Proceedings of the Thirty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-26)
Special Track on AI and Social Good

PoemDirector: A Multi-Agent Context-Adaptive Instructional Mode Selection Framework for Chinese Classical Poetry Video Generation

Tengteng Cheng[1], **Xiaoli Zeng**[1*], **Jialu Huang**[1], **Mingliang Hou**[1], **Zitao Liu**[1], **Xiangyu Zhao**[2] and **Weiqi Luo**[1]
[1]Guangdong Institute of Smart Education, Jinan University, Guangzhou, China
[2]City University of Hong Kong, Hong Kong, China
chengt19980421@stu2024.jnu.edu.cn, zengxiaoli@stu.jnu.edu.cn,
neixin@stu2025.jnu.edu.cn, teemohold@outlook.com,
{liuzitao,lwq}@jnu.edu.cn, xianzhao@cityu.edu.hk

Abstract

Classical poetry is central to aesthetic education and cultural inheritance in China's K-12 education, and web-based instructional videos have become a primary learning resource. However, existing production pipelines either require costly teacher-guided editing or generate videos automatically without systematic instructional design, leading to unclear logic, shallow explanations, and limited contextual adaptation. We propose PoemDirector, a multi-agent framework developed with a national K-12 educational platform partner that unifies context-adaptive instructional mode selection, hierarchical explanation, and end-to-end video generation. A Director Agent analyzes each poem, selects a pedagogically grounded instructional mode from a teacher-co-designed library, and coordinates specialized agents to produce a complete instructional video from a poem title alone. We further establish a multidimensional evaluation framework, refined by literary and education experts, to assess instructional quality and poetic presentation. Experiments show that PoemDirector substantially outperforms commercial baselines and approaches human-crafted video quality across multiple dimensions. The resources are publicly available at https://github.com/ai4ed/PoemDirector.

1 Introduction

In China's K-12 curriculum, classical poetry is crucial to students' aesthetic education and cultural inheritance [Liu *et al.*, 2023]. However, in many under-resourced schools, poetry is frequently taught by rote memorization due to teachers' insufficient training or time constraints, preventing students from building a genuine understanding [Zhou *et al.*, 2025; Zheng *et al.*, 2023; Xie and Deng, 2023; Liu *et al.*, 2020]. Instructional videos on web-based platforms have become a primary learning resource and can help reduce the urban-rural quality gap [Yanghuang and Huang, 2024]. Yet producing high-quality videos demands significant effort [Brame, 2016; Manallack and Yuriev, 2016]. Even a 5-10 minute animated micro-lesson frequently involves dozens of hours of work [Guy and McNally, 2022; Xu *et al.*, 2025]. As a result, the supply of quality poetry video resources falls well short of curriculum demand, and students in under-resourced regions remain underserved. Figure 1 shows an example of an instructional video for classical poetry.

Recent advances in generative AI have made it increasingly feasible to automatically create videos that teach classical poetry. Existing approaches fall into two categories. Semi-automated pipelines split the production process into multiple stages, each handled by AI tools, and then require manual assembly to produce the final video [Weerakoon *et al.*, 2024]. In contrast, fully automated methods can generate complete videos directly from poem titles with minimal human intervention [Xu *et al.*, 2025; Kasneci *et al.*, 2023].

However, existing approaches face significant limitations. Semi-automated methods heavily rely on manual labor, leading to high production costs, while fully automated generation often lacks systematic instructional design, resulting in unclear logic and shallow explanations. Specifically, current end-to-end frameworks generally lack well-defined instructional objectives, presenting disconnected explanations rather than an organized understanding process. Furthermore, they lack teaching intelligence with contextual awareness, often employing a single narrative style that fails to engage students. Finally, hierarchical explanations aimed at deeper understanding are missing; most systems concentrate on superficial translations, preventing students from forming a genuine aesthetic appreciation of classical poetry.

To address these issues, we propose PoemDirector, a multi-agent framework that combines context-adaptive instructional mode selection, hierarchical explanation, and end-to-end generation. Inspired by real-world creative teams, PoemDirector employs a Director Agent as the cognitive core that analyzes the poem and produces a creative blueprint centered on educational objectives. The blueprint directs three specialized agents: the Visual Designer generates visuals, the Narrator synthesizes multi-role speech, and the Video Editor integrates multimodal assets along the timeline.

This work originates from a collaboration with a national

*Corresponding author.

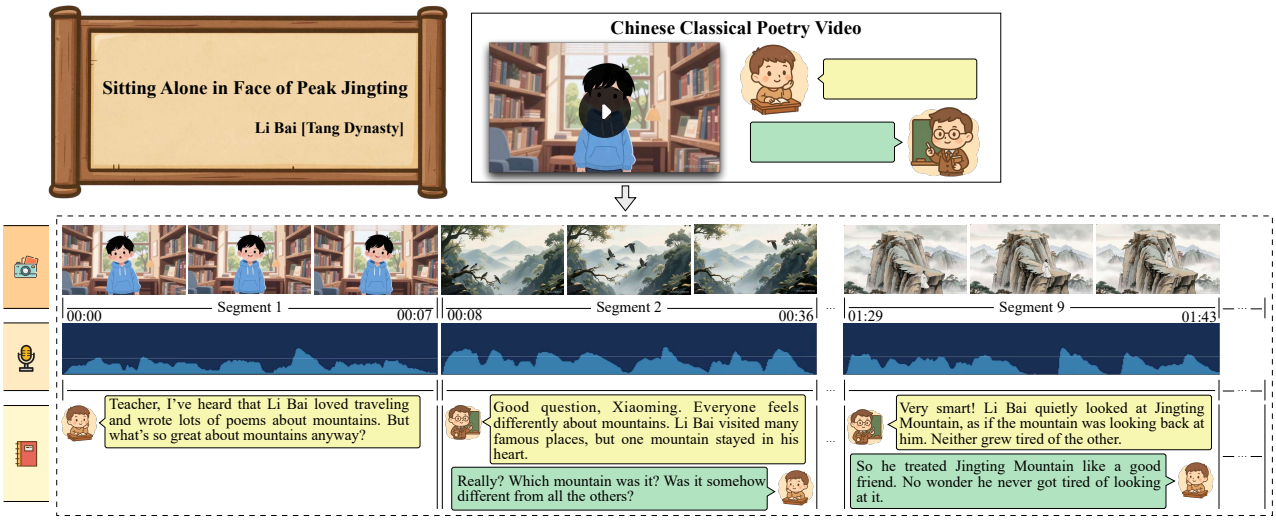

Figure 1: An example of an instructional video for Chinese classical poetry.

