

Bridging the Temporal Gap: A New Benchmark and Dynasty-aware Adaptation Framework for Ancient Chinese WSD

Yuzhi Liang* Xinrong Zhu Lixiang Ma Zhixiong Wan

School of Information Science and Technology

Guangdong University of Foreign Studies

yzliang@gdufs.edu.cn

{2022100319, 20241050018, 20251010002}@mail.gdufs.edu.cn

Abstract

Word sense disambiguation (WSD) is typically studied in synchronic settings, where the mapping from contextualized occurrences to sense labels is assumed to be temporally stable. This assumption may not hold for historical texts, in which meanings may emerge, shift, or disappear over time. We study diachronic WSD for Ancient Chinese, a challenging scenario involving long-term semantic evolution and substantial historical variation. We introduce Ancient Chinese Diachronic Word Sense Disambiguation Benchmark (AC-DIAWSD), a large-scale WSD dataset constructed from the *Twenty-Four Histories*, containing about 24K manually annotated instances across 12 dynasties from the Western Han to the Qing. We further propose DynaSense, a model-agnostic framework that enhances pre-trained language models with contrastive sense learning and dynasty-conditioned representation adaptation. Experiments with multiple PLMs and LLMs show that existing models struggle with diachronic WSD, while DynaSense consistently improves accuracy and F1 across backbones.

1 Introduction

Word sense disambiguation is a fundamental NLP task and an important component for applications such as knowledge graph construction, machine translation, and information extraction (Navigli, 2026; Simos and Makris, 2026; Cao et al., 2024). Recent PLM-based WSD systems achieve strong performance on standard benchmarks (Zhang and Li, 2025). Most of the WSD systems, however, are trained and evaluated on corpora treated as synchronic: the temporal origin of an occurrence is either unavailable or not explicitly modeled. This limitation becomes important for historical texts, where word meanings may emerge, shift, or disappear over time (Beelen et al., 2021).

In this paper, we study *diachronic WSD* for Ancient Chinese. Unlike synchronic WSD, diachronic WSD requires identifying the historically appropriate sense of a word in a specific temporal context. For example, the character “走” *zou* usually means “to walk” or “to move” in Modern Chinese, but often means “to run” or “to flee” in Ancient Chinese. In “弃甲曳兵而走”, “走” should be interpreted as “to flee”, showing that disambiguation must be grounded in both local context and historical time.

Ancient Chinese presents a particularly challenging setting. Its textual tradition extends across a long historical period, during which lexical usage and syntax evolved alongside institutions, cultural contexts, and writing styles. Yet existing supervised WSD research offers limited support for this setting for two main reasons. First, there is no large-scale Ancient Chinese WSD benchmark that combines fine-grained, occurrence-level sense annotations with explicit dynasty metadata. Second, existing PLM-based WSD methods rarely model diachronic semantic change explicitly, making them vulnerable to low-frequency historical senses, genre-induced variation, and differences in sense distributions across dynasties.

To address these challenges, we introduce AC-DIAWSD¹, a large-scale diachronic WSD dataset for Ancient Chinese constructed from the *Twenty-Four Histories* (Wang et al., 2025a). It covers 12 dynasties from the Western Han to the Qing dynasty, spanning more than 2,000 years, and contains approximately 24K manually annotated instances, with around 2,000 instances per dynasty. Since single characters often function as lexical units in Ancient Chinese, we use target characters as the main disambiguation units and select representative polysemous characters from diverse se-

*Corresponding author.

¹The data and code are available at <https://github.com/PapersCode/DynaSense>.

semantic categories.

We further propose DynaSense, a model-agnostic framework for enhancing PLM-based diachronic WSD. DynaSense combines contrastive sense learning, which sharpens sense boundaries, with dynasty-conditioned representation adaptation, which injects period-specific semantic bias into the representation space. This design enables PLMs to model dynasty-specific sense variation while sharing information across periods with controlled parameter overhead.

Our contributions are summarized as follows:

- We introduce AC-DIAWSD, a large-scale Ancient Chinese diachronic WSD dataset with approximately 24K annotated instances across 12 dynasties and more than 2,000 years. To the best of our knowledge, it is the first Ancient Chinese WSD benchmark with explicit dynasty-level temporal information.
- We propose DynaSense, a model-agnostic framework that improves supervised PLM-based WSD by modeling diachronic semantic variation through contrastive sense learning and dynasty-conditioned representation adaptation.
- We conduct extensive experiments with PLMs and LLMs, establishing strong baselines for Ancient Chinese diachronic WSD. Results show that DynaSense consistently improves accuracy and F1 across different backbone models.

2 Related Work

2.1 Lexical Semantic Change Detection

Lexical Semantic Change Detection (LSCD) is closely related to diachronic WSD, as both concern the interaction between word meaning and historical context. Whereas diachronic WSD assigns predefined senses to individual occurrences using temporal information, LSCD typically determines whether, and to what extent, a lexeme’s semantics has changed across periods by comparing its usages in diachronic corpora (Schlechtweg et al., 2020). Existing methods can be grouped into three broad families. Representation-based approaches encode usages with static or contextual embeddings and measure distances between period-level representations (Cassotti et al., 2023; Periti and Tahmasebi, 2024). Sense-distribution approaches induce or annotate senses and quan-

tify changes in their prevalence across periods; diachronic sense annotations further show that improving word-sense induction also benefits LSCD (Tang et al., 2023; Schlechtweg et al., 2025). Interpretable approaches generate definitions or lexical substitutes for individual usages and compare the resulting semantic descriptions (Giulianelli et al., 2023; Periti et al., 2024).

2.2 WSD benchmarks

Early WSD research was largely shaped by the Senseval and SemEval campaigns, which standardized sense annotation and established lexical-sample and all-words task settings (Raganato et al., 2017). Most modern WSD benchmarks are built on WordNet and use SemCor as the primary training corpus (Agirre et al., 2004; Edmonds and Cotton, 2001). Although these benchmarks serve as de facto standards for English and have been extended to multilingual settings, e.g., XL-WSD (Pasini et al., 2021), they are generally formulated in synchronic settings, where sense inventories and sense distributions are treated as relatively stable and diachronic semantic change is not explicitly modeled.

Chinese WSD resources remain limited and mainly target modern texts. Ancient Chinese presents additional challenges due to its concise expressions, flexible syntax, and substantial historical variation (Wei et al., 2025). Shu et al. (2021) developed a valuable sense-annotated corpus for Ancient Chinese and demonstrated its application to BERT-based WSD. However, this resource was not designed for systematic cross-dynasty evaluation with explicit temporal metadata. ChiWUG provides a graph-based evaluation dataset for Chinese LSCD (Chen et al., 2023), but focuses on lexeme-level change in modern Chinese rather than token-level sense assignment in historical texts. Existing resources therefore leave a clear need for an Ancient Chinese WSD benchmark with explicit and systematic dynasty-level coverage.

2.3 WSD Methods

WSD research has evolved from traditional supervised methods to PLM- and LLM-based approaches. Early statistical methods rely heavily on large annotated corpora and struggle with diachronic semantic shifts (Iacobacci et al., 2016; Masethe et al., 2024). PLMs are now the dominant paradigm: models such as GlossBERT (Huang et al., 2019) and ancient Chinese encoders, includ-

ing GuwenBERT and SikuRoBERTa, improve semantic representation but generally assume static word meanings and transfer poorly across historical periods (Arrington et al., 2025; Ghinassi et al., 2024). Recent LLM-based methods show few-shot potential through prompting and multi-agent collaboration (Sumanathilaka et al., 2025; Zhang et al., 2025), but their reliability in historical contexts remains underexplored (Mosolova et al., 2025).

Overall, existing WSD research mainly targets modern languages under static semantic assumptions. Ancient Chinese WSD still lacks systematic diachronic frameworks, making it important to compare PLM fine-tuning and LLM few-shot generalization under low-resource and diachronic settings.

