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 reading texts can lead to disengagement and an unpleasant learning 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 for poetry understanding. We used a pretrained diffusion model for storyboard generation and trained a prosody-based poem-to-melody generator with a Transformer model. Through a within-subjects study with 25 non-native Chinese language learners, we compared the learning outcomes from textbook reading and musical storyboards using standardised reading comprehension tests. Participants' learning experience was measured by Self-Assessment Manikin (SAM) and thematically analysed based on their open-ended feedback. Experimental results show that musical storyboards retained the learning outcomes of textbooks, while more effectively engaged participants and created a more pleasant learning experience[1].

1 Introduction

Textbook reading has long been the dominant method for learning classical poetry in Chinese-speaking communities [Pan and Chen, 2020]. It typically relies on plain-language interpretations to help learners understand archaic vocabulary and poetic techniques that are less familiar in modern Mandarin [Lee and sum Wong, 2012]. While textbooks effectively explain the meaning of classical poetry, research suggests that extensive text reading may lead to boredom [Pawlak *et al.*, 2020; Li, 2022], disengagement [Guthrie and Davis, 2003], and negative learning outcomes over time [Feng *et al.*, 2013].

To improve the experience of poetry learning, researchers in learning sciences have explored alternative or auxiliary multimodal methods to textbook reading, such as viewing static images [Pujadas and Muñoz, 2023] 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 [Chung *et al.*, 2023; Bajrami and Ismaili, 2016], vocabulary acquisition [Muñoz *et al.*, 2023; Pratama and Hadi, 2023] and imagery comprehension [Pujadas and Muñoz, 2023; Yuzela *et al.*, 2023]. Moreover, music serves as a potential auxiliary method for poetry learning, as Zhang et al. uncover that musical training relates to the neural processing of tones and vowels in classical Chinese poetry [Zhang *et al.*, 2023], underscoring the benefits of music to language education.

Despite these potential benefits, creating audiovisual media could be costly, time-consuming, and require specialised expertise for human workers. To address this need, 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 improve the experience of classical poetry learning by providing an engaging type of media for education.

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

We validated the system using standard computational metrics from image and music generation studies, demonstrating its ability to produce high-quality content. However, recognising that technical measures alone do not fully capture the

[1]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 a smoother animation, with the melody synthesised as singing voice.

human learning outcomes and experience, we further evaluated the system through a within-subjects study with 25 non-native Chinese language learners. The study assessed the preliminary learning outcomes, engagement, and satisfaction of learners while collecting their opinions on the system compared to textbook-based learning. The results show that musical storyboards retain learners' test performance while making the learning experience substantially more engaging and pleasant. We also analyse feedback from participants that exposes the benefits and drawbacks of this learning approach, specifically for linguistic education materials.

In summary, our work contributes the following:

(1) LivePoem framework for generating musical storyboards 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

Audiovisual Media For Language Learning

From lectures and standard language tests to autodidactism, textbook reading has long played a central role in various language learning scenarios. However, recent research has identified some limitations of textbook-based learning [Fletcher and Tobias, 2005], including its overemphasis on the form and accuracy of linguistic knowledge, and the lack of a realistic and meaningful context for language learners [King, 2002]. Besides, textbook-centric classes can lead to over-involvement of teachers and under-involvement of students [O'Neill, 1982], resulting in boredom [Li, 2022; Pawlak *et al.*, 2020], disengagement [Guthrie and Davis, 2003] and negative learning outcomes over time [Feng *et al.*, 2013]. To address these issues, multimodal learning materials, especially audiovisual media [Tahmina, 2023], have become increasingly popular to complement or improve textbook-based learning. For example, Pujadas and Muñoz expose language learners to TV series with words and their images co-occurring, which have a positive impact on vocabulary expansion [Pujadas and Muñoz, 2023]. Lee and Révész find that on-screen texts (e.g.,video subtitles) can marginally improve the grammatical ability of language learners [LEE and RÉVÉSZ, 2018]. Recently, Zhang et al. observe the benefits of musical training to language learning, in that music audio input facilitates speech tone and vowel processing abilities of language learners [Zhang *et al.*, 2023]. Hnatyshyn et al. use melody sequences to metaphorically represent DNA, showing that this *musification* improves the engagement of learning cancer-related knowledge, compared to text materials [Hnatyshyn *et al.*, 2024]. These studies highlight the benefits of audiovisual materials in various learning scenarios.

