ChinaRxiv

AI translation · View original & related papers at
chinarxiv.org/items/chinaxiv-202506.00091

A Retrieval-augmented Analysis Framework for Target Galaxy Search

Postprint

Authors: Chao Tang, Yihan Tao, Dongwei Fan, Shirui Wei, Chenzhou Cui and Changhua Li

Date: 2025-06-13T16:50:50+00:00

Abstract

The identification of specific galaxy populations in large-scale spectroscopic surveys represents an essential yet challenging task, particularly for rare or anomalous galaxies that deviate from the typical galaxy distributions. Traditional methods based on template-fitting or predefining spectral features face challenges in addressing the complexity and scale of modern astronomical data sets. To overcome these limitations, we propose GalSpecEncoder-KB, a modular and flexible framework that combines deep learning with knowledge base retrieval to enable efficient and interpretable analysis of galaxy spectra. The framework integrates a Transformer-based feature encoder, GalSpecEncoder, pre-trained with masked-modeling strategy to capture semantically rich and context-aware spectral representations. By leveraging a Retrieval-Augmented Analysis approach, the knowledge base constructed from catalogs enables metadata retrieval and weighted voting for target galaxy identification. Using the Sloan Digital Sky Survey as a comprehensive case study, we demonstrate the capabilities of the framework for target galaxy search. Experimental results demonstrate the exceptional generalizability and adaptability across diverse galaxy search tasks, including identification of LINERs, Strong Gravitational Lenses, and detection of Outliers, while maintaining robust performance and interpretable spectral analysis capabilities.

Full Text

Abstract

The identification of specific galaxy populations in large-scale spectroscopic surveys represents an essential yet challenging task, particularly for rare or anomalous galaxies that deviate from typical galaxy distributions. Traditional meth-

ods based on template-fitting or predefining spectral features face challenges in addressing the complexity and scale of modern astronomical data sets. To overcome these limitations, we propose GalSpecEncoder-KB, a modular and flexible framework that combines deep learning with knowledge base retrieval to enable efficient and interpretable analysis of galaxy spectra. The framework integrates a Transformer-based feature encoder, GalSpecEncoder, pre-trained with a masked-modeling strategy to capture semantically rich and context-aware spectral representations. By leveraging a Retrieval-Augmented Analysis approach, a knowledge base constructed from catalogs enables metadata retrieval and weighted voting for target galaxy identification. Using the Sloan Digital Sky Survey as a comprehensive case study, we demonstrate the capabilities of the framework for target galaxy search. Experimental results demonstrate exceptional generalizability and adaptability across diverse galaxy search tasks, including identification of LINERs, Strong Gravitational Lenses, and detection of Outliers, while maintaining robust performance and interpretable spectral analysis capabilities.

Key words: catalogs –galaxies: general –methods: data analysis

1. Introduction

The study of galaxy spectra is crucial for unraveling the physical processes occurring within galaxies and tracing their evolution across cosmic time. By analyzing spectral features, we can infer a galaxy' s composition, star formation history, and the activity of its central black hole, offering insights into the mechanisms driving galaxy formation and evolution. Among these pursuits, a particularly significant task is the search for target galaxies with specific characteristics, such as low-ionization nuclear emission-line region (LINER) galaxies or strong gravitational lens systems. This task holds immense scientific value, as it enables the identification of rare or extreme galaxy populations and their role in the broader cosmic ecosystem. However, the search for target galaxies is inherently challenging due to the complexities of galaxy spectra. Spectra often exhibit a range of intricate features, including strong emission lines indicative of dynamic astrophysical processes, redshift-induced distortions, and noise from observational limitations. The diversity and overlap in spectral characteristics further complicate the task, particularly when differentiating between subtle variations in galaxy types. Therefore, effectively learning representations from these spectra and extracting meaningful features are critical steps in addressing these challenges.

Traditional analytical methods, rooted in theoretical models [**?**, **?**, **?**], often fall short in capturing the nuanced details of galaxy spectra. While these methods provide a solid foundation, they lack the sophistication required to reproduce the intricate features of high signal-to-noise ratio (SNR) spectra accurately. This limitation is especially pronounced when dealing with strong emission lines, which serve as vital indicators of active processes within galaxies [**?**, **?**]. To address these limitations, a growing body of work has turned to data-driven

methodologies, particularly machine learning (ML), to extract and analyze the rich information encoded in galaxy spectra [**?**]. Specifically, these methods can be broadly categorized into two different classes: supervised learning and unsupervised learning approaches.

Supervised learning approaches, which leverage labeled data sets, are often employed for regression and classification tasks, such as galaxy spectral classification [**?**]. By learning the relationships between input features and labeled outputs, these models have achieved remarkable success in data-rich tasks, accurately identifying galaxy properties and types. However, the reliance on large, high-quality labeled data sets poses significant limitations. Labeling astronomical data is labor-intensive and costly, and many regions of parameter space remain underexplored due to a lack of sufficient training samples. Moreover, supervised learning models are inherently task-specific, requiring bespoke architectures and extensive retraining for each new analysis. This results in inefficiencies and reduced flexibility in the data analysis pipeline.

In contrast, unsupervised learning eliminates the dependency on labeled data by uncovering latent structures and intrinsic patterns directly from the data. Classical techniques like Principal Component Analysis (PCA) have been widely used to reduce the dimensionality of galaxy spectra and extract dominant features [**?**]. Clustering algorithms, such as k-means and hierarchical clustering, have further enabled the grouping of galaxy spectra based on intrinsic similarities, aiding in the identification of distinct populations [**?**]. However, these methods often rely on oversimplified assumptions, such as linearity in PCA or fixed cluster numbers in clustering, which may fail to capture the intricate variability of galaxy spectra. While unsupervised models have proven useful, their inability to effectively handle the high complexity and diversity of galaxy spectra highlights the need for more advanced approaches tailored to these challenges.

In response to the limitations of both supervised and unsupervised learning, self-supervised learning has emerged as a promising alternative, leveraging the power of learned embeddings to address diverse tasks in galaxy spectral analysis. For instance, [**?**] employed a variational autoencoder (VAE) to reduce the dimensionality of galaxy spectra, creating latent space representations that proved effective for downstream tasks such as outlier detection, redshift estimation, and galaxy classification. Building on this foundation, [**?**] introduced convolutional elements into the autoencoder architecture, enhancing the model' s ability to capture correlated spectral features. Further advancements by [**?**] integrated attentive convolutional encoders with physical modeling, yielding embeddings that facilitated anomaly detection and other tasks [**?**, **?**]. These self-supervised approaches highlight the potential of data-driven embeddings to extract meaningful representations from galaxy spectra, enabling flexible and efficient analysis without reliance on extensive labeled data sets.