3 The AC-DIAWSD Benchmark

3.1 Dataset Construction

Constructing a diachronic WSD dataset for Ancient Chinese is challenging due to compressed syntax, archaic usages, and subtle semantic shifts across long time spans. We build AC-DIAWSD through corpus preparation, target character selection, instance sampling, and human-AI collaborative annotation.

Corpus Preparation and Sampling. We derive the raw corpus from the *Twenty-Four Histories*, covering 12 dynasties from Western Han to Qing. To fit PLM input limits, we split the lengthy, unpunctuated Ancient Chinese texts into segments of at most 256 tokens.

Each instance consists of a text segment, a target character occurrence, and its dynasty label. Since one segment may contain multiple target characters, it can yield multiple WSD instances. We sample 2,000 target character instances per dynasty to ensure comparable temporal coverage.

To capture diverse semantic changes, we select 63 polysemous target characters from six linguistic categories: high-frequency content words, abstract function words, institutional governance terms, social and kinship terms, modal and evaluative words, and counter-intuitive archaic words. The same target character set is used across all 12 dynasties, and each target character has samples in every dynasty. Sense inventories are obtained from the Ancient Chinese Word Sense Annotated Corpus (Shu et al., 2021). Details are provided in Table 8 in Appendix A.

Human-AI Collaborative Annotation. We adopt a three-stage human-AI annotation pipeline. First, TongGu translates Ancient Chinese segments into modern Chinese, and Qwen3-8B prepares candidate sense information for each target character instance based on the predefined sense inventory. Second, a prompted LLM performs initial WSD using the original segment, the generated translation, and the candidate senses. Third, two domain experts independently verify the LLM outputs and correct translation or sense-labeling errors. The annotators were members of our research team and received no additional compensation for annotation. Disagreements are resolved by a senior expert, and the adjudicated labels are used as the final annotations in AC-DIAWSD. Annotation prompts are provided in Appendix H.

3.2 Dataset Statistics

Table 1 presents the statistics of AC-DIAWSD across 12 dynasties. Each dynasty contains 2,000 target character instances. The same 63 target characters are used in every dynasty, and each target character has samples in each dynasty. Candidate sense inventories contain 3 to 35 senses per target character, with the average number of candidate senses ranging from 14.01 to 15.00 across dynasties.

3.3 Quality Evaluation

We evaluate annotation reliability using Cohen’s Kappa. Because WSD labels are target-specific and each target character has its own sense inventory (Artstein and Poesio, 2008; Navigli, 2009), we report Global Pooling Kappa (κ_{pool}) and Word-level Weighted Average Kappa (κ_{weighted}), with definitions in Appendix B.

As shown in Table 2, human experts achieve high agreement, with $\kappa_{\text{pool}} = 0.8766$ and $\kappa_{\text{weighted}} = 0.7976$, confirming the reliability of the final annotations. LLM predictions show lower agreement with gold labels, especially under the word-level metric, highlighting the need for expert verification in diachronic WSD annotation.

3.4 Diachronic Characteristics

We quantify diachronic semantic shift using Jensen-Shannon Divergence (JSD) between a target character’s sense distributions in the earliest dy-

Dynasty	Segments	Instances	Target Chars	Avg. Len	Min Len	Avg. Senses
Western Han	300	2000	63	200.88	156	14.01
Eastern Han	300	2000	63	197.83	95	14.20
Western Jin	297	2000	63	197.86	135	14.44
Liu Song	296	2000	63	200.13	159	14.50
Southern Liang	270	2000	63	198.44	151	14.56
Northern Qi	293	2000	63	200.74	69	15.00
Tang	295	2000	63	199.19	22	14.26
Later Jin	294	2000	63	199.55	49	14.45
Song	293	2000	63	198.59	21	14.42
Yuan	298	2000	63	197.75	25	14.64
Ming	291	2000	63	196.01	63	14.99
Qing	297	2000	63	201.91	22	14.76

Table 1: Dataset statistics across different dynasties. Each dynasty contains 2,000 target character instances and the same set of 63 target characters. Candidate sense inventories contain 3 to 35 senses per target character.

Comparison	N	κ_{pool}	κ_{weighted}
Human vs. Human	23,995	0.8766	0.7976
LLM vs. Gold	23,995	0.7716	0.6360

Table 2: Cohen’s Kappa scores for annotation agreement.

nasty P and the latest dynasty Q :

$$\text{JSD}(P \parallel Q) = \frac{1}{2}D_{\text{KL}}(P \parallel M) + \frac{1}{2}D_{\text{KL}}(Q \parallel M), \quad (1)$$

where $M = \frac{1}{2}(P + Q)$. JSD ranges from 0 to 1, with larger values indicating stronger distributional divergence.

Category	Average JSD
High-Frequency Content Words	0.1881
Abstract Function Words	0.0974
Institutional Governance Words	0.1529
Societal Kinship Terms	0.2236
Modal Evaluative Words	0.2203
Counter-Intuitive Archaic Words	0.1774

Table 3: Average JSD across different target character categories.

Table 3 shows that different target character categories exhibit different degrees of diachronic semantic shift. Societal kinship terms and modal evaluative words have the highest average JSD scores, suggesting stronger distributional changes over time, whereas abstract function words show the lowest average JSD and thus relatively stable sense distributions. Representative semantic-shift examples for the six target character categories and the overall JSD distribution are provided in Appendix C.

4 Dynasty-aware Contrastive Adaptation Framework

Although Transformer-based models have achieved strong performance in synchronic WSD (Saidi and Jarray, 2026; Yae et al., 2025), their application to diachronic WSD remains challenging. In historical texts, long-term semantic drift may cause the contextual embeddings of different senses of the same polysemous character to lie close together in the representation space. Such overlap makes it difficult for models to learn clear and stable decision boundaries across dynasties (Kober et al., 2017). Moreover, under low-resource settings, standard supervised learning often struggles to capture fine-grained sense distinctions.

To address these challenges, we propose DynaSense, a model-agnostic framework readily integrated into Transformer-based WSD models. It consists of two complementary components. Hypernetwork-based Dynasty Adaptation (HDA) injects dynasty-conditioned adaptation vectors to capture macro-level temporal and domain shifts. Diachronic Contrastive Alignment (DCA) uses a contrastive objective to improve fine-grained sense separation. The overall process of DynaSense is demonstrated in Figure 1.

Given an input instance (x_i, w_i, c_i) , where x_i denotes the context, w_i is the target character, and $c_i \in \mathcal{C}$ is the dynasty label, a backbone PLM first encodes the context and produces a contextual representation $h_i \in \mathbb{R}^m$ for the target character. When the target character corresponds to multiple tokens, we obtain h_i by pooling over the target span. During training and inference, the dynasty label c_i is used by HDA to obtain the dynasty-adapted representation. DCA is applied only during training as

