Large-Scale Multimodal Knowledge Graph about Classical Chinese Poetry: Fine-grained Method and Comprehensive Evaluation

Shuo Wang[1], Qing Zhu[1], Yang Xiao[1]*, Minglong Lei[1] ,

[1]College of Computer Science, Beijing University of Technology, China
shuow0920@gmail.com, {ccgszq, xiaoyang, leiml}@bjut.edu.cn,

Abstract

Classical Chinese poetry is a treasured cultural heritage of humanity, attracting extensive research interest. However, the study of classical Chinese poetry is hindered by the lack of open, large-scale, and fine-grained multimodal datasets. Prior datasets are either limited by modality constraints, dataset size, or the level of dataset refinement, making them inadequate for effectively supporting studies and the development of applications in classical Chinese poetry. To address these issues, we propose a method for constructing a large-scale and fine-grained multimodal knowledge graph of classical Chinese poetry. We first design an informative ontology graph for classical Chinese poetry and comprehensively collect knowledge about poetry based on it. Furthermore, the method leverages knowledge augmentation, prompt optimization, and text-image alignment to acquire comprehensive, fine-grained knowledge. Both qualitative and quantitative evaluations are conducted on the Multimodal Knowledge Graph of Classical Chinese Poetry (CPMK), highlighting its comprehensiveness and high quality. We also conduct downstream evaluations on four tasks: poetry question answering, poetry theme classification, poetry-image retrieval, and rigid-formats poetry generation. Significant results are achieved across all four tasks, demonstrating CPMK's effectiveness in supporting research on Chinese poetry. CPMK will be released to promote research in Chinese culture[1].

1 Introduction

Classical Chinese poetry is a treasured cultural heritage that passes down ancient literature and fosters cross-cultural understanding between the East and the West. As times change, understanding Chinese poetry has become increasingly difficult. Differences between ancient poetry and modern Chinese,

Figure 1: Illustration of poetry understanding via AP(Ancient Poetry), MCT(Modern Chinese Translation), PI(Poetry Imagery), IM(Imagery Meaning), II(Imagery Image), and APBG(Ancient Poetry Background).

the evolution of poetry imagery meanings, and factors such as the poetry background all affect our understanding of Chinese poetry.

For example, SuShi(苏轼)'s poem "水光潋滟晴方好, 山色空蒙雨亦奇 " describes the beauty of nature in diverse weather conditions. However, the evolving meanings of poetry imagery, such as KongMeng(空蒙), alongside the background of ancient poetry, have made it difficult to fully understand the poem. This difficulty not only hinders understanding of the poetry but also impedes the study and development of applications for classical Chinese poetry. Presenting the semantics of poetry imagery through both textual and visual modalities(Figure 1) can help people from different cultural backgrounds understand Chinese poetry.

Although classical Chinese poetry is a cornerstone of Chinese cultural heritage and has been extensively studied from various scholarly per-

* Corresponding author.
[1]https://github.com/Sure-aa/CPMK

Findings of the Association for Computational Linguistics: ACL 2026, pages 14646–14669
July 2-7, 2026

spectives, most existing research remains predominantly focused on the textual modality (Wang et al., 2023; Wei et al., 2024). The scarcity of multimodal datasets hinders research beyond the textual modality. Therefore, constructing a MultiModal Knowledge Graph (MM-KG) of classical Chinese poetry is essential for advancing research in this area.

To facilitate the discussion, we introduce several key concepts relevant to classical Chinese poetry, which are explained in the following sections. Ancient Poetry refers to the content of classical Chinese poetry, characterized by its traditional form and archaic language, which differ significantly from modern Chinese. Poetry Imagery denotes specific objects or concepts that poets use to express emotions and thoughts. Imagery Meaning is the modern Chinese meaning of the Poetry Imagery, while Imagery Image refers to the visual representation of the Imagery Meaning. Modern Chinese Translation refers to the translation of classical Chinese poetry into modern Chinese.

To the best of our knowledge, the currently available MM-KG of classical Chinese poetry is limited to PKG(Li et al., 2022), which is the only text-vision modality knowledge graph in this field. However, it has many shortcomings, making it difficult to support downstream tasks.

(1) PKG focuses solely on imagery-related knowledge, neglecting other critical aspects such as poetry-related knowledge and author-related knowledge, both of which are essential for poetry research. For instance, Jiang et al. (2024) uses poetry appreciation to generate images of ancient poetry. **(2)** Imagery images in the PKG are represented as URLs in website[2], but many of these images are no longer available. Among 59,721 randomly sampled URLs, 12,854(21%) are found to be invalid, severely impacting the knowledge graph's utility for downstream tasks. **(3)** Auditory elements are crucial components of classical Chinese poetry. These elements are mandated in many poetry forms, such as five-character and seven-character poems. However, the auditory data are overlooked in PKG.

To address the issues above, this paper proposes a method for constructing a large-scale, fine-grained MM-KG of classical Chinese poetry that integrates textual, visual, and auditory modalities. To obtain comprehensive knowledge of classical Chinese poetry, we constructed an ontology graph encompassing multiple aspects of poetry-related knowledge. Guided by this graph, we systematically collected knowledge related to its concepts. To ensure the completeness of textual knowledge, we employ a poetry knowledge augmentation strategy. For visual data in the ontology graph, we use generative models to generate images rather than traditional web scraping, thereby enhancing the correlation between text and images. During image generation, origin prompts are processed using a prompt optimization technique to improve image quality. For the obtained text-image pairs, text-image alignment is used to filter out high-quality text-image pairs. For auditory data, we gathered auditory knowledge for characters found in classical Chinese poetry. The proposed method leads to the construction of an MM-KG of Chinese Poetry (CPMK) that spans textual, visual, and auditory modalities.

Qualitative and quantitative evaluations demonstrate that CPMK is more comprehensive and accurate than existing poetry datasets. To further validate the effectiveness of CPMK in downstream tasks, we incorporate it into poetry question answering task, poetry theme classification task, poetry-image retrieval task and rigid-formats poetry generation tasks. Experimental results demonstrate that CPMK significantly improves the performance of downstream tasks. Through qualitative and quantitative research, as well as validation in downstream tasks, it has been demonstrated that CPMK can effectively support the study and development of applications related to classical Chinese poetry. Our contributions are listed below:

- We propose a method for constructing a large-scale and fine-grained MM-KG of classical Chinese poetry. We first design an ontology of classical Chinese poetry to gather comprehensive knowledge, and then adopt knowledge augmentation, prompt optimization, and text-image alignment to acquire a large-scale and fine-grained MM-KG.
- Using this method, we construct a multimodal knowledge graph of classical Chinese poetry with **6,834,825** textual nodes, **211,467** visual nodes, and **82,679** auditory nodes. Qualitative and quantitative evaluations, along with downstream task validation, collectively confirm its quality and effectiveness in the field of classical Chinese poetry.

[2]https://unsplash.com

- We propose a knowledge-enhanced poetry-image retrieval model. By establishing connections between classical Chinese poetry and images through modern Chinese translation of poetry, the model achieves state-of-the-art results on two datasets in the multimodal task of poetry-image retrieval. It uses a large amount of textual data and only a small amount (or even no) visual data, providing insights for other multimodal tasks.

- We construct two datasets for the classical Chinese poetry-image retrieval task using manual collection and automated generation methods. As far as we know, this is the first benchmark for this task, and it will be released to advance research in classical Chinese poetry.

- We validate the effectiveness of CPMK across four tasks, including poetry question answering, poetry theme classification, poetry-image retrieval, and rigid-formats poetry generation. Results in all tasks demonstrate the effectiveness of CPMK in supporting research on classical Chinese poetry.

2 RELATED WORKS

2.1 Knowledge Graph Construction

Due to advancements in LLMs, many studies have utilized them to construct knowledge graphs. Wang et al. (2025) leverages LLMs for triple extraction, relational embedding, and schema-based normalization, which supports multi-domain construction without retraining or fine-tuning. FolkScope(Yu et al., 2023) leverages the generative power of LLMs and human-in-the-loop annotation to semi-automatically construct the knowledge graph. MMGraphRAG (Wan and Yu, 2025) utilizes the Yolo model to extract image entities and leverages large vision models to obtain textual descriptions of the image entities, facilitating the construction of multimodal knowledge graphs. However, in the field of classical Chinese poetry, the lack of a large-scale knowledge base like Wikipedia makes it difficult to collect substantial amounts of data, rendering existing methods difficult to apply directly.

In the field of classical Chinese poetry, there have also been studies focused on constructing knowledge graphs. KnowPoetry (Hong et al., 2020) proposes a framework to extract poems, poets, and their relationships from Tang poetry, thereby constructing a domain ontology and a knowledge graph. SKG-Poetry (Zhao et al., 2022) constructs a sememe knowledge graph of classical Chinese poetry, linking classical and modern Chinese vocabularies to enhance semantic understanding. These knowledge graphs are either constrained by their modalities or suffer from quality deficiencies, which makes it difficult for them to support downstream tasks effectively.

2.2 Classical Chinese Poetry Data

Research on classical Chinese poetry data primarily focuses on textual data, with limited exploration of visual and audio modalities. The ancient corpora of text include four main datasets: Poetry(Werneror, 2017), CCPM(Li et al., 2021), ACP-Corpus(Liu et al., 2025), Chinese-poetry-and-prose(VMIJUNV, 2023).

There is limited attention to vision and audio modalities in the study of classical Chinese poetry. In terms of vision modality, the PKG (Li et al., 2022) compiles knowledge related to poetry imagery, and (Liu et al., 2020) maps poems to specific categories and collects images corresponding to those categories. Regarding the audio modality, to our knowledge, no relevant knowledge graph has been identified.

3 Method for constructing CPMK

We analyse the data requirements from recent studies on classical Chinese poetry, such as Li et al. (2022); Jiang et al. (2024); Li et al. (2021), to construct an ontology graph. This graph serves as the guidance for the construction of the MM-KG of classical Chinese poetry. The ontology graph is in Appendix I. Guided by the ontology graph, this method overcomes the limitations of previous studies, which lacked comprehensive coverage of Chinese poetry knowledge. The construction of the Chinese poetry MM-KG consists of several stages, with a schematic overview of the pipeline provided in Appendix J.

3.1 Acquisition of Raw Data.

Knowledge related to ancient poetry and the author is crawled from the authoritative poetry website SouYun[3]. We extract words that appear more than 5 times and all the characters that have appeared in ancient poetry. These words and characters are

[3]https://sou-yun.cn/

used to crawl for their semantic meanings on the website HanDian[4]. For words, if their semantic meaning exists, they are categorized as poetry imagery, and their meaning serves as imagery meaning. For characters, in addition to their semantic meanings, we also crawl their auditory knowledge and visual knowledge in HanDian. Characters are visually represented as GIFs or SVGs to demonstrate stroke order. Pinyin and Zhuyin are offered as audio to illustrate pronunciation.

When dealing with imagery images, manual collection of extensive images is exhausting, and web scraping poses significant challenges due to the unique characteristics of classical Chinese poetry. (1) There is a lack of comprehensive image databases for Chinese literature, as existing large-scale image websites primarily focus on modern elements and offer limited coverage of ancient Chinese literature. (2) Some imagery meanings are relatively abstract, making it challenging to find images that basically convey their visual meaning when using web scraping.

ChenMeng(尘梦) ChiXiao(赤霄)

Figure 2: Imagery image for the "ChenMeng" and "ChiXiao", both generated using a generative model.

Generative models trained on large-scale datasets can effectively address the issues mentioned. For instance, the poetry imagery ChiXiao(赤霄)'s imagery meaning refers to the legendary ancient sword of LiuBang(刘邦), the poetry imagery ChenMeng(尘 梦)'s imagery meaning symbolizes the illusion of the mortal world. As shown in Figure 2, generative models can generate content related to ancient legends and abstract concepts with relatively effective results. Therefore, this paper uses a generative model to create imagery images. The selection of the generative model and the prompt setting can be found in Appendix E.