In this work, we introduce GalSpecEncoder-KB, a self-supervised framework that fundamentally advances galaxy spectral analysis through two synergistic innovations. First, our GalSpecEncoder employs a Transformer architecture pre-

trained with masked spectral modeling, overcoming the limited receptive fields of prior convolutional neural network-based (CNN-based) autoencoders by capturing long-range dependencies across galaxy spectra through self-attention mechanisms. Second, we pioneer a Retrieval-Augmented Analysis (RAA) paradigm that integrates a catalog-derived knowledge base with similarity search, enabling flexible target galaxy identification without task-specific retraining. This contrasts with conventional end-to-end models that require complete architectural redesign for new tasks. The framework operates through two phases: (1) self-supervised pre-training of the spectrum encoder to generate context-aware embeddings, and (2) knowledge base construction from public catalogs using these embeddings, followed by metadata retrieval (comprising similarity search and metadata association) and weighted voting for analysis. Initial validation on Sloan Digital Sky Survey (SDSS) spectra demonstrates that this combination of Transformer-based representation learning and knowledge base system effectively preserves physically meaningful spectral features while maintaining task adaptability.

The outline of this paper is as follows. In Section 2, we provide a comprehensive description of the GalSpecEncoder-KB framework, including the detailed structure and training of the GalSpecEncoder model, as well as the strategy of the downstream knowledge base analysis pipeline. Section 3 outlines the experimental setup used to validate the effectiveness of the GalSpecEncoder-KB framework, using SDSS galaxy spectra as a case study. Section 4 presents the experimental results, while Section 5 discusses the impact of technical details on performance. Finally, the conclusions of this work are summarized in Section 6.

2. Methodology

In this section, we introduce GalSpecEncoder-KB, a spectral analysis framework tailored for target galaxy search tasks. Inspired by the Retrieval-Augmented Generation (RAG) technique [**?**], which integrates retrieval mechanisms with generative models for knowledge-intensive inference, our framework adapts this paradigm to galaxy spectral analysis—a domain requiring rich contextual information and expert-level feature interpretation. As illustrated in Figure 1

, the framework operates through two interconnected components: a deep learning (DL)-based encoder and a modular knowledge base system.

2.1. GalSpecEncoder

2.1.1. Model Structure An overview of GalSpecEncoder is depicted in Figure 2 [FIGURE:2]. The standard Transformer [**?**] receives a one-dimensional (1D) sequence of token embeddings as input. While spectral data are inherently a 1D signal, they consist of S data points. Treating each data point as an individual token in the Transformer would be computationally prohibitive due to the model' s quadratic time complexity, which scales poorly with increasing

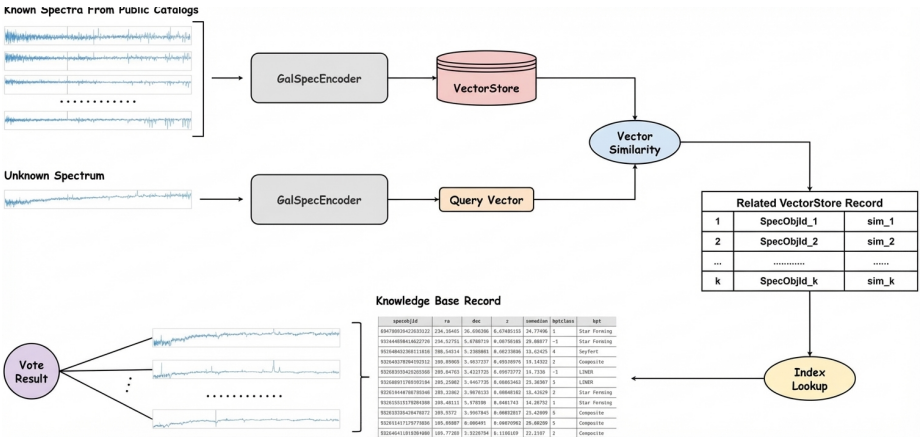

Figure 1: Figure 1

sequence length. To handle this issue, we reshape the spectrum into a sequence of fixed-size patches, where S is the resolution of the original spectrum, P is the resolution of each spectrum patch, O is the number of overlapping data points between two adjacent patches, and N = (S - P)/(P - O) + 1 is the resulting number of patches, which also serves as the effective input sequence length for the Transformer. The Transformer maintains a constant latent vector size D across all its layers. To align with this structure, we map each patch to D dimensions with a trainable linear projection. Additionally, position embedding is added to each spectrum patch to retain positional information. These embeddings are standard learnable 1D vectors. The resulting sequence of embedding vectors is then used as input to the encoder. We refer to the output of this projection as the patch embeddings.

The encoder module in our model is based on the standard Transformer architecture, specifically utilizing the encoder stack. It processes patch embeddings through L=6 identical layers. Each layer contains multi-head self-attention (MSA) and feed-forward networks (FFN). Standard implementations of these components are provided in Appendix A.1. In this work we employ h = 6 parallel attention heads, with embedding processing dimension Dh = D/h = 768/6 = 128.

2.1.2. Training Strategy Inspired by the work of [**?**], we use a masked-modeling self-supervised learning strategy (Figure 3 [FIGURE:3]) to pre-train the GalSpecEncoder model. Masked-modeling self-supervised learning strategies, widely employed in Natural Language Processing [**?**] and Computer Vision [**?**], have proven highly effective for pre-training spectral encoders [**?**]. It enables the model to learn context-aware representations by predicting missing elements in input sequences. The rich features in galaxy spectra, such as emission and absorption lines and continuum variations, make this approach particularly well-

suited.

In this work, we randomly mask 25% of the spectrum patch sequences to train the GalSpecEncoder. This masking strategy compels the model to capture both global relationships, such as correlations between features at different wavelengths, and local patterns, including detailed line shapes. By reconstructing the masked patches, the GalSpecEncoder learns to develop robust embeddings that encapsulate comprehensive spectral information. This approach aligns seamlessly with the self-supervised learning paradigm, effectively leveraging the abundance of unlabeled data to maximize the utility of available spectral features.

