LivePoem: Improving the Learning Experience of Classical Chinese Poetry with AI-Generated Musical Storyboards

Qihao Liang, **Xichu Ma**, **Torin Hopkins** and **Ye Wang**
School of Computing, National University of Singapore
qihao.liang@u.nus.edu, ma_xichu@nus.edu.sg, torinhopkins@gmail.com, wangye@comp.nus.edu.sg

Abstract

Textbook reading has long dominated classical poetry education in Chinese-speaking communities. However, research has shown that extensive text-based learning can lead to learner disengagement and a less pleasant experience. This paper aims to improve the experience of classical Chinese poetry learning by introducing LivePoem–a system that generates *musical storyboards* (storyboards with background music) as audiovisual aids to support poetry comprehension. We employ a pre-trained diffusion model for storyboard generation and train a prosody-based poem-to-melody generator using a Transformer model, both validated by standard objective metrics to ensure generation quality. Through a within-subjects study involving 25 non-native Chinese learners, we compared learning outcomes from textbook reading and musical storyboard viewing through standardised reading comprehension tests. Additionally, the learning experience was assessed by the Self-Assessment Manikin (SAM) and an inductive thematic analysis of learners' open-ended feedback. Experimental results show that musical storyboards retained the learning outcomes of textbooks, while more effectively engaging learners and providing a more pleasant learning experience.

1 Introduction

Textbook reading has long been the dominant method for classical poetry education in Chinese-speaking communities [Zhao and Li, 2024a]. It typically employs plain-language interpretations to help learners understand the archaic vocabulary and poetic techniques that are less common in modern Mandarin [Zhao and Li, 2024b]. While textbooks effectively convey the meaning of classical poetry, research suggests that extensive text reading can lead to boredom [KrukMariusz, 2021], disengagement [Guthrie and Davis, 2003], and negative learning outcomes over time [Feng *et al.*, 2013].

To improve the experience of poetry learning, researchers in the learning sciences have explored alternative or auxiliary multimodal methods to textbook reading, such as listening to music [Tilwani *et al.*, 2022], viewing pictures and videos [Perez and Rodgers, 2019; Tahmina, 2023]. These methods highlight the role of audiovisual media in fostering various language skills [Khasawneh, 2023], including speech fluency [Hwang *et al.*, 2024], vocabulary acquisition [Muñoz *et al.*, 2023; Pratama and Hadi, 2023], and imagery comprehension [Pujadas and Muñoz, 2023]. Moreover, music has emerged as another promising modality for language learning, as it relates to the neural processing of tones and vowels in classical Chinese poetry [Zhang *et al.*, 2023]. This underscores the positive impact of music on language education.

Despite these potential benefits, creating such audiovisual media can be costly, time-consuming, and require specialised expertise for human workers. To address this challenge, we develop LivePoem, a generative AI system that converts poetry into *musical storyboards*–storyboards with background music–as audiovisual learning materials. A musical storyboard includes a chain of images that visualise the poem's content, accompanied by the singing of poem lines (as in Figure 1). With AI-generated musical storyboards, we aim to enhance the learning experience of classical poetry by offering a more engaging form of educational media.

The system includes two phases: (A) storyboard generation, and (B) poem-to-melody generation. In (A), we employ a pre-trained language model (LM) and a latent diffusion model (LDM) [Huang *et al.*, 2023] to generate storyboards for poetry. The LM expands a poem into a script that describes the poem's content in plain language, transforming the connotative and poetic language into more understandable descriptions. This script then prompts the LDM to generate visualisations of the poem's content. In (B), we train a prosody-based melody generator with a Transformer model [Lewis *et al.*, 2020]. The poem is scanned and converted into a prosody template, conditioning the Transformer model to generate a melody that rhythmically aligns with the poem. This ensures that the output melody is singable, matching the poem's syllabic and rhythmic pattern [Liang *et al.*, 2024]. Finally, the generated music is automatically aligned with storyboards by grouping musical phrases and images according to poem lines, enabling synchronised playback.

We validated the system using standard computational met-

Supplementary materials are available at: https://github.com/lqhac/LivePoem

Figure 1: Example of a musical storyboard for a Chinese poem *Jing Ye Si* by *Li Bai*. A musical storyboard consists of a sequence of visual frames with a background melody. The storyboard is interpolated into smoother animation, with the melody synthesised as singing voices.

rics from image and music generation studies, demonstrating its ability to produce high-quality content. However, recognising that automatic metrics alone cannot fully capture the human learning outcomes and experience, we further conducted a within-subjects study with 25 non-native Chinese language learners. This study (1) examined learners' preliminary learning outcomes, engagement, and satisfaction with musical storyboards versus textbooks, and (2) collected their opinions on both learning approaches. The results show that musical storyboards retained learners' test performance while making the learning experience substantially more engaging and pleasant. We also analysed learners' feedback that exposed the benefits and limitations of the traditional textbook-based and multimodal learning approaches.

In summary, our work contributes the following:

(1) The LivePoem framework for musical storyboard generation to support classical Chinese poetry learning;

(2) A two-part human-grounded study evaluating the effects of AI-generated musical storyboards in poetry learning.

2 Related Work

Textbook-based and Multimodal Learning

From lectures and standard tests to self-learning, textbooks have beeing playing a central role in various learning scenarios. Despite their ubiquity and effectiveness, recent research has identified some limitations of textbooks [Zhao and Li, 2024b], such as their overemphasis on vocabulary form and accuracy [Brown, 2011], and the lack of a meaningful context [King, 2002]. Furthermore, textbook-centric classes may lead to teacher over-involvement and student under-involvement [O'Neill, 1982], resulting in boredom [KrukMariusz, 2021], disengagement [Guthrie and Davis, 2003], and negative learning outcomes over time [Feng *et al.*, 2013]. To address these issues, multimodal learning materials–especially audio-visual media [Tahmina, 2023]–have been shown to complement or enhance textbook-based learning. For example, Lee and Révész find that textually enhanced video captions can facilitate the acquisition of grammatical knowledge [Lee and Révész, 2018]. Zhang et al. observe the benefits of musical training for language learning, as music audio input facilitates tone and vowel processing abilities of language learners [Zhang *et al.*, 2023]. Hnatyshyn et al. use melody alterations to symbolise cancerations in DNA, showing that knowledge *musification* improves the pleasantness of learning, compared to reading text materials [Hnatyshyn *et al.*, 2024].