[4]https://www.zdic.net/

3.2 Poetry Knowledge Augmentation

Since the pre-Qin period (before 1000 BCE), the long-term transmission of Chinese poetry has introduced significant textual variations. To the best of our knowledge, existing studies (Wei et al., 2024; He et al., 2023) often overlook these nuances, resulting in incomplete datasets. Furthermore, as most data is sourced from the internet, its reliability is compromised by varying website quality and unresolved historical inconsistencies. This paper adopts a cross-augmentation strategy, which integrates variations from multiple knowledge bases to provide the most comprehensive and reliable knowledge. We focus on two core aspects: knowledge of ancient poetry and the author.

We collect knowledge about ancient poetry and authors from GuShiWen[5] and GuoXueHui[6]. For ancient poetry-related knowledge, we employ a two-phase deduplication strategy inspired by (Liu et al., 2025): global alignment removes redundant poems, while local alignment segments poems by punctuation and evaluates overlaps between text chunks. Similar ancient poems are clustered rather than overwritten, with their relevant knowledge integrated to ensure a comprehensive representation. Details of the process are provided in Appendix F. For author-related knowledge, we determine entity consistency by verifying the author's name and dynasty, and then aggregate the relevant information.

3.3 Prompt Optimization for Imagery Image Generation

Maintaining the consistency between the input text and generated images is a challenge. Many generative models use CLIP's text encoder, freezing its parameters while only the diffusion process is trained (Ramesh et al., 2022; Rombach et al., 2022). However, research from Zhang et al. (2024) shows that CLIP's text encoder effectively handles fewer than 20 tokens, leading to hallucinations when processing longer texts.

Table 1 demonstrates that raw imagery meanings sourced from the internet exceed token limits(21.72). This internet-derived textual data often contains irrelevant noise, which affects the performance of downstream tasks. Additionally, much poetry imagery has multiple meanings that are difficult to distinguish using scripts. The above issues PKG has not yet resolved. In PKG, multi-

[5]https://www.gushiwen.cn/
[6]https://www.gushicimingju.com/

Table 1: The average token distribution. Abbreviations —IM: Imagery Meaning; SP: Supplementary; Desc: Description.

Type	Token Length	Total Num
Raw IM	21.72	177,664
Refined IM	6.70	257,028
SP Knowledge	23.72	58,827
Visual Desc	15.30	135,720

ple imagery meanings and textual noise are combined into a single query, leading to a weak correspondence between the meanings and the retrieved images. Some studies (Liu et al., 2025) attempt to remove textual noise using regular expressions, but they struggle to cover all cases in large-scale datasets. Therefore, this paper uses LLMs to filter noise, separate complex raw imagery meanings into distinct meanings, and retain useful auxiliary information as supplementary knowledge, producing refined imagery meanings.

Table 1 shows that the refined imagery meanings often become overly concise(6.70), failing to achieve the optimal token length. Intuitively, providing detailed descriptions within the model's comprehension range enhances the accuracy of the generated images. For example, prompts like "The sea god" are too concise, whereas "The majestic sea god stands above the waves" provides clearer, more interpretable context for the generative model. Additionally, there is a significant difference between the semantics of the text and the visual descriptions, and using textual semantics directly as prompts often fails to generate images that meet expectations(Betker et al., 2023). Drawing inspiration from Retrieval Augmented Generation(RAG) technologies, we utilize LLMs to rewrite refined imagery meanings into visual descriptions suitable for generative models. It enhances clarity and relevance while keeping the prompt length within a manageable 20 tokens, effectively reducing the likelihood of hallucinations.

LLMs are also used to determine if an imagery meaning can be visually represented, discarding inputs like stopwords that lack visual significance. This ensures that only visually meaningful data is processed by the generative models, enhancing efficiency. The instruction is shown in Appendix G.

3.4 Imagery Meaning-Imagery Image Alignment.

When handling large-scale text-image pairs, accurately aligning them is a significant challenge. It is common to use the CLIPScore(Hessel et al., 2021) to evaluate the relevance between text and images. For a given image encoding v, caption encoding c, and scaling factor w, the CLIPScore formula is as follows:

$$\text{CLIPScore}(c, v) = w \cdot \max(\cos(c, v), 0)$$

CLIPScore has certain limitations (Schuhmann et al., 2022): a high threshold may lead to the omission of entities, while a low threshold can weaken alignment between text and images, particularly with large-scale text-image pairs. Inspired by GLIDE (Nichol et al., 2022), which evaluates image generation quality through classification, we abandon the traditional threshold-setting approach. Instead, we propose leveraging an image-to-text retrieval task to address this alignment challenge.

In the image-to-text retrieval task, text perturbations are introduced. Specifically, for each generated imagery image, the imagery image is used to retrieve imagery meaning along with the two text perturbations. The first perturbation is a randomly selected imagery meaning from the total set, while the second perturbation is composed of a random character selected from the tokenization vocabulary in BERT(Devlin et al., 2019). If all generated imagery images correctly retrieve the imagery meaning, the imagery images and imagery meaning are considered aligned, and the text-image pair with the highest CLIP score is selected as the final match. Otherwise, those text-image pairs are deemed mismatched and discarded.

4 Qualitative and Quantitative Evaluations

To assess the quality of CPMK, we conduct both qualitative and quantitative evaluations. The details of qualitative evaluations are in Appendix D.

4.1 Quantitative Evaluations

(1)We counted the number of entities in each dataset, with the results presented in Table 2. To our knowledge, CPMK is the first dataset to integrate text, vision, and audio modalities within classical Chinese poetry. According to the table, CPMK significantly exceeds prior research in the number of entities. Large-scale datasets facilitate

research and application development in classical Chinese poetry.

Table 2: Modal entity statistics across datasets. Results with * are inferred from their papers due to dataset unavailability.

Corpus	#Text	#Vision	#Audio
CCPM[17]	136,090	-	-
RPG*[10]	215,227	-	-
VMIJUNV[33]	1,515,463	-	-
ACP-Corpus*[20]	2,159,920	-	-
PKG[18]	1,115,143	96,049	-
Image2Poem*[19]	117,867	1,036	-
CPMK	6,834,825	211,467	82,679

(2)To evaluate the effectiveness of the proposed prompt optimization and text-image alignment methods, we design a comparative evaluation. The raw imagery meanings are optimized using a heuristic approach rather than prompt optimazation as prompts to generate new images, which are then refined through our text-image alignment method. The image generation part is the same as that used in this paper. We record the average token length of the text processed by the heuristic methods and calculate the CLIPScore[7] of the generated images. The details of the heuristic approach are shown in Appendix H.

Table 3: Average CLIPScore for semantic alignment. Abbreviations —Heu: Heuristic; PO: Prompt Optimization; Align: Our Proposed Method; Max: Maximum score among all generated pairs.

Data	CLIPScore	Total Num
Heu	0.912	752,772
Heu + Align	1.056	306,270
Heu + Align + Max	1.068	102,090
PO	1.022	407,160
PO + Align	1.136	319,419
PO + Align + Max	1.191	106,473

The results in Table 3 demonstrate that images processed by prompt optimization have a higher CLIPScore compared to the heuristic approach. By simply reducing the number of imagery meaning-imagery image pairs from 407,160 to 319,419, text-image alignment significantly improved CLIPScore, demonstrating its effectiveness. Notably, the final counts of imagery meaning-imagery image pairs obtained through the heuristic approach (102,090) and prompt optimization (106,473) are very close. It suggests that LLMs with visual capabilities can effectively determine whether an imagery meaning is visually representable, enhancing computational efficiency. It also demonstrates that LLMs can rewrite text prompts for image generation.

5 Downstream Task Validation

To validate the effectiveness of CPMK in downstream tasks, we conduct a preliminary experiment on four downstream tasks. We conduct a poetry theme classification task and a poetry question answering task to validate the effectiveness of textual knowledge in CPMK. the poetry-image retrieval task is validated for visual knowledge. To represent phonological knowledge, this paper uses Pinyin as a textual proxy for audio during validation. This approach is validated through the rigid-format poetry generation task. Since both poetry question answering and poetry theme classification adopt RAG techniques, without loss of generality, we introduce poetry question answering in Appendix A.

5.1 Poetry Theme Classification

The Poetry Theme Classification task aims to categorize poems into the appropriate theme categories. We use the RAG technique to apply CPMK and PKG in this task and evaluate their performance using the TCCP[8] dataset. TCCP is a theme classification dataset for Chinese classical poetry, containing 3,247 poems across 9 categories. We use poetry imagery as a query to retrieve relevant imagery meanings from CPMK and PKG, then combine this knowledge with the original poem and input it into different models for poetry theme classification. Given that poetry imagery in CPMK has multiple meanings, we use an LLM to select the appropriate one based on context. However, because PKG does not distinguish between these meanings, all interpretations are directly input into the model as retrieval knowledge. Details are in Appendix B.

The results indicate that both CPMK and PKG enhance the model's classification capability. However, CPMK demonstrates a more significant improvement, achieving state-of-the-art results. This suggests that when inputting the same type of knowledge, CPMK provides both higher accuracy and broader coverage than PKG.

[7] In this paper, we calculate CLIPScore using the CN-CLIP-1B model (Yang et al., 2022).

[8] https://github.com/shuizhonghaitong/classification_GAT/tree/master/data

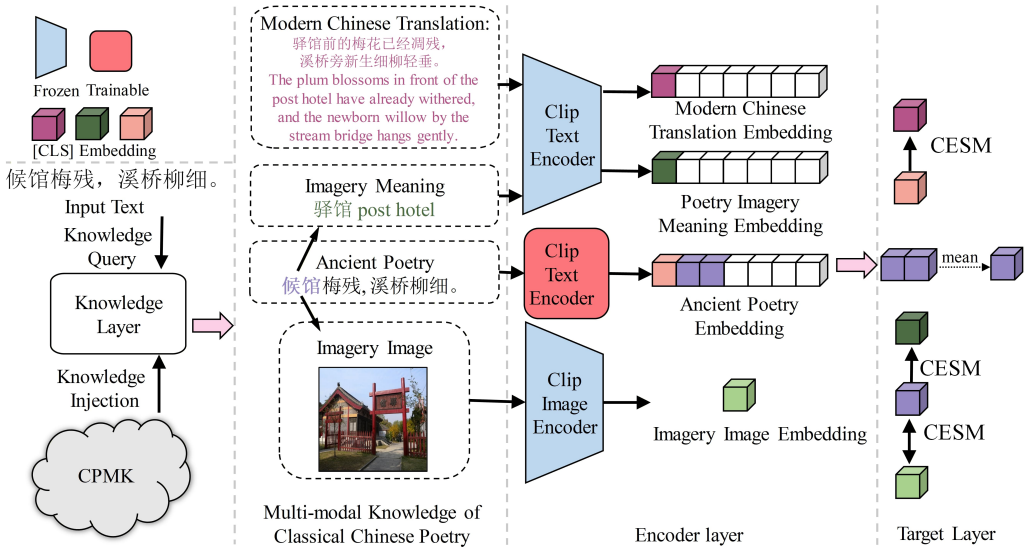


Figure 3: The overview framework of KPIR.

Table 4: The performance of different LLMs with PKG and CPMK.

Models	Micro-F1	Macro-F1
gemma2:2B[32]	24.28	23.36
gemma2:2B + PKG[32]	26.75	22.07
gemma2:2B + CPMK[32]	31.28	34.23
qwen2.5:3B[44]	39.51	42.63
qwen2.5:3B + PKG[44]	41.98	44.15
qwen2.5:3B + CPMK[44]	48.15	50.15
deepseek-r1:7B[6]	73.66	68.85
deepseek-r1:7B + PKG[6]	71.60	68.89
deepseek-r1:7B + CPMK[6]	75.72	69.90

5.2 Poetry-Image Retrieval

Existing retrieval models struggle with classical Chinese poetry due to domain-specific knowledge gaps and limited multimodal datasets. To address this, we propose KPIR (Knowledge-enhanced Poetry-Image Retrieval, Fig 3). Because the current retrieval model establishes a correspondence between modern Chinese and images, we leverage this by associating the encoding of the ancient poetry with its MCT to bridge the semantic gap between poem and image(Image $\leftrightarrow$ MCT $\leftarrow$ Ancient Poetry).