Furthermore, the masked-modeling strategy significantly enhances the encoder' s ability to handle observational gaps and noise, which are prevalent in astronomical data. By reconstructing missing spectral regions, the GalSpecEncoder is trained to infer information from contextual cues, fostering a deeper understanding of the intrinsic structures and variabilities of the data. This process equips the GalSpecEncoder with the flexibility required to adapt to the complexities of real-world spectra, ensuring reliable and consistent representation learning from diverse and imperfect inputs.

2.2. Knowledge Base System

2.2.1. Encoding and Retrieval The pre-trained GalSpecEncoder generates dense vector representations for known galaxy spectra from public catalogs. These vectors are stored in a vector database optimized through indexing techniques, balancing storage efficiency with retrieval accuracy. When analyzing an unknown spectrum, the same encoder maps it into this vector space as a query vector. We employ Euclidean distance (L2) to measure geometric proximity in the high-dimensional space, using the Flat algorithm to retrieve the Top_K most similar vectors. Each vector maintains linkage to its original specobjid identifier from source catalogs.

2.2.2. Metadata Association For target galaxy search applications, we associate spectral vectors with classification metadata critical for galaxy identification. This includes: (1) spectroscopic classifications from standard diagnostic diagrams (BPT classification), and (2) catalog-level types (e.g., Strong Gravitational Lens, Outliers). The object specobjid linkage supports direct retrieval of these identifiers through federated cross-matching with source catalogs.

2.2.3. Weighted Inference The conversion from L2 distances (xi) to voting weights (wi) prioritizes two objectives: (1) assigning higher weights to closer neighbors and (2) ensuring stable weight differentiation while mitigating the effects of extreme values. We evaluate four candidate transformations (see Appendix A.2 for details) and select the optimal function through comparative analysis of their weight distributions. The exponential transformation e^{-x} demonstrates superior performance, as quantified by its alignment with the

knowledge base' s distance patterns and robustness metrics. The final weighting scheme applies softmax normalization to ensure the weights sum to one.

3. Experiments

This section provides an overview of the experimental setup to validate the effectiveness of the proposed GalSpecEncoder-KB framework. To evaluate its performance, we use SDSS galaxy spectra as a practical use case. The section is divided into two main parts: Dataset Construction & Data Pre-processing Operations (Sections 3.1 and 3.2 respectively), and Experiment Setting (Section 3.3).

3.1. Dataset Construction

In this work, the data set serves two primary purposes: self-supervised learning of the GalSpecEncoder and the construction of a foundational knowledge base. For self-supervised learning, explicit labels are not required to compute the loss function or guide model training, as the model derives insights directly from the data without supervision. In contrast, constructing the knowledge base—intended for downstream applications identifying specific galaxy types—necessitates categorical information to ensure its accuracy and practical relevance.

To achieve these objectives, we utilize spectroscopic data from the SDSS Data Release 16 (DR16; [**?**]). Data acquisition was carried out as part of the SDSS and the Baryon Oscillation Spectroscopic Survey (BOSS; [**?**]), with wavelength coverage of (3800, 9200) Å and (3600, 10400) Å, respectively. The overall data were drawn from the SpecObj catalog within SDSS DR16, a comprehensive resource encompassing spectroscopic measurements for 2,963,274 galaxies. The catalog provides spectral data with a resolution of approximately R 2000.

To construct the knowledge base for the galaxy search pipeline proposed in this work, accurate galaxy classification information is essential. We utilize data labels from two SDSS emission-line VACs: galSpecExtra (SDSS DR8; [**?**]) and emissionLinesPort (SDSS DR12; [**?**]). To ensure label reliability, only samples with consistent classification results across both catalogs were included. Additionally, the knowledge base is enriched with entries from the eBOSS Strong Gravitational Lens Detection Catalog (SDSS DR16; [**?**]) for Strong Gravitational Lenses and data from [**?**] to incorporate Outliers. These curated resources form the foundation for robust spectral analysis and galaxy search tasks.

3.2. Data Pre-processing Operations

To standardize the spectral data, we resampled the SDSS spectra from logarithmic wavelength space to a fixed linear wavelength range of 3850–9000 Å using interpolation. This process ensured a uniform resolution of 3688 data points across all spectra, facilitating precise alignment and comparability of spectral features. By addressing gaps and smoothing irregularities, this standardization

enhances the continuity of spectral profiles, providing a robust foundation for subsequent analysis and model training. Such pre-processing is critical for preserving data set integrity and ensuring the extraction of reliable and scientifically meaningful insights.

To normalize the spectra, we computed the median flux within the rest-frame wavelength range rest = 5300–5850 Å. This wavelength window was chosen for its relative insensitivity to redshift-induced variations, thereby reducing potential amplitude biases. This approach effectively mitigates distortions linked to redshift-dependent flux scaling, which is essential when handling large data sets with diverse redshift distributions. This normalization aligns spectral amplitudes across the data set, reducing variability and ensuring uniformity. It establishes a consistent framework for identifying subtle astrophysical patterns, thereby enabling reliable comparative analysis and robust data interpretation.

To adapt spectral data for the Transformer while addressing its quadratic self-attention complexity $O(n^2)$, we segmented the spectra into fixed-size patches, each comprising 15 spectral points with an overlap of 10 points. This patch-based approach encodes local spectral features, enabling the Transformer to effectively process non-text sequential data and focus on localized information. By reducing the sequence length appropriately and leveraging GPU parallelism, this method significantly decreases computational costs and accelerates training. This pre-processing step ensures compatibility with the Transformer, facilitating efficient and scalable analysis of large-scale spectral data sets like SDSS.

In the end, we obtain 293,392 SDSS spectra for pre-training, with each sample comprising 3688 data points. Additionally, 46,054 labeled spectra are used for sampling during the construction of the knowledge base.

3.3. Experiment Setting

To verify the effectiveness of the proposed GalSpecEncoder-KB framework, we conduct a set of experiments leveraging pre-processed SDSS galaxy spectra. In the first phase, we examine the representation capabilities of the GalSpecEncoder by performing comparative studies, which assess its ability to encode spectral data into high-quality embeddings that accurately capture the structural and informational characteristics of the input spectra. Building upon this evaluation, the second phase focuses on the framework' s target galaxy search functionality. In this work, we design a series of search tasks based on SDSS data to evaluate the pipeline' s robustness and adaptability under varying application scenarios, highlighting its practical applicability to real-world spectral analysis challenges.

The proposed model is implemented using the PyTorch framework [**?**]. All experiments in this work are conducted in the same hardware and software environment. The hardware environment is based on a CentOS Linux release 7.9.2009 (Core) operating system installed on a 64-bit Intel(R) Xeon(R) Platinum 8163 CPU @ 2.50 GHz with four cores and four NVIDIA Tesla V100 SXM2

GPUs, supported by Alibaba Cloud Elastic Compute Service. The software environment is Python 3.10.16, torch 2.5.1 and torchvision 0.20.1.

