Modeling Diachronic Semantics in Classical Chinese via Multi-Scale Fourier Analysis

Xuanning Liu
School of Computer Science
Beijing University of Posts and Telecommunications
Beijing, China
liuxuanning@bupt.edu.cn

Bin Wu
School of Computer Science
Beijing University of Posts and Telecommunications
Beijing, China
wubin@bupt.edu.cn

***Abstract*—Modeling diachronic semantics in Classical Chinese is challenging due to the millennial span and the intricate balance between cultural continuity and linguistic shift. Existing methods often rely on discrete temporal labels, failing to capture the inherently continuous nature of semantic evolution. To bridge this gap, we propose the Multi-Scale Fourier Transformer (MSFT), which reformulates diachronic representation learning from a frequency-domain perspective. By decoupling textual features into modulus and phase components, MSFT utilizes the Semantic Inheritance Intensifier (SII) and Evolutionary Trajectory Gating (ETG) modules to independently model stable semantic inheritance and dynamic linguistic drift. Specifically, low-frequency signals capture long-term thematic stability, while high-frequency signals encode localized semantic innovations. A Cross-Scale Resolution Attention (CRA) mechanism further enables the model to adaptively integrate evolutionary patterns across different historical resolutions. Experiments on the Temporal CCL-WSD dataset demonstrate that MSFT consistently outperforms state-of-the-art baselines. Furthermore, qualitative analysis confirms the structural interpretability of our framework, showing that internal phase dynamics strongly correlate with external semantic distance, providing a principled methodology for quantifying the millennial metamorphosis of the Chinese language.**

***Index Terms*—Diachronic Semantic Representation, Fourier Transform, Semantic shift.**

I. Introduction

The three-millennium evolution of Classical Chinese represents a unique linguistic phenomenon where semantic shifts embody the interplay between cultural continuity and historical transformation. Unlike modern standardized languages, Classical Chinese undergoes profound semantic metamorphosis across dynasties. These changes manifest in four primary patterns: expansion (e.g., “河” expanding from the Yellow River to all rivers), narrowing (e.g., “禽” narrowing from all animals to birds), shifting (e.g., “钱” shifting from tools to currency), and emotional variation (e.g., “诽” shifting from neutral censure to derogatory slander) . Such diachronic fluidity poses significant challenges for automated understanding.

Pre-training techniques have substantially advanced textual semantic understanding by learning co-occurrence patterns from large-scale corpora [3], [11]. While some models have been adapted for Classical Chinese through fine-tuning on historical texts [18], [20], [21], they typically follow a static paradigm. By treating all texts as temporally homogeneous, these models overlook the dynamic evolution across eras, limiting their ability to capture diachronic shifts.

Diachronic word representation has emerged to address this by explicitly incorporating temporal signals. One prevalent approach aligns separate embedding spaces trained on different time intervals [6], [15], [17], which often entails high computational complexity. Another approach directly integrates temporal embeddings or time-masking strategies into Transformer architectures [8], [9], [23]. However, these methods frequently discretize history into coarse-grained dynastic labels, failing to capture the continuous nature of semantic evolution and the underlying mechanisms of inheritance and dissemination [4], [13]. Consequently, existing models struggle to balance era-stable thematic continuity with era-specific semantic innovations.

To address these limitations, we propose the **M**ulti-**S**cale **F**ourier **T**ransformer (**MSFT**), which reformulates diachronic modeling from a frequency-domain perspective. In this framework, frequency reflects the rate of semantic change: low-frequency signals capture stable thematic continuity across eras, while high-frequency signals encode rapid contextual or stylistic shifts. MSFT decouples textual features into modulus and phase, employing the **S**emantic **I**nheritance **I**ntensifier (**SII**) and **E**volutionary **T**rajectory **G**ating (**ETG**) modules to independently model semantic inheritance and dynamic evolution. Furthermore, a **C**ross-Scale **R**esolution **At**tention (**CRA**) mechanism is introduced to adaptively integrate multi-scale spectral features across different historical resolutions .

