3 = v_{32}$g	type(D3) = Q	f compare, {$D_1 = v_1$, $D_2 = \{v_{21}, v_{22}\}$}, sign = sign_value[, degree_type = degree_value]g	f value, {$D_1 = v_1$, $D_2 = \text{range}(v_{21}, v_{22})$}, "compare of D_3" = sign_valueg f value, {$D_1 = v_1$, $D_2 = \text{range}(v_{21}, v_{22})$}, "degree_type of D_3" = degree_valueg	A Blue Whale Weighs Over 200 Times More Than an Elephant.
		$v_{31} \neq v_{32}$ type(D3) = C	f compare, {$D_1 = v_1$, $[(D_2 = v_{21}, D_3 = v_{31}), (D_2 = v_{21}, D_3 = v_{32})]$}g	-	Avocado Sales Continue to Rise While Grapefruit Sales Decline Steadily.
Merge	f value, {$D_1 = v_1$, $D_2 = v_{21}$}, $D_3 = v_3$g f value, {$D_1 = v_1$, $D_2 = v_{22}$}, $D_3 = v_3$g		f merge, {$D_1 = v_1$, $D_2 = \{v_{21}, v_{22}\}$}, $D_3 = v_3$g	f value, {$D_1 = v_1$, $D_2 = \{v_{21}, v_{22}\}$}, $D_3 = v_3$g	Prices of Avocados and Blueberries Soar.
Combine	f value, {$D_1 = v_1$, $D_2 = v_{21}$}, $D_3 = v_{31}$g f value, {$D_1 = v_1$, $D_2 = v_{22}$}, $D_3 = v_{32}$g		f combine, {$D_1 = v_1$, $[(D_2 = v_{21}, D_3 = v_{31}), (D_2 = v_{21}, D_3 = v_{32})]$}g	-	Stable Increase from 2015 to 2020 Followed by a Sudden Crash in 2020: A Tale of Stock Market.

R3. Diversity in natural language expressions. For a given fact, diverse natural language expressions should be generated. This diversity and fluency can be achieved through the extraction of corresponding expression patterns from user data and the use of back-translation.

Dataset construction. Each item in the training dataset consists of a fact-title pair. For example, the fact may be "compare f obj 1: the price of beef, obj 2: the price of pork, sign: larger, more times: 1-fold g", and the corresponding title may be "Beef is twice as expensive as pork". The common method for constructing a high-quality struct-to-text dataset (e.g., WikiSQL [42]) typically involves initial synthesis based on templates, paraphrasing, and quality checking. To support the model's ability to handle diversity in facts, attributes, and natural language presentations, we use the following steps to construct the dataset:

Initial dataset synthesizing. Our dataset should cover diverse data domains and facts (R1 and R2). We construct initial sentences using real-world tables with various meaningful data attributes. We choose the data tables from the Spider dataset [43], which contains 876 real-world data tables from various domains, each with categorical, quantitative, and temporal attributes that are semantically connected. Examples of these attributes include the height and weight of students, the capacity of buildings, and the credits of courses. A straightforward method for converting these facts into sentences is to fit words into templates. For instance, the fact "compare f reference 1: the price of beef, reference 2: the price of pork, sign: larger, more times: 1-fold g" can be converted to "The price of beef is larger than the price of pork by 1-fold" using the template "[reference 1] is [sign] than [reference 2] by [more times]".

Back-translation. Back translation methods [44] are often used for data augmentation in natural language processing tasks. In these methods, sentences generated by template-based methods are translated into another language and then translated back into English. In our scenario, the back translation method helps eliminate grammatical errors and provides more idiomatic and diverse expressions (R3). For example, consider the following three sentences generated by the same template with different data attributes, along with their back translations: (1) The USA's GDP is larger than Mexico's GDP. → The USA's GDP is higher than Mexico's. (2) The USA's area is larger than Mexico's area. → The USA is bigger than Mexico. (3) The USA's population is larger than Mexico's Population. → The USA has more population than Mexico. As can be seen, the sentences become more natural and fluent after the back translation process.

Manual check and editing. Since back-translation may introduce semantic bias in sentences, we manually check the generated sentences and remove or edit those with incorrect semantic meanings. As a result, we obtained 6,000 fact-title pairs.