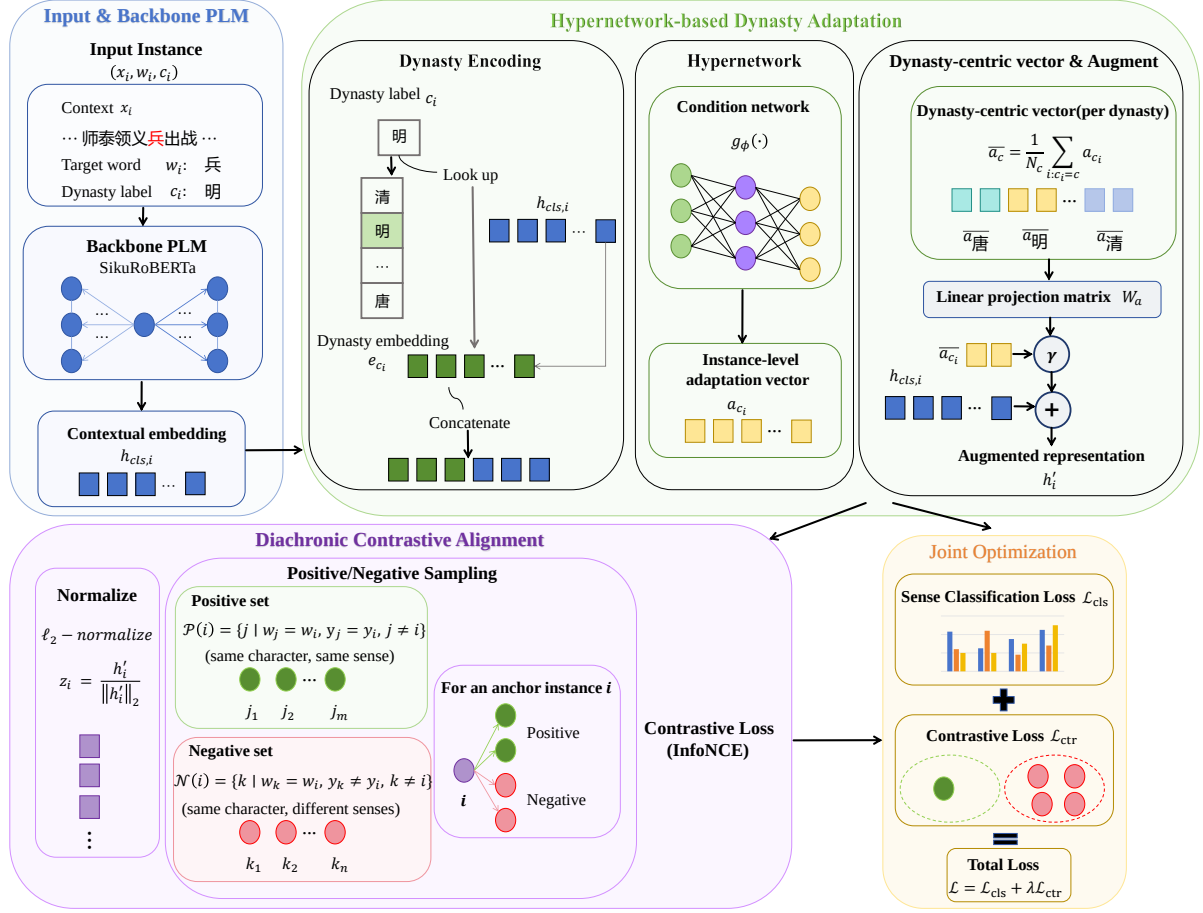


Figure 1: Overview of DynaSense. A backbone PLM produces a target character representation, which is augmented with a hypernetwork-generated dynasty adaptation vector. The model is trained with both sense classification and diachronic contrastive alignment.

an auxiliary objective.

4.1 Hypernetwork-based Dynasty Adaptation

Standard PLM-based WSD models usually lack explicit temporal conditioning, and therefore tend to map the same character from different historical periods into a shared static representation space. This makes it difficult to model dynasty-specific sense distributions. To mitigate this issue, DynaSense treats different dynasties as distinct semantic domains and uses a hypernetwork to generate dynasty-conditioned adaptation vectors. These vectors are then injected into the PLM representation space to compensate for dynasty-specific semantic bias.

Specifically, each dynasty label c_i is mapped to a learnable condition embedding e_{c_i} . We concatenate this condition embedding with the contextual representation h_i , and feed the result into a shared hypernetwork $g_\phi(\cdot)$ to obtain an instance-

level adaptation vector:

$$a_i = g_\phi([e_{c_i}; h_i]), \quad (2)$$

where $[\cdot]$ denotes vector concatenation.

To obtain a stable dynasty-level adaptation signal, we aggregate instance-level adaptation vectors within each dynasty:

$$\bar{a}_c = \frac{1}{N_c} \sum_{i:c_i=c} a_i, \quad (3)$$

where N_c is the number of training instances from dynasty c . After training the hypernetwork, we compute the instance-level adaptation vectors for all training instances and aggregate them by dynasty to obtain a fixed dynasty-level vector $\bar{a}_c$. The resulting dynasty-level adapters are pre-computed and stored for downstream WSD training and inference.

We then inject the corresponding dynasty adaptation vector into the original PLM representation

through a residual connection:

$$h'_i = h_i + \gamma W_a \bar{a}_{c_i}, \quad (4)$$

where W_a is a linear projection matrix and γ is a learnable gating coefficient that controls the strength of dynasty-specific adaptation. The resulting representation h'_i preserves the general semantic knowledge encoded by the PLM while incorporating dynasty-aware semantic information.

4.2 Diachronic Contrastive Alignment

Diachronic WSD often suffers from data sparsity and ambiguous decision boundaries between senses. To better structure the representation space, we introduce Diachronic Contrastive Alignment, which applies contrastive learning to the adapted representations. The intuition is to pull together instances that share the same target character and sense, while pushing apart instances of the same character of different senses. This encourages the model to learn clearer sense boundaries and reduces representation overlap between different senses.

We first apply ℓ_2 -normalization to the adapted representation:

$$z_i = \frac{h'_i}{\|h'_i\|_2}. \quad (5)$$

Let y_i denote the sense label of instance i . Ideally, for an anchor instance i , we define the positive set $\mathcal{P}(i)$ and the negative set $\mathcal{N}(i)$ as:

$$\begin{aligned} \mathcal{P}(i) &= \{j \mid w_j = w_i, y_j = y_i, j \neq i\}, \\ \mathcal{N}(i) &= \{k \mid w_k = w_i, y_k \neq y_i, k \neq i\}. \end{aligned} \quad (6)$$

Since DCA operates on the dynasty-adapted representations produced by HDA, the contrastive objective is applied in a representation space that already incorporates dynasty-specific information.

For computational efficiency and training stability, we randomly sample at most M_p positive instances and M_n negative instances for each anchor. Anchors without valid positive instances are excluded from the contrastive objective. We further analyze the alternative positive and negative sampling strategies in Appendix G.

We then use an InfoNCE-style objective on the normalized representations. For compactness, we define the temperature-scaled similarity as

$$s_{i,j} = \exp(\langle z_i, z_j \rangle / \tau), \quad (7)$$

where τ is the temperature hyperparameter. For each positive pair (i, j) , the pairwise contrastive loss is

$$\ell_{i,j} = -\log \frac{s_{i,j}}{s_{i,j} + \sum_{k \in \mathcal{N}(i)} s_{i,k}}. \quad (8)$$

Let $\mathcal{I}_{\text{valid}}$ denote the set of anchors with valid positive instances. The overall diachronic contrastive alignment loss is then computed by averaging over valid anchors and their positive instances:

$$\mathcal{L}_{\text{ctr}} = \frac{1}{|\mathcal{I}_{\text{valid}}|} \sum_{i \in \mathcal{I}_{\text{valid}}} \frac{1}{|\mathcal{P}(i)|} \sum_{j \in \mathcal{P}(i)} \ell_{i,j}. \quad (9)$$

This objective provides an additional structural constraint for distinguishing dynasty-specific sense usage.

4.3 Joint Optimization

Given the adapted representation h'_i , we use a classifier to predict the sense label and denote the classification loss as $\mathcal{L}_{\text{cls}}$. The final objective combines sense classification with diachronic contrastive alignment:

$$\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda \mathcal{L}_{\text{ctr}}, \quad (10)$$

where λ controls the contribution of the contrastive loss.

5 Experiments

We evaluate DynaSense on AC-DIAWSD using both PLM-based and LLM-based baselines. Our experiments examine overall performance, generality across PLM backbones, the contribution of each module, hyperparameter sensitivity, and qualitative model behavior. We use an 80%/20% train-test split and report results averaged over three independent runs. The evaluation metrics include accuracy and macro-averaged precision, recall, and F1. Further implementation details are provided in Appendix D.

5.1 Baselines

To comprehensively evaluate our proposed method, we compare it against a diverse set of baselines, encompassing both PLMs and various families of LLMs.