Cutting-Edge Technical Underpinnings for Musical Storyboard Generation

A musical storyboard includes a chain of images (video) and the singing of poem lines. We thus divide musical storyboard generation into two core technologies: text-to-video generation, and melody generation from poetry.

Text-based Video Generation. Text-based video generation creates image frames from texts that describe the expected content in the resulting video, which is often modelled as generating temporally coherent static images iteratively. In this field, diffusion-based models [Song *et al.*, 2021; Kim *et al.*, 2024] have become popular due to their high-quality synthesis and semantic controllability. This success extends to video generation, where a promising direction is tuning pre-trained text-to-image generators in zero- or few-shot settings [Xu *et al.*, 2023; Clark and Jaini, 2024]. One of the most recent works, Free-Bloom, uses pre-trained large language models to generate frame-level text descriptions and latent diffusion models for frame generation, achieving strong performance without additional training [Huang *et al.*, 2023].

Generating Melodies for Poetry. Poetry and melody have long been closely related, both sharing a rhythmic and musical nature. Research has shown that early Chinese poems were composed for singing [Hu, 2023], and that singing in return contributed to the emergence of diverse poetic genres [Minli, 2024]. Generating melodies for poetry can be viewed as a subtask of lyrics-to-melody generation. Early deep learning methods mainly utilise end-to-end models (e.g., LSTM-GAN [Yu *et al.*, 2021]) on paired melody-lyrics datasets, but the size and quality of such datasets do not suffice to make large models converge well. Some researchers address this problem by pre-training transformer models on unpaired datasets [Sheng *et al.*, 2021], which has spurred a wave of using attention-based methods for melody generation (e.g., [Liang and Wang, 2024]). Recently, research has explored the interpretability of AI melody generators, examining melody-lyrics relationships in attention-based models [Duan *et al.*, 2023]. Liang et al. highlight the importance of prosody in singability, which measures the rhythmic compatibility between melodies and lyrics during singing [Liang *et al.*, 2024].

3 LivePoem System Architecture

3.1 System Overview

LivePoem is a generative AI framework that automatically creates *musical storyboards* from classical Chinese poetry. A

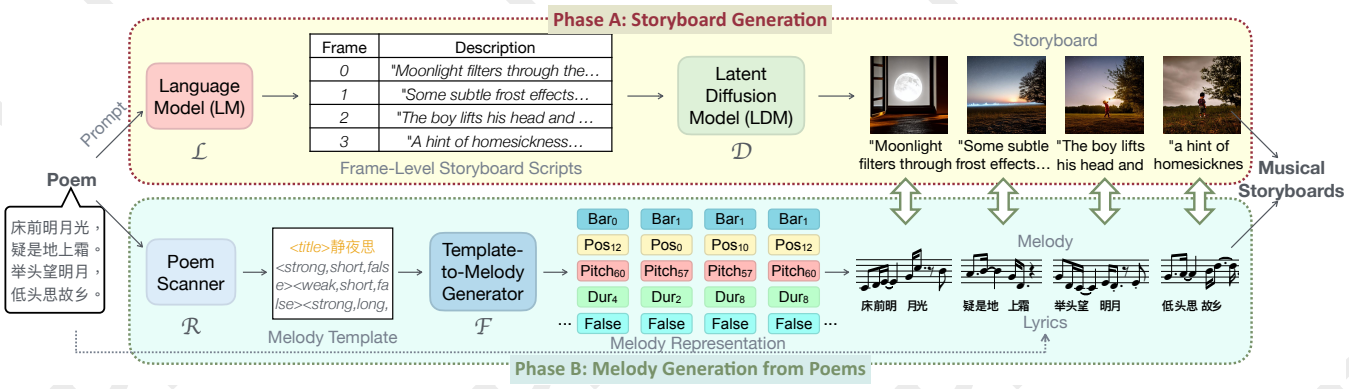

Figure 2: The two-phase musical storyboard generation framework of LivePoem. Phase A uses a language model to expand a poem into a script, which describes the connotative poetic content in plain language. These scripts then prompt a latent diffusion model to generate visual frames. Phase B extracts the poem's prosody to create a melody template, enabling singable music generation. Finally, the generated storyboards and melodies are aligned by sentence boundaries, combining visuals and music for enhanced learning.

musical storyboard for a poem contains (1) *a sequence of images visually depicting the poem's content*, paired with (2) *the singing of the poem lines*. As in Figure 2, this framework includes two phases: (A) the storyboard generation, and (B) the poem-to-melody generation.

3.2 Phase A: Storyboard Generation

In the storyboard generation phase (top panel of Figure 2), a language model first expands an input poem into a script. This expansion converts the connotative and metaphorical poetic language into more straightforward, plain-language descriptions, helping generative models capture the meaning of classical poetry. Next, a latent diffusion model [Huang *et al.*, 2023] is prompted by this script to generate a sequence of temporally coherent images as the storyboard.

Frame-Level Script Generation

Let $X = \{x_1, x_2, \ldots, x_n\}$ denote a classical Chinese poem with n lines. A language model $\mathcal{L}$ is employed to automatically generate a corresponding script $Y = \{y_1, y_2, \ldots, y_n\}$, where each y_i is a plain-language description of the poetic scene depicted in x_i, for all $i \in \{1, \ldots, n\}$. To obtain Y, a system prompt X_0 is provided to prime the input poetry X, specifying the generation task and guiding the model to focus on scene interpretation and descriptive fidelity.

$$Y = \mathcal{L}([X_0, X]) \tag{1}$$

where $[X_0, X]$ denotes the concatenation of X_0 and X.