In this work, we adopt the mean-squared error (MSE) loss to measure reconstruction performance and optimize the model with the AdamW optimizer [**?**]. The initial learning rate is set to 1×10 and managed using a cosine annealing schedule with a warmup phase. During the first 2000 steps, the learning rate gradually increases, after which a cosine annealing schedule is applied over the maximum training duration of 500,000 steps. To mitigate overfitting, we adopt a model checkpointing strategy that retains only the best-performing model based on validation loss. The batch size is set to 32, ensuring that each training iteration processes 32 data samples, balancing computational efficiency and performance.

4. Results

In this section, we present the results of our experiments conducted on SDSS galaxy spectra to evaluate the proposed GalSpecEncoder-KB framework. First, we perform comparative studies to assess the encoding quality of the GalSpecEncoder and validate its ability to effectively represent SDSS spectral data. Next, we focus on the core application of the pipeline—target galaxy search—by designing and analyzing multiple search experiments on SDSS data. These results highlight the framework' s effectiveness in identifying specific galaxy types and demonstrate its potential for advancing galaxy spectral analysis.

4.1. Experimental Evaluation of Encoding Quality

In the analysis of SDSS galaxy spectra, the search for target galaxies is fundamentally a feature identification and classification task. Specifically, it involves identifying galaxies with specific characteristics or attributes from vast observational data, where these features may relate to the galaxies' physical properties, evolutionary stages, or environmental factors. Therefore, to evaluate whether the proposed analysis pipeline, combining the GalSpecEncoder with a knowledge base, is effective for target galaxy search tasks, we design a comparative experiment involving a three-class classification task: star-forming, active galactic nucleus (AGN), and normal galaxy. This experiment serves to validate the utility of our approach. The labels for the test set and the prior information in the knowledge base are derived from the two emission-line catalogs provided by SDSS, as described in Section 2: galSpecExtra and emissionLinesPort. Table 1 summarizes the sample data used in the comparative experiments.

To comprehensively evaluate the effectiveness of our approach, this section compares it with two representative benchmarks: a classical ML method, random forest (RF), and a DL model. RF [**?**] is a well-established supervised ML technique widely applied to automate spectral classification tasks for stars and galaxies [**?**, **?**]. Notably, [**?**] demonstrated that RF outperformed other methods, including k-nearest neighbors (KNN), support vector classifiers, and multi-

layer perceptron (MLP) networks, in classifying intermediate-redshift emission-line galaxies, showcasing its robust distinguishing capability. Meanwhile, GalSpecNet [**?**] is a 1D CNN model, which leverages supervised learning to perform end-to-end training for classifying emission-line galaxy spectra into star-forming, composite, AGN, and normal galaxy cases, demonstrating state-of-the-art (SOTA) performance. Therefore, we select these two representative models for our comparative experiments. To further evaluate whether the Transformer architecture can capture meaningful physical semantics, we performed supervised training on the three-class classification task using the same training data, loss function, and optimizer as in GalSpecNet.

To evaluate the performance of our method in the comparative experiments, we adopt commonly used classification metrics, including Accuracy (Acc), Precision (P), Recall (R), and F1-score, to assess the results for each galaxy type. In this framework, each galaxy is treated as a binary classification problem, where the target type is labeled as the "true type" and all other types are collectively considered the "false type". For example, when evaluating star-forming galaxies, true positives (TP) represent cases where both the label and the model' s prediction correctly identify the galaxy as star-forming. True negatives (TN) indicate cases where the model correctly predicts a galaxy as not star-forming, aligning with the ground truth. False positives (FP) occur when the model incorrectly classifies a galaxy as star-forming, while the ground truth designates it as not star-forming. Conversely, false negatives (FN) represent cases where the model fails to classify a galaxy as star-forming when it is labeled as star-forming. We here use the macro-averaged F1-score.

Table 2 presents a comprehensive comparison of performance metrics across galaxy types among the evaluated models mentioned above, highlighting their relative strengths and limitations. When comparing GalSpecEncoder-KB with the supervised GalSpecEncoder, we find that although the self-supervised strategy uses a large amount of unlabeled data for pre-training, its learning goals differ from the final classification task. Self-supervised learning focuses on capturing general features, while supervised learning directly targets the classification task and better captures the relevant features. As a result, for specific classification tasks, the supervised method may perform better. Additionally, the similarity-based retrieval used in GalSpecEncoder-KB depends heavily on the quality and completeness of the vector knowledge base, which can limit its ability to handle ambiguous galaxy types.

Furthermore, the overall classification performance of GalSpecNet is better than that of the supervised GalSpecEncoder. This is probably attributed to the translation invariance inherent in CNNs, an inductive bias that mitigates redshift effects and enhances the robustness of spectral representations. However, supervised ML/DL approaches require building task-specific data sets and models, which limits their reuse and transfer. The knowledge base approach mitigates these constraints by enabling similarity-based retrieval without additional training, significantly reducing computational costs and time while enhancing

adaptability to other tasks.

This classification experiment validates the efficacy and generalizability of GalSpecEncoder embeddings, while demonstrating the effectiveness of our knowledge-based RAA framework. These empirical results establish a robust foundation for downstream classification tasks, sustaining a comprehensive evaluation of our methodology. Considering the balance between accuracy and flexibility, we adopt the proposed GalSpecEncoder-KB framework in subsequent experimental investigations.

4.2. Experimental Results for Target Galaxy Search

The comparative experiments presented in Section 4.1 validate the effectiveness of GalSpecEncoder-KB in spectral classification tasks. Motivated by these results, we extend its application to galaxy search tasks, specifically targeting the identification of certain types of galaxies within large-scale observational data sets. The GalSpecEncoder-KB framework achieves this by decoupling the feature extraction and classification modules, leveraging a knowledge base and similarity search technique to achieve classification outcomes. As a result, the method allows for the rapid adaptation to galaxy search tasks by simply reconfiguring the knowledge base to include prior information about the target galaxies. This flexibility ensures that the model can efficiently utilize the robust context-aware representations learned during pre-training, enabling seamless transfer to search tasks for any galaxy type.