K-12 educational platform partner, which identified scalable poetry-video production as a practical need. Frontline K-12 language teachers contributed classroom interaction scenarios for different poem types and co-designed the instructional mode library used by the Director Agent. Literary experts iteratively refined the evaluation rubric, and education experts conducted annotation quality control throughout evaluation. By combining pedagogical expertise with automated generation, PoemDirector aims to provide low-cost, high-quality instructional resources for under-resourced educational settings. Our contributions are summarized as follows:

- We propose PoemDirector, a multi-agent framework that integrates context-adaptive instructional mode selection, hierarchical explanation, and end-to-end video generation through multi-role cooperation.
- We establish an expert-informed multidimensional evaluation framework that assesses both instructional quality and poetic presentation in classical poetry videos.
- We demonstrate through comparative studies that PoemDirector outperforms commercial baselines and approaches human-crafted video quality, and we report low-cost deployment on a real-world platform.

2 Related Work

To contextualize our approach, we review two lines of research most relevant to the automatic generation of instructional videos for classical poetry: AI-driven video generation and automated educational content creation.

2.1 AI-Driven Video Content Generation

Text-to-video generation has advanced rapidly, driven by diffusion models and flow-matching-based frameworks [Lipman *et al.*, 2023; Ho *et al.*, 2020]. Representative systems such as Lumiere and Make-A-Video demonstrate strong potential in generating short, high-fidelity video clips [Bar-Tal *et al.*, 2024; Singer *et al.*, 2022]. However, two major limitations persist for practical instructional use. First, video duration is severely constrained. State-of-the-art models typically produce clips of about 10 seconds, or roughly 240 frames, due to the computational demands of training [Xie *et al.*, 2025]. Second, content controllability remains challenging. Video generation models are highly sensitive to prompt quality, and prompt optimization strategies are rarely evaluated for their effect on final output [Wang and Yang, 2024; Cheng *et al.*, 2025]. These constraints make current models unsuitable for directly producing multi-minute instructional videos with coherent pedagogical structure.

2.2 Automated Educational Content Creation

Recent work has explored using generative AI to automate instructional material development and multimedia educational tools. Advances in multimodal generation, such as unified audio-text frameworks that jointly model speech and language, have expanded the technical foundation for building such systems [Liu *et al.*, 2026; Li *et al.*, 2021]. Chen and Wu propose an automated framework that converts Tang poetry into multimedia videos through text comprehension, image creation, and video synthesis [Chen and Wu, 2024]. Other studies show that generative AI can improve instructional design and the quality of micro-learning resources [Bolick and da Silva, 2024; Moore *et al.*, 2025]. However, existing automated frameworks face two persistent limitations. First, interaction modes are largely passive. Most systems generate narrated slides or voiceovers with no dialogic elements. Research in learning sciences has shown that dialogic and narrative approaches can enhance student engagement and learning [Chi *et al.*, 2018; Kuhn *et al.*, 2014]. Second, content generation is rarely integrated with pedagogically sound multimodal planning [Peng *et al.*, 2019; Paramythis and Loidl-Reisinger, 2004]. Large language model (LLM)-powered agentic systems offer a promising direction for coordinating multi-step generation pipelines through role-based cooperation [Hong *et al.*, 2024; Qian *et al.*, 2024]. However, no ex-

isting framework integrates pedagogically grounded instructional planning with end-to-end multimodal video generation for classical poetry education. This gap motivates the multi-agent design adopted in this work.

3 Methodology

This section presents the framework of PoemDirector, as illustrated in Figure 2. The Director Agent converts a poem title into a unified creative blueprint through contextualized, hierarchical script design (Section 3.3). The Narrator synthesizes multi-role speech (Section 3.4), the Visual Designer generates visual assets (Section 3.5), and the Video Editor integrates multimodal outputs along the timeline (Section 3.6).

3.1 Problem Statement

The goal is to automatically generate an instructional video $y \in \mathcal{Y}$ for K-12 students based solely on a poem title $t \in \mathcal{T}$. Unlike current approaches that prioritize visual aesthetics or rely on fixed narration templates, we explicitly model instructional objectives to guide explanation depth, narrative structure, and multimodal presentation. We organize this process as a collaborative endeavor among specialized agents to ensure instructional coherence and progressive explanation.

3.2 System Overview

PoemDirector decomposes this task into a set of specialized agents collaborating on a unified creative blueprint. We introduce the creative blueprint space $\mathcal{B}$, the audio asset space $\mathcal{A}$, and the visual asset space $\mathcal{C}$. The four core agents are represented as parameterized mappings:

$$\begin{aligned} B &= \mathrm{Director}(t;\theta_{\mathrm{D}}) \quad B \in \mathcal{B} \\ A &= \mathrm{Narrator}(B;\theta_{\mathrm{N}}) \quad A \in \mathcal{A} \\ C &= \mathrm{VisualDesigner}(B;\theta_{\mathrm{V}}) \quad C \in \mathcal{C} \\ y &= \mathrm{VideoEditor}(B,A,C;\theta_{\mathrm{VE}}) \quad y \in \mathcal{Y} \end{aligned}$$

where θ_{D} denotes the parameters of the LLM used by the Director Agent. θ_{N} represents the speech foundation model parameters for the Narrator Agent. θ_{V} contains the combined parameters of the large visual model and the low-rank adaptation (LoRA) components used by the Visual Designer Agent. θ_{VE} denotes the program configurations in the Video Editor Agent for timeline construction and asset arrangement.