LLM-based generation. Natural language generation (NLG) from structured data has been studied for many years [18]. Recently, large language models [24], [26] (LLM) that have been pre-trained on large-scale natural language corpus have shown the ability to be generalized to various new tasks through fine-tuning on relatively small datasets. Our task is a translation task, thus we chose the T5 model [26] (short for text to text translation transformer) as the base model and fine-tuned it using the fact-to-title dataset. The trained model is able to generate titles given the input of facts. The input facts and output sentences are treated as word sequences for training and deployment. The training loss was observed to converge after 20 epochs, as measured by the mean squared error on a validation dataset, as displayed in Figure 4. The trained model uses a sequence-to-sequence architecture with attention mechanisms to generate fluent sentences given a fact as input. For example, given the fact "compare f obj1: Tangorodrim's age; obj2: Black Flame's age; sign: smaller g" as input, the trained model outputs the sentence "Tangorodrim is younger than Black Flame". The trained model determines the most appropriate term for expressing the comparison, and in this case, it chose the term "younger" because it is more natural and idiomatic than other possible options such as "the age is smaller". The average BLEU [45] scores on the test dataset are 0.67 (BLEU-1) and 0.72 (BLEU-3), indicating that the model is able to generate sentences that are highly similar to the reference titles in terms of content and structure.

Fig. 4: The training loss converged at 20 epochs.

4.3 Quantifying Metrics for Facts and Titles

In subsection 3.2, we summarized the factors for measuring a title. However, without proper quantifying metrics for these factors, it is still difficult for users to understand and explore the title space. Therefore, we develop quantifying metrics for each factor. Quantifying the amount of information in a natural language sentence is a non-trivial problem that has been studied for a long time. D'Alfonso [32] proposed a method for quantifying semantic informativeness according to the atomic facts in a natural language sentence. Inspired by D'Alfonso [32], we proposed a multi-dimensional quantification method according to the atomic facts of the visualization. In the following text, we describe the quantification for factors of a title S and its corresponding fact F mentioned in subsection 3.2, namely, feature importance I, conciseness C, generic information richness G, preciseness P, coverage (broadness) B and non-technicality T.

Feature Importance is determined by the sub-facts and the importance of the feature type. The salient facts, such as extremes, outliers, and sudden changes in trends, are more important. A fact is more important if the sub-facts are more important or the feature type is important. $I(S) = \sum_{i=1}^{n} I(S_{s_i}) \cdot I_{t_{fact}}$, where S_{s_i} is the sub-facts of the title and $I_{t_{fact}}$ is the importance of the feature type. Unless there are extreme values, the feature importance of atomic facts is set to the same value. When dealing with a chart that contains multiple temporal lines, it's crucial to analyze the trends and compare and combine them. For temporal stacked charts, it's important to pay attention to the trend of the summation. Non-temporal charts should emphasize the comparison of categories and the identification of extremes.

Conciseness decreases when the number of words increases. We set the conciseness metrics similar to the brevity penalty of BLEU [45]. The conciseness is defined as $C(S) = e^{(1 - l/m)}$, where l is the number of words. The minimum value of m is set as the minimum length of all generated titles to ensure that the value is constrained to the range of 0 to 1.

Generic Information Richness measures how much generic information like attribute name and feature type is involved. We count the number of range, attributes, and feature types as the generic information richness. $G(F) = n_{range} + n_{attr} + n_{feature}$.

Preciseness refers to the extent to which a title accurately reflects the truth. In the context of titles, preciseness measures how much the numbers or values mentioned in the title deviate from the ground truth. Using approximations may introduce a loss in preciseness. For instance, if the ground truth value is 2.11 times, but the title uses "doubles" to describe it, the preciseness can be calculated based on the difference between the described value and the ground-truth value. The precision score is presented as: $P(F) = 1 - \frac{|n_d - n_r|}{|n_r|}$, where n_d is the described value and n_r is the real value. When it comes to trend analysis, the difference between the ideal trend and the ground-truth data values (real trend) can affect the preciseness of the title. In practice, it is assumed that when the value increases, the ideal trend will increase linearly. As shown in Figure 5, the impreciseness can be measured by the difference between the ideal interpolated trend and the ground truth trend.

Fig. 5: Trend preciseness calculation. Given a trend and a sentence, the trend preciseness can be calculated according to the difference between the real trend and the ideal trend.

Coverage measures the broadness of atomic facts covered by the sentence. Each atomic fact has a coverage of 1. The title of high-level facts have coverage of the summation of the sub-facts: $B(S) = \sum_{i=1}^{n} B(S_{s_i})$, where S_{s_i} is the sub-facts of the title.

Non-Technicality refers to avoid specialized knowledge or background to comprehend the visualization. According to a formative study, the number of words and complexity of words are the two most important factors affecting non-technicality, and they have been used in previous approaches to calculate non-technicality [46], [47]. For instance, Flesch [46] defined non-technicality by considering the number of words and the average number of syllables in a word, as the syllable count indicates complexity. We adopt Flesch's approach to calculate non-technicality using the formula $T(S) = 206.835 - 1.015(l_w) - 84.6(l_s)$, where l_w represents the number of words, and l_s is the average number of syllables in a word. For common sentences, the non-technicality score ranges from 0 (practically unreadable) to 100 (easy for a literate person). Regarding numbers, we first convert them to their English form and then count the word and syllable numbers to calculate the value. For instance, "2007" is converted to "two thousand and seven."