PLM Baselines. The selected PLMs are categorized into three groups: (1) *General Chinese PLMs*, represented by BERT-base-Chinese (Devlin et al., 2019); (2) *Ancient Chinese-specific PLMs*, including GuwenBERT², SikuBERT³, and

²<https://huggingface.co/ethanyt/guwenbert-base>

³<https://huggingface.co/SIKU-BERT/sikubert>

SikuRoBERTa⁴; and (3) *Task-specific WSD Models*, comprising PLM-WSD (Ghinassi et al., 2024), a supervised classifier originally evaluated on historical Latin that predicts candidate senses from contextualized target representations⁵; PolyBERT (Xia et al., 2025) which integrates local and global semantic representations, LTRS (Wang et al., 2025b) based on learning-to-rank for senses, and RS-BERT (Zhou et al., 2024) which incorporates Chinese radical features.

LLM Baselines. We select a broad spectrum of LLMs with varying architectures and parameter scales, including Qwen3-1.7B (Team, 2025), Qwen3-4B (Team, 2025), Qwen2.5-7B (Yang et al., 2024), DeepSeek-Qwen-1.5B and DeepSeek-Llama-8B (DeepSeek-AI, 2025), Llama-3B and Llama-8B (Team, 2024b), Gemma-7B (Team, 2024a), and Qwen3.6-35B⁶. Except for Qwen3.6-35B, all LLM baselines are fine-tuned on the training set using LoRA; Qwen3.6-35B is evaluated without parameter updates. The prompt templates include a persona instruction and a task query, with details provided in Appendix H.4 and Appendix H.2.

5.2 Main Results

As shown in Table 4, we evaluate PLMs with full-parameter fine-tuning and fine-tune LLMs for comparison. When integrated with DynaSense, SikuRoBERTa achieves the best overall accuracy of 70.41% and F1-score of 62.47%, outperforming the standard SikuRoBERTa baseline by 4.44 and 5.17 points, respectively. Domain-specific Ancient Chinese PLMs, however, do not consistently outperform BERT-base-Chinese, possibly because their pretraining corpora do not align uniformly with the dynasty distribution of our dataset. For example, SikuBERT is pretrained on *Siku Quanshu*, whose coverage is uneven across historical periods and may favor later transmitted texts. Consistent with this explanation, our per-dynasty results show that SikuBERT yields clearer gains over BERT-base-Chinese on the Yuan, Ming, and Qing dynasties. Fine-tuned LLMs, such as Qwen2.5-7B, also achieve competitive performance with an F1-score of 58.77%, but require substantially larger parameter scales and higher training cost. These results suggest that dynasty-aware structural

adaptation enables compact PLM-based models to perform strongly on Ancient Chinese diachronic WSD. We provide parameter-size comparisons in Appendix F, and sensitivity analyses on λ and contrastive sampling strategies in Appendix G.

5.3 Generality of DynaSense

To validate the generality and model-agnostic nature of DynaSense, we integrate our framework across diverse PLM architectures. As shown in Table 5, our method consistently yields positive F1-score gains ($\Delta F1$) across all evaluated baselines. Most notably, ancient Chinese-specific models such as GuwenBERT (+9.82) and SikuBERT (+8.91) exhibit the most substantial improvements. Furthermore, the consistent gains observed on task-specific WSD models (e.g., PolyBERT and RS-BERT) demonstrate that our diachronic adaptation mechanism successfully captures temporal semantic shifts that existing task-specific features fail to address, serving as a highly effective, plug-and-play enhancement.

5.4 Ablation Study

Table 6 presents the ablation results using SikuBERT and SikuRoBERTa as backbones. Removing either HDA or DCA consistently degrades performance, demonstrating their complementary contributions. Without HDA, the model loses explicit adaptation to dynasty-specific temporal variation; without DCA, fine-grained senses of the same target are less effectively separated in the representation space. The larger F1 drop observed after removing DCA suggests that contrastive representation structuring plays a particularly important role in fine-grained sense discrimination.

5.5 Case Study

Beyond quantitative metrics, we perform a qualitative case study (Table 7) to analyze common failure patterns and model behaviors in our WSD task. The following categories highlight the observed challenges and the specific improvements brought by our approach:

Fine-grained Confusion. Models frequently struggle when the semantic boundaries between candidate senses are highly similar or ambiguous. We observe this in two typical scenarios. First, in “累朝...” (Successive courts relied on...), the baseline model misclassifies the target “朝” as “dynasty” (朝代) rather than the contextually accurate “imperial court” (朝廷). Second, in “而谓公羊...”

⁴<https://huggingface.co/SIKU-BERT/sikuroberta>

⁵We replace the Perseus parallel corpus used in the original study with translations of Ancient Chinese texts.

⁶<https://huggingface.co/Qwen/Qwen3.6-35B-A3B-FP8>

Setting		Model	Acc	Precision	Recall	F1-score
PLM	FFT	BERT-base-Chinese	60.40	50.62	56.32	51.81
		GuwenBERT	56.34	45.93	52.56	47.40
		PLM-WSD	63.25	54.01	61.47	55.46
		PolyBERT	54.91	42.58	49.75	44.40
		LTRS	53.72	43.91	50.07	45.17
		RS-BERT	60.42	50.00	56.74	51.55
		SikuBERT	61.37	50.02	56.92	51.62
		SikuRoBERTa	65.97	55.99	62.17	57.30
LLM	Fine-tuned	Qwen3-4B	54.98	45.91	50.35	47.48
		Qwen3-1.7B	43.97	33.48	40.24	35.13
		Qwen2.5-7B	66.62	57.03	63.88	58.77
		DeepSeek-Qwen-1.5B	42.04	31.76	38.05	33.19
		DeepSeek-Llama-8B	61.60	51.64	58.59	53.25
		Llama-3B	42.20	31.72	38.02	33.22
		Llama-8B	59.63	48.45	55.19	50.05
		Gemma-7B	57.67	45.94	53.15	47.76
	Non-fine-tuned	Qwen3.6-35B	61.80	43.20	44.95	40.46
	SikuRoBERTa (Best PLM) + DynaSense			70.41	61.23	66.75

Table 4: Main experimental results on AC-DIAWSD. All metrics are in percentages (%).

Model	F1-score	Δ F1 ($\uparrow$)
BERT-base-Chinese	54.84	3.03
GuwenBERT	57.22	9.82
PolyBERT	49.59	5.19
LTRS	49.42	4.25
RS-BERT	55.13	3.58
SikuBERT	60.53	8.91
SikuRoBERTa	62.47	5.17

Table 5: Performance of different PLM-based models with DynaSense. Δ F1 denotes the F1-score improvement brought by integrating both HDA and DCA modules over the original backbone model.

(And to call Gongyang a pancake maker), the baseline incorrectly predicts "family/household" (家庭、人家) for the character "家", missing the nuance of a "tradesman" or artisan (经营某种行业的人). By integrating DynaSense, the contrastive learning module effectively sharpens the semantic boundaries between these closely related senses, enabling the model to correctly disambiguate both cases.

High-frequency Bias. Faced with rare or context-specific historical usages, uncertain models tend to default to the most common modern senses. For instance, in "长兴三年" (The third year of the Changxing era), the baseline incorrectly predicts the highly frequent sense "length" (长度) for the character "长". By incorporating DynaSense, the explicit injection of dynasty meta-data allows the model to successfully recognize

Model Variant	Acc	P	R	F1
SikuBERT (Full)	67.92	59.11	65.41	60.53
–HDA	64.95	54.30	60.69	55.63
–DCA	62.73	51.87	58.05	53.14
SikuRoBERTa (Full)	70.41	61.23	66.75	62.47
–HDA	68.16	57.61	63.90	59.08
–DCA	67.19	55.64	62.45	57.29

Table 6: Ablation study of the proposed DCA and HDA modules on DynaSense.

"Changxing" as a specific era name from the Tang dynasty, thus correctly predicting the label "era name" (年号).