Our main contributions are as follows:

- We propose a frequency-domain model that decouples textual features into modulus and phase components, utilizing **SII** and **ETG** modules to capture both stable thematic inheritance and dynamic semantic evolution.
- We introduce an attention mechanism that adaptively integrates representations across diverse historical resolutions to track continuous semantic drift .
- We provide a comprehensive interpretability analysis of frequency-domain representations, empirically demonstrating that MSFT's internal phase dynamics closely align with recognized linguistic shifts.

This work was supported by the National Natural Science Foundation of China (Grant No. 62372060).

II. Related Work

A. Pre-training for Classical Chinese

Mainstream Chinese pre-trained language models are primarily developed on modern corpora, which significantly constrains their performance on classical Chinese texts. To bridge this linguistic and stylistic gap, recent studies have fine-tuned general models on extensive historical datasets. Representative efforts include AnchiBERT [18], trained on texts from 1000 BCE to 200 CE; SikuBERT [21], built upon the *Si Ku Quan Shu* corpus; and GujiRoBERTa [20], which integrates over 13,000 classical works from the *Dai Zhi Ge* collection. Beyond corpus-level adaptation, research has explored modeling linguistic divergence through structured knowledge graphs [22], contrastive learning on parallel corpora [24], and syntactic tree-based analyses [27]. However, these methods largely adhere to static pre-training paradigms that overlook the inherently dynamic and diachronic nature of classical Chinese, thereby limiting their ability to model long-term semantic evolution.

B. Diachronic Word Representation

Diachronic word representation aims to capture subtle meaning variations across historical periods by explicitly incorporating temporal information. Existing approaches generally follow two paradigms: the first trains independent embeddings for separate time intervals and aligns them [5], [7], [15], which effectively reveals semantic drift but entails high computational complexity ; the second integrates temporal signals directly into the architecture via specialized temporal functions or embeddings [12], [19], which relies on discretized time labels, failing to capture intrinsic multi-scale temporal dynamics. With the rise of Transformers, several studies have introduced time-masking strategies or learnable temporal parameters into self-attention mechanisms [4], [13], [14]. More recently, models such as DPLM [23] and Ticktack [9] have been tailored to trace semantic continuity across adjacent historical periods. Despite these advances, most methods rely on discretized temporal labels, which forces semantics into predefined eras. Such discretization undermines the continuous nature of semantic evolution and weakens the model's capacity to capture intrinsic temporal dynamics.

C. Fourier Transform

The Fourier transform has gained prominence in deep learning for its ability to capture global and multi-scale representations efficiently. Frequency-domain features possess broad receptive fields and model long-range dependencies by leveraging the convolution theorem, where convolution in the time domain corresponds to element-wise multiplication in the frequency domain. Building on this, FNet [10] replaced self-attention with the Fast Fourier Transform (FFT) to reduce complexity, while FFC [2] utilized frequency-domain perception for efficient convolutional modules. Although Fourier-based features have shown success in other domains like image restoration [28] and sequential recommendation [25], their potential in natural language processing, particularly for modeling the multi-scale metamorphosis of diachronic semantics, remains insufficiently explored.

III. Method

A. Method Overview

The MSFT architecture aims to disentangle stable semantic inheritance from dynamic evolution by modeling textual features in the frequency domain. As illustrated in Figure 1, the model consists of two primary components: **Spectral Evolutionary Block**(left) for spectral feature extraction and transformation and a **Cross-Scale Resolution Attention**(right) for adaptive feature fusion. Given an input sentence $X = [x_1, x_2, \ldots, x_n]$, the model generates multi-scale representations through stacked layers and dense feature connections, allowing for a comprehensive capture of diachronic patterns across different historical resolutions.

B. Spectral Evolutionary Block

The core of MSFT lies in treating semantic evolution as a signal processing task. In this view, frequency reflects the rate of semantic change: low-frequency signals capture stable thematic continuity, while high-frequency signals encode rapid localized shifts or contextual innovations.

For each word vector x_i, we first apply the Fast Fourier Transform (FFT) to convert it into the frequency domain, decomposing the signal into modulus (M) and phase (ϕ) components with decomposition function $\mathbb{D}$:

$$M_i, \phi_i = \mathbb{D}(FFT(x_i)) \tag{1}$$

To refine these two components, we propose two specialized modules:

1. **Semantic Inheritance Intensifier (SII)**: The modulus M characterizes the energy distribution of semantic components, representing the stable inheritance of language. To reinforce these era-invariant features and filter out temporal noise, we employ the SII module as $\hat{M} = \sigma(W_s * M)$, where W_s is a scaling weight that prioritizes dominant semantic signals, and σ is activation function.