KPIR leverages expertise from CPMK, encompassing poetry imagery, imagery meaning, imagery image, and MCT. The architecture employs three encoders: a frozen text encoder for MCT (f_{mct}) and imagery meaning (f_{im}), a trainable text encoder for ancient poetry (f_{ap}) and poetry imagery (f_{pi}), and a frozen image encoder for imagery image (f_{ii}). Knowledge injection is achieved via Cross-Entropy Similarity Matching (CESM), which aligns bimodal embedding distributions by integrating similarity scores into a cross-entropy framework.

Given a mini-batch containing N bimodal (X, Y) pairs, where Y includes poetry knowledge from CPMK (MCT, imagery meaning, imagery image), based on either the ancient poetry or the poetry imagery in X. We form representation pairs $\{(f_i^x, f_j^y), y_{i,j}\}$ with labels $y_{i,j}$: 1 for matching pairs and 0 for non-matching ones. In a mini-batch, the CESM loss from modality X to Y is:

$$p_{i,j} = \frac{\exp(\text{sim}(f_i^x, f_j^y))}{\sum_{k=1}^{N} \exp(\text{sim}(f_i^x, f_k^y))}$$

$$\text{CESM}(X, Y) = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{N} y_{i,j} \log(p_{i,j})$$

We applied the CESM in four stages: $\mathcal{L}_{ap2mct} = \text{CESM}(f_{ap}, f_{mct})$, $\mathcal{L}_{pi2im} = \text{CESM}(f_{pi}, f_{im})$, and $\mathcal{L}_{pi2ii} = \text{CESM}(f_{pi}, f_{ii})$,$\mathcal{L}_{ii2pi} = \text{CESM}(f_{ii}, f_{pi})$. Knowledge is injected through MCT($\mathcal{L}_{ap2mct}$) and imagery meaning($\mathcal{L}_{pi2im}$). We utilize $\mathcal{L}_{ii2pi}$ and $\mathcal{L}_{pi2ii}$ to preserve the correspondence between text and image. The final loss function is:

$$\mathcal{L} = \mathcal{L}_{ap2mct} + \mathcal{L}_{pi2im} + \mathcal{L}_{pi2ii} + \mathcal{L}_{ii2pi}$$

Dataset and Evaluation Metrics. We extract 30,000 pairs of ancient poetry and MCT from CPMK as the training dataset, with all poetry

Table 5: Performance of Different Models on the Poetry-Image Retrieval Task

Models	PI-Manual		PI-Generate					
	t2i	i2t	t2i			i2t		
	R@3	R@3	R@5	R@10	R@20	R@5	R@10	R@20
Taisu-0.2B[21]	0.714	0.700	0.167	0.217	0.260	0.290	0.340	0.375
AltClip-0.9B[5]	0.601	0.671	0.236	0.293	0.375	0.249	0.331	0.418
R2D2-0.4B[41]	0.586	0.814	0.124	0.161	0.224	0.269	0.359	0.444
CN-CLIP-0.4B[43]	0.686	0.771	0.179	0.236	0.323	0.282	0.365	0.452
KPIR-0.4B	**0.914**	**0.857**	**0.404**	**0.507**	**0.602**	**0.348**	**0.445**	**0.546**

imagery-related knowledge sourced from CPMK. Due to the lack of existing datasets for poetry-image retrieval tasks, we construct two datasets for evaluation: 1) We manually collected 70 high-quality pairs of ancient poetry and image (PI-Manual) from the internet. 2) We generate 1000 images corresponding to ancient poetry using a generative model (PI-Generate), with specific generation details provided in the appendix C.2. In addition, we generated a validation set of 500 images from ancient poems, following the aforementioned methodology but using an alternative generative model. This study uses recall as the evaluation metric, counting the number of correct answers within the retrieved set.

Main results. The experimental results on poetry-image retrieval demonstrate that KPIR achieves state-of-the-art performance across two datasets. Our model, KPIR-0.4B initialized using the CN-CLIP(Yang et al., 2022) model, significantly surpasses previous methods. To explore the role of each component in CPMK and the effectiveness of the KPIR framework, we conducted additional experiments in Appendix C.

5.3 Rigid-Formats Poetry Generation

Rigid-Formats Poetry Generation involves creating poems that adhere to rigid rhythmic formats. However, the scarcity of audio knowledge for ancient Chinese poetry limits the development of prosody research. To evaluate current resources, we test three popular Pinyin conversion tools on the CPMK dataset (containing 50,769 Pinyin characters). The coverage results are summarized in Table 6.

Due to CPMK's rich audio knowledge, we can easily expand the training set for the Rigid-Formats Poetry Generation task in (Li et al., 2020), thereby enhancing the model's rhythm perception. We supplement the original 19,244 training entries with an additional 20,000 ancient poems annotated with pinyin and evaluated on their test set. The results are shown in Table 7. The results indicate that augmenting the training data with CPMK's audio knowledge improves the model's performance, demonstrating the effectiveness of CPMK's phonological knowledge.

Table 6: Coverage of Pinyin conversion tools on the CPMK dataset.

Tool	Count	Coverage (%)
Pypinyin[25]	37,684	74.22
Pinyin[24]	26,227	51.66
Xpinyin[23]	25,368	49.97

Table 7: Performance comparison between SongNet and SongNet-Aug.

Category	Metric	SongNet	SongNet-Aug
Format ↑	MA-F1	99.16	**99.84**
	MI-F1	99.17	**99.82**
Rhyme ↑	MA-F1	81.66	**84.41**
	MI-F1	81.19	**83.55**
Diversity ↑	MA-D-1	79.70	**80.67**
	MI-D-1	3.50	**3.75**
	MA-D-2	97.60	**98.15**
	MI-D-2	40.00	**40.81**

6 Conclusion

This paper proposes a method for constructing an MM-KG for classical Chinese poetry, integrating textual, visual, and auditory modalities. By introducing knowledge augmentation, we ensure textual data completeness. We enhance the correlation between text and images through prompt optimization and text-image alignment. Qualitative evaluation, quantitative evaluation, and downstream tasks evaluation validate the quality and effectiveness of CPMK.

7 Limitation

While this study makes strides in alleviating the data scarcity challenges associated with constructing MM-KGs for classical Chinese poetry, several limitations persist.

(1) Due to constraints in human labor and computational power, constructing a large-scale, ground-truth image dataset comparable to ImageNet remains currently infeasible. Consequently, we rely on synthetic data, though we acknowledge that such an image is inherently constrained by the generative models' internal representations and parameter spaces. This may introduce a cascading bias—where generated content reflects model-specific interpretations rather than the true essence of the imagery—potentially limiting the generalization boundaries of multi-modal representations. To mitigate these biases, we utilized the highest-performing models available and invested substantial computational resources, including a generation phase exceeding one month, to ensure optimal representational accuracy.

(2)While the effectiveness of the audio modality is evidenced by improvements in phonetic tools and data augmentation, a direct evaluation of its impact on TTS(Text-to-Speech) performance or human perception remains a limitation of this work. Nevertheless, the fine-grained, character-level audio data in CPMK makes such a direction highly feasible. This granularity is essential for modeling continuous speech from discrete characters—a task that would be significantly more challenging without our dataset.

(3) The textual data primarily rely on open-domain internet resources, which may lack the rigorous proofreading and authority found in official classical archives. This is a pervasive challenge in contemporary knowledge graph engineering that may impact the absolute factual integrity of the knowledge base.

(4) Although we have validated the effectiveness of CPMK through four tasks, these benchmarks remain relatively fundamental. Future research will focus on designing more sophisticated downstream applications to further explore the high-level cognitive value of the CPMK.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, and 1 others. 2023. Improving image generation with better captions. Technical report, OpenAI.

Jiahuan Cao, Yang Liu, Yongxin Shi, Kai Ding, and Lianwen Jin. 2024. Wenmind: A comprehensive benchmark for evaluating large language models in chinese classical literature and language arts. *Advances in Neural Information Processing Systems*, 37:51358–51410.

Zhongzhi Chen, Guang Liu, Bo-Wen Zhang, Qinghong Yang, and Ledell Wu. 2023. AltCLIP: Altering the language encoder in clip for extended language capabilities. In *Findings of the Association for Computational Linguistics*, pages 8666–8682.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv: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*, pages 4171–4186.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. GLM: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 320–335.

Zirui Guo, Lianghao Xia, Yanhua Yu, Tu Ao, and Chao Huang. 2024. Lightrag: Simple and fast retrieval-augmented generation. *arXiv preprint arXiv:2410.05779*.

Ming He, Yan Chen, Hong-Ke Zhao, Qi Liu, Le Wu, Yu Cui, Gui-Hua Zeng, and Gui-Quan Liu. 2023. Composing like an ancient chinese poet: Learn to generate rhythmic chinese poetry. *Journal of Computer Science and Technology*, 38(6):1272–1287.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528.

Liang Hong, Wenjun Hou, and Lina Zhou. 2020. Knowpoetry: A knowledge service platform for tang poetry research based on domain-specific knowledge graph. *Library Trends*, 69:101–124.

IDEA-CCNL. 2024. Idea-ccnl ziya-llama-13b-v1.1. https://huggingface.co/IDEA-CCNL/Ziya-LLaMA-13B-v1.1.

Jing Jiang, Yiran Ling, Binzhu Li, Pengxiang Li, Junming Piao, and Yu Zhang. 2024. Poetry2image: An iterative correction framework for images generated from chinese classical poetry. *arXiv preprint arXiv:2407.06196*.

Diederik P Kingma. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Piji Li, Haisong Zhang, Xiaojiang Liu, and Shuming Shi. 2020. Rigid formats controlled text generation. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 742–751.

Wenhao Li, Fanchao Qi, Maosong Sun, Xiaoyuan Yi, and Jiarui Zhang. 2021. Ccpm: A chinese classical poetry matching dataset. *arXiv preprint arXiv:2106.01979*.

Yuqing Li, Yuxin Zhang, Bin Wu, Ji-Rong Wen, Ruihua Song, and Ting Bai. 2022. A multi-modal knowledge graph for classical chinese poetry. In *Findings of the Association for Computational Linguistics*, pages 2318–2326.

Lixin Liu, Xiaojun Wan, and Zongming Guo. 2018. Images2poem: Generating chinese poetry from image streams. In *Proceedings of the 26th ACM international conference on Multimedia*, pages 1967–1975.

Yang Liu, Lan Lan, Jiahuan Cao, Hiuyi Cheng, Kai Ding, and Lianwen Jin. 2025. Large-scale corpus construction and retrieval-augmented generation for ancient chinese poetry: New method and data insights. In *Findings of the Association for Computational Linguistics*, pages 779–817.

Yulong Liu, Guibo Zhu, Bin Zhu, Qi Song, Guojing Ge, Haoran Chen, GuanHui Qiao, Ru Peng, Lingxiang Wu, and Jinqiao Wang. 2022. Taisu: A 166m large-scale high-quality dataset for chinese vision-language pre-training. *Advances in Neural Information Processing Systems*, 35:16705–16717.

Yusen Liu, Dayiheng Liu, Jiancheng Lv, and Yongsheng Sang. 2020. Generating chinese poetry from images via concrete and abstract information. In *2020 International Joint Conference on Neural Networks*, pages 1–8.

lxneng. 2025. xpinyin. https://pypi.org/project/xpinyin/.

lxyu. 2016. pinyin. https://pypi.org/project/pinyin/.

mozillazg. 2025. pypinyin. https://pypi.org/project/pypinyin.

Alexander Quinn Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob Mcgrew, Ilya Sutskever, and Mark Chen. 2022. GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models. In *Proceedings of the 39th International Conference on Machine Learning*, pages 16784–16804.

PeterH0323. 2024. Peterh0323 ancient-chat-llm. https://github.com/PeterH0323/ancient-chat-llm.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, and 1 others. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294.

shenzhi wang. 2025. shenzhi-wang llama3-8b-chinese-chat. https://huggingface.co/shenzhi-wang/Llama3-8B-Chinese-Chat.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, and 179 others. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

VMIJUNV. 2023. Chinese-poetry-and-prose. https://github.com/VMIJUNV/Chinese-poetry-and-prose.