Script-to-Storyboard Generation

Prompted by the descriptive script $Y = \{y_1, y_2, \ldots, y_n\}$, a latent diffusion model $\mathcal{D}$ generates a sequential storyboard $S = \{s_1, s_2, \ldots, s_n\}$, where s_i denotes the set of frames depicting x_i. To ensure semantic and temporal coherence across storyboard frames, we incorporate attention shift and sampling strategies proposed in [Huang *et al.*, 2023]. Formally, this generation process can be expressed as:

$$\{s_i\}_{i=1}^{t} = \{\mathcal{D}(y_i)\}_{i=1}^{t} \tag{2}$$

3.3 Phase B: Poem-to-Melody Generation

To synthesise a singable background melody for the input poem, we train a poem-to-melody generator that produces a melody sequence $M = \{m_1, m_2, \ldots, m_n\}$ of n musical phrases, where m_i corresponds to the i-th line of the poem. The generation of M is guided by the prosody of X, as the prosodic melody-poem alignment ensures that the resulting melody is both musically coherent and rhythmically compatible with the poetic structure [Liang *et al.*, 2024]. Technically, this phase (bottom of Figure 2) includes: 1) a poetry scanner $\mathcal{R}$, which scans[1] a poem X to extract its prosody structure and create a melody template; 2) a prosody-to-melody generator $\mathcal{F}$, which employs an end-to-end Transformer model [Lewis *et al.*, 2020] to generate a melody that matches the format and prosodic pattern in the template.

Poem Scansion

In classical Chinese poetry, the prosody of a poem is related to the poem's *form*, which stipulates the metrical and tonal format during composition. The prosody of poems is set by counting and grouping the syllables (characters) along with caesuras (in-line pauses) and a long pause at the end of the line [Birrell, 2022]. For instance, a five-character rhythmic poem consists of five characters (`ch`) per line. Each line is commonly segmented as either `[2 ch|2 ch |1 ch]` or `[2 ch|1 ch |2 ch]`, where the first syllable of each segment is typically emphasised, and the final syllable tends to be lengthened [Kuo, 1971]. Based on this concept, we design a poem scanner that derives a melody template by analysing poetic structure, segmenting syllables, and annotating each syllable with `strong`/`weak`, `short`/`long`, and a binary `LC` marker for the last character of each line.

For example, in 静夜思 (Jìng Yè Sī), the second line is segmented as: 疑是|地上|霜 (Yí shì|dì shàng|shuāng), with the underlined characters stressed. Additionally, 霜 (shuāng) is also the last syllable of this line and can thus be represented

[1]Scansion, or scanning, is the analysis of the metrical patterns of a poem by organising its lines into feet of stressed and unstressed syllables and showing the major pauses, if any.

as <strong,long,true>. When input to the model, the embeddings of all three symbols are concatenated and linearly projected to the embedding dimension of $\mathcal{F}$.

Template-to-Melody Generation

The melody generator $\mathcal{F}$ aims to generate a sequence of melody notes given a melody template. Each melody note n corresponds to a single time step in the sequence and is symbolically represented as a concatenation of five musical attributes (Figure 2, mid-bottom area): bar number $\texttt{Bar}_\texttt{b}$, position in bar $\texttt{Pos}_\texttt{x}$, pitch value $\texttt{Pitch}_\texttt{p}$, note duration $\texttt{Dur}_\texttt{d}$, and a phrase boundary indicator $\texttt{Phrase}_\texttt{h}$ denoting whether n is the last note in a melody phrase. All subscripts (b, x, p, d, h) represent the actual value of each attribute. This representation minimises the length of melody sequences, reducing training costs while preserving melody structure. The model is trained, validated, and tested on LMD[2] and POP909 datasets [Wang* *et al.*, 2020] (8:1:1 split), with the following objective function:

$$L = CE_{Bar}+CE_{Pos}+CE_{Pitch}+CE_{Dur}+CE_{Phrase} \quad (3)$$

where L denotes the total loss; CE_α denotes the cross entropy loss for an attribute α.

To ensure the alignment between melodies and poems, a sampling strategy is performed on bar and position symbols during inference: A penalty of $r = -10^8$ is added to the logits of the bar and position symbols that (1) have been previously sampled or (2) contradict the prosody pattern of poetry. This strategy ensures the monophonicity[3] of melody and the prosodic match between melody and poem.

With the generated storyboard S and the melody M, the storyboard frames and melody phrases that depict the same poem line are grouped together as a musical storyboard $MS = \{\{s_1, m_1\}, \{s_2, m_2\}, ..., \{s_n, m_n\}\}$. This synchronises melody and storyboard, enabling simultaneous playback. For better presentation, we follow [Huang *et al.*, 2023] to interpolate the generated storyboard frames, thereby creating smoother animations, and use X Studio[4] to synthesise the generated melody into a singing voice.

3.4 System Validation

We validated the LivePoem system against state-of-the-art image and melody generators using a poetry corpus published by the Ministry of Education in China (with 137 poems)[5]. For storyboard generation, we used FIFO [Kim *et al.*, 2024] as the baseline. The CLIP scores [Radford *et al.*, 2021], which measure the similarity between an image and a piece of descriptive text, were 0.30 for LivePoem and 0.28 for FIFO. For poem-to-melody generation, we selected SongMASS [Sheng *et al.*, 2021] and TeleMelody [Ju *et al.*, 2022] as baselines. We first measured the singability of melodies using Prosody-BLEU (PB) [Liang *et al.*, 2024]. Then, we evaluated the music quality of these models against a test dataset with 379 songs randomly sampled from the LMD and Pop909 datasets. The metrics included Pitch Count (PC), Pitch Class Transition Matrix (PCTM), Pitch Range (PR), and Note Length Transition Matrix (NLTM) [Yang and Lerch, 2020]. The results (Table 1) demonstrate the competitive performance of LivePoem. On all metrics, a higher score indicates better performance.