2. **Evolutionary Trajectory Gating (ETG)**: The phase ϕ captures the orientation and drift of semantic vectors, providing a principled way to represent linguistic variation. The ETG unit refines these trajectories by modeling directional shifts $\phi_S = \phi + \delta$ and semantic reversals $\phi_R = -\phi$, computing the final phase via $\hat{\phi} = Gate(\phi, \phi_R, \phi_S)$. This gating mechanism allows the model to capture both gradual drift (e.g., sense expansion) and abrupt semantic metamorphosis (e.g., emotional variation).

The enhanced components are reconstructed via the Inverse Fast Fourier Transform (IFFT) to obtain the refined textual feature $\hat{c}_i$. To prevent vanishing gradients and share information globally, we implement dense feature connections between layers:

$$c_i = IFFT(\mathbb{C}(\hat{M}_i, \hat{\phi}_i)) \tag{2}$$

$$\hat{x_i} = \sigma(c_i + x_i) \tag{3}$$

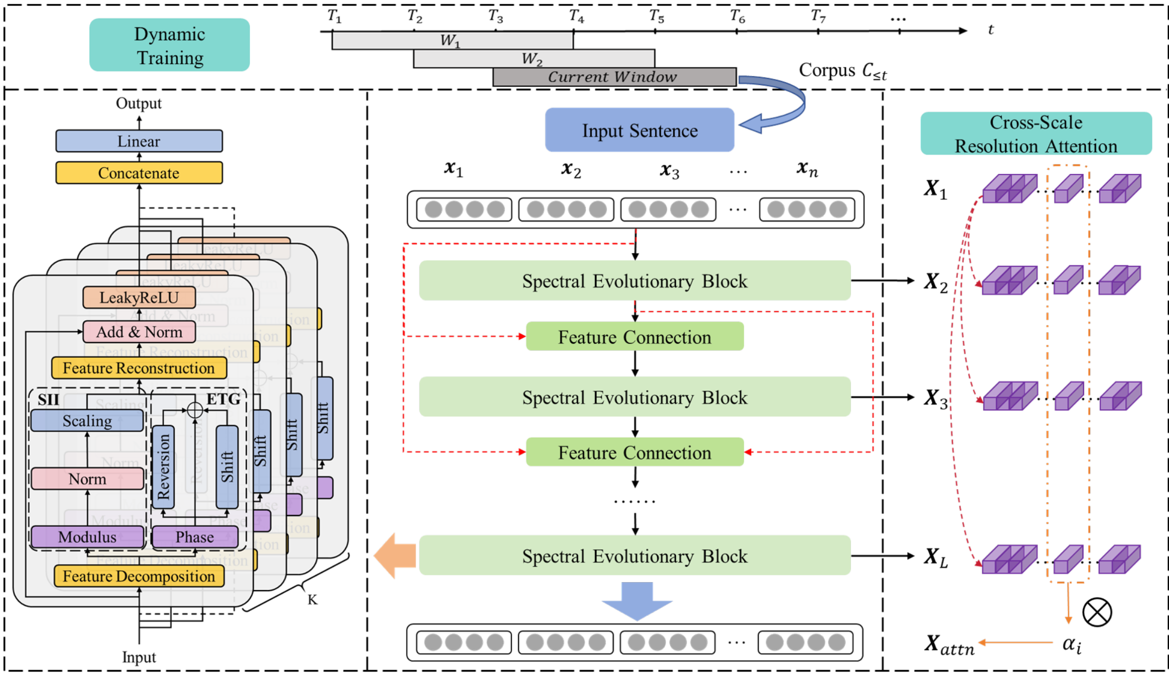

Fig. 1: Framework of MSFT. Red dotted lines indicate the feature connections between layers. Left part is the illustration of **S**pectral **E**volutionary **B**lock(**SEB**). Right part is the illustration of **C**ross-Scale **R**esolution **A**ttention(**CRA**).

where $\mathbb{C}$ is reconstruction function.