Xueyao Wan and Hang Yu. 2025. Mmgraphrag: Bridging vision and language with interpretable multimodal knowledge graphs. *arXiv preprint arXiv:2507.20804*.

Chengyu Wang, Zhongjie Duan, Bingyan Liu, Xinyi Zou, Cen Chen, Kui Jia, and Jun Huang. 2024. PAI-diffusion: Constructing and serving a family of open Chinese diffusion models for text-to-image synthesis on the cloud. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, pages 1–8.

Qing Wang, Weiping Liu, Xiumei Wang, Xinghong Chen, Guannan Chen, and Qingxiang Wu. 2023. A spatial – temporal graph model for pronunciation feature prediction of chinese poetry. *IEEE Transactions on Neural Networks and Learning Systems*, 34(12):10294–10308.

Qingwang Wang, Chaohui Li, Yi Liu, Qiubai Zhu, Jian Song, and Tao Shen. 2025. An adaptive framework embedded with llm for knowledge graph construction. *IEEE Transactions on Multimedia*, 27:2912–2923.

Yuting Wei, Linmei Hu, Yangfu Zhu, Jiaqi Zhao, and Bin Wu. 2024. Knowledge-guided transformer for joint theme and emotion classification of chinese classical poetry. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 32:4783–4794.

Werneror. 2017. Poetry. https://github.com/Werneror/Poetry.

Xiaojun Wu, Dixiang Zhang, Ruyi Gan, Junyu Lu, Ziwei Wu, Renliang Sun, Jiaxing Zhang, Pingjian Zhang, and Yan Song. 2024. Taiyi-diffusion-xl: advancing bilingual text-to-image generation with large vision-language model support. *arXiv preprint arXiv:2401.14688*.

Chunyu Xie, Heng Cai, Jincheng Li, Fanjing Kong, Xiaoyu Wu, Jianfei Song, Henrique Morimitsu, Lin Yao, Dexin Wang, Xiangzheng Zhang, and 1 others. 2023. CCMB: A large-scale chinese cross-modal benchmark. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 4219–4227.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, and 1 others. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

An Yang, Junshu Pan, Junyang Lin, Rui Men, Yichang Zhang, Jingren Zhou, and Chang Zhou. 2022. Chinese clip: Contrastive vision-language pretraining in chinese. *arXiv preprint arXiv:2211.01335*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 43 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Fulong Ye, Guang Liu, Xinya Wu, and Ledell Wu. 2024. Altdiffusion: A multilingual text-to-image diffusion model. In *Proceedings of the AAAI conference on artificial intelligence*, pages 6648–6656.

Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Guoyin Wang, Heng Li, Jiangcheng Zhu, Jianqun Chen, and 1 others. 2024. Yi: Open foundation models by 01. ai. *arXiv preprint arXiv:2403.04652*.

Changlong Yu, Weiqi Wang, Xin Liu, Jiaxin Bai, Yangqiu Song, Zheng Li, Yifan Gao, Tianyu Cao, and Bing Yin. 2023. FolkScope: Intention knowledge graph construction for E-commerce commonsense discovery. In *Findings of the Association for Computational Linguistics*, pages 1173–1191.

Beichen Zhang, Pan Zhang, Xiaoyi Dong, Yuhang Zang, and Jiaqi Wang. 2024. Long-clip: Unlocking the long-text capability of clip. In *European Conference on Computer Vision*, pages 310–325.

Jiaxing Zhang, Ruyi Gan, Junjie Wang, Yuxiang Zhang, Lin Zhang, Ping Yang, Xinyu Gao, Ziwei Wu, Xiaoqun Dong, Junqing He, and 1 others. 2022. Fengshenbang 1.0: Being the foundation of chinese cognitive intelligence. *arXiv preprint arXiv:2209.02970*.

Jiaqi Zhao, Ting Bai, Yuting Wei, and Bin Wu. 2022. Poetrybert: Pre-training with sememe knowledge for classical chinese poetry. In *International conference on data mining and big data*, pages 369–384.

Jialong Zuo, Haoyou Deng, Hanyu Zhou, Jiaxin Zhu, Yicheng Zhang, Yiwei Zhang, Yongxin Yan, Kaixing Huang, Weisen Chen, Yongtai Deng, Rui Jin, Nong Sang, and Changxin Gao. 2025. Is nano banana pro a low-level vision all-rounder? a comprehensive evaluation on 14 tasks and 40 datasets. *arXiv preprint arXiv:2512.15110*.

Appendices

A Poetry Question Answering Task

To validate the comprehensiveness of the data in CPMK, we applied it to the poetry question answering task. Following the traditional RAG framework, we use regular expressions to extract the author, title, and ancient poetry as keywords from the original query. Like LightRAG(Guo et al., 2024), we use keywords as query conditions to retrieve related knowledge from CPMK. The retrieved knowledge is combined with the original query and fed into the LLM to answer questions.

We evaluate five tasks related to classical poetry in WenMind(Cao et al., 2024), including Basic Q&A (T1), Ancient Poetry Translation (T2), Sentiment Classification (T3), Ancient Poetry to English (T4), and Poet Introduction (T5), totaling 1,310 questions. T1 involves questions about basic knowledge of ancient poetry, such as identifying the title and author based on the content. T2 is about translating ancient poetry into modern Chinese. T3 deals with sentiment classification of the poetry. T4 involves translating the poetry into English. T5 provides an introduction to the poet. Since the knowledge in PKG covers only poetry imagery and cannot support most of the

Table 8: The performance of different models, with * indicating results cited from original paper.

Models	T1	T2	T3	T4	T5
ChatGLM3-6B*[8]	10.6	55.5	43.0	44.9	33.3
Ancient-Chat-7B*[27]	14.7	52.7	36.0	28.6	23.9
LLaMA3-Chinese-8B*[31]	1.7	62.4	42.5	52.3	23.4
Baichuan2-13B-Chat*[42]	20.1	66.9	43.0	51.3	55.4
Ziya-LLaMA-13B*[13]	6.4	57.5	40.5	40.2	31.6
Qwen1.5-32B-Chat*[2]	32.0	67.9	64.0	54.7	58.1
Yi-1.5-34B-Chat*[46]	30.5	69.0	53.5	54.2	64.6
ChatGPT-4[1]	23.6	**75.9**	61.3	66.3	44.1
ChatGPT-4-RAG	**73.4**	75.8	**64.3**	**71.2**	**65.9**

tasks mentioned above, this study uses only CPMK for the experiments. We conduct experiments on ChatGPT-4 using the same model-scoring metric as WenMind. In this paper, the LLM used is kept at default settings without any additional adjustments to the model's temperature or context length, unless specifically emphasized otherwise.

As shown in Table 8, our ChatGPT-4-RAG demonstrates strong performance across most tasks, thanks to the high-quality CPMK, which enhances the model's understanding of ancient poetry. However, its performance on T2 is inferior to ChatGPT-4's, likely due to overlap between WenMind's internet-based dataset and the extensive datasets used to train current LLMs. We think the results are sufficient to illustrate that the CPMK encompasses a broad range of knowledge. This study employed a simple RAG framework without task-specific adjustments and still achieved significant performance improvements across most tasks. The experimental results demonstrate that CPMK is of high quality and can provide the model with better knowledge related to classical Chinese poetry.

B Details of Poetry Theme Classification Task

B.1 Implement Details

For the TCCP dataset, we split it into training, validation, and test sets at 7:2:1. Its theme is divided into nine categories: homesickness, chanting things, landscape, missing someone, meditating on the history, pastoral, frontier war, boudoir resentment, and farewell. To enhance the model's ability to comprehend and analyze problems, we ask it not only to answer questions but also to provide explanations for its answers. Given that the TCCP dataset contains nine categories, we provide one learning example per category, leveraging few-shot learning to improve the model's performance. The instruction to LLM with RAG is

You are an expert in Chinese classical poetry. The user will provide a series of questions related to themes of classical poems. Your answer categories should be within ["homesickness", "chanting things", "landscape", "missing someone", "meditating on the history", "pastoral", "frontier war", "boudoir resentment", and "farewell"]. The answer should only include the category and an explanation for your choice. For example, respond in the format: {"Answer": "boudoir resentment", "Explanation": "This poem expresses longing for a loved one far away, which is typical of the "boudoir resentment" category."}

The instruction is followed by nine learning examples, each corresponding to one of the nine categories. To establish a standard evaluation, we use ChatGPT-4 to generate explanations for few-shot learning. To investigate how differences in imagery knowledge between CPMK and PKG affect poetry theme classification, our prompts specifically instruct the LLM to analyze the relationship between poetry imagery, imagery meaning, and the theme classification results. The instruction to ChatGPT-4 is

As an expert in Classical Chinese poetry, you need to analyze the poetry theme classification problem provided by the user, along with the classification result. By considering the relevant poetry imagery and imagery meaning in the poem, explain why the poem fits the given theme and summarize the final result in a single paragraph.

B.2 Case Study

We present case studies for both CPMK and PKG using DeepSeek-R1:7B in Figures 4 and 5. Small-scale LLMs (2B, 3B, and 7B) are selected for

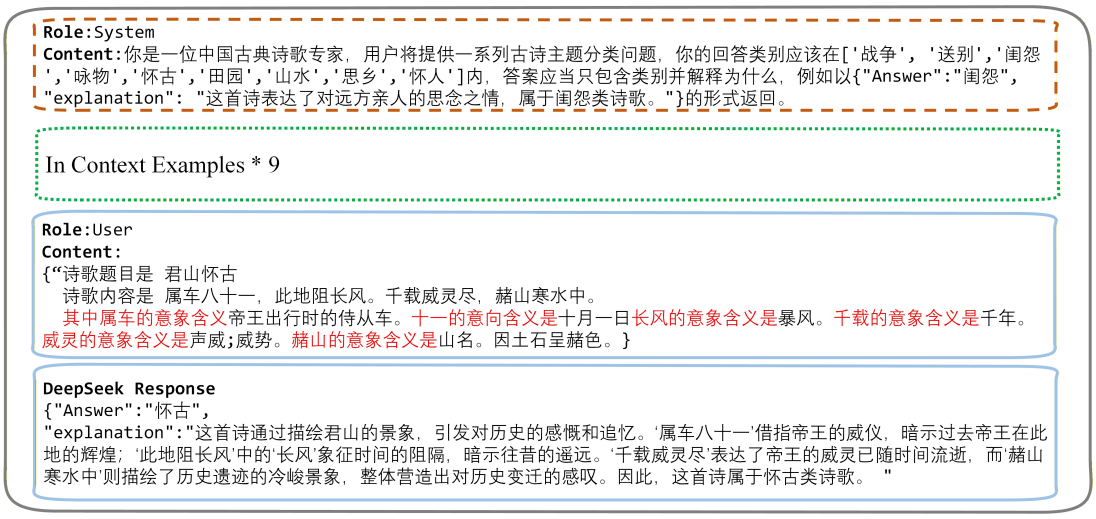


Figure 4: Case study of DeepSeek-R1:7B classification using CPMK.