Models	PB	PC	PCTM	PR	NLTM
LivePoem	0.88	78.20	81.68	91.20	74.34
SongMass	0.40	57.24	30.43	80.83	39.05
TeleMelody	0.61	31.35	11.66	58.77	21.15

Table 1: System validation results on singability and music quality

4 Experiment

The experiment investigated two research questions (RQs):

RQ1: Can musical storyboards retain textbooks' effectiveness in improving learners' comprehension of classical Chinese poetry?

RQ2: Do musical storyboards provide a more engaging and pleasant learning experience compared to textbooks?

4.1 Study Overview

To address both RQs, we conducted a two-part user study: The first part used standardised reading comprehension tests to measure participants' poetry understanding using musical storyboards and textbooks (RQ1). The second used the self-assessment manikin (SAM) and free-response questions to explore participants' learning experiences with musical storyboards and textbooks (RQ2).

4.2 Participants

We recruited 25 Chinese language learners from our institution's mailing list, following these criteria: (1) non-native Chinese speakers; (2) aged over 18; (3) interested in classical Chinese poetry; (4) are not hard of hearing or experiencing limited vision. Participants were aged 18–35 (14 men, 10 women, 1 undisclosed). They self-reported their Chinese language proficiency using the Inter-agency Language Roundtable (ILR) scale[6], a validated standard used by federal agencies in the U.S. for grading language proficiency: one at Level 0 (no proficiency), 14 at Level 1 (elementary), and 10 at Level 2 (limited working) ($M = 1.4, SD = 0.6$).

4.3 Part I: Evaluating the Effectiveness of Musical Storyboards in Learning Classical Poetry

The first part employed a within-subjects design and included the following two test conditions::

Textbook (TB): Participants read textbooks to answer reading comprehension questions. The textbooks contained the author's biography, annotations of important words, the translation and background of the poem, etc.

Storyboard (SB): Participants viewed musical storyboards to answer reading comprehension questions. The storyboards included visualisations and singing of the poems.

[2]https://colinraffel.com/projects/lmd/

[3]A melody is *monophonic* if it consists of a single, unaccompanied melodic line, with no overlapping or simultaneous pitches.

[4]https://xstudio.music.163.com

[5]http://www.moe.gov.cn/srcsite/A26/s8001/202204/W020220420582344386456.pdf; pages 58–63

[6]https://www.govtilr.org/Skills/ILRscale2.htm

To select poems for the reading comprehension tests, we referred to a corpus published by the Ministry of Education (MoE) of China[7], which lists all compulsory poems in the national curriculum. The corpus was categorised into four difficulty levels: 1st–2nd, 3rd–4th, 5th–6th, and 7th–9th grades[8]. Regarding the number of questions, our pilot test showed that four poems, each with five multiple-choice questions, were optimal given time and cognitive load constraints. Each question included a correct answer, three incorrect answers, and an "I don't know" choice. This question set took 30–60 minutes to complete, aligning with the standard format of MoE poetry comprehension tests. To control order effects, both test conditions and poem samples were computationally randomised [Davies *et al.*, 2014]. Specifically, we randomly selected four poems, one from each of the four difficulty levels. Two poems were randomly paired with musical storyboards, the other two with textbooks. The order of the four poems was also randomised when presented to participants. Participants were reimbursed at USD 7.59 per half hour.

The reading comprehension question sheet was created with Qualtrics[9] and followed these steps:

1. Participant Consent and Background: Participants read an introduction, signed a consent form approved by the ethics committee, and provided demographic details. They self-reported their Chinese proficiency using the Inter-agency Language Round-table (ILR) scale[10].

2. Task Familiarisation: Participants completed a familiarisation session that simulated the formal study. They first read a poem (excluded from the formal study) and answered two test questions. They then read the same poem again with textbooks and storyboards and answered the same questions.

3. Pre-test: Participants read four poems sequentially. For each poem, they first read only the poem, without additional materials, and answered all questions sequentially. This step assessed their initial understanding and controlled prior knowledge differences.

4. Post-test: Participants reread each poem with either its musical storyboard or textbook and answered the same questions. Throughout the study, each test question was presented individually on a separate page. After submitting an answer, participants were not allowed to revisit previous questions.

5. Experience Rating: After completing the test, participants rated their experience with both materials on the Self-Assessment Manikin, a widely used measurement assessing engagement and pleasantness of users' experience [Hnatyshyn *et al.*, 2024; Robinson and Clore, 2002].

Two metrics were computed to measure participants' performance: (1) **Accuracy:** the percentage of correct answers in the test; (2) **Improvement:** the difference in accuracy between pre-tests and post-tests.

[7]http://www.moe.gov.cn/srcsite/A26/s8001/202204/W020220420582344386456.pdf; pages 58–63

[8]Grades here refer to school years in the Chinese education system, viz., 年级(nián jí).

[9]https://www.qualtrics.com

[10]https://www.govtilr.org/Skills/ILRscale2.htm

Predictor	β	SE	z	p	95% CI
Intercept	0.96	0.20	4.87	$< .001$	[0.57, 1.35]
Condition (TB)	0.01	0.09	0.12	.90	[-0.17, 0.19]
Proficiency (2)	-0.25	0.14	-1.81	.07	[-0.52, 0.02]
Proficiency (3)	-0.19	0.14	-1.35	.18	[-0.47, 0.09]
Pre-Acc.	-0.78	0.14	-5.68	$< .001$	[-1.04, -0.51]
Condition × Pre-Acc.	0.13	0.15	0.87	.38	[-0.16, 0.42]
Difficulty	-0.02	0.01	-1.56	.12	[-0.04, 0.00]
Order of Test Poems	-0.03	0.02	-1.24	.22	[-0.08, 0.02]
Random Effects Var.	0.01	0.04			

Table 2: Results of a linear mixed-effects model (LMM) [Meteyard and Davies, 2020] predicting learners' improvement in test performance. The formula is `Improvement ~ Condition× Pre-Acc + Difficulty + Proficiency + Order + (1 | Participant)`. The variables $\boldsymbol{\beta}$, $\boldsymbol{SE}$, $\boldsymbol{z}$, and $\boldsymbol{p}$ represent the estimated coefficient, standard error, z-statistics, and p-value in the results of the LMM, respectively.