In order to capture more and richer frequency domain semantic information, we set up a multi-head mechanism in each layer. Each head can decompose, transform and reconstruct features. For head h, we got reconstructed feature X_h. And for all heads, we concatenated outputs of each head to get the final output of the l-th Spectral Evolutionary Block:

$$\hat{X}_l = W_l * cat(X_l^1, X_l^2, ..., X_l^H) + B_l \tag{4}$$

where $cat(\cdot)$ means concatenation operation, W_l and B_l are learnable weight matrices.

C. Cross-Scale Resolution Attention

While feature connections allow downstream layers to access upstream information, the model must effectively utilize these features across different scales. We propose a Cross-Scale Resolution Attention mechanism to adaptively integrate these multi-scale features. Specifically, for each layer l, the mechanism computes a resolution indicator s_i^l that reflects the spectral response at position i:

$$s_i^l = \frac{1}{d} \sum_{j=1}^{d} |\boldsymbol{x}_i^l(j)| \tag{5}$$

$$\boldsymbol{\alpha}_i = softmax(\omega^\top \boldsymbol{s_i}) \tag{6}$$

where ω denotes a learnable scoring vector that projects the hidden gate feature into a scalar layer score, which is shared across layers.

The final representation x_i^{atten} is a weighted integration of features from L layers, each representing a distinct historical scale:

$$\boldsymbol{x}_i^{atten} = \sum_{l=1}^{L} \alpha_i^l \boldsymbol{x}_i^l \tag{7}$$

$$\boldsymbol{X}_{atten} = [\boldsymbol{x}_1^{atten}, \boldsymbol{x}_2^{atten}, ..., \boldsymbol{x}_n^{atten}] \tag{8}$$

This design ensures that the model can dynamically focus on long-term thematic stability or localized semantic innovations depending on the linguistic context.

D. Dynamic Pre-training Mechanism

To bridge the gap between discrete temporal encoding and the continuous nature of linguistic change, we implement a dynamic pre-training strategy centered on Incremental Knowledge Transfer. Unlike traditional methods that treat historical periods as isolated buckets, our approach ensures semantic continuity through two key mechanisms:

1. **Overlapping Temporal Sliding Windows**: We define a sequence of overlapping windows $\{W_1, W_2, \ldots, W_T\}$. Each window W_t covers a specific duration with a predefined stride, ensuring that adjacent windows share a portion of the corpus. This overlap allows the MSFT to capture the gradual transition of word meanings, effectively modeling the fluid metamorphosis of semantics across long-term spans.

2. **Temporal Causal Pre-training**: We adopt a multi-stage training paradigm governed by a temporal causal constraint, where stage t only accesses corpus $\mathcal{C}_{\leq t}$. Parameters from stage $t-1$ serve as the initialization for stage t, forcing the model

to encode semantic trajectories based on cumulative historical context.

IV. Experiments

A. Experiment Settings

We evaluated MSFT on the Temporal CCL-WSD dataset, which covers ancient to modern Chinese periods with approximately 70 million characters. Following the procedures outlined in [16], [23], we apply disambiguation thresholds ranging from 2 to 10 to address sense sparsity. Each threshold specifies the minimum number of sentences required for a sense to be included in the evaluation, where smaller values correspond to more challenging settings. For thresholds ≤ 4, we use a single-instance test set; for ≥ 5, an 8:2 training-to-testing ratio is applied.

B. Implementation Details

We implement MSFT in PyTorch, initialized with SikuRoBERTa ($d = 768$). Based on validation, we configure 6 SEBs with 8 Fourier heads each. Our dynamic pre-training employs a stage-wise paradigm, where stage t accesses a localized window $\{D_{t-2}, D_{t-1}, D_t\}$. The model is optimized on 4 NVIDIA 2080Ti GPUs using AdamW ($lr = 2 \times 10^{-5}$, batch size 64, weight decay 0.01).

Importantly, while the corpus is partitioned by dynasties, we introduce no explicit temporal embeddings. This forces the model to capture diachronic evolution entirely through frequency-domain features, significantly distinguishing MSFT from methods relying on discrete time markers.