3.3 The Director Agent

The main challenge in generating instructional poetry videos is designing a systematic blueprint that achieves clear educational objectives. Most existing methods focus on visual aesthetics but lack a central module to coordinate poem understanding, instructional planning, and video presentation. PoemDirector addresses this challenge by assigning the Director Agent as the central coordinating hub.

Information Completion and Structured Analysis

The Director maps a user-provided poem title to a uniquely identified poem and its supporting materials. Let the system's curated poetry knowledge base be:

$$\mathcal{D} = \{(t_k, p_k, meta_k, r_k)\}_{k=1}^{N}$$

where t_k is the poem title, p_k is the full poem text, $meta_k$ contains meta-information such as the author and dynasty, and r_k denotes supporting references such as annotations, translations, and cultural allusions.

Candidate Retrieval with Abstention and User Confirmation. Given an input title string t (possibly partial), the system retrieves all title-matched candidates from $\mathcal{D}$:

$$\mathcal{K}(t) = \{\, k \mid \mathrm{Match}(t, t_k) = 1,\ 1 \le k \le N \,\}$$

where $\mathrm{Match}(\cdot)$ implements title-based matching including partial-title matching after normalization. The retrieved candidates are presented to the user as a list for confirmation. If no candidate matches, the system abstains and no downstream generation is triggered:

$$\hat{k} = \begin{cases} \varnothing & |\mathcal{K}(t)| = 0 \\ \mathrm{Select}(\mathcal{K}(t)) & |\mathcal{K}(t)| \ge 1 \end{cases} \tag{1}$$

where $\mathrm{Select}(\cdot)$ denotes user selection among the candidate list. Once the user confirms a unique poem, the Director retrieves the complete poem package as grounded evidence:

$$(p, meta, r) = (p_{\hat{k}}, meta_{\hat{k}}, r_{\hat{k}}) \quad \hat{k} \neq \varnothing$$

Structured Analysis. The retrieved evidence $(p, meta, r)$ is then organized into a structured representation. The Director employs the frozen LLM to produce a poem profile:

$$\mathcal{P}(p, meta, r) = H(p, meta, r; \theta_{\mathrm{D}})$$

where $H(\cdot;\theta_{\mathrm{D}})$ is implemented by the frozen Director LLM with a prompt template. Here, $\mathcal{P}(p, meta, r)$ is a structured poem profile covering themes, imagery, emotions, key cultural allusions, and pedagogical highlights.

Instructional Modes and Contextual Scripts

The Director selects a presentation form that best matches the current poem and its context from a finite library of instructional modes $\mathcal{M} = \{m_1, m_2, \ldots, m_K\}$. Each mode m_j specifies (i) a context type such as parent–child dialogue, teacher–student instruction, or peer discussion, (ii) a role configuration and interaction style, and (iii) a script structure template together with mode-specific constraints. The mode library was co-designed with frontline K-12 language teachers, who contributed classroom interaction scenarios suitable for different poem types.

LLM-Based Mode Selection. The Director receives the structured poem profile $\mathcal{P}(p, meta, r)$, the mode library $\mathcal{M}$, and a natural-language description of each mode template. It then selects a single mode $m^* \in \mathcal{M}$ that best fits the current poem:

$$m^* = \mathcal{H}_{\text{select}}\big(\mathcal{P}(p, meta, r), \mathcal{M}; \theta_{\mathrm{D}}\big)$$

where $\mathcal{H}_{\text{select}}(\cdot)$ is implemented by the Director LLM with a prompt template to select exactly one mode.

Template Instantiation. After choosing m^*, the Director instantiates its script structure template T_{m^*} to build an explanation skeleton, represented as an ordered set of segments:

$$T_{m^*} = \big\{(\tau_\ell, f_\ell, \Delta t_\ell)\big\}_{\ell=1}^{L}$$

where τ_ℓ is the segment position in the narrative, f_ℓ is the pedagogical function label, and Δt_ℓ is the expected duration. This skeleton provides a structural interface for hierarchical content planning and time pacing.

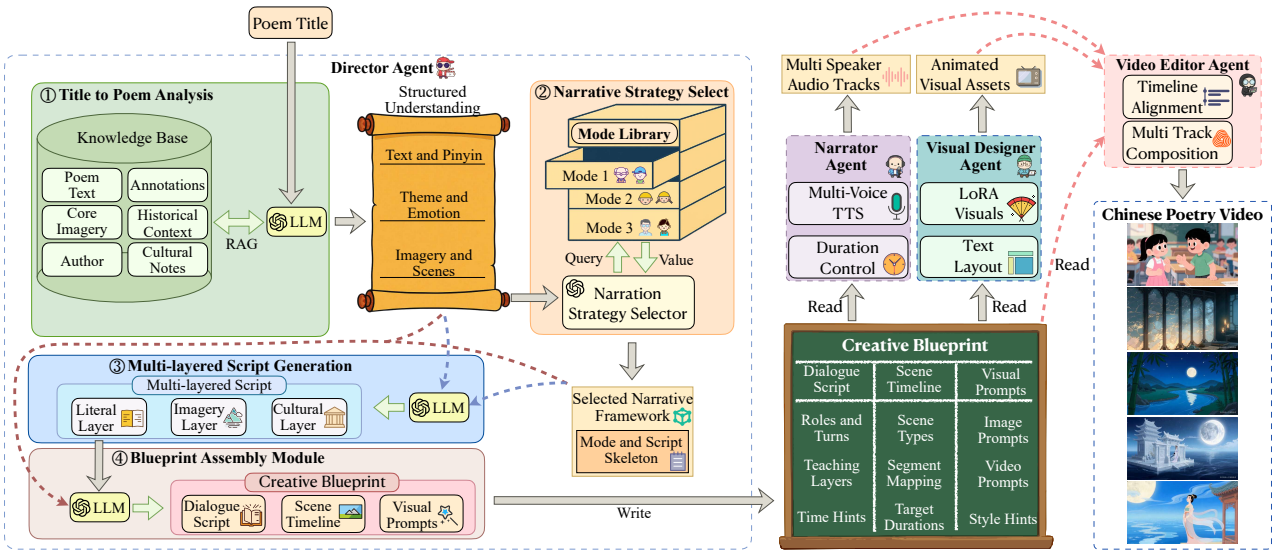

Figure 2: Overall methodological framework of PoemDirector.