4.4 Part II: Investigating the Learning Experience of Musical Storyboards and Textbooks

The second part qualitatively investigates participants' experience with musical storyboards and textbooks. Apart from their SAM ratings, participants answered free-response questions about their preferences, perceived effectiveness, and additional insights or suggestions they wished to provide. A thematic analysis [Braun *et al.*, 2019] was performed to summarise key insights from their responses.

5 Results

We first address RQ1 by quantitatively analysing participants' performance in the reading comprehension tests. Next, we investigate RQ2 by examining participants' ratings on SAM and analysing their free responses to the interview questions.

5.1 Can Musical Storyboards Retain Textbooks' Effectiveness in Improving Learners' Comprehension of Classical Chinese Poetry?

To measure the effects of textbooks and musical storyboards on poetry understanding, we compared participants' test accuracy and improvement under both conditions. Specifically, we fitted a linear mixed-effects model (LMM) [Meteyard and Davies, 2020] to predict participants' improvement in test performance (Table 2). The model included fixed effects of test condition (TB vs. SB), pre-test accuracy (Pre-Acc.), language proficiency (Proficiency), the difficulty and order of test poems, and an interaction between condition and pre-test accuracy. Participant ID was included as a random effect.

The model intercept revealed a significant overall improvement in test accuracy ($\beta = 0.96$, $z = 4.87$, $p < .001$) when SB was the reference level. This also held when TB was set to the reference level ($\beta = 0.97$, $z = 5.59$, $p < .001$), indicating that the reading comprehension performance was significantly improved by both TB (pre-test: $N = 100$, $M = .60$, $SD = 0.25$; post-test: $N = 100$, $M = .80$, $SD = 0.23$) and SB (pre-test: $N = 100$, $M = .65$, $SD = 0.20$; post-test: $N = 100$, $M = .75$, $SD = 0.12$). However, the Condition predictor (Table 2, row 2) showed that the difference in improvement between TB and SB was insignificant ($\beta = 0.01$,

$z = 0.12$, $p = .90$). These results suggest that musical storyboards can facilitate the understanding of poems and retain the effectiveness of textbooks (Figure 3(A)(B)).

Pre-test accuracy showed a strong negative association with improvement ($\beta = -0.78$, $z = -5.68$, $p < .001$), indicating that participants who initially scored higher showed relatively less improvement. Poem difficulty ($\beta = -0.02$, $z = -1.56$, $p = .12$), language proficiency levels, and the order of test poems ($\beta = -0.03$, $z = -1.24$, $p = .22$) were not significantly related to improvement. Additionally, the interaction between test condition and pre-test accuracy was not significant ($\beta = 0.13$, $z = 0.87$, $p = .38$), suggesting a consistent relationship between prior knowledge and improvement across conditions. Overall, these findings indicate that learners' test performance was similarly improved by TB and SB, with limited influence of other factors such as language proficiency, poem difficulty, and test order (all $p > .05$).

The estimated variance for participants' random intercepts was 0.01 ($SE = 0.04$), indicating minimal individual differences in improvement after accounting for the fixed effects. The assumption diagnostics of the linear mixed-effects model showed no major violations of normality (Shapiro-Wilk test, $W = .98$, $p = .10$) or heteroscedasticity of the residuals.

5.2 Do Musical Storyboards Provide a More Engaging and Pleasant Learning Experience Compared to Textbooks?

To assess participants' learning experiences with TB and SB, paired t-tests were performed on their SAM ratings. The results (Figure 3(C)) show that musical storyboards received significantly higher ratings on **pleasure** (SB: $M = 6.68$, $SD = 0.93$; TB: $M = 5.72$, $SD = 1.64$; $t(24) = 2.87$; $p < .01$) and **arousal** (SB: $M = 5.56$, $SD = 1.20$; TB: $M = 4.76$, $SD = 1.63$; $t(24) = 2.38$, $p < .05$). No significant difference was observed for **dominance** (SB: $M = 6.04$, $SD = 1.89$; TB: $M = 5.84$, $SD = 2.07$; $t(24) = 0.53$, $p = .30$). These findings suggest that participants, as language learners, perceive musical storyboards as more pleasant and engaging compared to traditional textbook-based learning.

We then performed an inductive thematic analysis [Braun *et al.*, 2019] on participants' responses (numbered P1 to P25). Two raters from the research group independently coded the responses and concurred on four themes: *poetry understanding through multimodality*, *guided and open interpretation*, *clarity of information conveyance*, and *enhanced learning experience and complementary methodologies*. The Cohen's kappa yielded $\kappa = 0.94$ on all themes ($M = .93$, $SD = 0.13$, $max = 1.00$, $min = 0.62$), indicating an almost perfect inter-rater reliability [Landis and Koch, 1977]. The following sections present findings from the thematic analysis.

Poetry Understanding Through Multimodality. Almost all participants (23/25) mentioned that their general understanding of poems was improved by musical storyboards. Their responses highlighted the role of visualisations and music in their poetry comprehension. Specifically, 20 participants said that the visuals helped them better understand the poems. For example, "*It directly illustrates the meaning of the difficult descriptions*" (P9). Some also found the background singing helpful, as P5 commented, "*I can't read the words but if narrated to me I probably can understand it.*".