C. Baseline Models

We compare MSFT against three categories of baselines. For **Classic Chinese PLMs**, we evaluate SikuRoBERTa [21] and GujiRoBERTa [20], which are trained on massive ancient corpora. For **Diachronic Representation Models**, we include TempoBERT [13], Temporal Attention [14], T-TempoBERT [17], and DPLM [23], which integrate temporal signals into Transformers. Finally, for **LLMs**, we evaluate SikuGPT [1] and Qwen2.5-7B [26] using zero-shot and LoRA fine-tuning.

D. Quantitative Results

As shown in Table I, where MSFT consistently outperforms all baselines on the Temporal CCL-WSD dataset. Notably, at Threshold 10, it surpasses the strongest diachronic baseline (DPLM) by over 1.0%. Such robustness originates from decoupling features into modulus and phase via the SII and ETG modules. By integrating these multi-scale patterns through the CRA mechanism, MSFT jointly captures stable inheritance and dynamic drift.

In contrast, static PLMs perform less competitive, underscoring the necessity of temporal awareness. By treating corpora as temporally static, they fail to disentangle "temporal noise", introducing distortions that severely weaken their capacity to track semantic evolution.

Among diachronic baselines, TempoBERT achieves solid results via explicit time tokens, while Temporal Attention struggles to capture continuous, multi-scale semantic shifts. T-TempoBERT's focus on short-term adaptation often overlooks early-stage semantics due to its emphasis on recent data. Conversely, DPLM achieves more balanced performance across all historical periods by explicitly integrating temporal information.

Finally, LLM baselines exhibit suboptimal average performance, unexpectedly lagging behind specialized static PLMs. This indicates that merely scaling up parameters or employing LoRA fine-tuning is insufficient for resolving complex diachronic shifts. Without explicit continuous modeling, LLMs rely heavily on surface-level distributional cues, leaving them susceptible to temporal noise.

E. Ablation Study

To evaluate the impact of modules SEB, SII, ETG, and dynamic pre-training strategy, we conduct ablation studies at $threshold = 10$. We compare five model variants:

MSFT w/o multi-head: Replaces the multi-head SEB with a single-head Fourier Transformation block.

MSFT w/o SII: Removes the scaling processes within the SII module.

MSFT w/o ETG: Replaces phase gating mechanism with a standard linear filtering layer in ETG module.

MSFT w/o connection: Eliminates dense feature connections between SEB layers.

MSFT w/o dynamic: Employs static pre-training on the entire corpus rather than our multi-stage strategy.

The results in Table II reveal several critical insights:

(1) Removing multi-head extraction leads to a significant 3.83% drop, confirming that integrating features from diverse frequency bands is essential for robust text modeling in the frequency domain.

(2) Accuracy declines upon removing either SII or ETG, indicating that both modulus and phase are vital for capturing diachronic semantics . Notably, SII's removal causes a larger drop, suggesting that stabilizing semantic inheritance (modulus) is slightly more influential in Classical Chinese modeling than tracking trajectory drift (phase).

(3) The performance loss without dense connections demonstrates their role in sharing multi-scale spectral information across layers to enhance global dependencies.

(4) Eliminating the dynamic strategy essentially reverts MSFT to a static model. Its success confirms that our multi-stage approach effectively captures the implicit temporal evolution required for long-term semantic understanding.

F. Interpretability and Qualitative Analysis

To validate the structural interpretability of our framework, we performed a detailed qualitative analysis focusing on the decoupling of diachronic semantic stability and evolution.

Experimental Setup and Metrics Our evaluation aims to determine whether the magnitude and phase components within MSFT correspond to established linguistic shifts over a 3000-year span (from Zhou Dynasty to Qing Dynasty). We selected two representative word groups:

model	Thresholds									
	2	3	4	5	6	7	8	9	10	Avg
SikuRoBERTa	66.63	68.15	70.91	72.62	73.66	75.07	75.80	77.02	77.05	72.99
GujiRoBERTa	70.21	72.52	74.41	75.95	77.21	77.40	80.15	79.97	80.60	76.49
TempoBERT	71.00	72.64	75.36	75.68	78.11	78.20	80.24	80.53	81.26	77.00
TempoAtt	65.76	67.39	70.53	71.61	73.22	73.19	75.72	75.94	76.66	72.22
T-TempoBERT	71.27	73.28	75.85	75.66	78.30	78.57	80.09	80.66	81.35	77.23
DPLM	**71.58**	73.49	75.91	76.34	**78.73**	78.94	80.54	81.23	81.46	77.58
SikuGPT	66.75	67.31	68.59	70.60	73.57	75.22	76.84	78.44	80.60	73.10
Qwen	65.67	65.73	66.34	67.91	70.13	74.80	77.58	80.33	81.97	72.34
MSFT	71.04	**73.68**	**76.17**	**76.72**	78.17	**79.65**	**81.08**	**82.35**	**82.47**	**77.93**

TABLE I: Overall performance on Temporal CCL-WSD with different thresholds. Bold and underlined results indicate the optimal and sub-optimal results, respectively.

Model	Acc
MSFT w/o multi-head	78.64
MSFT w/o SII	80.92
MSFT w/o ETG	81.26
MSFT w/o connection	79.88
MSFT w/o dynamic	79.70
MSFT	82.47

TABLE II: Ablation experiment results.

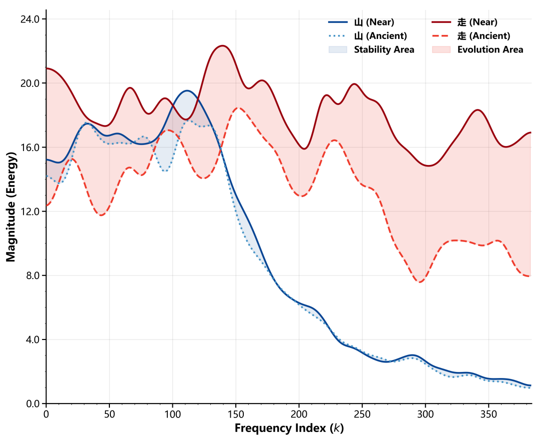

Fig. 2: Magnitude spectra of stable and evolving terms.

- **Stable Group**: Terms with highly consistent meanings across history, typified by “山”(mountain), to serve as the control baseline.
- **Evolving Group**: Terms that underwent significant semantic metamorphosis, specifically “走”(ancient: to run; modern: to walk), to analyze the model’s sensitivity to semantic drift.

To provide a solid empirical grounding for our interpretability analysis, we designed a multi-dimensional experimental framework that aligns model-internal frequency dynamics with established linguistic baselines. The following metrics were utilized to quantify semantic metamorphosis from the frequency-domain perspective:

- **Average Spectral Magnitude** (M): For a given word, we extract the magnitude component M from the output of the Semantic Inheritance Intensifier (SII) module. This metric measures the distribution of semantic energy across different frequency indices.

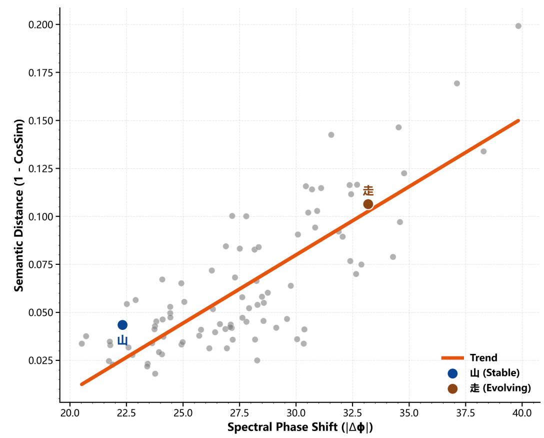

Fig. 3: Correlation between the internal phase shift and external semantic distance.

- **Internal Phase Drift** ($|\Delta\phi|$): The L_2 norm of the difference between the mean phase vectors $\overline{\phi}$ of two distinct historical periods: $|\Delta\phi| = ||\overline{\phi}_1 - \overline{\phi}_2||_2$. This metric explicitly measures the phase drift of a word at two different time points, capturing how far the representation of a word has shifted from its original state.
- **External Semantic Distance** (D_{sem}): The semantic distance based on the cosine similarity of the word embeddings given in two distinct historical periods: $D_{sem} = 1 - \frac{E_1 \cdot E_2}{||E_1||||E_2||}$. This metric intuitively illustrates the semantic changes of a particular word during this time period.