Cutting-Edge Technical Underpinnings for Musical Storyboard Generation

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

Text-based Video Generation aims to create 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.*, 2020; 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 pretrained 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 correlated as they both have a rhythmic nature. Early studies have shown that poems were originally created for singing [Pohl, 1992] in ancient times. Generating melodies for poetry can be viewed as a subtask of lyrics-to-melody generation. Early methods mainly train end-to-end models (e.g., LSTM-GAN [Yu *et al.*, 2021]) on paired melody-lyrics datasets, though the dataset size is insufficient to make the models converge well. Some scholars solve this data shortage problem by pre-training transformer models on unpaired datasets [Sheng *et al.*, 2021a], which spurs a wave of using attention-based methods for melody generation [Gurunath Reddy *et al.*, 2022]. 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. highlighted the importance of prosody in singability, which measures the compatibility between melodies and lyrics in 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

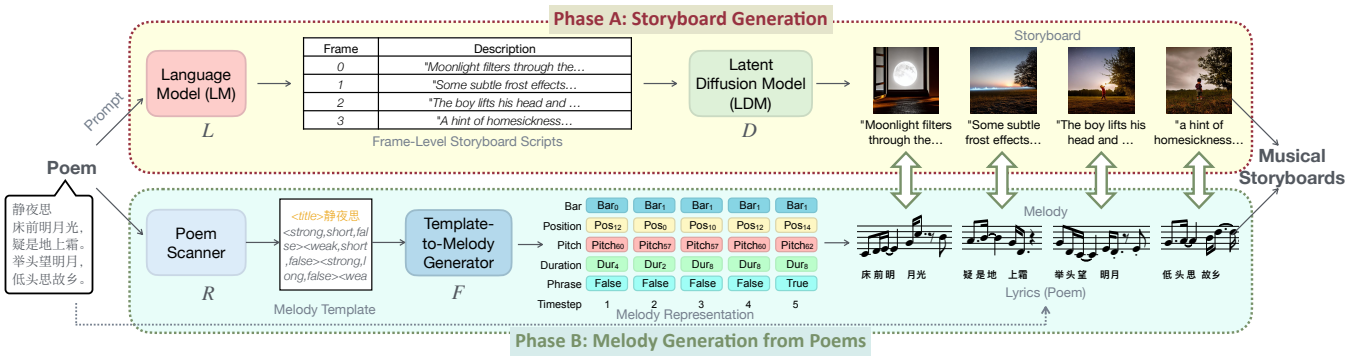

Figure 2: The two-phase musical storyboard generation framework of LivePoem. Phase A uses a language model to expand a poem into a script, which describes the connotative poetic content in plain language. These scripts then prompt a latent diffusion model to generate visual frames. Phase B extracts the 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 includes a sequence of images visually depicting the poem's content, paired with the singing of the poem lines. We build a two-phase framework (Figure 2), including (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 the 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 detailed meaning of classical poetry. Next, a latent diffusion model [Huang *et al.*, 2023] uses this script as prompts to generate a sequence of temporally coherent images as the storyboard.

Frame-Level Script Generation

Given a poem $X = \{x_1, x_2, ..., x_n\}$ with n lines, a language model LM is prompted to automatically generate a script $Y = \{y_1, y_2, ..., y_n\}$, where each y_n describes the poetic scenes of x_n in plain language. x_n in X represents the n-th line in the poem X; To obtain Y, we used a system prompt X_0 to warm up the language model $\mathcal{L}$, clarifying its task scope and fully utilising its capability to describe scenes.