Three-Level Instructional Interpretation

The Director then incorporates a three-level pedagogical interpretation structure to coordinate progressive learning. This prevents the explanation from degenerating into a shallow, linear narration. Each segment receives an interpretation-level label $h_\ell \in \mathcal{H}$, where

$$\mathcal{H} = \{\text{literal}, \text{imagery}, \text{thematic}\}$$

The label h_ℓ is assigned by the Director LLM based on the segment function f_ℓ and the poem profile $\mathcal{P}(p, meta, r)$:

$$h_\ell = \mathcal{H}_{\text{assign}}\big(f_\ell, \mathcal{P}(p, meta, r); \theta_{\text{D}}\big)$$

where $\mathcal{H}_{\text{assign}}(\cdot)$ is implemented by the frozen Director LLM with a prompt template that outputs exactly one label from $\mathcal{H}$. The script skeleton is accordingly updated as:

$$T_{m^*} = \big\{(\tau_\ell, f_\ell, \Delta t_\ell, h_\ell)\big\}_{\ell=1}^{L}$$

In this structure, the literal layer targets word-level and phrase-level understanding. The imagery layer focuses on key poetic imagery that conveys emotion. The thematic layer addresses cultural connotations and deeper themes.

Creative Blueprint Generation

The Director expands the skeleton with the hierarchical labels to create a creative blueprint:

$$B = (S, L, Q)$$

Here, S denotes the dialogue script, in which each line carries a role identifier, an interpretation-level label, and an expected duration. L defines the scene timeline that maps script segments to scene types and target durations. Q specifies image and video prompts for each scene, covering scene-related elements such as characters and locations as well as style constraints. Together, these three components ensure consistency in instructional logic, narrative structure, and visual style.

First, the Director uses the LLM to expand each skeleton segment into multiple dialogue lines:

$$S = \{(u_i, r_i, c_i, \hat{\tau}_i)\}_{i=1}^{N_S}$$

where u_i is the text of the i-th line, r_i is the speaking role identifier, $c_i \in \mathcal{H}$ is the pedagogical interpretation level, and $\hat{\tau}_i$ is the expected duration.

Next, the Director aligns script segments with scene types to obtain the timeline:

$$L = \big\{\,(\ell_j, z_j, T_j)\,\big\}_{j=1}^{\mathcal{J}}$$

where z_j is the scene type, ℓ_j is the set of covered segment indices, and T_j is the target duration interval.

Finally, the Director generates a visual prompt per scene:

$$Q = \Big\{\,\big(q_j^{\text{img}}, q_j^{\text{vid}}\big)\,\Big\}_{j=1}^{\mathcal{J}}$$

where each $\big(q_j^{\text{img}}, q_j^{\text{vid}}\big)$ specifies the scene-related elements (e.g., characters, locations) and style constraints for image and video generation. This design ensures consistency in instructional logic, narrative structure, and visual style.

Several design choices collectively reduce the risk of factual errors in the generated content. The curated poetry knowledge base contains over 1.5 million records covering poem texts, metadata, annotations, translations, and cultural allusions. The Director retrieves grounded evidence from this knowledge base before generating any explanation, and the system explicitly abstains when no valid title match is found (Eq. 1). Structured mode templates and hierarchical interpretation labels further constrain the output space, limiting the LLM's generation to pedagogically scoped content. These safeguards are consistent with the strong human-rated knowledge accuracy reported in Section 4.4.

3.4 The Narrator Agent

The Narrator Agent generates multi-character voice tracks from the creative blueprint to replace monotonous single-speaker narration. It assigns a vocal configuration $v(r_i)$ to

each role r_i (e.g., teacher, child, or elder) and synthesizes speech segments using a speech foundation model:

$$a_i = \mathrm{TTS}\big(u_i, v(r_i); \theta_{\mathrm{N}}\big)$$

This auditory differentiation helps students identify speakers and creates a more immersive learning environment.

3.5 The Visual Designer Agent

The Visual Designer Agent generates visuals through two specialized modules guided by the creative blueprint: a LoRA-based module for stylistic atmosphere and a layout-driven rendering module for textual readability. For each scene j, it uses the prompt pair $(q_j^{\mathrm{img}}, q_j^{\mathrm{vid}})$ from the blueprint to direct the static composition and dynamic motion.

LoRA-Based Classical Visuals

Generic large vision models often struggle with the specific aesthetic requirements of Chinese classical poetry. We apply standard LoRA adaptation to bias the diffusion backbone toward a consistent classical visual style, using over 100 curated images of characters and scenes in the Chinese classical style produced by a commercial text-to-image platform[1] and reviewed by human annotators. This lightweight adaptation restricts the visual output to a consistent historical style without the high cost of full-model retraining. During inference, the agent generates a static keyframe I_j, which is then expanded into a video clip clip_j by an image-to-video model:

$$I_j = \mathrm{Diffusion}\big(q_j^{\mathrm{img}};\ \theta_{\mathrm{base}}, \theta_{\mathrm{lora}}\big)$$
$$\mathrm{clip}_j = \mathrm{I2V}\big(I_j, q_j^{\mathrm{vid}};\ \theta_{\mathrm{i2v}}\big)$$

Poem Layout Rendering

To avoid character distortion and layout confusion in end-to-end generation, the poem text is decomposed into character-level display units with pinyin annotations and rendered by an explicit layout engine. This engine maps each unit to spatial coordinates and style attributes and produces a text video track with strict time alignment to the audio narration.