To facilitate a distinct contrastive analysis of semantic evolution, we sampled corpora from the representative temporal endpoints (here we chose the earliest and latest historical periods) for each target word. This sampling strategy aims to capture the maximal diachronic trajectory and highlight the divergence in frequency-domain representations.

Analysis of Spectral Evolution Figure 2 illustrates the temporal spectral energy distribution for the selected samples. For the stable term “山”, the magnitude spectra of two periods exhibit nearly perfect overlap, resulting in a negligible Stability Area. In contrast, the evolving term “走” shows a marked energy surge in the mid-to-high frequency bands ($k > 150$), forming a significant Evolution Area. These findings suggest that MSFT successfully disentangles semantic components. Stable terms maintain their energy in low-frequency bands, representing thematic inheritance, while evolving terms generate "additional energy" in high-frequency bands, which quantifies the increased contextual complexity and semantic innovation over time.

Correlation Analysis To verify the structural interpretability of the frequency-domain components, we analyze the correlation between the internal phase shift $|\Delta\phi|$ and the external semantic distance D_{sem}. As illustrated in Fig. 3, the scatter plot, comprising the 100 most high-frequency words from the corpus, exhibits a significant positive linear correlation between these two variables. Notably, the term "山", representing semantic stability, exhibits minimal phase displacement and restricted semantic distance, thus situating it within the low-drift/low-distance region in the bottom-left quadrant. In contrast, the term "走" displays a substantial semantic metamorphosis, resulting in its distribution in the high-drift/high-distance region in the top-right quadrant. These results empirically demonstrate that the ETG's phase-gating mechanism effectively captures temporal linguistic shifts by mapping abstract semantic changes to observable phase rotations.

V. Conclusion

We introduced MSFT to tackle the challenge of modeling diachronic semantics in Classical Chinese. By decoupling textual features into modulus and phase, our framework systematically isolates stable thematic inheritance from dynamic linguistic drift and adaptively integrates multi-scale historical dynamics. Empirically, this continuous frequency-domain approach yields strong predictive performance and structural interpretability, where internal phase drift directly correlates with historical semantic distances. Future work will extend these diachronic features to complex ancient Chinese NLP tasks, such as historical event extraction and evolutionary knowledge graph reasoning.

References

[1] Chang, L., Dongbo, W., Zhixiao, Z., Die, H., Mengcheng, W., Litao, L., Si, S., Bin, L., Jiangfeng, L., Hai, Z., et al.: Sikugpt: A generative pre-trained model for intelligent information processing of ancient texts from the perspective of digital humanities. arXiv preprint arXiv:2304.07778 (2023)

[2] Chi, L., Jiang, B., Mu, Y.: Fast fourier convolution. Advances in Neural Information Processing Systems **33**, 4479–4488 (2020)

[3] Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers). pp. 4171–4186 (2019)

[4] Dhingra, B., Cole, J.R., Eisenschlos, J.M., Gillick, D., Eisenstein, J., Cohen, W.W.: Time-aware language models as temporal knowledge bases. Transactions of the Association for Computational Linguistics **10**, 257–273 (2022)

[5] Di Carlo, V., Bianchi, F., Palmonari, M.: Training temporal word embeddings with a compass. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 6326–6334 (2019)

[6] Drinkall, F., Rahimikia, E., Pierrehumbert, J., Zohren, S.: Time machine gpt. In: Findings of the Association for Computational Linguistics: NAACL 2024. pp. 3281–3292 (2024)

[7] Dubossarsky, H., Hengchen, S., Tahmasebi, N., Schlechtweg, D.: Time-out: Temporal referencing for robust modeling of lexical semantic change. In: Annual Meeting of the Association for Computational Linguistics. pp. 457–470. ACL (2019)

[8] Goel, R., Kazemi, S.M., Brubaker, M., Poupart, P.: Diachronic embedding for temporal knowledge graph completion. In: Proceedings of the AAAI conference on artificial intelligence. vol. 34, pp. 3988–3995 (2020)