Guided and Open Interpretation. More than half of the participants (14/25) mentioned that their understanding was guided by textbooks or musical storyboards. For example, "*it (traditional learning) is more direct in conveying the message*" (P15). P9 felt guided by musical storyboards: "*It provides storyboards to help people visualise descriptions that are hard to understand.*". Moreover, some participants said musical storyboards inspired their own interpretation and imagination. "*(storyboards) help to paint the image and clarify the context of what the poem is describing, while still leaving enough room for interpretation by the viewer themselves*" (P13). These responses suggest that while both approaches can guide learners' understanding, they focus on different modalities: textbooks rely on verbal explanation, whereas storyboards offer visual context without constraining open interpretation.

Clarity of Information Conveyance. 16 out of 25 participants felt that musical storyboards effectively communicated the meaning of poems. P20 commented, "*as a beginner it (storyboard) was a lot clearer and easier to grasp the essence of the poem*". Nine participants supported the clarity of textbooks. For example, "*... words translate better to me than pictures and music in the context of better understanding the poems*" (P12). Three participants expressed that both approaches were effective in conveying the meanings of poems. For instance, "*Storyboard-based learning can be as effective as the traditional learning for people who find it hard to grasp harder concepts with complex vocabulary.*" (P23). These insights suggest that musical storyboards can clearly convey poetic meaning, while participants' preferences for different modalities during learning reflect their individual learning styles. Beginners or visual learners lean towards visual elements that simplify textual descriptions, whereas others who enjoy reading more may prefer textual explanations.

Enhanced Learning Experience and Complementary Methodologies. Regarding learning experience, some participants noted that compared to musical storyboards, textbooks tended to be boring, confusing, and distracting. For instance, "*I find it much more engaging with the visuals and the audio in storyboards. When I am faced with a wall of text, I often get lost and forgot what I have read earlier*" (P17). Even some participants who preferred textbook-based learning acknowledged that "*it (textbook) was boring*" (P3). These comments reveal the limitations of traditional textbook-based learning and underscore the greater pleasure and engagement of storyboard-based poetry learning. Furthermore, some participants suggested that combining textbooks and musical storyboards could be ideal. For example, "*Ideally, a blend of both would be good.*" (P19). These comments indicate that a complementary approach of both textbooks and musical storyboards can accommodate diverse learning preferences and further enhance the overall learning experience.

5.3 Discussion: The Role of Different Modalities

From the responses in Part II of our user study, we found that multimodal materials enhanced both engagement and pleas-

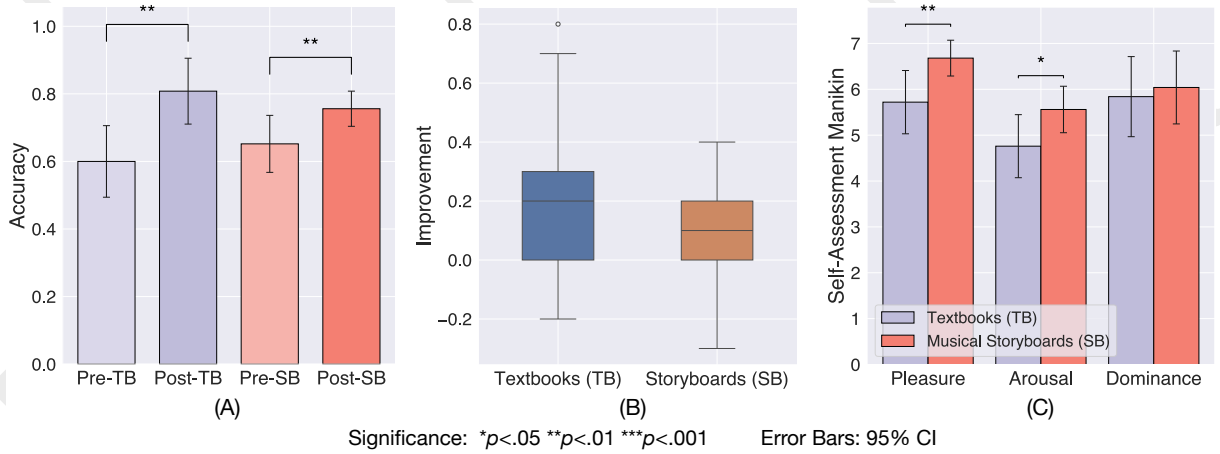

Figure 3: (A) Accuracy of participants in pre-tests and post-tests. (B) Improvement in accuracy from pre-test to post-test. (C) Participant ratings of textbook-based learning and musical storyboard-based learning on the Self-Assessment Manikin (SAM) scale.

antness in effective poetry learning. To further investigate the roles of individual modalities, we conducted a survey with 11 native Chinese speakers (5 men, 6 women; aged 25–57, $M = 27.7$, $SD = 9.4$; labelled A1–A11), all of whom had studied classical Chinese poetry extensively through formal curricula. Participants reviewed learning materials for two poems presented in five single modalities: (1) textbooks, (2) static images, (3) animation, (4) singing (audio), and (5) reading aloud (audio). They were free to review them at their own pace. Participants then commented on what they liked, disliked, and felt could be improved in each modality, comparing their effectiveness in supporting poetry understanding. Key insights from their feedback are summarised below.

Textbook Reading Was Considered the Most Effective Modality. Most participants (8/11) identified textbook reading as the most effective and widely used method, highlighting its accessibility (A5), comprehensiveness (A1), information density (A4, A6, A7), and depth (A6). For example, "*It provides the most well-rounded, comprehensive information*" (A1). However, the experience was often considered "boring" and "daunting for beginners" (A5). As A3 noted, "*If I weren't familiar with Chinese or interested in Chinese poetry, it would scare me away*". These comments suggest that, while offering in-depth explanations, textbooks can be less engaging and deterring for beginners.