Diachronic Correction: Traditional PLMs are highly susceptible to interference from dominant basic senses when processing terminology related to social identities and political institutions. In "太宗..." (Taizong personally campaigned...), the baseline misclassifies "宗" as its most common basic sense, "clan" (宗族). Conversely, the integrated Hypernetwork module in DynaSense leverages dynasty-conditioned information to correctly identify the specific historical institutional sense "temple name" (帝王庙号). This demonstrates the model's enhanced capability to capture historical contexts, effectively mitigating disambiguation challenges posed by diachronic semantic shifts.

6 Conclusion

We presented AC-DIAWSD, a large-scale diachronic WSD benchmark for Ancient Chinese with dynasty-level temporal information and fine-

Case Type	Context Snippet	Label	Baseline	+DynaSense
Fine-grained Confusion	累朝赖多士之用 (Successive courts relied on the service of numerous scholars)	Imperial court	Dynasty	Imperial court
Fine-grained Confusion	而谓公羊为卖饼家 (And to call Gongyang a pancake maker)	Tradesman	Family	Tradesman
High-frequency Bias	长兴三年 (The third year of the Changxing era)	Era name	Length	Era name
Diachronic Correction	太宗亲征辽东 (Taizong personally campaigned against Liaodong)	Temple name	Clan	Temple name

Table 7: Qualitative and error analysis cases. The target characters in the context snippets are highlighted in red.

grained sense annotations. Based on this benchmark, we proposed DynaSense, a model-agnostic framework that combines contrastive sense learning with hypernetwork-based dynasty adaptation to model semantic variation across historical periods. Experiments with PLM and LLM baselines show that diachronic WSD in Ancient Chinese remains challenging, and that DynaSense consistently improves PLM-based models by capturing dynasty-specific semantic shifts. We hope AC-DIAWSD and DynaSense will support future research on historical language understanding and diachronic semantics.

Limitations

Several aspects of this work warrant further investigation. First, AC-DIAWSD is derived from the *Twenty-Four Histories*, which primarily consist of official historiography and emphasize political, military, and biographical narratives. The dataset may therefore not capture the full diversity of Ancient Chinese usage. Second, the current implementation of DynaSense assumes access to dynasty metadata: HDA uses dynasty labels during both training and inference, while DCA operates on the resulting dynasty-adapted representation during training. Extending the framework to texts with unavailable or uncertain temporal information would broaden its applicability. Finally, performance remains lower for highly polysemous targets with numerous and subtly overlapping senses. Although DCA improves sense separation, such fine-grained semantic distinctions remain challenging. Future work will explore retrieval-augmented methods that incorporate richer external historical context to further improve disambiguation in these cases.

Acknowledgements

This research is supported by the Joint Project of Philosophy and Social Sciences Planning Discipline in Guangdong Province (Grant No.: GD23XZY07), and the Ordinary University Characteristic Innovation Project of Guangdong Province (Grant No.: 2023KTSCX031).

References

- Eneko Agirre, Itziar Aldabe, Mikel Lersundi, David Martínez, Eli Pociello, and Larraitz Uria. 2004. [The basque lexical-sample task](#). In *Proceedings of the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, SEN-SEVAL@ACL 2004, Barcelona, Spain, July 25-26, 2004*. Association for Computational Linguistics.
- Clare Arrington, Mauricio Gruppi, and Sibel Adali. 2025. [ConShift: Sense-based language variation analysis using flexible alignment](#). In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 167–181, Albuquerque, New Mexico. Association for Computational Linguistics.
- Ron Artstein and Massimo Poesio. 2008. [Inter-coder agreement for computational linguistics](#). *Comput. Linguistics*, 34(4):555–596.
- Kaspar Beelen, Federico Nanni, Mariona Coll Ardanuy, Kasra Hosseini, Giorgia Tolfo, and Barbara McGillivray. 2021. [When time makes sense: A historically-aware approach to targeted sense disambiguation](#). In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL-IJCNLP 2021 of *Findings of ACL*, pages 2751–2761. Association for Computational Linguistics.
- Yukun Cao, Chengkun Jin, Yijia Tang, and Ziyue Wei. 2024. [Word sense disambiguation combining knowledge graph and text hierarchical structure](#). *ACM Trans. Asian Low Resour. Lang. Inf. Process.*, 23(12):161:1–161:16.

- Pierluigi Cassotti, Lucia Siciliani, Marco de Gemmis, Giovanni Semeraro, and Pierpaolo Basile. 2023. [XL-LEXEME: wic pretrained model for cross-lingual lexical semantic change](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, ACL 2023, Toronto, Canada, July 9-14, 2023, pages 1577–1585. Association for Computational Linguistics.
- Jing Chen, Emmanuele Chersoni, Dominik Schlechtweg, Jelena Prokic, and Chu-Ren Huang. 2023. [Chiwug: A graph-based evaluation dataset for chinese lexical semantic change detection](#). In *Proceedings of the 4th Workshop on Computational Approaches to Historical Language Change, LChange@EMNLP 2023, Singapore, December 6, 2023*, pages 93–99. Association for Computational Linguistics.
- DeepSeek-AI. 2025. [Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning](#). *CoRR*, abs/2501.12948.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.
- Philip Edmonds and Scott Cotton. 2001. [SENSEVAL-2: overview](#). In *Proceedings of Second International Workshop on Evaluating Word Sense Disambiguation Systems, SENSEVAL@ACL 2001, Toulouse, France, July 5-6, 2001*, pages 1–5. Association for Computational Linguistics.
- Iacopo Ghinassi, Simone Tedeschi, Paola Marongiu, Roberto Navigli, and Barbara McGillivray. 2024. [Language pivoting from parallel corpora for word sense disambiguation of historical languages: A case study on latin](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC/COLING 2024, 20-25 May, 2024, Torino, Italy*, pages 10073–10084. ELRA and ICCL.
- Mario Giulianelli, Iris Luden, Raquel Fernández, and Andrey Kutuzov. 2023. [Interpretable word sense representations via definition generation: The case of semantic change analysis](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL 2023, Toronto, Canada, July 9-14, 2023, pages 3130–3148. Association for Computational Linguistics.
- Luyao Huang, Chi Sun, Xipeng Qiu, and Xuanjing Huang. 2019. [GlossBERT: BERT for word sense disambiguation with gloss knowledge](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3509–3514, Hong Kong, China. Association for Computational Linguistics.
- Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli. 2016. [Embeddings for word sense disambiguation: An evaluation study](#). In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 897–907, Berlin, Germany. Association for Computational Linguistics.
- Thomas Kober, Julie Weeds, John Wilkie, Jeremy Reffin, and David J. Weir. 2017. [One representation per word - does it make sense for composition?](#) *CoRR*, abs/1702.06696.
- Hlaudi Daniel Masethe, Mosima Anna Masethe, Sunday Olusegun Ojo, Fausto Giunchiglia, and Pius Adewale Owolawi. 2024. [Word sense disambiguation for morphologically rich low-resourced languages: A systematic literature review and meta-analysis](#). *Information*, 15(9).
- Anna Mosolova, Marie Candito, and Carlos Ramisch. 2025. [In the LLM era, word sense induction remains unsolved](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 17161–17178, Vienna, Austria. Association for Computational Linguistics.
- Roberto Navigli. 2009. [Word sense disambiguation: A survey](#). *ACM Comput. Surv.*, 41(2):10:1–10:69.
- Roberto Navigli. 2026. [Is word sense disambiguation dead in the LLM era?](#) In *Fortieth AAAI Conference on Artificial Intelligence, Thirty-Eighth Conference on Innovative Applications of Artificial Intelligence, Sixteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2026, Singapore, January 20-27, 2026*, pages 39753–39762. AAAI Press.
- Tommaso Pasini, Alessandro Raganato, and Roberto Navigli. 2021. [XL-WSD: an extra-large and cross-lingual evaluation framework for word sense disambiguation](#). In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 13648–13656. AAAI Press.
- Francesco Periti, Pierluigi Cassotti, Haim Dubossarsky, and Nina Tahmasebi. 2024. [Analyzing semantic change through lexical replacements](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL 2024, Bangkok, Thailand, August 11-16, 2024, pages 4495–4510. Association for Computational Linguistics.
- Francesco Periti and Nina Tahmasebi. 2024. [A systematic comparison of contextualized word embeddings](#)