4.4 System Interface

Figure 6 shows the interface of AutoTitle, allowing users to upload a visualization and generate a desired title. Once uploaded, parsed attributes and color mapping are displayed in (g). Users can modify data attribute information in (g) if parsed errors occur. The back-end system generates titles, calculates quantifying metrics, and sends them back to the front end.

The interface helps users comprehend the title space as well as find desired titles from the vast design space. AutoTitle offers two methods for exploring the space of titles: the Radar view (e) and the RadViz view (f). Both views have six axes representing metrics. In the Radar view (e), each title is a line connecting the data values for each factor. The radar graph enables users to filter on each axis, obtain cross-filter results, and select a title by hovering or clicking the line. The RadViz view [48] (f) plots all titles as points among the six metrics using the force-directed method. The larger the value in a metric is, the stronger the attraction force is between the metric and the point. The RadViz view enables users to hover over and click to select the desired metric range. For instance, if users prefer preciseness, they can move closer to the related dimension. The representative title view (c) shows representative titles that have a high score in at least one dimension. Once the range of factors is determined, the system sorts the titles that meet the requirements in the title view (d). Users can specify a metric to sort the selected titles. The title with

Fig. 6: The interface of our controlled title generation. (a) allows users to upload a visualization. (b) presents the currently selected title. (c) the representative title view. (d) the alternative title view presents a list of alternative titles. (e) and (f) are title selection panels that can be switched using the toggle. The radar view (e) shows the value of the six metrics of the titles and allows users to select the desired range. The Radviz view (f) maps a title as a point. The data information view (g) shows the generic information of data attributes.

the highest value in the speci ed metric is displayed above the visualization in (b).

5 Case Study

In this section, we discuss several real-world cases where our title generation system was used. Figure 7 shows three charts crawled online, including an area chart, a line chart, and a bar chart.

Change of Tennis Backhand Type.[1] Figure 7 (a) shows the stacked area chart of one-handed and two-handed backhands of the top 100 male tennis from 1973 to 2021. Among the generated titles, we have chosen the three most representative ones. Auto Title can generate titles that describe generic information, such as (1), which has high coverage. Title (2) describes the growth in the proportion of two-handed backhands, covering half of the atomic facts. It is more precise and has more feature importance than Title (1) because it describes the trend. Title (3) describes the trend for both categories, which has broader coverage, and higher feature importance than the others. However, the preciseness of Title (3) is slightly lower than that of Title (2) because (2) provides the degree of the trend.

Daily Visualization Views on Datawrapper. Figure 7 (b) is a line chart crawled online[2] showing the daily number of visualization views created on the Datawrapper website from Oct. 2019 to Jul. 2020. Figure 7 shows three typical generated titles for the chart. Title (1) presents generic information about the chart, which has high coverage because it covers all atomic but low feature importance and preciseness because it does not provide any data features. Title (2) mentions an “exponential surge,” which has high feature importance because it describes a critical feature of the chart. However, its coverage is low because it only covers a small proportion of the time range, namely, the atomic facts. Title (3) describes the maximum value of daily views, which has high feature importance. Among these titles, Title (1) has the highest conciseness and non-technicality.

Global Sea Levels. Figure 7 (c) shows a bar chart[3] of the annual global sea levels from the 1700s to 2000s. Titles (1) and (3) are the generic titles showing the basic information, which have a high coverage but low importance. Title (3) has higher general information richness than (1) because it includes the time range. Title (2) is more precise than (1) and (3), as it describes the trend feature. Among these sentences, Title (1) has the highest conciseness, and Title (3) has the lowest.

Downloads Growth in Online Meeting Tools. Figure 8 illustrates the growth rate in downloads of online meeting tools. The chart features a categorical attribute and a quantitative attribute. we showcase four titles that capture the essence of the data. Title (1) provides a straightforward expression of general information, with low feature importance and precision. On the other hand, title (2) highlights the maximum value, held by Houseparty, and has high feature importance. It also boasts a concise and non-technical title, making it highly accessible. Titles (3) and (2) convey similar meanings (with compare form), but provide more general information and preciseness. Lastly, title (4) accurately conveys the numerical value of Houseparty's growth rate, with high precision but limited coverage.

1. How has male tennis changed from 1973 to 2021: https://observablehq.com/@unkleho/how-has-mens-tennis-changed-from-1973-2021
2. Datawrapper daily visualization views: https://blog.datawrapper.de/coronavirus-data-visualization-effect-datawrapper/
3. https://observablehq.com/@terezaif/annual-global-sea-levels