Visual Materials With Motion Enhance the Engagement of Poetry Understanding. Six participants expressed a preference for animations, describing them as "intuitive" (A9) and "engaging" (A3, A4), with A7 noting, "*Even without reading the poem, I would know the vibe of the poem at first glance*". Seven participants felt that animations were more dynamic than static images, enhancing their overall engagement. As A10 explained, "*Animations tend to be more dynamic and vivid*". This shows that motion in visual materials can significantly increase engagement with poetic content.

Singing Enhances Pleasantness, While Reading Aloud Preserves Tonal Accuracy. Several participants noted that singing made the learning experience more pleasant, with A6 stating, "*I love music. It enhances the pleasantness of listening to poetry*". Some mentioned that singing motivated their learning, with A5 adding, "*It makes me want to know more about the poetry*". However, A3 and A7 expressed concerns that singing could sacrifice some tonal accuracy, as it prioritised musicality over the accurate tonal information in Chinese. In contrast, audio recordings of poetry reading can preserve the correct tones, but were often considered "boring" (A9) and "plain" (A8). This suggests that while both speech and music convey the poem's content, a balance of both–such as offering both reading and singing–can maintain tonal accuracy while enhancing the pleasantness of the experience.

These findings reveal a trade-off between comprehensiveness and experience in multimodal learning. Learners' preferences are shaped by their prior knowledge, learning styles and objectives. While textbooks provide rich content, they may deter beginners due to their information density and cognitive load. In contrast, multimodal materials, such as music videos, can enhance the learning experience but may lack sufficient depth for advanced learners. Therefore, it is essential that educators with access to AI tools like LivePoem carefully select materials that align with learners' profiles and set personalised instructional goals.

6 Conclusion

This paper explores the possibility of applying AI-generated audiovisual media to Chinese language learning. We specifically focus on AI-generated musical storyboards and their effects on classical Chinese poetry education. To this end, we propose and implement a new generative AI system, LivePoem, which automates high-quality musical storyboard generation. Through a human-subjects study with Chinese language learners, we demonstrate that musical storyboards significantly improve the pleasure and engagement of classical Chinese poetry learning, while retaining the learning outcomes of textbooks. Based on these findings, we recommend integrating both traditional textbooks and multimedia materials more frequently in Chinese poetry teaching to enhance learner engagement and effectiveness.

Ethical Statement

This work involved human subjects in its research. All ethical and experimental procedures have been approved by the Departmental Ethics Review Committee (DERC), National University of Singapore.

Acknowledgments

We thank all reviewers for their input. This work is funded by a research grant MOE-MOESOL2021-0005 from the Ministry of Education of Singapore.

References

[Birrell, 2022] Anne Birrell. *Popular songs and ballads of Han China*. Routledge, 2022.

[Braun *et al.*, 2019] Virginia Braun, Victoria Clarke, Nikki Hayfield, and Gareth Terry. Thematic analysis. In Pranee Liamputtong, editor, *Handbook of Research Methods in Health Social Sciences*, pages 843–860. Springer Singapore, Singapore, 2019.

[Brown, 2011] Dale Brown. What aspects of vocabulary knowledge do textbooks give attention to? *Language Teaching Research*, 15(1):83–97, 2011.

[Clark and Jaini, 2024] Kevin Clark and Priyank Jaini. Text-to-image diffusion models are zero-shot classifiers. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2024. Curran Associates Inc.

[Davies *et al.*, 2014] Matthew E. P. Davies, Philippe Hamel, Kazuyoshi Yoshii, and Masataka Goto. Automashupper: Automatic creation of multi-song music mashups. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(12):1726–1737, 2014.

[Duan *et al.*, 2023] Wei Duan, Yi Yu, Xulong Zhang, Suhua Tang, Wei Li, and Keizo Oyama. Melody generation from lyrics with local interpretability. *ACM Trans. Multimedia Comput. Commun. Appl.*, 19(3), feb 2023.

[Feng *et al.*, 2013] Shi Feng, Sidney D'Mello, and Arthur C Graesser. Mind wandering while reading easy and difficult texts. *Psychonomic bulletin & review*, 20:586–592, 2013.

[Guthrie and Davis, 2003] John T. Guthrie and Marcia H. Davis. Motivating struggling readers in middle school through an engagement model of classroom practice. *Reading & Writing Quarterly*, 19(1):59–85, 2003.

[Hnatyshyn *et al.*, 2024] Rostyslav Hnatyshyn, Jiayi Hong, Ross Maciejewski, Christopher Norby, and Carlo C. Maley. Capturing cancer as music: Cancer mechanisms expressed through musification. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery.

[Hu, 2023] Zhe Hu. Analysis of the historical origin of ancient chinese poetry and art songs. *Journal of Innovation and Development*, 3(1):76–78, May 2023.

[Huang *et al.*, 2023] Hanzhuo Huang, Yufan Feng, Cheng Shi, Lan Xu, Jingyi Yu, and Sibei Yang. Free-Bloom: Zero-Shot Text-to-Video Generator with LLM Director and LDM Animator. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 26135–26158. Curran Associates, Inc., 2023.

[Hwang *et al.*, 2024] Gwo-Jen Hwang, Masoud Rahimi, and Jalil Fathi. Enhancing EFL learners' speaking skills, foreign language enjoyment, and language-specific grit utilising the affordances of a MALL app: A microgenetic perspective. *Computers Education*, 214:105015, 2024.

[Ju *et al.*, 2022] Zeqian Ju, Peiling Lu, Xu Tan, Rui Wang, Chen Zhang, Songruoyao Wu, Kejun Zhang, Xiang-Yang Li, Tao Qin, and Tie-Yan Liu. TeleMelody: Lyric-to-melody generation with a template-based two-stage method. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5426–5437, Abu Dhabi, United Arab Emirates, December 2022.

[Khasawneh, 2023] Mohamad Ahmad Saleem Khasawneh. Development of audio-visual media of language learning for children with autism. *Journal of Southwest Jiaotong University*, 58(2), 2023.