- for lexical semantic change. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, NAACL 2024, Mexico City, Mexico, June 16-21, 2024, pages 4262–4282. Association for Computational Linguistics.
- Alessandro Raganato, José Camacho-Collados, and Roberto Navigli. 2017. [Word sense disambiguation: A unified evaluation framework and empirical comparison](#). In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017, Valencia, Spain, April 3-7, 2017, Volume 1: Long Papers*, pages 99–110. Association for Computational Linguistics.
- Rakia Saidi and Fethi Jarray. 2026. [Arabic word sense disambiguation: a survey in the era of transformer-based models](#). *Int. J. Data Sci. Anal.*, 22(1):10.
- Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. 2020. [Semeval-2020 task 1: Unsupervised lexical semantic change detection](#). In *Proceedings of the Fourteenth Workshop on Semantic Evaluation, SemEval@COLING 2020, Barcelona (online), December 12-13, 2020*, pages 1–23. International Committee for Computational Linguistics.
- Dominik Schlechtweg, Frank D. Zamora-Reina, Felipe Bravo-Marquez, and Nikolay Arefyev. 2025. [Sense through time: diachronic word sense annotations for word sense induction and lexical semantic change detection](#). *Lang. Resour. Evaluation*, 59(2):1431–1465.
- Lei Shu, Yiluan Guo, Huiping Wang, Xuetao Zhang, and Renfen Hu. 2021. [古汉语词义标注语料库的构建及应用研究 \(the construction and application of Ancient Chinese corpus with word sense annotation\)](#). In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 549–563, Huhhot, China. Chinese Information Processing Society of China.
- Michael Angelos Simos and Christos Makris. 2026. [Word sense disambiguation with wikipedia entities: A survey of entity linking approaches](#). *Entropy*, 28(2).
- Deshan K. Sumanathilaka, Nicholas Micallef, and Julian Hough. 2025. [Glossgpt: GPT for word sense disambiguation using few-shot chain-of-thought prompting](#). In *The 16th International Conference on Ambient Systems, Networks and Technologies (ANT 2025) / The 8th International Conference on Emerging Data and Industry 4.0 (EDI40 2025) / Affiliated Workshops, April 22-24, 2025, Patras, Greece*, volume 257 of *Procedia Computer Science*, pages 785–792. Elsevier.
- Xiaohang Tang, Yi Zhou, Taichi Aida, Procheta Sen, and Danushka Bollegala. 2023. [A word sense distribution-based approach for semantic change prediction](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, volume EMNLP 2023 of *Findings of ACL*, pages 3575–3590. Association for Computational Linguistics.
- Gemma Team. 2024a. [Gemma: Open models based on gemini research and technology](#). *CoRR*, abs/2403.08295.
- Llama Team. 2024b. [The llama 3 herd of models](#). *CoRR*, abs/2407.21783.
- Qwen Team. 2025. [Qwen3 technical report](#). *CoRR*, abs/2505.09388.
- Hai Wang, Yuzhi Liang, and Han Ren. 2025a. [Temporal-aware soft prompt tuning for automatic text dating](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long Papers, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pages 3975–3987. Association for Computational Linguistics.
- Hansi Wang, Yue Wang, Qiliang Liang, and Yang Liu. 2025b. [LTRS: improving word sense disambiguation via learning to rank senses](#). In *Proceedings of the 31st International Conference on Computational Linguistics, COLING 2025, Abu Dhabi, UAE, January 19-24, 2025*, pages 1934–1942. Association for Computational Linguistics.
- Yuting Wei, Meiling Li, Yangfu Zhu, Yuanxing Xu, Yuqing Li, and Bin Wu. 2025. [A diachronic language model for long-time span classical chinese](#). *Inf. Process. Manag.*, 62(1):103925.
- Linhan Xia, Mingzhan Yang, Guohui Yuan, Shengnan Tao, Yujing Qiu, Guo Yu, and Kai Lei. 2025. [Polybert: Fine-tuned poly encoder bert-based model for word sense disambiguation](#). In *Knowledge Science, Engineering and Management - 18th International Conference, KSEM 2025, Macao, China, August 4-7, 2025, Proceedings, Part IV*, Lecture Notes in Computer Science, pages 433–443. Springer.
- Jung H. Yae, Nolan C. Skelly, Neil C. Ranly, and Phillip M. LaCasse. 2025. [Leveraging large language models for word sense disambiguation](#). *Neural Comput. Appl.*, 37(6):4093–4110.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jixai Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 22 others. 2024. [Qwen2.5 technical report](#). *CoRR*, abs/2412.15115.
- Junwei Zhang and Xiaolin Li. 2025. [Quantum-inspired non-homologous representation constraint mechanism for long-tail senses of word sense disambiguation](#). In *Thirty-Ninth AAAI Conference on Artificial Intelligence*.

cial Intelligence, Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence, Fifteenth Symposium on Educational Advances in Artificial Intelligence, AAAI 2025, Philadelphia, PA, USA, February 25 - March 4, 2025, pages 25877–25885. AAAI Press.

Kaiyuan Zhang, Qian Liu, Luyang Zhang, Chaoqun Zheng, Shuaimin Li, Bing Xu, Muyun Yang, Xinxiao Qiao, and Wenpeng Lu. 2025. [MADAWSD: Multi-agent debate framework for adversarial word sense disambiguation](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 22283–22302, Suzhou, China. Association for Computational Linguistics.

Xiaofeng Zhou, Heyan Huang, Zewen Chi, Mucheng Ren, and Yang Gao. 2024. [RS-BERT: pre-training radical enhanced sense embedding for chinese word sense disambiguation](#). *Inf. Process. Manag.*, 61(4):103740.

A Target Characters and Semantic Categorization

This appendix presents the complete set of 63 target characters used in our diachronic WSD task. We selected these characters to cover phenomena central to the interpretation of Ancient Chinese texts: high-frequency content characters convey core meanings, function characters mark syntactic relations, governance and social characters encode historical institutions and social structures, and modal and evaluative characters express attitudes and judgments. We also include characters with counterintuitive archaic senses to capture substantial differences between ancient and modern usage. As shown in Table 8, the characters are organized into six categories.

B Cohen’s Kappa

We provide detailed definitions of the two Kappa-based agreement measures used to evaluate annotation reliability.

- **Global Pooling Kappa** (κ_{pool}): treats all sense IDs as a unified label space and computes agreement over all evaluated instances.
- **Word-level Weighted Average Kappa** (κ_{weighted}): computes Cohen’s Kappa separately for each target character and aggregates the scores using a frequency-weighted average, thereby preserving target-specific sense boundaries.

C Diachronic Semantic-Shift Analysis

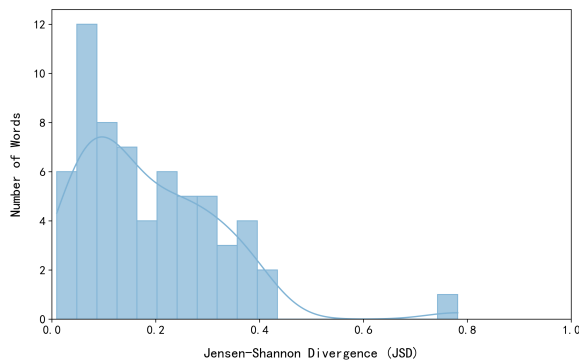


Figure 2: Distribution of Jensen-Shannon Divergence (JSD) scores across target characters.