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

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

Script-to-Storyboard Generation

Using the script Y, a latent diffusion model $\mathcal{D}$ generates a sequential storyboard $S = \{s_1, s_2, \ldots, s_t\}$, where s_t represents the t-th frame in S. To ensure semantic coherence across frames, we apply the sampling and attention mechanisms in [Huang *et al.*, 2023] to the diffusion model D. This phase can be formulated as:

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

3.3 Phase B: Poem-to-Melody Generation

The poem-to-melody generation phase creates a melody $M = \{m_1, m_2, \ldots, m_n\}$ with n musical phrases, aligned with the n lines of the input poem $X = \{x_1, x_2, \ldots, x_n\}$, respectively. The melody is generated based on the prosody of classical Chinese poetry, because the prosodic alignment between melodies and poems ensures that the melodies are well-suited for singing poetry. [Liang *et al.*, 2024]. This phase (bottom of Figure 2) includes: 1) a poetry scanner R, which scans[2] a poem X to extract its prosody structure and create a melody template. 2) a prosody-to-melody generator 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 poem composition. The prosody of a Chinese poem is set by counting and grouping the syllables (characters) along with pauses within the line (caesuras) and a long pause at the end of the line [Birrell, 2022]. For instance, a five-character rhythmic poem (one of the most important forms in Chinese poetry) consists of five Chinese characters per line. Each line can be segmented as either [2 syllables|2 syllables|1 syllable] or [2 syllables|1 syllable|2 syllables], where the first syllable of each segment tends to be emphasised; the last syllable of each line tends to be longer [Kuo, 1971]. Based on this concept, we designed a poem scanner that generates a melody template from a poem. The scanner analyses the structure of poetry, counts and groups the syllables, and annotates each syllable with `strong`/`weak` and `short`/`long` symbols. A binary symbol `LC` is used to mark if a character is the last one in a line. For example, in 静夜思 (Jìng Yè Sī), the first line is segmented as: 床前|明月|光 (Chuáng qián|míng yuè|guāng), with the underlined characters stressed. Besides, 光(guāng) is also the last syllable of this line and can thus be represented as (i.e., `<strong,long,true>`). When input to the model, the embeddings of all three symbols are concatenated and lin-

[2]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.

early projected to the embedding dimension of F.

Template-to-Melody Generation

F generates a sequence of melody notes from the prosody pattern in the template. To represent melody, we used five attributes to symbolically represent a musical note (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 binary phrase boundary indicator $\texttt{Phrase}_\texttt{h}$ notating if $\texttt{n}$ is the last note in a melody phrase. This strategy minimises the sequence length of the music representation and reduces training costs, ensuring musicality while preserving the structure of the melodies. We trained the model on POP909 dataset [Wang* *et al.*, 2020] after transposing all songs to C Major. The objective function was a summation of the cross-entropy losses of all attributes:

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

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

To ensure alignment between melodies and poems, we performed a sampling strategy on bar and position symbols during inference: A penalty of $r = -10^8$ is added to the logits of bar and position symbols that (1) have been previously sampled or (2) contradict the prosody pattern of poetry. This strategy ensures the monophonicity 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 the 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 followed [Huang *et al.*, 2023] to interpolate the generated storyboard frames, thereby creating smoother animations, and used ACE Studio[3] to synthesise the generated melody into a singing voice.

3.4 System Validation

We quantitatively compared the LivePoem system with several state-of-the-art image and melody generators on a poetry corpus (with 137 poems) published by the Ministry of Education in China[4]. 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 text, were 0.30 for LivePoem and 0.28 for FIFO. In the poem-to-melody generation phase, we selected SongMASS [Sheng *et al.*, 2021b] and TeleMelody [Ju *et al.*, 2022] as baselines, both specialised in lyrics-to-melody generation. 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 [5] 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 competitive performance of LivePoem. For all metrics, a higher score indicates better performance.

Table 1: System validation results on singability and music quality

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

4 Experiment

The experiment investigates 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 pleasant and engaging learning experience compared to textbooks?

4.1 Study Overview

We conducted a two-part study to address the RQs. The first part used standardised reading comprehension tests to measure the effects of musical storyboards and textbooks on participants' poem understanding (RQ1). The second used 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) Having no hard of hearing or limited vision. Participants were aged 18–35 (14 men, 10 women, 1 undisclosed). They self-reported their Chinese proficiency using the Inter-agency Language Round-table (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

Part one is a within-subjects study under two conditions:

Textbook (TB): Participants used textbooks as reference materials to answer reading comprehension questions. The books included a biography of author, 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.