3.6 The Video Editor Agent

The Video Editor Agent constructs a multi-track timeline based on the scene plan L. It aligns video segments with the synthesized speech duration and integrates all tracks with a unified resolution and frame rate to produce the final video:

$$y = \mathrm{VideoEditor}\big(\{a_i\}_{i=1}^{N_S}, \{\mathrm{clip}_j\}_{j=1}^{\mathcal{J}}, L; \theta_{\mathrm{VE}}\big)$$

4 Experiments

4.1 Dataset

For evaluation, we create a dataset of 100 poems spanning 7 dynasties, 46 poets, 6 genres, and 11 thematic categories such as pastoral landscapes, frontier warfare, friendship and farewell, and seasonal imagery. Approximately 95% of the poems come from prescribed K-12 curriculum texts, and the final selection was constrained by the availability of matched handcrafted animation videos (HAV) examples. Each test item provides only the poem title as input to all methods, and no external hints are supplied at inference time to ensure a fair comparison. The full dataset, automated evaluation code, and prompts are available in our public repository[2].

4.2 Baselines

We compare our method with four baselines for Chinese classical poetry video generation. From TikTok/Douyin[3], we collect three types of videos: explanatory short videos (ESV), poetry recitation videos (PRV), and HAV. ESV and PRV were selected using objective popularity indicators such as view counts and comment volume. HAVs are meticulously crafted by professional human teams and serve as a high-quality human reference. We also include Mootion[4], a general-purpose AI video production platform that can directly generate a poetry explanation video from a poem title alone. All reported results are averaged over the full 100-poem test set.

We exclude general video generation models such as Sora and Kling as baselines because they require manual prompt engineering, high-quality reference photos, and scene-level editing, and do not satisfy the title-only, end-to-end, multi-minute instructional setting used here. The human-generated baselines (ESV, PRV, HAV) serve as static reference data without any human alteration during assessment.

4.3 Evaluation Procedures

Evaluation Metrics

We developed six evaluation metrics through collaborative expert design and iterative refinement. Three literary experts first defined the initial metrics, then scored a sample set of videos to identify scoring biases, which were addressed through iterative revisions until high consistency was achieved. The final suite comprises depth of explanation (DoE), knowledge accuracy (KA), instructional clarity (IC), aesthetic coherence (AC), poetic imagery restoration (PIR), and emotional resonance (ER). The scoring rubric for all six metrics is presented in Table 1.

Annotation

Human evaluation was conducted by three trained annotators with backgrounds in teacher education and expertise in Chinese poetry. Prior to formal annotation, a calibration session was conducted with iterative practice scoring on 10 samples until inter-annotator consistency (Cohen's kappa) exceeded 0.65. As shown in Table 2, the final agreement ranged from 0.69 to 1.00. Education experts randomly re-evaluated 10% of the annotations after each round for quality control.

4.4 Overall Performance

The overall performance is reported in Table 2. Commercial baselines show clear weaknesses: PRV scores below 1.02 in depth and clarity, while ESV and Mootion achieve only moderate performance. HAV, representing professional human-crafted videos, leads in most metrics. Our method closely approaches HAV in depth of explanation (4.07 vs

[1]https://jimeng.jianying.com

[2]https://github.com/ai4ed/PoemDirector

[3]https://www.tiktok.com/explore

[4]https://www.mootion.com/

Metric	What it measures	Scoring rubric
DoE	Depth beyond literal paraphrase	**5:** Comprehensive interpretation of meaning, imagery, background, and techniques; **4:** Fairly in-depth, covering major background and key imagery; **3:** Basic meaning and some background, but limited analysis; **2:** Limited extension beyond surface meaning; **1:** Literal translation only; no deeper interpretation
KA	Factual correctness of meaning, background, and allusions	**5:** Accurate and reliable throughout; **4:** Accurate with only minor gaps or missing details; **3:** Generally accurate with occasional minor issues; **2:** Contains multiple inaccuracies or misleading statements; **1:** Clear factual or interpretive errors
IC	Clarity, coherence, and age-appropriate explanation	**5:** Clear, coherent, and easy for students to follow; **4:** Reasonable structure, fluent expression, key points highlighted; **3:** Generally understandable but somewhat cluttered; **2:** Logic not fully clear; expression not concise enough; **1:** Disorganized and hard to follow
AC	Consistency of visual/audio style with the poem's mood	**5:** Highly unified and expressive audiovisual style; **4:** Coordinated and consistent; aligns well with the poem; **3:** Mostly coherent with minor mismatches; **2:** Noticeable inconsistencies in style or visual treatment; **1:** Inconsistent style that disrupts the experience
PIR	Completeness and correctness of key imagery	**5:** Key imagery faithfully and vividly restored; **4:** Largely complete; generally corresponds to the poem; **3:** Main imagery captured but simplified; **2:** Only some imagery shown or presented inaccurately; **1:** Key imagery missing or incorrect
ER	Ability to convey the poem's emotion and engage viewers	**5:** Strong and nuanced emotional resonance; **4:** Emotions expressed fully; likely to elicit student resonance; **3:** Basic emotional tone conveyed but lacks depth; **2:** Weak emotional presentation with limited impact; **1:** Little or no emotional impact

Table 1: Compact rubric for the six human-evaluation metrics. The full annotation handbook with detailed scoring guidelines is available in the public repository.

Methods	DoE	KA	IC	AC	PIR	ER
ESV	3.73 (κ=0.928)	3.37 (κ=1.000)	3.04 (κ=0.798)	2.80 (κ=0.851)	2.83 (κ=0.922)	3.36 (κ=0.951)
PRV	1.00 (κ=1.000)	2.21 (κ=0.961)	1.02 (κ=0.789)	2.19 (κ=0.899)	2.22 (κ=0.970)	1.94 (κ=0.840)
Mootion	3.04 (κ=0.832)	3.03 (κ=1.000)	2.79 (κ=0.933)	2.57 (κ=0.948)	2.38 (κ=0.801)	2.44 (κ=0.962)
HAV (Human)	**4.72** (κ=0.917)	**4.19** (κ=0.949)	**3.92** (κ=0.696)	**3.91** (κ=0.774)	3.90 (κ=1.000)	3.74 (κ=0.914)
PoemDirector	4.07 (κ=0.796)	4.10 (κ=0.801)	3.90 (κ=0.689)	3.49 (κ=0.881)	**3.97** (κ=0.903)	**3.83** (κ=0.756)

Table 2: Human evaluation results. The consistency is measured by Cohen's kappa coefficient (κ) shown in parentheses.