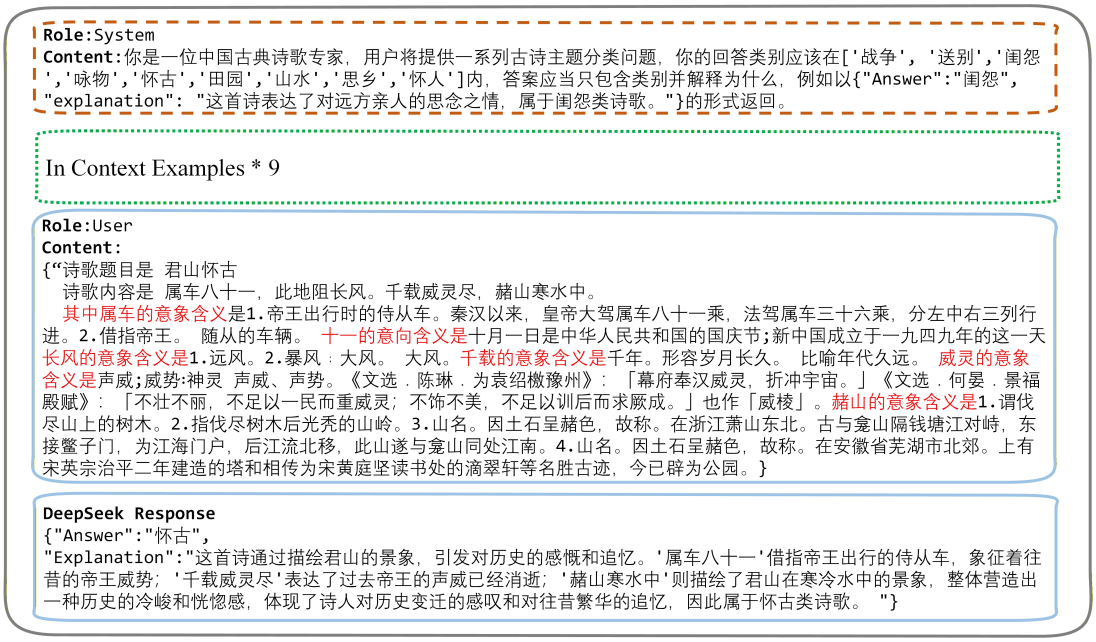


Figure 5: Case study of DeepSeek-R1:7B classification using PKG.

theme classification, as their relatively lower noise robustness should theoretically highlight the quality disparities between the two methods. Interestingly, we observed that these smaller models still possess a surprising capacity to extract meaningful insights from irrelevant text, which narrows the observed performance gap. Nonetheless, CPMK consistently delivers more precise imagery knowledge, thereby facilitating a more accurate understanding of poetry themes.

C Details of Poetry-Image Retrieval Task

C.1 Implement Details

KPIR is initialized using the CN-CLIP (Yang et al., 2022) model. We evaluate the performance of the baseline CN-CLIP and our KPIR across three scales, including 0.1B, 0.4B, and 1B parameter versions(in Figure 9:*Comparison Across Different Parameter Sizes*).

We use the Adam optimizer (Kingma, 2014) with a weight decay rate of 0.01 and a learning rate of 2e-5. The random seed is set to 123. The batch size is set to 128. Since our CESM model involves mini-batch comparisons, we shuffle the training set at the end of each epoch. Our exper-

As an expert in Chinese classical poetry, please help me extract scene and emotion description from the translation and appreciation of Chinese classical poetry to use as prompts for image generation.
Input: {"translation":"Translation of the poetry",
"appreciation":"Appreciation of the poetry"}
Output: {"scene_description":"Scene description from the poetry",
"emotion_ description ":"Emotion description from the poetry"}
Requirements:
1. The scene and emotion description should be concise and clear.
2. The scene and emotion should be suitable as prompts for image generation models.
Example:
Input: {"translation":"自古以来，人终不免一死！倘若能为国尽忠，死后仍可光照千秋，青史留名。",
"appreciation":"此句悲壮激昂、掷地有声，以磅礴的气势、高亢的语调显示了诗人的民族气节和舍生取义的生死观，表达了诗人赤诚的爱国情怀和视死如归的崇高精神，激励了无数的爱国之士为了民族大业而抛头颅、洒热血。"}
Output: {
"scene_ description ":"壮烈的战场，战士奋勇牺牲，历史长河中的英雄光辉."
“emotion_ description " : "悲壮激昂的爱国激情，视死如归的无畏精神。"}

Figure 6: Instructions for generating emotion description and scene description.

iments are conducted on an RTX 5090 GPU with 32 GB of memory.

C.2 PI-Generate dataset Construction

We detail the process for using LLMs to generate the poetry-image retrieval dataset(PI-Generate). Given the critical role of emotion in ancient poetry, we distinguish between scene and emotion descriptions to ensure the generated images accurately reflect the poems' intended meanings. Specifically, this study retrieves modern Chinese translations and poetry appreciations of ancient poetry from CPMK. These texts are then refined using ChatGPT-4 (Achiam et al., 2023), which generates tailored prompts that encapsulate both scene and emotion descriptions for image generation. To mitigate potential biases introduced by a single model architecture, we employed distinct generative models for the validation and test sets. The validation set consists of 500 poem-image pairs generated via DALL • E 3 (Betker et al., 2023), while the test set comprises 1,000 pairs generated using Nano Banana Pro (Zuo et al., 2025). The detailed instructions for extracting scene and emotion descriptions are illustrated in Figure 6.

The scene and emotion descriptions are combined to form the final prompt for image generation, following the instruction provided below: *{“scene_description”: “scene_description”, “emotion_description”: “emotion_description”}*

C.3 Poetry-Image Retrieval Task Ablation Study

We conduct an ablation study on KPIR-0.1B to demonstrate the impact of each loss function on the poetry-image retrieval task. The experimental results are shown in Table 9: *Ablation Study on Loss Functions*. Without loss of generality, we conduct ablation experiments on the CN-CLIP-0.1B model.

The results indicate that incorporating both global knowledge of poetry translation and local knowledge of poetry imagery significantly improves the model's performance, particularly enhancing its ability to retrieve text from images while maintaining its poetry-image retrieval capability as much as possible. Even when trained solely on text, the model's performance also improves. We attribute this improvement to preserving text-image correspondence in CN-CLIP during fine-tuning, as well as to our focus on fine-tuning the textual side, which correctly outputs poetry embeddings. Additionally, $\mathcal{L}_{pi2im}$ plays a crucial role in the knowledge injection process. Given the abundant presence of poetry imagery, aligning it with its meaning enables the model to understand poetry at a fine-grained level accurately.

C.4 Experiment on PKG

To investigate whether the superior performance of the KPIR framework stems from its architectural design or the high-quality knowledge within

Table 9: Experimental results of KPIR on poetry retrieval: Ablation study, task-specific evaluation (PKG & MCT), and model scaling comparison. **Bold** and underlined values denote the best and second-best performance.

Method / Configurations				PI-Manual		PI-Generate					
				t2i	i2t	t2i			i2t		
				R@3	R@3	R@5	R@10	R@20	R@5	R@10	R@20
$\mathcal{L}_{pi2ii}$	$\mathcal{L}_{ii2pi}$	$\mathcal{L}_{pi2im}$	$\mathcal{L}_{ap2mct}$	*Ablation Study on Loss Functions*							
		✓	✓	0.914	0.728	0.333	0.420	0.515	0.209	0.300	0.379
	✓	✓	✓	0.857	0.743	0.319	0.410	0.511	0.214	0.283	0.387
✓		✓	✓	0.914	0.743	0.327	0.442	0.525	0.211	0.287	0.391
✓	✓		✓	0.914	0.743	0.322	0.433	0.537	0.213	0.301	0.391
✓	✓	✓		0.786	0.814	0.257	0.336	0.433	0.232	0.302	0.403
✓	✓	✓	✓	0.886	0.771	0.328	0.414	0.510	0.240	0.324	0.441
$\mathcal{L}_{pi2ii}$	$\mathcal{L}_{ii2pi}$	$\mathcal{L}_{pi2im}$	$\mathcal{L}_{ap2mct}$	*Experiment on PKG(MCT is from CPMK)*							
✓	✓	✓		0.686	0.743	0.193	0.273	0.357	0.201	0.252	0.327
		✓	✓	0.823	0.700	0.315	0.403	0.495	0.203	0.290	0.376
✓	✓	✓	✓	0.786	0.757	0.260	0.357	0.449	0.216	0.289	0.370
MCT-retrieval Task											
CN-CLIP-0.1B [43]				0.871	**0.900**	0.300	0.381	0.465	0.324	0.409	0.505
CN-CLIP-0.4B [43]				0.886	0.886	0.328	0.400	0.494	**0.377**	0.454	0.541
CN-CLIP-1B [43]				0.857	**0.900**	0.330	0.403	0.491	0.374	**0.463**	0.550
Comparison Across Different Parameter Sizes											
CN-CLIP-0.1B [43]				0.714	0.729	0.171	0.233	0.298	0.216	0.282	0.387
CN-CLIP-0.4B [43]				0.686	0.771	0.179	0.236	0.323	0.282	0.365	0.452
CN-CLIP-1B [43]				0.743	0.714	0.201	0.268	0.339	0.279	0.347	0.429
Investigation of Single Point of Failure											
KPIR-0.1B_type1				0.826	0.714	0.316	0.389	0.482	0.218	0.303	0.388
KPIR-0.1B_type2				0.771	0.828	0.249	0.327	0.430	0.233	0.309	0.430
KPIR-0.1B (Ours)				0.886	0.771	0.328	0.414	0.510	0.240	0.324	0.441
KPIR-0.4B (Ours)				0.914	0.857	0.404	0.507	0.602	0.348	0.445	0.546
KPIR-1B (Ours)				**0.928**	**0.900**	**0.435**	**0.531**	**0.627**	0.356	0.460	**0.573**

CPMK, we conduct comparative experiments using the PKG. All experiments are initialized with CN-CLIP-0.1B, and the results are presented in Table 9: *Experiment on PKG*. Since PKG only contains poetry imagery, imagery meanings, and imagery images, it lacks the modern Chinese translation component—a core element of the KPIR framework. To ensure a fair comparison, we supplement PKG with translation knowledge from CPMK. For imagery images, due to server constraints and link decay among the 96,049 URLs, we successfully crawled 46,914 valid images.

Results show that:(1) Even when integrated with PKG knowledge, the model outperforms the CN-CLIP-0.1B, validating the effectiveness of the KPIR framework in multi-modal knowledge integration. (2) Under identical experimental settings, the performance using CPMK imagery knowledge consistently exceeds that of PKG. Even when supplemented with the same translation knowledge, the PKG-based variant still underperforms compared to the CPMK ablation results. This confirms that the CPMK knowledge offers superior semantic quality and better alignment for poetry retrieval.

C.5 Modern Chinese Translation-Image Retrieval Task

The core objective of KPIR is to bridge the semantic gap between ancient poetry and images by leveraging Modern Chinese Translation (MCT) as a pivot. To validate the effectiveness of this knowledge injection framework, we conduct a comparative experiment between KPIR's poetry-image retrieval and CN-CLIP's direct MCT-image retrieval. Specifically, we use modern translations of the poems as queries to perform bidirectional image retrieval using the base CN-CLIP model. The experimental results, as shown in Table 9:*MCT-retrieval Task*, reveal several key insights On the PI-Manual dataset, KPIR's poetry-image retrieval performance is superior to CN-CLIP's MCT-image retrieval. On the PI-Generate dataset: KPIR outperforms MCT-based retrieval in the Poetry-to-Image task, while in the Image-to-Poetry task, it

achieves comparable (though slightly lower) performance. These findings are particularly counterintuitive. Given that CN-CLIP is pre-trained on massive contemporary Chinese datasets, one would reasonably expect its performance on the MCT-image retrieval task to serve as a performance upper bound for KPIR. However, the fact that KPIR—using ancient poetry—can match or even surpass the performance of CN-CLIP using modern text indicates that its success is not solely reliant on MCT. Instead, it stems from the synergistic integration of both MCT and imagery-related knowledge, thereby confirming the rationality and superior capability of our proposed KPIR framework in complex cross-modal alignment.

C.6 Investigation of Single Point of Failure

Since KPIR utilizes Modern Chinese Translation (MCT) to bridge the gap between poetry and images, this study evaluates the translation quality to determine whether it constitutes a single point of failure. Specifically, we employed an LLM to rewrite the original MCTs of the 30,000 training samples into two distinct variants:

Type1(Mechanical): Translates metaphors as literal physical objects via character-by-character mapping.

Type2 (Unconstrained): Provides translations using an expert-persona prompt.

By comparing these two variants, we aim to quantify the model's sensitivity to the quality of semantic bridging and its robustness against potential translation inaccuracies. Both translation variants were generated using Qwen2.5-7B. We acknowledge that the overall translation quality is constrained by the model's mid-range parameter scale and its limited domain-specific expertise in classical Chinese poetry. Notably, the quality of Type1 was further suppressed by our intentional requirement for mechanical rigidity. To quantify this, we measured the semantic similarity between the generated translations and the ground truth using BGE-M3 embeddings. The cosine similarity scores for Type1 and Type2 were 0.712 and 0.740, respectively. These relatively moderate scores reflect a realistic 'sub-optimal' translation scenario, allowing us to rigorously evaluate how the KPIR performs when the semantic bridge is imperfect.