To select test poems, we referred to a corpus published by the Ministry of Education (MoE) of China[7], which lists all

[3]https://acestudio.ai

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

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

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

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

compulsory poems in the national curriculum. The corpus was categorised into four difficulty levels: 1st–2nd, 3rd–4th, 5th–6th, and 7th–9th grades. 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 one correct answer, three incorrect answers, and an "I don't know" choice. This question set took 30–60 minutes to complete, aligning with the format of MoE's standard poetry tests. To control order effects, both test conditions and poem samples were computationally randomised [Davies *et al.*, 2014]. Specifically, we first randomly selected four poems, one from each of the four levels. Two poems were randomly selected to be 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[8] 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[9].

2. Task Familiarisation: Participants completed a familiarisation session simulating 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. Once submitting an answer, they could not 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.

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

The second part qualitatively investigated participants' experience with musical storyboards and textbooks. In addition to 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.

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

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

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 fully assess the effectiveness of textbooks and musical storyboards in improving poetry understanding, we compared the test accuracy and improvement under both test conditions (TB vs. SB). Specifically, we applied a linear mixed-effects model [Meteyard and Davies, 2020], with improvement as the dependent variable. The fixed effects included test condition, pre-test accuracy, language proficiency, music proficiency, poem difficulty, and poem order. Participant number was included as a random effect. Results are in Table 2.

Table 2: Results of the linear mixed-effects model (LMM) [Meteyard and Davies, 2020] predicting the improvement in test accuracy. $\boldsymbol{\beta}$, $\boldsymbol{SE}$, $\boldsymbol{z}$, and $\boldsymbol{p}$ represent the estimated coefficient, standard error, z-statistics, and p-value in the results of the LMM, respectively. (SB: storyboard; TB: textbook)

Predictor	$\boldsymbol{\beta}$	$\boldsymbol{SE}$	$\boldsymbol{z}$	$\boldsymbol{p}$
Intercept	0.96	0.20	4.87	$< .001$
Condition (TB vs. SB)	0.01	0.09	0.12	.90
Language Proficiency (2)	-0.25	0.14	-1.81	.07
Language Proficiency (3)	-0.19	0.14	-1.35	.18
Pre-test Accuracy	-0.78	0.14	-5.68	$< .001$
Condition × Pre-test Accuracy	0.13	0.15	0.87	.38
Difficulty	-0.02	0.01	-1.56	.12
Order of Test Poems	-0.03	0.02	-1.24	.22
Intercept Variance	0.01	0.04		

The model intercept revealed a significant improvement in accuracy ($\beta = 0.96$, $z = 4.87$, $p < .001$) at the reference condition (SB). This also holds when TB was set to the reference condition ($\beta = 0.97$, $z = 5.59$, $p < .001$), 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: $N = 100$, $M = .75$, $SD = 0.12$). The difference in the improvement between TB and SB was insignificant ($\beta = 0.01$, $z = 0.12$, $p = .90$). These results demonstrate that musical storyboards can facilitate the understanding of poems and retain the effectiveness of textbooks (Figure 3(A)(B)). Besides, the pre-test accuracy had a strong negative effect on improvement ($\beta = -0.78$, $z = -5.68$, $p < .001$), suggesting that participants who initially scored higher showed relatively less improvement. The difficulty level of the poems did not notably affect improvement ($\beta = -0.02$, $z = 1.56$, $p = .12$).

Language proficiency levels were not significantly related to improvement. The order in which the poems were presented had no significant impact on improvement ($\beta = -0.03$, $z = -1.24$, $p = .22$). The interaction effects between test condition and pre-test accuracy were insignificant ($\beta = 0.13$, $z = 0.87$, $p = .38$), suggesting that the relationship between prior knowledge and improvement was similar for both TB and SB. The estimated variance for participants' intercepts was 0.01 ($SD = 0.04$), indicating minimal individual differences in improvement.

The assumption diagnostic of the statistical model above exhibited no major violations of normality (Shapiro-Wilk test, $W = .98, p = .30$) or heteroscedasticity (Breusch-Pagan test, $\chi^2 = 12.73$, $p = 0.18$). Residuals were nearly normally distributed, and the Durbin-Watson statistic (2.38) indicated no strong autocorrelation.