4.72), knowledge accuracy (4.10 vs 4.19), and instructional clarity (3.90 vs 3.92). Notably, PoemDirector surpasses HAV in poetic imagery restoration (3.97 vs 3.90) and emotional resonance (3.83 vs 3.74), which we attribute to the Director's structured content analysis and the Visual Designer's LoRA-based classical visuals. The largest remaining gap appears in aesthetic coherence (3.49 vs 3.91), reflecting the inherent limitations of current diffusion models in maintaining visual consistency across multi-minute video sequences. The gap in depth of explanation (4.07 vs 4.72) suggests that human experts still provide richer cultural context than the current retrieval-augmented pipeline. Despite these gaps, PoemDirector achieves this performance through fully automated generation at a marginal cost of $0.35-$0.45 per video, whereas HAV requires professional teams and substantially greater production effort.

4.5 Automated Evaluation with LLMs

To explore scalable evaluation, we develop an automated evaluator using multimodal LLMs. Video narrations are transcribed with Whisper large v3 and evaluated by Claude Sonnet 4.5[5] for DoE, KA, and IC [Radford *et al.*, 2023]. Uniformly sampled keyframes are assessed by Claude Sonnet 4.5 for AC and PIR. Complete videos are analyzed by Gemini 2.5 Flash for ER [Team *et al.*, 2024].

[5]https://www.anthropic.com/news/claude-sonnet-4-5

Methods	DoE	KA	IC	AC	PIR	ER
ESV	2.667	3.167	2.633	3.129	2.323	4.688
PRV	1.000	2.387	1.000	3.111	1.000	4.161
Mootion	2.867	2.933	2.967	3.200	2.500	4.750
HAV (Human)	**3.179**	3.321	3.036	3.032	2.226	4.774
PoemDirector	3.032	**3.452**	**3.935**	**3.935**	**2.903**	**5.000**

Table 3: Automated evaluation results.

As shown in Table 3, our method achieves competitive performance across most dimensions, particularly in KA (3.452), IC (3.935), and AC (3.935). The pipeline correctly identifies quality gaps such as PRV's deficiencies (scores of 1.000). However, LLMs exhibit compressed scoring ranges compared to human evaluators. For instance, ER scores fall between 4.161-5.000, whereas the human range is 1.94-3.83 in Table 2. These systematic differences motivate our multi-source validation study (Section 4.6).

4.6 Multi-Source Validation Study

To address the limitations of single-modality evaluation, we conduct a multi-source validation study using VideoScore and cross-source consistency analysis [He *et al.*, 2024]. As shown in Figure 3, our method achieves the highest dynamic degree score (3.41), attributed to the Visual Designer's image-to-video generation pipeline, while HAV and PRV rely on static imagery with minimal temporal variation. Figure 4

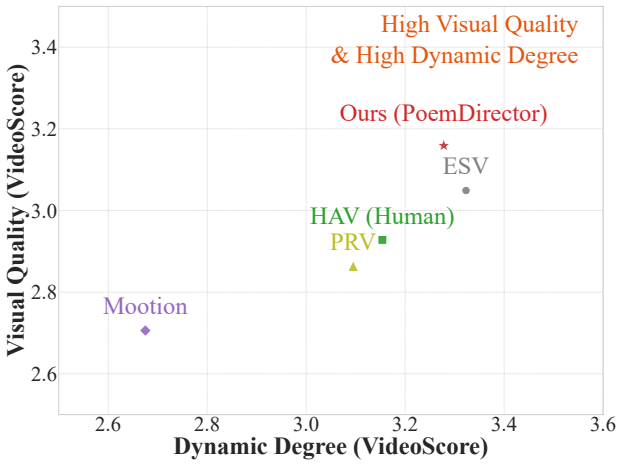

Figure 3: VideoScore visual quality-dynamics trade-off.

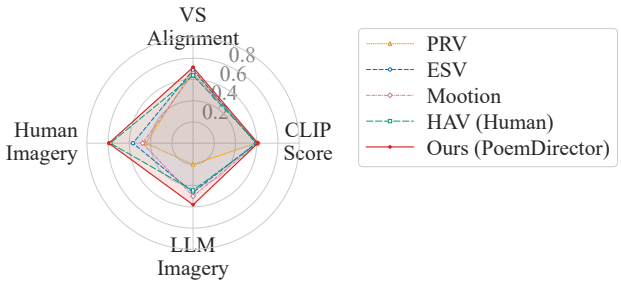

Figure 4: Cross-source consistency of poetic imagery restoration.

validates poetic imagery restoration across four independent sources: human expert annotations, LLM semantic extraction, CLIP-based vision-language alignment, and VideoScore text-video alignment. Our method ranks highest across all four sources, particularly in human annotations and CLIP scores, providing convergent evidence that the pipeline captures authentic poetic imagery.

4.7 Case Study

We conduct a qualitative case study on the Tang Dynasty poem "Farewell to Da Dong" by Shi Gao, with a visual comparison across three methods shown in Figure 5. Our method generates a video with three key strengths: (1) contextualized instruction through narrative storytelling enabled by the Director's contextual scripts, as shown at t=02s, (2) culturally authentic Tang Dynasty visual styling throughout the video, and (3) faithful poetic imagery restoration guided by the three-level instructional interpretation, including a flying goose in snowy mountains at t=85s and a traveler pointing toward distant horizons at sunset at t=140s. In contrast, HAV shows simplistic visual design with flat colors and cartoonish characters (t=275s), while Mootion relies on generic pastoral scenes (t=02s, t=45s) with text-heavy presentation.

4.8 Deployment and Collaboration

Stakeholder Collaboration. This project was initiated under the AI-for-Education program of a national K-12 educational platform partner, which identified scalable poetry-video production as a practical need. Frontline K-12 language teachers co-designed the instructional mode library by contributing classroom interaction scenarios for different poem types. Literary experts iteratively refined the evaluation rubric through pilot scoring, and education experts conducted annotation quality control throughout evaluation. The platform partner also provided deployment infrastructure and integration requirements.