To further evaluate the model' s performance in target galaxy search tasks, we design a series of experiments:

1. We further refine the classification by designating LINER galaxies as the target for the search task. To test our approach' s transferability, we carefully configure the knowledge base using data summarized in Table 1, ensuring sample diversity and representativeness. Specifically, the samples for star-forming, composite, Seyfert, LINER, and normal galaxies are sourced from the intersection of the galSpecExtra and emissionLinesPort catalogs.

2. The second task focuses on searching for small-sample galaxies, specifically selecting Strong Gravitational Lenses from the eBOSS Strong Gravitational Lens Detection Catalog. This catalog contains 838 "likely," 448 "probable"and 265 "possible"strong lens candidates, all identified within the final data release of eBOSS (part of SDSS DR16). These candidates were spectroscopically identified using methodologies derived from the BOSS Emission-Line Lens Survey (BELLS) and Sloan Lens ACS (SLACS) surveys [**?**], with enhanced inspection tools introduced by [**?**] to refine and expand upon earlier detection techniques.

3. In the final experiment, we aim to further investigate the quality of prior knowledge retrieved from the knowledge base. To achieve this, we refer

to the work of [**?**], who utilized an unsupervised RF-based anomaly detection algorithm to analyze 2,379,168 galaxy spectra from SDSS DR12. Their study identified over 400 samples with the highest anomaly scores, subsequently categorizing them into 16 distinct classes. Building on this foundation, we design the Outliers as the search targets and examine the similarity of the samples retrieved through our approach.

In summary, this study incorporates three distinct galaxy search experiments: LINER, Strong Gravitational Lens, and Outlier. To streamline the experimental setup, a single mixed test set is designed, containing both target and non-target galaxies, simulating real-world search scenarios. For instance, in the LINER galaxy search task, the goal is to extract LINER samples from the mixed test set. Detailed information on the experimental data is presented in Table 3 . The primary objective of these experiments is to evaluate whether our GalSpecEncoder-KB framework can be directly adapted to various galaxy search tasks without retraining. Performance metrics include P, R, and F1-score, with particular emphasis on R. This focus reflects the critical importance of minimizing false negatives in galaxy search tasks, where missing target galaxies could significantly impact the effectiveness of the analysis.

Table 4 presents the detailed summary of the P, R, and F1-score for the galaxy search tasks. The results highlight both the strengths and challenges of applying the GalSpecEncoder-KB framework across diverse search scenarios.

In the LINER galaxy search, our model achieves a precision of 0.8727, indicating a relatively low proportion of false positives. The recall is notably high at 0.9460, reflecting the model' s strong capability to correctly identify LINER galaxies within the mixed test set. The resulting F1-score of 0.9079 underscores a balanced performance in detecting LINER galaxies. However, the precision being lower than recall suggests some confusion between LINER galaxies and other types, likely due to the overlap of spectral characteristics within AGN subtypes. This overlap complicates the task of differentiating LINER galaxies from other AGN-like systems, especially star-forming or Seyfert galaxies, which share similar emission line characteristics.

The search for Strong Gravitational Lens achieves remarkable results, with near-perfect precision of 0.9937 and F1-score of 0.9701, demonstrating the model' s ability to effectively distinguish these rare but unique systems from non-target samples. While the recall (0.9476) is slightly low, it remains sufficiently high to ensure reliable identification of strong lens systems. This strong performance can be attributed to the distinct spectral features of lensing galaxies, which are effectively captured by our GalSpecEncoder-KB framework.

The Outlier galaxy search task yields the most variable performance among the three experiments, with a precision of 0.9847, recall of 0.8377, and an F1-score of 0.9053. The high precision indicates that most retrieved galaxies are genuine Outliers, but the lower recall reveals challenges in detecting a significant portion of the Outlier population. This result reflects the inherent difficulty in

identifying rare and diverse galaxy types, as Outliers are by definition characterized by atypical features that do not consistently align with the dominant patterns in the knowledge base. The lower recall may also be influenced by the relatively small representation of Outlier galaxies in the training set, which limits the model' s exposure to their unique characteristics during knowledge base construction.

After analyzing the recall failure samples (see Appendix B.1 for details), we determine that the low recall rate primarily stems from the absence of consistent spectral characteristics among the Outliers. With 16 distinct Outlier subtypes —each possessing unique spectral properties—each subtype effectively functions as an independent target during similarity retrievals. The limited sample sizes for certain subtypes hinder the development of a robust "search focus" within the knowledge base. In contrast, although there is one instance of the "Outliers on BPT diagram" among the recall failures, this subtype' s overall recall rate is an impressive 0.9778. This clearly indicates that when a target galaxy can establish a distinctive feature cluster in the knowledge base, achieving a high recall rate becomes feasible.

The consistently high recall across all tasks underscores the robustness of our GalSpecEncoder-KB framework in minimizing false negatives, a critical priority in galaxy search tasks to ensure scientific completeness. Variations in precision reflect the challenges posed by the spectral diversity and rarity of certain targets, such as LINER and Outlier galaxies, compared to the distinctiveness of Strong Gravitational Lens. Importantly, the ability of the knowledge base to retrieve relevant priors, as demonstrated in the Outlier galaxy search experiment, suggests that these retrieved priors can offer valuable insights into the physical mechanisms or observational conditions underlying anomalous samples. This capability highlights the potential of the framework not only as a classifier but also as a tool for helping us to understand and interpret the nature of their data.

Overall, the results demonstrate the adaptability of our GalSpecEncoder-KB framework to a wide range of search scenarios without retraining, emphasizing its suitability for real-world astronomical applications (see Appendix B.2 for details). Future efforts could focus on enriching the knowledge base with more diverse and representative samples, as well as enhancing methods to better capture the complexity and diversity of Outlier galaxies.

5. Discussions

This part delves into the critical factors shaping the performance of our GalSpecEncoder-KB framework in target galaxy search tasks. We explore the influence of the knowledge base configuration, focusing on the proportion of target galaxies, and assess how it affects retrieval accuracy and coverage. Furthermore, we analyze the role of Top_K, the number of prior samples retrieved, in balancing precision, recall, and F1-score. These insights are

instrumental in refining the framework for varied application scenarios and improving its adaptability to diverse search tasks.

5.1. Knowledge Base Configuration

The configuration of the knowledge base is a pivotal factor influencing the performance of target galaxy search tasks. To investigate this, we design an experiment to analyze the effect of varying the proportion of target galaxies within the knowledge base. Specifically, we focus on Strong Gravitational Lens as the target and systematically adjust their proportion in the knowledge base to observe its impact on key performance metrics such as P, R, and F1-score.