Figure 2 shows the distribution of Jensen-Shannon Divergence (JSD) scores across target characters. Most characters fall within a low-to-moderate JSD range, while a long tail indicates

that some characters undergo stronger diachronic semantic shifts.

Figure 3 presents representative semantic-shift examples from the six target character categories. These examples show that semantic change in Ancient Chinese is not uniform across lexical types: some characters exhibit gradual changes in dominant senses, while others undergo sharper shifts associated with institutional, social, or evaluative usage.

D Experimental Setup

Implementation Details. For PLM backbones, we set the batch size to 64 and the learning rate to 5×10^{-4} . For LLMs, we use LoRA fine-tuning with rank $r = 8$, $\alpha = 16$, and a dropout rate of 0.1. The LLM learning rate is 1×10^{-4} , the maximum sequence length is 2048, and the inference temperature is 0.5.

For DynaSense, the dynasty embedding dimension is set to 64 and the hidden dimension of the hypernetwork to 256. For hypernetwork pretraining, we sample up to 200 instances per dynasty. The adaptation scaling factor γ is initialized to 0.5. In the contrastive alignment module, we cap the number of positive and negative samples per anchor at $|\mathcal{P}(i)| = 10$ and $|\mathcal{N}(i)| = 40$, respectively. The loss balancing coefficient is set to $\lambda = 0.5$.

Evaluation Metrics. We report Accuracy, Precision, Recall, and F1-score. Precision, Recall, and F1 are computed using macro averaging to provide a balanced evaluation across sense categories.

E Dynasty-Level Results

As shown in Table 9, performance varies across dynasties, with later periods generally achieving higher F1 scores. Although the trend is not strictly monotonic, it suggests that the Modern Chinese knowledge encoded in the backbone may transfer more effectively to later historical texts, whose lexical usage is generally closer to Modern Chinese. Differences in pretraining coverage and dynasty-specific sense distributions may also contribute to the observed variation.

F Model Size Comparison

Figure 4 compares model parameter sizes and WSD accuracy. PLM-based models achieve competitive or better performance than several LLM-based baselines with substantially smaller parameter sizes, suggesting that task-specific diachronic

Category	Target Characters (with Pinyin)
High-Frequency Content Words	上 (<i>shàng</i>), 下 (<i>xià</i>), 中 (<i>zhōng</i>), 为 (<i>wéi</i>), 事 (<i>shì</i>), 任 (<i>rèn</i>), 出 (<i>chū</i>), 前 (<i>qián</i>), 去 (<i>qù</i>), 发 (<i>fā</i>), 名 (<i>míng</i>), 后 (<i>hòu</i>), 少 (<i>shǎo</i>), 当 (<i>dāng</i>), 得 (<i>dé</i>), 来 (<i>lái</i>), 治 (<i>zhì</i>), 理 (<i>lǐ</i>), 生 (<i>shēng</i>), 立 (<i>lì</i>), 行 (<i>xíng</i>), 长 (<i>zhǎng</i>)
Abstract Function Words	与 (<i>yǔ</i>), 乃 (<i>nǎi</i>), 即 (<i>jí</i>), 及 (<i>jí</i>), 既 (<i>jì</i>), 而 (<i>ér</i>), 自 (<i>zì</i>), 至 (<i>zhì</i>), 若 (<i>ruò</i>)
Institutional Governance Words	令 (<i>lìng</i>), 兵 (<i>bīng</i>), 制 (<i>zhì</i>), 君 (<i>jūn</i>), 国 (<i>guó</i>), 朝 (<i>cháo</i>), 王 (<i>wáng</i>)
Societal Kinship Terms	姓 (<i>xìng</i>), 宗 (<i>zōng</i>), 家 (<i>jiā</i>), 族 (<i>zú</i>), 老 (<i>lǎo</i>)
Modal Evaluative Words	义 (<i>yì</i>), 信 (<i>xìn</i>), 善 (<i>shàn</i>), 宜 (<i>yí</i>), 察 (<i>chá</i>), 忠 (<i>zhōng</i>), 恶 (<i>è</i>), 明 (<i>míng</i>), 欲 (<i>yù</i>), 正 (<i>zhèng</i>), 知 (<i>zhī</i>), 能 (<i>néng</i>), 诚 (<i>chéng</i>), 非 (<i>fēi</i>)
Counter-Intuitive Archaic Words	乘 (<i>chéng</i>), 亡 (<i>wáng</i>), 居 (<i>jū</i>), 见 (<i>jiàn</i>), 走 (<i>zǒu</i>), 闻 (<i>wén</i>)

Table 8: Taxonomy of the target characters selected for diachronic WSD. Characters are grouped into six linguistic dimensions with their corresponding Pinyin provided in parentheses.

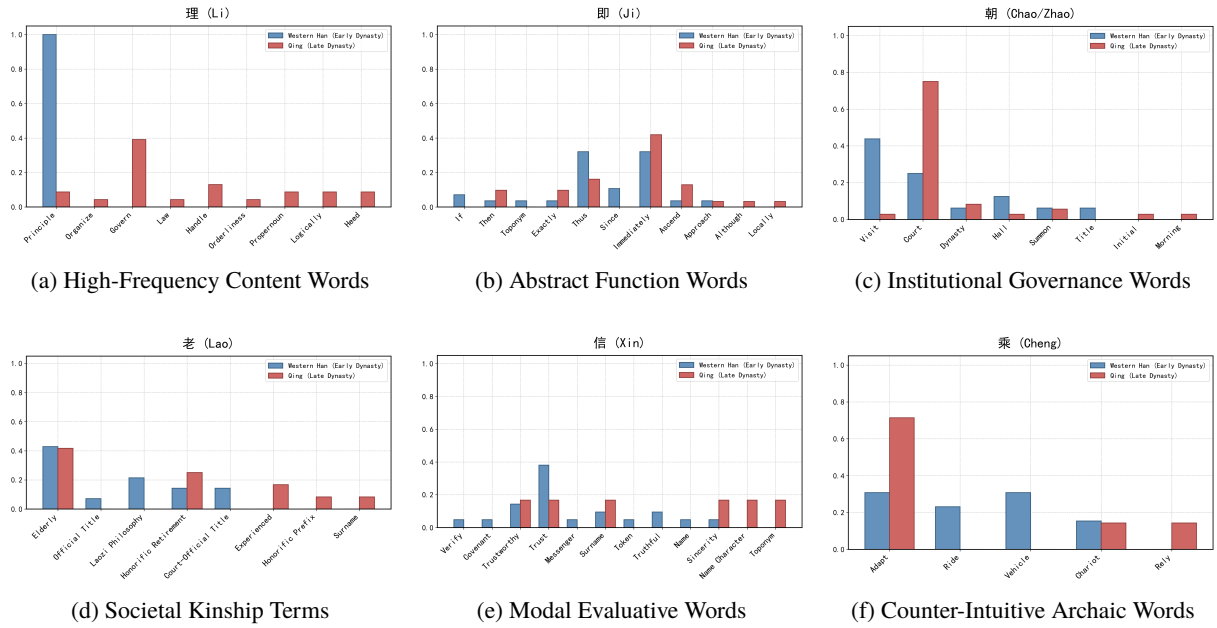


Figure 3: Representative semantic-shift examples across the six target character categories.

adaptation can be more resource-efficient than simply scaling model size.

G Hyperparameter Sensitivity and Sampling Analysis

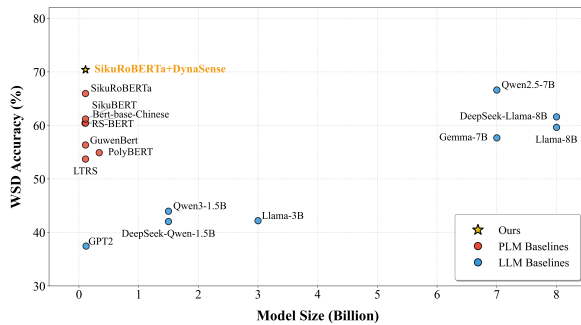


Figure 4: Accuracy comparison of models with varying parameter sizes.