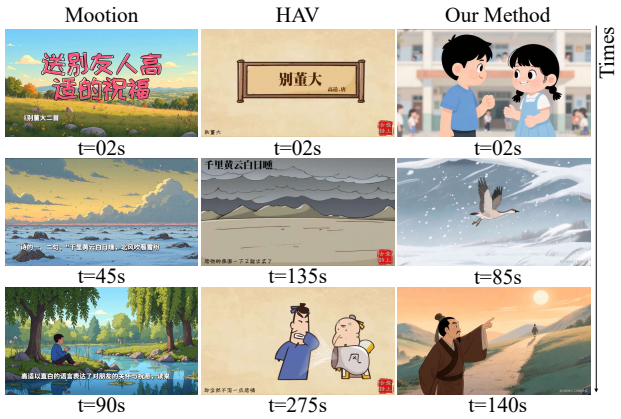

Figure 5: Visual comparison for "Farewell to Da Dong" by Shi Gao.

Deployment and Cost. We developed a web-based interface where users select a poem title and receive an automatically generated instructional video with no manual editing required. The Director Agent uses the GPT-4o API, while the Visual Designer employs locally deployed LoRA-fine-tuned FLUX and FramePack-1 models [Batifol *et al.*, 2025; Zhang *et al.*, 2025]. Only the Director incurs external API charges, and the marginal cost is approximately $0.35-$0.45 per video. The system was deployed in December 2025 on the partner platform, demonstrating practical feasibility as a low-cost educational content production tool.

Reproducibility. Representative demo videos, the dataset, evaluation scripts, the algorithm overview, and the detailed evaluation rubric are publicly available at https://github.com/ai4ed/PoemDirector.

5 Conclusion

In this paper, we introduced PoemDirector, a multi-agent framework designed to automate the generation of instructional videos for Chinese classical poetry. Systematic experiments show that PoemDirector substantially outperforms commercial baselines while approaching human-crafted video quality. The low-cost deployment on a real-world K-12 educational platform demonstrates practical feasibility for alleviating the shortage of quality poetry teaching resources in under-resourced schools.

This work has several limitations. Our evaluation relies on expert assessments rather than direct student learning gains, the framework lacks personalized adaptation, and it has only been applied to classical poetry. In future work, we aim to address these through controlled classroom experiments, cognitive diagnostic techniques for individual adaptability, and extension to other literary genres.

Acknowledgements

This work was supported in part by NSFC under Grant No. 62477025; in part by Beijing Natural Science Foundation under Grant No. L252010; in part by Key Laboratory of Smart Education of Guangdong Higher Education Institutes, Jinan University (2022LSYS003) and in part by the Guangdong University Student Science and Technology Innovation Cultivation Special Fund under Grant No. pdjh2026ak027.

References

[Bar-Tal *et al.*, 2024] Omer Bar-Tal, Hila Chefer, Omer Tov, Charles Herrmann, Roni Paiss, Shiran Zada, Ariel Ephrat, Junhwa Hur, Guanghui Liu, Amit Raj, Yuanzhen Li, Michael Rubinstein, Tomer Michaeli, Oliver Wang, Deqing Sun, Tali Dekel, and Inbar Mosseri. Lumiere: A space-time diffusion model for video generation. In *Proceedings of the 17th ACM SIGGRAPH Conference on Computer Graphics and Interactive Techniques Asia*, Japan, December 2024.

[Batifol *et al.*, 2025] Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv e-prints*, 2025.

[Bolick and da Silva, 2024] Alexis Danielle Bolick and Rafael Leonardo da Silva. Exploring artificial intelligence tools and their potential impact to instructional design workflows and organizational systems. *TechTrends*, 68(1):91–100, 2024.

[Brame, 2016] Cynthia J. Brame. Effective educational videos: Principles and guidelines for maximizing student learning from video content. *CBE-Life Sciences Education*, 15(4):1–6, 2016.

[Chen and Wu, 2024] Xu Chen and Di Wu. Automatic generation of multimedia teaching materials based on generative AI: Taking Tang poetry as an example. *IEEE Transactions on Learning Technologies*, 17:1327–1340, 2024.

[Cheng *et al.*, 2025] Jiale Cheng, Ruiliang Lyu, Xiaotao Gu, Xiao Liu, Jiazheng Xu, Yida Lu, Jiayan Teng, Zhuoyi Yang, Yuxiao Dong, Jie Tang, et al. Vpo: Aligning text-to-video generation models with prompt optimization. *arXiv preprint arXiv:2503.20491*, 2025.

[Chi *et al.*, 2018] Michelene TH Chi, Joshua Adams, Emily B Bogusch, Christiana Bruchok, Seokmin Kang, Matthew Lancaster, Roy Levy, Na Li, Katherine L McEldoon, Glenda S Stump, et al. Translating the icap theory of cognitive engagement into practice. *Cognitive science*, 42(6):1777–1832, 2018.

[Guy and McNally, 2022] Julia Guy and Michael B. McNally. Ten key factors for making educational and instructional videos. *Scholarly and Research Communication*, 13(2):1–18, 2022.

[He *et al.*, 2024] Xuan He, Dongfu Jiang, Ge Zhang, Max Ku, Achint Soni, Sherman Siu, Haonan Chen, Abhranil Chandra, Ziyan Jiang, Aaran Arulraj, et al. Videoscore: Building automatic metrics to simulate fine-grained human feedback for video generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Miami, Florida, November 2024.

[Ho *et al.*, 2020] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pages 6840–6851, Virtual, December 2020.

[Hong *et al.*, 2024] Sirui Hong, Xiawu Zheng, Jonathan Chen, Yuheng Cheng, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. Metagpt: Meta programming for a multi-agent collaborative framework. In *Proceedings of the 12th International Conference on Learning Representations*, Vienna, Austria, May 2024.