The prompt of Type1:

Act as a "Mechanical Literal Translator." Your sole objective is to translate Classical Chinese poetry into Modern Chinese by replacing each character with its most basic, primary dictionary definition.
Requirements:
1. Treat each Chinese character as an independent unit and translate it with its most fundamental Chinese equivalent.
2. Translate allusions and metaphors as literal physical objects. For example, "金雀" must be translated as "金色的鸟," not "头饰"; "朱雀桥" must be translated as "红色的雀鸟桥."
"Please output the result strictly in JSON format: {"translation": "mechanical literal Translation result here"}"

The prompt of Type2:

You are an expert in classical Chinese poetry. Your task is to translate Classical Chinese poetry into Modern Chinese.
"Please output the result strictly in JSON format: {"translation": "Modern Chinese Translation"}"

Despite the suboptimal quality of MCTs, KPIR still achieves performance gains. This resilience is primarily attributed to our dual-stream knowledge injection strategy, which combines Global (MCT) and Local (Imagery) knowledge. Specifically, while global translations may contain semantic biases, the granular, local poetry imagery-related knowledge acts as a corrective anchor, enabling the model to grasp the poem's meaning even when the MCT is inaccurate.

This synergy effectively mitigates the error propagation from flawed translations, validating the architectural rationality of the KPIR framework and the quality of the CPMK dataset. Ultimately, the integration of global and local perspectives ensures that the KPIR is not reliant on a single point of failure (MCT), but rather benefits from a robust, multi-level semantic understanding.

C.7 Cross-Style Generalization Analysis

While our dataset primarily consists of synthetic imagery generated via diffusion models, a potential concern is whether the model overfits to generative styles at the expense of real-world or traditional artistic styles. We conducted an Out-Of-Distribution (OOD) experiment using a manually curated test set of 106 pairs of ink paintings by Feng Zikai. We provide an example of the test set in Figure 8. Feng Zikai is a preeminent Chinese artist known for a minimalist and lyrical style that differs significantly from the textures produced by modern diffusion models. As shown in Table 10, although the model selection for KPIR-0.4B was

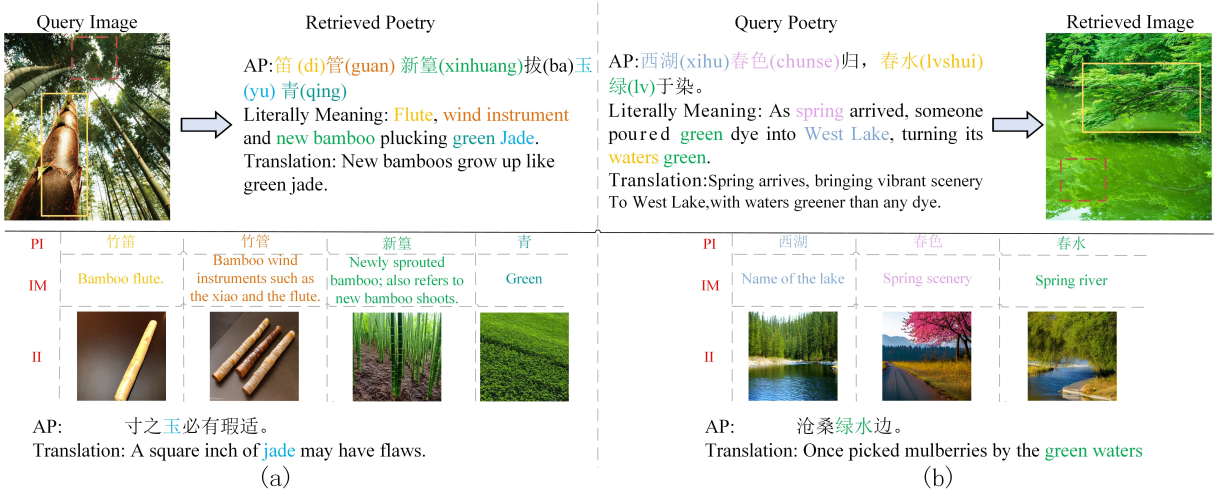
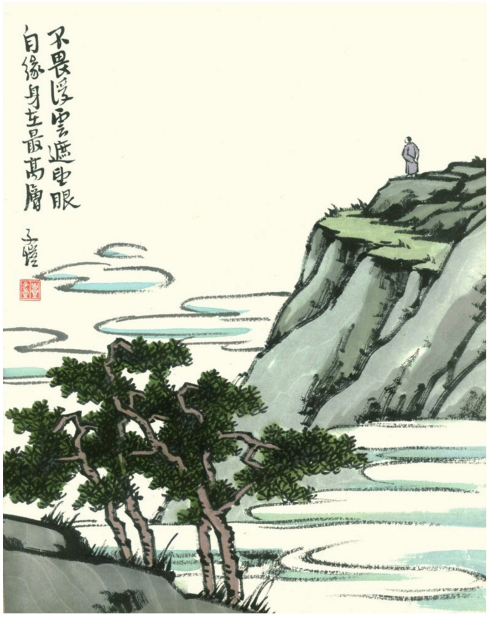


Figure 7: Two case studies of KPIR in the poetry-image retrieval task.



不畏浮云遮望眼，自缘身在最高层
Floating clouds cannot block my distant view, for I am standing on the highest peak.

Figure 8: An example of the OOD test set.

initially performed using a synthetic validation set, it still achieved performance gains in T2I tasks. To further isolate the impact of distributional bias, we evaluated the best-performing checkpoint directly on the OOD test set (denoted as *Optimal*). The significant performance improvement in this setting confirms that the model's gains stem from its ability to map poetry imagery to core semantic concepts rather than a superficial reliance on generative styles, demonstrating robust generalization across disparate visual domains.

Table 10: OOD evaluation results.

Model	T2IR@3	I2TR@3
Alt-CLIP-0.9B [5]	0.37	0.23
TaiSu-0.2B [21]	0.29	0.17
R2D2-0.4B [41]	0.24	0.27
CN-CLIP-0.4B [43]	0.57	0.38
KPIR-0.4B	0.62	0.37
KPIR-0.4B-Optimal	**0.65**	**0.48**

C.8 Poetry-Image Retrieval Case Study

Figure 7 presents two case studies for the poetry-image retrieval task. In these cases, a significant discrepancy exists between the literal meaning and the poetry translation, posing a challenge to the model's retrieval abilities. KPIR overcomes this challenge by leveraging both poetry imagery-related and MCT knowledge to achieve a more accurate understanding of the ancient poetry. In the figure, areas associated with poetry imagery-related knowledge are highlighted with yellow solid lines, while those related to MCT-knowledge are marked with brown dashed lines.

For example, as shown in Figure 7 (a), poetry imagery-related knowledge such as Di(笛) and Guan(管) is often translated as flute or other musical instruments in Modern Chinese. However, in the context of ancient poetry, they are more literally associated with bamboo. poetry imagery like XinHuang(新篁) are rarely used in contemporary language. These linguistic factors increase the difficulty of comprehending ancient poetry. By learning poetry imagery-related knowledge, KPIR over-

comes these challenges and correctly interprets the meaning. In Figure 7 (b), MCT knowledge provides a crucial semantic supplement for the poetry imagery-related knowledge. Although the water corresponding to XiHu(西湖) and ChunShui(春水) is blue, the MCT knowledge introduces the concept of "green water" through poetry translation, thereby deepening the model's understanding of the ancient poetry.

D Qualitative Evaluation of Imagery Meaning-Imagery Image Pairs

D.1 Questionnaire Design and Evaluation Process

In the qualitative evaluation, we designed a questionnaire to evaluate two aspects: the relevance between imagery meaning and imagery image, and whether imagery meaning is reasonably split on CPMK and PKG. The relevance and coverage scores are used, both ranging from 0 to 5.

Relevance Score: This metric evaluates the connection between the imagery meaning and imagery image, factoring in the imagery image's quality. Full points are given if the imagery image captures any essential meaning of the imagery meaning, with deductions for discrepancies. If there are two meanings and one is perfectly captured, full points 5 are awarded.

Coverage Score: Ranging from 0 to the relevance score of the current image, this metric measures the image's coverage of imagery meaning. Full score indicates complete coverage, while partial coverage results in proportional deductions. If one of two meanings is perfectly captured, a score of 2.5 is given.

The relevance score minus the coverage score can help determine whether the segmentation of the imagery meaning is reasonable. We invited five university students knowledgeable about classical Chinese poetry to evaluate 500 imagery meaning-imagery image pairs randomly selected from each dataset, totaling 1000 pairs.

As shown in Figure 9, CPMK significantly outperforms PKG in both relevance and coverage scores, with a smaller gap between the two compared to CPMK, indicating that CPMK achieves more reasonable image segmentation. Additionally, the CLIPScore between imagery meaning and imagery image further validates the higher similarity in CPMK. We provide a questionnaire example in Appendix D.2.

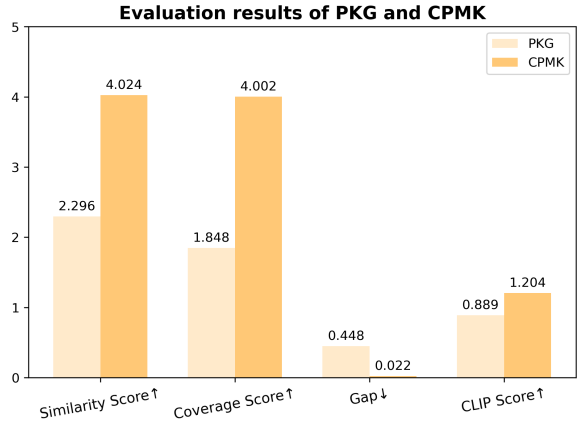


Figure 9: The evaluation results of imagery meaning-imagery image for PKG and CPMK. Gap represents the Similarity Score minus the Coverage Score.

D.2 Questionnaire Example

We provide a questionnaire example in Figure 10, illustrating the poetry imagery QuShu(氍毹) related knowledge in CPMK and PKG. In PKG, different meanings are distinguished by different colors, and grey indicates areas representing the origin of imagery meaning, which are treated as textual noise. Due to the influence of textual noise and improper segmentation of imagery meaning, the text-image correspondence in PKG is weak. The image barely reflects carpet, resulting in a relevance score of 2. Therefore, the coverage score for this text-image pair ranges from 0 to the relevance score (2), but it does not effectively convey the intended meaning of the stage, resulting in a compromised coverage score of 1. In CPMK, the correlation between imagery meaning and imagery image is high, with both relevance and coverage scores of 4.5 for imagery meaning-imagery image pairs.

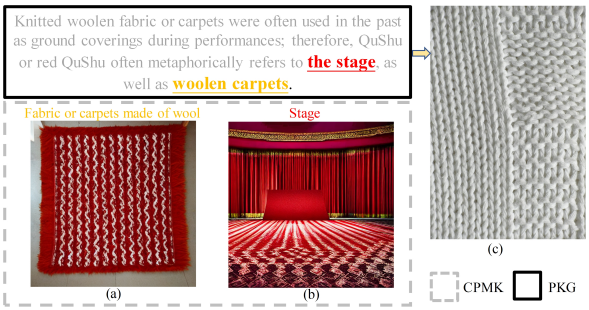


Figure 10: An example poetry imagery QuShu in the questionnaire.

D.3 Questionnaire Guidance

This section outlines the instructions provided to participants for evaluating the quality of knowledge graphs in our questionnaire. We provide guidelines for assessing two types of scores: similarity score and coverage score. As illustrated in Figure 14, we instruct participants to assign the highest similarity score (ranging from 0 to 5) if an image fully matches either the surface meaning (physical description) or the deep meaning (emotional significance) of any given meaning. Partial matches require point deductions based on the degree of alignment. Figure 14 explains the scoring for coverage, where the similarity score is constrained to fall within the range of 0 to the relevance score assigned to the image. If the image only covers a subset of interpretations, deductions are made proportionally depending on the number and significance of the uncovered interpretations. To ensure clarity, we provide three illustrative examples for each type of score, addressing common scenarios.