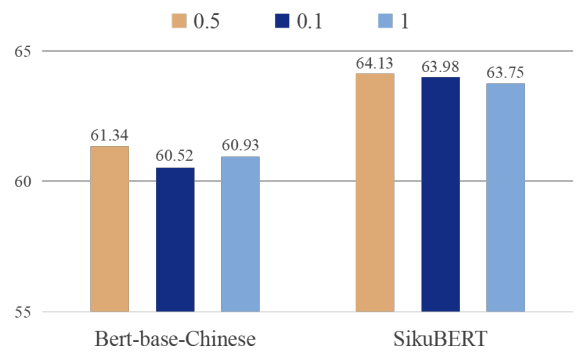


Figure 5: The sensitivity of λ

Sensitivity Analysis of λ . We evaluate the balancing coefficient λ , which controls the trade-off

Dynasty	Macro-F1
Western Han	65.00
Eastern Han	58.70
Western Jin	61.67
Liu Song	59.16
Southern Liang	64.20
Northern Qi	62.19
Tang	62.20
Later Jin	60.41
Song	65.24
Yuan	63.49
Ming	62.46
Qing	65.88
Average	62.55

Table 9: Dynasty-level macro-F1 scores (%) of SikuRoBERTa + DynaSense on AC-DIAWSD.

between the classification and contrastive losses. As shown in Figure 5, the model achieves its optimal accuracy (61.34%) at $\lambda = 0.5$. A lower value (e.g., $\lambda = 0.1$) provides insufficient contrastive signals to structurally constrain the representation space, whereas an excessively high value (e.g., $\lambda = 1.0$) interferes with the primary classification objective, resulting in slight performance degradation.

Analysis of Contrastive Sampling Strategies.

We investigate the impact of different positive and negative sampling pools on the contrastive learning module. Our default strategy constructs positive and negative pairs based on character identity and sense labels: positive samples share the same target character and sense as the anchor, whereas negative samples share the same target character but have different senses. We compare this default setting against four alternative pools (Figure 6).

For positive sampling, modifying the pool—either by selecting different senses within the same dynasty (Pos-SDyn) or sampling indiscriminately across all dynasties and senses of the target character (Pos-DSense)—causes a distinct performance drop on both SikuBERT and BERT backbones. This highlights that relaxing temporal or semantic consistency in positive pairs introduces detrimental noise that blurs fine-grained sense boundaries.

Conversely, expanding the negative sampling pool yields marginal gains. Drawing negatives from different dynasties (Neg-DDyn) or selecting different senses across all eras (Neg-DSense) slightly improves SikuBERT accuracy (from 66.07% to 66.33% and 66.59%, respectively). This indicates that cross-era hard negatives can further regularize the representation space.

However, we retain the default strategy because it provides a simple and efficient balance between sense discrimination and sampling complexity.

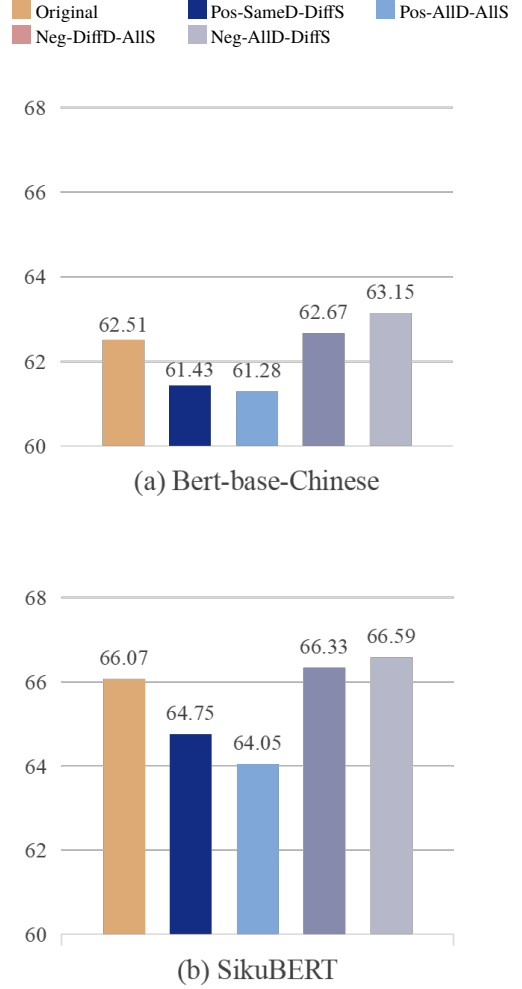


Figure 6: Comparison of WSD accuracy under different contrastive sample construction strategies.

H Prompts

H.1 Prompt for Ancient-to-Modern Chinese Translation

H.2 Task Query

H.3 Semantic Mapping Prompt

H.4 Persona Instruction

Ancient-to-Modern Chinese
System Instruction: You are TongGu, a large language model specialized in ancient Chinese, trained by the DLVCLab at South China University of Technology. You possess extensive knowledge of ancient Chinese texts and provide users with helpful and accurate responses.
Interaction Format: <User> Translate into modern Chinese: {text} <Tonggu>

Table 10: The system and user prompt utilized for translating Ancient Chinese text into Modern Chinese.

Task Query
Ancient Text: "{text}" Target Character: {word} All Senses for this Character: (Format: Sense ID: Sense Explanation)
Output Format: {"sense_id": "Sense ID or null", "confidence": <value between 0.0 and 1.0>}

Table 11: The user query format and expected output for the sense matching task.

Semantic Mapping Prompt
You are an expert in Ancient Chinese. Your task is to accurately extract the terms from the modern translation that correspond to the specified characters in the ancient text. Please strictly follow the steps below: 1. Identification and Analysis - Carefully read both the ancient text and its modern translation. - Locate all occurrences of the specified character within the ancient text. - Analyze the context for each occurrence. - Distinguish between content words and function words, adopting the corresponding extraction method: - Content words (nouns, verbs, adjectives): Extract meanings from the translation. - Function words (prepositions, conjunctions): State their grammatical roles. - Proper nouns: Directly output the proper nouns. 2. Extraction and Mapping - Content word extraction: - Extract the corresponding modern meaning from the translation for each occurrence. - Ensure definitions are complete, accurate, and based solely on the translation. - Function word processing: State grammatical function (e.g., “indicates...”). - Proper nouns: Extract directly if applicable. - Special cases: Infer based on context if there is no direct equivalent. 3. Core Principles - Supplement based on translation when a direct definition cannot be extracted. - Meaning must be concise and accurate (1-4 characters). - Do not extract words from the original ancient text. - Content must appear in or be a direct expansion of the translation. - Output “unknown” if the translation provides no corresponding meaning.
Output Specifications: - ['sense1', 'sense2', ...] - Output must be a list of meanings arranged in the text’s occurrence order. - Strictly format the output as JSON.
Output style: - Ancient Text: {text} - Modern Translation: {translation} - Target Character: {word}

Table 12: The detailed system prompts and output style for the Semantic Mapping task.

Persona Instruction
System Prompt You are an expert in Ancient Chinese linguistics, specializing in analyzing character meanings and sense mapping. Based on the provided ancient text and the target character, along with all its possible senses, determine whether the model’s given choice matches a specific sense. Task Requirements: - Carefully analyze the context of the original ancient text. - Accurately determine the specific meaning of the target character within the context. - Precisely match the model’s choice against the provided list of senses. - If the choice directly matches a sense, return that sense’s ID. - If the choice is a synonym of a sense, return that sense’s ID. - If neither applies but the choice is contextually reasonable, assess whether a new sense needs to be added. - If the choice does not match any sense and is unreasonable, return null.
Output Format (Strictly JSON, no additional text): <pre>{ "sense_id": "Sense ID or null", "confidence": <value between 0.0 and 1.0> }</pre>

Table 13: The persona instruction prompt for the sense matching evaluation.