[Kim *et al.*, 2024] Jihwan Kim, Junoh Kang, Jinyoung Choi, and Bohyung Han. FIFO-Diffusion: Generating Infinite Videos from Text without Training. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 89834–89868. Curran Associates, Inc., 2024.

[King, 2002] Jane King. Using DVD Feature Films in the EFL Classroom. *Computer Assisted Language Learning*, 15(5):509–523, 2002.

[KrukMariusz, 2021] KrukMariusz. Investigating the experience of boredom during reading sessions in the foreign language classroom. *Journal of Language and Education*, 7(3):89–103, Sep. 2021.

[Kuo, 1971] Ta-hsia Kuo. *A study of metre in Chinese poetry*. The University of Wisconsin-Madison, 1971.

[Landis and Koch, 1977] J. Richard Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977.

[Lee and Révész, 2018] Minjin Lee and Andrea Révész. Promoting grammatical development through textually enhanced captions: An eye-tracking study. *The Modern Language Journal*, 102(3):557–577, 2018.

[Lewis *et al.*, 2020] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online, July 2020.

[Liang and Wang, 2024] Qihao Liang and Ye Wang. Drawlody: Sketch-based melody creation with enhanced usability and interpretability. *IEEE Transactions on Multimedia*, 26:7074–7088, 2024.

[Liang *et al.*, 2024] Qihao Liang, Xichu Ma, Finale Doshi-Velez, Brian Lim, and Ye Wang. XAI-lyricist: Improving the singability of ai-generated lyrics with prosody explanations. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 7877–7885, 8 2024.

[Meteyard and Davies, 2020] Lotte Meteyard and Robert A.I. Davies. Best practice guidance for linear mixed-effects models in psychological science. *Journal of Memory and Language*, 112:104092, 2020.

[Minli, 2024] ZHAO Minli. On the relationship between singing and the development of early chinese poetic genres. *Frontiers of Literary Studies in China*, 18(4):379, 2024.

[Muñoz *et al.*, 2023] Carmen Muñoz, Geòrgia Pujadas, and Anastasiia Pattemore. Audio-visual input for learning l2 vocabulary and grammatical constructions. *Second Language Research*, 39(1):13–37, 2023.

[O'Neill, 1982] Robert O'Neill. Why use textbooks? *ELT journal*, 36(2):104–111, 1982.

[Perez and Rodgers, 2019] Maribel Montero Perez and Michael PH Rodgers. Video and language learning, 2019.

[Pratama and Hadi, 2023] Syahroni Syahrul Pratama and Muhamad Sofian Hadi. The vocabulary building audio-visual media: An innovation in vocabulary expertise. *Jurnal Studi Guru dan Pembelajaran*, 6(1):1–8, Apr. 2023.

[Pujadas and Muñoz, 2023] Geòrgia Pujadas and Carmen Muñoz. Measuring the visual in audio-visual input: The effects of imagery in vocabulary learning through tv viewing. *ITL-International Journal of Applied Linguistics*, 174(2):263–290, 2023.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[Robinson and Clore, 2002] Michael D Robinson and Gerald L Clore. Episodic and semantic knowledge in emotional self-report: evidence for two judgment processes. *Journal of personality and social psychology*, 83(1):198, 2002.

[Sheng *et al.*, 2021] Zhonghao Sheng, Kaitao Song, Xu Tan, Yi Ren, Wei Ye, Shikun Zhang, and Tao Qin. SongMASS: Automatic song writing with pre-training and alignment constraint. 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 13798–13805. AAAI Press, 2021.

[Song *et al.*, 2021] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

[Tahmina, 2023] Tania Tahmina. Students' perception of the use of YouTube in English language learning. *Journal of Languages and Language Teaching*, 11(1):151–159, 2023.

[Tilwani *et al.*, 2022] Shouket Ahmad Tilwani, Fatemeh Amini MosaAbadi, Sajad Shafiee, and Zeinab Azizi. Effects of songs on implicit vocabulary learning: Spoken-form recognition, form-meaning connection, and collocation recognition of iranian english as a foreign language learners. *Frontiers in Education*, Volume 7 - 2022, 2022.

[Wang* *et al.*, 2020] Ziyu Wang*, Ke Chen*, Junyan Jiang, Yiyi Zhang, Maoran Xu, Shuqi Dai, Guxian Bin, and Gus Xia. POP909: A Pop-song Dataset for Music Arrangement Generation. In *Proceedings of 21st International Conference on Music Information Retrieval, ISMIR*, 2020.

[Xu *et al.*, 2023] Jiale Xu, Xintao Wang, Weihao Cheng, Yan-Pei Cao, Ying Shan, Xiaohu Qie, and Shenghua Gao. Dream3D: Zero-shot text-to-3d synthesis using 3D shape prior and text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20908–20918, 2023.

[Yang and Lerch, 2020] Li-Chia Yang and Alexander Lerch. On the evaluation of generative models in music. *Neural Computing and Applications*, 32(9):4773–4784, 2020.

[Yu *et al.*, 2021] Yi Yu, Abhishek Srivastava, and Simon Canales. Conditional lstm-gan for melody generation from lyrics. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, 17(1):1–20, 2021.

[Zhang *et al.*, 2023] Zhenghua Zhang, Hang Zhang, Werner Sommer, Xiaohong Yang, Zhen Wei, and Weijun Li. Musical training alters neural processing of tones and vowels in classic chinese poems. *Brain and Cognition*, 166:105952, 2023.

[Zhao and Li, 2024a] Jiejing Zhao and Yan Li. Current Situation and Countermeasures of Teaching Chinese Ancient Poetry. *International Journal of Social Sciences and Public Administration*, 4(1):429–437, Aug. 2024.

[Zhao and Li, 2024b] Jiejing Zhao and Yan Li. Study on the Current Situation of Poetry Education in China's Compulsory Education Stage. *International Journal of Education and Humanities*, 15(3):321–323, Aug. 2024.