[9] Han, X., Hu, Q., Wang, Y., Gao, W., Zhang, L., Wang, Q., Mei, L., Deng, C., Feng, J.: Ticktack: Long span temporal alignment of large language models leveraging sexagenary cycle time expression. arXiv preprint arXiv:2503.04150 (2025)

[10] Lee-Thorp, J., Ainslie, J., Eckstein, I., Ontanon, S.: Fnet: Mixing tokens with fourier transforms. In: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 4296–4313 (2022)

[11] Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al.: Improving language understanding by generative pre-training (2018)

[12] Rosenfeld, A., Erk, K.: Deep neural models of semantic shift. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). pp. 474–484 (2018)

[13] Rosin, G.D., Guy, I., Radinsky, K.: Time masking for temporal language models. In: Proceedings of the fifteenth ACM international conference on Web search and data mining. pp. 833–841 (2022)

[14] Rosin, G.D., Radinsky, K.: Temporal attention for language models. In: Findings of the Association for Computational Linguistics: NAACL 2022. pp. 1498–1508 (2022)

[15] Rudolph, M., Blei, D.: Dynamic embeddings for language evolution. In: Proceedings of the 2018 world wide web conference. pp. 1003–1011 (2018)

[16] Shu, L., Guo, Y., Wang, H., Zhang, X., Hu, R.: The construction and application of ancient chinese corpus with word sense annotation(in chinese). In: Proceedings of the 20th Chinese national conference on computational linguistics. pp. 549–563 (2021)

[17] Tang, X., Zhou, Y., Bollegala, D.: Learning dynamic contextualised word embeddings via template-based temporal adaptation. In: The 61st Annual Meeting Of The Association For Computational Linguistics (2023)

[18] Tian, H., Yang, K., Liu, D., Lv, J.: Anchibert: A pre-trained model for ancient chinese language understanding and generation. In: 2021 International Joint Conference on Neural Networks (IJCNN). pp. 1–8. IEEE (2021)

[19] Wang, B., Di Buccio, E., Melucci, M.: Word2fun: Modelling words as functions for diachronic word representation. Advances in Neural Information Processing Systems **34**, 2861–2872 (2021)

[20] Wang, D., Liu, C., Zhao, Z., Shen, S., Liu, L., Li, B., Hu, H., Wu, M., Lin, L., Zhao, X., et al.: Gujibert and gujigpt: Construction of intelligent information processing foundation language models for ancient texts. arXiv preprint arXiv:2307.05354 (2023)

[21] Wang, D., Liu, C., Zhu, Z., Liu, J., Hu, H., Shen, S., Li, B.: Construction and application of pre-trained models of *Siku Quanshu* in orientation to digital humanities. Library Tribune **42**(6), 30–43 (6 2022)

[22] Wang, Y., Wang, J., Zhao, D., Zheng, Z.: Rethinking dictionaries and glyphs for chinese language pre-training. In: Findings of the association for computational linguistics: ACL 2023. pp. 1089–1101 (2023)

[23] Wei, Y., Li, M., Zhu, Y., Xu, Y., Li, Y., Wu, B.: A diachronic language model for long-time span classical chinese. Information Processing & Management **62**(1), 103925 (2025)

[24] Xiang, J., Liu, M., Li, Q., Qiu, C., Hu, H.: A cross-guidance cross-lingual model on generated parallel corpus for classical chinese machine reading comprehension. Information Processing & Management **61**(2), 103607 (2024)

[25] Xu, L., Tian, Z., Li, B., Zhang, J., Wang, D., Wang, H., et al.: Sequence-level semantic representation fusion for recommender systems. In: Proceedings of the 33rd ACM International Conference on Information and Knowledge Management. pp. 5015–5022 (2024)

[26] Yang, A., Yang, B., Hui, B., Zheng, B., Yu, B., Zhou, C., et al.: Qwen2 technical report. arXiv preprint arXiv:2407.10671 (2024)

[27] Zhang, S., Wang, P., Li, Z., Hou, J., Hu, Q.: Confidence-based syntax encoding network for better ancient chinese understanding. Information Processing & Management **61**(3), 103616 (2024)

[28] Zhou, M., Huang, J., Guo, C.L., Li, C.: Fourmer: An efficient global modeling paradigm for image restoration. In: International conference on machine learning. pp. 42589–42601. PMLR (2023)