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

To assess participants' impressions on the experience with TB and SB, paired t-tests were performed on the ratings on self-assessment manikin. The results 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, 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. 20 out of 25 participants said that the visuals helped them better understand the poems. For example, "*It provides storyboards to help people visualise descriptions that are hard to understand. It directly illustrates the meaning of the difficult descriptions*" (P9). Some 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 participants (14/25) mentioned that their understanding was guided by textbooks or musical storyboards. For example, "*I prefer traditional learning as it 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.*". Besides, 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). The feedback suggests that, while both approaches can guide learners in understanding poetry, storyboards could leave more room for learners' open interpretations of poems.

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 (storyboards) 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 clearly conveyed 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. However, I personally feel that the traditional learning explanations of the poems evoked more emotional connotations*" (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 may gravitate towards visual elements that simplify textual descriptions, whereas others who enjoy reading more may prefer textual explanations that directly convey the meaning of poetry.

Enhanced Learning Experience and Complementary Methodologies Regarding the experiences with textbooks and musical storyboards, some participants noted that textbooks were boring, confusing, and distracting compared to musical storyboards. 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 was boring*"(P3). These comments reveal the limitations of traditional textbook-based learning and underscore the greater pleasure and engagement brought by musical storyboards for poetry learning. Furthermore, several participants suggested that a combination of 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 Method Analysis: Role of Different Modalities

The free responses by language learners reveal that the joint use of multiple modalities play an important role in learning traditional poem. To explore the specific contributions of these modalities, we conducted a survey with 11 native Chinese speakers (5 men, 6 women; numbered A1-A11; aged 25–57, $M = 27.7, SD = 9.4$), who have studied classical Chinese poetry extensively through formal curriculum standards. They reviewed with two poems and materials in vari-

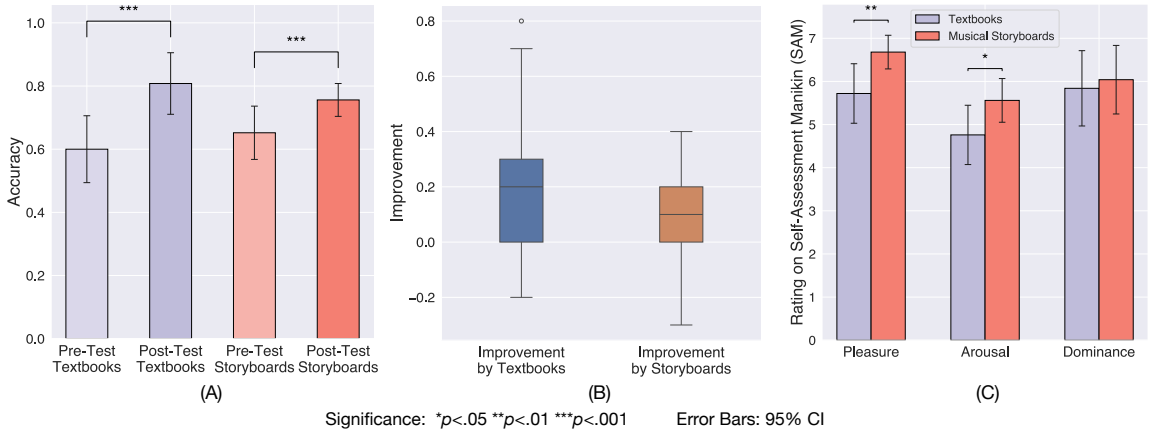

Figure 3: (A) Accuracy of participants in pre-tests and post-tests. (B) Improvement in accuracy from pre-test to post-test (C) Participant ratings of textbook-based learning and musical storyboard-based learning on the Self-Assessment Manikin (SAM) scale.

ous modalities, including: (1) textbooks, (2) static images, (3) animation, (4) singing (audio), and (5) reading aloud (audio). They were given the freedom to review the materials at their own pace. Participants provided feedback on each modality, commenting on aspects they liked, disliked, and felt could be improved. They also compared the modalities based on their effectiveness in enhancing 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, its experience was often considered "boring" and "daunting for beginners" (A5). For example, "*If I weren't familiar with Chinese or interested in Chinese poetry, it would scare me away*" (A3). The feedback indicates that while textbooks offer thorough explanations, the learning experience they provide may less engaging and unpleasant.

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 pointed out 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 visuals significantly improves the engagement with poetry content.