[Kasneci *et al.*, 2023] Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, Stephan Krusche, Gitta Kutyniok, Tilman Michaeli, Claudia Nerdel, Jürgen Pfeffer, Oleksandra Poquet, Michael Sailer, Albrecht Schmidt, Tina Seidel, Matthias Stadler, Jochen Weller, Jochen Kuhn, and Gjergji Kasneci. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103:102274, 2023.

[Kuhn *et al.*, 2014] Deanna Kuhn, Laura Hemberger, and Valerie Khait. Trialogues: The contexts that support argumentative thinking. In *Rethinking Knowledge*, pages 155–179. Routledge, 2014.

[Li *et al.*, 2021] Hang Li, Wenbiao Ding, Yu Kang, Tianqiao Liu, Zhongqin Wu, and Zitao Liu. Ctal: Pre-training cross-modal transformer for audio-and-language representations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, Online and Punta Cana, Dominican Republic, November 2021.

[Lipman *et al.*, 2023] Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *Proceedings of the 11th International Conference on Learning Representations*, Kigali, Rwanda, May 2023.

[Liu *et al.*, 2020] Zitao Liu, Guowei Xu, Tianqiao Liu, Weiping Fu, Yubi Qi, Wenbiao Ding, Yujia Song, Chaoyou Guo, Cong Kong, Songfan Yang, and Gale Yan Huang. Dolphin: A spoken language proficiency assessment system for elementary education. In *Proceedings of The Web Conference 2020*, Taipei, Taiwan, April 2020.

[Liu *et al.*, 2023] Dawei Liu, Ping He, and Huifen Yan. An empirical study of schema-associated mnemonic method for classical Chinese poetry in primary and secondary education based on cognitive schema migration theory. *Scientific Reports*, 13:20720, 2023.

[Liu *et al.*, 2026] Tianqiao Liu, Xueyi Li, Hao Wang, Haoxuan Li, Zhichao Chen, Weiqi Luo, and Zitao Liu. From text to talk: Audio-language model needs non-autoregressive joint training. In *Proceedings of the 14th International Conference on Learning Representations*, Rio de Janeiro, Brazil, April 2026.

[Manallack and Yuriev, 2016] David T. Manallack and Elizabeth Yuriev. Ten simple rules for developing a mooc. *PLOS Computational Biology*, 12(10):e1005061, 2016.

[Moore *et al.*, 2025] Steven Moore, Lydia Eckstein, Christine Kwon, and John Stamper. Generative AI in instructional design education: Effects on novice microlesson quality. In *Proceedings of the 27th International Conference on Artificial Intelligence in Education*, Tokyo, Japan, July 2025.

[Paramythis and Loidl-Reisinger, 2004] Alexandros Paramythis and Susanne Loidl-Reisinger. Adaptive learning environments and e-learning standards. *Electronic Journal of E-learning*, 2(1):181–194, 2004.

[Peng *et al.*, 2019] Hongchao Peng, Shanshan Ma, and Jonathan Michael Spector. Personalized adaptive learning: An emerging pedagogical approach enabled by a smart learning environment. *Smart Learning Environments*, 6(9), 2019.

[Qian *et al.*, 2024] Chen Qian, Wei Liu, Hongzhang Liu, Nuo Chen, Yufan Dang, Jiahao Li, Cheng Yang, Weize Chen, Yusheng Su, Xin Cong, Juyuan Xu, Dahai Li, Zhiyuan Liu, and Maosong Sun. Chatdev: Communicative agents for software development. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, Bangkok, Thailand, August 2024.

[Radford *et al.*, 2023] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, Honolulu, Hawaii, USA, July 2023.

[Singer *et al.*, 2022] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.

[Team *et al.*, 2024] Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

[Wang and Yang, 2024] Wenhao Wang and Yi Yang. Vidprom: A million-scale real prompt-gallery dataset for text-to-video diffusion models. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, Vancouver Convention Center, December 2024.

[Weerakoon *et al.*, 2024] Oshani Weerakoon, Ville Leppänen, and Tuomas Mäkilä. Enhancing pedagogy with generative AI: Video production from course descriptions. In *CompSysTech '24: Proceedings of the International Conference on Computer Systems and Technologies 2024*, pages 249–255, Ruse, Bulgaria, June 2024. Association for Computing Machinery.

[Xie and Deng, 2023] Heping Xie and Sue Deng. Drawing as a strategy for children to learn ancient Chinese poetry. *Acta Psychologica*, 240:104039, 2023.

[Xie *et al.*, 2025] Desai Xie, Zhan Xu, Yicong Hong, Hao Tan, Difan Liu, Feng Liu, Arie Kaufman, and Yang Zhou. Progressive autoregressive video diffusion models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 6322–6332, 2025.

[Xu *et al.*, 2025] Tao Xu, Yuan Liu, Yaru Jin, Yueyao Qu, Jie Bai, Wenlan Zhang, and Yun Zhou. From recorded to AI-generated instructional videos: A comparison of learning performance and experience. *British Journal of Educational Technology*, 56(4):1463–1487, 2025.

[Yanghuang and Huang, 2024] Yuling Yanghuang and Yong Huang. Advancing whole-book reading teaching in primary and secondary schools by regional instructional cooperation: A case study of Yangzhong city in Jiangsu province of China. *Science Insights*, 45(1):1425–1429, 2024.

[Zhang *et al.*, 2025] Lvmin Zhang, Shengqu Cai, Muyang Li, Gordon Wetzstein, and Maneesh Agrawala. Frame context packing and drift prevention in next-frame-prediction video diffusion models. In *Proceedings of the 39th International Conference on Neural Information Processing Systems*, Hilton Mexico City Reforma, December 2025.

[Zheng *et al.*, 2023] Lei Zheng, Xiang Qi, and Chongjiu Zhang. Can improvements in teacher quality reduce the cognitive gap between urban and rural students in China? *International Journal of Educational Development*, 100:102781, 2023.

[Zhou *et al.*, 2025] Yan Zhou, Chutima Vatanakhiri, and Somchart Adulvajarameta. A study on characteristics of learning interests in the teaching of chanting ancient poetry in the first grade of primary schools in Shizuishan city, China. *St. John's Journal (Humanities and Social Sciences)*, 28(42):55–73, 2025.