The motivation behind this experiment lies in the challenges posed by real-world astronomical data sets, where target galaxies often constitute a small fraction of the total sample. Understanding how the proportion of target galaxies affects search performance is essential for optimizing the knowledge base to balance recall and precision. Recall is particularly critical in galaxy search tasks, where minimizing false negatives ensures the comprehensive identification of target galaxies.

The results presented in Figure 4 [FIGURE:4] reveal a clear trade-off between precision and recall as the proportion of Strong Gravitational Lens increases. Higher proportions of target galaxies in the knowledge base enhance recall, as more similar samples improve the likelihood of correctly identifying target galaxies. However, this improvement comes at the expense of precision. The increased proportion of target galaxies introduces a class imbalance, leading to more false positives. This occurs because the model is biased toward predicting Strong Gravitational Lenses, even when their similarity scores are not sufficiently high. This trade-off is reflected in the F1-score, which balances the two metrics.

These findings emphasize the importance of carefully curating the knowledge base to achieve optimal performance. In applications where high recall is critical, such as identifying rare phenomena, a higher proportion of target galaxies may be preferable. Conversely, for tasks requiring high precision, reducing the proportion of target galaxies might help mitigate false positives. This experiment underscores the need for task-specific knowledge base configurations to ensure the robustness and adaptability of galaxy search models in diverse astronomical scenarios.

5.2. Top_K

To explore the impact of the number of retrieved priors (Top_K) on the performance of galaxy search tasks, we conduct experiments using multiple galaxy types as the target. By varying K while analyzing P, R, and F1-score, we aim to evaluate the influence of retrieval depth on search outcomes and identify an optimal K for balancing performance metrics.

Contrary to the typical trade-off observed in many retrieval tasks, the results

displayed in Figure 5 [FIGURE:5] indicate that precision, recall, and F1-score exhibit consistent trends as K increases. All three metrics gradually decrease with larger K values, suggesting that an increase in the number of retrieved priors does not significantly enhance recall but instead introduces marginally more noise, leading to reductions in precision and overall performance.

This behavior can be attributed to the characteristics of the data set and the model's similarity computation. When K is small, the retrieved priors are highly relevant to the target class, resulting in strong performances across all metrics. As K increases, the additional priors retrieved are likely to include less relevant or ambiguous samples, diluting both precision and recall without positive contributions. Based on this analysis, we select $K = 3$ as the optimal configuration. At this value, the model achieves favorable performance across precision, recall, and F1-score. This choice not only balances key metrics but also enhances the model's tolerance for errors, ensuring reliable results in galaxy search tasks. Furthermore, retrieving three priors can improve the interpretability of the search process by providing users with relevant samples to better understand the spectrum in question.

6. Conclusions

In this work, we propose a galaxy search framework, which integrates a Transformer-based spectrum encoder—GalSpecEncoder, along with a catalog knowledge base, similarity search techniques, and a weighted voting algorithm. Using the SDSS spectral data as a comprehensive case study, we demonstrate the effectiveness and generalizability of the framework. The primary contributions of our work can be summarized as follows:

1. The results tested on SDSS spectra highlight the effectiveness of our GalSpecEncoder, which is pre-trained using a masked-modeling strategy. This approach leverages the self-attention mechanism to model long-range dependencies and capture robust spectral features. As a result, GalSpecEncoder achieves enhanced generalization performance in downstream tasks and establishes a scalable framework for robust galaxy spectral analysis.

2. By relying on prior knowledge embedded in the knowledge base, our approach bypasses the need for extensive task-specific retraining, instead focusing on the retrieval of context-aware spectral embedding and corresponding metadata. Furthermore, the modularity and flexibility of this framework allow for seamless adaptation to a wide range of target galaxy search tasks. By simply reconfiguring the knowledge base to include spectra relevant to the desired targets, the pipeline can quickly and efficiently pivot to new applications, demonstrating exceptional transferability and efficiency in diverse spectral analysis challenges.

3. Our proposed GalSpecEncoder-KB framework improves model interpretability by using retrieved prior knowledge to analyze unknown

spectra. It automates target galaxy searches while integrating additional knowledge base fields to help users better understand spectral properties. Appendix B.2 provides a case study illustrating the framework' s workflow and decision-making process. Unlike traditional ML methods that rely on opaque weight matrices, our framework enhances transparency by leveraging retrieved metadata.

In conclusion, our proposed GalSpecEncoder-KB framework offers a robust and scalable solution for galaxy spectral analysis by combining Transformer-based encoding, metadata retrieval, and a weighted voting algorithm. However, there are some limitations. The method' s performance is somewhat unstable because it relies on high-quality, diverse training data and a comprehensive, representative knowledge base to accurately learn and generalize spectral features. Also, noise is currently mixed into the spectral features, so spectra with different SNR may harm retrieval and affect performance. Moreover, the similarity retrieval algorithm scans the entire vector database, and its computational cost grows linearly with the size of the database, which can become high when the database exceeds 500,000 entries.

To overcome these challenges, we plan to explore several optimization strategies in future work. We aim to adopt a Transformer-CNN hybrid model that preserves the Transformer' s advantage in modeling long-range dependencies while better capturing local features. We also plan to implement a learnable weight transformation strategy, replacing the fixed e^{-x}, to improve the weighted voting algorithm by considering additional context such as the spectral SNR. Finally, as indicated in Appendix C, the current exact similarity retrieval algorithm shows inefficiencies when handling large-scale knowledge bases. Therefore, we will consider using a Hierarchical Navigable Small World (HNSW; [**?**]) algorithm to boost retrieval efficiency when the knowledge base becomes very large, even if this requires more storage. Furthermore, we plan to extend the application of the GalSpecEncoder-KB framework to other downstream tasks, such as redshift estimation and galaxy physical parameter prediction, to further explore its potential and versatility.

Acknowledgments

This work is supported by the National Key R&D Program of China (2022YFF0711500), National Natural Science Foundation of China (NSFC, 12273077,12403102, 12373110, and 12103070), and Strategic Priority Research Program of the Chinese Academy of Sciences (XDB0550101). Data resources are supported by China National Astronomical Data Center (NADC), CAS Astronomical Data Center and Chinese Virtual Observatory (China-VO). This work is supported by Astronomical Big Data Joint Research Center, co-founded by National Astronomical Observatories, Chinese Academy of Sciences and Alibaba Cloud.