Singing enhances pleasantness, while reading aloud preserves tonal information. Several participants highlighted that the pleasantness of learning was fostered by singing voices, with A6 stating, "*I love music. It enhances the pleasantness of listening to poetry*". Others noted that singing motivates their learning, with A5 adding, "*It makes me want to know more about the poetry*". However, A3 and A7 raised that singing could sacrifice some tonal accuracy, as it might prioritise musicality over the precise tonal information. In contrast, audio-only reading can preserve the correct tones, but was 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 overall pleasantness of the experience.

These insights reveal a trade-off between comprehensiveness and experience regarding multimodal learning materials. Learners' impressions on different learning methods can be affected by their prior knowledge, preferences, and learning objectives. For example, while textbooks offer comprehensive information, they may undermine learners' motivation due to the lack of engagement for beginners. In contrast, multimodal materials like music videos can enhance engagement but may not always provide enough depth for advanced learners. Therefore, it is essential that educators with access to AI tools like LivePoem carefully select materials that align with learners' profile and their instructional goals.

6 Conclusion

This paper explores the effectiveness of AI-generated audio-visual media in Chinese language learning. We specifically focus on AI-generated musical storyboards and their effects on classical Chinese poetry learning. To this end, we propose and implement a new generative AI system, LivePoem, which automates 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

[Bajrami and Ismaili, 2016] Lumturie Bajrami and Merita Ismaili. The role of video materials in efl classrooms. *Procedia - Social and Behavioral Sciences*, 232:502–506, 2016. International Conference on Teaching and Learning English as an Additional Language, GlobELT 2016, 14-17 April 2016, Antalya, Turkey.

[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*, pages 843–860. Springer Singapore, Singapore, 2019.

[Chung *et al.*, 2023] Yu-Jung Chung, Chen-Wei Hsu, Meng-Hsun Chan, and Fu-Yin Cherng. Enhancing esl learners' experience and performance through gradual adjustment of video speed during extensive viewing. In *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems*, CHI EA '23, New York, NY, USA, 2023.

[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.

[Fletcher and Tobias, 2005] J. D. Fletcher and Sigmund Tobias. *The Multimedia Principle*, page 117–134. Cambridge Handbooks in Psychology. Cambridge University Press, 2005.

[Gurunath Reddy *et al.*, 2022] M Gurunath Reddy, Zhe Zhang, Yi Yu, Florian Harscoet, Simon Canales, and Suhua Tang. Deep attention-based alignment network for melody generation from incomplete lyrics. In *2022 IEEE International Symposium on Multimedia (ISM)*, pages 236–239. IEEE, 2022.

[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.

[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.

[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. *arXiv preprint arXiv:2405.11473*, 2024.

[King, 2002] Jane King. Using dvd feature films in the efl classroom. *Computer Assisted Language Learning*, 15(5):509–523, 2002.

[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.

[Lee and sum Wong, 2012] John Sie Yuen Lee and Tak sum Wong. Glimpses of ancient china from classical chinese poems. In *International Conference on Computational Linguistics*, 2012.

[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.

[Li, 2022] Chengchen Li. Foreign language learning boredom and enjoyment: The effects of learner variables and teacher variables. *Language Teaching Research*, 0(0):13621688221090324, 2022.

[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.

[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.

[Pan and Chen, 2020] Li Pan and Ping Chen. Research on language-teaching materials-an evaluation of extensive reading textbooks. *Theory and Practice in Language Studies*, 10(12):1628–1633, 2020.

[Pawlak *et al.*, 2020] Mirosław Pawlak, Mariusz Kruk, Joanna Zawodniak, and Sławomir Pasikowski. Investigating factors responsible for boredom in english classes: The case of advanced learners. *System*, 91:102259, 2020.

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

[Pohl, 1992] Karl-Heinz Pohl. Ye xie's" on the origin of poetry""(yuan shi)". a poetic of the early qing. *T'oung Pao*, pages 1–32, 1992.

[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.*, 2021a] 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.

[Sheng *et al.*, 2021b] 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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13798–13805, May 2021.

[Song *et al.*, 2020] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.

[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.

[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.

[Yuzela *et al.*, 2023] Anita Yuzela, Agus Kristiyanto, and Slamet Riyadi. The effect of audio and audio visual imagery exercises on the level of creativity of aerobic gymnastics instructors. *International Journal of Human Movement and Sports Sciences*, 11(2):292–298, 2023.

[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.