E Generative Model Selection

In this section, we discuss the selection of the image generation model and prompt settings. We chose the Taiyi-1B model (Zhang et al., 2022) for its balance of quality and efficiency. As shown in Table 11, Taiyi-1B performs comparably to larger models while offering faster inference due to its smaller parameter size. However, due to computational overhead, the entire image generation process required approximately one month of processing time on an NVIDIA RTX 4090 (24GB) GPU.

Table 11: Comparison of different models in Chinese(COCO-CN) datasets. Data is cited from (Wu et al., 2024)

Models	CLIPSim(↑)	FID(↓)	IS(↑)
Taiyi-1B[49]	0.197	69.226	21.060
Alt-1.5B[45]	0.220	68.488	22.126
Pai-1B[35]	0.196	72.572	19.145
Taiyi-3.5B[40]	0.225	67.675	22.965

The text prompts were primarily set based on the model's recommendations. To generate realistic images reflecting common real-world scenes, we included the word "realistic(现实)" in positive prompts. However, as the model often generated anime-style images, we added "anime(动漫)" to steer it toward the desired output. The final positive prompt words are "{}，现实 " where the imagery meaning is inserted into the {} position, while the negative prompt words are "广告, ，, ！, 。, ；, 资讯, 新闻, 水印, 动漫 ".

F Knowledge Augmentation

In this section, we detail the process of merging relevant data. First, ancient poetry is mapped to hash values to identify duplicates by comparing these hashes. Next, ancient poems are segmented by punctuation marks(，。!？), and similar ancient poems are merged based on the number of matching text segments. Duplicate or similar ancient poems are merged, and their associated knowledge is integrated. This process is carried out according to Algorithm 1.

Algorithm 1 Similarity Matching for the ancient poetry

Input: Raw databases $\{D_i\}$, num of databases l.
Output: Deduplicated databases D'.
$P \leftarrow \bigcup_{i=1}^{l}$ Extract ancient poetrys from D_i.
$D' \leftarrow \bigcup_{i=1}^{l}$ data from D_i.
for each pair $(P_s, P_l) \in P \times P$ **where** $P_s \neq P_l$ **and** $|P_l| \geq |P_s|$ **do**
  $N_s \leftarrow$ number of sentence segments in P_s
  $N_l \leftarrow$ number of sentence segments in P_l
  $C \leftarrow$ number of identical sentence blocks between P_s and P_l
  if $C > \frac{N_s}{2}$ **then**
    RK_s = GetRelevantKnowledge(P_s)
    $D'[l] \leftarrow D'[l] \cup \{RK_s\}$
    Delete($D'[s]$)
  end if
end for

G Prompt Optimization for imagery meaning

We introduce the instructions about prompt optimization for imagery image generation, which is shown in Figure 11. We use DeepSeek-Chat(DeepSeek-AI et al., 2025) for processing raw imagery meanings.

H Heuristic Approach

This section introduces the heuristic approaches used to process raw imagery. The primary objective is to separate distinct meanings associated with the same poetry imagery, eliminate textual

Input consists of keywords from classical Chinese poetry and their explanations. Each explanation may include multiple meanings as well as information about the poem's source.
Input format: Keyword&&Explanation
Task:
1. Data Cleaning:
- Identify and extract the main meaning from the explanation.
- Extract the poem's source, background explanation, and other related information as " Supplementary Knowledge". This information should retain its original format.
2. Determine if each main meaning has visual characteristics.
- Visual characteristics refer to elements that can be visualized, such as natural landscapes, animals, plants, specific scenes, etc.
3. If a main meaning has visual characteristics, set " Visual Information " to true and generate a visual description no longer than 20 words.
- The visual description should include details such as shape, color, action, background, etc., suitable for image generation.
4. If a main meaning does not have visual characteristics, set "visual information" to false and leave "visual description" empty.
Return format should be as follows:
Keyword:[
{"Meaning": "Extracted main meaning 1", "Supplementary Knowledge": "Related background or source 1", "Visual Information": true, "Visual Description": "Detailed visual description"},
{"Meaning": "Extracted main meaning 2", "Supplementary Knowledge": "Related background or source 2", "Visual Information": false, "Visual Description": ""},]
Example:
Input: 剥床&&1.语出《易剥》:"剥床以足，以灭下也。"陈梦雷浅述:"侵灭正道，自下而上也。"又:"剥床以肤，切近灾也。"陈梦雷浅述:"阴祸已迫其身也。"后用"剥床"称残害忠良或迫身之祸。
Output: [
{"Meaning": "残害忠良或迫身之祸", "Supplementary Knowledge": "语出《易剥》:"剥床以足，以灭下也。"陈梦雷浅述:"侵灭正道，自下而上也。"又:"剥床以肤，切近灾也。"陈梦雷浅述:"阴祸已迫其身也。", "Visual Information": false, "Visual Description": ""}]

Figure 11: Instructions for processing raw imagery meanings.

noise in poetry imagery, and retain valuable information from it as supplementary knowledge. Because raw imagery can have multiple meanings across various formats, it is challenging to develop heuristic rules that can be universally applied to split meanings. As a result, we split imagery meanings based on the structure of the crawled data. Regarding textual noise, the data we gathered contains a large amount of noise. We use regular expressions to remove as much as possible. Regarding supplementary knowledge, we find that a significant amount is marked by (《》) (book title). However, due to the large volume of similar information, it is difficult to fully represent them with regular expressions alone. Here are three regular expression examples:

(1) 语本《.*?》?：“.*?”
(2) 王逸注:.*?。
(3)《.*?》：.*?”

I Ontology Graph

We present the ontology graph in Figure 12, which classifies concepts into key concepts and attributes based on their significance.

J Pipeline of the method for constructing MM-KG of classical Chinese poetry

Figure 13 depicts the proposed methodology for constructing the classical Chinese poetry MM-KG. The pipeline distinguishes between different modalities through color coding: red, green, and blue represent textual, visual, and audio data, respectively. Grey-shaded blocks delineate the systematic stages of knowledge processing and graph synthesis.

K Examples of CPMK

In this section, we provide examples of the CPMK.

K.1 Example of poetry imagery and Character Related Knowledge in CPMK

In Figure 15, we present an example of poetry imagery and character-related Knowledge in CPMK. For poetry imagery (谢墅), we provide its imagery meaning, Supplementary Information, and imagery image. For the character (墅), we provide its Pinyin, ZhuYin, Stroke Count, and Explanation.

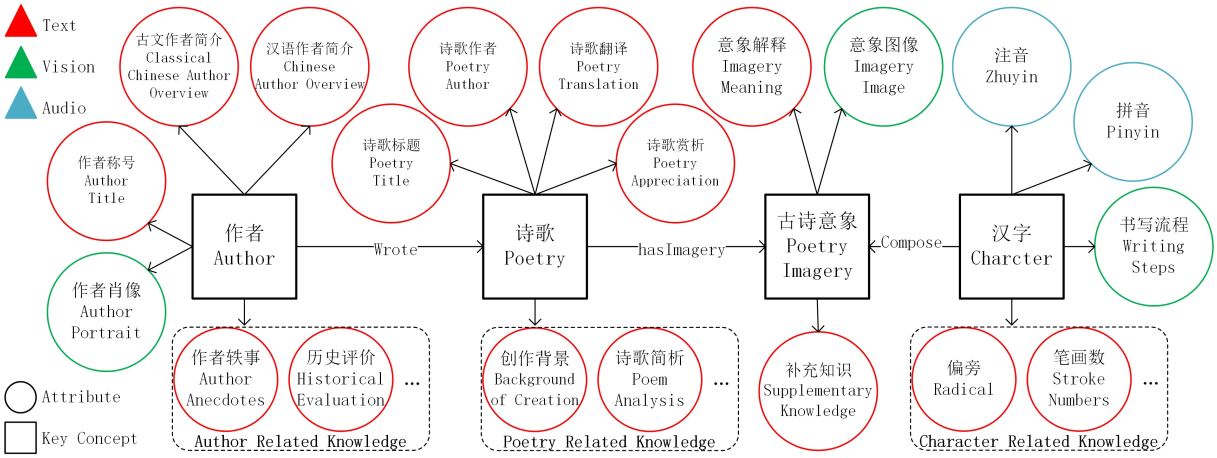


Figure 12: The ontology graph of classical Chinese poetry, where shapes represent types and colors represent modalities.

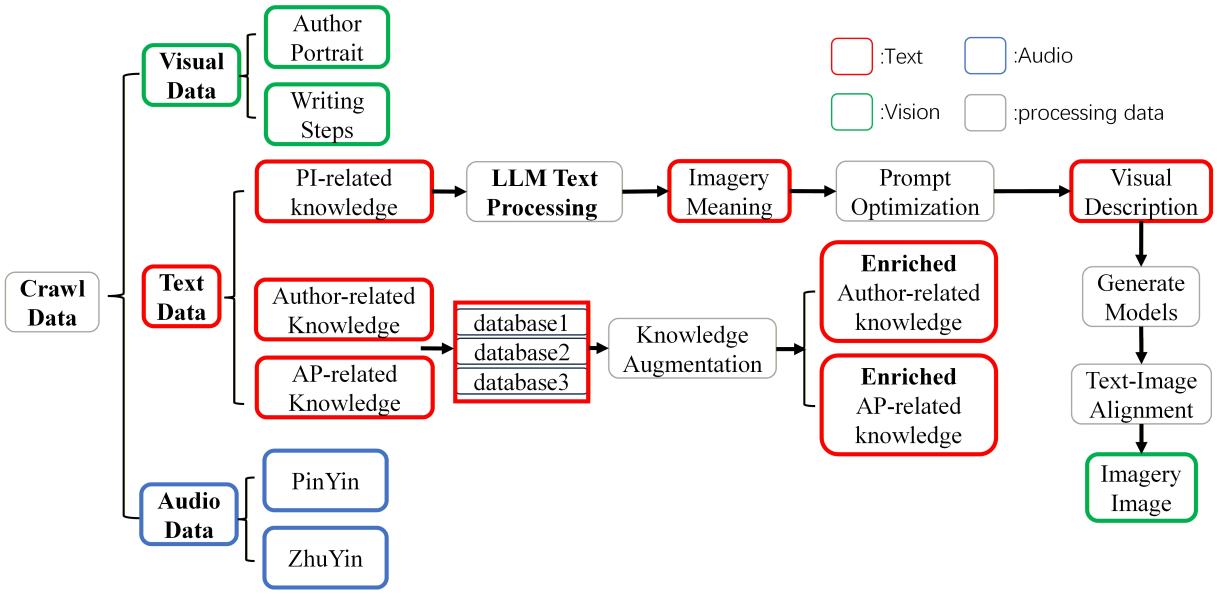


Figure 13: Poetry MM-KG Construction Pipeline

K.2 Example of Poetry-Related Knowledge in CPMK

In Figure 16, we present an example of poetry-related knowledge in CPMK. We integrate knowledge from SouYun, GuShiWen, and GuoXueHui, striving to present as complete a representation of poetry-related knowledge as possible through the consolidation of various databases. Notably, the version of the classical Chinese poem obtained from SouYun (两岸猿声啼不尽) differs from those found in GuShiWen and GuoXueHui (两岸猿声啼不住). By preserving these textual variations and their original sources, CPMK effectively captures the philological richness and historical evolution of classical works, which is often overlooked in single-source datasets. Through the use of knowledge augmentation, we provide a comprehensive representation of the knowledge related to these poems.

K.3 Example of Author-Related Knowledge in CPMK

In Figure 17, we present an example of author-related knowledge in CPMK. This includes the poet's name(李白), the historical era, a brief introduction, and an image of the poet. Besides knowledge from SouYun, we also integrate various data sources from SouYun to provide a comprehensive introduction to the authors.

词语:鸡 解释:家禽
Word: Chicken **Definition**: Poultry

Figure 1

词语:爱 解释:1.对人或事有深挚的感情 2.容易
Word: Love **Definition**:
1.A deep affection for someone or something.
2.Easy to be prone to or inclined toward something

Figure 2

词语:鸳鸯 解释:是传说牛郎、织女分居天河两岸，每年七夕，喜鹊飞临天河，……事见《岁华纪丽．七夕》注引汉应劭《风俗通》。后因以"鸳鸯"为七夕的典实。
Word: Ride the Magpie **Definition**: Refers to the legend of the Cowherd and Weaver Girl….reuniting on Qixi Festival via a magpie bridge.