Appendix A Technical Specifications

A.1. Transformer Component Details

The encoder module in our model is based on the standard Transformer architecture [**?**], specifically utilizing the encoder stack. It is composed of a stack of L = 6 identical layers. Each layer has two sub-layers. The first is an MSA mechanism (Equation (A1)), which allows the model to jointly attend to information from multiple representation subspaces across different regions of the galaxy spectrum.

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)W^O$$

where head_i is standard qkv self-attention, which is widely used to build a block in neural network architectures. In our work, the self-attention mechanism generates query (q), key (k), and value (v) vectors for each spectral patch and computes attention weights based on the dot product between the queries and keys. These weights are then used to compute a weighted sum of the value vectors, resulting in a context-aware representation for each patch. This process enables the self-attention mechanism to capture global relationships among spectrum patches, effectively combining local and global information to extract richer spectral features.

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{D_h}}\right)V$$

Here the projections are parameter matrices $W^Q, W^K, W^V \in \mathbb{R}^{D\times D_h}$ for the queries, keys, and values, while $W^O \in \mathbb{R}^{D\times D}$ represents a trainable linear projection combining the heads' outputs. In addition, $X \in \mathbb{R}^{N\times D}$ represents an input sequence and $D_h = D/h$.

A.2. Selection of the Weight Transformation Function

To determine an appropriate transformation, understanding the distribution of pairwise distances within the knowledge base is critical to optimizing the weighting scheme and ensuring robust voting performance. Considering that this study uses SDSS galaxy spectra as a use case, we calculate and visualize the distribution of pairwise L2 distances across the SDSS-derived knowledge base (as depicted in Figure A1), which reveals the following characteristics:

1. **Skewness ( ).** The L2 distance distribution exhibits a slight right skew with a skewness value of 0.44. This indicates the presence of a few relatively large distance values that deviate from the mean.
2. **Kurtosis ( ).** The distribution shows a kurtosis of -0.40, suggesting it is flatter than a normal distribution. This implies that data points tend to be more concentrated around the mean with fewer extreme values.

3. **Mean ( ) and Median.** The mean distance is 2.84, while the median is 2.71, indicating a relatively symmetric distribution around the central tendency.
4. **Standard Deviation ( ).** The standard deviation is 1.18, providing a measure of the dispersion of distances around the mean.

Overall, the L2 distance distribution lacks fat-tail or sharp-peak characteristics, instead showing a slight skew and flat kurtosis, with data concentrated around the mean and few extreme values. These properties are essential considerations for selecting an appropriate transformation function.

To convert L2 distances into weights, we evaluated four candidates: inverse (1/(x +) with = 10), inverse square (1/(x ² +)), linear (-x), and exponential (e^{-x }) functions. The term prevents division-by-zero errors in inverse transformations. The impact of each function on the weight distribution was analyzed through their resulting skewness and kurtosis (as shown in Table A1). For our task, the ideal weight distribution should exhibit symmetry to minimize the impact of extreme values, distinctiveness to effectively differentiate weights across distances, and stability to ensure smoothness and robustness against outliers.

1. **Symmetry.** The e^{-x} function yields a weight distribution with a skewness of 1.69, which, while not perfectly symmetric, is a significant improvement over the highly skewed distributions produced by 1/(x +) and 1/(x ² +). The latter function introduces extreme values that could destabilize model performance.
2. **Distinctiveness.** With a kurtosis of 3.43, the e^{-x} transformation approaches the ideal kurtosis of 3 for a normal distribution. This indicates a more optimal balance between peakedness and tail weight, allowing for better differentiation of weights, particularly in high-density regions where distances are smaller.
3. **Stability.** The exponential function provides a smooth and gradual decay in weights as distances increase, avoiding the abrupt changes and extreme values associated with the other transformation functions. This stability is crucial for maintaining consistent model performance across various distance ranges.

In summary, the exponential function e^{-x} was selected as the weight transformation function due to its superior balance of symmetry, distinctiveness, and stability in the weight distribution. This choice is expected to enhance the overall performance and robustness of our model in processing L2 distances within the knowledge base.

Appendix B Case Study

B.1. Analysis of Recall Failure Samples in Outlier Galaxy Search Tasks

The overall Outliers sample comprises only 377 instances (as shown in Table 3). Since an additional 100 samples are randomly selected as the test set, the remaining pool becomes too limited to establish a robust "search focus" within the knowledge base. Our similarity search analysis identified two primary failure modes among the 17 undetected samples:

1. **Prototype Starvation.** 58.82% (10/17, corresponding to the first 10 rows in Table B1) of the failed samples belong to five rare subtypes (each represented by fewer than 10 samples in the knowledge base). The sparse priors associated with these subtypes hinder the formation of distinctive feature clusters in the vector space.
2. **Narrow Feature Window.** 35.29% (6/17, specifically referring to Sodium excess galaxies and Weak H emission) of the failed samples have a limited range of distinct spectral features, differing from normal galaxies by only a few key lines. This scarcity of unique features results in over-smoothed spectral representations during similarity comparisons, leading to increased misclassification.

These findings highlight the intrinsic challenges in detecting Outliers. Moreover, the low recall rate may be attributed to the relatively scant representation of Outliers in the training set, which restricts the model' s capacity to encode their unique spectral features.

B.2. Interpretability Verification of LINER Spectral Retrieval and Classification

In this study, we selected the SDSS spectrum spec-0881-52368-0036 (redshift z = 0.0668) as the target for case analysis. The spectrum was manually verified as a LINER using the BPT diagnostic diagram. This study aims to evaluate whether the GalSpecEncoder-KB framework can accurately identify such spectra and, through its retrieval results, clarify the basis for classification, thereby providing empirical support for the method' s interpretability.

Figure B1 compares the target galaxy spectrum with the top_3 retrieved samples. LINER galaxies are characterized by relatively strong low-ionization emission lines, such as [O I] 6300, [N II] 6583, and [S II] 6716,6731, while the [O III] 5007 line remains comparatively weak. This pattern suggests that LINER galaxies generally have a lower overall ionization state, distinguishing them from other types of AGNs. However, the comparison shows that the target spectrum and the three retrieved samples exhibit striking similarities in these key spectral features (highlighted in Figure B1), demonstrating the high quality of embedding and the effectiveness of the similarity search algorithm.

In Table B2, we performed a provenance analysis on the metadata of these four

samples. These parameters help us understand the physical meaning behind the retrieval results and the criteria for classification. For example, the weight values show the importance of each sample in the decision-making process, while the emission-line intensity ratios serve as key indicators in the BPT diagnostic diagram.

1. **Physical Consistency Verification.** According to [**?**], the dividing line between LINERs and Seyfert 2 galaxies is defined by Equation (B1). Our calculations show that both samples spec-1445-53062-0466 and spec-1009-52644-0170 fall below this line and are located very close to the target sample on the BPT diagram. This supports the reliability of the retrieval results.
$$\log([\text{O III}]/\text{H}\beta) = 1.18 \log([\text{N II}]/\text{H}\alpha) + 0.77$$

2. **Explanation of Anomalous Sample.** Although sample spec-1009-52644-0170 is labeled as a Seyfert galaxy, its parameter values indicate that it lies near the LINER/Seyfert 2 boundary and actually falls within the LINER region. This finding agrees with the known transitional types between LINER and Seyfert galaxies, where the boundary is not strict [**?**]. This observation further supports the validity of the retrieval outcomes.

3. **Decision Credibility.** Even with the inclusion of a mismatched sample, the weighted voting mechanism—with a weight ratio of 68.12%—produced a classification that matches manual verification. This confirms that the retrieval weighting method is effective and that the GalSpecEncoder-KB framework is reliable in its classification decisions.

In summary, this case study illustrates the workflow and decision-making criteria of the GalSpecEncoder-KB framework in the target galaxy search task. Through detailed visualization of retrieval results, metadata provenance analysis, and interpretability assessment, it provides strong empirical support for the method' s interpretability.

Appendix C Algorithm Efficiency Experiment

The absolute size of the knowledge base is a critical factor influencing the practicality of the GalSpecEncoder-KB framework, particularly in terms of its impact on the efficiency of our similarity retrieval process. To quantify this relationship, this study designed a systematic experiment. The experimental results (as depicted in Figure C1) demonstrate that, under the current configuration using the exact similarity retrieval algorithm, the retrieval time exhibits a strict linear growth with respect to the size of the knowledge base.

To assess the efficiency of our similarity retrieval process, we randomly select 10 test samples and measure the average retrieval time across knowledge bases of different sizes: 1000, 5000, 10,000, and up to 500,000 entries. Each retrieval is repeated 100 times to reduce randomness, and the average time per sample is recorded. The results show a linear growth in retrieval time, with a regression

function of y = 0.0058x + 0.0490, where x is the knowledge base size and y is the retrieval time.

ORCID iDs Yihan Tao https://orcid.org/0000-0002-3143-9337

References

Ahumada, R., Prieto, C. A., Almeida, A., et al. 2020, ApJS, 249, 3 Aihara, H., Prieto, C. A., An, D., et al. 2011, ApJS, 193, 29 Alam, S., Albareti, F. D., Prieto, C. A., et al. 2015, ApJS, 219, 12 Ansel, J., Yang, E., He, H., et al. 2024, in 29th ACM Int. Conf. on Architectural Support for Programming Languages and Operating Systems, 2 (New York: ACM) ASPLOS '24 Baron, D., & Poznanski, D. 2017, MNRAS, 465, 4530 Breiman, L. 2001, Machine Learning, 45, 5 Brownstein, J. R., Bolton, A. S., Schlegel, D. J., et al. 2011, ApJ, 744, 41 Dawson, K. S., Schlegel, D. J., Ahn, C. P., et al. 2012, AJ, 145, 10 Devlin, J., Chang, M., Lee, K., & Toutanova, K. 2019, in NAACL, Volume 1 (Minneapolis, MN: Association for Computational Linguistics), 4171 Dosovitskiy, A., Beyer, L., Kolesnikov, A., et al. 2021, in ICLR Fraix-Burnet, D., Bouveyron, C., & Moultaka, J. 2021, A&A, 649, A53 He, K., Chen, X., Xie, S., et al. 2022, in IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR) (Piscataway, NJ) 15979 Huertas-Company, M., & Lanusse, F. 2023, PASA, 40, e001 Ivezić, Ž., Connolly, A. J., VanderPlas, J. T., & Gray, A. 2020, Statistics, Data Mining, and Machine Learning in Astronomy: A Practical Python Guide for the Analysis of Survey Data, Vol. 8 (Princeton, NJ: Princeton Univ. Press) Kewley, L. J., Dopita, M. A., Sutherland, R., Heisler, C., & Trevena, J. 2001, ApJ, 556, 121 Kewley, L. J., Groves, B., Kauffmann, G., & Heckman, T. 2006, MNRAS, 372, 961 Kyritsis, E., Maravelias, G., Zezas, A., et al. 2022, A&A, 657, A62 Lewis, P., Perez, E., Piktus, A., et al. 2020, in NeurIPS (Red Hook, NY) Li, X.-R., Lin, Y.-T., & Qiu, K.-B. 2019, RAA, 19, 111 Liang, Y., Melchior, P., Hahn, C., et al. 2023a, ApJL, 956, L6 Liang, Y., Melchior, P., Lu, S., Goulding, A., & Ward, C. 2023b, AJ, 166, 75 Loshchilov, I., & Hutter, F. 2019, in ICLR Malkov, Y. A., & Yashunin, D. A. 2018, ITPAM, 42, 824 Melchior, P., Liang, Y., Hahn, C., & Goulding, A. 2023, AJ, 166, 74 Parker, L., Lanusse, F., Golkar, S., et al. 2024, MNRAS, 531, 4990 Portillo, S. K., Parejko, J. K., Vergara, J. R., & Connolly, A. J. 2020, AJ, 160, 45 Talbot, M. S., Brownstein, J. R., Dawson, K. S., Kneib, J.-P., & Bautista, J. 2021, MNRAS, 502, 4617 Teimoorinia, H., Archinuk, F., Woo, J., Shishehchi, S., & Bluck, A. F. 2022, AJ, 163, 71 Tojeiro, R., Heavens, A. F., Jimenez, R., & Panter, B. 2007, MNRAS, 381, 1252 Tojeiro, R., Percival, W. J., Heavens, A. F., & Jimenez, R. 2011, MNRAS, 413, 434 Vaswani, A., Shazeer, N., Parmar, N., et al. 2017, in NeurIPS, 30 Veilleux, S., & Osterbrock, D. E. 1987, ApJS, 63, 295 Wu, Y., Tao, Y., Fan, D., Cui, C., & Zhang, Y. 2024, MNRAS, 527, 2022 Yip, C.-W., Connolly, A., Berk, D. V., et al. 2004, AJ, 128, 2603 Zhang, K., Schlegel, D. J., Andrews, B. H., et al. 2019, ApJ, 883, 63

Source: ChinaXiv —Machine translation. Verify with original.