Figure 3

相似性得分:
如果图片符合词语的某一个解释的**表层含义-物理描述**或**深层含义-情感**，则相似性得分打满分。分值为0~5
如图1符合表层含义打5分，如图2玫瑰符合解释1的深层含义，虽不满足解释2"容易"，但较为满足解释1，打4分.
图3既不符合七夕的表层含义与深层含义打0分，若勉强符合可以凭感觉折中打分
Similarity Score:
If an image matches either the **surface meaning (physical description)** or **deeper meaning (emotional aspect)** of a word, it gets a full similarity score (0–5).
Examples:
•Image 1 matches the surface meaning —score: 5.
•Image 2 (e.g., a rose) fits the deeper meaning of Definition 1 but not Definition 2 ("easiness") —score: 4.
•Image 3 doesn't match Qixi's surface or deeper meanings —score: 0.
•If barely relevant, score based on intuition.

覆盖性得分:
覆盖性得分的打分范围为(0，当前图像的相似性得分)
如果图片能够覆盖该词语的所有解释上则赋予当前图像的相似性得分，否则根据覆盖度酌情减分
如图1符合所有解释打5分 ,图2因相关性得分为4，**得分范围为(0,4),因为只满足一个减半最终打2分**，
图3因相似性得分为0分则覆盖性也为0分。
Coverage Score:
The coverage score **ranges from 0 to the current similarity score of the image**. If the image covers all interpretations of the word, it gets the full similarity score. Otherwise, points are deducted based on the degree of coverage.
Examples:
•Image 1 fully covers all interpretations —score: 5.
•Image 2, with a similarity score of 4, has a coverage range of (0, 4). Since it only satisfies one interpretation, the score is halved to 2.
•Image 3, with a similarity score of 0, also gets a coverage score of 0.

Figure 14: Instructions for evaluating the Similarity Score and Coverage Score.

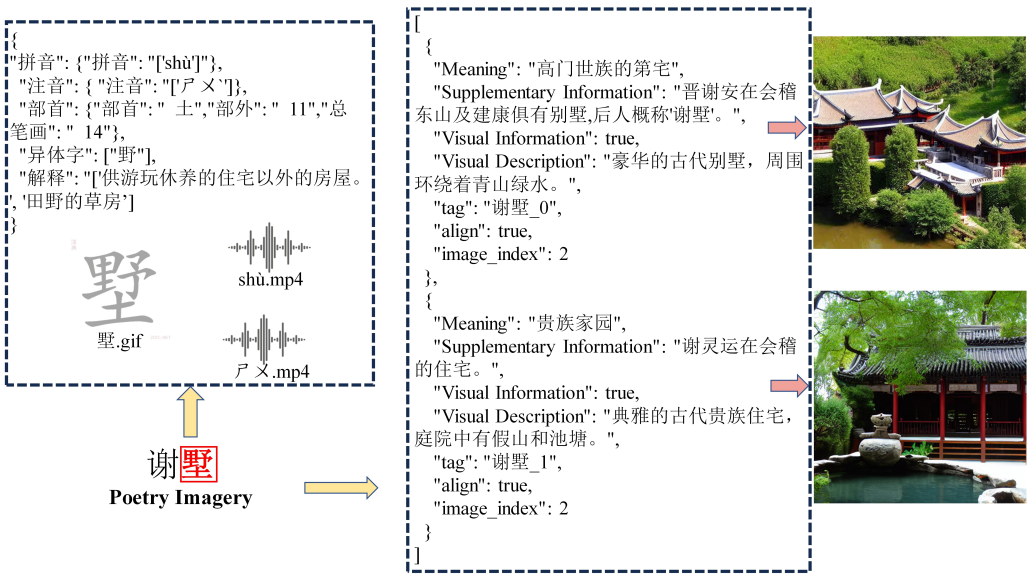


Figure 15: An example of character-related knowledge and poetry imagery-related knowledge in CPMK.

[{ "标题": "早发白帝城（一作白帝下江陵）",
"作者": "李白",
"朝代": "唐朝",
"内容": "朝辞白帝彩云间，千里江陵一日还。两岸猿声啼不尽，轻舟已过万重山。“,
“来源”: “搜韵"},
{"标题": "早发白帝城",
"内容": "朝辞白帝彩云间，千里江陵一日还。两岸猿声啼不住，轻舟已过万重山。",
"朝代": "唐朝",
"作者": "李白",
"古诗简介": "《早发白帝城》是唐代伟大诗人李白在流放途中遇赦返回时所创作的一首七言绝句，是李白诗作中流传最广的名篇之一。诗人是把遇赦后愉快的心情和江山的壮丽多姿、顺水行舟的流畅轻快融为一体来表达的。全诗无不夸张和奇想，写得流丽飘逸，惊世骇俗，但又不假雕琢，随心所欲，自然天成。",
"翻译/译文": "清晨我告别高入云霄的白帝城；江陵远在千里船行只一日时间。两岸猿声还在耳边不停地啼叫；不知不觉轻舟已穿过万重青山。",
"注释": "发：启程。白帝城：故址在今重庆市奉节县白帝山上。朝：早晨。辞：告别。彩云间：因白帝城在白帝山上，地势高耸，从山下江中仰望，仿佛耸入云间。白帝：今四川省奉节。江陵：今湖北荆州市。一日还：一天就可以到达；还：归；返回。猿：猿猴。啼：鸣、叫。住：停息。万重山：层层叠叠的山，形容有许多。",
"赏析": " 《早发白帝城》是唐代伟大诗人李白在流放途中遇赦返回时所创作的一首七言绝句，是李白诗作中流传最广的名篇之一。诗人是把遇赦后愉快的心情和江山的壮丽多姿、顺水行舟的流畅轻快融为一体来表达的。全诗无不夸张和奇想，写得流丽飘逸，惊世骇俗，但又不假雕琢，随心所欲，自然天成。……",
"古诗分类": " ['唐诗三百首', '小学生必背古诗70首', '七言绝句', '长江','叙事诗', '山水诗']",
“来源”: ”: “国学荟"},
{"标题": "早发白帝城",
"朝代": "唐朝",
"作者": "李白",
"内容": "朝辞白帝彩云间，千里江陵一日还。两岸猿声啼不住，轻舟已过万重山。",
"相关信息": {
"译文及注释": [
{ "译文": "清晨，我告别高入云霄的白帝城，江陵远在千里之外，船行只需要一天时间便能返回。两岸猿声还在耳边不停地回荡，轻快的小舟已驶过万重青山。" },
{"译文二": "清晨，我告别高入云霄的白帝城，江陵远在千里，船行只需一日。两岸猿声，还在耳边不停地啼叫，不知不觉，轻舟已穿过万重青山。" },
{ "注释": "发：启程。白帝城：故址在今重庆市奉节县白帝山上。杨齐贤注：“白帝城，公孙述所筑。初，公孙述至鱼复，有白龙出井中，自以承汉土运，故称白帝，改鱼复为白帝城。”王琦注：“白帝城，在夔州奉节县，与巫山相近。所谓彩云，正指巫山之云也。”朝：早晨。辞：告别。彩云间：因白帝城在白帝山上，地势高耸，从山下江中仰望……"}
],
"创作背景": "公元759年（唐肃宗乾元二年）春天，李白因永王李璘案被流放夜郎，途经重庆。行至白帝城的时候，忽然收到赦免的消息，惊喜交加，随即乘舟东下江陵。此诗即是作者回到江陵时所作，所以诗题一作《下江陵》。",
"赏析": [
{ "赏析": "唐代安史之乱初期，唐玄宗奔蜀，太子李亨留讨安禄山，不久，李亨既位，史唐肃宗。玄宗又曾命令儿子永王李璘督兵平叛，永王李璘在江陵，召兵万人，自树一帜，肃宗怀疑他争夺帝位，已重兵相压，李璘兵败被杀。李白曾经参加过永王李璘的幕府，被加上“附逆”罪流放夜郎（今贵州遵义），……"}],
“简析”:“《早发白帝城》是一首七言绝句。此诗意在描摹自白帝至江陵一段长江水急流速、舟行若飞的情况。首句写白帝城之高；次句则描述了江陵路遥，舟行迅速；三句以山影猿声为背景，衬托行舟飞进；四句写行舟轻如无物，点明水势如泻。诗人遇赦后愉快的心情和顺水行舟的流畅轻快、……。”},
“来源”: ”: “古诗文"},
]

Figure 16: An example of poetry-related knowledge in CPMK

[{ “诗人": "李白",
“朝代": "唐代",
“称号”: [“饮中八仙”, “大李杜",],
“介绍": "李白（701年－762年），字太白，号青莲居士，又号“谪仙人”，祖籍陇西成纪（今甘肃省秦安县），出生于蜀郡绵州昌隆县（一说出生于西域碎叶）。唐代伟大的浪漫主义诗人，被后人誉为“诗仙”，与杜甫并称为“李杜”，为了与另两位诗人李商隐与杜牧即“小李杜”区别，杜甫与李白又合称“大李杜”。据《新唐书》记载，李白为兴圣皇帝（凉武昭王李暠）九世孙，与李唐诸王同宗。其人爽朗大方，爱饮酒作诗，喜交友。李白深受黄老列庄思想影响，有《李太白集》传世，诗作中多为醉时写就，代表作有《望庐山瀑布》《行路难》《蜀道难》《将进酒》《早发白帝城》等。",
“其他知识": {
"轶事典故": [{"友挚情": "…… }, {"生死考证": "....."}],
"家庭成员": [{ "家人": “……" },{ "配偶": “……"}, {"子女": “……"}],
“后世纪念”: [{"墓地": “……" }, {"纪念馆": “……}],
"主要成就": [{"主要成就": “……"}, {"歌": “……"}, { “代表作品”: "……"}, {"剑术": “……"}, {"道经": “……" }, {"思想": “……"}],
"人物生平": [{"早年天才": “……"}, {"辞亲远游": “……" }, {"蹉跎岁月": “……"}, {"西游献赋": “……"}, {"李杜相识": “……"}, {"安史入幕": “……"}, {"溘然病逝": “……"}],
"来源": "古诗文"
},
{“其他介绍":[
{"中国历代人名大辞典": "【生卒】：701—762\n【介绍】：\n唐陇西成纪人，其先人隋末流寓西域，故生于安西都护府所属碎叶城。中宗神龙初，迁居蜀之绵州昌隆县青莲乡，又尝寓居山东，故亦称山……" },
{ "唐诗大辞典修订本": "【生卒】：701—762\n字太白，号青莲居士，排行十二，陇西成纪(今甘肃秦安西北)人，其先隋末窜于碎叶(今吉尔吉斯斯坦托克马克附近)，李白即出生于此。中宗神龙元年(705)随家……"},
{"唐诗汇评": "李白（701-762），字太白，号青莲居士。祖籍陇西成纪（今甘肃秦安）。出生地有蜀中、西域、长安诸说，迄无定论。少时居绵州彰明县清廉乡（今属四川江油），读书吟诗，遍观百家……"},
{ "词学图录": "李白（701-762）字太白，号青莲居士。祖籍陇西成纪（今甘肃秦安东），隋末其先人流寓西域，白出生于安西大都护府碎叶城，五岁随父迁居绵州昌隆（今江油）青莲乡。天宝初供奉翰林。……"},
{ "黄鹤楼志·人物篇": "李白（701—762）唐代诗人。字太白，号青莲居士，世人又称谪仙、诗仙。祖籍陇西成纪（今甘肃静宁西南），先世流迁中亚，5岁随父定居绵州昌隆县（今四川江油县）青莲乡。……"},
{"全唐文·卷三百四十七": "白字太白。兴圣皇帝九世孙。白生梦长庚星。因以命之。举有道不应。天宝初至长安。贺知章言于元宗。召见金銮殿。论当世事。奏颂一篇。诏供奉翰林。忤高力士。摘其诗激杨……"}
]}

Figure 17: An example of author-related knowledge in CPMK